Gesellschaft für Informatik e.V. (GI)

publishes this series in order to make available to a broad public recent findings in informatics (i.e. computer science and information systems), to document conferences that are organized in cooperation with GI and to publish the annual GI Award dissertation.

Broken down into
- seminars
- proceedings
- dissertations
- thematics

current topics are dealt with from the vantage point of research and development, teaching and further training in theory and practice. The Editorial Committee uses an intensive review process in order to ensure high quality contributions.

The volumes are published in German or English.

Information: http://www.gi.de/service/publikationen/lni/

ISBN 978-3-88579-416-5

This book presents outstanding dissertations in informatics of the year 2011. Every year, the German Gesellschaft für Informatik (GI), the Swiss Informatics Society (SI), the Austrian Computer Society (OCG) and the German Chapter of the ACM (GChACM) jointly give an award for an excellent dissertation that represents an important advance in informatics. The award winner is chosen in a careful selection procedure from candidates proposed by Austrian, Swiss and German universities, where each university may suggest at most one dissertation per year. The series "Ausgezeichnete Informatikdissertationen" presents these outstanding dissertations to the informatics community, as well as to the public in order to support knowledge transfer from universities to industry and society. This volume is in German.

GI-Dissertationspreis 2011

12

GI-Edition

Lecture Notes in Informatics

Steffen Hölldobler et al. (Hrsg.)

Ausgezeichnete Informatikdissertationen 2011

Dissertations

Steffen Hölldobler et al. (Hrsg.)

Ausgezeichnete Informatikdissertationen 2011

Im Auftrag der GI herausgegeben durch die Mitglieder des Nominierungsausschusses

Abraham Bernstein, Universität Zürich
Steffen Hölldobler (Vorsitzender), Technische Universität Dresden
Klaus-Peter Löhr, Freie Universität Berlin
Paul Molitor, Martin-Luther-Universität Halle-Wittenberg
Gustaf Neumann, Wirtschaftsuniversität Wien
Rüdiger Reischuk, Universität zu Lübeck
Myra Spiliopoulou, Otto-von Guericke-Universität Magdeburg
Harald Störrle, Technical University of Denmark
Dorothea Wagner, Universität Karlsruhe (TH)

Gesellschaft für Informatik e.V. (GI)

Lecture Notes in Informatics (LNI) - Dissertations
Series of the Gesellschaft für Informatik (GI)

Volume D-12

ISBN 978-3-88579-416-5

printed by Köllen Druck+Verlag GmbH, Bonn

Vorwort

Die Gesellschaft für Informatik e.V. (GI) vergibt gemeinsam mit der Schweizer Informatik Gesellschaft (SI), der Österreichischen Computergesellschaft (OCG) und dem German Chapter of the ACM (GChACM) jährlich einen Preis für eine hervorragende Dissertation im Bereich der Informatik. Hierzu zählen nicht nur Arbeiten, die einen Fortschritt in der Informatik bedeuten, sondern auch Arbeiten aus dem Bereich der Anwendungen in anderen Disziplinen und Arbeiten, die die Wechselwirkungen zwischen Informatik und Gesellschaft untersuchen. Die Auswahl dieser Dissertationen stützt sich auf die von den Universitäten und Hochschulen für diesen Preis vorgeschlagenen Dissertationen. Jede dieser Hochschulen kann jedes Jahr nur eine Dissertation vorschlagen. Somit sind die im Auswahlverfahren vorgeschlagenen Kandidatinnen und Kandidaten bereits „Preisträger“ ihrer Hochschule.

Die 30 Einreichungen zum Dissertationspreis 2011 belegen die zunehmende Bedeutung und auch die Bekanntheit des Dissertationspreises. Wie jedes Jahr wurden die vorgeschlagenen Arbeiten im Rahmen eines Kolloquiums im Leibniz-Zentrum für Informatik Schloss Dagstuhl von den Nominierten vorgestellt. Für die Mitglieder des Nominierungsausschusses war das persönliche Zusammentreffen mit den Nominierten der Höhepunkt der Auswahlarbeit, und für die Nominierten hat das Kolloquium sicher eine Reihe neuer Erfahrungen und wissenschaftlicher Kontakte geboten. Das wissenschaftlich sehr hohe Niveau der Vorträge, die regen Diskussionen und die angenehme Atmosphäre in Schloss Dagstuhl wurde von allen Teilnehmerinnen und Teilnehmern des Kolloquiums sehr begrüßt.

Wie in jedem Jahr fiel es dem Nominierungsausschuss sehr schwer, eine einzige Dissertation auszuwählen, die durch den Preis besonders gewürdigt wird. Mit der Präsentation aller vorgeschlagenen Dissertationen in diesem Band wird die Ungerechtigkeit, eine aus mehreren ebenbürtigen Dissertationen hervorzuheben, etwas ausgeglichen. Dieser Band soll zudem einen Beitrag zum Wissenstransfer innerhalb der Informatik und von den Universitäten und Hochschulen in die Bereiche Technik, Wirtschaft und Gesellschaft leisten.

Die beteiligten Gesellschaften zeichnen Dr. rer. nat. Johannes Textor, der an der Universität zu Lübeck promovierte, für seine hervorragende Dissertation „Search and Learning in the Immune System: Models of Immune Surveillance and Negative Selection“ mit dem Dissertationspreis 2011 aus.

Zum Verständnis molekularbiologischer Vorgänge im menschlichen Immunsystem entwickelt Herr Textor eine analytische Modellierung als stochastische Suchprozesse sowie als Klassifikationsprobleme. Er analysiert diese rigoros und detailliert und entwickelt effiziente algorithmische Lösungsverfahren. Dabei sind seine Resultate zum Teil überraschend und widerlegen bisherige Erwartungen. Darüber hinaus zeigt Herr Textor eindrucksvoll, dass die gefundenen Ergebnisse unmittelbar in der Molekularbiologie angewendet werden können und dort zu neuen, den bisherigen Stand der Kunst erheblich verbessernden Methoden und Resultaten führen.

Mit dieser Preisverleihung würdigen die beteiligten Gesellschaften – die Gesellschaft für Informatik e.V. (GI), die Schweizer Informatik Gesellschaft (SI), die Österreichische Computergesellschaft (OCG) und das German Chapter of the ACM (GChACM) – eine herausragende wissenschaftliche Arbeit, die zeigt, wie mit Hilfe kreativer informatischer Modellierung und Analyse auch entfernte Disziplinen – in diesem Fall die Molekularbiologie – entscheidend vorangebracht werden können.

Ein besonderer Dank gilt dem Nominierungsausschuss, der sehr effizient und konstruktiv zusammengearbeitet hat. Bei Frau Julia Koppenhagen, Frau Sylvia Wünsch und Herrn Christoph Wernhard möchte ich mich für die Unterstützung bei der Entgegennahme der vorgeschlagenen Dissertationen, für die Organisation des Kolloquiums sowie für die Zusammenstellung und Anpassung der Beiträge an das Format der GI-Edition Lecture Notes in Informatik (LNI) bedanken. Für die finanzielle Unterstützung des Nominierungskolloquiums sei den beteiligten Gesellschaften gedankt. Die Gastfreundlichkeit und die hervorragende Bewirtung in Dagstuhl trugen zum Erfolg des Kolloquiums bei, wofür ich mich an dieser Stelle ebenfalls herzlich bedanke.

Steffen Hölldobler
Dresden im August 2012

Kandidaten für den GI-Dissertationspreis 2011

Dr. Martin Bader	Universität Ulm
Dr. Stephan Bode	TU Ilmenau
Dr. Ralf Dreesen	Universität Paderborn
Dr. Dominik Freydenberger	Goethe-Universität Frankfurt am Main
Dr. Oliver Friedmann	Ludwig-Maximilians-Universität
Dr. Stephan Heckmüller	Universität Hamburg
Dr. Tobias Heer	RWTH Aachen
Dr. Bernhard Kainz	Technische Universität Graz
Dr. Dominik Karch	Ruprecht-Karls-Universität Heidelberg
Dr. Peter Kiefer	Otto-Friedrich-Universität Bamberg
Dr. Marius Kloft	Technische Universität Berlin
Dr. Steffen Kopecki	Universität Stuttgart
Dr. Heiko Paulheim	Technische Universität Darmstadt
Dr. Martin Potthast	Bauhaus-Universität Weimar
Dr. Erik Rodner	Friedrich-Schiller-Universität Jena
Dr. Ignaz Rutter	Karlsruher Institut für Technologie (KIT)
Dr. Oliver Schaudt	Universität zu Köln
Dr.-Ing. Fabian Scheler	Friedrich-Alexander-Universität Erlangen-Nürnberg
Dr. Marco Schmidt	Julius-Maximilians-Universität Würzburg
Dr. Hella Seebach	Universität Augsburg
Dr.-Ing. Sebastian Stober	Otto-von-Guericke-Universität Magdeburg
Dr. Torsten Stüber	Technische Universität Dresden
Dr. Johannes Textor	Universität zu Lübeck
Dr. Nora Christina Toussaint	Eberhard Karls Universität Tübingen
Dr. Christian Wachinger	Technische Universität München
Dr. Matthias Weidlich	Universität Potsdam
Dr. Ralf Wimmer	Albert-Ludwigs-Universität Freiburg
Dr.-Ing. Carola Winzen	Universität des Saarlandes
Dr. Thomas Würthinger	Johannes Kepler Universität Linz
Dr. Christine Zarges	TU Dortmund

Mitglieder des Nominierungsausschusses für den GI-Dissertationspreis 2011

Von links nach rechts:

Prof. Dr. Paul Molitor	Martin-Luther-Univ. Halle-Wittenberg
Prof. Dr. Gustaf Neumann	Wirtschaftsuniversität Wien
Prof. Dr. Harald Störrle	Technical University of Denmark
Prof. Dr. Myra Spiliopoulou	Otto-von-Guericke-Univ. Magdeburg
Prof. Dr.-Ing. Klaus-Peter Löhr	Freie Universität Berlin
Prof. Dr. Abraham Bernstein	Universität Zürich
Prof. Dr. Steffen Hölldobler (Vorsitzender)	Technische Universität Dresden

Nicht im Bild:

Prof. Dr. Rüdiger Reischuk (Fotograf)	Universität zu Lübeck
Prof. Dr. Dorothea Wagner	Universität Karlsruhe (TH)

Inhaltsverzeichnis

Genome Rearrangement Algorithmen

Martin Bader

Vector Informatik GmbH
Ingersheimer Straße 24
70499 Stuttgart
martin.bader@vector.com

Abstract: Durch die steigende Anzahl von vollständig sequenzierten Genomen gewinnt der Vergleich von Spezies auf Basis der sequenzierten Genome immer mehr an Bedeutung. Im Gegensatz zum klassischen Ansatz, welcher ein Distanzmaß basierend auf Punktmutationen verwendet, lassen *Genome Rearrangement Algorithmen* kleinere Mutationen ausser acht und berücksichtigen nur größere Umstrukturierungen, welche die Reihenfolge der Gene auf den Chromosomen verändern. Dadurch sind diese Algorithmen ein wertvolles Hilfsmittel zum Vergleich von Spezies, welche seit Millionen von Jahren divergieren, sowie von sich schnell verändernden Genomen, wie sie zum Beispiel in Krebszellen vorkommen.
Die Untersuchung dieser Genomumstrukturierungen führt zu einer Vielzahl von algorithmischen Fragestellungen, wie die Berechnung von evolutionären Distanzen, die Rekonstruktion evolutionärer Ereignisse und der Genreihenfolge in hypothetischen Vorfahren, bis hin zur Berechnung ganzer phylogenetischer Bäume. In der hier vorgestellten Dissertation [Bad11] werden einige dieser Problemstellungen sowohl von der theoretischen als auch von der praktischen Seite untersucht. So wird gezeigt, dass das Median-Problem, bei welchem ein möglichst plausibler Vorfahr für drei gegebene Genome gesucht wird, sowohl für die *Transpositionsdistanz* als auch für die *gewichtete Reversal- und Transpositionsdistanz* zu den NP-vollständigen Problemen gehört. Dazu wird ein Algorithmus vorgestellt, welcher in der Praxis vorkommende Instanzen dieser Probleme dennoch lösen kann. Desweiteren werden ein neuer Algorithmus zur phylogenetischen Rekonstruktion basierend auf diesen Distanzen sowie erstmals Algorithmen zum paarweisen Genomvergleich, welche Duplikationen und Deletionen beliebiger Länge zulassen, vorgestellt. Ein weiteres Ergebnis der Dissertation, auf welches in dieser Zusammenfassung jedoch nicht näher eingegangen werden soll, ist eine effiziente Implementierung eines Vor- und Nachverarbeitungsschritt, welcher für viele Genome Rearrangement Algorithmen benötigt wird.

1 Einleitung

Im Laufe der Evolution wird ein Genom immer wieder durch größere Umstrukturierungen, den sogenannten *Genome Rearrangements*, verändert. Diese evolutionären Operationen können sowohl die Reihenfolge der Gene auf den Chromosomen als auch die Orientierung der Gene verändern. Unter der Orientierung eines Genes versteht man hierbei die Information, auf welchem Strang der Doppelhelix sich die codierende Nukleotidsequenz des Genes befindet. Bei einem *Reversal* zum Beispiel wird ein Stück eines Chromosoms

ausgeschnitten und umgedreht an der selben Stelle wieder eingefügt. Hierbei wird der Rückwärtsstrang des ausgeschnittenen Abschnittes in den Vorwärtsstrang des Chromosoms und der Vorwärtsstrang des ausgeschnittenen Abschnittes in den Rückwärtsstrang des Chromosoms eingefügt. Dadurch ändert sich sowohl die Reihenfolge aller Gene im ausgeschnittenen Abschnitt, als auch deren Orientierung. Ein beispielhaftes Reversal ist in Abb. 1 dargestellt.

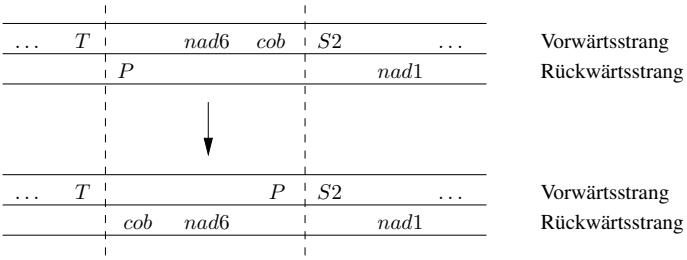


Abbildung 1: Beispiel für ein Reversal. Die drei Gene P, $nad6$ und cob ändern sowohl ihre Reihenfolge als auch den Strang des Chromosoms.

Für die algorithmische Betrachtung von Genome Rearrangements sind lediglich die unterschiedliche Reihenfolge und Orientierung der Gene auf den Chromosomen in verschiedenen Organismen von Interesse, jedoch nicht die genaue Bezeichnung der Gene. Deshalb kann hier eine abstraktere Schreibweise verwendet werden: In einem Genom, dem Referenzgenom, werden die Gene von 1 bis n durchnummeriert, ein Pfeil über der Zahl gibt die Orientierung des Gens an (wobei $\overrightarrow{x}$ der Vorwärtsstrang und $\overleftarrow{x}$ der Rückwärtsstrang ist). In den anderen Genomen werden die Gene durch die entsprechenden Nummern aus dem Referenzgenom ersetzt. In obigen Beispiel könnte also das erste Genom als $(\dots \overrightarrow{1}\overleftarrow{2}\overrightarrow{3}\overrightarrow{4}\overrightarrow{5}\overleftarrow{6} \dots)$ und das Ergebnis des Reversals als $(\dots \overrightarrow{1}\overleftarrow{4}\overleftarrow{3}\overrightarrow{2}\overrightarrow{5}\overleftarrow{6} \dots)$ geschrieben werden. Im Folgenden werden wir zwei weitere vereinfachende Annahmen machen, welche für bakterielle und mitochondriale Genome weitestgehend zutreffen.

1. Jedes Genom besteht aus genau einem zirkulären Chromosom (d.h. das Chromosom hat eine ringförmige Struktur und keinen fest definierten Anfangspunkt).

2. Jedes der betrachteten Gene kommt in jedem Genom exakt einmal vor.

Dadurch ergibt sich folgende mathematische Definition eines Genoms:

Definition 1 *Ein* Genom $\pi = (\pi_1 \dots \pi_n)$ *ist ein Wort über dem Alphabet* $\Sigma_n = \{1, \dots, n\}$*, in welchem die Indizes zyklisch sind (d.h. der Nachfolger von* n *ist* 1*) und jedes Element* x *eine positive oder negative Orientierung besitzt (gekennzeichnet durch* $\overrightarrow{x}$ *oder* $\overleftarrow{x}$*).*

Ein Teilwort $\pi_i \dots \pi_j$ eines Genoms bezeichnen wir als *Segment*. Auf einem Genom lassen sich jetzt die folgenden drei evolutionären Operationen definieren:

Definition 2 *Ein* Reversal $rev(i,j)$ *angewandt auf ein Genom* $\pi = (\pi_1 \dots \pi_n)$ *schneidet das Segment* $\pi_i \dots \pi_{j-1}$ *aus einem Genom, invertiert die Reihenfolge und Orientierung aller Gene des Segments, und fügt das Segment wieder an der selben Stelle in das Genom ein. Das heißt,*

$$rev(i,j) \cdot \pi = (\pi_1 \dots \pi_{i-1} - \pi_{j-1} \dots - \pi_i \pi_j \dots \pi_n),$$

wobei $-\pi_k$ *das Gen* π_k *mit invertierter Orientierung ist.*
Eine Transposition $tp(i,j,k)$ *angewandt auf ein Genom* $\pi = (\pi_1 \dots \pi_n)$ *schneidet das Segment* $\pi_i \dots \pi_{j-1}$ *aus einem Genom und fügt es vor dem Element* π_k *wieder ein. Das heißt,*

$$tp(i,j,k) \cdot \pi = (\pi_1 \dots \pi_{i-1} \pi_j \dots \pi_{k-1} \pi_i \dots \pi_{j-1} \pi_k \dots \pi_n).$$

Eine invertierte Transposition $itp(i,j,k)$ *angewandt auf ein Genom* $\pi = (\pi_1 \dots \pi_n)$ *ist die Konkatenation des Reversals* $rev(i,j)$ *und der Transposition* $tp(i,j,k)$ *angewandt auf* π*. Das heißt,*

$$itp(i,j,k) \cdot \pi = tp(i,j,k) \circ rev(i,j) \cdot \pi = (\pi_1 \dots \pi_{i-1} \pi_j \dots \pi_{k-1} - \pi_{j-1} \dots - \pi_i \pi_k \dots \pi_n).$$

Tatsächlich sind diese drei Operationen die drei in der Evolution am häufigsten vorkommenden Genome Rearrangements bei unichromosomalen Genomen. Anhand dieser Operationen lassen sich nun Genome Rearrangement Distanzen definieren:

Definition 3 *Die* Reversal- und Transpositionsdistanz $d_{rt}(\pi^1, \pi^2)$ *zwischen zwei Genomen* π^1 *und* π^2 *ist die minimale Anzahl an Reversals, Transpositionen und invertierten Transpositionen, welche benötigt wird, um* π^1 *in* π^2 *zu transformieren.*

Die *Reversaldistanz* $d_{rev}(\pi^1, \pi^2)$ und die *Transpositionsdistanz* $d_{tp}(\pi^1, \pi^2)$ sind analog definiert, nur dass hier die Operationen auf Reversals bzw. Transpositionen beschränkt sind. Bei der *gewichteten Reversal- und Transpositionsdistanz* $d_{wrt}(\pi^1, \pi^2)$ sind die verschiedenen Operationen unterschiedlich gewichtet, die Distanz ist hier das minimale Gewicht einer Sequenz von Operationen, welche π^1 in π^2 transformiert. Eine Sequenz von Operationen, welche π^1 in π^2 transformiert, nennt man eine sortierende Sequenz. In Abb. 2 ist eine sortierende Sequenz dargestellt, welche das Genom $\pi^1 =$

$$rev(2,5) \cdot (\overrightarrow{1}\ \overleftarrow{2}\ \overrightarrow{9}\ \overleftarrow{5}\ \overrightarrow{3}\ \overrightarrow{7}\ \overrightarrow{4}\ \overleftarrow{8}\ \overleftarrow{6}\ \overrightarrow{10}) = (\overrightarrow{1}\ \overrightarrow{5}\ \overleftarrow{9}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{7}\ \overrightarrow{4}\ \overleftarrow{8}\ \overleftarrow{6}\ \overrightarrow{10})$$

$$itp(6,7,9) \cdot (\overrightarrow{1}\ \overrightarrow{5}\ \overleftarrow{9}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{7}\ \overrightarrow{4}\ \overleftarrow{8}\ \overleftarrow{6}\ \overrightarrow{10}) = (\overrightarrow{1}\ \overrightarrow{5}\ \overleftarrow{9}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overleftarrow{8}\ \overleftarrow{7}\ \overleftarrow{6}\ \overrightarrow{10})$$

$$tp(2,4,7) \cdot (\overrightarrow{1}\ \overrightarrow{5}\ \overleftarrow{9}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overleftarrow{8}\ \overleftarrow{7}\ \overleftarrow{6}\ \overrightarrow{10}) = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5}\ \overleftarrow{9}\ \overleftarrow{8}\ \overleftarrow{7}\ \overleftarrow{6}\ \overrightarrow{10})$$

$$rev(6,10) \cdot (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5}\ \overleftarrow{9}\ \overleftarrow{8}\ \overleftarrow{7}\ \overleftarrow{6}\ \overrightarrow{10}) = (\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8}\ \overrightarrow{9}\ \overrightarrow{10})$$

Abbildung 2: Das Genom $(\overrightarrow{1}\ \overleftarrow{2}\ \overrightarrow{9}\ \overleftarrow{5}\ \overrightarrow{3}\ \overrightarrow{7}\ \overrightarrow{4}\ \overleftarrow{8}\ \overleftarrow{6}\ \overrightarrow{10})$ wird durch 4 Operationen in das Genom $(\overrightarrow{1}\ \overrightarrow{2}\ \overrightarrow{3}\ \overrightarrow{4}\ \overrightarrow{5}\ \overrightarrow{6}\ \overrightarrow{7}\ \overrightarrow{8}\ \overrightarrow{9}\ \overrightarrow{10})$ transformiert.

$(\overrightarrow{1}\overleftarrow{2}\overrightarrow{9}\overleftarrow{5}\overrightarrow{3}\overrightarrow{7}\overrightarrow{4}\overleftarrow{8}\overleftarrow{6}\overrightarrow{10})$ in das Genom $\pi^2 = (\overrightarrow{1}\overrightarrow{2}\overrightarrow{3}\overrightarrow{4}\overrightarrow{5}\overrightarrow{6}\overrightarrow{7}\overrightarrow{8}\overrightarrow{9}\overrightarrow{10})$ mit zwei Reversals, einer Transposition und einer invertierten Transpositionen transformiert. Es lässt sich zeigen, dass diese Sequenz minimal ist, also ist $d_{rt}(\pi^1, \pi^2) = 4$.
Ein weiteres häufig verwendetes Distanzmaß ist die *Breakpointdistanz*. Diese ist jedoch nicht über die Länge einer sortierenden Sequenz definiert, sondern es wird lediglich die Anzahl der unterschiedlichen Genübergänge in den Genomen gezählt.

2 Median-Probleme

Bei einem Medianproblem wird nach einem Genom gesucht, welches „möglichst mittig“ zwischen drei anderen Genomen liegt. Formal besteht die Eingabe aus drei Genomen π^1, π^2 und π^3, und es wird nach einem Genom σ gesucht, welches unter einem gegebenen Distanzmaß $d(\cdot, \cdot)$ die Summe der Distanzen $\Sigma_{i=1}^{3} d(\sigma, \pi^i)$ minimiert. Ein Programm, welches das Medianproblem entweder exakt oder approximativ löst, nennt man Mediansolver. Solche Mediansolver werden in allen aktuellen Programmen zur phylogenetischen Rekonstruktion basierend auf Genome Rearrangements verwendet. Allerdings sind die Medianprobleme ungleich schwerer zu berechnen als die korrespondieren paarweisen Distanzfunktionen. So ist die Reversaldistanz in linearer Zeit berechenbar [BMY01], Caprara konnte jedoch 1999 die NP-Vollständigkeit des *Reversal-Medianproblems* (kurz *RMP*) beweisen [Cap99, Cap03]. Ausgangspunkt dieses Beweises ist eine Variante des *Eulerian Cycle Decomposition Problem* (kurz *ECD*), bei dem gefragt ist, ob sich ein Graph vollständig in kantendisjunkte Dreiecke zerlegen lässt. Die NP-Vollständigkeit von ECD wurde bereits 1981 von Holyer bewiesen [Hol81]. Caprara reduziert in seinem Beweis zunächst ECD auf das *Zyklen-Medianproblem*, wobei für drei Genome π^1, π^2 und π^3 ein Genom σ gesucht wird, welches die Anzahl der Zyklen $\Sigma_{i=1}^{3} c(\sigma, \pi^i)$ im *multiplen Breakpointgraph* maximiert. Diese Zyklenanzahl ist eng mit der Reversaldistanz verknüpft, es gilt folgende Ungleichung:

$$\Sigma_{i=1}^{3} d_{rev}(\sigma, \pi^i) \geq 3n - \Sigma_{i=1}^{3} c(\sigma, \pi^i)$$

Anschließend wird gezeigt, dass der Reversalmedian gleichzeitig auch der Zyklenmedian ist, da im konstruierten Graph für den Zyklenmedian in obiger Formel die Gleichheit gilt. Dieser Schritt vervollständigt also die Reduktion von ECD auf das RMP.
Für die Transpositionsdistanz war der Status für sowohl die paarweise Distanz als auch für das Medianproblem lange Jahre offen. In meiner Dissertation gelang es mir zu beweisen, dass zumindest das *Transpositions-Medianproblem* (kurz *TMP*) ebenfalls NP-vollständig ist[1]. Dazu musste der Beweis von Caprara an verschiedenen Stellen abgeändert werden. So musste sichergestellt werden, dass beim Zyklusmedian alle Gene dieselbe Orientierung haben. Dieser Schritt war für das RMP nicht notwendig, da Reversals die Orientierung von Genen verändern. Bei der Transpositiondistanz wird jedoch nie die Orientierung eines Gens geändert, der Zyklusmedian könnte also nicht durch Transpositionen in die Ausgangsgenome transformiert werden und wäre somit keine gültige Lösung für das TMP.

[1] Mittlerweile konnte sogar die NP-Vollständigkeit der paarweisen Transpositionsdistanz bewiesen werden [BFR11].

Eine weitere Schwierigkeit beim TMP ist, dass für die untere Schranke die Länge der Zyklen im multiplen Breakpointgraph beachtet werden muss. Es gilt:

$$\Sigma_{i=1}^{3} d_{tp}(\sigma, \pi^i) \geq \frac{1}{2} \cdot (3n - \Sigma_{i=1}^{3} c_{odd}(\sigma, \pi^i))$$

Hierbei ist $c_{odd}(\sigma, \pi^i)$ die Anzahl der Zyklen ungerader Länge zwischen σ und π^i im multiplen Breakpointgraph. Schließlich muss noch gezeigt werden, dass für den Zyklusmedian bei obiger Formel die Gleichheit gilt. Dies ist jedoch nicht ohne weiteres möglich, da kein polynomieller Algorithmus für die Transpositionsdistanz zur Verfügung steht. Daher ist ein weiterer Reduktionsschritt nötig, welcher die Genome so modifiziert, dass die Transpositionsdistanzen zwischen diesen Genomen einfach zu berechnen sind.
Der Beweis lässt sich auch einfach auf das *gewichtete Reversal- und Transpositions-Medianproblem* (kurz *wRTMP*) übertragen, dieses ist also ebenfalls NP-vollständig. Für die Berechnung dieser Probleme in der Praxis habe ich einen Branch-and-Bound Algorithmus entwickelt, welcher ebenfalls auf der Arbeit von Caprara beruht. Der Algorithmus kann den TMP oder den wRTMP wahlweise exakt oder approximativ berechnen. Experimentelle Ergebnisse zeigen, dass der Algorithmus sowohl schneller ist als auch eine bessere Approximationsgüte besitzt als die einzige andere verfügbare Implementierung eines Mediansolvers für das TMP [YZT08] (welcher jedoch auf einer anderen Technik beruht). Für das wRTMP gibt es keine andere Implementierung, die praxisrelevante Instanzen des Problems lösen kann.

3 Phylogenetische Rekonstruktion

Bei der phylogenetischen Rekonstruktion geht es darum, zu einer Menge von Organismen einen möglichst plausiblen Abstammungsbaum zu ermitteln. Im Zusammenhang mit Genome Rearrangements werden die Spezies durch ihre Genome repräsentiert und die Topologie des phylogenetischen Baumes sowie Genreihenfolgen für innere Knoten des Baumes berechnet. Die Genreihenfolgen der inneren Knoten entsprechen also den Genomen hypothetischer Vorfahren. Die Güte eines phylogenetischen Baumes wird über die Distanzen benachbarter Knoten berechnet, wobei als Distanzmaß eine festgelegte Genome Rearrangement Distanz verwendet wird (also z.B. die Reversaldistanz oder die Transpositionsdistanz). Formal lässt sich dies wie folgt definieren:

Definition 4 *Ein* phylogenetischer Baum *einer Menge von Genomen $P = \{\pi^1, \ldots, \pi^k\}$ (den Eingabegenomen) ist ein Baum $T = (V, E)$, wobei V die Knoten und E die Blätter des Baumes sind. Jeder Knoten ist mit einem Genom π^i markiert, und es besteht eine Bijektion zwischen den Eingabegenomen und den Blättern des Baumes (d.h. jedes Element aus P ist Markierung von genau einem Blatt). Das Gewicht einer Kante (π^i, π^j) ist die Distanz $d(\pi^i, \pi^j)$. Das Gewicht eines Baumes $w(T)$ ist die Summe der Gewichte seiner Kanten.*

Beim *Phylogenie-Problem* geht es nun darum, zu einer Menge von Eingabegenomen einen Baum mit möglichst geringem Gewicht zu finden. Da dieses Probelm für alle bekannten

Genome Rearrangement Distanzen NP-hart ist, werden hierfür in der Praxis heuristische Algorithmen eingesetzt. Die Approximationsgüte dieser Algorithmen lässt sich rechnerisch nur sehr grob bestimmen, deshalb werden die Algorithmen anhand von Benchmark-Datensätzen getestet und verglichen. Prinzipiell finden sich in der Literatur zwei verschiedene algorithmische Strategien für das Phylogenie-Problem:

Strategie 1: Erstelle zuerst die Topologie des Baumes und berechne dann Genome für die inneren Knoten

Diese Strategie wurde erstmals im Tool `BPAnalysis` [SB98] verwendet, welches über alle möglichen Topologien iteriert und eine schnelle Heuristik zur initialen Berechnung der Genome der inneren Knoten verwendet. Anschließend werden die Genome der inneren Knoten mit Hilfe eines Mediansolvers iterativ verbessert. Dieser erste Ansatz war relativ langsam, hauptsächlich wegen der Iteration über alle möglichen Topologien. Deshalb wurde dieser Schritt durch eine Heuristik ersetzt, ebenfalls wurde die restliche Implementierung des Algorithmus fortlaufend verbessert [CJM$^+$00, MWB$^+$01, MSTL02]. Der Schwachpunkt dieser Strategie ist jedoch, dass die Topologie nicht mit der eigentlich verwendeten Genome Rearrangement Distanz, sondern mit der leichter zu berechnenden *Breakpoint Distanz* berechnet wird.

Strategie 2: Beginne mit einem leeren Baum und füge die Eingabegenome iterativ hinzu. Markiere in jedem Schritt neu hinzugekommene innere Knoten

Diese Strategie wurde zuerst in einem Softwaretool namens `MGR` implementiert [BP02]. Die Heuristik zum Hinzufügen eines Genoms verwendet einen Mediansolver für das RMP (der Einsatz von Mediansolvern für andere Distanzmaße ist auch möglich, siehe z.B. [AS08, XM10]). Eine weitere Verbesserung namens `amGRP` nutzt die Tatsache, dass der Median im Allgemeinen nicht eindeutig ist, und wählt aus allen Medianen mittels einer Heuristik denjenigen aus, welcher für den weiteren Verlauf des Algorithmus möglichst gut ist [BMM07]. Der Vorteil der Strategie ist, dass der phylogenetische Baum direkt bezüglich dem gewünschten Distanzmaß aufgebaut wird. Allerdings müssen in jeder Iteration sehr viele Mediane berechnet werden. Dies ist akzeptabel für die Reversaldistanz, die Mediansolver für das TMP und das wRTMP sind jedoch für diesen Ansatz trotz der im letzten Kapitel angesprochenen Verbesserungen zu langsam.

Um das Phylogenie-Problem bezüglich der gewichteten Reversal- und Transpositionsdistanz zu lösen, sollte ein Algorithmus also zwei Eigenschaften erfüllen. Erstens sollte der phylogenetische Baum direkt bezüglich dieser Distanz aufgebaut werden. Zweitens sollte der Einsatz von Mediansolvern weitestgehend vermieden werden. Der von mir entwickelte Algorithmus verfolgt Strategie 2, er baut also den Baum iterativ auf. Anstatt neue innere Knoten mit einem Mediansolver zu bestimmen, wird für jede Kante (π^i, π^j) eine Menge von Knoten $cloud(\pi^i, \pi^j)$ gespeichert, welche „zwischen“ π^i und π^j liegen. Formal gilt für die Knoten $\pi^s \in cloud(\pi^i, \pi^j) : \; d_{wrt}(\pi^i, \pi^s) + d_{wrt}(\pi^s, \pi^j) \leq d_{wrt}(\pi^i, \pi^j) + \delta$ für eine kleine Konstante δ. Anstatt beim Hinzufügen von einem Knoten $\pi^p \in P$ zum bestehenden Baum den Median von π^i, π^j und π^p zu berechnen, wird für alle Knoten $\pi^s \in cloud(\pi^i, \pi^j)$ die Distanz $d_{wrt}(\pi^s, \pi^p)$ berechnet. Derjenige Knoten, welcher die-

se Distanz minimiert, wird als Median verwendet. Diese Wahl ist in den meisten Fällen nicht viel schlechter als der tatsächliche Median, durch eine Beschränkung der Größe von $cloud(\pi^i, \pi^j)$ ist dieses Verfahren jedoch wesentlich schneller. Der so erstellte phylogenetische Baum durchläuft abschliessend noch zwei Verbesserungsschritte. Im ersten Schritt wird durch systematisches Entfernen und Neueinfügen die Topologie des Baumes verbessert. Dieser Schritt bedient sich abermals der „Clouds“ zwischen den Knoten. Im zweiten Schritt werden die Genome der inneren Knoten wie bei Strategie 1 beschrieben mit einem Mediansolver optimiert. Da der Baum schon vor diesem Schritt relativ gut ist, müssen hierbei nur wenige Iterationen durchlaufen werden, so dass der Einsatz des Mediansolvers nicht zu kostspielig ist.
Mangels eines anderen Tools, welches phylogenetische Bäume bezüglich der gewichteten Reversal- und Transpositionsdistanz berechnet, war nur eine indirekte Evaluierung des Algorithmus möglich. Dazu wurden phylogenetische Bäume bezüglich der Reversaldistanz mit den State-of-the-Art-Tools `GRAPPA` [MWB$^+$01], `MGR` [BP02] und `amGRP` [BMM07] berechnet und für diese Bäume das Gewicht bezüglich der gewichteten Reversal- und Transpositionsdistanz ermittelt. Dabei zeigte sich, dass der von mir gewählte direkte Ansatz zu besseren Ergebnissen führt. Anschließend wurde bei meinem Algorithmus das Distanzmaß auf die Reversaldistanz umgestellt. Hier schnitt mein Algorithmus bei ähnlicher Laufzeit wie `amGRP` nur unwesentlich schlechter als dieser ab, `GRAPPA` und `MGR` wurden sowohl in Laufzeit als auch in der Qualität der Ergebnisse deutlich geschlagen.

4 Genome Rearrangements mit Duplikationen

In den vorigen Kapiteln haben wir angenommen, dass jedes Genom aus nur einem Chromosom besteht, und dass jedes Gen in jedem Genom exakt einmal vorkommt. Diese Annahme ist für bakterielle und mitochondriale DNA praktikabel, trifft jedoch nicht für die Genome höherer Organismen zu. Insbesondere bei pflanzlicher DNA sind lange duplizierte Abschnitte der DNA häufig. Ebenso ist in der Krebsforschung die Betrachtung von Duplikationen wichtig, da hier viele der Mutationen aus solchen bestehen. Aus algorithmischer Sicht sind Duplikationen jedoch problematisch. So ist zum Beispiel die Reversaldistanz in linearer Zeit berechenbar, falls keine duplizierten Gene vorliegen. Mit duplizierten Genen ist dieses Problem jedoch NP-vollständig [CZF$^+$05]. Falls die Anzahl der Vorkommen eines Gens in den zu vergleichenden Genomen unterschiedlich sind, müssen ausserdem zusätzliche Operationen wie *Duplikationen* und *Deletionen* im Algorithmus erlaubt werden. Durch diese algorithmischen Schwierigkeiten wurden Duplikationen nur wenig untersucht. Die meisten Algorithmen, welche Duplikationen erlauben, beschränken die Länge der Duplikation auf ein einzelnes Gen oder ein Segment beschränkter Länge (siehe z.B. [EM02]). Zum Zeitpunkt meiner Dissertation gab es nur einen Ansatz, welcher Duplikationen beliebiger Länge erlaubte, allerdings auf einem stark vereinfachten Genommodell [OFS08]. In meiner Dissertation habe ich einen Algorithmus entwickelt, welcher eine sortierende Sequenz zwischen zwei Genomen π^1 und π^2 berechnet und dabei eine Vielzahl an verschiedenen Operationen berücksichtigt, unter anderem *Tandemduplikationen* und *Deletionen* beliebiger Länge. Obwohl das Startgenom π^1 der Restrik-

tion unterliegt, dass seine Chromosomen entweder identisch oder elementendisjunkt sein müssen (d.h. hier können duplizierte Gene nur bei einer exakten Kopie eines Chromosoms auftreten), kann dieses Modell in der Praxis eingesetzt werden, z.B. zur Untersuchung von Mutationen in Krebszellen. Der Algorithmus basiert auf einer Erweiterung des *Breakpoint-Graphen* (für Details siehe [BP93]). Während dieser im „klassischen Fall" (d.h. ohne duplizierte Gene) in Zyklen zerfällt, erhält man mit Duplikationen Zusammenhangskomponenten. Es lässt sich zeigen, dass die Anzahl der Komponenten maximiert wird, falls $\pi^1 = \pi^2$ gilt und π^1 oben angesprochene Restriktion erfüllt. Desweiteren kann die Anzahl der Komponenten durch jede Operation maximal um 1 erhöht werden. Daraus lässt sich eine Greedy-Strategie entwickeln: Sortiere π^2 rückwärts nach π^1 (π^1 darf nicht verändert werden, damit es die Restriktion immer erfüllt) und wende (falls möglich) eine Operation an, welche die Anzahl der Komponenten erhöht. Leider ist dies nicht in jedem Schritt möglich. Deshalb wurde ein weiteres Maß $\tau(\pi^1, \pi^2)$ eingeführt, welches die „Unordnung" zwischen den Genomen bestimmt. Dieses Maß berücksichtigt einerseits die Anzahl an Breakpoints, andererseits die absolute Anzahl der Elemente, welche in π^2 eingefügt oder gelöscht werden müssen. Falls es keine Operation gibt, welche die Anzahl der Komponenten vergrößert, wird diejenige Operation genommen, welche $\tau(\pi^1, \pi^2)$ minimiert. Ausserdem wird $\tau(\pi^1, \pi^2)$ verwendet, falls mehrere Operationen die Anzahl der Komponenten vergrößert, d.h. von diesen wird ebenfalls immer diejenige gewählt, welche $\tau(\pi^1, \pi^2)$ minimiert. Um die Terminierung des Algorithmus zu garantieren, ist ausserdem ein einfacher Fallback-Algorithmus notwendig, der in der Praxis aber nur sehr selten zum Einsatz kommt.
Zur praktischen Evaluation des Algorithmus habe ich zufällige Sequenzen von Operationen auf ein festes Genom π^1 angewandt, um ein Zielgenom π^2 zu erhalten. Nun wurde mein Algorithmus verwendet, um eine sortierende Sequenz zwischen π^1 und π^2 zu berechnen, und diese mit der zur Generierung verwendeten Sequenz verglichen. Die Ergebnisse zeigen, dass die Längen dieser Sequenzen sehr ähnlich sind, solange diese nicht zu lang sind. Ebenfalls entsprechen sich hier die relativen Häufigkeiten der verschiedenen Operationen in der berechneten und generierten Sequenz. Weitere Experimente wurden auf der *Mitelman Database of Chromosome Abberations and Gene Fusions in Cancer* [MJM10] durchgeführt, welche eine Sammlung von Krebskaryotypen beinhaltet, die von Hand aus der Literatur der letzten 20 Jahre zusammengetragen wurde. Alle beschriebenen Genome konnten von meinem Algorithmus sehr schnell analysiert werden, die resultierenden sortierenden Sequenzen entsprechen sehr gut den in der Literatur beschriebenen Sequenzen.

5 Zusammenfassung

In meiner Dissertation habe ich verschiedene Probleme im Bereich Genome Rearrangements betrachtet. Die Hauptergebnisse meiner Dissertation sind folgende:

- Eine effiziente Implementierung eines für viele Genome Rearrangement Algorithmen benötigten Vor- und Nachverarbeitungsschrittes (nicht in dieser Zusammenfassung beschrieben).

- Der Beweis der NP-Vollständigkeit des *Transpositions-Medianproblems*, sowie ein praktisch einsetzbarer Branch-and-Bound Algorithmus für eben dieses Problem und für das (ebenfalls NP-vollständige) *gewichtete Reversal- und Transpositions-Medianproblem*. Der Algorithmus bringt eine deutliche Verbesserung sowohl in der Laufzeit als auch in der Qualität der Ergebnisse im Vergleich zum bisher einzigen verfügbaren Transpositions-Mediansolver.
- Ein Algorithmus zur phylogenetischen Rekonstruktion basierend auf der *gewichteten Reversal- und Transpositionsdistanz*. Dies ist der bisher einzige Algorithmus, der direkt auf dieser Distanz arbeitet. Die Qualität der vom Algorithmus berechneten Bäume ist wesentlich besser als bei einem indirekten Ansatz über die Reversaldistanz. Selbst bei der phylogenetischen Rekonstruktion basierend auf der Reversaldistanz, für die der Algorithmus eigentlich nicht entwickelt wurde, ist er nur unwesentlich schlechter als der momentan hierfür beste verfügbare Algorithmus.
- Ein Algorithmus zum paarweisen Vergleich von Genomen mit duplizierten Genen. Der Algorithmus betrachtet eine Vielzahl von verschiedenen evolutionären Operationen und ist einer der wenigen Algorithmen, welche Duplikationen und Deletionen beliebiger Länge zulassen.

Literatur

[AS08] Z. Adam und D. Sankoff. The ABCs of MGR with DCJ. *Evolutionary Bioinformatics*, 4:69–74, 2008.

[Bad11] M. Bader. *Genome Rearrangement Algorithms*. Dissertation, Universität Ulm, 2011.

[BFR11] L. Bulteau, G. Fertin und I. Rusu. Sorting by Transpositions is Difficult. In L. Aceto, M. Henzinger und J. Sgall, Hrsg., *Proc. 38th Int. Coll. on Automata, Languages and Programming, Part I*, Seiten 654–665. Springer-Verlag, Heidelberg, 2011.

[BMM07] M. Bernt, D. Merkle und M. Middendorf. Using Median Sets for Inferring Phylogenetic Trees. *Bioinformatics*, 23:e129–e135, 2007.

[BMY01] D.A. Bader, B.M.E. Moret und M. Yan. A Linear-Time Algorithm for Computing Inversion Distance between Signed Permutations with an Experimental Study. *J. of Computational Biology*, 8:483–491, 2001.

[BP93] V. Bafna und P.A. Pevzner. Genome Rearrangements and Sorting by Reversals. In *Proc. 34th IEEE Symposium on Foundations of Computer Science*, Seiten 148–157. IEEE Computer Society Press, 1993.

[BP02] B. Bourque und P.A. Pevzner. Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Research*, 12(1):26–36, 2002.

[Cap99] A. Caprara. Formulations and Hardness of Multiple Sorting by Reversals. In S. Istrail, P. Pevzner und M.S. Waterman, Hrsg., *Proc. 3rd Annual Int. Conf. on Computational Molecular Biology*, Seiten 84–93. ACM Press, New York, 1999.

[Cap03] A. Caprara. The Reversal Median Problem. *INFORMS J. on Computing*, 15(1):93–113, 2003.

[CJM+00] M.E. Cosner, R.K. Jansen, B.M.E. Moret, D.A. Raubeson, L.-S. Wang, T. Warnow und S.K. Wyman. A New Fast Heuristic for Computing the Breakpoint Phylogeny and Experimental Phylogenetic Analyses of Real and Synthetic Data. In P. Bourne, M. Gribskov, R. Altman, N. Jensen, D. Hope, T. Lengauer, J. Mitchell, E. Scheeff, C. Smith, S. Strande und H. Weissig, Hrsg., *Proc. 8th Int. Conf. on Intelligent Systems for Molecular Biology*, Seiten 104–115. AAAI Press, Menlo Park, 2000.

[CZF+05] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi und T. Jiang. The Assignment of Orthologous Genes via Genome Rearrangement. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(4):302–315, 2005.

[EM02] N. El-Mabrouk. Reconstructing an Ancestral Genome Using Minimum Segments Duplications and Reversals. *J. of Computer and System Sciences*, 65:442–464, 2002.

[Hol81] I. Holyer. The NP-Completeness of Some Edge-Partition Problems. *SIAM J. on Computing*, 10(4):713–717, 1981.

[MJM10] F. Mitelman, B. Johansson und F. Mertens, Hrsg. *Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer*, http://cgap.nci.nih.gov/Chromosomes/Mitelman, 2010.

[MSTL02] B.M.E. Moret, A. Siepel, J. Tang und T. Liu. Inversion Medians Outperform Breakpoint Medians in Phylogeny Reconstruction from Gene-order Data. In R. Guigó und D. Gusfield, Hrsg., *Proc. 2nd Workshop on Algorithms in Bioinformatics*, Seiten 521–536. Springer-Verlag, Heidelberg, 2002.

[MWB+01] B.M.E. Moret, S.K. Wyman, D.A. Bader, T. Warnow und M. Yan. A New Implementation and Detailed Study of Breakpoint Analysis. In R.B. Altman, A.K. Dunkerand L. Hunker, K. Lauderdale und T.E. Klein, Hrsg., *Proc. 6th Pacific Symposium on Biocomputing*, Seiten 583–594, 2001.

[OFS08] M. Ozery-Flato und R. Shamir. Sorting Cancer Karyotypes by Elementary Operations. In C. Nelson und S. Vialette, Hrsg., *Proc. 6th Annual RECOMB Satellite Workshop on Comparative Genomics*, Seiten 211–225. Springer-Verlag, Heidelberg, 2008.

[SB98] D. Sankoff und M. Blanchette. Multiple Genome Rearrangement and Breakpoint Phylogeny. *J. of Computational Biology*, 5(3):555–570, 1998.

[XM10] A.W. Xu und B.M.E. Moret. Genome Rearrangement Analysis on High-Resolution Data. Submitted, 2010.

[YZT08] F. Yue, M. Zhang und J. Tang. Phylogenetic Reconstruction from Transpositions. *BMC Genomics*, 9(Suppl 2):S15, 2008.

Dr. Martin Bader wurde am 4. Februar 1980 In Friedrichshafen geboren. Von 2000 bis 2005 studierte er Informatik an der Universität Ulm. Sein Diplom schloss er mit Gesamtnote 1,0 ab, seine Diplomarbeit zum Thema „Sorting by weighted transpositions and reversals“ wurde mit dem NRW Undergraduate Science Award 2005 in der Kategorie Bioinformatics and Genome Research ausgezeichnet. Nach seinem Abschluss als Diplominformatiker promovierte er im Institut für Theoretische Informatik an der Universität Ulm. Am 6. Mai 2011 schloss er seine Promotion zum Thema „Genome Rearrangement Algorithms“ mit dem Gesamtprädikat „summa cum laude“ ab. Seitdem arbeitet er für die Vector Informatik GmbH in Stuttgart als Software Development Engineer.

Qualitätsziel-orientierter Architekturentwurf und Traceability für weiterentwickelbare Software-Systeme

Zusammenfassung zur Dissertation von

Stephan Bode

Technische Universität Ilmenau
Stephan.Bode@arcor.de

Abstract: Die Evolution von Softwaresystemen erfordert häufige Anpassungen z. B. aufgrund sich ändernder Geschäftsprozesse oder Technologien. Bisherige Methoden unterstützen dies nur unzureichend aufgrund mangelnder Berücksichtigung von Qualitätszielen wie Weiterentwickelbarkeit sowie mangelnder Nachvollziehbarkeit von Architekturentwurfsentscheidungen. Das neue Konzept *Goal Solution Scheme*, das Qualitätsziele über Architekturprinzipien auf Lösungsinstrumente durch explizite Abhängigkeiten abbildet, hilft geeignete Architekturlösungen entsprechend ihrem Einfluss auf die Qualitätsziele wie Weiterentwickelbarkeit auszuwählen und Entwurfsentscheidungen nachzuvollziehen. Das Schema ist in ein zielorientiertes Architekturentwurfsvorgehen eingebettet, das etablierte Methoden und Konzepte des Requirements Engineering und Architekturentwurfs verbessert und integriert. Dies wird ergänzt durch ein Traceability-Konzept, welches eine (halb-)automatische Erstellung von Traceability Links mit hoher Genauigkeit und Trefferquote ermöglicht. Die Realisierbarkeit des Entwurfsansatzes wurde mit einer Fallstudie eines Softwaresystems für mobile Serviceroboter gezeigt. Ein prototypisches Werkzeug names EMFTrace zeigt die Anwendbarkeit der Konzepte.

1 Einleitung

Softwaresysteme werden heute mit häufigen Forderungen nach Veränderungen konfrontiert, z. B. aufgrund sich ändernder Geschäftsprozesse oder Technologien. Die Software und speziell ihre Architektur muss mit diesen häufigen Änderungen zurecht kommen, um dauerhaft nutzbar zu bleiben, da eine Ablösung existierender Softwaresysteme durch Neuentwicklungen wegen der damit verbundenen Risiken gewöhnlich keine akzeptable Lösung darstellt.

Während der Software-Evolution können Änderungen zu einer Strukturverschlechterung der Architektur führen, der Architekturerosion. Dies erschwert weitere Änderungen aufgrund von Inkonsistenzen oder fehlendem Programmverstehen oder verhindert sie gar. Um Änderungen zu unterstützen und die Erosion zu vermeiden, müssen speziell Qualitätsziele wie Weiterentwickelbarkeit, Performanz oder Gebrauchstauglichkeit sowie die Nachvollziehbarkeit von Entwurfsentscheidungen beim Architekturentwurf berücksichtigt werden. Dies wird jedoch oft vernachlässigt.

Existierende Entwurfsmethoden unterstützen den Übergang von den Qualitätzielen zu geeigneten Architekturlösungen nur unzureichend, da immer noch eine Lücke zwischen Methoden des Requirements Engineering und Architekturentwurfs existiert. Insbesondere fehlt Unterstützung für die Weiterentwickelbarkeit und die Nachvollziehbarkeit von Entwurfsentscheidungen durch explizite Modellabhängigkeiten.

Diese Arbeit präsentiert drei neue miteinander verzahnte Konzepte. Zuerst wird das *Goal Solution Scheme* vorgestellt, das Qualitätsziele über Architekturprinzipien auf Lösungsinstrumente durch explizite Abhängigkeiten abbildet. Das Schema hilft geeignete Architekturlösungen entsprechend ihrem Einfluss auf die Qualitätsziele auszuwählen. Es wird besonders hinsichtlich Weiterentwickelbarkeit diskutiert. Zum zweiten wird das Schema in ein *zielorientiertes Architekturentwurfsvorgehen* eingebettet, das etablierte Methoden und Konzepte des Requirements Engineering und Architekturentwurfs verbessert und integriert. Drittens wird dies ergänzt durch ein *Traceability-Konzept, welches einen regelbasierten Ansatz mit Techniken des Information Retrieval verbindet*. Dies ermöglicht eine (halb-)automatische Erstellung von Traceability Links mit spezifischen Typen und Attributen für eine definierte Semantik sowie mit hoher Genauigkeit und Trefferquote.

Die Realisierbarkeit des Ansatzes wird an einer Fallstudie eines Softwaresystems für mobile Serviceroboter gezeigt. Ein prototypisches Werkzeug names EMFTrace wurde als eine erweiterbare Plattform basierend auf Eclipse-Technologie implementiert, um die Anwendbarkeit der Konzepte zu zeigen. Es integriert Entwurfsmodelle von externen CASE-Werkzeugen mittels XML-Technologie in einem gemeinsamen Modell-Repository, wendet Regeln zur Linkerstellung an und bietet Validierungsfunktionen für Regeln und Links.

2 Bewertung des Standes der Technik

Im folgenden wird ein kurzer Abriss zur Bewertung von Methoden und Konzepten des Standes der Technik gegeben, die die Entwicklung der drei neuen Konzepte dieser Arbeit maßgeblich beeinflusst haben.

Beschreibung funktionaler und qualitativer Anforderungen Für die Analyse, Verfeinerung und grafische Repräsentation von Qualitätszielen und ihren Beziehungen untereinander können z. B. das NFR Framework [CNYM00], das i* Framework [Yu95] oder die standardisierte User Requirements Notation (URN) [ITU08] genutzt werden. Obwohl diese Ansätze des Anforderungsmanagements gut zur Darstellung der Beziehungen zwischen Qualitätszielen geeignet sind, haben sie aus Architektursicht einige Nachteile und Beschränkungen. Aufgrund ihres Ursprungs zielen sie eindeutig auf die Anforderungsanalysephase. Daher berücksichtigen sie nur unzureichend Architekturprinzipien und technische Randbedingungen und helfen wenig beim zielgerichteten Architekturentwurf mittels Vorschlag geeigneter Lösungsinstrumente. Aber die Ansätze können gut in Architekturentwurfsmethoden eingebettet werden, um Nutzen aus ihnen zu ziehen.

Evolutionsunterstützung durch Architekturentwurfsmethoden Für den Softwarearchitekturentwurf existieren einige etablierte Methoden wie QASAR [Bos00], ADD [BKB02] oder Globale Analyse [HNS00], welche in ein allgemeines Modell mit den Phasen Analyse, Synthese und Evaluierung [HKN+07] passen. Die Methoden versuchen gewöhnlich Qualitätsziele und funktionale Anforderungen mittels Architekturmustern [HA07] oder sogenannten Taktiken [BKB02] auszubalancieren. Jedoch fehlt ein systematischer, zielgerichteter Ansatz mit Anleitung zur Auswahl geeigneter Architekturentwurfselemente insbesondere für die Synthesephase und ausgerichtet auf Weiterentwickelbarkeit. Darüber hinaus fehlt die Integration mit den zielorientierten Ansätzen des Requirements Engineering (RE). Dies resultiert in einer immer noch vorhandenen konzeptionellen Lücke zwischen den Ansätzen der RE-Gemeinschaft sowie der Architekturentwurfsmethoden und schließlich in den wohl bekannten Evolutionsproblemen.

Entwurfs-Traceabiltiy Traceability, Rückverfolgbarkeit, ist ein Konzept, das das Nachvollziehen von Entwurfsentscheidungen ermöglichen kann. Sogenannte Traceability Links können verschiedene Softwareartefakte in Beziehung zueinander setzen und ermöglichen so die Verfolgung von Entwurfsentscheidungen sowie die Durchführung einer Änderungsauswirkungsanalyse. Ansätze der Requirements Traceability, z. B. [GF94, RJ01], erlauben es die Herkunft von Anforderungen und ihre Änderungen zurückzuverfolgen. Leider gibt es nur wenige Traceability-Ansätze für Entwurfsartefakte. Information Retrieval-basierte Ansätze, z. B. [ACC+02, CHSB+05], nutzen Ähnlichkeitsmaße auf Bezeichnern und können automatisiert werden. Allerdings ist die Genauigkeit und Trefferquote ihrer nachgelagerten Linkerstellung stark beschränkt. Konsequenterweise sollten Traceability Links bereits während des Entwurfs und der Änderung von Architekturmodellen erstellt werden. Daher müssen die Schritte zur Linkerstellung in die Entwurfsmethoden eingebettet werden. Regelbasierte Ansätze, z. B. [SZPMK04, MGP08], können Links halb-automatisch während der Durchführung von Entwurfsaktivitäten erzeugen, sodass der Entwickler nur bei Konflikten und Mehrdeutigkeit eingreifen muss. Die Korrektheit und Vollständigkeit der Links hängt hier hauptsächlich von der Qualität der Regeln ab. Leider gibt es bisher keine Ansätze, die in den Entwurfsprozess integriert sind und Links zwischen verschiedenartigen Artefakten, insbesondere Qualitätszielen und Entwurfselementen, erstellen können.

3 Begriff Weiterentwickelbarkeit

Obwohl es bereits einige Definitionen für Weiterentwickelbarkeit (engl. evolvability) gibt (z. B. [RLL98, BCE08]), existiert noch kein standardisiertes Qualitätsmodell und kein einheitliches Verständnis der Weiterentwickelbarkeit und insbesondere der Differenzierung zu Wartbarkeit. Breivold et al. [BCE08] diskutieren Weiterentwickelbarkeit im Detail, aber ihre Verfeinerung in Untermerkmale ist unvollständig und sie bieten keine Anleitung für eine zielgerichtete Berücksichtigung von Weiterentwickelbarkeit beim Architekturentwurf. Da der Begriff von zentraler Bedeutung ist für diese Arbeit und insbesondere die Diskussion im Rahmen des Goal Solutions Schemes, erfolgt hier eine Definition des

Qualitätsziels und eine Abgrenzung insbesondere zu Wartbarkeit, Wartung sowie Evolution (siehe auch [RB09]). Demnach ist Weiterentwickelbarkeit die Fähigkeit eines Softwaresystems sich während seiner Lebenszeit an Änderungen und Verbesserungen in Anforderungen und Technologien anzupassen, die die architekturelle Struktur des Systems beeinflussen, mit den geringsten Kosten unter Beibehaltung der Architekturintegrität.

4 Das Goal Solution Scheme

Wie bereits oben erwähnt ist die systematische Unterstützung für die Behandlung von Qualitätszielen wie Weiterentwickelbarkeit, Performanz oder Gebrauchstauglichkeit durch existierende Konzepte mangelhaft. Zur Unterstützung einer qualitätszielorientierten Entwicklung ist eine Abbildung zwischen Problemraum und Lösungsraum notwendig. Das Goal Solution Scheme (GSS) wurde entwickelt, um eine Abbildung zwischen Elementen beider Räume explizit modellieren und somit systematisch geeignete Architekturlösungen aus Qualitätszielen ableiten zu können. Aufgrund der verschiedenen Konzepte des Problem- und Lösungsraums besteht dazwischen eine konzeptuelle Lücke, welche zu Schwierigkeiten bei der Erstellung und Pflege der Beziehungen zwischen beiden führt. Um die Lücke zu schließen, wurden zwei zusätzliche Ebenen zwischen den Zielen und den Lösungsinstrumenten eingeführt, wie in Abbildung 1 zu sehen.

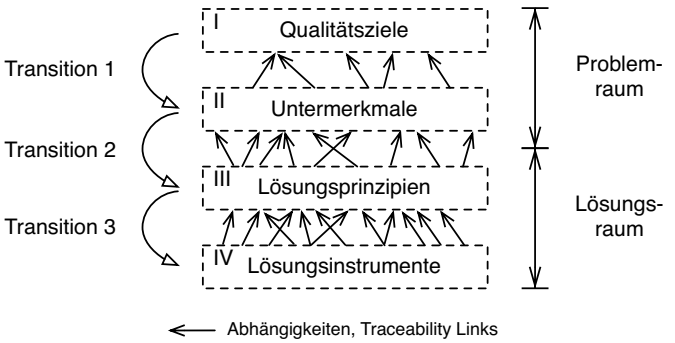


Abbildung 1: Struktur des Goal Solution Schemes.

Die Ebenen des Schemas entsprechen den Phasen während des Entwicklungsprozesses und enthalten die Elemente dieser Phasen. Ebene I und II repräsentieren den Problemraum, während die Ebenen III und IV den Lösungsraum darstellen. Von oben nach unten zeigt das Schema mögliche Verfeinerungen und Entscheidungen während des Architekturentwurfs und der Implementierung von Qualitätszielen. Die Beziehungen zwischen den Elementen des Schemas bilden einen Graph mit einer baumähnlichen Struktur. Jede Beziehung drückt eine Abhängigkeit aus, z. B. eine Verfeinerung von Zielen. Die Änderung eines Elementes erfordert die Änderung der abhängigen Elemente. Außerdem repräsentieren die Beziehungen zwischen den Ebenen einen positiven oder negativen Einfluss der Elemente

des Lösungsraums auf die Elemente des Problemraums. Gewichte auf den Beziehungen drücken den Einfluss quantitativ aus und werden für die Entscheidungsfindung verwendet. Die Abhängigkeiten können außerdem als Traceability Links repräsentiert werden, welche während des Entwurfs aufgezeichnet werden und die Entscheidungen reflektieren.

Ebene I ist Teil des Problemraums und beinhaltet die Qualitätsziele. Da es gewöhnlich schwieriger ist, die Qualitätsziele eines Softwaresystems zu realisieren als seine Funktionalität, sollten diese Qualitätsziele explizit modelliert werden, wie dies auch von den zielorientierten RE Ansätzen vorgeschlagen wird. Ebene I enthält die hochrangigen Qualitätsziele wie Weiterentwickelbarkeit, Performanz und Gebrauchstauglichkeit.

Ebene II ist auch noch Teil des Problemraums und enthält die Untermerkmale der hochrangigen Qualitätsziele, wie z. B. Änderbarkeit, Wiederverwendbarkeit und Testbarkeit. Diese Ebene wurde eingeführt, um die erwähnte Lücke zwischen Problem- und Lösungsraum mittels Zielverfeinerung zu verringern. Die Beziehungen der Transition 1 zwischen Ebene I und II repräsentiert eine Abbildung von hochrangigen Qualitätszielen zu Unterzielen ähnlich einem Qualitätsmodell. Zur Modellierung dieser Beziehungen kann die URN verwendet und auf Qualitätsmodelle wie ISO 9126 zurückgegriffen werden. Später können auch Präferenzen des Kunden für die Ziele auf den beiden Ebenen vergeben werden. Somit dient Ebene II zur Ausbalancierung der Ziele und zur Konfliktlösung.

Ebene III mit den Lösungsprinzipien wurde mit der Intention eingeführt, die Lücke zwischen Problem- und Lösungsraum zu schließen. Es ist viel einfacher den Einfluss von Prinzipien auf Qualitätsziele zu bestimmen als direkt den Einfluss von Lösungen, denn die meisten Prinzipien wurden entwickelt mit der klaren Absicht zur Unterstützung von Qualitätszielen. Dementsprechend enthält diese Ebene Lösungsprinzipien mit bekanntem Einfluss auf Qualitätseigenschaften einer Software. Lösungsprinzipien existieren in verschiedenen Forschungsgebieten. Typischerweise sind sie in Fachbüchern zu finden. Beispiele für Lösungsprinzipien der Softwaretechnik sind lose Kopplung oder Trennung von Zuständigkeiten. Der Übergang von Problem- zu Lösungsraum manifestiert sich in Transition 2 von Ebene II zu III. Hier erfolgt die Abbildung von Qualitätszielen oder Unterzielen zu Lösungsprinzipien. Die Beziehungen drücken den Einfluss der Prinzipien auf die Ziele aus. Da dieser Übergang zum Kern jeder Entwurfsmethode gehört, können entsprechende Konzepte auch dort gefunden werden, wie z. B. die Taktiken von ADD. Die Methode QASAR erwähnt die Notwendigkeit Qualitätsziele durch funktionale (technische) Lösungen umzusetzen, bietet aber kein Konzept zur Modellierung dessen. Die Einführung von Ebene III stellt eine wesentliche Leistung des GSS dar, weil Prinzipien eine bedeutende Rolle beim Entwurf spielen. Verglichen mit bisherigen Arbeiten wie den zielorientierten RE Ansätzen bedeutet die Prinzipienebene eine signifikante Verbesserung.

Ebene IV deckt die Lösungsinstrumente verschiedener Abstraktionsstufen ab, wie z. B. Architekturmuster, -stile, aber auch Frameworks und fertige Komponenten. Die Lösungsinstrumente sind beschrieben mit Vorbedingungen für ihre Anwendbarkeit und einer Menge von Einflusswerten bezüglich der Lösungsprinzipien und Qualitätsziele. Die Einflusswerte der Lösungsinstrumente werden mit den Beziehungen der Transition 3 zwischen Ebene III und IV abgebildet.

Als ein Ergebnis bietet das GSS die Ausrichtung von Entwurfsprinzipien und funktiona-

len Lösungen an Qualitätszielen. Mit den explizit modellierten Beziehungen klassifiziert es Lösungsinstrumente entsprechend ihrem Einfluss auf die Ziele. Das Schema erleichtert die Identifikation von Zielkonflikten sowie deren Lösung durch Priorisierung und Verfeinerung sowie durch die Abbildung auf Lösungsprinzipien und -instrumente. Die Lösungsinstrumente dienen als Quelle für Vorschläge von Entwurfsalternativen während des Entscheidungsprozesses basierend auf einer quantitativen Bewertung der Instrumente entsprechend ihres Einflusses. Das GSS fördert darüber hinaus die Erstellung von Traceability Links zwischen Zielen und Entwurfsartefakten. Das Goal Solution Scheme repräsentiert die signifikanteste Neuerung und Leistung der Arbeit. Es basiert auf Prinzipien des modellbasierten Entwurfs und kombiniert Ideen des zielorientierten RE mit Architekturentwurfsprinzipien und -aktivitäten. Es wurde in mehreren Fallstudien entwickelt und angewendet und im Detail für Weiterentwickelbarkeit beschrieben. Das GSS für Weiterentwickelbarkeit stellt eine konsolidierte Sicht auf das Qualitätsziel und seine Untermerkmale dar und bestimmt den Einfluss fundamentaler Entwurfsprinzipien auf das Ziel. Außerdem wird eine Menge von Entwurfsmustern bezüglich ihres Einflusses auf Weiterentwickelbarkeit bewertet. Somit wird das abstrakte Ziel Weiterentwickelbarkeit greifbarer und kann während des Entwurfs leichter umgesetzt werden.

5 Zielorientierter Architekturentwurf

Im Rahmen der Dissertation wurde das Goal Solution Scheme in eine zielorientierte Architekturentwurfsmethode (GOAD – Goal-oriented Architectural Design) eingebettet. Die GOAD-Methode fokussiert auf eine vorwärts gerichtete, iterativ inkrementelle Entwicklung mit Backlog-Konzept und folgt der allgemeinen Aufteilung in die drei Phasen Architekturanalyse, Architektursynthese und Architekturevaluierung (vgl. [HKN+07]). Die Methode überwindet Beschränkungen existierender Methoden und Konzepte indem es die besten Teile weiterentwickelt und zusammen mit dem Goals Solution Scheme integriert.

GOAD beginnt nach der Anforderungserhebung mit einer Zielmodellierung, welche von den zielorientieren Requirements Engineering Ansätzen adaptiert wurde, und nutzt die Zielpriorisierung zur Fokussierung des Entwurfs. Auf diese Weise wird Transition 1 des GSS unterstützt. Für die Architekturanalyse wurde die Methode Globale Analyse von Hofmeister et al. weiterentwickelt und integriert. Globale Analyse nutzt sogenannte Faktorentabellen und Themenkarten zur Analyse des Einflusses von Qualitätszielen und funktionalen Anforderungen sowie zur Ermittlung geeigneter Architekturstrategien anhand von Lösungsprinzipien. Für die Architekturstrukturierung während der Architektursynthese bietet GOAD Aktivitäten, die von der Attribute Driven Design (ADD) Methode adaptiert wurden. Dabei wird auf die Ausnutzung des GSS im Detail eingegangen.

Die qualitätszielorientierte Architekturentwurfsmethode und das Goal Solution Scheme gemeinsam ermöglichen einen leichteren Übergang vom Problemraum zum Lösungsraum (Tansition 2 von Ebene II zu III und weiter zu IV des GSS). Die Methode bietet Unterstützung für die Auswahl von Entwurfslösungen unter zu Hilfenahme der Themenkarten der Globalen Analyse sowie mittels der Einflusswerte der explizit modellierten Abhängigkeiten im GSS. Dieses Vorgehen berücksichtigt explizit Architekturprinzipien

und Entwurfsbeschränkungen und es unterstützt die Entscheidungsfindung mit einem Vorrat an Lösungsinstrumenten, welche mittels Fallstudien quantitativ hinsichtlich ihres Einflusses auf das Qualitätsziel Weiterentwickelbarkeit bewertet wurden. Für die Integration von ausgewählten Lösungsinstrumenten in die Gesamtarchitektur finden die von der QASAR Methode vorgeschlagenen Architekturtransformationsschritte Anwendung. Außerdem werden in GOAD zur Detaillierung der Architektur die Softwarekategorien der Quasar-Methode [Sie04] für eine verbesserte Trennung von Zuständigkeiten durch explizite Abhängigkeiten integriert. Weiterhin erfolgt in der Arbeit eine Diskussion über die bottom-up vs. top-down Strukturierung der Architektur bezüglich der Unterstützung der Weiterentwickelbarkeit der resultierenden Software. Die Anwendbarkeit der zielorientierten Architekturentwurfsmethode wird mittels einem Fallstudienprojekt nachgewiesen.

6 Traceability-Konzept und Toolunterstützung

Inspiriert durch die existierenden Ansätze von Cleland-Huang et al. [CHSB+05] und Spanoudakis et al. [SZPMK04], wurde ein Konzept zur (halb-)automatischen und regelbasierten Erstellung von Traceability Links entwickelt, was zusätzlich die Information Retrieval Technik n-Gram-Matching anwendet. Abbildung 2 zeigt das Konzept im Überblick.

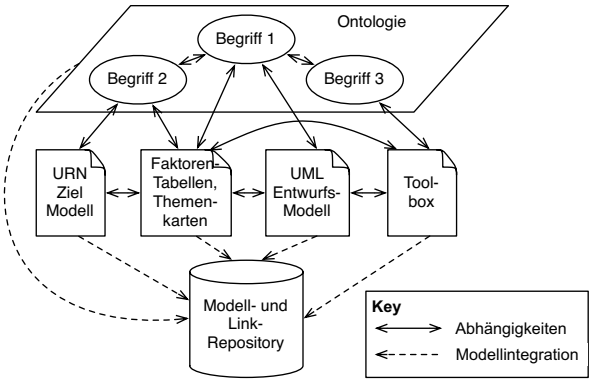


Abbildung 2: Überblick über das Traceability-Konzept

Artefakte die durch das Konzept verlinkt werden, sind a) Zielmodelle, die in URN notiert sind, b) Faktorentabellen und Themenkarten der Globalen Analyse, c) UML Modelle für die Entwurfsspezifikation, sowie d) Ontologien modelliert mit der Web Ontology Language (OWL) als semantisches Netz wichtiger Begriffe aus den verschiedenen Entwurfsphasen. Die Abhängigkeiten zwischen den Entitäten dieser Artefakte werden als Links mit einem spezifischen Typ zur Verbesserung der semantischen Ausdruckskraft der Links repräsentiert. Zu diesem Zweck wurden ein Traceability Metamodell entwickelt und Linktypen analysiert, geclustert und in einem Katalog gesammelt. Die Links werden zusammen mit den Entwurfsmodellen in einem gemeinsamen Modell-Repository abgelegt.

Das Traceability-Konzept trägt zum zerklüfteten Forschungsgebiet der Traceability-Ansätze bei, indem es verschiedene bereits existierende Ideen zu einem umfassenden Ansatz integriert und mehrere Entwurfsartefakte, die durch die GOAD-Methode verwendet werden, verknüpft, anstatt sich nur auf einen kleinen Ausschnitt zu beschränken. Insbesondere erfolgt die Verlinkung von Qualitätszielen bis hin zu Entwurfselementen wie UML-Komponenten über die konzeptuelle Lücke zwischen Problem- und Lösungsraum hinweg. Ziel des Konzeptes ist es die Entwurfsschritte zu verfolgen, die zu den Entwurfsartefakten geführt haben, um eine Auswirkungsanalyse für zukünftige evolutionäre Änderungen zu ermöglichen. Daher werden alle Modelle in einem Repository integriert und als eine Neuheit des Konzeptes mit Ontologiebegriffen verknüpft, um mehr explizite Abhängigkeiten nicht nur innerhalb sondern auch zwischen den Modellen zu ermöglichen.

Das erstellte Traceability Metamodell ist die Basis für Toolunterstützung. Ein Katalog konsolidierter Linktypen wurde ebenfalls modelliert, um ihn im Repository abzulegen. Als eine Neuheit des Konzeptes wird ein regelbasierter Ansatz für die (halb-)automatische Identifikation und Aufzeichnung von Links kombiniert mit der n-Gram-Matching Technik. Dies führt auf der einen Seite zu einem vertretbaren Aufwand für die Regeldefinition und resultiert auf der anderen Seite in einer hohen Korrektheit und Vollständigkeit der identifizierten Links. Die Regeln werden mittels XML Schema Definition (XSD) definiert und ebenfalls als Modell im Repository vorgehalten. N-Gram-Matching wird verwendet als Information Retrieval Technik zur Ermittlung von Ähnlichkeiten zwischen Modellelementbezeichnern und somit zur Erhöhung der Trefferwahrscheinlichkeit der Regeln.

Um die Anwendbarkeit des Traceabiilty Konzeptes zu zeigen, wurde ein Tool namens EMFTrace entworfen und implementiert. EMFTrace ist eine erweiterbare Plattform basierend auf Eclipse Technology. Das Tool setzt das Traceability-Konzept vollständig um und verknüpft existierende CASE-Tools, die der Erstellung der in der GOAD Methode verwendeten Artefakten dienen, wie jUCMNav für Zielmodelle, Visual Paradigm für UML-Modelle und Protegé für Ontologien. Zur Integration der Artefakte greift EMFTrace auf das Modell-Repository EMFStore zurück und bietet EMF-basierte Metamodelle für jedes Artefakt. Als wichtiger Vorteil gegenüber anderen Ansätzen ist hier die Verwendung von standardisierten Modellierungssprachen wie URN und UML zu nennen, denn die Standards sind eher stabil und selten Gegenstand von Änderungen. Die Traceability Links und Regeln werden ebenfalls mit EMF-basierten Metamodellen verwaltet. EMFTrace verwendet einen Regelinterpreter zur Anwendung der Regeln, einen Linkmanager zur Erzeugung von Traceability Links und bietet die Fähigkeit Ketten von Traceability Links zu identifizieren und zu verwalten sowie Konsistenzprüfungen zur Pflege der Links durchzuführen.

7 Evaluierung des Ansatzes

Zur Evaluierung der vorgestellten Konzepte wurde eine Fallstudie durchgeführt am Fachgebiet Neuroinformatik und Kognitive Robotik der Technischen Universität Ilmenau, Dort wurde in einem evolutionären Szenario über mehrere Jahre eine Softwareplattform für mobile Serviceroboter entwickelt, die einem Reegineering unterzogen wurde. Ein Teil der Softwareplattform, das Framework für die Kommunikation der einzelnen Softwarekompo-

nenten des Roboters, das neu entwickelt wurde, war Objekt der Fallstudie. Das Goal Solution Scheme und die GOAD-Methode half den Entwicklern in der Fallstudie die relevanten Qualitätsziele zu identifizieren, diese zu priorisieren und geeignete Lösungsinstrumente für einen weiterentwickelbaren Entwurf auszuwählen. Darüber hinaus wurden zur Evaluierung des Traceability-Konzeptes 1716 Modellelemente mit CASE-Tools erfasst. Die im Anschluss mit EMFTrace und 76 definierten Regeln erstellten Traceability Links wiesen eine Genauigkeit von 86,4% und eine Trefferquote von 84,6% auf.

8 Fazit und Ausblick

Abschließend kann gesagt werden, dass die vorgestellten Konzepte der Dissertation, das Goal Solution Scheme, die zielorientierte Architekturentwurfsmethode GOAD und das Traceability-Konzept samt Toolunterstützung durch EMFTrace, einen umfassenden und praktikablen Ansatz zur Entwicklung weiterentwickelbarer Software und zur Nachvollziehbarkeit von Entwufsentscheidungen darstellen. Der Ansatz fokussiert systematisch auf die Erfüllung der Qualitätsziele beim Entwurf, insbesondere auf Weiterentwickelbarkeit. Mittels der Traceability Links können häufige Softwareänderungen besser analysiert werden, was eine der größten Herausforderungen heutiger Softwareentwicklungsprojekte darstellt. Wo immer möglich, wurden etablierte Methoden und Konzepte aufgegriffen und kombiniert. Außerdem schlägt der integrierende Ansatz eine Brücke vom Requirements Engineering zum Architekturentwurf mit Konzepten aus beiden Forschungsgebieten. In zukünftigen Arbeiten sollte das Goal Solution Scheme für weitere Qualitätsziele und Lösungsprinzipien sowie -instrumente untersucht und der gesamte Ansatz für weitere evolutionäre Projekte hinsichtlich Reengineering angewendet werden. Das Traceability-Konzept könnte zur weiteren Verbesserung durch zusätzliche Ansätze ergänzt sowie um eine Änderungsauswirkungsanalyse erweitert werden.

Literatur

[ACC+02] G. Antoniol, G. Canfora, G. Casazza, A. De Lucia und E. Merlo. Recovering Traceability Links between Code and Documentation. *IEEE TSE*, 28(10):970–983, 2002.

[BCE08] H. Breivold, I. Crnkovic und P. Eriksson. Analyzing Software Evolvability. In *Proc. COMPSAC 2008*, Seiten 327–330. IEEE, July 2008.

[BKB02] L. J. Bass, M. Klein und F. Bachmann. Quality Attribute Design Primitives and the Attribute Driven Design Method. In *Revised Papers from the 4th International Workshop on Software Product-Family Engineering*, Seiten 169–186. Springer, 2002.

[Bos00] J. Bosch. *Design and use of software architectures: Adopting and evolving a product-line approach*. ACM Press/Addison-Wesley, 2000.

[CHSB+05] J. Cleland-Huang, R. Settimi, O. BenKhadra, E. Berezhanskaya und S. Christina. Goal-Centric Traceability for Managing Non-Functional Requirements. In *Proc. ICSE '05*, Seiten 362–371. IEEE, May 2005.

[CNYM00] L. Chung, B. A. Nixon, E. Yu und J. Mylopoulos. *Non-functional Requirements in Software Engineering*. Kluwer, 2000.

[GF94] O. C. Z. Gotel und A. C. W. Finkelstein. An Analysis of the Requirements Traceability Problem. In *Proceedings of the First International Conference on Requirements Engineering, Colorado Springs, CO, USA*, Seiten 94–101. IEEE, April 1994.

[HA07] N. Harrison und P. Avgeriou. Pattern-Driven Architectural Partitioning: Balancing Functional and Non-functional Requirements. In *Second International Conference on Digital Telecommunications, 2007 (ICDT '07)*, Seiten 21–26. IEEE, July 2007.

[HKN+07] C. Hofmeister, P. Kruchten, R. L. Nord, H. Obbink, A. Ran und P. America. A general model of software architecture design derived from five industrial approaches. *Journal of Systems and Software*, 80(1):106–126, January 2007.

[HNS00] C. Hofmeister, R. Nord und D. Soni. *Applied Software Architecture*. Addison-Wesley Longman, Boston, MA, USA, 2000.

[ITU08] ITU-T. Recommendation ITU-T Z.151 User requirements notation (URN) – Language definition, Nov 2008.

[MGP08] P. Mäder, O. Gotel und I. Philippow. Rule-Based Maintenance of Post-Requirements Traceability Relations. In *Proc. RE '08*, Seiten 23–32, USA, 2008. IEEE.

[RB09] M. Riebisch und S. Bode. Software-Evolvability. *Informatik-Spektrum*, 32(4):339–343, August 2009.

[RJ01] B. Ramesh und M. Jarke. Toward Reference Models for Requirements Traceability. *IEEE Trans. Softw. Eng.*, 27(1):58–93, 2001.

[RLL98] D. Rowe, J. Leaney und D. Lowe. Defining systems architecture evolvability - a taxonomy of change. In *Proc. ECBS'98*, Seiten 45–52, Jerusalem, Israel, 1998. IEEE.

[Sie04] Johannes Siedersleben. *Moderne Software-Architektur: Umsichtig planen, robust bauen mit Quasar*. dpunkt.verlag, Heidelberg, Germany, 2004.

[SZPMK04] G. Spanoudakis, A. Zisman, E. Perez-Minana und P. Krause. Rule-Based Generation of Requirements Traceability Relations. *JSS*, 72(2):105–127, 2004.

[Yu95] E. Siu-Kwong Yu. *Modelling Strategic Relationships for Process Reengineering*. Dissertation, University of Toronto, Toronto, Ontario, Canada, 1995.

Stephan Bode wurde geboren am 5. Februar 1983 in Mühlhausen/Thüringen. Von der Technischen Universität Ilmenau erhielt er 2008 einen Abschluss als Diplom-Informatiker. Danach begann er seine Dissertation an der TU Ilmenau, gefördert durch ein Graduiertenstipendium des Landes Thüringen. Außerdem war er als wissenschaftlicher Mitarbeiter an der Universität tätig. Seine Promotion zum Doktor-Ingenieur im September 2011 erfolgte mit dem Prädikat summa cum laude. Seine Forschungsinteressen betreffen Softwarearchitekturentwurfsmethoden, Software-Evolution, Softwarequalität sowie Traceability. Er ist Autor mehrerer Veröffentlichungen auf nationalen und internationalen Workshops und Konferenzen. Seit Oktober 2011 arbeitet er als Technical Consultant bei Senacor Technologies AG.

Generierung von Prozessoren aus Instruktionssatzbeschreibungen

Ralf Dreesen
Institut für Informatik, Universität Paderborn
rdreesen@uni-paderborn.de

Abstract: Die Automatisierung der Prozessorentwicklung ist ein seit Langem untersuchtes Thema, das durch den zunehmenden Einsatz anwendungsspezifischer Prozessoren (ASIPs) in Ein-Chip-Systemen (SoCs) erheblich an Bedeutung gewonnen hat. Bei bisherigen Ansätzen werden Prozessoren auf mikroarchitektonischer Ebene beschrieben, wodurch ein Entwickler viele komplexe und fehleranfällige Aspekte definieren muss. Die Dissertation [Dre11] verfolgt einen bisher kaum untersuchten Ansatz, bei dem nur der Instruktionssatz eines Prozessors modelliert wird. Die Mikroarchitektur wird vollständig und automatisch durch neu entworfene Methoden erzeugt. Durch Variation eines Generatorparameters wird so ein Satz kompatibler Prozessoren mit verschiedenen physikalischen und dynamischen Eigenschaften erzeugt. Abhängig vom Einsatzgebiet kann aus diesem Entwurfsraum eine passende Implementierung ausgewählt werden.

1 Einleitung und Motivation

Zur Entwicklung anwendungsspezifischer Prozessoren werden heute Werkzeuge eingesetzt, die entweder eine Hardwarebeschreibung oder eine Prozessorbeschreibung verlangen. Bei einer *Hardwarebeschreibung* in Sprachen wie VHDL oder Verilog wird ein Prozessor auf mikroarchitektonischer Ebene definiert. Der Anwender muss sowohl Funktionseinheiten, als auch die Instruktions-Pipeline und ihre Kontrolle beschreiben. Letzteres ist dabei sehr zeitaufwendig und fehleranfällig.

Einzelne Instruktionen sind auf dieser Ebene der Beschreibung nicht mehr erkennbar. Die Semantik aller Instruktionen ist vereint und auf eine Vielzahl von Ressourcen verteilt. Eine Additionsinstruktion trägt zum Beispiel zum Decodierer, zur arithmetisch-logischen Einheit (ALU), zum Forwarding und zum Interlocking bei. Auf der anderen Seite vereint die ALU die Semantik aller arithmetischen und logischen Instruktionen. Der Instruktionssatz eines so beschriebenen anwendungsspezifischen Prozessors kann daher nur schwer nachträglich geändert werden.

Um diesen Problemen zu begegnen, bieten *Prozessorbeschreibungssprachen* wie zum Beispiel nMl [PLGG08] und Lisa [Inc08] spezielle Konzepte, welche helfen die Spezifikation eines Prozessors zu strukturieren und zu vereinfachen. Ein Prozessor wird in diesen Sprachen aber nach wie vor auf mikroarchitektonischer Ebene beschrieben. Da-

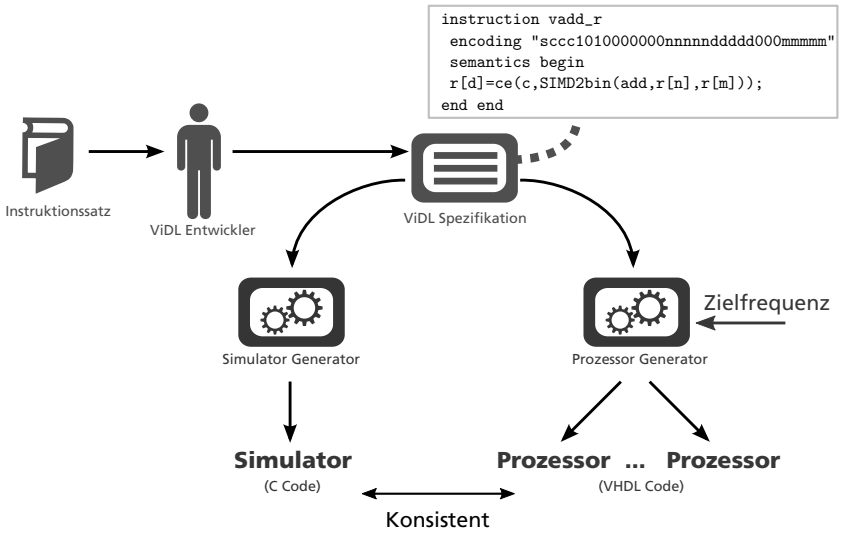


Abbildung 1: Überblick des Generator-Systems.

mit werden dynamische und physikalische Eigenschaften des Prozessors schon zu einem sehr frühen Zeitpunkt festgelegt. Eine späte Exploration dieses Entwurfsraums ist damit stark eingeschränkt. Zudem muss der Anwender selber Daten-, Kontroll- und Ressourcenkonflikte in der Instruktions-Pipeline auflösen. Diese Aufgabe ist fehleranfällig und ein aufwendiges Testen des Prozessors daher ratsam.

Im Gegensatz zu Hardwarebeschreibungen und Prozessorbeschreibungen modelliert eine *Instruktionssatzbeschreibung* in ViDL ausschließlich den Instruktionssatz eines Prozessors, nicht den Prozessor an sich. Dieser wird stattdessen, wie in Abbildung 1 dargestellt, generiert. Alle Aspekte der Mikroarchitektur werden dabei automatisch aus der Semantik des Instruktionssatzes und einer vorgegebenen Zielfrequenz des Prozessors hergeleitet. Dazu wurden im Rahmen der Arbeit eine Reihe von Analysen, Transformationen und Optimierungen entwickelt. Weitere Methoden wurden entworfen, um einen schnellen Instruktionssatzsimulator aus derselben Spezifikation zu erzeugen. Für die Sprache ViDL wurden domänenspezifische Sprachkonzepte entworfen, um Instruktionssätze prägnant und verständlich zu beschreiben.

2 Die Spezifikationssprache ViDL

Beim Entwurf der Sprache ViDL [Dre12b] wurden drei Ziele verfolgt. Erstes sollen typische Entwurfsabläufe für anwendungsspezifische Prozessoren unterstützt werden. Insbesondere die Exploration des Instruktionssatz-Entwurfsraums, sowie die Erweiterung

und Änderung bestehender Instruktionssätze ist von Relevanz. Zweitens sollen bewährte Richtlinien des Sprachentwurfs [Wat04, Seb09] befolgt werden, um eine hohe Sprachqualität zu erreichen. Eine Sprache die diese Richtlinien befolgt verringert typischerweise die Entwicklungszeit sowie die Fehleranfälligkeit und erhöht die Wartbarkeit. Drittens soll die Sprache mächtig genug sein, um realistische Instruktionssätze kompakt zu beschreiben. Eine Sprache, die hingegen nur eine kleine Klasse künstlicher Instruktionssätze abdeckt, würde sicherlich nicht vonseiten der Industrie akzeptiert. Um diese Ziele zu erreichen, wurde ein kleiner Satz kombinierbarer Sprachkonzepte so entworfen, dass sich auch die Eigenarten realistischer Instruktionssätze elegant beschreiben lassen. Diese Konzepte und Prinzipien werden im Folgenden kurz vorgestellt.

Die Semantik von Instruktionen wird in ViDL in einem Dialekt der funktionalen Sprache SML definiert. Eine ViDL Spezifikation ist dadurch frei von Seiteneffekten, was zum einen die Verständlichkeit erhöht und zum anderen die Fehleranfälligkeit verringert. ViDL unterstützt viele funktionale Konzepte, wie Funktionen höherer Ordnung (Funkionale), Lambda Ausdrücke, Polymorphie, Tupel und Closures. Die Konzepte ermöglichen einen hohen Grad an Wiederverwendbarkeit. So kann ein SIMD-Berechnungsmuster einfach als Funktional definiert werden. Die eigentliche arithmetische Operation einer konkreten SIMD-Instruktion wird dann wie in Abbildung 1 zu sehen in Form einer benannten Funktion oder eines Lambda Ausdrucks an das Funktional übergeben. Im Gegensatz zu anderen Ansätzen, kann das Prinzip SIMD also mit Sprachmitteln formuliert werden, was zu einem schlanken Sprachdesign beiträgt und Erweiterungen zulässt. Funktionen und Funktionale sind in der Regel polymorph, d.h. sie können auf beliebige n-Bit Daten angewendet werden, wodurch die Wiederverwendbarkeit weiter gesteigert wird. Zum Beispiel kann ein einmalig definierter Adressierungsmodus in mehreren Instruktionssätzen unterschiedlicher Bitbreite wiederverwendet werden.

Instruktionssätze verwenden häufig Strukturen wie virtuelle Adressräume, Sub-Wort Zugriffe auf Speicher und dedizierte Register für verschiedene Prozessormodi. Solche Strukturen sind generell Sichten auf physikalische Speicherelemente und I/O Schnittstellen. In ViDL werden solche Sichten durch architektonische Schnittstellen modelliert. Eine architektonische Schnittstelle beschreibt einen virtuellen Speicher mit einer definierten Anzahl an Elementen einer bestimmten Bitbreite. Die Relation zwischen virtuellen und physikalischen Elementen wird durch Abbildungen für Lese- und Schreibzugriffe spezifiziert. Die Relation selbst kann dynamisch sein, also zum Beispiel vom Prozessormodus abhängen. Architektonische Schnittstellen sind allgemein anwendbar und decken viele spezielle Konzepte anderer Sprachen ab. So zum Beispiel die Byte-Ordnung bei Speicherzugriffen, die sich bei manchen Ansätzen durch eine spezielle Direktive oder gar nicht definieren lässt.

Die meisten Instruktionssätze beinhalten Instruktionen, die Speicherelemente teilweise oder bedingt schreiben. Zum Beispiel ändert eine arithmetische Instruktion in der Regel nur bestimmte Bits des Status-Registers. Ein bedingter Sprungbefehl kann als ein bedingter Schreibzugriff auf den Programmzähler aufgefasst werden. Ähnliches gilt für die bedingte Befehlsausführung bei ARM und CoreVA. Um solches Verhalten einfach und

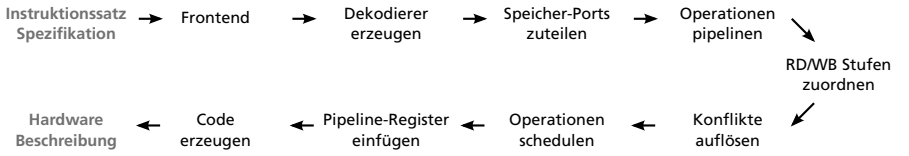


Abbildung 2: Überblick der Methoden zur Erzeugung eines Prozessors.

effektiv zu beschreiben, integriert ViDL eine dreiwertige Logik, ähnlich der Tri-State Logik für Buszugriffe. Anstelle des hochohmigen Zustandes "Z" drückt der Zustand "ϵ" jedoch aus, dass ein Bit eines Speichers unverändert bleibt. Epsilon-Logik harmoniert sehr gut mit ViDLs funktionalen Konzepten und kann sowohl in Funktionen, als auch in architektonischen Schnittstellen verwendet werden. Die Definition der bedingten Befehlsausführung bei ARM kann zum Beispiel in einer Funktion gekapselt werden. Beim CoreVA Instruktionssatz kann die bedingte Ausführung sogar auf SIMD-Ebene heruntergebrochen werden — eine Eigenschaft, die sich mit anderen Ansätzen nicht ausdrücken lässt.

ViDL operiert ausschließlich auf Bitketten, deren Breite nicht explizit durch den Entwickler definiert wird. Statt dessen wird die Bitbreite jeder Operation implizit durch den Generator hergeleitet. Dadurch wird die Spezifikation wesentlich vereinfacht und die Wartbarkeit erhöht. Zudem erhöht sich durch Polymorphie der Grad der Wiederverwendbarkeit. Um Bitbreiten herzuleiten, werden sie als Typen modelliert [DTK11]. ViDLs Typverband definiert dazu mehrere parametrisch polymorphe Typen, die in Subtyp-Relation stehen. Der Parameter ist dabei eine beliebige natürliche Zahl, die die Bitbreite beschreibt. Der Zusammenhang zwischen den Bitbreiten der Parameter und des Ergebnisses jeder Operation wird durch eine polymorphe Signatur beschrieben. Diese Form ist wesentlich prägnanter und verständlicher als zum Beispiel der Inferenz-Algorithmus der Hardware-Beschreibungssprache Verilog [Soc01]. Durch diese Modellierung können im Generator bewährte Methoden der Typinferenz angewendet werden, um die Parameter der polymorphen Typen, also die Bitbreiten, zu bestimmen. Widersprüchliche oder mehrdeutige Ausdrücke in ViDL führen dabei dank der statischen Typisierung zur Generierungszeit zu einem Typfehler. Dadurch wurde eine Mehrdeutigkeit im Pseudocode des ARM Handbuches aufgedeckt.

3 Methoden zur Prozessorgenerierung

ViDL abstrahiert stark von der Implementierungsebene eines Prozessors. Ein Prozessorgenerator bedarf daher vieler Analysen und Transformationen, die in diesem Abschnitt kurz vorgestellt werden.

Das Frontend des Generators führt die üblichen syntaktischen und statisch-semantischen

Analysen, wie Namensanalyse und Typanalyse durch. Zudem werden dort die meisten Sprachkonzepte in Datenfluss transformiert. Das Ergebnis ist eine Zwischendarstellung in Form eines Datenflussgraphen, der die Semantik aller Instruktionen vereint. Dieser Graph wird schrittweise durch viele Methoden transformiert, bis schließlich die Hardwarebeschreibung einer Prozessorimplementierung erzeugt wird. Abbildung 2 zeigt einen Überblick dieser Methoden, von denen im Folgenden einige kurz erläutert werden.

Eine Registerbank eines Prozessors hat in der Regel mehrere Lese- und Schreib-Ports, um gleichzeitig auf die Inhalte der Register zuzugreifen. Weil ViDL strikt von der Mikroarchitektur abstrahiert, wird dieser Aspekt nicht durch den Entwickler beschrieben, sondern durch den Generator hergeleitet. Dabei müssen den Zugriffen aus der Spezifikation so Ports zugeteilt werden, dass keine Ressourcenkonflikte bestehen und die Anzahl der Ports minimiert wird. Die Minimierung ist wichtig, weil Ports eine “teuere” Ressource in einer Hardware-Implementierung darstellen. Die in der Arbeit beschriebene Methode reduziert das Problem auf Graphfärbung und hat für alle evaluierten Instruktionssätze stets die minimale Anzahl an Ports erzeugt.

Um eine hohe Taktfrequenz zu erreichen, muss der Datenpfad eines Prozessors gepipelined werden. In der Dissertation werden zu diesem Zweck mehrere Methoden beschrieben, die auch erfolgreich implementiert wurden. Die Struktur der erzeugten Pipeline hängt dabei von drei Faktoren ab: dem Datenflussgraphen, der vorgegebenen Zielfrequenz und der Zieltechnologie. Letztere wird dabei durch eine Datenbank ihrer abgeschätzten Gatterlaufzeiten beschrieben. Bisher muss ein Entwickler all diese Faktoren berücksichtigen, um die Aspekte der Mikroarchitektur zu einem frühen Zeitpunkt selbst festzulegen. Diese Entscheidungen verlangen viel Erfahrung und eine gute Abschätzung von Signallaufzeiten. Die folgenden Methoden automatisieren diesen Prozess. Die Pipeline wird so konstruiert, dass die Zielfrequenz voraussichtlich erreicht wird, ansonsten aber die Pipeline-Tiefe und Instruktionslatenzen weitestgehend minimiert werden. Dadurch werden Ressourcen gespart, die andernfalls zu einem erhöhten Flächen- und Energiebedarf führen. Durch geringe Instruktionslatenzen wird der Instruktionsdurchsatz erhöht, bzw. die Anzahl der Zyklen pro Instruktion (CPI) verringert.

In einem ersten Schritt werden die Read und Write-Back Stufen der Pipeline separat für jedes Speicherelement bestimmt, also zum Beispiel für die freie (GP) Registerbank, das Status-Register und den Hauptspeicher. Die Methode berücksichtigt dabei die Gatterlaufzeiten der Operationen auf dem Datenpfad und die Zielfrequenz. Um unnötige Ressourcen für Bypässe zu vermeiden, versucht die Methode die Distanz zwischen Read und Write-Back Stufen zu minimieren. Für übliche Instruktionssätze wird nur ein Bypass für die GP Registerbank benötigt. Bei einer geringen Zielfrequenz fällt selbst dieser in der Regel weg. Die Mikroarchitektur wird also genau auf die Instruktionssemantik und die Zielfrequenz zugeschnitten.

Für den Fall, dass die Write-Back Stufe nicht der Read Stufe folgt, können Datenkonflikte auftreten. Diese Konflikte werden im generierten Prozessor wenn möglich durch Forwarding und andernfalls durch Interlocking behoben. Die Methode zur Erzeugung des Forwardings analysiert dabei, unter welchen Umständen ein Ergebnis bereits zu einem

früheren Zeitpunkt bekannt ist. Sie basiert auf einer Analyse kritischer Pfade im Datenflussgraphen, einer Analyse partiell-definierter Ausdrücke und einem Satz von Termersetzungregeln, um den Datenflussgraphen zu normalisieren. Die Methode liefert für jede Stufe zwischen Read und Write-Back das frühzeitige Ergebnis und eine Signalleitung, mit der das Interlocking gesteuert wird. Für den Sonderfall, dass ein Ergebnis nie frühzeitig bekannt ist, entfällt der entsprechende Bypass automatisch.

In ähnlicher Weise wird bestimmt, wann ein Kontrollflusswechsel durch einen (bedingten) Sprung ausgelöst wird. Zudem werden Steuersignale erzeugt, um spekulativ ausgeführte Instruktionen im Fall einer falschen Sprungvorhersage abzubrechen. Dabei ist sichergestellt, dass solche Instruktionen stets abgebrochen werden können, also noch keinen Einfluss auf den Prozessorzustand genommen haben.

Die vorgestellten Methoden kapseln das Expertenwissen eines Schaltungstechnikers. Der Entwurfsablauf ist dadurch vollständig automatisiert, ein Entwickler braucht also keinen Aspekt der Mikroarchitektur beisteuern. Wie die Auswertung im nächsten Abschnitt zeigt, treffen die Methoden gute Entscheidungen, die sich in den physikalischen und dynamischen Eigenschaften des erzeugten Prozessors widerspiegeln.

4 Evaluation

Zur Evaluation der Sprache ViDL wurden vier praktische Instruktionssätze (ARM, MIPS, Power und CoreVA) und zwei akademische Instruktionssätze (OISC und SRC) spezifiziert. Dank ViDLs mächtiger Konzepte, wie Epsilon-Logik, architektonischer Schnittstellen und Typinferenz, konnten die Instruktionssätze weitestgehend beschrieben werden. Das schließt Instruktionen mit ungewöhnlichen Bitbreiten, verzögerten Effekten und anspruchsvollen Adressierungsmodi ein. Durch die starke Abstraktion der Sprache und bewährte funktionale Konzepte zur Kapselung gemeinsamen Verhaltens, konnten die Instruktionssätze effizient beschrieben werden. So lag der Zeitaufwand eines erfahrenen Entwicklers zwischen 90 Minuten für SRC und einem Monat für ARM. Unerfahrene Benutzer (Studenten) haben den CoreVA und Power Instruktionssatz in 2 bzw. 3 Monaten spezifiziert. Dank der einfachen Struktur der Sprache, konnten sie sich schnell einarbeiten.

ViDL eignet sich gut zur Exploration des Entwurfsraumes eines Instruktionssatzes. Zur Evaluation wurden Instruktionssätze in kurzer Zeit erweitert und geändert. Zum Beispiel wurde eine bestehende 32-Bit ARM Spezifikation nachträglich in eine generische n-Bit Spezifikation überführt, wobei n ein statischer Parameter ist. Zwei Stellen in der Spezifikation, die dabei erweitert werden mussten, wurden durch die Typanalyse des Generators automatisch identifiziert.

Der ebenfalls generierte Instruktionssatzsimulator erreicht für übliche Instruktionssätze im Mittel eine Geschwindigkeit von 60 Mips auf einem typischen 3 GHz Rechner. Selbst der Simulator eines 256-Bit breiten ARM Instruktionssatzes erreicht noch eine Geschwin-

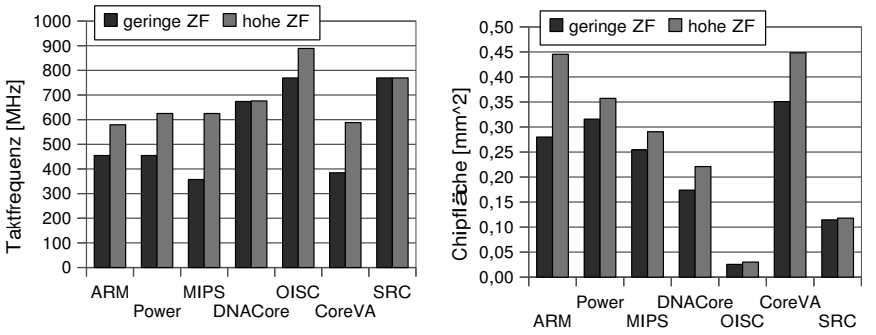


Abbildung 3: Taktfrequenz und Flächenbedarf generierter Prozessoren für verschiedene Zielfrequenzen (ZF).

digkeit von ca. 20 Mips. Die Evaluation hat gezeigt, dass die Optimierungen im Generator einen erheblichen Einfluss auf die Geschwindigkeit des Simulators haben.

Um nachzuweisen, dass die Methoden des Generators gute Prozessoren erzeugen, wurden für 13 Instruktionssätze (einschließlich Varianten) insgesamt 83 Prozessoren generiert. Für jeden Instruktionssatz wurden ca. 6 Prozessoren für unterschiedliche Zielfrequenzen erzeugt. Die Prozessoren werden automatisch generiert, also ohne Beitrag des Entwicklers. Die so erzeugten Prozessoren in Form von VHDL-Code wurden für eine Standardzellentechnologie[1] synthetisiert[2].

In Abbildung 3 ist die erreichbare Taktfrequenz und der Flächenbedarf generierter Prozessoren für verschiedene Instruktionssätze und zwei Zielfrequenzen (ZF) zu sehen. Mit steigender Zielfrequenz wird der Datenpfad auf eine zunehmende Anzahl an Pipeline-Stufen verteilt. Der kritische Pfad wird verringert und die erreichbare Frequenz des Prozessors steigt. Allerdings erhöht sich damit auch der Bedarf an Ressourcen (Chipfläche), zum Beispiel für Bypässe und Pipeline-Register. Es fällt auf, dass die DNACore und SRC Prozessoren auch für eine geringe Zielfrequenz schon eine hohe Taktrate erzielen. Beide Instruktionssätze enthalten keine Multiplikation, wodurch auch bei einer sehr einfachen Pipeline Struktur der kritische Pfad eine geringe Signallaufzeit hat.

Die Syntheseergebnisse liegen in etwa in der Größenordnung einer Handimplementierung. Für diesen Vergleich wurden generierte CoreVA Prozessoren mit einer optimierten VHDL Implementierung eines erfahrenen Schaltungstechnikers verglichen. Bei gleicher Taktfrequenz (ca. 350 MHz) hat der generierte Prozessor eine ca. 40% höhere Leistungsaufnahme bei doppeltem Flächenbedarf. Der Entwicklungsaufwand eines unerfahrenen Benutzers liegt mit zwei Monaten für ViDL jedoch deutlich unter dem Aufwand von einem Jahr für die VHDL Implementierung. Hinzu kommt, dass die Konfliktauflösung des

[1] 65 nm ST Microelectronics Low Power für Worst-Case ($1.1V$; $125°C$)
[2] Cadence RTL Compiler

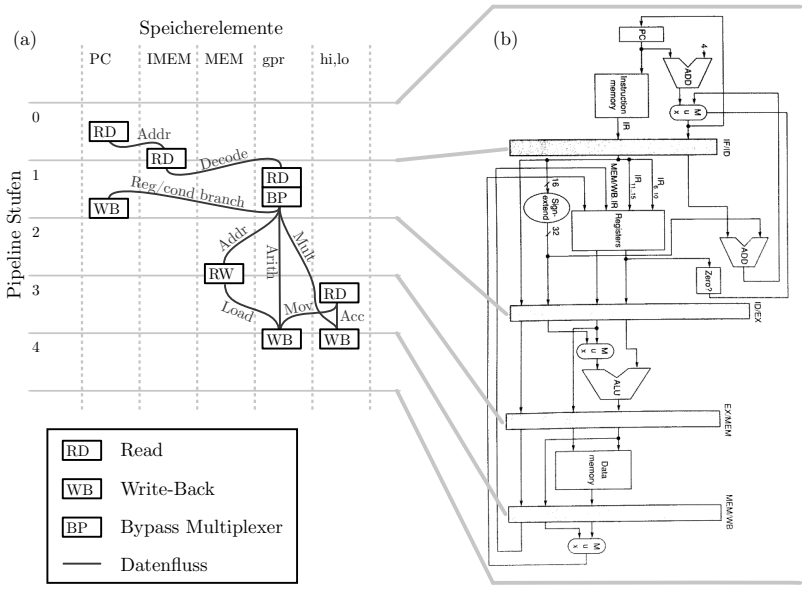


Abbildung 4: Vergleich der (a) generierten MIPS-Pipeline für eine Zielfrequenz von 400 MHz mit (b) der Pipeline aus dem Buch *Computer Architecture* [HP06].

generierten Prozessors nicht getestet[3] werden braucht, weil sie automatisch erzeugt wurde.

Die Struktur generierter Pipelines wird in der Dissertation am Beispiel des MIPS Instruktionssatzes eingehend untersucht. Abhängig von der Zielfrequenz ergeben sich dabei Pipeline-Tiefen zwischen zwei und sieben Stufen, wobei ein Prozessor mit zwei Pipelinestufen einem Prozessor ohne Pipelining entspricht. Wie in Abbildung 4 zu sehen ist, entspricht für eine Zielfrequenz von 400 MHz die automatisch hergeleitete Pipeline jener, die durch Hennessy und Patterson [HP06] beschrieben wird. Folglich ist das dynamische Verhalten (CPI) gleich. Wie üblich, haben nur Lade-Instruktionen und einige Instruktionen im Umfeld der Multiplikation eine Latenz von einem Zyklus. Durch eine geschickte Organisation der Pipeline kommen Multiply-Accumulate (MAC) Instruktionen sogar ohne Latenz aus. Auch diese Eigenschaft wird automatisch erzielt, ohne dass ein Entwickler Aspekte der Mikroarchitektur vorgeben muss. Falsch vorhergesagte Sprünge erzeugen in der hergeleiteten Pipeline zwei Strafzyklen, korrekt vorhergesagte unterbrechen die Ausführung hingegen nicht. Insgesamt setzt der generierte Prozessor die hohe Taktfrequenz also auch in einen hohen Instruktionsdurchsatz um. Für geringere Zielfrequenzen werden kürzere Pipelines erzeugt und das CPI verbessert sich entsprechend. So können für eine hergeleitete dreistufige MIPS-Pipeline keine Konflikte mehr auftreten. Alle In-

[3] Unter der Annahme, dass der Prozessorgenerator korrekt ist.

struktionen haben also ein CPI von 1 und falsch vorhergesagte Sprünge verursachen keine Strafzyklen. Zudem werden bei dieser Pipeline automatisch alle Bypässe und Steuerleitungen weggelassen, wodurch sich die Implementierung erheblich vereinfacht.

Um die Stärken des Systems zu demonstrieren, wurde im Rahmen der Dissertation der DNACore [Dre12a] Prozessor entwickelt. Der DNACore Instruktionssatz ist eine Erweiterung des MIPS Instruktionssatzes zur Beschleunigung des Smith-Waterman Algorithmus. Dieser Algorithmus wird in der Bioinformatik eingesetzt, um DNA, RNA und Proteinsequenzen zu vergleichen. Der Algorithmus berechnet eine sogenannte Scoring-Matrix, wodurch er eine hohe Komplexität hat, jedoch inhärent datenparallel ist. Diese Parallelität wird durch einen Satz anwendungsspezifischer SIMD-Instruktionen ausgenutzt. Zudem werden viele Speicherzugriffe durch einen kleinen Satz interner Register vermieden. Zusammen bilden SIMD-Instruktionen und interne Register konzeptuell ein Systolisches-Array, durch das ein Streifen der Matrix gepumpt wird. Die Spezifikation der Erweiterung in ViDL hat nur einen halben Tag gedauert. In der Praxis wird eine Auslastung der SIMD-Einheit von 97% erreicht, d.h. es werden in einem Takt durchschnittlich 3.8 Elemente der Scoring-Matrix berechnet. Zusammen mit den Syntheseergebnissen des generierten Prozessors ergibt sich so eine Geschwindigkeit von $2.6\,\mathrm{GCUPS}$[4] bzw. eine Energieeffizienz von $58\,\mathrm{GCUPS}/W$. Letztere liegt erheblich über den Ergebnissen alternativer Ansätze.

5 Zusammenfassung und Ausblick

In der Dissertation wurde gezeigt, wie effiziente Prozessoren aus einer abstrakten Instruktionssatzspezifikation automatisch generiert werden können. Im Gegensatz zu bestehenden Ansätzen abstrahiert die Sprache ViDL konsequent von der mikroarchitektonischen Ebene. Dadurch wird zum einen die Entwicklung erheblich vereinfacht und zum anderen die Fehleranfälligkeit in einem sensiblen Bereich deutlich gesenkt. Zudem können konsistente Simulatoren und Prozessoren mit sehr unterschiedlichen physikalischen und dynamischen Eigenschaften ohne zusätzlichen Aufwand aus derselben Instruktionssatzspezifikation generiert werden. Dies ist bei bestehenden Ansätzen in dieser Form nicht möglich. Zur Herleitung der Mikroarchitektur wurde Expertenwissen aus der Schaltungstechnik generalisiert und in Methoden gekapselt. Die Wirksamkeit der Methoden wurde an praktischen Instruktionssätzen belegt.

Es ist geplant, ViDL in Zukunft in zwei Richtungen voranzutreiben. Zum einen soll der Prozessorgenerator um weitere Optimierungen ergänzt werden. Vor allem Hardware-Sharing birgt hier ein bedeutendes Optimierungspotenzial. Zum anderen sollen Generatoren für Übersetzerwerkzeuge entwickelt werden. Ich habe bereits solche Generatoren für eine andere Beschreibungssprache mitentwickelt und schätze, dass sich einige der Konzepte gut übertragen lassen. Ein Entwurf für eine Erweiterung ViDLs wurde bereits fertiggestellt.

[4] Billion Cell Updates per Second

Literatur

[Dre11] Ralf Dreesen. *Generating Processors from Specifications of Instruction Sets*. Dissertation, University of Paderborn, Germany, 2011.

[Dre12a] Ralf Dreesen. DNACore: An Application Specific Processor for the Smith-Waterman Algorithm. In *39th International Symposium on Computer Architecture (ISCA 2012)*, 2012. (submitted).

[Dre12b] Ralf Dreesen. ViDL: A Versatile ISA Description Language. In *19th Annual IEEE International Conference and Workshops on the Engineering of Computer Based Systems (ECBS-19)*, April 2012. (accepted for publication).

[DTK11] Ralf Dreesen, Michael Thies und Uwe Kastens. Type Analysis on Bitstring Expressions. In *Proceedings of the 9th Workshop on Optimizations for DSP and Embedded Systems (ODES-9)*, April 2011.

[HP06] John Hennessy und David Patterson. *Computer Architecture - A Quantitative Approach*. Morgan Kaufmann, 2006.

[Inc08] CoWare Inc. *LISA Language Reference Manual*, 2008.

[PLGG08] Johan Van Praet, Dirk Lanneer, Werner Geurts und Gert Goossens. nML: A Structural Processor Modeling Language for Retargetable Compilation and ASIP Design. In Prabhat Mishra und Nikil Dutt, Hrsg., *Processor Description Languages*, Seiten 65–93. Morgan Kaufmann, 2008.

[Seb09] Robert W. Sebesta. *Concepts of Programming Languages*. Addison-Wesley Publishing Company, USA, 9th. Auflage, 2009.

[Soc01] IEEE Computer Society. *IEEE Std 1364-2001 - IEEE Standard Verilog Hardware Description Language*. The Institute of Electrical and Electronics Engineers, Inc, 2001.

[Wat04] David A. Watt. *Programming Language Design Concepts*. John Wiley & Sons, 2004.

Ralf Dreesen wurde am 20. Oktober 1980 in Soest geboren. Nach dem Abitur studierte er von 2001 bis 2006 in Paderborn Informatik mit Nebenfach Elektrotechnik. Im Anschluss, arbeitete er als Doktorand am Lehrstuhl für Programmiersprachen und Übersetzer von Professor Kastens, wo er 2011 mit Auszeichnung promovierte. Neben der Arbeit an seiner Dissertation entwickelte er in dieser Zeit im Rahmen einer Kooperation mit Infineon eine vollständige Compiler-Werkzeugkette für den CoreVA Mobilprozessor, die heute intensiv genutzt wird. Seine Forschungsinteressen liegen sowohl im Bereich des Übersetzerbaus, als auch im Gebiet der Schaltungstechnik.

Inclusion of Pattern Languages and Related Problems

Dominik D. Freydenberger

Goethe-Universität, Frankfurt am Main
freydenberger@em.uni-frankfurt.de

Abstract: Patternsprachen sind ein einfacher und eleganter Mechanismus zur Beschreibung von Sprachen, deren Wörter über Wiederholungen definiert sind. Trotz dieser Einfachheit sind viele der kanonischen Fragestellungen für Patternsprachen überraschend schwer zu lösen. Die vorliegende Arbeit befasst sich mit verschiedenen Aspekten des Inklusionsproblems für Patternsprachen. Neben Beweisen zur Unentscheidbarkeit dieses Problems, selbst für verschiedene stark eingeschränkte Unterklassen, werden die Resultate auf *regex*, eine in modernen Programmiersprachen weit verbreitete Erweiterung der regulären Ausdrücke übertragen. Ein weiterer Schwerpunkt der Untersuchungen sind die Existenz und Berechnung deskriptiver Pattern, welche inklusionsminimale Verallgemeinerungen beliebiger Sprachen durch Patternsprachen darstellen.

1 Pattern und ihre Sprachen

Pattern, d. h. endliche Wörter aus Variablen und Terminalsymbolen, stellen eine kompakte, elegante und natürliche Methode dar, gewisse Sprachen mit Wörtern mit Wiederholungen zu repräsentieren. Ein Pattern erzeugt ein Wort durch eine Substitution, die alle Variablen im Pattern durch beliebige endliche Wörter über einem festen Terminalalphabet ersetzt. Die *Patternsprache* eines Pattern ist somit die Menge aller Wörter, die durch die Substitution aus dem Pattern gebildet werden können; etwas formaler ausgedrückt ist eine Patternsprache also die Menge aller Bilder des Pattern unter beliebigen terminalerhaltenden Homomorphismen. Wenn wir beispielsweise das Pattern $\alpha{:=}x_1\,\mathtt{a}\,x_2\,\mathtt{b}\,x_1$ (mit Variablen x_1, x_2 und Terminalsymbolen $\mathtt{a}, \mathtt{b}$) betrachten, liegen folglich (unter anderem) die Wörter $w_1{:=}\mathtt{aabbba}$, $w_2{:=}\mathtt{abababab}$ und $w_3{:=}\mathtt{aaabaa}$ in der von α erzeugten Patternsprache $L(\alpha)$, wogegen die Beispielwörter $w_4{:=}\mathtt{ba}$, $w_5{:=}\mathtt{babbba}$ und $w_6{:=}\mathtt{abba}$ nicht von α erzeugt werden können.

Da jedes Vorkommen einer mehrfach auftretenden Variablen durch eine (feste) Substitution stets gleich ersetzt werden muss, können Patternsprachen als eine einfache Formalisierung des Wiederholungsoperators verstanden werden. Die Untersuchung von unvermeidbaren Pattern in Wörtern lässt sich bis zu [Thu06] zurückverfolgen und zählt mittlerweile zu den etablierten Teilgebieten der Wortkombinatorik (siehe [Cas02]), während die explizite Verwendung von Pattern als Sprachgeneratoren von Angluin in [Ang80] eingeführt wurden. In Angluins Definition ist es nicht erlaubt, Variablen löschend zu ersetzen, daher wird diese Klasse von Patternsprachen in der Literatur im Allgemeinen als

NE-Patternsprachen bezeichnet („NE“ steht in diesem Fall für „non-erasing“). Die NE-Patternsprache eines Patterns α über einem Alphabet Σ bezeichnen wir mit $L_{\mathrm{NE},\Sigma}(\alpha)$.

Ebenfalls häufig betrachtet werden die von [Shi82] eingeführten sogenannten *E-Patternsprachen* („E“ wie „erasing“ oder „extended“); bei dieser Sprachklasse ist die löschende Substitution gestattet. (Analog zu $L_{\mathrm{NE},\Sigma}(\alpha)$ verwenden wir die Bezeichnung $L_{\mathrm{E},\Sigma}(\alpha)$ für die E-Patternsprache von α.) Im obigen Beispiel ist somit das Wort w_3 zwar in der E-Patternsprache $L_{\mathrm{E},\Sigma}(\alpha)$, jedoch nicht in der NE-Patternsprache $L_{\mathrm{NE},\Sigma}(\alpha)$ enthalten. .

Trotz des – scheinbar kleinen – definitorischen Unterschieds zwischen E- und NE-Patternsprachen besitzen die beiden Sprachklassen sehr unterschiedliche Eigenschaften. Beispielsweise lässt sich das *Äquivalenzproblem* (d. h. die Frage, ob zwei Pattern die gleiche Sprache erzeugen) für *NE*-Patternsprachen trivial in Polynomialzeit entscheiden, während die Entscheidbarkeit des Äquivalenzproblems für *E*-Patternsprachen ein schweres und seit mehr als 25 Jahren offenes Problem darstellt (siehe [OU97, Rei07]).

Beide Sprachklassen haben ihren Ursprung in der Induktiven Inferenz, einem Teilgebiet der (mathematischen) Lerntheorie. Ziel dieser Betrachtungen war es, zu gegebenen Mengen von Wörtern effektiv Pattern zu finden, die all diese Wörter beschreiben. Inzwischen wurden Patternsprachen auch in der Theorie der formalen Sprachen ausgiebig untersucht. Wegen ihrer einfachen Definition kommen Pattern und ihre Sprachen in einer Vielzahl anderer Gebiete der Informatik und der diskreten Mathematik vor: Die in modernen Programmiersprachen verwendeten *erweiterten regulären Ausdrücke* (auch *regex* genannt, siehe Abschnitt 3), *Wortgleichungssysteme*, verschiedene *Datenbankanfragesprachen* – kurz, nahezu alle Modelle, die einen Wiederholungsoperator verwenden – können als Erweiterung oder Anwendung von Patternsprachen verstanden werden.

Aufgrund ihrer einfachen Definition ließe sich vermuten, dass die meisten interessanten Eigenschaften einer Patternsprache direkt an dem sie erzeugenden Pattern zu erkennen sind. Insbesondere sollte sich daher eigentlich leicht entscheiden lassen, ob zwei Pattern die gleiche oder vergleichbare Sprachen erzeugen (d. h. ob eine Äquivalenz- oder Inklusionsbeziehung vorliegt), und ob ein Wort in der Sprache eines Patterns ist (das sogenannte *Wortproblem*). Tatsächlich ist diese Vermutung aber in den meisten Fällen ein Trugschluss – beispielsweise ist das Wortproblem NP-vollständig und das Inklusionsproblem im Allgemeinen unentscheidbar.

Diese unteren Schranken für Probleme zu Patternsprachen gelten selbstverständlich auch unmittelbar für alle mächtigeren Modelle, die Patternsprachen erweitern. Häufig folgt in der theoretischen Informatik auf das Feststellen einer unteren Schranke für ein Problem die Frage nach geeigneten Einschränkungen, mit denen diese Schranke überwunden werden kann. Hierbei bieten sich Patternsprachen als „Versuchsgelände“ für eine Vielzahl dieser Einschränkungen an, denn eine Einschränkung, die für Patternsprachen nicht zu besseren unteren Schranken führt, kann auch bei den stärkeren Modellen nicht greifen.

Jede Erkenntnis über die Schwierigkeiten im Umgang mit Patternsprachen lässt sich daher unmittelbar auf die mächtigeren Klassen übertragen. Gleichzeitig lassen sich die für Patternsprachen entwickelten Beweise in mächtigeren Modellen oft erweitern, so dass deutlich stärkere Resultate erzielt werden können. Auf diese Art stellen Patternsprachen eine Art „Werkzeugkasten“ für andere Modelle zur Verfügung. Zwei Beispiele für die Anwen-

dung dieses Prinzips werden in den Abschnitten 3 und 7 vorgestellt.

Die hier zusammengefasste Dissertation [Fre11b] beschäftigt sich hauptsächlich mit verschiedenen Aspekten des Inklusionsproblems. Den eigentlichen Hauptteil der Arbeit bilden die Kapitel 3 bis 7, die sich grob in zwei Teile unterteilen lassen.

Der erste Teil, der aus den Kapiteln 3 und 4 besteht, befasst sich direkt mit dem Inklusionsproblem für Patternsprachen und wendet eine in diesem Kontext entwickelte Technik auf eine in der Praxis weit verbreitete Erweiterung der regulären Ausdrücke an.

Der zweite Teil, bestehend aus den Kapiteln 5 bis 7, untersucht Fragestellungen zu sogenannten *deskriptive Pattern*, die aufgrund ihrer Definition eng mit dem Inklusionsproblem verwandt sind.

Im Folgenden wird der Inhalt der einzelnen Kapitel genauer vorgestellt und jeweils explizit erwähnt, in welcher Form die genannten Resultate bereits veröffentlicht wurden. Aus Platzgründen wird auf die meisten Definitionen verzichtet.

2 Inklusion von Patternsprachen (Kapitel 3)

Die Entscheidbarkeit des Inklusionsproblems für Patternsprachen wurde bereits bei der Einführung der NE-Patternsprachen durch [Ang80] als offenes Problem benannt. Sowohl der NE- als auch der E-Fall dieses Problems erwiesen sich als überraschend schwer zu lösen, bis schließlich in [JSSY95] der Beweis für die Unentscheidbarkeit beider Fälle gelang.

Allerdings basiert der in [JSSY95] angegebene Beweis auf der Annahme, dass die darin verwendeten Patternsprachen auf Terminalalphabeten von unbeschränkter Größe definiert werden können. Im Gegensatz dazu verwenden aber die meisten Anwendungen von Patternsprachen Alphabete von beschränkter Größe. Trotz des wegweisenden Resultats von [JSSY95] blieb die Entscheidbarkeit des Inklusionsproblems über festen Alphabeten offen. Das erste Hauptresultat der vorliegenden Arbeit beantwortet diese Frage negativ:

Theorem 3.3 *Sei Σ ein endliches Alphabet, $|\Sigma| \geq 2$. Das Inklusionsproblem für E-Patternsprachen über Σ ist unentscheidbar.*

Mittels einer Reduktion aus [JSSY95] folgt hieraus außerdem die Unentscheidbarkeit des Inklusionsproblems für NE-Patternsprachen über endlichen Alphabeten mit mindestens vier Buchstaben. Theorem 3.3 beantwortet die entsprechenden offenen Fragen von [Rei06] und [Sal06] und wurde bereits kurz nach seiner Veröffentlichung von [BHLW10] in der Datenbanktheorie angewendet.

Im restlichen Teil von Kapitel 3 wird die Entscheidbarkeit der Inklusion für stärker eingeschränkte Klassen von Patternsprachen untersucht. Sowohl der ursprüngliche Beweis von Jiang et al., als auch der darauf aufbauende Beweis von Theorem 3.3 setzen voraus, dass die verwendeten Pattern unbeschränkt viele verschiedene Variablen enthalten dürfen.

Daher ist es naheliegend, als weitere Einschränkung die Anzahl der in den Pattern vorkommenden Variablen zu betrachten. Eine Verfeinerung des Beweises von Theorem 3.3

zeigt, dass das Inklusionsproblem für E-Patternsprachen mit einer beschränkten (wenn auch vergleichsweise hohen) Anzahl von Variablen weiterhin unentscheidbar ist:

Theorem 3.10. *Ist $|\Sigma| = 2$, so ist das Inklusionsproblem $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\beta)$ für Pattern mit beschränkter Variablenanzahl unentscheidbar, wenn*

1. *α auf 3 und β auf 2854 Variablen beschränkt ist, oder*
2. *α auf 2 und β auf 2860 Variablen beschränkt ist.*

Um für weiter reduzierte Variablenanzahlen interessante Resultate unterhalb der Unentscheidbarkeit zu finden, wird (analog zu einer Technik aus der Untersuchung kleiner Turingmaschinen) außerdem ein sehr einfaches Berechnungsmodell auf die Inklusion von Patternsprachen reduziert. Hierzu wird die Iteration der *Collatzfunktion $\mathcal{C}$* (siehe [Lag09a, Lag09b]) betrachtet[1]. Die *Collatzvermutung* besagt, dass die wiederholte Iteration von $\mathcal{C}$ für alle Startwerte zum *trivialen Zyklus* 4, 2, 1 führt.

Hier stellt sich heraus, dass bereits eine verhältnismäßig geringe Anzahl von Variablen genügt, um die Iteration der Collatzfunktion in Patternsprachen zu codieren:

Theorem 3.11. *Sei $|\Sigma| = 2$. Jeder Algorithmus, der die Inklusion $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\beta)$ für Pattern mit beschränkter Variablenanzahl entscheidet, wobei α auf 2 und β auf 74 Variablen beschränkt ist, kann effektiv in einen Algorithmus konvertiert werden, der für jedes $n \geq 1$ entscheidet, ob die Iteration von $\mathcal{C}$ mit Startwert n zum trivialen Zyklus führt.*

Dieses Resultat lässt sich leicht zu dem folgenden mächtigeren Theorem erweitern:

Theorem 3.12. *Sei $|\Sigma| = 2$. Jeder Algorithmus, der die Inklusion $L_{\mathrm{E},\Sigma}(\alpha) \subseteq L_{\mathrm{E},\Sigma}(\beta)$ für Pattern mit beschränkter Variablenanzahl entscheidet, wobei α auf 4 und β auf 80 Variablen beschränkt ist, kann effektiv in einen Algorithmus konvertiert werden, der die Existenz nicht-trivialer Zyklen von $\mathcal{C}$ entscheidet.*

Ein entsprechender Entscheidungsalgorithmus könnte also in endlicher Zeit entweder die Collatzvermutung widerlegen, oder aber die Nichtexistenz einer von zwei möglichen Klassen von Gegenbeispielen beweisen.

Auch wenn aus diesen Resultaten keine Unentscheidbarkeit der Inklusionsprobleme von Patternsprachen mit den entsprechenden Variablenanzahlen folgt, zeigen die Theoreme 3.11 und 3.12, dass die entsprechenden Probleme wenn nicht unentscheidbar, so doch zumindest schwer zu lösen sind. Alle hier aufgeführten Resultate lassen sich mit ähnlichen Variablenanzahlen auf größere (endliche) Alphabete und auf NE-Patternsprachen übertragen.

Die Inhalte dieses Kapitels wurden in den Arbeiten [FR10a] (zuerst erschienen als [FR08]) und [BF10] veröffentlicht.

3 Erweiterte reguläre Ausdrücke (Kapitel 4)

Reguläre Ausdrücke zählen zu den am weitesten verbreiteten Beschreibungsmechanismen und werden sowohl in der theoretischen, als auch in der praktischen Informatik auf ver-

[1] Die Funktion $\mathcal{C}$ ist definiert durch $\mathcal{C}(n) := 3n + 1$ für ungerade n, und $\mathcal{C}(n) := \frac{1}{2}n$ für gerade n.

schiedenste Arten angewendet. Allerdings haben sich im Lauf der Jahrzehnte in Theorie und Anwendung zwei unterschiedliche Interpretationen dieses Konzepts entwickelt.

Während die Theorie weitestgehend der klassischen Definition folgt und reguläre Ausdrücke betrachtet, die genau die Klasse der regulären Sprachen beschreiben, erlauben die meisten modernen Implementierungen von regulären Ausdrücken (so zum Beispiel in Perl, Java, C#) die Spezifikation von Wiederholungen mittels *Variablen* (oder *Rückreferenzen*). Die daraus resultierenden *erweiterten regulären Ausdrücke*, auch *regex* genannt, können dank dieser Wiederholungen auch nichtreguläre Sprachen beschreiben.

Beispielsweise erzeugt der erweiterte reguläre Ausdruck $((\mathtt{a} \mid \mathtt{b})^*)\%x\, x$ die nichtreguläre Sprache $\{ww \mid w \in \{\mathtt{a},\mathtt{b}\}^*\}$. Hierbei kann der Teilausdruck $((\mathtt{a} \mid \mathtt{b})^*)\%x$ ein beliebiges Wort $w \in \{\mathtt{a},\mathtt{b}\}^*$ erzeugen, während zudem dieses Wort w in der Variablen x gespeichert wird. Weitere Vorkommen von x erzeugen exakt das gleiche Wort w.

Andererseits führt die Verwendung von Variablen nicht zwangsläufig zu Nichtregularität; beispielsweise erzeugt (für jedes $n \geq 1$) der Ausdruck

$$\alpha_n := (\underbrace{(\mathtt{a} \mid \mathtt{b}) \ldots (\mathtt{a} \mid \mathtt{b})}_{n \text{ mal } (\mathtt{a} \mid \mathtt{b})})\%x\, x$$

die endliche (und daher reguläre) Sprache aller Wörter $ww \in \{\mathtt{a},\mathtt{b}\}^*$, die die Länge $2n$ haben. Hierbei fällt auf, dass *klassische reguläre Ausdrücke* (d. h. Ausdrücke, die keine Variablen enthalten) für diese Sprachen exponentiell länger sind als die Ausdrücke α_n.

Kapitel 4 befasst sich hauptsächlich mit den folgenden zwei Fragen: Erstens, können erweiterte reguläre Ausdrücke – in Hinsicht auf ihre Länge oder auf ihre Variablenanzahl – effektiv minimiert werden? Und zweitens, um wie viel kompakter ist die Beschreibung von regulären Sprachen durch erweiterte reguläre Ausdrücke im Vergleich zu klassischen regulären Ausdrücken?

Die Klasse der Patternsprachen ist eine Teilklasse der von erweiterten regulären Ausdrücken erzeugten Sprachen, da sich die erzeugenden Pattern ohne großen Aufwand direkt in die entsprechenden Ausdrücke konvertieren lassen. Allerdings erhöht die Verwendung des Alternierungsoperators $\mid$ die Ausdruckskraft so sehr, dass deutlich schärfere Nichtentscheidbarkeitsresultate als die in Kapitel 3 enthaltenen Resultate zur Inklusion für Patternsprachen erzielt werden können.

Die Konstruktion zum Beweis von Theorem 3.10 lässt sich mit signifikantem Zusatzaufwand zu einer mächtigeren Konstruktion für erweiterte reguläre Ausdrücke umbauen (Theorem 4.14), mit deren Hilfe in Theorem 4.15 mehrere Nichtentscheidbarkeitsresultate gewonnen werden:

Theorem 4.15. *Für erweiterte reguläre Ausdrücke ist die Allspracheneigenschaft nicht entscheidbar; Regularität sowie Endlichkeit des Komplements sind weder semi-entscheidbar, noch co-semi-entscheidbar.*

Mittels dieser Resultate können die beiden weiter oben genannten Fragen beantwortet werden: Erweiterte reguläre Ausdrücke können nicht effektiv minimiert werden, und der Größenunterschied zwischen erweiterten und klassischen regulären Ausdrücken ist durch keine berechenbare Funktion beschränkt.

Besondere Erwähnung verdient hierbei die Tatsache, dass die genannten negativen Eigenschaften bereits bei erweiterten regulären Ausdrücken mit einer einzigen Variable gelten.

Als praktische Konsequenz lässt sich feststellen, dass selbst die Verwendung einer einzigen Variable zwar deutlich kompaktere Ausdrücke erlauben kann (was sich natürlich auch positiv auf die Geschwindigkeit des Matchens der Ausdrücke auswirkt), dass aber andererseits dieser Vorteil aufgrund der Unmöglichkeit einer effektiven (oder gar effizienten) Minimierung nicht generell nutzbar ist.

Es ist zu hoffen, dass geeignete Teilklassen der erweiterten regulären Ausdrücke gefunden werden können, die diese enormen Kompressionsmöglichkeiten möglichst umfassend ausnutzen und sich trotzdem effektiv (und vorzugsweise auch effizient) aus klassischen regulären Ausdrücken berechnen lassen.

Die Inhalte dieses Kapitels wurden in der Arbeit [Fre12] (zuerst erschienen als [Fre11a]) veröffentlicht.

4 Existenz deskriptiver Pattern (Kapitel 5)

Ein Pattern α ist *konsistent* mit einer Menge $S \subseteq \Sigma^*$, wenn jedes Wort von S in der von α erzeugten Sprache $L(\alpha)$ enthalten ist[2], d. h. wenn $L(\alpha) \supseteq S$ gilt. So sind beispielsweise die Pattern $\alpha_0{:=}x$, $\alpha_1{:=}xyxyx$ und $\alpha_2{:=}x\,\mathtt{a}\,\mathtt{b}\,y$ konsistent mit der Menge $S_0{:=}\{\mathtt{ababa}, \mathtt{ababbababbab}, \mathtt{babab}\}$. Konsistente Pattern liefern also eine kompakte und leicht verständliche Darstellung von Gemeinsamkeiten der Wörter einer Menge.

Wie das oben stehende Beispiel zeigt, gibt es zu einer Menge von Wörtern im Allgemeinen eine Vielzahl von konsistenten Pattern, die intuitiv eine sehr unterschiedliche Güte haben können; beispielsweise ist das Pattern α_0 mit jeder Sprache konsistent und dürfte daher für nahezu alle Anwendungen als eine triviale und uninteressante Approximation gelten.

Es ist daher vorteilhaft, konsistente Pattern hoher Qualität formal zu fassen. Ein in der Literatur häufig verwendetes Qualitätsmaß ist das Konzept der *deskriptiven* Pattern für eine Menge S. Ein Pattern δ dessen Sprache $L(\delta)$ in einer gegebenen Klasse P von Patternsprachen liegt, ist P-deskriptiv für eine Sprache S, wenn δ mit S konsistent ist und außerdem $L(\delta) \supset L(\gamma) \supseteq S$ für kein Pattern γ mit $L(\gamma) \in P$ gilt.

Anschaulich formuliert kann ein P-deskriptives Pattern als Erzeuger einer kleinsten innerhalb der Klasse P möglichen Generalisierung der Zielsprache S verstanden werden. Die vorliegende Arbeit betrachtet hierbei vor allem die Klassen der NE-Patternsprachen, der E-Patternsprachen, und die Klasse der *terminalfreien E-Patternsprachen* (also derjenigen E-Patternsprachen, die von Pattern erzeugt werden, die keine Terminalsymbole enthalten). In den ersten beiden Fällen sprechen wir gewöhnlich von *NE-deskriptiven* beziehungsweise *E-deskriptiven* Pattern.

Die wahrscheinlich naheliegendste Frage bei der Suche nach deskriptiven Pattern für beliebige Sprachen ist, ob zu jeder Sprache mindestens ein deskriptives Pattern existiert. Für

[2] Hierbei ist natürlich von Fall zu Fall festzulegen, ob E- oder NE-Patternsprachen betrachtet werden.

NE-Patternsprachen wurde diese Frage bereits von [Ang80] positiv beantwortet. Ebenso wurde von Jiang et al. [JKS+94] bewiesen, dass alle endlichen Sprachen ein E-deskriptives Pattern besitzen, während der Fall von E-deskriptiven Pattern für unendliche Sprachen offen blieb. Das Hauptresultat des Kapitels beantwortet diese Frage negativ:

Theorem 5.18. *Zu jedem Alphabet Σ mit $|\Sigma| \geq 2$ existiert eine unendliche Sprache $L_\Sigma \subset \Sigma^*$, für die kein Pattern E-deskriptiv ist.*

Um die durch die Nichtentscheidbarkeit des Inklusionsproblems entstehenden Schwierigkeiten zu umgehen, verwendet der Beweis von Theorem 5.18 terminalfreie E-Patternsprachen, da für diese Unterklasse die Inklusion durch ein einfaches syntaktisches Kriterium charakterisiert wird.

Die meisten Inhalte dieses Kapitels wurden in der Arbeit [FR10b] (zuerst erschienen als [FR09]) veröffentlicht.

5 Deskriptive Generalisierung (Kapitel 6)

Im Gegensatz zu Kapitel 5, in dem die Existenz deskriptiver Pattern im Vordergrund steht, befasst sich dieses Kapitel mit der Frage nach ihrer effektiven Auffindbarkeit.

Hierzu wird das Konzept der *deskriptiven Generalisierung (anhand von positiven Daten)* eingeführt (kurz: das DG-Modell), das an Golds klassisches Modell (siehe [Gol67]) der *Identifikation von Sprachen im Limes (anhand von positiven Daten)*, das LIM-TEXT-Modell angelehnt ist. Das LIM-TEXT-Modell wurde in der einschlägigen Literatur intensiv untersucht, ein recht aktueller Übersichtsartikel ist [NS08].

Anschaulich (und vereinfacht) ausgedrückt untersuchen wir die Existenz von Lernstrategien, die Positivbeispiele der zu lernenden Sprache L wortweise einlesen und, wann immer ein bisher ungesehenes Wort eingelesen wird, ein Pattern als *Hypothese* ausgeben.

Ist P eine Klasse von Patternsprachen, so ist eine Klasse $\mathcal{L}$ von Sprachen *P-deskriptiv generalisierbar*, wenn eine berechenbare *Generalisierungsstrategie* S existiert, so dass für jede Sprache $L \in \mathcal{L}$ die Folge der von S ausgegeben Hypothesen gegen ein Pattern δ konvergiert, das P-deskriptiv für L ist.

Es muss also keine exakte Repräsentation der Zielsprachen erlernt werden, sondern nur eine Approximation. Hierbei stellt sich heraus, dass die Korrespondenz zwischen zu generalisierenden Sprachen und korrekten Hypothesen im Allgemeinen schwächer ist als bei den üblicherweise in der Literatur betrachteten Lernmodellen: Ein Pattern δ kann gleichzeitig deskriptives Pattern (und korrekte Hypothese) für zwei unvergleichbare Sprachen L_1, L_2 sein, und eine Sprache L kann durch zwei unterschiedliche deskriptive Pattern δ_1 und δ_2 generalisiert werden.

Um Probleme mit der Unentscheidbarkeit der Inklusion zu vermeiden, wird das Modell anhand der Klasse $\mathrm{ePAT}_{\mathrm{tf}}$ der teminalfreien E-Patternsprachen untersucht. Im Mittelpunkt der Untersuchung steht hierbei das Konvergenzverhalten der *kanonischen Strategie* Canon. Diese berechnet zu jeder Menge von Positivbeispielen aus der Zielsprache ein de-

skriptives Pattern. Die Klasse der Sprachen, auf denen Canon korrekt konvergiert, lässt sich wie folgt charakterisieren:

Theorem 6.26. *Sei Σ ein Alphabet mit $|\Sigma| \geq 2$. Für jede Sprache $L \subseteq \Sigma^*$, und jede Aufzählung $t : \mathbb{N} \to L$ von Positivbeispielen von L konvergiert* Canon *korrekt auf t genau dann, wenn L ein telling set besitzt.*

Ein *telling set* einer Sprache L ist hierbei eine endliche Teilmenge $T \subseteq L$, so dass ein terminalfreies Pattern δ existiert, dass sowohl für T als auch für L $\mathrm{ePAT}_{\mathrm{tf}}$-deskriptiv ist. Mittels dieser Charakterisierung lässt sich eine große deskriptiv generalisierbare Klasse von Sprachen definieren, wodurch tiefere Einsichten zur Stärke des DG-Modells gewonnen und ein Vergleich zum $\mathrm{LIM\text{-}TEXT}$-Modell ermöglicht werden.

Die Inhalte dieses Kapitels wurden in der Arbeit [FR12] (zuerst erschienen als [FR10c]) veröffentlicht.

6 Über eine Vermutung zu deskriptiven Pattern (Kapitel 7)

Dieses Kapitel befasst sich mit einer Vermutung, die der Autor beim Versuch aufstellte, den Beweis von Theorem 5.18 zu einer Charakterisierung derjenigen Sprachen zu erweitern, für die kein Pattern in Bezug auf die Klasse der terminalfreien E-Pattersprachen deskriptiv ist.

Dazu wird ein größeres technisches Instrumentarium eingeführt und anschließend zur Konstruktion von Gegenbeispielen zu dieser Vermutung verwendet. Darüber hinaus werden verschiedene Phänomene diskutiert, die die Schwierigkeit einer Charakterisierung der genannten Sprachen verdeutlichen.

Als Fazit dieses Kapitels lässt sich feststellen, dass die Nichtexistenz deskriptiver Pattern selbst in Bezug auf die Klasse der terminalfreien E-Patternsprachen komplex und wahrscheinlich nur schwer zu charakterisieren ist.

7 Zusammenfassung und neuere Resultate

Trotz ihrer einfachen Definition führen Pattern zu komplizierten Problemen, von denen einige durch die in der hier zusammengefassten Dissertation vorgestellten Resultate besser verstanden werden können. Auch wenn für viele dieser Probleme nur festgestellt werden konnte, dass sie beweisbar schwer sind, ist der Autor überzeugt, dass der Nutzen dieser Ergebnisse über das tiefere Verständnis von Patternsprachen hinausgeht. Zwei Beispiele, die diesen Standpunkt unterstützen, sind die in Kapitel 3 vorgestellte Erweiterung der Resultate auf ein in der Praxis weitverbreitetes Modell, sowie die in Abschnitt 2 erwähnte Anwendung in der Datenbanktheorie aus [BHLW10].

Zwei weitere Entwicklungen, die sich nach Einreichen der Dissertation ergeben haben, sind in diesem Zusammenhang erwähnenswert: Erstens wurde der in der Einleitung erwähnte

(und in Kapitel 4 angewandte) Ansatz, Resultate zu Patternsprachen in mächtigere Modelle zu übertragen und zu erweitern, inzwischen in [FS11] erfolgreich angewendet. Hierdurch konnten mehrere in [BHLW10] offen gelassene Fragen beantwortet und darüber hinausgehende Resultate erzielt werden.

Außerdem konnte der Autor bei einem Forschungsbesuch am MPII in Saarbrücken zusammen mit Timo Kötzing das Konzept der deskriptiven Generalisierung von Patternsprachen auf Unterklassen von regulären Ausdrücken übertragen. Ein Artikel über die dabei gewonnenen effizienten Algorithmen zum Approximieren von XML DTDs ist derzeit in Arbeit.

Literatur

[Ang80] D. Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, 21:46–62, 1980.

[BF10] J. Bremer und D. D. Freydenberger. Inclusion Problems for Patterns With a Bounded Number of Variables. In *Proc. DLT 2010*, Jgg. 6224 of *LNCS*, Seiten 100–111, 2010.

[BHLW10] P. Barceló, C. A. Hurtado, L. Libkin und P. T. Wood. Expressive languages for path queries over graph-structured data. In *Proc. PODS 2010*, Seiten 3–14, 2010.

[Cas02] J. Cassaigne. Unavoidable Patterns. In M. Lothaire, Hrsg., *Algebraic Combinatorics on Words*, Kapitel 3, Seiten 111–134. Cambridge University Press, Cambridge, New York, 2002.

[FR08] D. D. Freydenberger und D. Reidenbach. Bad news on decision problems for patterns. In *Proc. DLT 2008*, LNCS 5257, Seiten 327–338, 2008.

[FR09] D. D. Freydenberger und D. Reidenbach. Existence and Nonexistence of Descriptive Patterns. In *Proc. DLT 2009*, LNCS 5583, Seiten 228–239, 2009.

[FR10a] D. D. Freydenberger und D. Reidenbach. Bad news on decision problems for patterns. *Information and Computation*, 208(1):83–96, 2010.

[FR10b] D. D. Freydenberger und D. Reidenbach. Existence and nonexistence of descriptive patterns. *Theoretical Computer Science*, 411(34-36):3274 – 3286, 2010.

[FR10c] D. D. Freydenberger und D. Reidenbach. Inferring Descriptive Generalisations of Formal Languages. In *Proc. COLT 2010*, Seiten 194–206, 2010.

[FR12] D. D. Freydenberger und D. Reidenbach. Inferring Descriptive Generalisations of Formal Languages. *Journal of Computer and System Sciences*, 2012. Angenommen.

[Fre11a] D. D. Freydenberger. Extended Regular Expressions: Succinctness and Decidability. In *Proc. STACS 2011*, LIPIcs 9, Seiten 507–518, 2011.

[Fre11b] D. D. Freydenberger. *Inclusion of Pattern Languages and Related Problems*. Dissertation, Fachbereich Informatik und Mathematik, Johann Wolfgang Goethe-Universtität, Frankfurt am Main, 2011. Logos Verlag, Berlin.

[Fre12] D. D. Freydenberger. Extended Regular Expressions: Succinctness and Decidability. *Theory of Computing Systems*, 2012. Angenommen.

[FS11] D. D. Freydenberger und N. Schweikardt. Expressiveness and Static Analysis of Extended Conjunctive Regular Path Queries. In *Proc. AMW 2011*, CEUR Workshop Proceedings 749, 2011.

[Gol67] E. M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[JKS+94] T. Jiang, E. Kinber, A. Salomaa, K. Salomaa und S. Yu. Pattern Languages With and Without Erasing. *International Journal of Computer Mathematics*, 50:147–163, 1994.

[JSSY95] T. Jiang, A. Salomaa, K. Salomaa und S. Yu. Decision Problems for Patterns. *Journal of Computer and System Sciences*, 50:53–63, 1995.

[Lag09a] J. C. Lagarias. The 3x+1 problem: An annotated bibliography (1963–1999), Aug 2009. `http://arxiv.org/abs/math/0309224`.

[Lag09b] J. C. Lagarias. The 3x+1 Problem: An Annotated Bibliography, II (2000–2009), Aug 2009. `http://arxiv.org/abs/math/0608208`.

[Nag77] T. Nagell, Hrsg. *Selected mathematical papers of Axel Thue*. Universitetsforlaget, Oslo, 1977.

[NS08] Y. K. Ng und T. Shinohara. Developments from enquiries into the learnability of the pattern languages from positive data. *Theoretical Computer Science*, 397:150–165, 2008.

[OU97] E. Ohlebusch und E. Ukkonen. On the equivalence problem for E-pattern languages. *Theoretical Computer Science*, 186:231–248, 1997.

[Rei06] D. Reidenbach. *The Ambiguity of Morphisms in Free Monoids and its Impact on Algorithmic Properties of Pattern Languages*. Dissertation, Fachbereich Informatik, Technische Universität Kaiserslautern, 2006. Logos Verlag, Berlin.

[Rei07] D. Reidenbach. An examination of Ohlebusch and Ukkonen's Conjecture on the equivalence problem for E-pattern languages. *Journal of Automata, Languages and Combinatorics*, 12:407–426, 2007.

[Sal06] K. Salomaa. Patterns, 2006. Vorlesung, 5th PhD School in Formal Languages and Applications, URV Tarragona.

[Shi82] T. Shinohara. Polynomial Time Inference of Extended Regular Pattern Languages. In *Proc. RIMS Symposia on Software Science and Engineering, Kyoto*, LNCS 147, Seiten 115–127, 1982.

[Thu06] A. Thue. Über unendliche Zeichenreihen. *Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl.*, 7, 1906. Nachgedruckt in [Nag77].

Dominik D. Freydenberger Geboren am 31. Oktober 1979 in Regensburg. Studium der Informatik an der TU Kaiserslautern, Abschluss als Diplom-Informatiker im März 2006. Anschließend Teilnahme an der *5th PhD School in Formal Languages and Applications* an der URV Tarragona. Von März 2007 bis Februar 2012 wissenschaftlicher Mitarbeiter am Institut für Informatik an der Goethe-Universität in Frankfurt am Main, Promotion zum Doktor der Naturwissenschaften am 27. Juni 2011. Ab März 2012 Akademischer Rat auf Zeit, ebenfalls an der Goethe-Universität.

Exponentielle untere Schranken zur Lösung infinitärer Auszahlungsspiele und linearer Programme

Oliver Friedmann

Ludwig-Maximilians-Universität
Institut für Informatik
Oliver.Friedmann@ifi.lmu.de

Abstract: Wir betrachten die Strategieverbesserungstechnik zur Lösung infinitärer Auszahlungsspiele sowie den Simplexalgorithmus zur Lösung damit assoziierter linearer Programme.

Unser Beitrag zur Strategieverbesserung und zum Simplexverfahren besteht in der Konstruktion exponentieller unterer Schranken für mehrere Verbesserungs- bzw. Pivotregeln. Für jede Verbesserungsregel, die wir in dieser Arbeit unter die Lupe nehmen, konstruieren wir Zweispieler-*Paritätsspiele*, zu deren Lösung der entsprechend parametrisierte Strategieverbesserungsalgorithmus eine exponentielle Anzahl an Iterationen benötigt. Anschließend übersetzen wir diese Spiele in Einspieler-*Markov-Entscheidungsprozesse*, die wiederum beinahe direkt in konkrete lineare Programme überführt werden können, zu deren Lösung der entsprechende parametrisierte Simplexalgorithmus dieselbe Anzahl an Iterationen benötigt. Zusätzlich zeigen wir, wie sich die unteren Schranken auf expressivere Spieleklassen wie Auszahlungs- und stochastische Auszahlungsspiele übertragen lassen.

Ausführliche Zusammenfassung

Infinitäre Auszahlungsspiele Die Arbeit betrachtet eine Vielzahl naheverwandter Klassen von Spielen, die als *infinitäre Auszahlungsspiele* bezeichnet werden. Hierbei handelt es sich um Nullsummenspiele mit perfekter Information, die von einem oder zwei Spielern gespielt werden und darüber hinaus manchmal einen zusätzlichen randomisierten Spieler enthalten, der von der "Natur" kontrolliert wird. Das Spielfeld wird als gerichteter, totaler Graph modelliert, wobei jeder Knoten des Graphs zu einem der Spieler gehört.

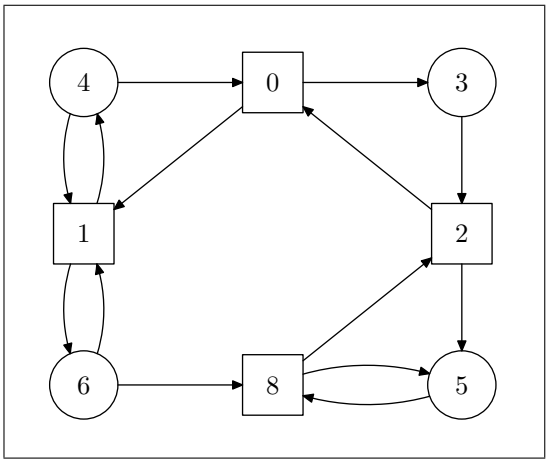
Um eine Partie in so einem Spiel zu spielen, wird ein einzelner Spielstein auf einen Knoten (zum Beispiel auf einen ausgezeichneten Anfangsknoten) des Graphs gelegt und anhand der ausgehenden Kanten zu einem Nachfolgeknoten gezogen. Der Spieler, zu dem der aktuelle Knoten gehört, entscheidet hierbei, welche ausgehende Kante gewählt werden soll. Dieser Prozess wird unendlich lange fortgeführt, woraus sich eine unendliche Knotensequenz ergibt. Nun hängt es von der spezifischen Klasse von Spiel ab, um entscheiden zu können, welche Auszahlung die Spieler erhalten oder welcher Spieler die Partie gewinnt.

Es ist die Aufgabe jedes Spielers die eigene Auszahlung zu maximieren bzw. gegen den anderen Spieler zu gewinnen. Eine *Strategie* eines Spielers ist ein vollständiger Aktions-

plan für alle möglichen Situationen, die in einer Partie gegen einen beliebigen Gegner auftreten können. Eine Strategie spezifiziert für jeden Knoten, der zu dem entsprechenden Spieler gehört, welcher Nachfolger auszuwählen ist; dies kann im Allgemeinen von der gesamten *Historie* der Partie bis zu diesem Punkt abhängen. Eine Strategie, die nicht von der Historie abhängig ist, nennt man *positional*.

Alle Spiele, die in dieser Arbeit betrachtet werden, sind *positional determiniert*, d.h. positionale Strategien genügen, um die Entscheidungsprobleme zu beantworten, die mit den Spielen verbunden sind. Dies ist aus vielen Gründen angenehm, da beispielsweise die Anzahl der positionalen Strategien endlich ist (sofern das Spiel endliche Größe hat).

Paritätsspiele sind infinitäre 2-Spieler-Auszahlungsspiele, die auf gerichteten Graphen gespielt werden, deren Knoten mit natürlich Zahlen, genannt *Prioritäten*, beschriftet sind. Die zwei Spieler, genannt *Even* und *Odd*, konstruieren einen unendlichen Pfad im Spielegraph. Even gewinnt, falls die größte Priorität, die unendlich oft in der Partie zu sehen ist, gerade ist. Odd gewinnt andernfalls. Ein Paritätsspiel kann wie folgt visualisiert werden (runde Knoten gehören Spieler Even):



Das Problem, ein Paritätsspiel zu *lösen*, d.h. zu bestimmen, welcher der beiden Spieler eine *Gewinnstrategie* besitzt, ist bekanntermaßen äquivalent zum Model-Checking-Problem des modalen μ-Kalküls [EJ91]. Darüber hinaus wird die algorithmische Essenz des Paritätsspiellösens zur Lösung vielfältiger Probleme in der Computer-unterstützten Verifikation [FLL10] sowie in der Kontrollersynthese [VAW03] eingesetzt.

Paritätsspiele bilden eine sehr spezielle Unterklassen der sog. *Mean Payoff Spiele* [Pur95, ZP96], die selbst eine Unterklasse der sog. *Discounted Payoff Spiele* bilden, die wiederum eine sehr spezielle Unterklasse der rundenbasierten *stochastischen Spiele* [Con92, AM09] bilden. Allgemeinere *stochastische Spiele* wurden zuvor bereits von Shapley [Sha53] untersucht.

Eine weitere, überaus wichtige Klasse infinitärer "Auszahlungspiele" stellen die sog. *Markoventscheidungsprozesse* (nach Andrey Markov) dar, die ein mathematisches Modell zur sequentiellen Entscheidungsfindung in unsicheren Situationen beschreibt. Das Interesse

an Markoventscheidungsprozessen begann mit der bahnbrechenden Arbeit von Bellman [Bel57]. Markoventscheidungsprozesse können als spezielle Unterklasse der rundenbasierten stochastischen Spiele aufgefasst werden, in der nur ein Spieler tatsächlich genutzt wird. Markoventscheidungsprozesse haben einen starken Bezug zur Praxis, beispielsweise in der Robotik, der automatisierten Kontrolle, der Ökonomie, im Operations Research sowie in der künstlichen Intelligenz.

Paritäts- und verwandte, ausdrucksstärkere Spieleklassen wie Payoff oder stochastische Spiele, sind bereits für sich alleine genommen ein hochinteressantes Thema aus komplexitätstheoretischer Sicht. Obwohl bekannt ist, dass die Entscheidungsprobleme, die mit diesen Spielen verbunden sind, zu $\mathtt{NP} \cap \mathtt{coNP}$ [EJS93, Pur95], ja sogar zu $\mathtt{UP} \cap \mathtt{coUP}$ [Jur98, ZP96], sowie zu $\mathtt{PLS}$ [BM08] gehören, ist es ein entscheidendes, offenes Problem, ob eine dieser Spieleklassen in deterministischer Polynomialzeit gelöst werden kann. Auf der anderen Seite ist bekannt, dass Markoventscheidungsprozesse durch spezielle Verfahren aus dem linearen Programmieren bereits in Polynomialzeit gelöst werden können.

Strategieverbesserung Die *Strategieverbesserung* oder *Strategieiteration* ist ein sehr umfassender Ansatz, der als Lösungsverfahren für infinitäre Auszahlungsspiele und verwandte Probleme angewandt werden kann. Er wurde von Howard [How60] im Jahre 1960 zur Lösung von Markoventscheidungsprozessen eingeführt und wurde von Hoffman und Karp im Jahre 1966 zur Lösung nichtterminierender stochastischer Spiele [HK66] angepasst. Ein paar Jahrzehnte später passte Condon den Algorithmus zur Lösung rundenbasierter stochastischer Spiele [Con92] an; Puri, Zwick und Paterson verwendeten die Methode zur Lösung von Discounted und Mean Payoff Spielen [Pur95, ZP96]. Schließlich formulierten Jurdziński und Vöge eine *diskrete* Variante der Strategieverbesserung zur Lösung von Paritätsspielen [VJ00].

Die Eleganz der Strategieiteration liegt in der Einfachheit ihrer abstrakten Beschreibung begründet. Sie basiert auf einer (Fixpunkt–)Iteration über eine spezielle, endliche Unterklasse der Strategien des ersten Spielers. In jeder Iteration wird die aktuelle Strategie auf eine sog. *Bewertung* abgebildet. Die Bewertung einer Strategie erlaubt es zu entscheiden, *ob* die Strategie bereits *optimal* für den ersten Spieler ist und falls nicht, *wie* die Strategie verändert werden kann, um eine *bessere* zu erhalten. Eine besonders ansprechende Eigenschaft der Strategieverbesserung ist, dass Bewertungen effizient berechnet werden können. Um nun eine optimale Strategie zu finden – die es erlaubt, eine Lösung für das Spiel abzuleiten –, wird das folgende algorithmische Schema angewendet, wobei mit einer beliebigen Strategie σ gestartet wird:

Algorithm 1 Strategieverbesserung

```
1: while σ ist nicht optimal do
2:     σ ← Verbessere(σ)
3: end while
```

Tatsächlich beschreibt die Strategieverbesserungstechnik eine ganze Klasse von Algorithmen, da im Allgemeinen mehr als ein verbesserter Strategiekandidat existiert, mit dem die Iteration fortgesetzt werden kann. Die Methode zur Auswahl der Nachfolgestrategie

wird dementsprechend *Verbesserungsregel* genannt. Unter der Annahme, dass nur effiziente Verbesserungsregeln betrachtet werden, lässt sich die Zeitkomplexität der Strategieverbesserung im Prinzip durch die Anzahl der benötigten Iterationen beschreiben, da eine einzelne Iteration in deterministischer Polynomialzeit ausgeführt werden kann.

Somit stellt sich die Frage, ob jede Verbesserungsregel immer zu einer kleinen Anzahl an Iteration führen sollte. Es ist nicht sonderlich überraschend, dass dem nicht so ist. Ein Beispiel ist bereits seit einiger Zeit bekannt, wonach eine absichtlich schlecht gewählte deterministische Verbesserungsregel eine exponentielle Anzahl an Iterationen erzeugen kann [BV07]. Auf der anderen Seiten stellt sich die Frage, ob es überhaupt theoretisch möglich ist, dass es eine Verbesserungsregel geben könnte, die immer für eine kleine Anzahl an Iterationen sorgen würde. Hier ist die Antwort positiv und der einfache Beweis dazu ist wohlbekannt. Allerdings zeigt der Beweis keine Hinweise auf, wie eine solche Verbesserungsregel aussehen könnte. Oder anders formuliert, die Verbesserungsregel, die aus dem Beweis herausfällt, ist selbst nicht effizient berechenbar.

Eine Vielzahl von sehr unterschiedlichen Verbesserungsregeln wurden bereits vorgestellt. Im Allgemeinen gibt es *deterministische*, *randomisierte* und *memorisierende* Verbesserungsregeln. The wichtigsten Regeln, die in der Literatur erwähnt werden, sind die deterministischen SWITCH-ALL [VJ00] und SWITCH-BEST [Sch08], die randomisierten Verbesserungsregeln SWITCH-HALF [MS99], RANDOM-FACET [Kal92, Kal97, MSW96] und RANDOM-EDGE (Folklore) sowie die memorisierende LEAST-ENTERED [Zad80] Regel. Keine nicht-trivialen unteren Schranken an die worst-case Laufzeitkomplexität all dieser Regeln waren bis zu dieser Dissertation bekannt.

Wie sich außerdem in dieser Arbeit gezeigt hat, ist die Strategieverbesserung darüber hinaus nahe mit dem sog. *Simplexverfahren* zur Lösung linearer Programme verwandt.

Lineare Programmierung Die lineare Programmierung ist eines der wichtigsten Gebiete der Optimierungstheorie und wird aktiv beforscht. Viele ökonomische und praktische Aufgaben können als lineare Programme formuliert werden und viele Unterprobleme der Informatik basieren auf der Lösung assoziierter linearer Programme.

Ein lineares Programm beschreibt die Aufgabe, eine gegebene *lineare Zielfunktion* zu maximieren (oder zu minimieren), wobei zusätzliche eine Menge linearer Gleichungen und Ungleichungen erfüllt werden muss. Formal ist ein lineares Programm die *Maximierung* einer Zielfunktion

$$c_1 x_1 + \ldots + c_n x_n$$

bezüglich einer Menge linearer (Un)Gleichungen, genannt *Bedingungen*,

$$\begin{aligned} a_{1,1} x_1 + \ldots + a_{1,n} x_n &= b_1 \\ a_{2,1} x_1 + \ldots + a_{2,n} x_n &= b_2 \\ &\vdots \\ a_{m,1} x_1 + \ldots + a_{m,n} x_n &= b_m \end{aligned}$$

wobei alle *Koeffizienten* c_i, alle b_i, sowie alle Koeffizienten $a_{i,j}$ reelle Zahlen sind.

Es zeigt sich, dass Markoventscheidungsprozesse als lineare Programme formuliert werden können, was auch der Grund dafür ist, dass sich diese Klasse von Spielen in polynomieller Zeit lösen lassen.

Es gibt drei grundlegende Algorithmen zur Lösung linearer Programme. Da ist zunächst der *Simplexalgorithmus*, der von Dantzig [Dan63] im Jahre 1963 vorgestellt wurde. Die genaue Laufzeitkomplexität dieses Algorithmus ist unbekannt und es ist ein wichtiges offenes Problem, diese Frage adäquat zu beantworten. Die anderen beiden Algorithmen zur Lösung linearer Programme, nämlich die sog. *Ellipsoidmethode* nach Khachiyan [Kha79] und die sog. *Interior-Point Methode* nach Karmarkar [Kar84], lösen lineare Programme sogar in Polynomialzeit, jedoch nicht in *starker* Polynomialzeit.

Obwohl bis jetzt nicht einmal geklärt ist, ob das Simplexverfahren tatsächlich ein Polynomialzeitalgorithmus sein könnte, so hat das Verfahren zumindest das Potenzial ein starker Polynomialzeitalgorithmus zu sein, was vermutlich auch einer der Gründe dafür ist, dass sich noch so viele Forscher für dieses Verfahren interessieren.

Simplex Algorithmus Der *Simplexalgorithmus* basiert auf der Beobachtung, dass der Raum der Punkte, die alle Bedingungen erfüllen, im Prinzip ein konvexes Polytop beschreibt (wenn man von ein paar Spezialfällen absieht) und dass die Zielfunktion einen optimalen Wert auf einer der Ecken annimmt. Folgerichtig ist die Idee der Methode, mit einer beliebigen Ecke zu starten, anschließend zu überprüfen, ob die Zielfunktion auf dieser Ecke bereits optimal ist, und falls nicht, zu einer benachbarten Ecke mit besserem Zielfunktionswert zu wechseln.

Ähnlich der Strategieverbesserung wird die Simplexmethode mit einer sog. *Pivotregel* parametrisiert, die bestimmt, welche erlaubte Nachbarecke auszuwählen ist. Gleichermaßen hängt die Laufzeitkomplexität des Simplexalgorithmus hauptsächlich von der Anzahl der *Pivotschritte*, d.h. von der Anzahl der besuchten Ecken ab, da alle anderen Operationen in (starker) Polynomialzeit ausgeführt werden können.

Die Frage, ob jede Pivotregel immer nur eine geringe Anzahl von Pivotschritten benötigen würde, um eine optimale Ecke zu finden, konnte von Klee und Minty [KM72] bereits kurz nachdem Dantzig den Simplexalgorithmus vorgestellt hatte, verneint werden. Im Gegensatz zur Strategieverbesserung ist es ein offenes Problem, ob es überhaupt theoretisch möglich sein könnte, eine kleine Anzahl an Pivotschritten zusichern zu können. Dies ist bekannt als die *Hirschvermutung* (vgl. dazu z.B. [Dan63], pp. 160,168).

Viele der Verbesserungsregeln der Strategieverbesserung können als Pivotregeln für den Simplexalgorithmus (und umgekehrt) formuliert werden, da die meisten bereits abstrakt genug definiert sind. Genau wie bei der Strategieverbesserung waren bis zu dieser Arbeit keine nicht-trivialen unteren Schranken an die worst-case Laufzeitkomplexität für viele von diesen Regeln bekannt.

Es bleibt die Frage, ob es eine tiefere Verbindung zwischen der Strategieverbesserung zur Lösung infinitärer Auszahlungsspiele und dem Simplexalgorithmus zur Lösung linearer Programme geben könnte. In der Tat wird sich zeigen, dass Markoventscheidungsprozesse das “missing link” darstellen, das die Strategieverbesserung mit der Simplexmethode in einer vernünftigen Art und Weise in Beziehung setzen kann.

Beitrag der Dissertation Die Dissertation beschreibt *exponentielle* (d.h. von der Form $2^{\Omega(n)}$) und *subexponentielle* (d.h. von der Form $2^{\Omega(n^c)}$ für ein $0<c<1$) untere Schranken für alle wesentlichen Verbesserungsregeln der Strategieiteration und alle wesentlichen Pivotregelen des Simplexalgorithmus mit bislang ungeklärtem Komplexitätsstatus.

Zunächst werden die Strategieverbesserung für Paritätsspiele unter die Lupe genommen und explizite Spielefamilien konstruiert, auf denen die verschiedenen Verbesserungsregeln exponentielle bzw. subexponentielle Zeit zur Lösung benötigen.

Genauer werden exponentielle untere Schranken für die deterministischen SWITCH-ALL und SWITCH-BEST Regeln zur Lösung infinitärer Auszahlungsspiele, subexponentielle untere Schranken für die randomisierten RANDOM-EDGE und RANDOM-FACET Regeln zur Lösung von Spielen und linearen Programmen, eine subexponentielle untere Schranke für die randomisierte SWITCH-HALF Regel zur Lösung von Spielen, sowie schließlich eine subexponentielle untere Schranke für Zadehs LEAST-ENTERED Regel zur Lösung von Spielen und linearen Programmen beschrieben.

Der Komplexitätsstatus all dieser Regeln war mehrere Jahrzehnte lang ungelöst. Es bestand bis zuletzt die Hoffnung, dass eine dieser Regeln zumindest Paritätsspiele in polynomieller Zeit hätte lösen können.

Zweitens wird gezeigt, dass sich alle Resultate für Paritätsspiele auf die ausdrucksstärkeren Spieleklassen wie Mean und Discounted Payoff Spiele sowie auf rundenbasierte stochastische Spiele übertragen lassen.

Drittens wird beschrieben, wie die Konstruktionen für Paritätsspiele so als Markoventscheidungsprozesse umgebaut werden können, dass sich dieselben Komplexitätsresultate für Strategieverbesserung auf Markoventscheidungsprozessen ergeben. Dies ist im Allgemeinen vermutlich nicht immer möglich, die speziellen Spiele, die in der Arbeit betrachtet werden, lassen sich jedoch alle übersetzen.

Viertens wird die Beziehung zwischen dem Strategieverbesserungsalgorithmus für Markoventscheidungsprozesse und dem Simplexalgorithmus zur Lösung linearer Programme formalisiert, was es erlaubt, alle anwendbaren unteren Schranken auf den Bereich der linearen Programme zu übertragen. Auch hier waren diese Probleme mehrere Jahrzehnte lang ungelöst. Es ist jedoch zu beachten, dass die Resultate dieser Dissertation keine Aussagen über die Gültigkeit der Hirschvermutung treffen können.

Schließlich werden in der Arbeit noch exponentielle untere Schranken für den Model-Checking-Algorithmus nach Stevens und Stirling, sowie für den rekursiven Algorithmus nach Zielonka zur Lösung von Paritätsspielen gezeigt.

Eingesetzte Techniken Alle Konstruktionen zu unteren Schranken, die in der Dissertation für die Strategieverbesserung und das Simplexverfahren beschrieben werden, basieren auf folgenden Schritten:

1. Zunächst werden Familien von Paritätsspielen konstruiert, die jeweils – für jede Verbesserungsregel gesondert – untere Schranken konstituieren. Hierbei beschränken sich die Konstruktionen auf eine in der Arbeit definierte Teilklasse der Paritätsspie-

le, die sog. *Senken-Paritätsspiele*.

Die Strategieverbesserung für Paritätsspiele wird als eine Art deterministisches (da beide Spieler deterministischer Natur sind) Berechnungsmodell aufgefasst, in dem sich verschiedene, funktionale Strukturen implementieren lassen. Alle Konstruktionen beschreiben die Implementierung von verschiedenen Varianten binärer Zähler.

Das spielentscheidende Gadget, das hier eingesetzt wird, um exponentiell viele Iterationen hervorzurufen, wird *einfacher Zykel* genannt.

Die Beweise, dass die Strategieverbesserung auf den Konstruktionen tatsächlich exponentielles bzw. subexponentielles Verhalten zeigt, basieren dann im Wesentlichen darin, zu zeigen, dass die Sequenz der Strategien binäres Zählen nachempfindet, oder zumindest mit hoher Wahrscheinlichkeit "robust genug" binär zählt, sofern randomisierte Verbesserungsregeln betrachtet werden.

Abbildung 1 zeigt exemplarisch die vollständige Konstruktion eines 3-Bit-Zählers, implementiert in Form eines Paritätsspiels, die exponentiell viele Iterationen zur Lösung benötigt, wenn der Algorithmus mit der Standardverbesserungsregel SWITCH-ALL parametrisiert wird.

2. Die Konstruktionen für Paritätsspiele werden auf die ausdrucksstärkeren Spieleklassen wie Mean und Discounted Payoff Spiele, sowie rundenbasierte stochastische Spiele übertragen.

   Hierzu wird genauer bewiesen, dass sich die Strategieverbesserung auf den sog. Senken-Paritätsspielen exakt gleich verhält, wie die Strategieverbesserung auf den ausdrucksstärkeren Spielklassen, wobei die entsprechenden Spiele durch die gängigen Reduktionen von Paritätsspielen konstruiert werden. Diese Beziehung wird unabhängig von der konkreten Verbesserungsregel bewiesen.

3. Die Resultate werden auf Markoventscheidungsprozesse übertragen. Da es (zumindest nach aktuellem Kenntnisstand) keine allgemeine Übersetzung von Paritätsspielen in Markoventscheidungsprozesse gibt, wird die Gültigkeit der Übersetzungen aller Konstruktionen in jedem Einzelfall bewiesen.

   Die einzigen Strukturen, die in den Graphen der Markoventscheidungsprozesse bei der Übersetzung abgeändert werden muss, sind jene Knoten, die von Spieler Odd kontrolliert werden. In den Spielen der Dissertation wird Spieler Odd glücklicherweise nur auf eine Weise eingesetzt, nämlich in Form der einfachen Zyklen und davon leicht abgewandelter Zyklen.

   Diese Struktur kann durch einen Randomisierungsknoten simuliert werden, wobei die Wahrscheinlichkeit dafür, aus dem Zykel hinauszuziehen, exponentiell klein gewählt wird.

   Dieses Verfahren wurde durch Fearnleys Arbeit [Fea10] inspiriert.

4. Die Übertragung der Resultate auf den Simplexalgorithmus zur Lösung linearer Programme erfolgt durch den allgemeinen Beweis, dass die Simplexmethode auf linearen Programmen, die durch die Konstruktionen auf Markoventscheidungsprozessen induziert werden, sich genau gleich verhält wie die Strategieverbesserung auf den

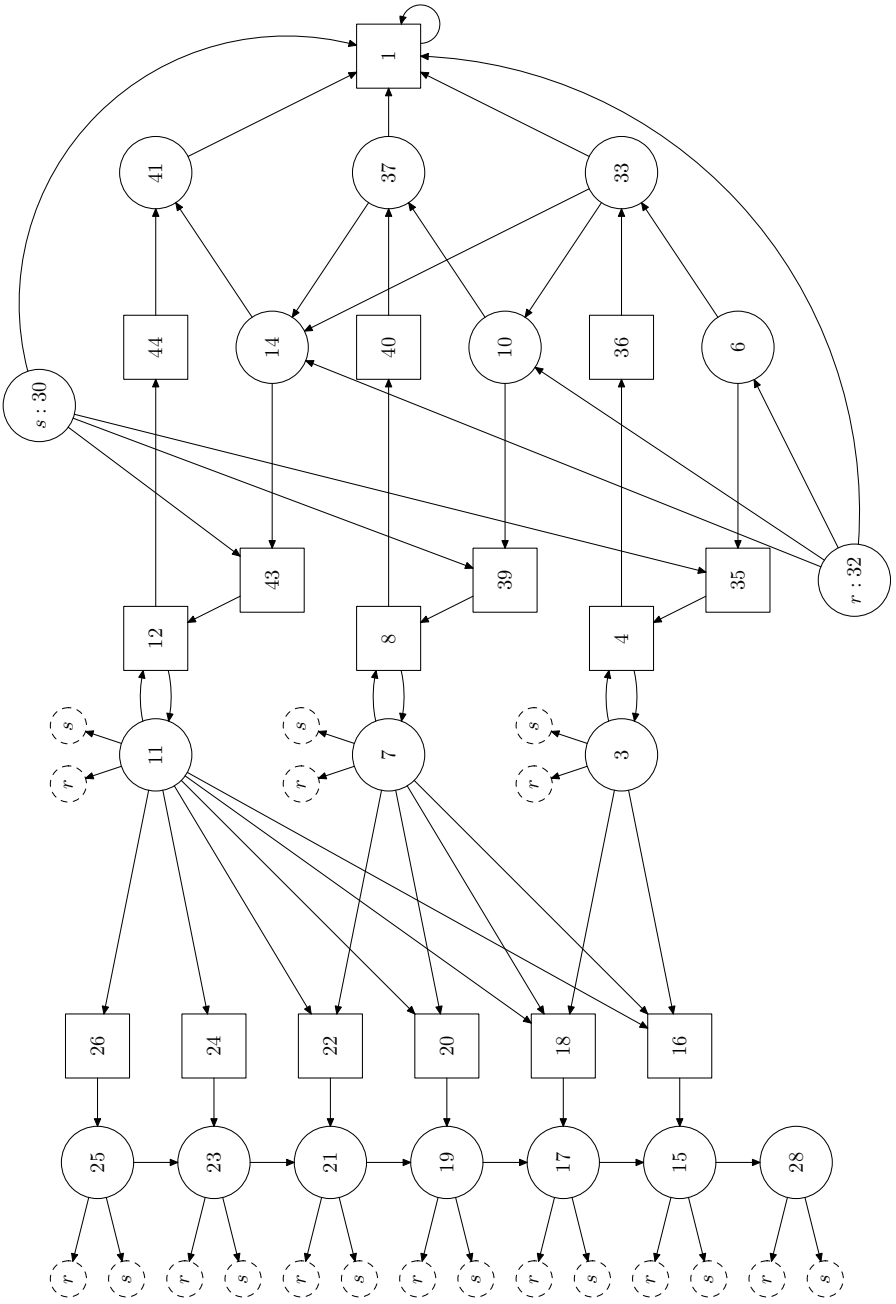


Abbildung 1: SWITCH-ALL Lower Bound Spiel G_3

originalen Markoventscheidungsprozessen. Dieses Verfahren wurde in Zusammenarbeit mit Thomas Dueholm Hansen und Uri Zwick entwickelt.

Die Bedingungen für optimale Werte (und Potenziale) in einem Markoventscheidungsprozess können als lineares Programm formuliert werden, in dem die Variablen den Werten und die Bedingungen des Programms den Kanten entsprechen.

Um nun die unteren Schranken für Markoventscheidungsprozesse auf den Simplexalgorithmus zur Lösung linearer Programme zu übertragen, wird gezeigt, dass (1) Basislösungen im induzierten linearen Programm den Strategien im originalen Spiel entsprechen, und dass (2) angrenzende Basislösungen mit verbesserten Kosten den Strategien entsprechen, die durch eine einzelne Verbesserungskante konstruiert werden. Dies erlaubt es, die unteren Schranken für Markoventscheidungsprozesse direkt auf den Simplexalgorithmus zu übertragen, sofern die entsprechende Verbesserungsregel dementsprechend als Pivotregel formuliert werden kann.

Literatur

[AM09] D. Andersson und P. B. Miltersen. The Complexity of Solving Stochastic Games on Graphs. In *ISAAC '09: Proceedings of the 20th International Symposium on Algorithms and Computation*, Seiten 112–121, Berlin, Heidelberg, 2009. Springer.

[Bel57] R. E. Bellman. *Dynamic programming*. Princeton University Press, 1957.

[BM08] A. Beckmann und F. Moller. On the complexity of parity games. In *Proceedings of the BCS Conference Visions of Computer Science*, 2008.

[BV07] H. Björklund und S. Vorobyov. A combinatorial strongly subexponential strategy improvement algorithm for mean payoff games. *Discrete Applied Mathematics*, 155(2):210–229, 2007.

[Con92] A. Condon. The Complexity of Stochastic Games. *Information and Computation*, 96:203–224, 1992.

[Dan63] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.

[EJ91] E. Emerson und C. Jutla. Tree Automata, μ-Calculus and Determinacy. In *Proceedings of the 32nd Symposium on Foundations of Computer Science*, Seiten 368–377, San Juan, 1991. IEEE.

[EJS93] E. Emerson, C. Jutla und A. Sistla. On Model-Checking for Fragments of μ-Calculus. In *Proceedings of the 5th Conference on CAV, CAV'93*, Jgg. 697 of *LNCS*, Seiten 385–396. Springer, 1993.

[Fea10] J. Fearnley. Exponential Lower Bounds For Policy Iteration. *CoRR*, abs/1003.3418, 2010.

[FLL10] O. Friedmann, M. Latte und M. Lange. A Decision Procedure for CTL* Based on Tableaux and Automata. In *IJCAR*, Seiten 331–345, 2010.

[HK66] A. J. Hofmann und R. M. Karp. On nonterminating stochastic games. *Management Science*, 12(5):359–370, 1966.

[How60] R. Howard. *Dynamic Programming and Markov Processes*. The M.I.T. Press, 1960.

[Jur98] M. Jurdziński. Deciding the winner in parity games is in UP ∩ coUP. *Information Processing Letters*, 68(3):119–124, 1998.

[Kal92] G. Kalai. A Subexponential Randomized Simplex Algorithm (Extended Abstract). In *Proceedings of the 24th STOC*, Seiten 475–482, 1992.

[Kal97] G. Kalai. Linear programming, the simplex algorithm and simple polytopes. *Mathematical Programming*, 79:217–233, 1997.

[Kar84] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of computing*, Seiten 302–311, New York, NY, USA, 1984. ACM.

[Kha79] L. Khachiyan. A Polynomial Algorithm in Linear Programming. *Soviet Mathematics Doklady*, 20:191–194, 1979.

[KM72] V. Klee und G. L. Minty. How good is the simplex algorithm? *Inequalities*, III:159–179, 1972.

[MS99] Y. Mansour und S. P. Singh. On the Complexity of Policy Iteration. In *Proceedings of the 15th UAI*, Seiten 401–408, 1999.

[MSW96] J. Matoušek, M. Sharir und E. Welzl. A subexponential bound for linear programming. *Algorithmica*, 16(4-5):498–516, 1996.

[Pur95] A. Puri. *Theory of Hybrid Systems and Discrete Event Systems*. Dissertation, University of California, Berkeley, 1995.

[Sch08] S. Schewe. An Optimal Strategy Improvement Algorithm for Solving Parity and Payoff Games. In *Proceedings of the 17th Annual Conference on Computer Science Logic, CSL'08*, Jgg. 5213 of *LNCS*, Seiten 369–384. Springer, 2008.

[Sha53] L. S. Shapley. Stochastic Games. *Proceedings of National Academy of Sciences USA*, 39:1095–1100, 1953.

[VAW03] A. Vincent, A. Arnold und I. Walukiewicz. Games for Synthesis of Controllers with Partial Observations. *Theoretical Computer Science*, 303(1):7–34, 2003.

[VJ00] J. Vöge und M. Jurdzinski. A Discrete Strategy Improvement Algorithm for Solving Parity Games. In *Proceedings of the 12th International Conference on Computer Aided Verification, CAV'00*, Jgg. 1855 of *LNCS*, Seiten 202–215. Springer, 2000.

[Zad80] N. Zadeh. What is the worst case behaviour of the simplex algorithm? Bericht, Department of Operations Research, Stanford, 1980.

[ZP96] U. Zwick und M. Paterson. The complexity of mean payoff games on graphs. *Theoretical Computer Science*, 158(1-2):343–359, 1996.

Oliver Friedmann wurde am 19. November 1984 in München gebohren. Er absolvierte ein Diplomstudium der Informatik mit Nebenfach Mathematik an der Universität München (LMU). 2011 promovierte er im Fach Informatik in München über algorithmische Spieltheorie und lineares Programmieren mit summa cum laude. Für seine Doktorarbeit und Forschungstätigkeiten erhielt er zahlreiche internationale Preise und Auszeichnungen.

Einsatz von Lasttransformationen und ihren Invertierungen zur realitätsnahen Lastmodellierung in Rechnernetzen

Stephan Heckmüller

Universität Hamburg, Fachbereich Informatik
SSI Schäfer Noell

heckmueller@informatik.uni-hamburg.de

Abstract: Die im Folgenden zusammengefasste Dissertation [Hec11] befasst sich mit der Charakterisierung von Lasten in Rechnernetzen. Da moderne Rechnernetze aus einer Vielzahl von Einzelkomponenten bestehen, welche die Charakteristika der Last verändern, wird besonderes Augenmerk auf die Abhängigkeit der Lasteigenschaften von der betrachteten Schnittstelle gelegt. Die Veränderung von Lasteigenschaften durch Auftragsverarbeitung wird durch das Konzept der Lasttransformation formalisiert. Hierbei ist Lasttransformation als Transformation einer Primärlast in eine Sekundärlast durch ein verarbeitendes System zu verstehen.

Aufbauend auf dem Konzept der Lasttransformation werden Transformationen, wie sie durch häufig eingesetzte Verarbeitungsmechanismen in heutigen Netzen vorgenommen werden, als Abbildungen auf markovschen Prozessen modelliert. Hierzu werden für solche Primärlasten, die sich als *Batch Markovian Arrival Process* (BMAP) charakterisieren lassen, Beschreibungen der Sekundärlast als BMAP angegeben. Es werden modellbasierte Transformationen für Fragmentierungsmechanismen, verlustbehaftete Übertragungen und Ratenkontrollmechanismen vorgeschlagen und diskutiert. Umfangreiche Validationsstudien bestätigen den hohen Grad an Realitätsnähe der vorgeschlagenen modellbasierten Transformationen.

Neben der Betrachtung der in Rechnernetzen auftretenden Lasttransformationen wird das hierzu inverse Problem der *inversen Lasttransformation* untersucht. Diesbezüglich wird die inverse Transformation von Auftragslängen untersucht. Darüber hinaus werden Verfahren vorgeschlagen, um die Charakteristika des Ankunftsprozesses eines zeitdiskreten Warteschlangensystems ausgehend von der Kenntnis des Abgangsprozesses zu rekonstruieren.

1 Einleitung

Die Verfügbarkeit von realistischen Lastmodellen ist für die Bewertung und Planung von Rechnernetzen unverzichtbar. Darüber hinaus spielt bei der Bereitstellung von Ressourcen – wie beispielsweise von Übertragungskapazitäten – die modellbasierte Lastprognose eine wichtige Rolle. Aufgrund der stetig zunehmenden Vielfalt von verteilten Anwendungen und der immensen Verbreitung von Rechnernetzen ist allerdings eine hohe Komplexität der statistischen Eigenschaften der zu modellierenden Lasten in Rechnernetzen zu verzeichnen, wodurch eine realitätsnahe Lastmodellierung sehr erschwert wird.

Die vorliegende Arbeit beschäftigt sich daher mit der Charakterisierung von Lasten in Rechnernetzen an der Vielzahl von in solchen Netzen existierenden Schnittstellen. Die Abhängigkeit der Lasteigenschaften von der jeweilig betrachteten Schnittstelle ist hierbei als Folge des Aufbaus moderner Rechnernetze besonders zu betonen. Diese Rechnernetze konstituieren einerseits verteilte Systeme, welche sich aus einer Vielzahl von Netzknoten zusammensetzen. Andererseits sind in jedem dieser Knoten die zur Kommunikation benötigten Protokolle in einer Schichtenarchitektur organisiert, wobei die jeweiligen Protokollimplementierungen im Allgemeinen nur mit den direkt angrenzenden Schichten interagieren.

Diese Vielzahl von Einzelkomponenten trägt jeweils einen Teil zur Auftragserfüllung (z. B. zum Verbindungsaufbau zwischen kommunizierenden Instanzen oder zur Übermittlung von Dateneinheiten) bei und verändert so die Eigenschaften der an sie übergebenen Last durch die vorgenommene Weiterverarbeitung. Diese modifizierte Last wird an die nachfolgende Komponente über die gemeinsame Schnittstelle zwischen den beiden Komponenten weitergegeben. Dies impliziert, dass die durch ein Einzelsystem wahrgenommene Last nicht mehr der originären Last entspricht, sondern bereits alle durch vorhergehende Module vorgenommenen Modifikationen in der Last enthalten sind. Daher ist es unabdingbar, diese schnittstellenspezifische Sequenz von Lastmodifikationen in die Betrachtung miteinzubeziehen, um so zu einem validen Modell der Last an einer gegeben Schnittstelle zu gelangen. Ein solches Modell stellt wiederum die Voraussetzung für eine realistische Leistungsbewertung der betrachteten Systeme dar.

Die Modellierung von Lastveränderungen wird in der vorliegenden Arbeit durch das Konzept der *Lasttransformation* formalisiert. Diese beschreibt die durch ein verarbeitendes System vorgenommene Transformation von der eingehenden *Primärlast* zu der ausgehenden *Sekundärlast*. Hierbei sind insbesondere Transformationen von Aufträgen und ihren Attributen und die Transformation von zeitlichen Eigenschaften eines Auftragsstroms zu unterscheiden. Es wird das Ziel verfolgt, analytische Modellierungstechniken zu entwickeln, welche sowohl die Transformation von zeitlichen Eigenschaften als auch von Auftragslängen umfassen.

Die Modellierung der Transformationsprozesse erfolgt als (transformationsspezifische) Abbildung auf stochastischen Prozessen. Ausgehend von der stochastischen Beschreibung der Primärlast und dem jeweiligen Transformationsprozess ist hierzu ein Modell der Sekundärlast anzugeben, welches wiederum ein stochastischer Prozess ist. Bei der Leistungsbewertung von Systemen, welche die transformierte Sekundärlast verarbeiten, kann so sichergestellt werden, dass die Auswirkungen der gegebenen Lasttransformation auf die Lasteigenschaften berücksichtigt werden. Somit kann die Realitätsnähe der Lastbeschreibung an nachfolgenden Schnittstellen erhöht werden. Aufgrund der hohen Flexibilität dieser Prozessbeschreibung operieren die vorzuschlagenden modellbasierten Transformationen auf *Batch Markovian Arrival Processes* (BMAP). BMAPs sind zeitkontinuierliche Markovsche Modelle und erlauben die Modellierung von sog. Batch-Ankünften – also der zeitgleichen Ankunft von mehreren Aufträgen.

Im Folgenden werden modellbasierte Transformationen für eine Reihe von Verarbeitungsmechanismen vorgeschlagen: Zunächst wird die Lasttransformation, wie sie sich durch das Hinzufügen von Headerdaten ergibt, modelliert. Daraufhin werden die Auswirkungen von

Fragmentierungsmechanismen betrachtet und diese als Abbildung auf BMAPs formuliert [HW07a]. Des Weiteren werden die Einflüsse von Paketverlusten auf die Lasteigenschaften betrachtet. Aufbauend auf den gewonnenen Modellen ist es möglich, in effizienter Weise den bei gegebenem Ankunftsprozess und Fehlermodell erzielbaren Durchsatz zu bestimmen. Dieser Sachverhalt kann zur Optimierung von Übertragungsparametern genutzt werden [HW07b].

Ein weiterer betrachteter Verarbeitungsmechanismus ist der Token-Bucket-Regulator. Solche Regulatoren werden in Rechnernetzen zur Lastglättung eingesetzt und stellen sicher, dass eine spezifizierte Paketankunftsrate langfristig nicht überschritten wird. Diesbezüglich wird der Abgangsprozess eines solchen Systems wiederum als markovscher Prozess beschrieben [HW08, HW09]. Die Güte aller Transformationsalgorithmen wird durch umfangreiche Validationsstudien untermauert. Zur Durchführung der vorgeschlagenen Transformationen wurde des Weiteren ein auf dem Konzept der Datenflussarchitektur basierendes Werkzeug entworfen [HSW08].

Die Betrachtung von markovschen, modellbasierten Transformationen wird schließlich abgeschlossen mit Validationsstudien, in denen transformierte Modelle mit realem, gemessenem Verkehr verglichen werden [HW10]. Es zeigt sich dabei, dass auch komplexe Transformationssequenzen im Realsystem mit Hilfe der vorgeschlagenen Methoden mit guter Genauigkeit nachgebildet werden können. Dies untermauert die Nützlichkeit dieser Methoden.

Neben der Frage nach der Prognostizierbarkeit von Lasten an einer gegebenen Schnittstelle aufbauend auf Lastbeschreibungen an vorgelagerten Schnittstellen ist ebenso die Rekonstruierbarkeit der ursprünglichen Lasteigenschaften von Interesse, wie sie vor der Durchführung gegebener Lasttransformationen vorlagen. In der vorliegenden Arbeit wird diese Fragestellung zunächst als inverses Problem der Lasttransformation formalisiert – der *inversen Lasttransformation*. Diesbezüglich werden zunächst Algorithmen zur Rekonstruktion von Auftragslängen vorgeschlagen. Es werden Aufträge betrachtet, welche durch kommunizierende Anwendungsinstanzen an das unterliegende Netz übergeben wurden [HMB+10].

Darüber hinaus werden inverse Transformationen im Kontext diskreter Warteschlangensysteme untersucht. Hierbei ist die Zielstellung, ausgehend von der Kenntnis des Abgangsstroms auf Charakteristika eines unbekannten Ankunftsstromes zu schließen. Diesbezüglich erfolgt der Einsatz von Tobit-Regressionsmodellen [Ame84]; diese Modelle sind insbesondere in den Sozial- und Wirtschaftswissenschaften populär. Ein Einsatz im Kontext der Rechnernetze ist dem Autor nicht bekannt. Mit Hilfe der vorgeschlagenen Methoden ist es möglich, Rückschlüsse auf Charakteristika des Ankunftsprozesses zu ziehen, die bisher nicht rekonstruierbar waren. Die erzielten Ergebnisse ermöglichen die Parametrisierung von Algorithmen zur Leistungsbewertung oder zur Lastprognose und können so zur Optimierung des Systemverhaltens genutzt werden [HW11].

2 Lasttransformation

Innerhalb von Netzen von Bedienstationen ist die Last, wie sie an einer einzelnen Station anliegt, im Allgemeinen von der Diensterbringung anderer Stationen abhängig. Dies gilt im Besonderen für gegenwärtige Rechnernetze, welche in Schichten und verteilten Einzelsystemen organisiert sind. Jede Station beeinflusst hier die Last der jeweilig nachfolgenden Station, wie in Abbildung 1 beispielhaft dargestellt. Die so hervorgerufene Veränderung der Lastcharakteristiken wird im Folgenden als *Lasttransformation* von Primär- zu Sekundärlast bezeichnet. In der vorgenannten Abbildung modellieren diese Transformationen beispielsweise Fragmentierung und Headergenerierung in einer *IP*-Protokollinstanz oder Paketverluste bei der drahtlosen Übertragung.

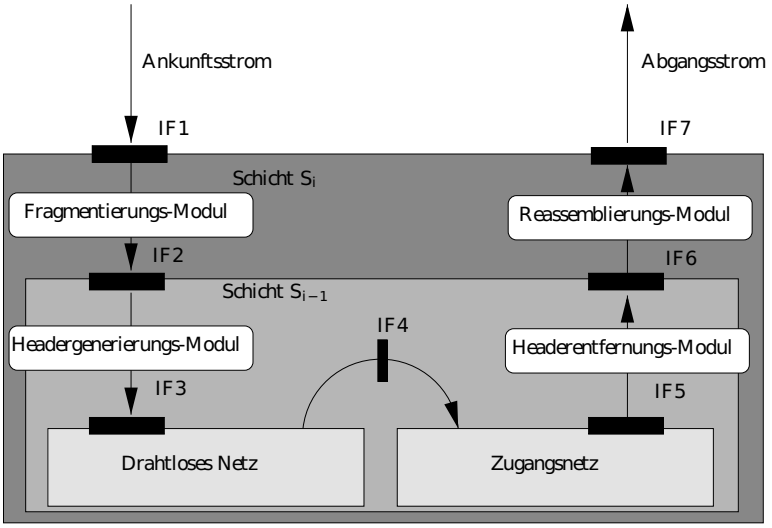


Abbildung 1: Übertragung eines Paketes als Ablauf von Transformationen; IF1,...,IF7 bezeichnen die Schnittstellen (*interfaces*)

Um diese Transformationen formal beschreiben zu können, wird Last, wie in Definition 1 [Wol99] dargestellt, definiert.

Definition 1 *Die* Last $L = L(E, S, IF, T)$ *wird definiert als eine Sequenz von Aufträgen, die während des Beobachtungsintervalls T an das Bediensystem S durch seine Umgebung E übergeben werden. Die Aufträge werden über die Schnittstelle IF übergeben, welche das Bediensystem von seiner Umgebung trennt.* ◊

Die Last kann somit durch eine Sequenz von Aufträgen a_i, die während des betrachteten Zeitintervalls T eintreffen, beschrieben werden. Für wohldefinierte Lasten sei der Ankunftsprozess definiert als Tupel aus Ankunftszeitpunkten t_i und den Aufträgen a_i, wobei

$\mathcal{A}$ die Wertemenge der Aufträge sei.

$$\{(a_i, t_i) | a_i \in \mathcal{A}, t_1 \leq t_2 \leq ... \leq t_N, t_{1,...,N} \in T\} \quad (1)$$

Einzelne Aufträge a_i können hierbei beispielsweise Datenübertragungs- oder Verbindungsaufbauwünsche repräsentieren. Das vorgestellte Konzept ist weiterhin nicht auf Modellierung von Rechnernetzen beschränkt. Zur Modellierung von Datenbanksystemen könnten beispielsweise Transaktionen als Aufträge modelliert werden, um so zu einer Lastbeschreibung gemäß Definition 1 zu gelangen. Als Lasttransformation bezeichnen wir hierauf aufbauend die Transformation der an ein verarbeitendes System übergebenen Primärlast in die abgehende Sekundärlast, wie sie durch die Verarbeitung erfolgt.

Im Kontext moderner, auf dem *Internet Protocol* basierender Netze lassen sich eine Reihe von verschiedenen Attributtypen identifizieren. Das für die Leistungsbewertung wohl wichtigste Attribut sind die Auftragslängen, weil diese in einer Vielzahl von Systemen unmittelbar den zur Verarbeitung notwendigen Aufwand bestimmen. Die Transformation von Längenattributen vollzieht sich zumeist bei der Verarbeitung von Aufträgen der nächsthöheren Schicht. Beispiele hierfür sind das Hinzufügen von Headerdaten, sowie die Fragmentierung, Segmentierung und die Kompression von zu versendenden Nutzdaten. Welche Attribute bei der Lastspezifikation zu berücksichtigen sind, hängt vom zu modellierenden Szenario ab. In der hier zusammengefassten Arbeit erfolgt aus vorgenannten Gründen eine Beschränkung auf das Längenattribut.

Von großer Wichtigkeit ist darüber hinaus das zeitliche Verhalten. Obwohl es prinzipiell möglich wäre, auch diese Eigenschaft direkt als Auftragsattribut zu repräsentieren, nimmt das zeitliche Verhalten in der gegebenen Lastdefinition ausdrücklich eine Sonderstellung ein. Dies spiegelt einerseits die Notwendigkeit wider, den Auftragsstrom als zeitbehafteten Vorgang zu betrachten. Andererseits ist eine Trennung zwischen Attributen und zeitlichem Verhalten nützlich, um zu einer klaren Unterscheidung von Transformation der Auftragsattribute und des zeitlichen Verhaltens zu gelangen.

Die Eigenschaften der Sekundärlast lassen sich oftmals nicht direkt beobachten. In solchen Fällen kann im Rahmen eines modellbasierten Ansatzes der Einfluss des Systems auf die induzierte Primärlast durch ein Transformationsmodell erfasst werden. Die hieraus resultierende modellbasierte Transformation ermöglicht die Abbildung der für die jeweilige Untersuchung relevanten Primärlasteigenschaften auf die korrespondierenden Sekundärlasteigenschaften. Unter der Voraussetzung hinreichender Validität kann das so gewonnene Sekundärlastmodell zur Lastprognose eingesetzt werden Die beschriebenen Zusammenhänge sind in Abbildung 2 illustriert.

In Abbildung 1 war bereits zu erkennen, dass oftmals mehrere Transformationsvorgänge sequentiell auf die initiale Primärlast wirken. Die Modellierung solcher Sequenzen ist mit Hilfe des Konzeptes modellbasierter Lasttransformationen auch aufgrund der klaren Trennung zwischen verarbeitendem System und der Umgebung gut handhabbar: Die zunächst isoliert betrachteten elementaren Transformationsschritte werden verknüpft, indem die jeweils prognostizierte Sekundärlast als Primärlast des nachfolgenden Systems aufgefasst wird.

Komplexe Transformationen können somit als Hintereinanderausführung mehrerer ele-

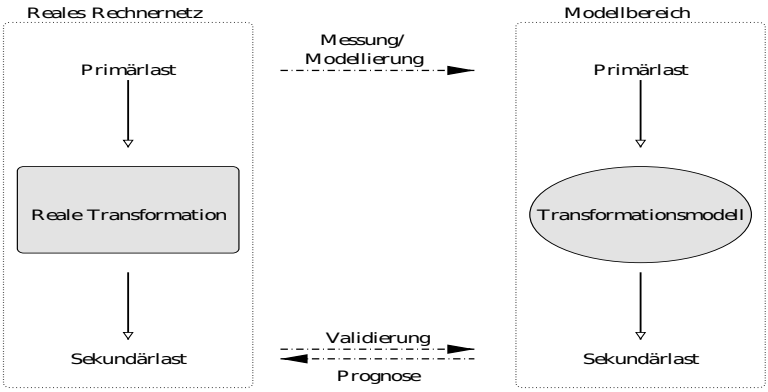


Abbildung 2: Lasttransformationen in realen Rechnernetzen und im Modellbereich

mentarer, modellbasierter Transformationen aufgefasst werden, welche zumeist einfacher zu charakterisieren sind. Dieses modulare Vorgehen ermöglicht es, Systeme zu modellieren, die ansonsten nicht oder nur sehr schwer charakterisierbar sind. Des Weiteren können durch dieses Vorgehen viele Transformationsmodelle wiederverwendet werden, weil bestimmte Mechanismen – wie Fragmentierung oder Headergenerierung – in einer Vielzahl von Systemen eingesetzt werden.

3 Lasttransformation auf BMAPs

Obwohl es natürlich möglich ist, Lasttransformationen in direkter, algorithmischer Art und Weise – d.h. durch direkte Transformation einer Folge von Primärlastaufträgen insbesondere mittels Simulation – vorzunehmen, birgt diese Herangehensweise einige Nachteile:

- Die Sekundärlastbeschreibung, die durch die Transformation konkreter Auftragssequenzen gewonnen wird, ist im Allgemeinen nicht der analytischen Behandlung zugänglich. Dies gilt insbesondere auch für warteschlangentheoretische Methoden (vgl. z. B. [DKL09, ST10]), was die Berechnung von Leistungskennwerten erschwert.
- Auch wenn für die Primärlast ein stochastisches Modell vorliegt, muss die Beschreibung des Sekundärlastmodells deterministisch sein. Dies folgt aus der Tatsache, dass simulative Transformationsmethoden nur auf konkrete Ankunftssequenzen – also auf Realisierungen der unterliegenden Prozessbeschreibung – angewendet werden können. Dies reduziert die Allgemeingültigkeit der Lastbeschreibung.

Aus den vorgenannten Gründen werden in der Dissertation Lasttransformationen vorgeschlagen, welche direkt auf die jeweiligen Prozessbeschreibungen angewendet werden können und so die analytische Handhabbarkeit und die Allgemeingültigkeit erhalten. Da aber gleichzeitig eine breite Anwendbarkeit der Modelle wünschenswert ist, gilt es, zu starke Beschränkungen durch die zu wählende Modellklasse zu vermeiden. Diese beiden

Ziele werden durch die verwendeten *Batch Markovian Arrival Processes (BMAP)* erreicht (vgl. [Luc93]). Diese gestatten die realitätsnahe Beschreibung der Last gemäß einer Vielzahl von Anwendungstypen und sind gleichzeitig analytisch handhabbar.

Die entwickelten Lasttransformationen lassen sich als Abbildungen auf $BMAPs$ verstehen

$$T_{BMAP} : \mathcal{BMAP}^p \rightarrow \mathcal{BMAP}^s. \tag{2}$$

In der hier zusammengefassten Dissertation werden Lasttransformationen in realen Systemen als Abbildungen auf *Batch Markovian Arrival Processes* (BMAPs) modelliert. Neben der Charakterisierung des zeitlichen Verhaltens erlauben BMAPs als zweidimensionale Beschreibungstechnik die Modellierung eines Auftragsattributes unter Verwendung der Batch-Größe. Prinzipiell ist es möglich, beliebige Attribute mit abzählbarer Wertemenge zu erfassen. Aufgrund der herausragenden Wichtigkeit der Auftragslänge gilt jedoch die Aufmerksamkeit diesem Attribut. Wird der Verarbeitungsaufwand durch abweichende Auftragscharakteristika bestimmt, kann die Batch-Größe alternativ als Indikator für die Komplexität des Auftrages verwendet werden. In beiden Fällen wird somit sowohl das zeitliche Verhalten als auch der für die einzelnen Aufträge notwendige Arbeitsaufwand erfasst, so dass davon ausgegangen werden kann, dass die Sekundärlast für die Leistungsbewertung nachfolgender verarbeitender Systeme hinreichend genau charakterisiert wird.

Es ergibt sich so das in Abbildung 3 dargestellte Procedere: Die durch eine der genannten Prozessbeschreibungen modellierte Primärlast wird als BMAP repräsentiert und kann dann mit Hilfe der modellbasierten Lasttransformation so modifiziert werden, dass die aus der gegebenen Lasttransformation resultierende Sekundärlast als BMAP-Modell gegeben ist. Zur Modellierung der Primärlast als BMAP existieren in der Literatur eine Vielzahl von Ansätzen (vgl. z. B. [KLL03, ODT09, CZS10]). Des Weiteren beinhaltet die Klasse der $BMAPs$ viele Prozessklassen, welche zur Modellierung von Lasten in Rechnernetzen verwendet werden (wie z. B. Markov-Modulierte Poisson-Prozesse oder Phase-Type-Verteilungen, vgl. oberen Teil von Abbildung 3).

Im Rahmen der Dissertation werden Transformationsalgorithmen vorgestellt, welche es erlauben, eine Reihe der wichtigsten Transformationsvorgänge in heutigen Rechnernetzen zu modellieren. Die Realitätsnähe wird durch Experimente in einem realen Netz demonstriert: Während der Übertragung eines MPEG-Videostroms wurde an mehreren Schnittstellen des Übertragungsweges die auftretende Last aufgezeichnet und deren Eigenschaften mit den Eigenschaften der mittels Transformationsalgorithmen erzeugten Sekundärlastmodelle verglichen. Es konnte durchweg ein hoher Grad an Übereinstimmung erzielt werden. Dies gilt auch für Szenarien, in denen Ratenkontrollmechanismen eingesetzt wurden.

4 Inverse Lasttransformation

Die bisherigen Betrachtungen galten dem Einfluss von Verarbeitungsmechanismen in Rechnernetzen auf die abgehende Sekundärlast. In vielen Fällen kann jedoch nur der bereits transformierte Verkehr beobachtet werden. Hier stellt sich oftmals eher die Frage, ob und wie sich die Eigenschaften der untransformierten Primärlast rekonstruieren lassen. Un-

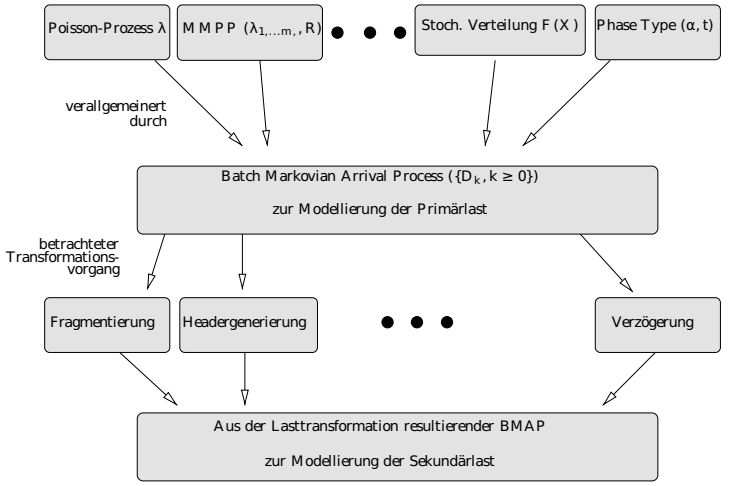


Abbildung 3: Schematische Darstellung der modellbasierten Lasttransformation auf BMAPs und den enthaltenen Prozessklassen

ter Nutzung der in dieser Arbeit verwendeten Begrifflichkeiten lässt sich dies als inverse Problemstellung zur Lasttransformation sehen. Wir bezeichnen dies daher als *inverse Lasttransformation von Sekundär- zu Primärlast*.

Wie bei allen inversen Problemstellungen stellt sich bei der inversen Lasttransformation aber auch die Frage der Durchführbarkeit. Im Gegensatz zur regulären Lasttransformation, deren Machbarkeit aufgrund ihres Einsatzes im realen Netz evident ist, kann diese hier nicht vorausgesetzt werden.

Es wird zunächst aufgezeigt, dass sich in modernen Rechnernetzen eine Vielzahl von inversen Lasttranformationen identifizieren lassen. Zunächst werden diese danach unterschieden, ob sie die Auftragsattribute selbst oder das zeitliche Verhalten der Aufträge der Sekundärlast betreffen. Bezüglich der Auftragsattribute lassen sich die inversen Transformationen weiter nach betroffenem Attributtyp (z. B. Adress- oder Längenattribut) unterscheiden.

Die Länge der Aufträge, welche durch kommunizierende Anwendungsinstanzen an das unterliegende Kommmunikationsnetz übergeben werden, ist zumeist nicht mehr direkt beobachtbar. Vielfach gilt diesem Attribut jedoch das eigentliche Interesse, z. B. im Kontext von Applikationsklassifikation oder der Modellierung des Nutzerverhaltens. Diese Rekonstruktion stellt insbesondere bei Verwendung von TCP – dem meist verwendeten Transportprotokoll – ein schwieriges Problem dar. Im Rahmen der Dissertation wurden Rekonstruktionsalgorithmen vorgeschlagen, welche sich Eigenschaften des TCP-Protokolls zu Nutze machen. Die Algorithmen nutzen hierbei Flags im Header der TCP-Pakete bzw. die Länge einzelner TCP-Segmente. Unter Verwendung realer Messdaten konnte demonstriert werden, dass für viele Applikationstypen eine Rekonstruktion der Auftragslängen mit guter Genauigkeit möglich ist.

Darüber hinaus werden Methoden zur Rekonstruktion von Ankunftsprozessen diskreter

Wartesysteme vorgeschlagen. In den betrachteten Systemen treffen pro Betrachtungsintervall eine variable Zahl von Aufträgen ein; die Anzahl der pro Intervall bedienbaren Aufträge wird als konstant angenommen. Durch Anwendung von aus der Literatur bekannten Regressionsmodellen – insbesondere dem Tobit-Modell [Ame84] – können auch für nicht unabhängig verteilte Ankunftsanzahlen weitgehende Aussagen über den Ankunftsprozess getroffen werden – ohne dass dieser beobachtet wird.

Zur Rekonstruktion ist lediglich die Kenntnis der Anzahl der Aufträge notwendig, die pro Intervall das System verlassen. Hierauf aufbauend werden Zeitpunkte identifiziert, an denen das Bediensystem leer ist. Für das folgende Intervall ist es nun möglich festzustellen, ob die Anzahl von Abgängen denen der Ankünfte des letzten Intervalls entspricht oder ob das System im entsprechenden Intervall vollständig ausgelastet ist. Die Anzahl der Abgänge in solchen Intervallen bildet die Basis für die Anwendung der vorgenannten Regressionsmodelle. Die vorgeschlagenen Methoden können zur Charakterisierung von Lasten in entfernten Netzknoten verwendet werden (z. B. im Kontext von Polling-Mechanismen) und sind auch im Falle von Tandemnetzen – also einer Sequenz von mehreren Bedienstationen – einsetzbar.

5 Fazit

Durch die im vorliegenden Beitrag zusammengefasste Dissertation konnten signifikante Fortschritte in den Forschungsgebieten Lastmodellierung und Lasttransformation erzielt werden. Mit Hilfe der entwickelten Lasttransformationen auf markovschen Ankunftsprozessen ist es möglich, den Einfluss von Verarbeitungsmechanismen auf die Lasteigenschaften sowohl in Bezug auf das zeitliche Verhalten als auch im Hinblick auf Auftragslängen zu modellieren. Dies erhöht die Validität der verwendeten Lastmodelle. Des Weiteren wurde das Konzept der inversen Lasttransformation eingeführt – also die Rekonstruktion von Primärlasteigenschaften ausgehend von einer bekannten Sekundärlast. Im Kontext der inversen Lasttransformation wurden Algorithmen entwickelt, welche es ermöglichen, Eigenschaften von nicht-beobachtbaren Lasten zu rekonstrieren. Die zahlreichen durchgeführten Studien zur Validierung der entwickelten (inversen) Lasttransformationsansätze unterstrichen, dass wir in nahezu sämtlichen betrachteten Szenarien eine erfreulich gute Realitätsnähe unserer Transformationsmodelle verzeichnen können und die Modelle somit auch eine gute Praxisrelevanz besitzen dürften.

Literatur

[Ame84] T. Amemiya. Tobit models: A survey. *Journal of Econometrics*, 24(1-2):3–61, 1984.

[CZS10] G. Casale, E.Z. Zhang und E. Smirni. Trace data characterization and fitting for Markov modeling. *Perf. Eval.*, 67(2):61 – 79, 2010.

[DKL09] A. Dudin, V. Klimenok und M. H. Lee. Recursive formulas for the moments of queue length in the BMAP/G/1 queue. *IEEE Comm. Letters.*, 13(5):351–353, 2009.

[Hec11] Stephan Heckmüller. *Einsatz von Lasttransformationen und ihren Invertierungen zur realitätsnahen Lastmodellierung in Rechnernetzen*. Dissertation, Univ. Hamburg, 2011.

[HMB+10] S. Heckmüller, G. Münz, L. Braun et al. Lasttransformation durch Rekonstruktion von Auftragslängen anhand von Paketdaten. *Zeitschrift Praxis der Informationsverarbeitung und Kommunikation (PIK)*, 11:113–120, 2010.

[HSW08] S. Heckmüller, M. Spork und B.E. Wolfinger. Load Transformation of Markovian Arrival Processes: Methods and Tool Support. In *SMCTools 2008*, Oct. 2008.

[HW07a] S. Heckmüller und B. E. Wolfinger. Load Transformations for Markovian Arrival Processes. In *Proceedings of ASMTA 2007*, Seiten 35–43, June 2007.

[HW07b] S. Heckmüller und B.E. Wolfinger. Modellierung verlustinduzierender Lasttransformationen für markovsche Ankunftsprozesse. In *MMBnet 2007 Workshop*, Seiten 36–49, 2007.

[HW08] S. Heckmüller und B. E. Wolfinger. Analytical Modeling of Token Bucket Based Load Transformations. In *Proceedings of SPECTS 2008*, Seiten 15–23, June 2008.

[HW09] S. Heckmüller und B. E. Wolfinger. Using Load Transformations for the Specification of Arrival Processes in Simulation and Analysis. *Simulation*, 85(8):485–496, 2009.

[HW10] S. Heckmüller und B. E. Wolfinger. Analytical Load Transformations of Video Streams: Validation Using Measured Traffic. In *Proceedings of SPECTS 2010*, Seiten 202–209, July 2010.

[HW11] S. Heckmüller und B. E. Wolfinger. Reconstructing arrival processes to discrete queueing systems by inverse load transformation. *Simulation*, 87(12):1033–1047, 2011.

[KLL03] A. Klemm, C. Lindemann und M. Lohmann. Modeling IP traffic using the batch Markovian arrival process. *Perform. Eval.*, 54(2):149–173, 2003.

[Luc93] D. M. Lucantoni. The BMAP/G/1 queue: a tutorial. In L. Doniatiello und R. Nelson, Hrsg., *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, Seiten 330–358. Springer-Verlag, New York, 1993.

[ODT09] H. Okamura, T. Dohi und K. S. Trivedi. Markovian Arrival Process Parameter Estimation With Group Data. *IEEE/ACM Trans. Netw.*, 17(4):1326–1339, 2009.

[ST10] Z. Saffer und M. Telek. Unified analysis of BMAP/G/1 cyclic polling models. *Queueing Syst. Theory Appl.*, 64(1):69–102, 2010.

[Wol99] B. E. Wolfinger. Characterization of Mixed Traffic Load in Service-Integrated Networks. *Systems Science Journal*, 25(2):65–86, 1999.

Stephan Heckmüller wurde 1979 in Homberg/Efze geboren. Er studierte Informatik an der Universität Hamburg und schloss das Diplomstudium im Jahr 2006 ab. Seine Diplomarbeit “Bereitstellung von Dienstgüte für aggregierte Multimedia-Ströme in lokalen Broadcast-Netzen” wurde durch die GI/ITG-Fachgruppe “Kommunikation und Verteilte Systeme” ausgezeichnet. Er war von 2006 bis 2011 als wissenschaftlicher Mitarbeiter an der Universität Hamburg tätig. Die Promotion erfolgte im Jahr 2011. Seit 2011 ist Herr Heckmüller als Software-Architekt im Bereich der Intralogistik tätig.

Direkte Ende-zu-Mitte Authentifizierung in kooperativen Netzen

Tobias Heer
COMSYS, RWTH Aachen University
heer@cs.rwth-aachen.de

Abstract: Kooperative Netze beruhen auf dem Prinzip der Zusammenarbeit von Benutzern auf Netzwerkebene. Sie ermöglichen dabei Kommunikation, wo andere Netzformen an wirtschaftliche oder technische Grenzen stoßen. Beispiele für kooperative Netzwerke sind dezentrale drahtlose Mesh-Netzwerke, Wi-Fi-Communities oder hybride Formen dieser Netzwerk-Typen. In kooperativen Netzen übernehmen Benutzergeräte Kernfunktionen des Netzes, wie z.B. die Weiterleitung von Paketen. Diese Kooperation bedingt eine opportunistische Offenheit, welche zu neuen Sicherheitsrisiken führt. Besonders die fehlende Möglichkeit zur Authentifizierung von Datenverkehr durch die weiterleitenden Geräte macht sie anfällig für Angriffe. Diese Arbeit schafft die Grundlagen für eine hocheffiziente Authentifizierung von Netzwerkverkehr durch die Knoten im Netz. Dies erlaubt es, die Identität eines Senders und die Integrität seiner Nachrichten effizient zu überprüfen, bevor die Nachrichten weitergeleitet werden. So können unauthentifizierte Datenströme und Angriffe effizient unterbunden werden.

1 Einleitung: Nutzen und Gefahren der Kooperation

Die Möglichkeit der spontanen drahtlosen Vernetzung von mobilen und stationären Geräten erlaubt es neue Netzwerkkonzepte zu etablieren, welche die Grenzen zwischen Netzwerkanbieter und Netzwerkbenutzer aufheben. Solche kooperativen Netze beruhen auf dem Prinzip der Zusammenarbeit von Benutzern auf Netzwerkebene, um Dienste, wie z.B. das Weiterleiten von Paketen oder den gemeinsamen Zugriff auf andere Netzwerkressourcen, wie Speicherplatz und Internetzugang, gemeinschaftlich zu erbringen. Beispiele für kooperative Netzwerke sind Ad-Hoc-Netze, dezentrale drahtlose Mesh-Netzwerke, Micro-Operator-Netzwerke, WLAN-Communities oder hybride Formen. Kooperative Netzwerke können dabei Lösungen schaffen wo andere Netzformen an technischen oder wirtschaftlichen Problemen scheitern. So gelingt es zum Beispiel Bewegungen wie Freifunk oder Funkfeuer mit einem minimalen Budget ganze Stadtteile mit Wi-Fi drahtlos zu vernetzen. Dabei kooperieren alle Benutzer, um sowohl gemeinsame Ziele (der Ausbau des Netzes) als auch egoistische Ziele (z.B. der Zugang zum Internet) zu erreichen.

Jedoch schafft das technische Konzept eines gemeinschaftlich organisierten Netzes auch neue Angriffsmöglichkeiten für egoistische und bösartige Benutzer. Zum Beispiel sind drahtlose Multi-Hop-Netzwerke (siehe Abb. 1) besonders anfällig gegenüber Angriffen, die auf dem Fluten des Netzwerks mit Schadpaketen oder der Manipulation und Fälschung von Paketen beruhen. Dabei lassen sich viele der möglichen Angriffe gegen kooperati-

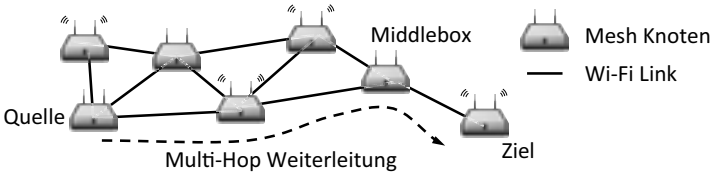


Abbildung 1: In kooperativen drahtlosen Mesh-Netzen kooperieren Knoten, um weite Distanzen drahtlos zu überbrücken. Dabei verschwimmt die Grenze zwischen Betreiber und Benutzer, da die Wi-Fi Router im Netz in der Regel ebenfalls Benutzern gehören.

ve Netze auf deren Offenheit gegenüber der Mitwirkung Unbekannter und dem Fehlen adäquater Authentifizierungsmechanismen auf verschiedenen technischen Ebenen zurückführen. Ein wichtiger Schritt zur Vermeidung von Missbrauch ist es daher, weiterleitenden Geräten, sogenannten *Middleboxen*, zu ermöglichen, sowohl die Identität der kommunizierenden Geräte als auch die Herkunft und Authentizität des Netzwerkverkehrs zu überprüfen. Protokolle, die dieses Ziel erreichen, fallen dabei in die Klasse der Ende-zu-Mitte Authentifizierungsprotokolle.

1.1 Ansätze zur Ende-zu-Mitte Authentifizierung

Effiziente Standardlösungen, um eine paketbezogene Authentifizierung zu erreichen, setzen typischerweise geteilte symmetrische Schlüssel zwischen den signierenden und verifizierenden Geräten voraus oder beruhen auf einer stets verfügbaren Verbindung zu einem Authentifizierungsserver. Diese Einschränkungen führen jedoch in kooperativen Netzen zu deutlichen Nachteilen bezüglich der Flexibilität, Effizienz, und Zuverlässigkeit. Daher ist oft eine *direkte* Authentifizierung zwischen den Endgeräten und Middleboxen durch alternative Authentifikationsmechanismen sinnvoll.

Eine weit verbreitete Methode zur Feststellung der Identität von Kommunikationspartnern bzw. der Authentizität und Integrität ihrer Nachrichten sind Public-Key Systeme. In der zugrundeliegenden Dissertation und den zugehörigen Veröffentlichungen wird ausführlich auf die Eigenschaften von Public-Key Verfahren zur Sicherung von *sporadischen* Authentfizierungsereignissen, wie z.B. der Authentifizierung eines Geräts beim Verbindungsaufbau eingegangen [HHK$^+$09]. Dieser Artikel setzt den Fokus jedoch auf die Authentifizierung von *hochfrequenten* Ereignissen, z.B. der paketweisen Verifikation von breitbandigem Nutzlastverkehr. Hier stellen die beachtlichen Anforderung der Public-Key Kryptographie an die Rechenleistung der Endgeräte und Middleboxen eine bedeutende Einschränkung dar. So benötigt die Verifikation eines Datenstroms mit einem Durchsatz von 30 Mbit/s bereits etwa 3.000 unabhängige Verifikationen, um jedes Nutzlastpaket zu authentifizieren. Gleichzeitig liegt die Leistung von typischen drahtlosen Wi-Fi Routern (z.B. mit AMD Geode 500 MHz CPU) für weit verbreitete asymmetrische Signaturen ohne spezielle Hardwarebeschleunigung, um zwei Größenordnungen *unter* diesem Wert (DSA 1024 Bits: 55 Verifikationen, ECDSA 160 Bits: 55 Verifikationen). Basierend auf diesen Werten könnte dieser Router daher nur schmalbandige Datenströme mit einem Volumen von weniger als 660 Kbit/s paketweise verifizieren.

Um eine ausreichend effiziente Ende-zu-Mitte Verifikation von Datenpaketen zu erreichen, müssen daher alternative Authentifizierungsmethoden geschaffen werden, welche die beschränkten Hardwareressourcen in kooperativen Netzen berücksichtigen. Im Gegensatz zu asymmetrischen Public-Key Verfahren zeichnen sich symmetrische Authentifizierungsverfahren durch eine hohe Effizienz aus. Jedoch lassen sich diese Verfahren in der Regel nicht ohne paarweise Schlüssel betreiben. Eine Ausnahme stellen hierbei Verfahren dar, die auf kryptographischen Einwegfunktionen, also Hash-Funktionen, und deren Verkettung zu Hash-Ketten beruhen. Basierend auf Hash-Ketten lassen sich Authentifizierungsprotokolle entwickeln, die keine gemeinsamen *geheimen* Schlüssel benötigen, sondern für die ein gesicherter *öffentlicher* Wert ausreichend ist.

Auf Hash-Ketten basierende Authentifizierungsprotokolle wurden bislang erfolgreich zur Authentifizierung auf Ende-zu-Ende-Basis und für die effiziente Sicherung von Multicast eingesetzt. Diese Arbeit erweitert und ergänzt bestehende Ansätze, um sie zur Authentifizierung in Ende-zu-Mitte Szenarien anwendbar zu machen. Dazu stellt dieser Artikel zwei geeignete Verfahren vor: Das “Adaptive and Lightweight Protocol for Hop-by-Hop Authentication” (ALPHA) verwendet effiziente Hash-Funktionen und Hash-Ketten, um eine schnelle Überprüfung der *Quelle* und *Integrität* eines Netzwerkpakets zu erreichen. Die “Stream-based Per-packet One-time Tokens for Cryptographic Source Authentication” (SPOTS) verzichten auf Ende-zu-Mitte Integritätsschutz, um die kryptographische Komplexität der Authentifizierung weiter zu senken. Dies bedeutet, dass SPOTS nur eine Überprüfung der *Paketquelle* durchführt, was SPOTS ermöglicht, noch effizientere Ansätze zu verfolgen. Beide Mechanismen lassen sich zusätzlich flexibel parametrisieren, um effiziente Ende-zu-Mitte Authentifizierung für ein breites Spektrum von Szenarien, innerhalb und außerhalb von kooperativen Netzwerken, zu ermöglichen.

2 Grundlagen: Hash-Ketten basierte Authentifizierung

Hash-Ketten nach Lamport [Lam81] sind flexible Konstrukte, die sich zur Verifikation der Quelle und Integrität eines Pakets eignen. Um eine Hash-Kette zu erzeugen wird ein Zufallswert s gewählt, auf den eine kryptographische Hash-Funktion H iterativ angewandt wird. Das Ergebnis $h_1 = H(s)$ bildet das erste Glied der Kette. Weitere Glieder werden durch wiederholte Anwendung von H auf das vorige Glied erzeugt: $h_i = H(h_{i-1}) = H^i(s)$. Das letzte Glied der Hash-Kette, h_n, wird dabei als Anker bezeichnet. Nach dem gegenseitigen gesicherten Austausch eines Ankers können sich zwei Kommunikationspartner zu einem beliebigen Zeitpunkt über die Preisgabe des nächst niedrigen Elements in der Hash-Kette h_{n-1} authentifizieren bzw. wiedererkennen.

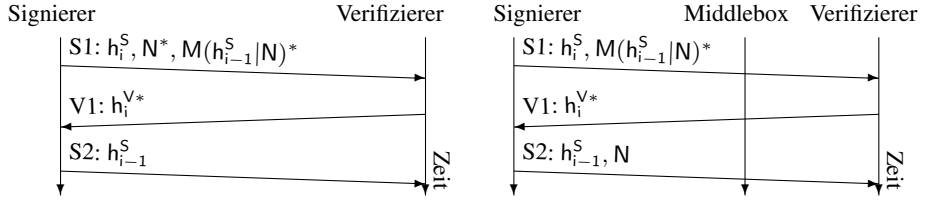
Um die Integrität von Nachrichteninhalten zu schützen, können noch nicht preisgegebene Elemente einer Hash-Kette als Schlüssel für ein symmetrisches Signaturverfahren (z.B. HMAC) verwendet werden. Diese Technik nennt sich ”Delayed Secret Disclosure”, also die zeitverzögerte Preisgabe eines vormals geheimen Schlüssels. Abbildung 2a veranschaulicht den Vorgang in einer leicht vereinfachten Variante. Ein Signierer übermittelt dabei eine geschützte Nachricht N an einen Verifizierer. Beide erzeugen zuvor eine Hash-Kette mit Elementen h_i^S und h_i^V und tauschen deren Ankerelemente aus. Die Superskripte

S und V deuten dabei die Herkunft an (Signierer, Verifizierer). Um die Nachricht N zu übermitteln, erzeugt der Signierer eine HMAC Signatur $M(h_{i-1}^S | N)$ der Nachricht, wobei er das nächste unveröffentlichte Hash-Ketten Element h_{i-1}^S als Schlüssel verwendet. Der Signierer übermittelt N und die Signatur in einem ersten Paket (S1) an den Verifizierer. Dieser speichert beide Werte und bestätigt den Empfang mit einem Element seiner Hash-Kette h_i^V in seiner Antwort (V1). Nach Erhalt der Bestätigung veröffentlicht der Sender das Hash-Ketten Element h_{i-1}^S in einem weiteren Paket (S2). Nach Erhalt des S2 Pakets kann der Empfänger die Authentizität von N prüfen. Zum Senden einer weiteren Nachricht beginnt der beschriebene Ablauf erneut. Durch die zeitliche Trennung zwischen Signaturerstellung bzw. Übermittlung und Preisgabe des Signaturschlüssels wird sichergestellt, dass nur der legitime Signierer zum Erstellungszeitpunkt der Signatur alle notwendigen Informationen besitzt. Aufgrund des interaktiven Charakters des Signaturablaufs werden diese Signaturen auch interaktive Hash-Ketten oder Interactive Hash Chain (IHC) Signaturen genannt.

IHC Signaturen zeichnen sich durch einen sehr geringen Berechnungsaufwand aus, da die Erstellung bzw. Verifikation nur wenige Anwendungen einer Hash-Funktion bedingt. Im Vergleich zu den zuvor genannten Werten für DSA und ECDSA kann der oben genannte Router ca. 100.000 Hash-Funktionen pro Sekunde berechnen. Wäre ausschließlich dieser Berechnungsaufwand ausschlaggebend, wäre die Rechenleistung des oben beschriebenen Routers mehr als ausreichend, um breitbandigen Verkehr mit hunderten Mbit/s zu verifizieren.

3 ALPHA: Effiziente Ende-zu-Mitte Authentifizierung

Die hohe Berechnungseffizienz von IHC Signaturen verspricht zwar eine effiziente Verifikation von Paketen im Netzwerk, jedoch besitzen diese Signaturen einige praktische Nachteile für den Einsatz in der Ende-zu-Mitte Authentifizierung. Dies lässt sich einfach an zwei Beispielen zeigen. Zum Einen soll die Authentifizierung das Netzwerk vor Angriffen wie dem unerlaubten Fluten mit Schadpaketen sichern. Zwar leistet die beschriebene Form der IHC Signatur eine Verifikation der Nachricht N, jedoch muss diese Nachricht zu-



(a) IHC Signatur ohne Commitment (b) IHC Signatur mit Commitment.

Abbildung 2: IHC Signaturen ohne Commitment (a) bedingen eine Zwischenspeicherung von Nachricht und Authentisierungscode und eignen sich daher nicht für den Einsatz in Ende-zu-Mitte Szenarien. Schemata mit Commitment (b) reduzieren die zu speichernden Daten stark. Der Stern (*) markiert Werte, die von Endsystemen und Middleboxen bis zum Empfang von S2 gespeichert werden müssen.

erst an den Verifizierer übermittelt werden, bevor deren Integrität überprüft werden kann. Dies bedeutet, dass Schadpakete nicht im Netzwerk aufgehalten werden können. Zum Anderen müssen alle Netzwerkteilnehmer, die an der Verifikation der Nachricht interessiert sind, die Nachricht bis zur Preisgabe des Schlüssels h_{i-1}^S im S2 Paket speichern. Im Ende-zu-Mitte Fall trifft dies auch auf die Middleboxen zu, was bedeutet, dass diese Netzwerkelemente große Datenmengen zwischenspeichern müssten und dadurch angreifbar für Denial of Service (DoS) Angriffe würden. Somit würden sich die beschriebenen IHC Signaturen zwar zur authentifizierten Übertragung von Informationen vom End-Gerät an die Middlebox eignen, jedoch könnten sie nicht größere Mengen an Netzwerkverkehr authentifizieren.

Eine leichte Abwandlung des Schemas kann beide Probleme beheben. Abbildung 2b zeigt eine IHC Variante, die zuerst nur ein "Commitment" im S1 Paket verschickt und später die Nachricht N während der Veröffentlichung des Schlüssels im S2 Paket versendet. Diese Variante der IHC Signaturen hat drei vorteilhafte Eigenschaften im Ende-zu-Mitte Fall:

Bandbreiteneffizienz: Das S1 Paket wird sehr klein, da die HMAC Signatur nur ca. 20 Byte (SHA-1) in Anspruch nimmt. Dies bedeutet, dass nur ein kleines Paket ohne Überprüfung an den Verifizierer ausgeliefert wird. Das große S2 Paket mit der Nachricht N kann vor der Weiterleitung durch die Middleboxen verifiziert werden.

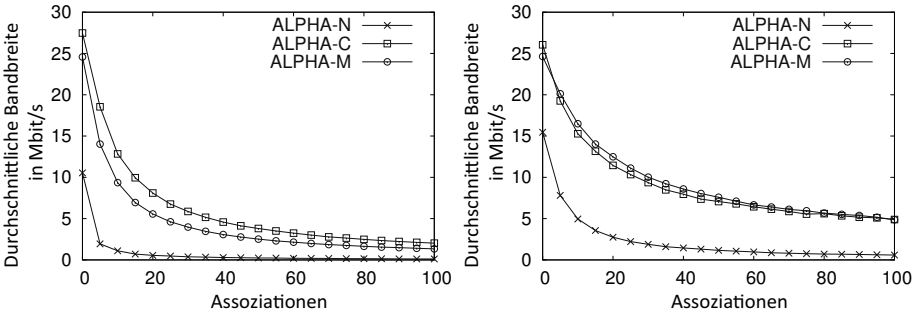
Speichereffizienz: Middleboxen müssen nur die kleinen Commitments anstatt der gesamten Nachricht N zwischenspeichern, was die Speicheranforderungen drastisch senkt und die Angreifbarkeit dieser Geräte deutlich verringert.

Senderausschluss: Ohne eine Bestätigung des Verifizierers durch ein V1 Paket kann der Sender keine großen S2 Pakete senden. Der Verifizierer kann so den Signierer daran hindern Datenpakete zu senden, wobei die Middleboxen dies, durch Weiterleitung oder Verwerfen des S2 Pakets, implizit umsetzen.

Aufgrund dieser Eigenschaften verwenden wir für das "Adaptive and Lightweight Protocol for Hop-by-Hop Authentication", ALPHA, diese zweite Variante der IHC Signaturen [HGGMW08]. Wir implementierten diese Variante und evaluierten sie für die oben genannten Router. Dabei konnte der Router bereits Durchsätze von 10 Mbit/s verifizieren, was den theoretischen Wert der genannten asymmetrischen Signaturen bereits um das 15-fache übersteigt. Jedoch zeigt sich bei genauerer Betrachtung des Routers, dass weder das Netzwerk, noch die CPU des Geräts der limitierende Faktor ist. Die offensichtliche Wurzel des Problems liegt dabei auf der Hand: Zwar senkt ALPHA den *Berechnungsaufwand* der Signaturen, jedoch verdreifacht es durch die IHC Signatur den *Kommunikationsaufwand*. Zur weiteren Erhöhung des möglichen Durchsatzes ist es daher notwendig, den Kommunikationsaufwand wieder zu senken.

3.1 Amortisierung des Kommunikationsaufwands

Die Tatsache, dass die Commitments (die HMAC Signaturen) in den S1 Paketen sehr klein sind, ermöglicht eine erste Optimierung des Verfahrens. Anstatt ein einziges Commitment für eine einzige Nachricht (ein S2 Datenpaket) im ersten S1 Paket zu senden, kann der Si-



(a) Eine Assoziation (Bidirektional)

(b) Fünf Assozationen (Bidirektional)

Abbildung 3: ALPHA Durchsatz bei 30 ms Netzwerklatenz.

gnierer mehrere Commitments für mehrere Nachrichten N_i gesammelt (kumuliert) versenden. Dieser Betriebsmodus von ALPHA nennt sich daher ALPHA-C (Cumulative Mode). Nach der Bestätigung durch den Verifizierer kann der Signierer dann alle Nachrichten N_i in S2 Paketen parallel versenden. ALPHA-C ermöglicht es, die Kommunikationskosten für den S1/V1 Austausch auf viele S2 Pakete zu verteilen. Um dies zu erreichen, müssen beim Senden des S1 Pakets bereits alle Signaturen der Datenpakete vorliegen. Da Transportprotokolle wie TCP stets mehrere Datenpakete gleichzeitig senden, um einen hohen Durchsatz zu erreichen, stellt dies in der Praxis keine Einschränkung dar. Die Anzahl der S2 Pakete pro S1 Paket ist bei dieser Optimierung jedoch durch zwei Faktoren begrenzt. Zum Einen müssen alle Signaturen Platz im S1 Paket finden. Je nach Protokollentwurf begrenzt dies die Anzahl der S2 Pakete in der Praxis auf ca. 70. Zweitens müssen Middleboxen alle Signaturen im S1 Paket bis zum Erhalt der S2 Pakete zwischenspeichern. Sollten die Middleboxen stark platzbeschränkt sein, kann dies eine Einschränkung darstellen und die Maximalzahl der S2 Pakete weiter senken.

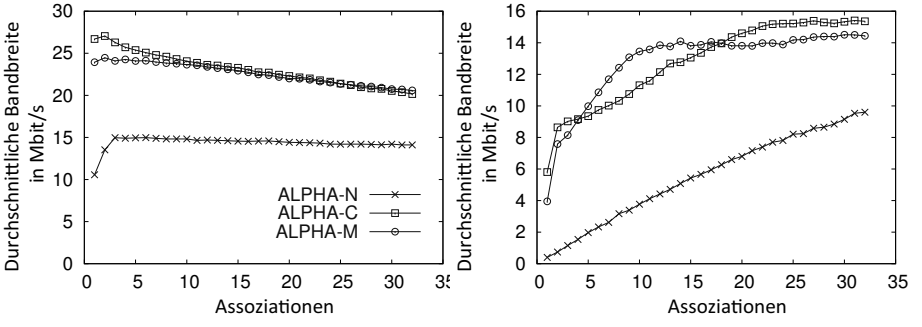
Um diese Einschränkungen zu umgehen bietet ALPHA einen weiteren Modus: Binäre Hash-Bäume, so genannte Merkle-Bäume [Mer90], erlauben es, eine große Zahl von Eingabewerten sicher auf einen einzigen Ausgabewert (die Wurzel) mit fester Größe abzubilden, sodass der Ausgabewert von allen Eingabewerten abhängt. Durch die Verwendung eines Merkle-Baums ist es möglich, eine beliebige Zahl von Nachrichten N_i durch einen kleinen Wert r zu repräsentieren. Der Sender kann somit r im S1 Paket authentisiert übermitteln und später nach Erhalt der Bestätigung alle N_i parallel versenden. Um ein S2 Paket zu verifizieren muss eine Middlebox nachvollziehen, ob die Nachricht N_i Teil der Eingabemenge des Baums war. Hierzu benötigt die Middlebox a) die Nachricht N_i, b) die Wurzel r und c) die Nachbarknoten des Pfads durch den Baum von N_i zu r. Da Merkle-Bäume balancierte Binärbäume sind, handelt es sich dabei um $log_2(n)$ Nachbarknoten für n Nachrichten. Dieser Betriebsmodus nennt sich aufgrund der Verwendung von Merkle-Bäumen ALPHA-M. ALPHA-M stellt konstante Speicherplatzanforderungen an Middleboxen zur vorübergehenden Speicherung von r. Der Platzbedarf in jedem S2 Paket wächst jedoch mit der Anzahl der parallel verschickten Nachrichten logarithmisch. Im Gegensatz dazu wächst bei ALPHA-C der Speicheraufwand der Middleboxen linear zur Anzahl der parallel versendeten S2 Paketen, wobei der Verifikationsaufwand konstant ist.

Durch Verwendung von ALPHA-C und ALPHA-M lässt sich der Durchsatz von ALPHA bereits auf 26 Mbit/s steigern, jedoch fällt bei genauerer Betrachtung auf, dass für größere Netzwerklatenzen der Durchsatz überproportional stark abfällt. Abbildung 3a stellt diesen Abfall für alle ALPHA Modi dar. ALPHA-N repräsentiert dabei den einfachen ALPHA Modus ohne parallele Übermittlung mehrerer S2 Pakete. Der Grund für diesen steilen Abfall ist, dass während des Austauschs der S1 und V1 Pakete keine Datenpakete gesendet werden können. Daher entstehen bei größerer Latenz lange Zeiten der Inaktivität. Um dieses Problem zu beheben, können mehrere unabhängige Signaturprozesse, also unabhängige ALPHA Assoziationen, ineinander verzahnt werden. Ein oder mehrere Assoziationen können so weiter Daten übertragen, während andere Assoziationen auf eine Bestätigung des S1 Pakets durch den Verifizierer warten. Abbildung 3a zeigt die Durchsatzwerte für fünf parallele Assoziationen. Insgesamt lässt sich der scharfe Abfall der Bandbreite durch die Verwendung mehrerer Assoziationen deutlich mindern. Abbildung 4 zeigt den kombinierten Effekt der vorgestellten Techniken. Abhängig vom Modus, von der Verbindungsverzögerung und dem Grad der Parallelität erreicht ALPHA dabei Werte von bis zu 25 Mbit/s auf dem drahtlosen Router und liegt damit bereits jenseits der Leistungsfähigkeit vieler drahtloser 802.11 Multi-Hop Verbindungen [WHBW11].

4 SPOTS: Token-basierte Quellauthentifizierung

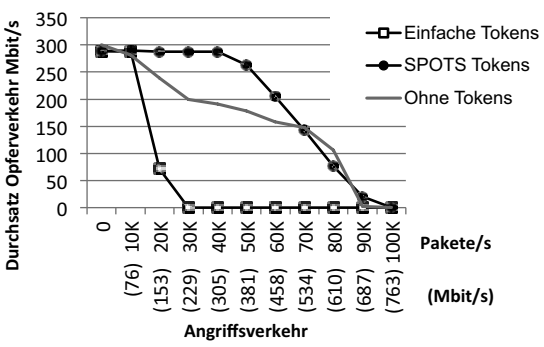
In einigen Anwendungsfällen ist für Middleboxen nur die Herkunft eines Pakets von Bedeutung. Zum Beispiel kann das Fluten des Netzwerks von Middleboxen bereits durch eine sichere Bestimmung der Quelle verhindert werden und bedingt keinen Integritätsschutz des Paketinhalts. ALPHA leistet daher durch seine Quell- *und* Integritätsprüfung für diese Szenarien zuviel und macht durch das IHC Verfahren die Kommunikation unnötig aufwändig. Eine einfachere Art der Quellauthentifizierung ohne eine Überprüfung der Paketinhalte ist daher notwendig, um weniger anspruchsvolle Szenarien effizient zu bedienen.

Eine auf Tokens basierende Quellauthentifizierung eignet sich für die einfache und sichere Bestimmung der Herkunft eines Pakets. Dabei wird jedem Paket ein *einmal* gültiges Token

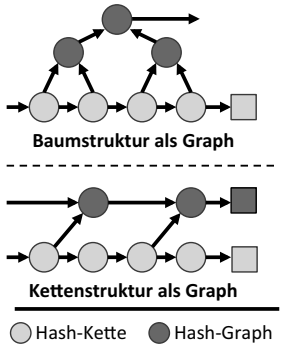


(a) 0,5 ms Netzwerklatenz (b) 20 ms Netzwerklatenz

Abbildung 4: Durchsatz der verschiedenen ALPHA-Modi.



(a) Angriff auf Tokens



(b) SPOTS Tokens verwenden Hash-Graphen

Abbildung 5: Die einfachen Token-Schemata stellen ein zusätzliches Risiko während DoS Angriffen dar. Die SPOTS Tokens ermöglichen effizienten Schutz gegen DoS Angriffe durch die Verwendung von Hash-Graphen. (PC Middlebox, 2.5GHz AMD Athlon 4800+).

hinzugefügt, welches der Empfänger effizient verifizieren kann. Hash-Ketten können als Basis für solche Tokens dienen, indem jedes Element der Kette als Token interpretiert wird. Die sukzessive Preisgabe der Hash-Ketten Elemente von h_n hin zu h_0 beschreibt dabei einen Strom von Tokens, in dem jedes Token durch Hashen einem zuvor empfangenen Token zugeordnet werden kann: $H(h_i) = h_{i+1}$. Da Hash-Ketten hocheffizient erzeugt und verifiziert werden können, verspricht dieser Ansatz einen hohen Durchsatz.

Bestehende Hash-Ketten-basierte Systeme [TO03, DHM05] sind jedoch entweder nur für geringe Bandbreiten geeignet oder lassen Netzwerkeffekte, wie Paketverlust und mögliche DoS Angriffe unberücksichtigt. Daher eignen sie sich nicht für den Schutz von breitbandigen Datenströmen in realen Szenarien. Das Grundproblem dieser Ansätze besteht dabei in der Annahme, dass die Tokens fast vollständig bei den verifizierenden Stellen (End-Systeme und Middleboxen) eintreffen. Bei nur geringem Verlust von wenigen Tokens kann dann durch mehrmalige Anwendung der Hash-Funktion die Lücke zwischen einem zuvor verifizierten Token h_i und einem neu eingetroffenen Token h_{i-n} geschlossen werden: $H^n(h_{i-n}) = h_i$. Jedoch steigt der Verifikationsaufwand eines Tokens dabei linear zur Anzahl der zuvor verlorenen Tokens, sodass die Verifikation nach längeren sequentiellen Verlusten deutlich aufwändiger wird. Dieses Verhalten macht den Einsatz solcher einfacher Token-Schemas unsicher, da ein Angreifer durch Fälschen von Paketzählern sehr einfach ungültige Tokens erzeugen kann, die einen hohen vorangegangenen Verlust andeuten. Um diese Tokens zu falsifizieren muss die Middlebox zahlreiche Iterationen der Hash-Funktion ausführen. Dies bedeutet, dass jedes Angriffspaket nicht nur Bandbreite, sondern in hohem Maße auch Rechenkapazität der Middlebox vergeudet. So stellt der scheinbare Schutz der Tokens in der Praxis eine Bedrohung dar, da ein Angreifer durch Ausnutzung der Tokens CPU-beschränkte Geräte deutlich einfacher überlasten kann. Abbildung 5a zeigt die Auswirkungen eines solchen Angriffs. Der durch einfache Tokens "geschützte" Opferstrom wird anfällig gegenüber DoS Angriffen. Dadurch ist die Leistungsfähigkeit der Middlebox bereits bei deutlich geringeren Angriffsraten erreicht und überschritten, sodass der Opferverkehr stark abnimmt.

4.1 Hash-Graphen für robuste Token-Verifikation

Um Hash-basierte Tokens in der Praxis einsetzbar zu machen ist es notwendig, Möglichkeiten zur effizienten Verifikation von Tokens nach sequentiellem Verlust voriger Pakete zu schaffen. Die "Stream-based Per-packet One-time Tokens for Cryptographic Source Authentication" (SPOTS) Schemata sind eine Alternative zur Verwendung von linearen Hash-Ketten. SPOTS verwendet dabei Hash-Graphen anstelle von Hash-Ketten. Ähnlich zu einem Merkle-Baum können beliebige azyklische gerichtete Graphen aus Verkettungen von Hash-Funktionen gestaltet werden. Jeder Knoten im Graph ist dabei das Ergebnis der Anwendung einer Hash-Funktion, während jede Kante die Anwendung einer Hash-Funktion darstellt. Die Herausforderung bei der Gestaltung von geeigneten Hash-Graphen besteht darin, im Normal- und im Maximalfall möglichst wenige Kanten für die Verbindung eines neuen Wertes (eines frischen Tokens) hin zu einem bereits verifizierten Wert zu durchschreiten. Gleichzeitig muss die Menge von Zusatzinformationen, die zur Durchschreitung der Kanten benötigt wird, gering gehalten werden. SPOTS verwendet Ketten- und Baumartige Graphen, um eine schnelle Verifikation zu ermöglichen. Wie in Abbildung 5b gezeigt, wird dazu eine reguläre Hash-Kette um einen Hash-Graphen erweitert. Falls kein Paketverlust angenommen wird (d.h. es liegt auch kein Angriff vor), kann entlang der originären Hash-Kette effizient verifiziert werden. Falls ein Paketverlust angezeigt wird, kann der Hash-Graph dazu verwendet werden, um ein Element der Hash-Kette mit wenigen Hash-Berechnungen zu verifizieren oder zu falsifizieren. Wie Abbildung 5a zeigt, lässt sich dadurch die Schwäche der einfachen Tokens nicht nur ausgleichen, sondern es kann zusätzlich ein effizienter Schutz vor DoS Angriffen erreicht werden. Gleichzeitig erzielen die mit SPOTS geschützten Datenströme in unserer Evaluation einen hohen Durchsatz von bis zu 69 Mbit/s für die oben genannten drahtlosen Router und bis zu 288 Mbit/s für eine Middlebox mit Athlon 4800+ Prozessor. Neuere Messungen mit aktueller PC Hardware zeigen sogar einen Durchsatz von bis zu 748 Mbit/s auf einem Intel i7-870 System. Dabei wird die Quelle jedes einzelnen Nutzlastpakets kryptographisch verifiziert, sodass gefälschte Pakete bereits früh im Netzwerk verworfen werden können.

5 Zusammenfassung

Kooperative Netze erlauben einen flexiblen Einsatz in Szenarien, in denen klassische geplante Netzwerkarchitekturen an technische oder wirtschaftliche Grenzen stoßen. Jedoch birgt der Aspekt der Kooperation neue Risiken, denen mit adäquaten Mitteln begegnet werden muss. Authentifizierung von Benutzern und Datenströmen auf Netzwerkebene kann dabei zu einer deutlichen Erhöhung der Widerstandsfähigkeit des Netzes gegen Angriffe führen, jedoch muss diese effizient erreichbar sein, um nicht die Leistungsfähigkeit des Netzes zu begrenzen und neue Schwachstellen zu schaffen. Dieser Artikel stellt zwei komplementäre Authentifizierungssysteme vor, die es Netzwerkkomponenten erlauben, sehr effizient die Authentizität von Paketen zu überprüfen. Die Systeme können als komplementär betrachtet werden, da sie maßgeschneiderte Lösungen für Szenarien bereithalten, die Quellauthentifizierung und/oder Integritätsschutz benötigen.

Literatur

[DHM05] Jing Deng, Richard Han und Shivakant Mishra. Defending Against Path-based DoS Attacks in Wireless Sensor Networks. In *Proceedings of the 3rd ACM workshop on Security of ad hoc and sensor networks, SASN*, Seiten 89–96. ACM Press, 2005.

[HGGMW08] T. Heer, S. Götz, O. Garcia Morchon und K. Wehrle. ALPHA: An Adaptive and Lightweight Protocol for Hop-by-hop Authentication. In *ACM CoNEXT: Proceedings of the 2008 ACM CoNEXT Conference, Madrid, Spain*. ACM, 2008.

[HHK+09] T. Heer, R. Hummen, M. Komu, S. Götz und K. Wehrle. End-host Authentication and Authorization for Middleboxes based on a Cryptographic Namespace. In *Proceedings of the IEEE International Conference on Communications 2009 (ICC 2009), Dresden, Germany*. IEEE, 2009.

[Lam81] L. Lamport. Password Authentication with Insecure Communication. *Communications of the ACM*, 24(11):770–772, November 1981.

[Mer90] RC. Merkle. A Certified Digital Signature. In *CRYPTO '89: Proceedings of the 9th Annual International Cryptology Conference on Advances in Cryptology*, Seiten 218–238, London, UK, 1990. Springer.

[TO03] H. Tewari und D. O'Mahony. Multiparty Micropayments for Ad Hoc Networks. In *Wireless Communications and Networking, WCNC 2003*, Jgg. 3, Seiten 2033–2040. IEEE, 2003.

[WHBW11] H. Wirtz, T. Heer, R. Backhaus und K. Wehrle. Establishing Mobile Ad-Hoc Networks in 802.11 Infrastructure Mode. In *ACM MobiCom 2011 Workshop on Challenged Networks (CHANTS'11)*, September 2011.

Tobias Heer studierte, nach einem zweijährigen Exkurs in die Erziehungswissenschaften, Informatik an der Universität Tübingen. Das Studium schloss er 2006 mit Auszeichnung ab. Seine Diplomarbeit verfasste er während eines (kalten, dunklen) Gastaufenthalts am Helsinki Institute for Information Technology in Finnland. Dabei befasste er sich mit leichtgewichtigen Sicherheitslösungen für mobile und ressourcenbeschränkte Geräte. Die Arbeit wurde 2008 mit dem KuVS Preis für die beste Diplomarbeit im Bereich Kommunikation und Verteilte Systeme ausgezeichnet. Er kehrte seither mehrmals, im Rahmen von Forschungsaufenthalten, nach Helsinki zurück. Seine Faszination für Sicherheit und Kommunikationsprotokolle konnte Tobias am DFG Graduiertenkolleg *Software für mobile Kommunikationssysteme* an der RWTH Aachen weiter ausleben. Seit 2009 ist er Projektleiter für die Netzwerkaspekte des *Mobile ACcess* Kooperationsprojekts und arbeitet am Entwurf sicherer kooperativer Wi-Fi Netzwerke. Neben seiner wissenschaftlichen Arbeit hat er sich, im Rahmen der Internet Engineering Task Force (IETF), an der Standardisierung von Mobilitäts- und Sicherheitsprotokollen beteiligt. Seine Promotion schloss Tobias im Dezember 2011 mit Auszeichnung ab. Er ist stolzer Vater von zwei Kindern: Jana und Jannik.

Strahlenbasierte Algorithmen für die medizinische Visualisierung vielfacher volumetrischer Datensätze

Bernhard Kainz
Technische Universität Graz
kainz@icg.tugraz.at

Abstract: Das Problem der gleichzeitigen Echtzeitdarstellung mehrerer volumetrischer Datensätze und deren automatischer Anpassung an reale Sachverhalte (zum Beispiel aus der Medizin/Radiologie) war bisher ungelöst. Seit der Entwicklung multimodaler und mehrdimensionaler bildgebender Verfahren in der Medizin und anderen Disziplinen wäre diese aber notwendig. Moderne PC Hardware ermöglicht nun eine effiziente und kostengünstige Umsetzung neuartiger Algorithmen, die dieses Problem für den allgemeinen Fall lösen und bisher nicht mögliche Anwendungen der direkten Volumsdarstellung erlauben. Die Ausnutzung aller Komponenten eines PC Systems, einschließlich der Grafikkarte (GPU) als leistungsfähigen Coprozessor, ist mittlerweile in der angewandten Informatik üblich. Die unterschiedlichen Speicher- und Prozessorarchitekturen machen allerdings Anpassungen bekannter und die Entwicklung vollkommen neuartiger und parallelisierbarer Algorithmen notwendig. Diese Arbeit zeigt, wie GPU-Algorithmen so effizient entwickelt werden können, dass zeitkritische Darstellungsaufgaben sogar mit garantierten Bilderzeugungsraten ermöglicht werden und positionskritische Aufgaben mit höchstmöglicher Genauigkeit durchgeführt werden können.

1 Einleitung

Seit etlichen Jahren kommen in der Humantechnologie die unterschiedlichsten volumetrischen bildgebenden Verfahren zum diagnostischen Einsatz, um pathologische Gegebenheiten so genau wie möglich einzugrenzen. Bisher war es allerdings nicht möglich, alle diese Verfahren in derselben virtuellen Szene darzustellen oder effizient genug vorzuverarbeiten, obwohl dies für viele Anwendungen zwingend nötig wäre. Volumetrische Bildsynthesealgorithmen werden üblicherweise direkt auf der Grafikkarte (GPU) eines PC-Systems ausgeführt und zeigen oft schwerwiegende Limitierungen, wenn mehrere Volumina gleichzeitig zu verarbeiten sind. Diese Daten müssen innerhalb einer Szene mit höchstmöglicher Genauigkeit und maximaler Interaktivität dargestellt werden. Da dies aber mit aktuellen Algorithmen nicht möglich ist, wird häufig auf eine rein zweidimensionale und sequentielle Untersuchung der vorhandenen Volumsdaten zurückgegriffen. Diese Art der Datenverknüpfung erfordert eine enorme geistige Anstrengung und einen dementsprechenden Zeitaufwand des behandelnden Arztes. Die zweidimensionale Methode ist zwar auf Detaillevel und für Messaufgaben sehr genau, es zeigt sich aber, dass mangels eines dreidimensionalen Überblicks die Gefahr besteht, kritische Gegebenheiten auf höherer Ebene zu übersehen und vor allem, dass wenn moderne multimodale Bildgebunsverfahren

zum Einsatz kommen, die mehr als nur einen Datensatz erzeugen, eine manuelle zweidimensionale Auswertung oft unmöglich wird.

2 Grundlegende Literatur

Seit in den 1980er Jahren die ersten direkten 3D Bildsynthesealgorithmen für medizinische volumetrische Datensätze eingeführt worden sind haben Forscher nützliche Anwendungsgebiete in der klinischen Praxis gefunden und Mediziner haben exzessiv versucht diese in ihren Arbeitsablauf einzubauen. Es ist daher naheliegend, dass zu diesem weit gesteckten Forschungsbereich auch eine Menge an klinischer Literatur existiert. Es haben sich allerdings nur sehr wenige Wissenschaftler, wie in der vorliegenden Dissertation, mit einer retrospektiven Analyse der vorhandenen direkten Volumsdarstellungsmethoden beschäftigt und deren Sinnhaftigkeit evaluiert. McCormick und DeFanti [McC88] haben in den späten 1980er Jahren bereits für die frühen Bildsynthese- und Visualisierungsmethoden zum damaligen Zeitpunkt ungeahnte Anwendungsmöglichkeiten vorhergesagt. Nichts desto trotz sind Fragestellungen wie sie McCormick und DeFanti stellten (vergleiche: *"a smooth integration of useful 3D images for diagnostics"* aus [McC88]) nur teilweise gelöst. Einer der wichtigsten Gründe dafür ist, dass die Auflösung und damit die Größe der generierten Daten in der Medizin mit der gleichen Geschwindigkeit wie die Leistungsfähigkeit moderner Computer wächst [UH91, NR02].

Diese Dissertation beschäftigt sich in erster Linie mit einem wichtigen Teil der medizinischen Datenvisualisierung: direkte dreidimensionale Bildsynthese und der Lösung spezieller inhärenter Probleme dieser Methoden. Neben früheren Verbesserungen der direkten Volumsdarstellung (zum Beispiel sogenannte "early-ray termination" – [Lev90, MIH05] oder "binary space partitioning" – [Lev90, LMK03, SHO$^+$08] Methoden) wurde in letzter Zeit vermehrt die Parallelisierbarkeit von strahlenbasierten Bildsynthesealgorithmen, die eine Unterklasse der direkten Volumsdarstellung definieren, ausgenutzt. Die Nutzung moderner Grafikkarten, die für den Algorithmenentwickler nichts anderes als hoch effiziente programmierbare Parallelprozessoren darstellen, hat bisher am meisten zur Nutzbarkeit der direkten Volumsdarstellung beigetragen [Wei06, AMHH08].

3 Hypothesen

Motiviert durch die immer noch übliche Praxis, medizinische Volumsdaten bevorzugt als Serie von Schichtbildern für die Diagnostik nacheinander zu betrachten, untersucht der motivierende Teil der vorliegenden Arbeit mittels einer Umfrage unter mehr als zwei Dutzend Medizinern, ob dreidimensionale Bildsynthese in der Klinik überhaupt erforderlich ist. Dafür wurden vier praxisorientierte Hypothesen evaluiert, die der Autor nach langjähriger Mitarbeit in der klinischen Praxis formulieren konnte. Im Wesentlichen stimmen die befragten Mediziner den Hypothesen zu, in denen erstens festgestellt wird, dass 3D Bildsynthese immer den diagnostischen Entscheidungsprozess beschleunigen muss,

um überhaupt von klinischem Personal für die Verwendung in Betracht gezogen zu werden, zweitens, dass wenn ein neues Visualisierungsverfahren einem bekanntem physikalischem Prinzip ähnelt, es schneller Akzeptanz findet, drittens, dass 3D Bildsynthese unerlässlich wird, wenn die Dimensionalität der zu untersuchenden Daten die normale menschliche Erfahrung übersteigt (z.B. Diffusion Tensor Imaging) und viertens, dass aktuelle 3D-Verfahren weder schnell genug, noch genau genug und nicht flexibel genug sind, um die übliche schichtbasierte Datensatzevaluierung zu ersetzen.

Ausgehend von diesen vier praxisorientierten Hypothesen konnten drei weitere technisch motivierte Hypothesen abgeleitet werden, die im Laufe der restlichen Arbeit näher untersucht werden. Zusammengefasst wird dabei festgestellt, dass die Klasse der volumetrischen strahlenbasierten Algorithmen so erweitert werden kann, dass deren Eingaberaum nicht mehr auf Eins beschränkt ist, dass durch die Ausnutzung der Eigenschaften des menschlichen visuellen Systems die Bilderzeugungsrate erhöht und überdies auch garantiert werden kann und dass diese Klasse von Algorithmen auch für hochgenaue Positionsbestimmungsaufgaben geeignet ist.

4 Methoden

Strahlenbasierte Algorithmen basieren auf dem Prinzip, dass virtuelle Strahlen in einer dreidimensionalen Szene verfolgt werden. Sollen mehr als nur ein einziges volumetrisches Objekt in derselben Szene dargestellt werden, muss für jedes dieser Objekte der Strahlenverlauf separat bestimmt werden. Wenn sich die darzustellenden Objekte auch noch überschneiden, wird aus der Bestimmung des Strahlenverlaufs eine aufwendige Sortieraufgabe aller auftretenden Strahlensegmente. Diese Sortierung ist im Allgemeinen zu langsam für Systeme mit Echtzeitanforderungen. Um dieses Problem zu lösen muss der übliche Ansatz, die Szene für ein Bild mehrmals zu rendern überdacht werden. Durch die Ausnutzung moderner GPU-Architekturen ist es möglich die zeitaufwendige Strahlenverlaufsbestimmung in einen einzigen Schritt und in einen GPU-Speicherbereich mit einem Durchsatz von mehreren Terabyte pro Sekunde zu verlagern. Dadurch ist es dem Autor nicht nur gelungen diese Klasse von Algorithmen so effizient zu gestalten, dass die Darstellung mehrerer beliebig angeordneter Volumina in Echtzeit möglich wurde, sondern auch, wie in Abbildung 4 gezeigt, die interaktive Manipulation der verwendeten volumetrischen Raumgeometrien.

Da für manche Applikationen hohe Bilderzeugungsraten alleine nicht ausreichend sein können, wird die Strahlverfolgung im Folgenden dahin gehend optimiert, dass nicht nur hohe Bilderzeugungsraten, sondern auch Bilderzeugungsraten mit einer garantierten Frequenz ermöglicht werden. Solche garantierten Bilderzeugungsraten werden bei interoperativen Applikationen benötigt, bei denen unter keinen Umständen ein Bildruckeln das kontinuierliche Arbeiten des durchführenden Arztes behindern darf. Im Gegensatz zu dem üblichen Verfahren, die Bilderzeugungsrate während der Szeneninteraktion zu Lasten der Qualität zu erhöhen, erhält der in der Arbeit vorgestellte Ansatz die wichtigen Details einer Szene und verhindert so eventuell lebensbedrohliche Fehlentscheidungen. Um Bilderzeugungsraten mit einer garantieren Frequenz zu ermöglichen, sollte idealerweise die

Abbildung 1: Konstruktive Raumgeometrieoperationen auf mehreren volumetrischen Datensätzen.

Wichtigkeit eines jeden einzelnen Bildpunktes vor der Strahlenverlaufsbestimmung bekannt sein. Da dies ein Henne-Ei-Problem beschreibt (das Berechnungsergebnis müsste dabei für jeden einzelnen Bildpunkt vor der eigentlichen Berechnung bekannt sein) hat der Autor zusammen mit seinen Kollegen eine Methode entwickelt um die erforderlichen Parameter für jedes einzelnes Bild vorab abschätzen zu können. Dabei werden zusätzliche Algorithmen eingeführt, die normalerweise dafür eingesetzt werden, ein Objekt nur zu skizzieren. Diese Algorithmen leiten sich von den Techniken ab, die Künstler verwenden um Skizzen zu erstellen, und sind im Allgemeinen sehr gut im Hinblick auf die menschliche Wahrnehmung untersucht. Außerdem beinhalten diese Methoden keine zeitaufwendigen Beleuchtungsberechnungen und können daher mit minimalem zeitlichen Mehraufwand in strahlenbasierte Algorithmen integriert werden. Ist die Skizze eines Objekts bekannt, kann diese auf die Bildebene projiziert werden und somit die Wichtigkeit jedes einzelnen Bildpunktes vorab abgeschätzt werden. Die bilderzeugenden Strahlen werden schlussendlich ihrer Wichtigkeit nach sortiert abgearbeitet und nicht berechnete Bildpunkte werden mit einer neu entwickelten Interpolationsmethode aufgefüllt. Wenn bei dieser Methode ein einzuhaltendes Zeitfenster für die Berechnung vorgegeben wird, kann davon ausgegangen werden, dass wenn die Bilderzeugung nach Ablauf der vorgegebenen Zeit gestoppt wird, die wichtigsten Bildpunkte bereits vorhanden sind und somit die maximal in dieser Zeit erzielbare Bildqualität erreicht wird. Abbildung 4 zeigt eine Illustration des beschriebenen Überganges von Skizzendarstellung zum rekonstruierten Ergebnisbild.

In einem weiteren Teil dieser Arbeit wird gezeigt, dass strahlenbasierte Algorithmen nicht nur bezüglich der Strahlenverlaufsbestimmung verbessert werden können, sondern dass auch deren Kern für bestimmte Aufgaben spezialisiert und damit optimiert werden kann. Viele medizinische Problemstellungen enthalten ein Positionsbestimmungsproblem. Dabei müssen zuvor bestimmte Stellen im menschlichen Körper so genau wie möglich in der aktuellen Lage des Patienten wiedergefunden werden. Üblicherweise werden dafür

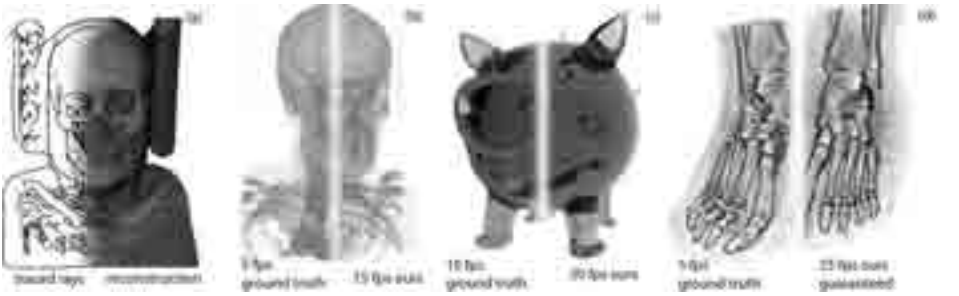


Abbildung 2: Illustration des Überganges von skizziertem Objekt zu rekonstruiertem Ergebnis (a). (b) zeigt wie die vorgestellten Methoden dazu verwendet werden kann die Bilderzeugungsrate zu erhöhen, indem unwichtige Teile interpoliert statt berechnet werden. Die gezeigten Algorithmen können sowohl für volumetrische Datensätze (b) wie auch geometrische photorealistische Darstellungen (c) benutzt werden. Eine zusätzliche Prioritätssortierung der wichtigen Bildbereiche kann sogar dazu genutzt werden um eine minimal erforderliche Bilderzeugungsrate bei bestmöglichem visuell wahrnehmbarem Ergebnis zu garantieren (d).

sogenannte Registrierungsverfahren eingesetzt, die die Lage und Verzerrung eines Bildes solange optimieren, bis der Unterschied zu einem anderen Bild minimal wird. Die sich daraus ergebenden Parameter können dann für die Berechnung der neuen Position eines Zielgebietes verwendet werden. Genaue Registrierungsverfahren sind allerdings häufig zu langsam für den Einsatz während eines realen Eingriffes und schnellere Verfahren sind oft zu ungenau für einen sicheren Einsatz. Die vorliegende Arbeit zeigt auch für dieses Problem einen möglichen Lösungsansatz am Beispiel eines speziellen 2D/3D Registrierungsproblems, so wie es häufig in der Praxis vorkommt. Registrierungsverfahren optimieren die Ähnlichkeit zwischen zwei Bildern häufig mittels Gradientenabstiegs. Dafür muss jeder neue Schritt zum vorherigen Ergebnis verglichen werden und die Differenz dazu in einem höherdimensionalen Raum gebildet werden. Betrachtet man dieses Problem mathematisch kann man feststellen, dass in der Gesamtberechnung des Gradienten die erste Ableitung eines spezialisierten strahlenbasierten Verfahrens ein integraler Bestandteil ist. Diese erste Ableitung kann auch mittels sogenannter automatischer Codeableitungsverfahren vor dem eigentlichen Einsatz bestimmt werden. Somit kann der Code eines GPU-basiertem Bilderzeugungsalgorithms direkt als spezialisierte Ableitung ausgeführt werden. Das Ergebnis der vorgeschlagenen Methode ist dann kein Bild mehr, sondern direkt der für die Parameteroptimierung benötigte Gradient der Zielfunktion. Mit diesem Verfahren erreichen wir eine sehr hohe Genauigkeit bei vernachlässigbarem Zeitaufwand und zeigen dessen Anwendungsmöglichkeit an einem konkreten Beispiel aus der Urologie. Mit diesem Verfahren sollte es in Zukunft möglich sein Organresektionen, wie sie zum Beispiel bei Erkrankungen der Prostata üblich sind, in bestimmten Fällen zu vermeiden und somit die Lebensqualität der Patienten zu erhöhen. Abbildung 4 zeigt den schematischen Ablauf eines solchen Verfahrens.

In den meisten Fällen ist allerdings das Wiederauffinden einer bestimmten Position im Körper nicht das alleinige Kriterium für einen erfolgreichen Eingriff. Bei minimal invasiven oder zerebralen Prozeduren ist es auch erforderlich, dass der behandelnde Arzt den

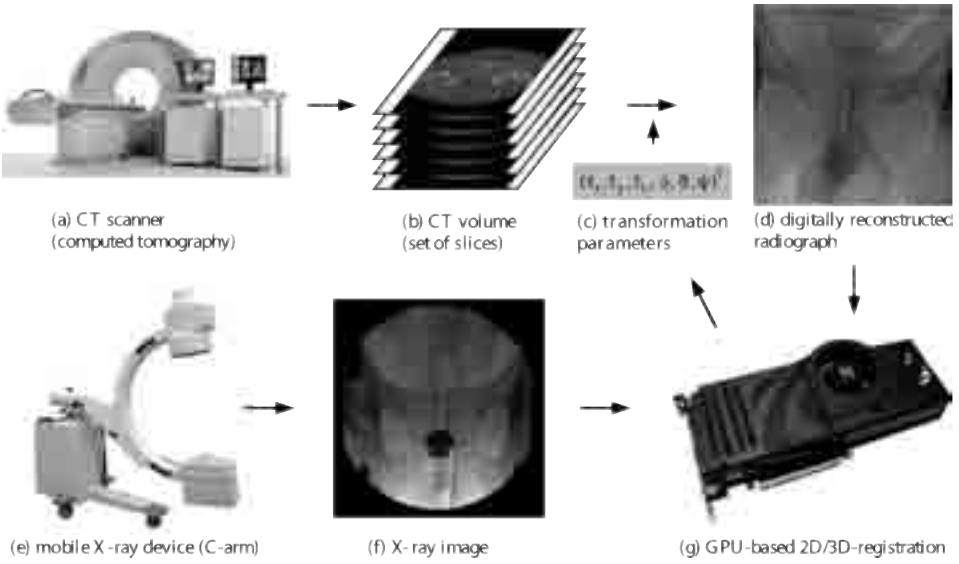


Abbildung 3: Schematischer Ablauf einer Registrierung mit spezialisiertem Bilderzeugungsalgorithmus am Beispiel eines Prostataeingriffs. Vor dem Eingriff wird eine Computertomographie des Patienten erstellt (a) woraus sich ein volumetrischer Datensatz ergibt (b). Mit einem speziellen Algorithmus können daraus die Transformationsparameter (c) für eine künstliche Röntgenbildprojektion (d) berechnet werden. Diese Transformationsparameter können anschließend auf der GPU (g) mit einem realen Röntgenbild (f) verglichen und hinsichtlich ihrer Position optimiert werden. Da während eines urologischem Eingriffes in den meisten Fällen ein C-Bogen (e) standardmäßig für die Erzeugung von Röntgenbildern zum Einsatz kommt, erfordert diese Methode keine neuen medizinischen Geräte um eine vorab bestimmte Position im Körper des Patienten wiederzufinden. Im Laufe dieser Arbeit wurde auch ein Zielobjekt entwickelt, dass die initiale Kalibrierung des C-Bogens erleichtert und beschleunigt.

sichersten Weg zum designierten Einsatzgebiet bestimmt. Im schlimmsten Fall kommt es bei falscher Planung dieser Eingriffe zu unbeabsichtigten Verletzungen von empfindlichen Strukturen, wie zum Beispiel Arterien oder Nervenfasern, und damit leider auch manchmal zu schweren Komplikationen oder zum Tod des Patienten. Um die Eingriffsplanung zu erleichtern hat der Autor dieser Arbeit mit seinen Kollegen eine volumetrische Berechnungsmethode entwickelt, um alle möglichen Zugangswege zu einem Bestimmten Zielgebiet und deren Sicherheit, inklusive dem möglichen Handlungsspielraum, zu bestimmen. Dabei kommen bildgebende Verfahren zum Einsatz, wie sie auch bei konventioneller Planung üblicherweise verwendet werden. Der Unterschied ist allerdings, dass mit der vorgestellten Methode nicht mehr jeder Datensatz einzeln evaluiert werden muss, sondern, dass das planungsrelevante Ergebnis aller nötigen Modalitäten in einer gemeinsamen Visualisierung dargestellt werden kann. Damit sinkt die kognitive Belastung des ausführenden Arztes erheblich und die zuvor oftmals nur geschätzten Handlungsspielräume können genau bestimmt werden. Im Unterschied zu bisher üblichen Verfahren, die nur einen Punkt als Ziel verwenden, wird bei unserer Methode das gesamte Volumen der Zielregion für die

Berechnung berücksichtigt und somit Pfade gezeigt, durch die das gesamte zu behandelnde Gebiet zu erreichen ist, ohne dass gefährdete Strukturen passiert werden. Dafür wird der zu behandelnde Bereich als volumetrische Lichtquelle in streuendem Medium (zum Beispiel Staub oder Nebel) gedacht. In der Natur bilden sich in so einer Situation sichtbare Lichtstrahlen. Diese Lichtstrahlen können zusammen mit den zu untersuchenden Volumsdatensätze durch die zuvor vorgestellten Methoden für die gleichzeitige Darstellung mehrerer volumetrischer Datensätze direkt im dreidimensionalen Raum und zusätzlich auf beliebigen Schnittebenen ohne Mehraufwand und in Echtzeit für die Planung eines Eingriffes angezeigt werden. Abbildung 4 zeigt die Anwendung dieser Methode für ein bereits jetzt im Einsatz befindliches Anwendungsszenario für die minimal invasive Behandlung von Leberkrebs. Abbildung 4 zeigt ein weiteres mögliches Szenario für nicht-medizinische Anwendungen.

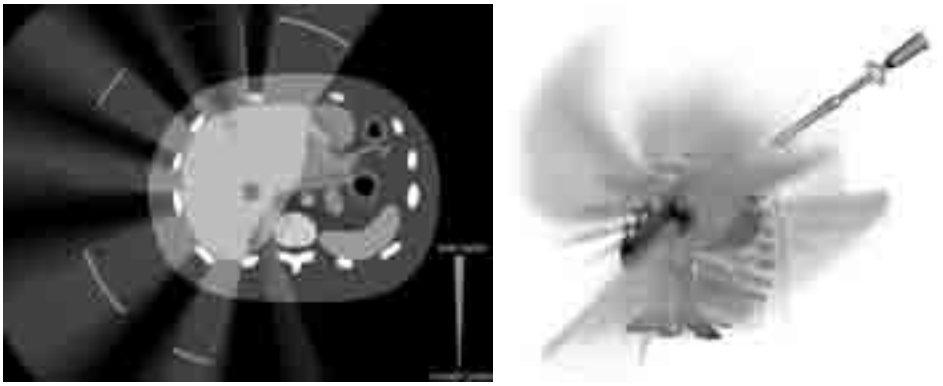


Abbildung 4: Ein Anwendungsbeispiel für die entwickelten Visualisierungs- und Darstellungsalgorithmen. Eine zu behandelnde pathologische Veränderung wird als helle Lichtquelle betrachtet und das umgebende Medium als streuend berechnet. Alle möglichen Zugangswege werden so als Volumendatensatz berechnet und zeigen die Sicherheit der Wege zu der zu behandelnden Strukturen auf einen Blick durch Abschattung und Farbgebung. Die hybride Visualisierung in 2D sowie in 3D erlaubt es außerdem auch den Weg zu finden, der für den Patienten am sichersten ist und gleichzeitig den größten Handlungsspielraum bietet.

5 Ergebnisse

Diese Dissertation beschreibt und evaluiert verschiedene vom Autor entwickelten Methoden um direkte dreidimensionale Bildsynthese in der klinische Praxis Nutzbar zu machen. Dabei konnten die getroffenen Annahmen durch folgende Forschungsergebnisse bestätigt werden.

Durch eine Vorabevaluierung mit mehr als zwei Dutzend Medizinern, die auf mehr als fünf Jahre Erfahrung in ihrem Fachgebiet zurückblicken können, konnte die Nützlichkeit und vor allem die zukünftige Unentbehrlichkeit von direkter dreidimensionaler Bildsynthese

Abbildung 5: Ein volumetrischer Datensatz wie er in Maschinenbaustudien verwendet wird. Gleichzeitig wird ein Volumen dargestellt, das die Zugangsmöglichkeiten zu einem schwer erreichbaren Bereich visualisiert. Beachten Sie die nicht auf der Hand liegende Zugangsmöglichkeit im unteren Bereich.

für viele medizinische Prozeduren gezeigt werden.

Auf technischer Ebene zeigt diese Arbeit als erste wie mehrere volumetrische Datensätze gleichzeitig in derselben Szene in Echtzeit dargestellt werden können. Es wird außerdem gezeigt wie dieser Algorithmus einfach erweitert werden kann um komplexe geometrische Operationen darauf auszuführen (wie zum Beispiel virtuelle Schnitte mit nichttrivialer Geometrie). Dieser Algorithmus ist von sich aus in der Lage, ohne sonst nötige berechnungsintensive Erweiterungen, zusätzlich zu Volumsdaten transparente Objekte ebenfalls in Echtzeit darzustellen. Die Flexibilität der vorgestellen Methoden wird außerdem durch eine Weiterentwicklung der Echtzeitberechnung von physikalisch approximierten Lichtstreuungseffekten gezeigt.

Um die Berechnungsgeschwindigkeit weiter zu erhöhen, wurden in dieser Arbeit geometrische Charakteristiken auf deren Relevanz für die Bildwahrnehmung im menschlichen visuellen System untersucht und nach deren Wichtigkeit gereiht. Die Ergebnisse dieser Evaluierung wurden dazu genutzt um bestimmte Bildteile nicht berechnen zu müssen sondern stattdessen aus benachbarten (wichtigen) Bereichen zu interpolieren. Damit wurde im ersten Schritt eine signifikante Erhöhung der Bilderzeugungsrate erreicht und im zweiten Schritt ein Algorithmus entwickelt, der es ermöglicht eine bestimmte minimale Bilderzeugungsfrequenz bei bestmöglichem visuell wahrnehmbarem Ergebnis zu garantieren.

Die vorgestellten Methoden wurden in konkreten Anwendungen für die klinische Praxis evaluiert und von klinischen Experten für gut befunden. So wurde zum Beispiel die Darstellung mehrerer Volumetrischer Datensätze in derselben Szene dazu genutzt, alle möglichen Zugangspfade minimal invasiver Eingriffe, die mehrere Bildgebungsmodalitäten erfordern, zusammen mit deren Sicherheit für den Patienten genutzt (vergleiche Abbil-

dung 4). Außerdem wurden die entwickelten Echtzeitbilddarstellungsmethoden dazu verwendet, die Position von zuvor als bösartig klassifiziertem Gewebe in einem späteren Eingriff mit sehr hoher Genauigkeit wiederzufinden. Es wurde auch gezeigt, dass es mit den entwickelnden Algorithmen möglich ist, Abhängigkeiten zwischen gleichzeitig dargestellten Datensätzen automatisch zu erkennen und entsprechend zu visualisieren.

6 Fazit

Alle Teile dieser Arbeit wurden vollständig von Experten aus den Bereichen Medizin und Informatik evaluiert. Dabei sind vorläufige Ergebnisse dieser Arbeit in renommierten internationalen Journalen und Konferenzen des jeweiligen Fachgebiets vorgestellt worden. Artikel in “ACM Transactions on Graphics” und “Circulation: Cardiovascular Imaging” sind nur zwei prominente Beispiele für über ein Dutzend von Experten begutachtete, wissenschaftliche Publikationen, die aus dieser Dissertation direkt hervorgegangen sind.

Die praktischen und theoretischen Ergebnisse bescheinigen den vorgestellten Methoden nicht nur deren Gültigkeit, sondern auch deren direkte Anwendbarkeit auf vielfältige Szenarien aus der klinischen Praxis. Durch die enge Zusammenarbeit mit Institutionen wie dem LKH Graz, dem ZMF Graz, der Universitätsklinik Leipzig, der Oxford University, der Fraunhofer-Gesellschaft, der Aalto University Helsinki, SIEMENS HealthCare und SIEMENS Corporate Research ist bereits jetzt die wissenschaftliche Weiterverwendung und wirtschaftliche Auswertung der entstandenen Algorithmen im Gange. In einem Folgeprojekt wird nun verstärkt die menschliche Wahrnehmung und deren Ausnutzung in Kombination mit den zu erwartenden Hardwareentwicklungen für noch effizientere strahlenbasierte Algorithmen erforscht werden.

Überblicksvideos der einzelnen vorgestellten Methoden stehen auf youtube zur Verfügung (Kommentar auf Englisch):

minimalinvasive Zugangspfadplanung:
`http://www.youtube.com/watch?v=mHO6gCm9EP4`

wahrnehmungsbasierte Optimierung:
`http://www.youtube.com/watch?v=HjZYv_FaPSI`

vielfache Volumen, Darstellungsalgorithmus:
`http://www.youtube.com/watch?v=2Ym0CoJ0pVk`

Literatur

[AMHH08] Tomas Akenine-Möller, Eric Haines und Natty Hoffman. *Real-Time Rendering 3rd Edition*. A. K. Peters, Ltd., Natick, MA, USA, 2008.

[Lev90] Marc Levoy. Efficient ray tracing of volume data. *ACM Trans. Graph.*, 9:245–261, July 1990.

[LMK03] Wei Li, Klaus Mueller und Arie Kaufman. Empty Space Skipping and Occlusion Clipping for Texture-based Volume Rendering. In *Proceedings of IEEE Visualization*, Seiten 317–324, 2003.

[McC88] B. H. McCormick. Visualization in scientific computing. *SIGBIO Newsl.*, 10:15–21, March 1988.

[MIH05] Manabu Matsui, Fumihiko Ino und Kenichi Hagihara. Parallel Volume Rendering with Early Ray Termination for Visualizing Large-Scale Datasets. In Jiannong Cao, Laurence Yang, Minyi Guo und Francis Lau, Hrsg., *Parallel and Distributed Processing and Applications*, Jgg. 3358 of *Lecture Notes in Computer Science*, Seiten 245–256. Springer Berlin Heidelberg, 2005.

[NR02] F. Natterer und E.L. Ritman. Past and future directions in x-ray computed tomography (CT). 12(4):175–187, 2002.

[SHO+08] Weiwei Song, Shungang Hua, Zongying Ou, Hu An und Kaifeng Song. Octree Based Representation and Volume Rendering of Three-Dimensional Medical Data Sets. In *Proceedings of the 2008 International Conference on BioMedical Engineering and Informatics - Volume 01*, Seiten 316–320, 2008.

[UH91] Jayaram K. Udupa und Gabor T. Herman, Hrsg. *3D imaging in medicine*. CRC Press, Inc., Boca Raton, FL, USA, 1991.

[Wei06] Daniel Weiskopf. *GPU-Based Interactive Visualization Techniques (Mathematics and Visualization)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Bernhard Kainz wurde am 27. Januar 1982 in Fürstenfeld (Steiermark, Österreich) geboren. Nach Absolvierung der Hochschulreife am BG/BRG Fürstenfeld und dem darauf folgenden Präsenzdienst (2001) absolvierte er das Studium der Telematik in den Fachrichtungen *Biomedical Engineering* und *Computer Graphics and Vision* an der Technischen Universität Graz. Er erhielt die akademischen Grade bakk.techn. (B.Sc.) 2005 und Dipl.-Ing. (DI, M.Sc.) 2007 mit Auszeichnung und begann danach das Doktoratsstudium der technischen Wissenschaften unter Professor Dieter Schmalstieg, das er 2011 ebenfalls mit Auszeichnung abschloss. Seine aktuellen Forschungsinteressen sind volumetrische medizinische Bildgebungsverfahren und deren automatische Verarbeitung und Visualisierung in Echtzeit, Parallelprogrammierung von Bildverarbeitungsalgorithmen und die Verwendung von *Augmented Reality* Methoden in der klinischen Praxis. Neben mehr als zwei Dutzend veröffentlichten wissenschaftlichen Publikationen hat er in den letzten Jahren zur Akquirierung von mehreren hunderttausend Euro an Forschungsmitteln beigetragen und zahlreiche Studentenprojekte betreut. Seine aktuelle Forschung über priorisierte Zeitplanung paralleler Algorithmen wird vom österreichischen Fonds zur Förderung der wissenschaftlichen Forschung unter der Projektnummer 23329 finanziert.

Quantitative Analyse der Spontanmotorik von Säuglingen für die Prognose der infantilen Cerebralparese

Dominik Karch

Institut für Medizinische Biometrie und Informatik
Universität Heidelberg
dominik.karch@googlemail.com

Abstract: Die subjektive, videogestützte Auswertung der kindlichen Spontanmotorik kann von erfahrenen Kinderneurologen als aussagekräftiges Verfahren zur Prognose von schweren Bewegungsstörungen wie der infantilen Cerebralparese (ICP) angewendet werden. Mit der hier vorgestellten Dissertation wurde eine computergestützte Methode entwickelt, die durch eine objektive Analyse die Anwendung dieses Verfahrens auch außerhalb spezialisierter Zentren ermöglichen soll. Die entwickelte Methode zur quantitativen Abbildung von Säuglingsbewegungen, die mit Hilfe eines Trackingsystems dreidimensional erfasst werden, ist durch ein sich automatisch rekalibrierendes biomechanisches Bewegungsmodell robust genug, um in der klinischen Praxis eingesetzt werden zu können. Auf Basis von Bewegungsaufnahmen aus einer Studie (58 Kinder ohne, 7 Kinder mit Outcome ICP) wurden Algorithmen entwickelt, die Eigenschaften anomaler Spontanmotorik (z.B. selbstähnliche Bewegungen) quantifizieren können. Damit konnten in der Stichprobe Säuglinge mit Outcome ICP identifiziert werden. Das Verfahren kann zu Trainingszwecken, zur Entscheidungsunterstützung und zur Erforschung von frühkindlichen Bewegungsstörungen eingesetzt werden.

1 Motivation

In den ersten Lebensmonaten zeigen Säuglinge noch keine zielgerichteten Bewegungen. Die Motorik, die das Kleinkind in diesem Alter äußert, entsteht vornehmlich nicht aufgrund äußerer Reize (ausgenommen Reflexe), sondern endogen im zentralen Nervensystem (ZNS) des Kindes. Aus diesem Grund können Bewegungsmuster Aufschluss über den Zustand des ZNS geben. Die Evaluation der Spontanmotorik kann daher potentiell für die Diagnose von neurologischen Erkrankungen und Anomalien genutzt werden.

Auf diesem Zusammenhang zwischen Spontanmotorik und der Integrität des neuromotorischen Systems fußt ein Prognoseverfahren, das u.a. vom Arzt Heinz F. R. Prechtl entwickelt wurde. Bei dieser Methode wird die Spontanmotorik aufgrund qualitativer Merkmale als normal oder anomal bewertet. In verschiedenen Studien wurde nachgewiesen, dass die Gruppe derjenigen Kinder, die solche anomalen Bewegungen im Alter von drei Monaten zeigen, ein sehr hohes Risiko aufweist, später eine infantile Cerebralparese (ICP) zu entwickeln [PEC+97]. Dabei handelt es sich um schwere Störungen des motorischen Systems, die bei etwa 7 von 100 extrem unreifen Frühgeborenen als Folge einer frühen Schädigung

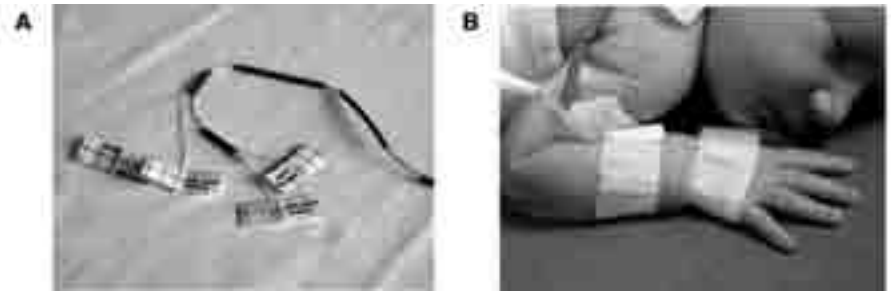


Abbildung 1: (A) Sensoren des Trackingsystems werden in hautfreundliche Pflaster eingebettet. (B) Diese werden auf die Segmente des Arms geklebt.

des ZNS auftreten. Eine ICP manifestiert sich erst im 2. Lebensjahr. Eine frühzeitige Prognosestellung ist im Hinblick auf die Einleitung einer gezielten Therapie zwar wichtig, mit der herkömmlichen Methodik (Bildgebung, neurologische Untersuchung, etc.) aber nicht möglich. Daher kommt der Evaluation der Spontanmotorik eine hohe Bedeutung zu. Um die Spontanmotorik eines Säuglings zu bewerten, werden dessen Bewegungen zunächst gefilmt. Der Säugling liegt dazu rücklings auf einer Matratze und kann sich für ca. 5 min ohne äußere Einwirkung frei bewegen. Anschließend werden die Bewegungen von einem Arzt beurteilt. Es zählt dabei nicht die Quantität, sondern die Qualität der Motorik. Zudem sollen ausdrücklich nicht Details (z.B. Handöffnung), sondern die Gestalt der Bewegungen in ihrer Gesamtheit beschrieben werden. Dabei spielen Begriffe wie Komplexität und Variabilität von Bewegungen eine wichtige Rolle. Da diese Methode auf abstrakten und eher unscharf definierten Konzepten beruht, ist ihre Anwendung abhängig von Kenntnisstand und Erfahrung des Arztes. Daher ist eine Methode zur objektiven Analyse wünschenswert.

Die hier präsentierte Dissertation verfolgte vornehmlich zwei Ziele. Zum einen sollten Methoden entwickelt werden, um aus Signalen von Bewegungssensoren, die am Säugling befestigt werden, die Bewegungen des Säuglings zu rekonstruieren. Die damit gewonnene quantitative Beschreibung der Spontanmotorik ist die Grundlage für deren Analyse. Zum anderen sollten Unterschiede zwischen den quantifizierten Säuglingsbewegungen von Kindern, die sich normal entwickelt haben, und Kindern, bei denen sich eine ICP ausgebildet hat, aufgefunden werden. Im Folgenden werden grundlegende Ideen und Ergebnisse der Dissertation dargelegt.

2 Studie und Versuchsaufbau

Am Universitätsklinikum Heidelberg wurde eine Studie durchgeführt, um die Spontanmotorik von Säuglingen mit Risiko für ICP quantitativ zu erforschen. Dafür wurden Bewegungen von Kindern einer Risiko- und einer Kontrollgruppe im Alter von drei Monaten sowohl gefilmt als auch mit einem Trackingsystem aufgezeichnet. Die Videoaufnahmen wurden von Experten befundet, d.h. die Bewegungsqualität wurde auf einer Skala von 1 bis 4 bewertet, deren letzte Stufe anomalen Bewegungen entspricht, die auf eine sich

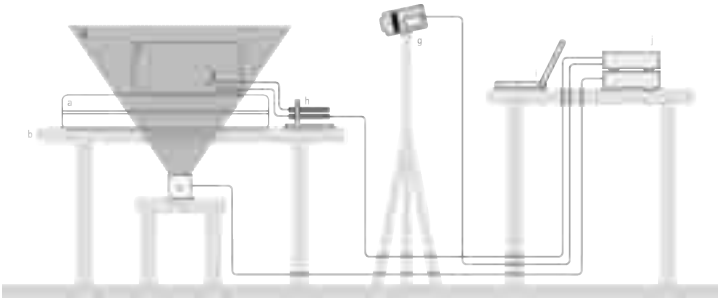


Abbildung 2: Versuchsaufbau: a) Matratze b) Holztisch c) Emitter d) Magnetfeld e) Bereich der Säuglingsbewegungen f) Sensoren g) Videokamera h) Vorverstärker i) Laptop j) Elektronikeinheit

entwickelnde ICP hinweisen. Im Alter von 2 Jahren wurde durch eine neurologische Untersuchung diagnostiziert, ob sich eine ICP entwickelt hat.

Für die Videoaufzeichnung liegen die Kinder rücklings auf einer Matratze und können sich frei bewegen. Um Bewegungen quantitativ zu erfassen, wird ein elektromagnetisches Trackingsystem verwendet, dessen Sensoren am Säugling mit Pflastern befestigt werden — jeweils 4 pro Extremität (s. Abb. 1). Unterhalb der Matratze befindet sich ein Emitter, der ein sehr schwaches Magnetfeld aufbaut (s. Abb. 2). Bewegen sich die Sensoren nun aufgrund der Säuglingsbewegungen in diesem Magnetfeld, so werden Ströme induziert und gemessen. Aus diesen Messungen werden Position und Orientierung eines jeden Sensors bestimmt. Diese Daten sind die Grundlage für die weitere Auswertung.

3 Biomechanisches Modell

Die Bewegungen des Säuglings werden von Sensoren erfasst, die von so genannten technischen Koordinatensystemen (TKS) repräsentiert werden. Die Darstellung anhand dieser Koordinatensysteme (KS) ist nicht dazu geeignet, um Säuglingsbewegungen zu beschreiben: zum einen sind Messungen nicht reproduzierbar, da sie von der exakten Positionierung des Sensors auf der Haut abhängen. Zum anderen stellen die Extremitäten des Körpers kinematische Ketten dar — Systeme aus Segmenten, die durch Gelenke verbunden sind. Dadurch sind Messungen der Sensoren, die an einem Segment wie z.B. dem Arm befestigt sind, stark korreliert. Daher lassen sich feine Bewegungen (z.B. Beugungen der Hand) nicht erkennen, wenn sie von Rotationen eines Gelenks weiter oben in der kinematischen Kette überlagert werden (z.B. im Schultergelenk). Aus diesen Gründen wird eine Abbildung der Sensorbewegungen auf die Bewegungen der Körpersegmente benötigt.

Der Zweck eines biomechanischen Modells ist es, Sensordaten in Bewegungsparameter zu transformieren, die die beobachteten Körperbewegungen möglichst akkurat abbilden. Unter dem Begriff „biomechanisches Modell“ werden im Folgenden die Beschreibung der Bewegungen der Körpersegmente mittels KS und deren Anordnung in kinematischen Ketten sowie Methoden, die aufgezeichnete Sensorbewegungen auf diese KS abbilden (die so

genannte *anatomische Kalibrierung*) zusammengefasst. Verfügt man über solche so genannten Segmentkoordinatensysteme (SKS), die die zeitlich veränderliche Position und Orientierung von Körpersegmenten wie Unterarm und Hand im Raum beschreiben, so lässt sich in einem ersten Schritt die Rotationsmatrix berechnen, die die relative Orientierung dieser Segmente zueinander ausdrückt. In einem zweiten Schritt kann man aus dieser Matrix Rotationswinkel um definierte Körperachsen extrahieren; so lassen sich aus der Rotationsmatrix, die die Orientierung des Unterarms relativ zum Oberarm angibt, Winkel für die Beugung/Streckung und für die Innen-/Außenrotation des Unterarms gewinnen. Damit verfügt man schließlich über eine Menge eindimensionaler Winkelzeitreihen, welche die *Freiheitsgrade* des Modells darstellen und so die Säuglingsbewegungen beschreiben. Für ältere Kinder existieren bewährte Methoden der anatomischen Kalibrierung. Bei Säuglingen gelten aber Randbedingungen aufgrund von *Anatomie* und *Verhalten*:

1. Die Segmente der Extremitäten sind sehr klein und weisen noch keine tastbaren Knochenpunkte auf, anhand derer man SKS ausrichten könnte; weiterhin können keine exakten Kalibrierungsbewegungen wie ein Anheben des Oberarms in definierten Ebenen durchgeführt werden, um eine funktionale Kalibrierung durchzuführen.

2. Während einer mehrminütigen Aufnahme kann es vorkommen, dass der Säugling durch abrupte Bewegungen wie Strampeln einen Zug auf ein Sensorkabel ausübt, wodurch sich die Orientierung des Sensors auf der Haut ändern kann. Dadurch wird aber die Kalibrierung ungültig und die berechneten Zeitreihen fehlerhaft.

Im Folgenden soll skizziert werden, wie durch die entwickelten Verfahren trotz dieser Randbedingungen eine andauernd gültige anatomische Kalibrierung erreicht wird.

3.1 Anatomische Kalibrierung und Kompensation von Bewegungsartefakten

Um Gelenkpositionen und -achsen eines Segments relativ zum TKS des Sensors zu schätzen, werden Annahmen zur mittleren zeitlichen Lage der zu schätzenden anatomischen Parameter getroffen, und es wird durch *zeitliche Integration* der aufgezeichneten Bewegungsdaten eine räumliche Beziehung zwischen diesen Parametern und dem TKS hergestellt. Dies soll an einem Beispiel erläutert werden: Die bauchwärts zeigende Achse des Oberkörpers soll relativ im TKS bestimmt werden. Das Kind liegt während der Aufnahme auf dem Rücken. Dieser Umstand wird für die Annahme genutzt, dass die gesuchte Achse ${}^{\mathrm{s}}\mathbf{z}_{\mathrm{seg}}$ im Mittel in Richtung der z-Achse $\mathbf{z}_{\mathrm{g}}$ des globalen Referenzsystems zeigt, also $\overline{{}^{\mathrm{s}}\mathbf{z}_{\mathrm{seg}}} = \mathbf{z}_{\mathrm{g}}$ gilt. Wenn mit R_{s} die Orientierung des Sensors im globalen Referenzsystem gemessen wird, lässt sich aus der aktuellen und der geschätzten mittleren Orientierung mit

$$ {}^{\mathrm{g}}\mathbf{z}_{\mathrm{seg}}(t) = R_{\mathrm{s}}(t)\overline{R_{\mathrm{s}}}\mathbf{z}_{\mathrm{g}} \tag{1} $$

die Richtung der z-Achse des Oberkörpers zum Zeitpunkt t bestimmen. Aus solcherlei definierten Achsen lassen sich die benötigten SKS der Segmente aufstellen, aus denen wie beschrieben die relativen Gelenkwinkel berechnet werden können.

Wenn durch äußere Einwirkung die Orientierung des Sensors auf der Haut geändert wird, so wird die anatomische Kalibrierung ungültig. Um derartige Störungen zu kompensieren, wird das anatomische SKS zunächst vom TKS des Sensors *entkoppelt*, indem letzteres auf ein so genanntes fixiertes technisches Koordinatensystem (FTKS) abgebildet wird, welches zwar an anatomischen Punkten ausgerichtet ist, jedoch jederzeit — und damit auch *nach* einer äußeren Einwirkung — neu aus Bewegungsdaten berechnet werden kann: TKS $\rightarrow$ FTKS. Um ein KS mit diesen gewünschten Eigenschaften aufstellen zu können, werden für das Segment die zwei Rotationszentren bestimmt, die das Segment an beiden Seiten begrenzen. Dabei handelt es sich um anatomische Punkte, deren Position in den TKS der entsprechenden Sensoren eigentlich zeitlich konstant sind. Rotationszentren lassen sich aus gemessenen Sensorbewegungen bestimmen: Die zeitlich konstante Position ${}^{\mathrm{s}_1}\mathbf{p}_{\mathrm{p}}$ eines Rotationszentrums im TKS des ersten, proximalen Sensors s_1 und die Position ${}^{\mathrm{s}_2}\mathbf{p}_{\mathrm{d}}$ im TKS des zweiten, distalen Sensors s_2 müssen so gewählt werden, dass im Mittel die Distanz ihrer zeitlich veränderlichen Positionen im globalen Referenzsystem minimal wird — denn sie sollten ja räumlich aufeinanderfallen. Wird das Rotationszentrum in den TKS der Sensoren aber als *zeitlich veränderlich* aufgefasst und wie folgt berechnet

$$\underset{{}^{\mathrm{s}_1}\mathbf{p}_{\mathrm{p}}(t),{}^{\mathrm{s}_2}\mathbf{p}_{\mathrm{d}}(t)}{\operatorname{argmin}} \frac{1}{w}\int_w ({}^{\mathrm{g}}\mathbf{p}_{\mathrm{p}}(t) - {}^{\mathrm{g}}\mathbf{p}_{\mathrm{d}}(t))^2 \mathrm{d}t \tag{2}$$

so passt sich seine Position nach einer äußeren Einwirkung der neuen Orientierung des Sensors auf der Haut an. Das auf Basis dieser anatomischen Punkte definierte FTKS ist folglich in der Lage, solch eine Störung zu kompensieren. Damit kann das anatomische SKS relativ zum FTKS kalibriert werden: TKS $\rightarrow$ FTKS $\rightarrow$ SKS.

Um das biomechanische Modell zu validieren, wird der *Modellfehler* in Form der Residuen zwischen gemessenen Sensorpositionen und -orientierungen und den anhand der Modellparameter rekonstruierten Sensorwerten berechnet. Diese Statistik akkumuliert Ungenauigkeiten aufgrund vereinfachender Modellannahmen. Im Mittel betrug der Fehler 18,7 mm und 4,3° für den Arm bzw. 14,3 mm und 4,1° für das Bein. Anhand dieser Richtwerte können fehlerhafte Aufnahmen identifiziert werden. Bei Aufnahmen ohne äußere Störung sollte das für die Kompensation berechnete FTKS seine Orientierung gegenüber dem TKS nicht ändern. Um nun festzustellen, wie zuverlässig die Bestimmung des FTKS ist, wurde für Aufnahmen ohne Störung bestimmt, wie hoch die Abweichung zwischen den beiden KS ist. Die mittlere Abweichung ist mit 4° sehr gering; im Vergleich dazu kamen bei Störungen Abweichungen von 90° vor.

4 Analyse der kindlichen Spontanmotorik

Typischen kindlichen Spontanbewegungen werden abstrakte Eigenschaften wie *Variabilität* und *Komplexität* zugeschrieben [HA04]. Es ist schwierig, solche Qualitäten *sprachlich konkret* zu definieren. Dies ist aber nötig, um die Methode der subjektiven Analyse der Spontanmotorik neuen Anwendern beizubringen und um sich über sie zu verständigen. Daher wird in der einschlägigen Literatur versucht, konkretere Aussagen zu treffen. Beispielsweise wird beschrieben, dass typische, normale Bewegungen durch eine Kombinati-

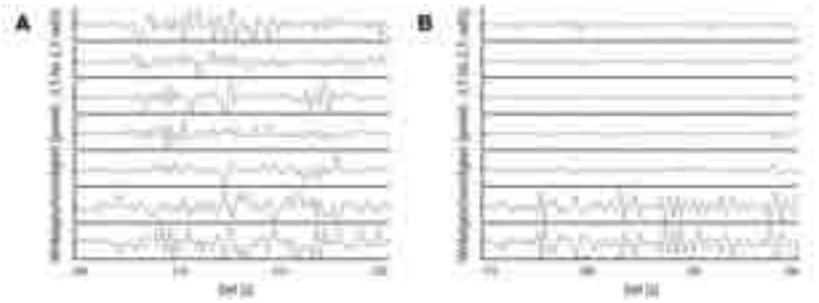


Abbildung 3: Heterogenität kindlicher Spontanmotorik: Zu sehen ist exemplarisch die Winkelgeschwindigkeit aller Modellfreiheitsgrade des Arms *eines* Kindes. Beide Abschnitte zeigen nach Aussage von Ärzten typische Spontanmotorik. (A) Alle Freiheitsgrade sind an der Bewegung beteiligt. (B) Nur in zwei Zeitreihen ist Aktivität zu verzeichnen.

on der Rotationen in den verschiedenen Gelenkfreiheitsgraden entstehen, und dass durch Rotationen entlang der Gelenkachsen der Eindruck von Eleganz entstände. Versuche, solche konkreten Beschreibungen direkt in objektive Merkmale auf Basis der quantitativen Bewegungsdaten zu überführen, waren nicht erfolgreich. Es konnten keine objektiven Merkmale gefunden werden, die den sprachlichen Beschreibungen entsprechen und sich zwischen den Gruppen signifikant unterscheiden. Diese Auswertungen konnten exemplarisch durch Fälle illustriert werden, die offensichtlich nicht diesen Umschreibungen entsprechen. Damit konnte exemplarisch gezeigt werden, wie heterogen „normale“ Bewegungen sind. So zeigt Abb. 3 zwei Abschnitte von Bewegungen eines gesunden Kindes, die von Ärzten als typisch befundet wurden. Im zweiten Abschnitt sind fast ausschließlich in den zwei Freiheitsgraden der Schulter Bewegungen zu erkennen; es fehlt also sowohl die beschriebene Kombination aus Bewegungen verschiedener Gelenke als auch die Rotationen entlang der Gelenkachsen. Als zielführend stellte sich die Formalisierung eigener Beobachtungen heraus: dem Autor waren bei intensiver Betrachtung von Videoaufnahmen einige ausgeprägte Bewegungen bei Kindern der ICP-Gruppe aufgefallen. Die Formalisierung dieser Beobachtungen geben einen Einblick hinsichtlich der Struktur der Säuglingsbewegungen und lassen sich als objektive Merkmale zur Identifikation von anomalen Bewegungen nutzen. Diese Merkmale lassen sich wiederum mit den abstrakten Konzepten *Variabilität* und *Komplexität* in Einklang bringen. Im Folgenden soll dargestellt werden, welche Besonderheiten anomale Spontanmotorik auszeichnen.

4.1 Zeitliche und räumliche Anordnung

Bewegungen sind sowohl zeitlich — es gibt Bewegungsphasen und Ruhephasen — als auch räumlich verteilt, d.h. Bewegungs- und Ruhephasen verteilen sich auf die Modellfreiheitsgrade entlang der kinematischen Kette. Um diese zeitliche und räumliche Anordnung zu untersuchen, wurde eine *Datenabstraktion* vorgenommen. Ziel dieser Abstraktion war es, die Zeitreihen in die genannten Bewegungs- und Ruhephasen aufzuteilen. Die Kodie-

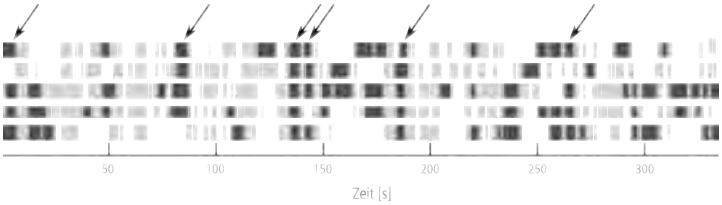


Abbildung 4: Aufteilung von Zeitreihen in Bewegungs- (rot) und Ruhephasen (grau). Jede Zeile stellt einen Modellfreiheitsgrad (z.B. Beugung des Knies) des Beins dar. Die Pfeile markieren Zeitpunkte, an denen synchrone Bewegungen in mehreren Freiheitsgraden auftreten.

rung funktioniert wie folgt: Ein Zeitabschnitt $[t_1, t_2]$ gehöre zur Menge der Intervalle $\mathbf{B}$, in denen Bewegungen vorkommen, wenn innerhalb des Abschnitts eine bestimmte Grenzgeschwindigkeit v_1 überschritten wird. Da nur diejenigen Abschnitte von Belang sind, in denen — im Gegensatz zu einem langsamen Trend — eine deutliche Bewegung zu sehen ist, soll innerhalb des Abschnitts mindestens ein mal eine höhere Grenzgeschwindigkeit v_2 überschritten werden.

$$\underset{t_1<t<t_2}{\forall} \mathrm{v}(t) > v_1 \wedge \underset{t}{\exists}\, \mathrm{v}(t) > v_2 \Rightarrow [t_1, t_2] \in \mathbf{B} \tag{3}$$

In Abb. 4 sind die Bewegungsabschnitte der Zeitreihen des Beins für ein Kind mit Outcome ICP rot markiert. Es ist zu erkennen, dass mehrmals in mehreren Gelenkfreiheitsgraden synchron Bewegungen auftreten. Ein Merkmal *isolierte Bewegungen*, das auf Basis dieser Beobachtung definiert wurde, zeigte bei vier von sieben Kindern der Stichprobe mit Outcome ICP sehr hohe Werte.

4.2 Stereotype Bewegungen

Variabilität gilt als wichtiges Merkmal typischer Säuglingsbewegungen. In Zusammenhang damit fiel beim Studium der Videoaufnahmen auf, dass Kinder mit Outcome ICP oft stereotypes Verhalten zeigten. Daher sollte die Hypothese geprüft werden, dass stereotype Bewegungen auf ICP hinweisen. Dafür wurde ein Verfahren entwickelt, um den Grad der Selbstähnlichkeit der quantifizierten Bewegungen zu beurteilen. Mit diesem Ansatz — es werden keine Muster vorgegeben — wird ein individuelles Bewegungsprofil erstellt.

Für die Bewertung der Selbstähnlichkeit zweier Abschnitte einer Winkelzeitreihe wurde das Distanzmaß Dynamic Time Warping (DTW) [SC78] verwendet, welches zwei Signalen dann einen niedrigen Wert zuweist, wenn diese eine ähnliche Form aufweisen. Dies entspricht eher dem menschlichen Ähnlichkeitsbegriff als die euklidische Distanz. Damit mittels dieses Distanzmaßes die Selbstähnlichkeit der Bewegungen einer Extremität wie dem Arm bewertet werden kann, müssen zwei Fragen beantwortet werden.

1. Welche Teilsegmente einer Zeitreihe werden miteinander verglichen?

2. Die Säuglingsbewegungen werden von mehreren Zeitreihen beschrieben. Wie kombiniert man deren Selbstähnlichkeitsmaße?

Ein Vergleich aller möglichen Teilsegmente einer Zeitreihe ist wegen der exponentiellen Laufzeit eines entsprechenden Algorithmus' unerwünscht, aber auch nicht sinnvoll, da so auch Segmente verglichen würden, die gar keine Bewegungen beschreiben. Daher wird die Heuristik verwendet, dass gerade die in Gl. 3 definierten Segmente verglichen werden. Um die Selbstähnlichkeit so zu beschreiben, dass man die Beschreibung mehrerer Winkelzeitreihen einer Extremität zu einer Gesamtbeschreibung der Selbstähnlichkeit kombinieren kann, wird eine Funktion mit folgenden Eigenschaften definiert:

- Jedem Paar von Zeitpunkten (t_1, t_2) wird ein Wert für Selbstähnlichkeit zugewiesen.
- Die Funktion wird auf einen festen Wertebereich normiert.
- Der Funktionswert wächst mit steigender Selbstähnlichkeit.

Die Funktion s bildet die Zeitpunkte (t_1, t_2) auf die exponentialtransformierte DTW-Distanz ab, wenn beide Zeitpunkte innerhalb von Bewegungssegmenten liegen, ansonsten auf 0.

$$\mathrm{s}_\iota(t_1, t_2) = \begin{cases} e^{-\mathrm{dtw}(B_1, B_2)} & \mid \underset{B_i \in \mathbf{B}}{\exists} t_k \in \mathbf{B}_k \underset{k=1,2}{\forall} \\ 0 & \text{sonst} \end{cases} \tag{4}$$

Durch Summation der mit dieser Funktion transformierten Winkelzeitreihen kann man so eine Gesamtbeschreibung der Selbstähnlichkeit einer Extremität erhalten, auf Basis derer sich statistische Maßzahlen bestimmen lassen. Abb. 5 zeigt eine Visualisierung dieses Maßes für Selbstähnlichkeit am Beispiel eines Kindes mit Outcome ICP.

4.3 Prognose der Cerebralparese auf Grund subjektiver und objektiver Analyse

Die Urteiler erreichten mit ihrer Prognose auf Basis der subjektiven Analyse auf der Stichprobe der Studie durchgehend eine sehr hohe Sensitivität; in der Regel befanden sich alle ICP-Fälle in der Gruppe der Kinder, deren Bewegungen als anomal befundet wurden. Allerdings entwickelten jeweils etwa die Hälfte der Kinder dieser Gruppe keine ICP, d.h. bezogen auf diesen Outcome waren etwa 50% falsch positiv. Zudem war die Urteilerübereinstimmung mit einem Wert von $0,55$ für Cohens κ nicht sehr hoch. Diese Ergebnisse zeigen nochmals auf, dass eine objektive Methode wünschenswert ist, die den Arzt bei der Prognosestellung unterstützt und die Zahl der falsch positiven Prognosen reduziert.

Die objektiven Merkmale wurden retrospektiv auf der Stichprobe ermittelt und erlauben daher keine unverzerrte Beurteilung ihres prognostischen Wertes. Da aber keine bewusste Optimierung bezüglich der Trennfähigkeit der Gruppen vorgenommen wurde, ermöglicht die Betrachtung der Verteilung der Merkmale in den Gruppen trotzdem eine Aussage über deren Zusammenhang mit dem Outcome. Mit dem Merkmal *isolierte Bewegungen* und einem weiteren Merkmal *v-moderat-Bein*, das ausdrückt, in wie weit in den Beingelenken

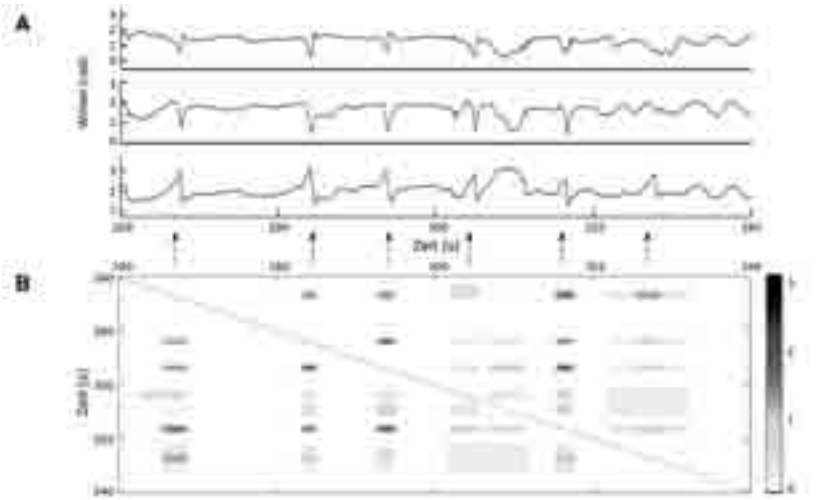


Abbildung 5: (A) Winkelzeitreihen von Ellenbogen und Schulter aus der Aufnahme eines Säuglings mit Outcome ICP. Es sind stereotype Abschnitte zu sehen (siehe Pfeile), die im Video als Armrudern zu erkennen sind. (B) Visualisierung der Selbstähnlichkeit dieser Aufnahme entsprechend Gl. 4.

Bewegungen in einem „mittleren“ Geschwindigkeitsbereich zu beobachten sind, lassen sich die Gruppen der Stichprobe durch Festlegen von Grenzwerten mit einer Sensitivität von 100%, einer Spezifität von 97% und einem positiv prädiktiven Wert von 78% trennen. Aufgrund der quantitativen Analyse lässt sich feststellen, dass sich die 7 Kinder der Stichprobe mit Outcome ICP durch isolierte Bewegungen, durch Abwesenheit ausgeprägter Beinbewegungen und durch stereotype Armbewegungen auszeichnen.

5 Diskussion und Ausblick

Das herkömmliche Vorgehen zur Erforschung der kindlichen Spontanmotorik beginnt damit, dass dem Arzt Bewegungsmuster oder Bewegungen einer bestimmten Qualität auffallen. Um zu prüfen, ob es sich dabei um *typische* Bewegungen handelt, die vielleicht mit dem Auftreten einer Pathologie korreliert sind, muss er nun unter diesem Gesichtspunkt zahlreiche Videoaufnahmen kindlicher Spontanmotorik anschauen. Dieses Vorgehen ist bewährt, aber sehr zeitaufwändig und deshalb oft nicht umsetzbar. Mit dem in dieser Arbeit entwickelten computergestützten Verfahren lassen sich Hypothesen dagegen sehr schnell und einfach überprüfen. Am Anfang dieses Prozesses steht immer noch die Beobachtung von auffälligen Bewegungen bzw. der visuelle Eindruck einer Bewegungsqualität. Diese Beobachtung muss nun aber in eine quantitative Beschreibung übersetzt werden, diese lässt sich dann automatisch auf alle Aufnahmen anwenden. Der kritische Schritt hierbei ist die *Formalisierung* dieser Beschreibung. Dies ist ein anspruchsvoller Vorgang, der ein gewisses Maß an Erfahrung und mathematische Kenntnisse verlangt. Ist die Formalisierung gelungen, so hat man allein damit schon ein tieferes Verständnis der Beobachtung erlangt.

Die Formalisierung hat nicht nur den Vorteil, dass Aufnahmen quantitativ bewertet werden können, sondern es können auch automatisiert Abschnitte einer Videoaufnahme ausgewählt werden, die die Bedingungen einer definierten anomalen Bewegung erfüllen. Dies könnte der Lehre und dem *Training* der subjektiven Analyse helfen, da die zeitliche Festlegung und die Konkretisierung der Bezeichnungen weniger Platz für Missverständnisse und Misskonzeptionen lassen als die Beschreibung einer ganzen Aufnahme mit abstrakten Begriffen. Bei einer subjektiven Methode ist es schwierig, das implizit vorhandene Wissen oder Fähigkeiten wie etwa die „Wahrnehmung der Bewegungsqualität“ an andere weiterzugeben, es also zu externalisieren. Das Mittel der Wahl ist oft „learning by doing“, so wird die subjektive Analyse hauptsächlich anhand von Beispielvideos vermittelt; diese Videos sind der erste Schritt zum Aufbau *eigener* Erfahrungen; ein Prozess, der Zeit kostet. Die Externalisierung mit dem Mittel der Sprache — die u.a. nötig ist, um in der wissenschaftlichen Gemeinschaft zu diskutieren — ist problematisch, da man mit Umschreibungen arbeiten muss. Dazu sind diese Beschreibungen kaum evaluierbar; es kann schlecht festgestellt werden, ob die eigenen Wahrnehmungen auf die richtigen beschreibbaren Eigenschaften reduziert wurden. Die formalisierten Beobachtungen dagegen *können* evaluiert werden — und stehen im Einklang mit den abstrakten Begriffen der subjektiven Analyse der Spontanmotorik. Dank der Formalisierung kann eine messbare und evaluierbare Externalisierung des Wissens erfolgen. Zu guter Letzt lässt sich die computergestützte Methode direkt im klinischen Einsatz verwenden. Die quantitative Auswertung kann Ärzte bei der Entscheidungsunterstützung unterstützen, um eine bessere Prognose zu erreichen.

Literatur

[HA04] M Hadders-Algra. General Movements: A Window for Early Identification of Children at High Risk for Developmental Disorders. *J Pediatr*, 145:12–18, 2004.

[PEC+97] HFR Prechtl, C Einspieler, G Cioni, AF Bos, F Ferrari und D Sontheimer. An early marker for neurological deficits after perinatal brain lesions. *Lancet*, 349:1361–63, 1997.

[SC78] H Sakoe und S Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE T Acoust Speech*, 26(1):43–49, 1978.

Dominik Karch studierte an der Ruprecht-Karls-Universität Heidelberg Medizinische Informatik mit dem Schwerpunkt Bild- und Signalverarbeitung. Während des Studiums verbrachte er ein Semester an der Università Politecnica delle Marche (Ancona, Italien), wo er eine Arbeit zur statistischen Analyse von Gangstörungen schrieb. Von 2007 bis 2011 arbeitete er als wissenschaftlicher Mitarbeiter am Institut für Medizinische Biometrie und Informatik der Medizinischen Fakultät Heidelberg, wo er u.a. in einem DAAD-Projekt mit der Universidad de Chile arbeitete, und promovierte bei Prof. Dr. Dickhaus mit summa cum laude. Seit 2011 arbeitet er bei der Berlin Heart GmbH in der Forschung und Entwicklung.

Mobile Intention Recognition

Peter Kiefer

Geoinformations-Engineering
Institut für Kartografie und Geoinformation*
ETH Zürich
pekiefer@ethz.ch

Abstract: Das Problem der mobilen Intentionserkennung besteht darin, aus dem raumzeitlichen Verhalten eines Agenten Rückschluss auf seine Intentionen zu ziehen. Ein mobiler Dienst, der stets über die aktuelle Intention des Nutzers Bescheid wüsste, könnte eine bessere mobile Assistenz bieten als ein herkömmlicher ortsabhängiger Dienst. Ein zentrales Problem in der mobilen Intentionserkennung besteht darin, den für die Verhaltensinterpretation relevanten räumlichen und zeitlichen Kontext zu bestimmen. Bisherige Verfahren nehmen diesbezüglich an, dass lediglich ein lokaler zusammenhängender Kontext relevant ist. Der Hauptbeitrag dieser Arbeit besteht in zwei neuen, auf formalen Grammatiken verschiedener Komplexität beruhenden Formalismen, mit denen sich eine größere Problemklasse modellieren und interpretieren lässt. Die größere Ausdrucksmächtigkeit wird an Hand eines in der mobilen Intentionserkennung häufig vorkommenden Verhaltensmusters gezeigt, des 'Return-to-region' Musters, das nicht auf einen lokalen Kontext beschränkt ist und mehrfach überlappend auftreten kann.

1 Einführung

Viele der heute auf dem Markt verfügbaren ortsbezogenen Dienste unterstellen, dass der aktuelle räumliche Kontext des Nutzers ausreichend Aufschluss auf sein Informationsbedürfnis gibt: Betritt man eine bestimmte Region, so erhält man automatisch den zur Region gehörenden Informationsdienst (vgl. Abb. 1, links). Diese Annahme ist häufig nicht zutreffend, wie sich leicht am sogenannten Raumdurchquerungsproblem erläutern lässt. Dieses Problem (orig.: 'room-crossing problem') wurde zuerst von [Sch05] im Kontext eines mobilen Museumsführers eingeführt und tritt dann auf, wenn das Betreten einer Region nicht mit einem Interesse an dieser Region verbunden ist. So durchquert beispielsweise der Nutzer des mobilen Tourismusführers in Abb. 1 (rechts) auf der Suche nach einem bestimmten Gebäude mehrere für ihn uninteressante Regionen. Ein einfacher ortsbezogener Dienst würde in dieser Situation ständig unbenötigte Informationsdienste starten und damit eher stören als helfen.

*Die zu Grunde liegende Dissertation ist entstanden an der Fakultät für Wirtschaftsinformatik und Angewandte Informatik, Otto-Friedrich-Universität Bamberg, Deutschland [Kie11b], und wurde im Anschluss veröffentlicht als [Kie11a]

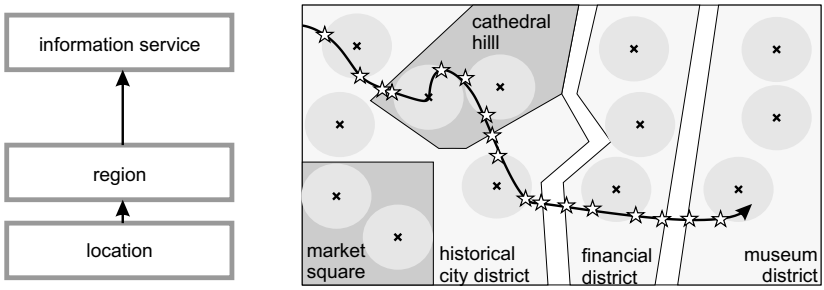


Abbildung 1: Links: Architektur eines einfachen ortsbezogenen Dienstes. Rechts: Durchqueren von Regionen bei der Wegsuche ('room-crossing problem'), vgl. [Kie11a, Abb. 1.1, 1.2].

Das Raumdurchquerungsproblem lässt sich auf zwei Defizite einfacher ortsbezogener Dienste zurückführen. Erstens bieten die meisten Orte mehr als eine Affordanz [JRGE98]. Eine eins zu eins Abbildung von Ort auf Informationsbedürfnis ist daher im Allgemeinen nicht möglich. Zweites Defizit ist die zeitliche und räumliche Beschränktheit. Ein einfacher ortsbezogener Dienst berücksichtigt nur den kleinsten räumlichen Kontext im kleinstmöglichen Zeitintervall. Das Verhalten des Touristen in Abb. 1 (rechts) zum Beispiel ließe sich vielleicht besser unter zu Hilfenahme des vorherigen Verhaltens und mit Bezug auf eine der Eltern-Regionen der Partonomie interpretieren.

Der vorliegende Text – eine Zusammenfassung der Dissertation zum gleichnamigen Thema [Kie11a] – schlägt einen Ansatz zur intelligenteren mobilen Assistenz durch *mobile Intentionserkennung* vor. Die mobile Intentionserkennung zielt auf die Erkennung des kognitiven Zustands des Nutzers ab und ist ein Spezialfall des seit den 80er Jahren in der künstlichen Intelligenz bekannten Problems der Planerkennung [Kau87], des inversen Problems des Planens.

Der Schwerpunkt dieser Arbeit liegt insbesondere auf zwei neuen Repräsentationsformalismen, mit denen sich auch solche Problemklassen der mobilen Intentionserkennung modellieren lassen, bei denen der bei der Interpretation zu berücksichtigende Kontext nicht lokal beschränkt und nicht notwendigerweise zusammenhängend ist. Dies ist in der mobilen Intentionserkennung häufig der Fall, beispielsweise wenn ein Nutzer zu einer vorher besuchten Region zurückkehrt ('Return-to-region' Muster). Die beiden Formalismen basieren auf formalen Grammatiken mit verschiedener Ausdrucksmächtigkeit – kontextfrei bzw. mild kontextsensitiv (Baumadjunktionsgrammatiken) – und sind für häufig auftretende Problemklassen der mobilen Intentionserkennung geeignet. Intentionserkennung wird dabei zu einem Parsingproblem, das für beide Formalismen effizient lösbar ist.

Die folgenden Abschnitte orientieren sich an den Kapiteln der Dissertation: Abschnitt 2 führt die intentionsabhängigen mobilen Dienste ein. Abschnitt 3 stellt einen Bezug zu vorherigen Arbeiten in der Planerkennung her. Die räumlich beschränkten Grammatiken werden in Abschnitt 4 vorgestellt. Die Anwendung der neuen Formalismen auf das 'Return-to-region' Muster wird in Abschnitt 5 zusammengefasst. Abschnitt 6 zieht Resumee und gibt einen Ausblick auf offene Fragen in der mobilen Intentionserkennung.

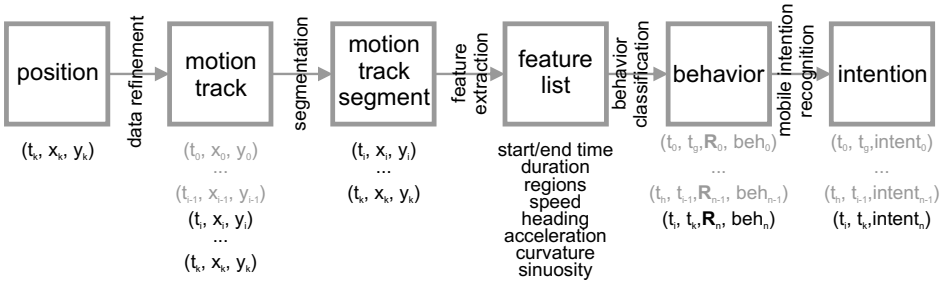


Abbildung 2: Architektur eines intentionsabhängigen mobilen Dienstes, vgl. [Kie11a, Abb. 2.2].

2 Mobile Intentionserkennung

Die Argumentation in Abschnitt 1 hat gezeigt, dass eine direkte Abbildung von Positionsdaten auf Informationsdienste zu keiner zufriedenstellenden mobilen Assistenz führt. Daher führt die Dissertation [Kie11a] in Kapitel 2 die *intentionsabhängigen mobilen Dienste* ein. Diese ermitteln zunächst die Intention des Nutzers und bieten dann einen dazu passenden Dienst. Zur Ermittlung der Intentionen ist eine sehr große semantische Lücke zwischen Sensordaten (Position) und kognitiven Prozessen (Intention) zu schließen [Kau87].

Die Arbeit stellt hierzu eine Mehrschichtenarchitektur vor, die vereinfacht in Abb. 2 dargestellt ist: Die Positionsdaten werden aufgenommen und nach vom Anwendungsfall abhängigen Kriterien inkrementell in Bewegungssegmente zerlegt (vgl. z.B. [BDvKS10]). Das Segment wird dann an Hand verschiedener geometrischer und raum-zeitlicher Eigenschaften klassifiziert, was als Ausgabe sogenannte Verhalten ('behaviors') ergibt, annotiert mit Region, Start- und Endzeitpunkt. Typische 'behaviors' wären beispielsweise *Suchen* (für langsame 'kreuz-und-quer' Bewegung) oder *Rennen* (für schnelle Bewegung geradeaus). Anzumerken ist, dass die Verfahren zur mobilen Intentionserkennung in den folgenden Abschnitten unabhängig von einer bestimmten Architektur oder einem bestimmten Algorithmus zur Segmentierung oder Klassifikation sind.

Das Problem der Intentionserkennung besteht darin, aus der (im Allgemeinen noch unvollständigen) raum-zeitlichen Verhaltenssequenz die jeweils aktuelle Intention zu berechnen. Wichtiger Beitrag von Kapitel 2 ist die Abgrenzung der mobilen Intentions- zur allgemeinen Planerkennung: Intentionserkennung will nicht die komplette Planstruktur erkennen, sondern gibt sich mit der aktuellen Intention zufrieden. In der allgemeinen Planerkennung sind die Eingaben außerdem nicht geordnet, wodurch im schlimmsten Fall jede Eingabe mit jeder anderen zu möglichen Planstrukturen kombiniert werden könnte. In der mobilen Intentionserkennung hingegen sind Eingaben geordnet und durch Regionen strukturiert, was gleichzeitig auch den Suchraum strukturiert.

Gruppiert man jedoch strikt nur benachbarte Eingaben, so lassen sich bestimmte raumzeitliche Abhängigkeiten in der mobilen Intentionserkennung nicht ausdrücken: Abbildung 3 zeigt Beispiele, in denen ein Nutzer zu einer Region zurückkehrt (rechtes Beispiel) bzw. eine Region betritt, die in einer bestimmten Relation (visibleSouthWest) zu einer

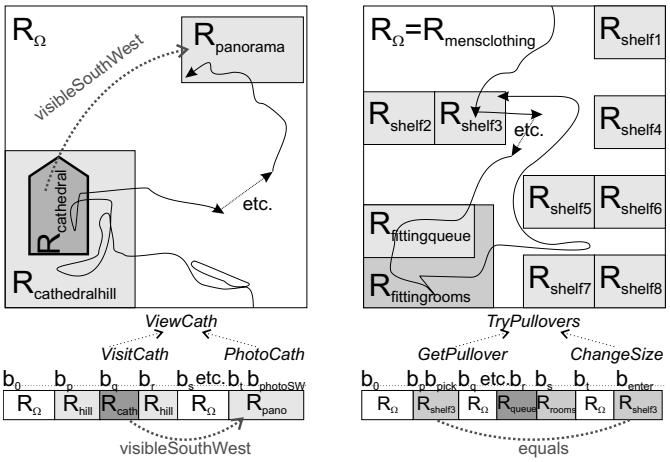


Abbildung 3: Beispiele für kontextübergreifende raum-zeitliche Abhängigkeiten in der mobilen Intentionserkennung (links: Tourismusführer, rechts: Einkaufsführer), vgl. [Kie11a, Abb. 2.19].

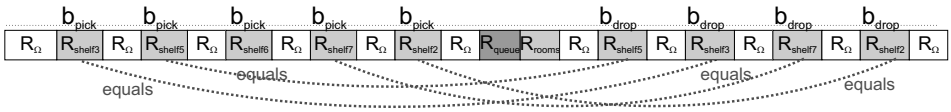


Abbildung 4: Überlappende kontextübergreifende raum-zeitliche Abhängigkeiten (Beispiel Einkaufsführer): pick_1 pick_2 pick_3 pick_4 drop_2 drop_1 drop_3 drop_4, vgl. [Kie11a, Abb. 2.20].

vorher besuchten Region steht (linkes Beispiel). Die zugehörigen Intentionen ‘PhotographiereKathedrale’ bzw. ‘TauscheGröße’ können nur erkannt werden, wenn zum Zeitpunkt des Betretens bekannt ist, dass (und welches) Verhalten vorher aufgetreten ist. Die Verhalten dazwischen sind hierbei nicht relevant. Der zu berücksichtigende raum-zeitliche Kontext ist hier also nicht zusammenhängend. Abbildung 4 zeigt ein Beispiel dafür, dass die kontext-übergreifenden Abhängigkeiten auch überlappend auftreten können. Die beschriebenen kontext-übergreifenden Abhängigkeiten und die bisher für sie in der Fachliteratur fehlenden Ansätze sind Hauptmotivation für die in Abschnitt 4 vorgestellten Verfahren. Ausschlaggebend für die Überlegungen ist auch, dass die Abhängigkeiten nur in beschränktem Maße auftreten, so dass die mobile Intentionskennung weit von der Komplexität der allgemeinen Planerkennung entfernt ist.

3 Relevante Fachliteratur

Das Kapitel 3 der Dissertation [Kie11a] bietet einen Überblick über relevante Fachliteratur im Bereich Planerkennung. Zusammenfassend zeigt sich, dass frühe Ansätze die zeitlichen und räumlichen Aspekte des Problems weitgehend ausgeblendet haben und eine zu hohe

algorithmische Komplexität aufweisen [SSG78, Kau87]. Neuere Ansätze aus dem Bereich 'Mobile Computing' arbeiten häufig auf einer tieferen semantischen Ebene, d.h. sie erkennen keine Intentionen, sondern vielmehr Aktivitäten oder die wahrscheinlichste nächste Position. [AS03] ist hierfür ein typisches Beispiel, und hat gleichzeitig auch – wie andere in der Arbeit beschriebene Ansätze – eine zu niedrige Ausdrucksmächtigkeit für die mobile Intentionserkennung (Endlicher Automat bzw. Markov Modell).

Hoher Popularität erfreuten sich lange Zeit auch probabilistische Ansätze, nämlich naive [CG93] bzw. dynamische Bayes Netzwerke (z.B. [LPFK07, Bui03]). Während erstere nicht für dynamische Prozesse geeignet sind, können letztere schon rein prinzipiell keine kontext-übergreifenden Abhängigkeiten abdecken. Grund sind die Unabhängigkeitsannahmen, die zwischen den Zeitscheiben in einem dynamischen Bayes-Netzwerk notwendig sind.

Mobile Intentionserkennung kann mit formalen Grammatiken modelliert und durch Parsingalgorithmen gelöst werden. So schlägt [Pyn99] 'Probabilistic State-Dependent Grammars' vor, bei denen Produktionen abhängig von einer allgemeinen Zustandsvariable sind. Auch hier gilt – wie bei den dynamische Bayes Netzen – dass sich ohne Explosion der Komplexität in diese allgemeine Zustandsvariable keine unbegrenzte Historie enkodieren lässt, wie sie für das 'Return-to-region' Muster notwendig wäre. In [Sch05] sind die Produktionen nur von einem Zustand abhängig, nämlich der räumlichen Region. Dadurch sind die vorgeschlagenen 'Spatially Grounded Intentional Systems' (SGISs) für den Modellierer einfach zu verstehen. Andererseits sind sie jedoch strikt kontextfrei, was eine Modellierung von 'Return-to-region' verhindert.

Ein Schritt in Richtung kontext-übergreifende Abhängigkeiten stellen [GS07] vor, die strukturelle Parallelen zwischen natürlicher Sprachverarbeitung und Planerkennung ziehen und daraus folgern, dass mild kontextsensitive Grammatiken zur Modellierung von Überkreuzabhängigkeiten gut geeignet wären. Ein konkreter Formalismus wird jedoch nicht vorgestellt und raum-zeitliche Aspekte spielen keine gesonderte Rolle. Eine mild kontextsensitive Grammatik speziell für die mobile Intentionserkennung wird im folgenden Abschnitt vorgestellt.

4 Räumliche Grammatiken

Die folgenden Formalismen zur mobilen Intentionserkennung bauen auf formalen Grammatiken auf. Gemeinsam ist ihnen, dass Intentionen als Nicht-Terminale verwendet werden, Paare aus Region und Verhalten (beh, R) werden zu Terminalen. Im Unterscheid zur allgemeinen Architektur aus Abb. 2 werden also die genauen Zeitpunkte ignoriert[1]. Die Regeln der Grammatiken definieren mögliche Beziehungen zwischen Intentionen, Sub-Intentionen und Verhalten. Algorithmisch gesehen ergibt sich aus der Formalisierung mobiler Intentionserkennung mit räumlichen Grammatiken ein inkrementelles Parsing Problem mit räumlichen Constraints (siehe hierzu Abschnitt 4.3).

[1] Eine temporale Erweiterung des Ansatzes ist in [KRS10] zu finden.

4.1 Spatially-Constrained Context-Free Grammars

Der erste neue Repräsentationsformalismus zur mobilen Intentionserkennung, mit dem man kontextübergreifende Abhängigkeiten zwischen Intentionsn modellieren kann, sind die räumlich beschränkten kontextfreien Grammatiken (Spatially-Constrained Context-Free Grammars, SCCFGs)[2]. Als Unterschied zu einer herkömmlichen kontextfreien Grammatik können in einer SCCFG zwei Arten räumlicher Constraints annotiert werden:

- Grounding Constraints (bezüglich einer Menge von Regionen **R**): Diese sagen aus, dass alle Symbole (d.h. Intentionen und Verhalten) der rechten Seite einer Regel in einer Region aus **R** stattfinden müssen.
- Non-Local Constraints (annotiert mit einem räumlichen Relationstyp *t*): Zwei Intentionen der rechten Hand werden verknüpft. Es ergibt sich eine räumliche Abhängigkeit vom Typ *t* zwischen den Regionen, in denen die Intentionen stattfinden.

Die folgende SCCFG-Regel beispielsweise legt fest, dass die Intention *BuyPullover* nur in der Region *Shop* auftreten kann. Sie setzt sich zusammen aus dem Nehmen eines Kleidungsstücks, einer (in weiteren Regeln zu spezifizierenden) Intention *ContinueShopping* sowie dem Ablegen eines Kleidungsstücks. Der nicht-lokale Constraint (0,2,equals) sagt aus, dass die Regionen von Pick und Drop in einer Relation equals stehen, d.h. dieselben sein müssen. Man legt ein Kleidungsstück im Idealfall also immer dort zurück, wo man es vorher genommen hat.

$$\mathrm{I}_{\mathrm{BuyPullover}} \quad \rightarrow \quad \mathrm{I}_{\mathrm{Pick}}\ \mathrm{I}_{\mathrm{ContinueShopping}}\ \mathrm{I}_{\mathrm{Drop}} \qquad \{\mathrm{R}_{\mathrm{Shop}}\},\ \{(0,\ 2,\ \mathit{equals})\}$$

4.2 Spatially-Constrained Tree-Adjoining Grammars

Die nicht-lokalen Constraints einer SCCFG ermöglichen die Modellierung kontextübergreifender Abhängigkeiten. In Abbildung 4 haben wir gesehen, dass diese auch überkreuzend auftreten können. Regelanwendungen in einer SCCFG können diese Überkreuzungen nicht dynamisch erzeugen. Daher reicht die Ausdrucksmächtigkeit der SCCFG für viele Anwendungsfälle der mobilen Intentionserkennung nicht aus.

Die Dissertation führt deshalb die räumlich beschränkten Baumadjunktionsgrammatiken (Spatially-Constrained Tree-Adjoining Grammars, SCTAGs) ein. Diese erweitern die in der natürlichen Sprachverarbeitung für Überkreuzabhängigkeiten verwendeten Baumadjunktionsgrammatiken um räumliche Constraints. Die Arten räumlicher Constraints entsprechen hierbei den oben erwähnten Grounding und Non-Local Constraints.

Baumadjunktionsgrammatiken [JS97] sind ein Mitglied der Familie der mild kontextsensitiven Grammatiken [Jos85], einer eigenen Klasse formaler Grammatiken, die in der Chomsky-Hierarchie zwischen kontextfreien und kontextsensitiven Grammatiken anzusiedeln sind. Sie haben eine Reihe formaler Eigenschaften, durch die sie für die mobile

[2] Eine formale Definition von SCCFGs gibt die Dissertation in Definition 12.

Intentionserkennung interessant werden, insbesondere polynomiale Parsbarkeit und die Ermöglichung von Überkreuzabhängigkeiten.

Baumadjunktionsgrammatiken unterscheiden sich von kontextfreien dadurch, dass die Bausteine der Grammatik keine Regeln, sondern Elementarbäume sind. Diese können nach bestimmten Regeln ineinander eingesetzt werden. Zentral ist die Regel der Adjunktion, durch die ein Baum B_2 in einen anderen Baum B_1 derart eingeschoben wird, dass der ursprüngliche Baum B_1 in einen linken, mittleren und rechten Kontext zerschnitten wird. Dies kann zu überkreuzenden Beziehungen im entstehenden Baum führen. Eine formale Definition von SCTAGs gibt die Dissertation in Definition 13, ein Beispiel einer SCTAG wird auch in Abschnitt 5 präsentiert (siehe auch Abb. 5).

4.3 Parsen räumlicher Grammatiken

Die Arbeit zeigt, wie herkömmliche mit dynamischer Programmierung arbeitende Chart Parser um eine Auflösung räumlicher Constraints erweitert werden können. Die Laufzeitkomplexität der Parser wird dabei nicht erhöht, sondern im Mittel durch die Reduktion der zu berücksichtigenden Hypothesen verringert. Wichtig zu beachten ist, dass nur solche Chart Parser als Basis für die mobile Intentionserkennung dienen können, die die Valid Prefix Property [JS97, p.50] einhalten, die also inkrementell entscheiden, ob es für eine gegebene unvollständige Inputfolge überhaupt ein mögliches Wort der Sprache geben kann. Die Arbeit stellt beispielhaft Chart Parser für SGISs, SCCFGs und SCTAGs vor, die bestehende Algorithmen aus der Literatur um räumliche Constraints erweitern. Die Komplexität der beiden ersteren ist hierbei $O(n^3)$, die des letzteren $O(n^9)$.

5 Evaluation

Die Ausdrucksmächtigkeit von SCCFGs und SCTAGs wird mit einem Beispielszenario evaluiert, in dem das Return-to-Region Muster offensichtlicherweise auftritt, nämlich dem in Abb. 3 (rechts) und Abb. 4 dargestellten Kleidungsgeschäft. Kunden nehmen eine gewisse Anzahl von Kleidungsstücken aus den Regalen (pick), probieren diese in den Umkleidekabinen an und legen im Anschluss optional einige der Kleidungsstücke wieder zurück in die Regale (drop). An diesem vereinfachten Beispiel eines Einkaufsvorgangs demonstriert die Arbeit mit Hilfe der im vorherigen Kapitel vorgestellten Parsing-Algorithmen, wie SCCFGs und SCTAGs eine Reduktion der Ambiguität im Vergleich zu der bereits existierenden räumlichen Grammatik SGIS erreichen.

Die erlaubten Verhaltensmuster im Kleidungsgeschäft lassen sich als pick^n try drop^m ($m \leq n$) darstellen, wobei die drop-Verhalten nur in Regionen stattfinden können, in denen auch ein pick war. Während SGISs diese Zusatzbedingung nicht prüfen und beim Betreten aller Regal-Regionen stets die Hypothese ‘drop’ aufstellen, bewahren SCCFGs und SCTAGs den räumlichen Kontext der ‘pick’-Verhalten (also die Information, in welchen Regalen etwas genommen wurde), und erzeugen später nur dann eine Hypothese ‘drop’,

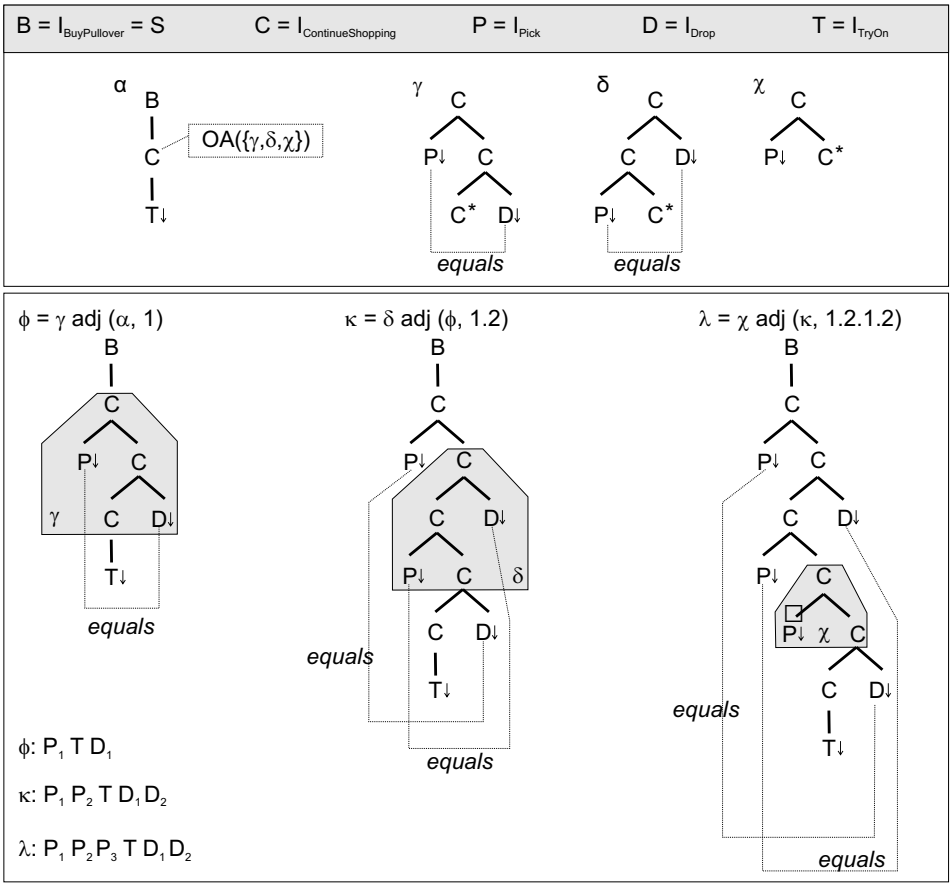


Abbildung 5: Eine SCTAG für überlappende Return-to-Region Abhängigkeiten im Beispiel des Kleidungsgeschäfts. Oben: Grammatik. Unten: Zwei Ableitungsschritte. Vgl. [Kie11a, Abb. 4.13, 4.14].

wenn eine der Regal-Regionen aus der ‘pick’-Phase betreten wird. Daher werden bei SCCFGs und SCTAGs weniger Hypothesen generiert (Desambiguierung). Der Vergleich von SCCFGs und SCTAGs zeigt nun weiterhin, dass das Parsen der SCCFG zu wenige Hypothesen generiert, da die Regionen der drop-Verhalten in genau umgekehrter Reihenfolge zu den Regionen der pick-Verhalten auftreten müssen (‘nesting’). Diese Zusatzbedingungen ist für das Kleidungsgeschäft nicht zutreffend, so dass SCTAGs wegen ihrer Unterstützung von Überkreuzabhängigkeiten in diesem Fall zu empfehlen wären.

Abbildung 5 zeigt das Prinzip, wie eine SCTAG für den Anwendungsfall des Kleidungsgeschäfts aussehen kann: Ein initialer Baum α beinhaltet einen Blattknoten für die Intention *TryOn*. Durch die C-Knoten wird sichergestellt, dass die beiden Möglichkeiten ‘pick und späteres drop’ (Auxiliarbäume γ und δ) ‘pick ohne drop’ (χ) adjungiert und miteinander kombiniert werden können. Zwei beispielhafte Adjunktionen zur Erzeugung einer Überkreuzabhängigkeit sind unter der Grammatik dargestellt.

Das Kapitel 5 der Dissertation stellt außerdem das Desktoptool ‘Intention Simulation Environment’ vor, mit dem verschiedene mögliche Algorithmen für die Ebenen des IAMS Frameworks getestet werden können.

6 Zusammenfassung und Ausblick

Die mobile Intentionserkennung versucht, aus dem Verhalten eines mobilen Nutzers Rückschluss auf seine Intentionen zu ziehen. Dadurch ermöglicht sie eine intelligentere mobile Assistenz als einfache ortsbezogene Dienste. Die Dissertation [Kie11a] beleuchtet die mobile Intentionserkennung als Spezialproblem des aus der künstlichen Intelligenz bekannten allgemeinen Planerkennungsproblems und führt zwei neue Formalismen zur mobilen Intentionserkennung ein. Diese beruhen auf formalen Grammatiken verschiedener Komplexität und erlauben die Interpretation in der mobilen Intentionserkennung häufig vorkommender Verhaltensmuster, die nicht auf einen lokalen Kontext beschränkt sind und überlappend auftreten können.

Der vorgestellte Ansatz wurde bereits erweitert durch die Integration zeitlicher Constraints in formale Grammatiken [KRS10]. Aktuelle Arbeiten des Autors beschäftigen sich damit, wie das durch Eye Tracking Technologien aufgenommene Blickverhalten des Nutzers in die Verhaltensinterpretation einbezogen werden kann [KSR12]. Für weitere offene Fragestellungen, wie beispielsweise probabilistische Erweiterungen, sei auf das letzte Kapitel der Dissertation verwiesen [Kie11a, Kapitel 6].

Literatur

[AS03] Daniel Ashbrook und Thad Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[BDvKS10] Maike Buchin, Anne Driemel, Marc van Kreveld und Vera Sacristan. An Algorithmic Framework for Segmenting Trajectories based on Spatio-Temporal Criteria. In *ACM-GIS 2010, 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seiten 202–211. ACM Press, 2010.

[Bui03] Hung Hai Bui. A general model for online probabilistic plan recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.

[CG93] Eugene Charniak und Robert P. Goldman. A Bayesian model of plan recognition. *Artificial Intelligence*, 64(1):53–79, 1993.

[GS07] Christopher W. Geib und Mark Steedman. On Natural Language Processing and Plan Recognition. In *Proc. of the 20th Int. Joint Conference on Artificial Intelligence*, Seiten 1612–1617, 2007.

[Jos85] A. K. Joshi. Tree Adjoining Grammars: How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions? In D. R. Dowty et al., Hrsg., *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, Seiten 206–250. Cambridge University Press, Cambridge, 1985.

[JRGE98] T. Jordan, M. Raubal, B. Gartrell und M. Egenhofer. An affordance-based model of place in GIS. In *Proc. 8th Int. Symposium on Spatial Data Handling*, Seiten 98–109, Vancouver, 1998. IUG.

[JS97] Aravind K. Joshi und Yves Schabes. Tree-Adjoining Grammars. In G. Rozenberg und A. Salomaa, Hrsg., *Handbook of Formal Languages*, Jgg. 3, Seiten 69–124. Springer, Berlin, New York, 1997.

[Kau87] Henry A. Kautz. *A Formal Theory of Plan Recognition*. Dissertation, University of Rochester, Rochester, NY, 1987.

[Kie11a] Peter Kiefer. *Mobile Intention Recognition*. Springer, New York, 2011. isbn 978-1461418535.

[Kie11b] Peter Kiefer. *The Mobile Intention Recognition Problem And An Approach Based On Spatially-Constrained Grammars*. Dissertation, Fakultät für Wirtschaftsinformatik und Angewandte Informatik, Otto-Friedrich-Universität Bamberg, Germany, 2011.

[KRS10] Peter Kiefer, Martin Raubal und Christoph Schlieder. Time Geography Inverted: Recognizing Intentions in Space and Time. In *ACMGIS 2010, 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seiten 510–513. ACM Press, 2010.

[KSR12] Peter Kiefer, Florian Straub und Martin Raubal. Location-Aware Mobile Eye Tracking for the Explanation of Wayfinding Behavior. In *Proceedings of the AGILE'2012 International Conference on Geographic Information Science*, 2012.

[LPFK07] Lin Liao, Donald J. Patterson, Dieter Fox und Henry Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5-6):311–331, 2007.

[Pyn99] David V. Pynadath. *Probabilistic Grammars for Plan Recognition*. Dissertation, The University of Michigan, 1999.

[Sch05] Christoph Schlieder. Representing the Meaning of Spatial Behavior by Spatially Grounded Intentional Systems. In *GeoSpatial Semantics, First International Conference*, Jgg. 3799 of *Lecture Notes in Computer Science*, Seiten 30–44. Springer, 2005.

[SSG78] C.F. Schmidt, N.S. Sridharan und J.L. Goodson. The Plan Recognition Problem: An Intersection of Psychology and Artificial Intelligence. *Artificial Intelligence*, 11(1-2):45–83, 1978.

Peter Kiefer ist seit Juli 2011 als PostDoc am Lehrstuhl für Geoinformations-Engineering an der ETH Zürich beschäftigt. Er promovierte im Mai 2011 am Lehrstuhl für Angewandte Informatik in den Kultur-, Geschichts- und Geowissenschaften der Otto-Friedrich-Universität Bamberg, wo er als wissenschaftlicher Mitarbeiter seit 2005 in die Lehre und Forschung an der Schnittstelle von semantischer Informationstechnologie und ortsbezogenen Diensten involviert war. Ebenfalls an der Otto-Friedrich-Universität Bamberg erhielt er im Jahr 2005 sein Diplom in Wirtschaftsinformatik. Peter Kiefer ist Autor und Ko-Autor von 30 Buchkapiteln, Konferenzbeiträgen und Artikeln. Seine aktuelle Forschung befasst sich mit Methoden der mobilen Blickerfassung ('Eye Tracking') zur Verbesserung mobiler ortsbezogener Dienste durch Raumkognitionsstudien und aufmerksamkeitsabhängige Assistenzsysteme.

Maschinelles Lernen mit multiplen Kernen

Marius Kloft

Abteilung Maschinelles Lernen, Technische Universität Berlin
kloft@tu-berlin.de

Abstract: Diese Arbeit gibt zunächst eine grundlegende Einführung in Theorie und Praxis des Maschinellen Lernens mit multiplen Kernen und skizziert den Stand der Forschung. Weiter entwickelt die Arbeit eine neue Methodologie des Lernens mit mehreren Kernen und beweist deren Effizienz und Effektivität. Sie entwickelt Algorithmen zur Optimierung des assoziierten mathematischen Programmes, die im Vergleich zu vorherigen Ansätzen um bis zu zwei Größenordnungen schneller sind. Unsere theoretische Analyse des Generalisierungsfehlers zeigt dabei Konvergenzraten mit Ordnungen von maximal $O(M/n)$, frühere Analysen präzisierend, die bisher nur $O(\sqrt{M/n})$ erreichten. In Anwendungen auf zentrale Fragestellungen der Bioinformatik und des Maschinellen Sehens werden Vorhersagegenauigkeiten erreicht, die den bisherigen Stand der Forschung signifikant übertreffen, wodurch eine Grundlage zur Erschließung neuer Anwendungsfelder und Forschungsansätze geschaffen wird.

1 Einführung

Ziel des Maschinellen Lernens ist das Erlernen des unbekannten Zusammenhanges zweier Variablen X und Y aus Daten $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$, um bei Beobachtung eines neuen Musters x eine möglichst präzise Vorhersage für dessen unbekanntes Konzept y abgeben zu können. Einen besonders eleganten Lösungsansatz hierfür bieten nicht-lineare, „kernbasierte“ Lernverfahren [MMR$^+$01]: Mit der Substitution aller Skalarprodukte $\langle \phi(x), \phi(\widetilde{x}) \rangle$ durch eine Nicht-Linearität $k(x, \widetilde{x})$ – dem sogenannten *Kern* – werden die Muster *implizit* in einen hoch-dimensionalen Merkmalsraum eingebettet, in welchem sie linear getrennt werden können. So erzeugen wir auf systematische Art und Weise aus einfacheren Lernmaschinen sehr viel komplexere und leistungsstärkere – im Merkmalsraum lineare – was den Lernschritt von der Datenrepräsentation modular entkoppelt. Aufgrund ihrer Ausdruckskraft und Leistungsstärke – bei gleichzeitig sehr geringer Lauf- und Ausführungszeit – stellen kernbasierte Lernverfahren in Anwendungsbereichen mit komplexen Problemstellungen und großen Datenmengen, wie beispielsweise der Bioinformatik und dem Maschinellen Sehen, den gegenwärtigen Standard dar.

Klassische kernbasierte Lernverfahren verwenden einen *einzelnen* Kern, der in der Regel aus einer im Vorfeld zu spezifizierenden Menge von Kandidatenkernen $\mathcal{K} = \{k_1, \ldots, k_M\}$ durch Kreuzvalidierung ausgewählt wird. Problematischerweise können die Kerne jedoch *komplementäre* Eigenschaften des Lernproblems charakterisieren: Zum Beispiel treten im Maschinellen Sehen eine Vielzahl von gegensätzlichen

Informationselementen auf, basierend auf der Farbverteilung eines Bildes oder den auftretenden Formen und Kanten et cetera. Nur einen einzelnen Kern, z. B. den „Farbkern“, auszuwählen, bedeutet daher zugleich auch wertvolle, komplementäre Information zu verwerfen! Beispielsweise mag Farbinformation hilfreich sein zur Erkennung von Stoppschildern, aber weniger hilfreich zur Annotation von Bildern, die Autos oder Luftballons enthalten.

Alle in dieser Dissertation entwickelten Methoden basieren daher auf einer optimierten Gewichtung mehrerer Kerne, *um die darin enthaltende Information zu fusionieren*:

$$k = \theta_1 k_1 + \cdots + \theta_M k_M \,.$$

Bis auf sehr kleine M (typischerweise $M \leq 3$) ist der Suchraum der θ_i allerdings zu groß für gewöhnliche Suchverfahren. Eine grundlegende Einsicht an dieser Stelle ist, dass viele Methoden des Maschinellen Lernens durch mathematische Programme definiert sind, wie beispielsweise die Support-Vektor-Maschine:

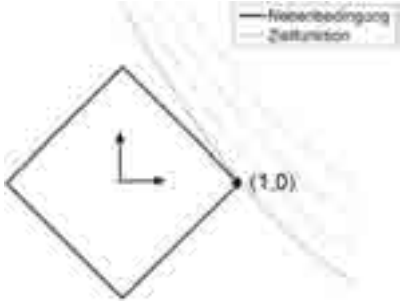
$$\begin{aligned} &\mathrm{SVM}(k, \mathcal{D}) \\ &:= \min_{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n} \max\Big(0, 1 - y_i\big(\langle \boldsymbol{w}, \phi(x_i)\rangle - b\big)\Big) \qquad (1) \\ &= \max_{\boldsymbol{\alpha}\in\mathbb{R}^n : \mathbf{0}\leq\boldsymbol{\alpha}\leq C, \boldsymbol{y}^\top\boldsymbol{\alpha}=0} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i, x_j)\,. \end{aligned}$$

Können wir die Parameter $\theta_1, \ldots \theta_M$ ebenfalls als Variablen in ein solches Optimierungsproblem mit aufnehmen?

In der klassischen Arbeit [LCG+04] wird der Suchraum auf positive Gewichte $\theta_m \geq 0$ eingeschränkt, die sich zu eins aufsummieren,

$$\sum_{m=1}^{M} \theta_m = 1\,, \qquad (2)$$

weil dies sonst zu trivialen Lösungen $\theta_1 = \cdots = \theta_M = \infty$ führen würde. Die Nebenbedingung (2) führt zu *dünn-besetzten* (oder „spärlichen“) Gewichten $\boldsymbol{\theta}$, wie man anhand der nebenstehenden Abbildung erkennen kann: In dem Optimum des mathematischen Programms (1) berührt eine der Höhenlinien der Zielfunktion die Nebenbedingung (2) in einer ihrer (dünn besetzten) Ecken.



Dünn besetzte Kerngewichte sind zwar leicht interpretierbar und können auch numerisch von Vorteil sein, jedoch erlauben wir uns, an dieser Stelle zu betonen, dass die Fokussierung auf dünn besetzte Kernmischungen zum Teil eher einer generellen Präferenz der gegenwärtigen Forschung für sogenannte *sparse* Methoden geschuldet ist. Dabei kann die Beschränkung auf dünn-besetzte Gewichte bei der *Fusion* von multiplen Kernen äußerst unlogisch und geradezu kontraintuitiv sein, insbesondere wenn die Kerne *komplementäre*

Eigenschaften des Lernproblems codieren. Dies ist beispielsweise in den Bereichen der Bioinformatik und des Maschinellen Sehens der Fall, wo – wie oben erwähnt – Farb-, Form- und Kanteninformationen synergetisch wirken und die bisher übliche Selektion die potenzielle Leistungsfähigkeit massiv einschränkt.

In der vorliegenden Dissertation positionieren wir uns gegen diesen vorherrschenden Trend und zeigen, dass ausgewogene, nicht-spärliche Kernkombinationen weit höhere Vorhersagegenauigkeiten ermöglichen können als sparse Methoden. Weiterhin beweisen wir theoretische Schranken, die weit präzisere Konvergenzraten aufweisen als bisher existierende. So lässt sich nun erklären, *warum* nicht-spärliche Kernmischungen oftmals effektiver sind. In numerischer Hinsicht leiten wir Algorithmen her, die schneller sind als die bisherigen und es erlauben, auch gewaltige Datenmengen, wie sie etwa in der Bioinformatik auftreten können, zu verarbeiten.

Die Hauptbeiträge der Arbeit [Klo11] können danach wie folgt zusammengefasst werden:

- Wir entwickeln eine neue *Methodologie* des Lernens mit multiplen Kernen, die zu nicht-spärlichen Kernkombinationen führt – effizienter und effektiver als vorherige Ansätze.
- Zur Lösung des mit der Methodologie assoziierten mathematischen Programms leiten wir Algorithmen her, die gleichzeitig Zehntausende von Trainingsbeispielen und Tausende von Kernen verarbeiten können, und beweisen deren Konvergenz, welche um bis zu zwei Größenordnungen schneller erfolgt als jene der überkommenen Algorithmen.
- *Theoretische* Analysen des Generalisierungsfehlers zeigen Konvergenzraten einer Ordnung von maximal $O(M/n)$ – was alle früheren Analysen präzisiert, die nur $O(\sqrt{M/n})$ erreichten. Auf Grundlage der theoretischen Schranken können wir erklären, *warum* nicht-spärliche Kerngewichte oftmals effektiver sind.
- In *Anwendungen* auf zentrale Fragestellungen der Bioinformatik und des Maschinellen Sehens werden Vorhersagegenauigkeiten erreicht, die den bisherigen Stand der Forschung weit übertreffen.

Im Folgenden gehen wir auf die Hauptergebnisse der Dissertation [Klo11] ausführlicher ein.

2 Lernen nicht-spärlicher Kernkombinationen

In der vorliegenden Dissertation verwerfen wir die Beschränkung auf dünn-besetzte Kerngewichte und definieren das Lernen mit multiplen Kernen einschränkungslos durch ein rigoroses, mathematisches Optimierungskriterium unter Verwendung völlig beliebiger Normen $\|\cdot\|_O$ und Lossfunktionen l [KRB10]:

$$\begin{aligned} \inf_{\boldsymbol{w},b,\boldsymbol{t}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_{2,O}^2 + C\sum_{i=1}^{n} l\big(t_i,\ y_i\big) \\ \text{s.t.} \quad & \forall i:\ \langle \boldsymbol{w}, \phi(x_i)\rangle + b = t_i\ . \end{aligned} \tag{P}$$

Diese das Feld vereinheitlichende Formulierung enthält alle existierenden Ansätze zum Lernen mit multiplen Kernen als Spezialfälle, die nun *gemeinsam* analysiert werden können. Beispielsweise leiten wir eine völlig allgemeine duale Repräsentation mit Hilfe einer eigens zu dem Zweck in [Klo11] von uns entwickelten Dualitätsmethode her:

$$\sup_{\boldsymbol{\alpha}:\, \mathbf{1}^\top \boldsymbol{\alpha}=0} -C \sum_{i=1}^{n} l^*\Big(-\frac{\alpha_i}{C}, y_i\Big) - \frac{1}{2}\Big\| \sum_{i=1}^{n} \alpha_i \phi(x_i)\Big\|^2_{2,O^*} . \tag{D}$$

Der Einfachheit halber konzentrieren wir uns in der hier angefertigten Zusammenfassung auf Minkowski ℓ_p-Normen $\|\boldsymbol{\theta}\|_p := \big(\sum_{m=1}^{M} |\theta_m|^p\big)^{1/p}$ und die SVM-Verlustfunktion $l(t) := \max(0, 1-t)$, d. h. wir betrachten das folgende mathematische Programm:

$$\min_{\boldsymbol{\theta}\geq \mathbf{0},\, \|\boldsymbol{\theta}\|_p=1} \ \mathrm{SVM}\Big(\sum_{m=1}^{M} \theta_m k_m, \mathcal{D}\Big) \tag{3}$$

3 Algorithmen

Im Rahmen der Dissertation werden drei effiziente Algorithmen zur Optimierung von (3) vorgestellt [KBLS08, KBS+09, KBSZ11]:

- ein Newton-Verfahren
- ein Block-Koordinaten-Abstiegs-Algorithmus
- ein Cutting-Plane-Algorithmus, basierend auf sequentieller, quadratisch-bedingter, quadratischer Programmierung mit Höhenlinien-Projektionen.

Jeder dieser Algorithmen existiert in zwei Varianten: als einfacher, zweischrittiger Algorithmus sowie als effizienter, fest in die SVM integrierter Algorithmus. In dieser Zusammenfassung konzentrieren wir uns darauf, den Block-Koordinaten-Abstiegs-Algorithmus darzustellen. Er basiert auf einer einfachen, analytischen Optimalitäts-Formel, die innerhalb von Mikrosekunden ausgewertet werden kann:

$$\forall\, m = 1, \ldots, M: \qquad \theta_m = \frac{\|\boldsymbol{w}_m\|_2^{2-p}}{\Big(\sum_{m'=1}^{M} \|\boldsymbol{w}_{m'}\|_2^p\Big)^{(2-p)/p}} . \tag{4}$$

In der einfacheren, modularen Version werden alternierend die Gleichungen (1) und (4) gelöst, so dass dieser Wrapper-Algorithmus sogar einfacher als SimpleMKL [RBCG08] ist, welches in jeder Iteration eine heuristische Line Search ausführt. In der zweiten, in Algorithmen-Tafel 1 beschriebenen Version ist das Mehr-Kern-Modul fest in die SVM eingebettet, um maximale Effizienz zu erreichen.

Alle Algorithmen sind in C++ programmiert und in die Shogun Machine Learning Toolbox [SRH+10] integriert worden, welche Schnittstellen zu MATLAB, Octave, Python und R beinhaltet. Die Konvergenz beider Algorithmen wird durch das folgende Theorem, dessen Beweis in Abschnitt 3.2.1 in [Klo11] geführt wird, sichergestellt:

Algorithm 1 In die SVM integrierter analytischer Block-Koordinaten-Algorithmus.

input: $p \in [1,\infty]\backslash\{2\}, \quad Q \in \mathbb{N}, \quad \epsilon > 0$
initialize: $\forall\, i,m:\ g_{m,i} = \hat{g}_i = \alpha_i = 0; \quad L = S = -\infty; \quad \theta_m := (1/M)^{(2-p)/p}$
iterate
  Select l variables $\alpha_{i_1},\dots,\alpha_{i_l}$ based on the gradient $\hat{\mathbf{g}}$ of SVM
  Store $\boldsymbol{\alpha}^{\text{OLD}} = \boldsymbol{\alpha}$ and then compute $\boldsymbol{\alpha} := \arg(\text{SVM}(K_{\boldsymbol{\theta}}))$ w.r.t. the selected variables
  Update gradient $\forall\, i,m:\ g_{m,i} := g_{m,i} + \sum_{q=1}^{Q}(\alpha_{i_q} - \alpha_{i_q}^{\text{OLD}})k_m(x_{i_q}, x_i)$
  Compute the quadratic terms $\forall m:\ S_m := \frac{1}{2}\sum_i g_{m,i}\alpha_i, \ \ \|\boldsymbol{w}_m\|_2^2 := 2\theta_m^2 S_m$
  $L_{\text{OLD}} = L, \quad L = \sum_i y_i\alpha_i, \quad S_{\text{OLD}} = S, \quad S = \sum_m \theta_m S_m$
  if $|1 - (L-S)/(L_{\text{OLD}} - S_{\text{OLD}})| \geq \varepsilon$
    Update $\boldsymbol{\theta}$ according to Eq. (4)
    if $p \in [1,2]$
      For all m compute $\|\boldsymbol{w}_m\|^2 := \theta_m^2 \boldsymbol{\alpha} K_m \boldsymbol{\alpha}$
    end if
  else
    break
  end if
  $\hat{g}_i = \sum_m \theta_m g_{m,i}$ for all $i = 1,\dots,n$
output: $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ as sparse vectors

Theorem 1. *Seien $K_1,\dots,K_M$ strikt positiv-definite Kernmatrizen. Dann ist jeder Häufungswert der Sequenz der von Algorithmus 1 zurückgegebenen Lösungen ein globaler, optimaler Punkt des mathematischen Programmes* (3).

Anhand der nebenstehend dargestellten Ergebnisse unserer Laufzeituntersuchungen erkennen wir, dass unsere Algorithmen – erstmals! – die effektive Verwendung von Zehntausenden von Datenpunkten und Tausenden von Kernen erlauben. Wie in Abbildung 1 dargestellt, erweisen sie sich um bis zu *zwei Größenordnungen* schneller als die State-of-the-Art Algorithmen SimpleMKL [RBCG08] und HessianMKL [CR08]. Während letztere bei ca. 10 000 Trainingsbeispielen und 1 000 Kernen „out of memory“ meldeten, kann unser Algorithmus durch on-the-fly-Berechnung von Kernen auch für größere Trainingsmengen eingesetzt werden.

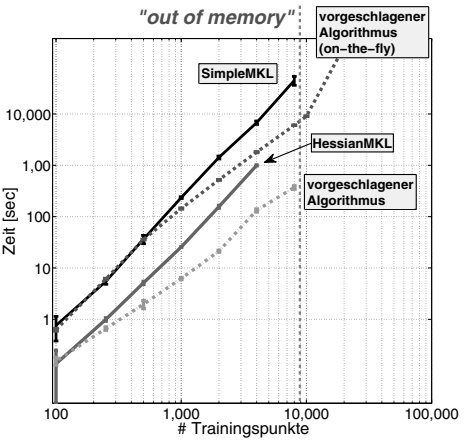


Abbildung 1: Laufzeit von Algorithmus 1 im Vergleich zu State-of-the-Art Verfahren.

4 Theoretische Analyse

Die vorgeschlagene Methodologie ist durch fundamentale Garantien der statistischen Lerntheorie untermauert [KB11, KB12] – wir zeigen die folgende obere Schranke auf die lokale Rademacher-Komplexität [BBM05] des Lernens mit multiplen Kernen:

Theorem 2 (Obere Rademacher-Schranke)**.** *Die lokale Rademacher-Komplexität des Lernens mit multiplen Kernen kann durch die folgende Schranke abgeschätzt*

werden:

$$R_r(H_p) \leq \min_{t\in[p,2]} \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rM^{1-\frac{2}{t^*}}, ceC^2t^{*2}\lambda_j^{(m)}\right)\right)_{m=1}^{M}\right\|_{\frac{t^*}{2}}} + \frac{\sqrt{Be}CM^{\frac{1}{t^*}}t^*}{n},$$

wobei $\lambda_j^{(m)}$ *den j-ten Eigenwert des m-ten Kerns (in absteigender Reihenfolge sortiert),* $t^* := \frac{t}{t-1}$ *den zu t konjugierten Exponenten und* $B^2 := \sup_x k(x,x)$ *bezeichnet.*

Beweisskizze. An dieser Stelle fassen wir die Schlüsselschritte des Beweises von Theorem 2 zusammen. Der vollständige Beweis ist auf den Seiten 51–59 in [Klo11] geführt.

1. Bestimmung der Komplexität der Originalklasse durch die zentrierte Klasse:
$$R_r(H_p) \leq R_r(\widetilde{H}_p) + n^{-\frac{1}{2}} \min\left(\sqrt{r}, C\sqrt{\left\|\left(\operatorname{tr}(J_m)\right)_{m=1}^{M}\right\|_{\frac{p^*}{2}}}\right)$$
2. Abschätzung der Komplexität der zentrierten Klasse:
$$R_r(\widetilde{H}_p) \leq \sqrt{\frac{r\sum_{m=1}^{M}h_m}{n}} + C\,E\left\|\left(\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle\widetilde{\boldsymbol{u}}_j^{(m)}\right)_{m=1}^{M}\right\|_{2,p^*}$$
3. Verwendung der Ungleichungen von Khintchine-Kahane und Rosenthal:
$$E\left\|\left(\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle\widetilde{\boldsymbol{u}}_j^{(m)}\right)_{m=1}^{M}\right\|_{2,p^*} \leq \sqrt{\frac{p^*}{n}}\,E\left(\sum_{m=1}^{M}\left(\sum_{j=h_m+1}^{\infty}\frac{1}{n}\sum_{i=1}^{n}\langle\widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle^2\right)^{\frac{p^*}{2}}\right)^{\frac{1}{p^*}}$$
4. Abschätzung der Komplexität der Originalklassse:
$$R_r(H_p) \leq \sqrt{\frac{4r\sum_{m=1}^{M}h_m}{n}} + \sqrt{\frac{4ep^{*2}C^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^{M}\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}CM^{\frac{1}{p^*}}p^*}{n}$$
5. Charakterisierung bezüglich der Trunkierung der Spektren der Kerne:
$$R_r(H_p) \leq \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rM^{1-\frac{2}{p^*}}, ep^{*2}C^2\lambda_j^{(m)}\right)\right)_{m=1}^{M}\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}CM^{\frac{1}{p^*}}p^*}{n} \qquad \square$$

Sind die oberen Schranken präzise oder möglicherweise verbesserbar? Diesbezüglich zeigen wir eine *untere* Schranke, deren Konvergenzrate mit jener der oberen Schranke übereinstimmt. Wir können daher zu dem Schluss kommen, dass die erzielten Raten nicht verbesserbar und theoretisch-optimal sind:

Theorem 3 (Untere Rademacher-Schranke)**.** *Seien die Kerne zentriert und unabhängig, identisch verteilt und* $c > 0$ *eine Konstante mit* $\lambda^{(1)} \geq \frac{1}{nD^2}$*. Dann gilt für alle* $r \geq \frac{1}{n}$ *und* $p \geq 1$*:*

$$R_r(H_p) \;\geq\; \sqrt{\frac{c}{n}\sum_{j=1}^{\infty}\min\left(rM, D^2M^{2/p^*}\lambda_j^{(1)}\right)}. \tag{5}$$

Da die Generalisierungsfähigkeit einer Lernmaschine durch die lokale Rademacher-Komplexität genauestens charakterisiert ist [BBM05], folgt aus Theorem 2 die folgende Generalisierungsschranke für ℓ_p-Norm MKL:

Theorem 4 (Theoretische Garantie)**.** *Angenommen* $\|k\|_\infty \leq B$ *und* $\exists d > 0$, $\alpha := \alpha > 1$, *so dass* $\forall m : \lambda_j^{(m)} \leq d_{\max} j^{-\alpha}$. *Dann gilt: Der Verlust des Lernens mit multiplen Kernen ist für jedes* $p \in [1, \ldots, 2]$ *und* $z > 0$ *mit Wahrscheinlichkeit größer gleich* $1 - e^{-z}$ *beschränkt durch*

$$P(l_{\hat{f}} - l_{f^*}) \leq \min_{t \in [p,2]} \; 186 \sqrt{\frac{3-\alpha_m}{1-\alpha_m}} \left(d_{\max} D^2 L^2 t^{*2}\right)^{\frac{1}{1+\alpha}} F^{\frac{\alpha-1}{\alpha+1}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}} + \frac{47\sqrt{B} D L M^{\frac{1}{t^*}} t^*}{n} + \frac{(22 B D L M^{\frac{1}{t^*}} + 27F)z}{n}.$$

Wir beobachten, dass durch die obige Schranke Konvergenzraten von maximal $O(M/n)$ erzielt, was eine wesentliche Verbesserung der genauesten aus der Literatur [CMR10] bekannten Rate von $O\big(\sqrt{M/n}\big)$ darstellt, denn typischerweise wird M als konstant angesehen und $n \to \infty$. Bei der typischen Anzahl von $M = 10$ Kernen und $n = 100\,000$ Trainingspunkten enthält die Schranke von [CMR10] einen Faktor von $\sqrt{M/n} = 1/100$, während unser Resultat $M/n = 1/10\,000$ erzielt – ein Unterschied von zwei Größenordnungen.

5 Anwendungen in der Bioinformatik und dem Maschinellen Sehen

Die entwickelte Methodologie wird in [Klo11] auf aktuelle Fragestellungen der Bioinformatik und des Maschinellen Sehens angewendet. Diese Bereiche der Informatik zeichnen sich durch das Auftreten vielfältiger, komplementärer Sichtweisen auf die Daten aus, was die Verwendung multipler Kerne sinnvoll macht. Sämtliche frühere Analysen – mit Ausnahme der Arbeit von [ZO07] zu subzellularer Lokalisierung von Proteinen – scheiterten darin, die Effektivität des Lernens mit multiplen Kernen nachzuweisen, z. B. [SZR06] in der Bioinformatik und [GN09] im Maschinellen Sehen. In dieser Dissertation wird gezeigt, dass unter Verwendung der neuen nichtspärlichen Methodologie die Vorhersagegenauigkeit in beiden Bereichen signifikant erhöht wird.

Visuelle Objekterkennung Dieser Bereich des Maschinellen Sehens beschäftigt sich mit dem Erkennen von Objekten in Bildern – ein schwieriges Unterfangen, denn Objekte können rotiert, verschoben, beleuchtet oder auch von anderen Objekten partiell verdeckt sein. Weiterhin können gewisse Merkmale relevant zum Erkennen einiger, aber wirkungslos zum Erkennen anderer Objekte sein. Beispielsweise ist Farbinformation hilfreich zur Erkennung von Stoppschildern, aber von wenig Nutzen bei der Erkennung von Autos oder Luftballons. Obwohl sich daher ein Einsatz mehrerer Kerne anbietet, zeigten frühere Analysen keinen Vorteil von klassischem Mehr-Kern-Verfahren (siehe z. B. [GN09]).

In [Klo11] analysieren wir den offiziellen, aus 8780 Bildern und 20 Objektklassen bestehenden, Datensatz der PASCAL VOC Challenge 2008. Wir verwenden multiple

Kerne basierend auf Histogrammen gerichteter Gradienten, visueller Wörter, Pixelfarben (letzteres in zwei Farbkanälen) und verschiedenen Pyramidenebenen. Dies resultiert in insgesamt 12 Kernen. Als Gütekriterium verwenden wir das offizielle Fehlermaß des VOC 2008 Wettbewerbs („durchschnittliche Präzision“) sowie den offiziellen Testdatensatz. Die Ergebnisse sind in Abbildung 2 dargestellt: Vertikale Balken geben den Unterschied der durchschnittlichen Präzisionen des Mehr-Kern-Lernens zu einer uniformen SVM an. Wir erkennen, dass unsere neue Methodologie in 18 der 20 Klassen vorteilhaft ist, während das klassische, dünn-besetzte Verfahren keine konsistente Verbesserung ergibt.

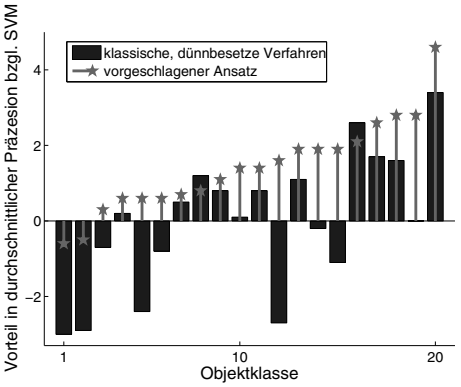


Abbildung 2: Empirische Ergebnisse bei der visuellen Objekterkennung.

Lokalisation von Genen in DNA Die Entdeckung von Transkriptionsstarts (TSS) von RNA Polymerase II bindenden Genen in genomischer DNA stellt einen kritischen Schritt zur Dechiffrierung von Transkription regulierenden Elementen dar. Demzufolge wurde in diesem aktiven Bereich der Bioinformatik eine große Anzahl von Lösungsansätzen vorgestellt. In der unabhängigen Studie [AdPS09] wurden 19 solcher State-of-the-Art Programme verglichen und das Programm ARTS von [SZR06] als genauestes Programm identifiziert. Wir zeigen, dass durch die Verwendung der vorschlagenen Methodologie die Vorhersagegenauigkeit weiter gesteigert wird - über jene des bisher besten Programms [SZR06] hinaus.

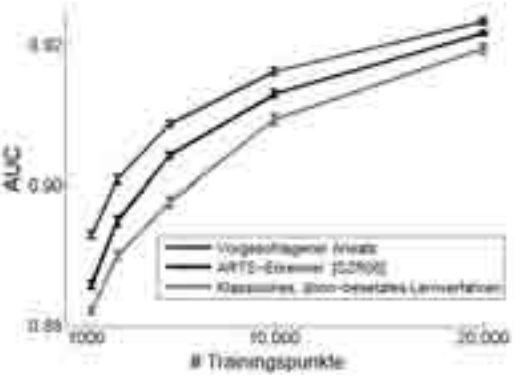


Abbildung 3: Empirische Ergebnisse in der Erkennung von Transkriptionsstarts.

Wie [SZR06] verwenden wir fünf verschiedene Kerne, die komplementäre Eigenschaften des Problems charakterisieren: das TSS Signal, die Promoter-Region, das 1. Exon, die bindende Energie und die Krümmung der DNA. In Übereinstimmung mit [SZR06] setzen wir die Fläche unter der ROC-Kurve (AUC) als Gütekriterium ein und experimentieren auf Grundlage der von [SZR06] bereitgestellten Datensätze. Die Ergebnisse der Analyse sind in Abbildung 3 dargestellt. Vertikale Balken stellen hierbei statistische Standardfehler dar. Wir beobachten, dass das klassische, dünn-besetzte Lernen dem Programm ARTS in Bezug auf alle Trainingsdatengrößen unterliegt – ARTS wiederum wird von der vorgeschlagenen, nicht-spärlichen Methodologie noch einmal deutlich übertroffen.

6 Zusammenfassung

Wir entwickelten eine neue Methodologie zum Lernen mit mehreren Nicht-Linearitäten (oder „Kernen“), einem hochaktuellen und bisher ungelösten Forschungsproblem des Maschinellen Lernens, die – im Gegensatz zu früheren Ansätzen – *keine* dünn besetzten Lösungen liefert. Unsere empirische Evaluierung zu herausfordernden Problemen aus den Bereichen Bioinformatik und Maschinelles Sehen zeigte, dass Vorhersagegenauigkeiten erreicht werden konnten, die den bisherigen Stand der Forschung weit übertreffen. Die vorgeschlagenen Algorithmen zur Optimierung erwiesen sich um bis zu zwei Größenordnungen schneller als existierende und erlaubten, zugleich Zehntausende von Trainingsbeispielen und Tausende von Kernen zu verarbeiten. Die entwickelten Techniken sind grundlegend untermauert durch die statistische Lerntheorie: Wir bewiesen Generalisierungsschranken der Ordnung $O(M/n)$, die weit höhere Konvergenzgeschwindigkeiten aufweisen als vorherige Schranken, welche bestenfalls $O(\sqrt{M/n})$ erzielten.[1]

Schließlich erlauben wir uns, zu bemerken, dass die aktuelle starke Präferenz von dünn besetzten Lernverfahren im Bereich Maschinelles Lernen – oder gar in den Wissenschaften im Allgemeinen – überdacht werden sollte und sich dem hier vorgeschlagenen Ansatz folgend bedeutende neue Perspektiven erschließen lassen. So kann bereits schwache Konnektivität in kausalen, grafischen Modellen dazu führen, dass *sämtliche* Variablen im optimalen Vorhersagemodell aktiv sind. Beispielsweise argumentiert Gelman [Gel11] in den Sozialwissenschaften: „Faktoren sind (fast) nie wirklich Null.“ Basierend auf nicht-spärlichen, multiplen kernbasierten Lernverfahren wurde durch die vorliegende Arbeit eine neue technologische Grundlage zur Fusion von Information geschaffen. Meine zukünftige Forschung wird sich auf deren Verwendung in bioinformatischen und technologischen Anwendungsbereichen konzentrieren, insbesondere in Bezug auf die Fragestellung abhängiger Datenströme.

Danksagung Mein herzlicher Dank gilt meinem Doktorvater Prof. Dr. Klaus-Robert Müller und meinen Mentoren Prof. Peter L. Bartlett, PhD und Prof. Dr. Gilles Blanchard sowie den Mitarbeiterinnen und Mitarbeitern der Abteilungen Maschinelles Lernen der TU Berlin und der UC Berkeley.

Literatur

[AdPS09] T. Abeel, Y. Van de Peer und Y. Saeys. Towards a gold standard for promoter prediction evaluation. *Bioinformatics*, 25(12):313–320, 2009.

[BBM05] P. L. Bartlett, O. Bousquet und S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[CMR10] C. Cortes, M. Mohri und A. Rostamizadeh. Generalization Bounds for Learning Kernels. In *Proceedings, 27th ICML*, Seiten 247–254, 2010.

[CR08] O. Chapelle und A. Rakotomamonjy. Second Order Optimization of Kernel Parameters. In *Proc. of the NIPS Workshop on Kernel Learning*, 2008.

[1] Weitere Beiträge aus [Klo11] wurden aus Platzgründen in dieser Zusammenfassung ausgespart: z. B. die Anwendung der vorgeschlagenen Methodologie in anderen Bereichen der Informatik, wie beispielsweise der Netzwerk Sicherheit [KBD+08, KNB09].

[Gel11] A. Gelman. Causality and Statistical Learning. *American Journal of Sociology*, 117(3):955–966, 2011.

[GN09] P. V. Gehler und S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, Seiten 221–228, 2009.

[KB11] M. Kloft und G. Blanchard. The Local Rademacher Complexity of Lp-Norm Multiple Kernel Learning. In *NIPS 2011, in press*, 2011.

[KB12] M. Kloft und G. Blanchard. On the convergence rate of multiple kernel learning. *Journal of Machine Learning Research (JMLR), accepted*, 2012.

[KBD+08] M. Kloft, U. Brefeld, P. Düssel, C. Gehl und P. Laskov. Automatic feature selection for anomaly detection. In *AISec*, Seiten 71–76. ACM, 2008.

[KBLS08] M. Kloft, U. Brefeld, P. Laskov und S. Sonnenburg. Non-Sparse Multiple Kernel Learning. In *Proc. NIPS Workshop on Kernel Learning*, 2008.

[KBS+09] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller und A. Zien. Efficient and Accurate Lp-Norm Multiple Kernel Learning. In *Advances in Neural Information Processing Systems 22*, Seiten 997–1005. MIT Press, 2009.

[KBSZ11] M. Kloft, U. Brefeld, S. Sonnenburg und A. Zien. Lp-norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.

[Klo11] M. Kloft. *ℓ_p-Norm Multiple Kernel Learning*. Dissertation, Technische Universität Berlin, Oct 2011.

[KNB09] M. Kloft, S. Nakajima und U. Brefeld. Feature Selection for Density Level-Sets. In *ECML/PKDD*, Seiten 692–704, 2009.

[KRB10] M. Kloft, U. Rückert und P. L. Bartlett. A Unifying View of Multiple Kernel Learning. In *ECML/PKDD*, Seiten 66–81, 2010.

[LCG+04] G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett und M. I. Jordan. Learning the kernel with semi-definite programming. *JMLR*, 5:27–72, 2004.

[MMR+01] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda und B. Schölkopf. An Introduction to Kernel-based Learning Algorithms. *IEEE N. Netw.*, 12(2):181–201, 2001.

[RBCG08] A. Rakotomamonjy, F. Bach, S. Canu und Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[SRH+10] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl und V. Franc. The SHOGUN Machine Learning Toolbox. *Journal of Machine Learning Research*, Seiten 1799–1802, 2010.

[SZR06] S. Sonnenburg, A. Zien und G. Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):472–480, 2006.

[ZO07] A. Zien und C. S. Ong. Multiclass multiple kernel learning. In *ICML*, Seiten 1191–1198. ACM, 2007.

Marius Kloft, geboren 1980, studierte von 2000 bis 2006 Mathematik, Physik und Informatik an der Philipps-Universität Marburg. Nach dem Diplom in Mathematik (2006) verfasste er zwischen 2007 und 2011 seine Dissertation an der Technischen Universität Berlin, dem Fraunhofer Institut FIRST und der University of California at Berkeley. Er hat Forschungsaufenthalte an dem Friedrich-Miescher-Laboratorium der Max-Planck Gesellschaft (Tübingen) und der Universität Tromso (Norwegen) verbracht und ist zur Zeit im Rahmen eines Projektes zur computergestützten Genomanalyse in Kooperation mit Laboratorien in Tübingen und New York an der Technischen Universität Berlin tätig.

Formalsprachliche Theorie der Haarnadelstrukturen

Steffen Kopecki

Department of Computer Science
The University of Western Ontraio, London, Canada
steffen@csd.uwo.ca

Abstract: Die (berenzte) Haarnadel-Vervollständigung und die Haarnadel-Verlängerung sind Operationen auf formalen Sprachen, welche die Modifikation von DNA Strängen durch Bildung von Haarnadelstrukturen während der Polymerase-Kettenreaktion modellieren. In dieser Arbeit befassen wir uns mit der formalsprachlichen Analyse dieser Operationen. Neben der Untersuchung der Abschlusseigenschaften von Sprachklassen unter den Operationen, beschäftigt sich die Arbeit mit der Lösung von Entscheidungsproblemen, die durch Haarnadel-Operationen gegeben sind.

1 Einführung

Die Haarnadelstruktur ist eine intramolekulare Basenpaarung, die in einsträngiger DNA oder RNA auftreten kann. Inspiriert durch dieses Phänomen wurden Operationen, wie die Haarnadel-Vervollständigung und Haarnadel-Verlängerung, auf formalen Sprachen definiert. Diese Arbeit beschäftigt sich ausschließlich mit der formalsprachlichen Untersuchung der Haarnadelstrukturen. Zunächst werden wir jedoch den biochemischen Ursprung der Haarnadelstruktur darlegen.

1.1 Haarnadelstrukturen in der Biochemie

Einsträngige DNA, im folgenden *DNA Strang* genannt, ist ein Polymer, bestehend aus Nukleotiden, welche sich durch ihre Nukleobasen A (Adenin), T (Thymin), G (Guanin) und C (Cytosin) unterscheiden. Jeder DNA Strang besitzt ein $5'$- und ein $3'$-Ende, welche nach der chemischen Struktur der Nukleotide benannt sind. Üblicherweise werden DNA Stränge in $5'$-nach-$3'$-Orientierung notiert. Ein DNA Strang kann abstrakt als ein Wort über dem vier-Buchstaben Alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ gesehen werden. Die Basen A und T, bzw. C und G, sind zueinander *Watson-Crick-komplementär*. Zwei DNA Stränge mit unterschiedlicher Orientierung können sich aneinander anlagern, wenn ihre Basen paarweise komplementär sind und können so einen *DNA Doppelstrang* bilden, die wohlbekannte Doppelhelix. In Abb. 1 ist ein Beispiel gegeben.

Für das Watson-Crick-Komplement und sein formalsprachliches Pendant verwenden wir die $\overline{\ }$ Notation, d. h. $\overline{\mathsf{A}} = \mathsf{T}$, $\overline{\mathsf{T}} = \mathsf{A}$, $\overline{\mathsf{C}} = \mathsf{G}$ und $\overline{\mathsf{G}} = \mathsf{C}$. Diese Notation erweitern wir

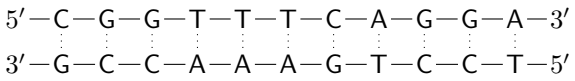


Abbildung 1: Anlagerung zweier komplementärer DNA Stränge

auf DNA Stränge in $5'$-nach-$3'$-Orientierung (bzw. Wörter) durch $\overline{a_1 \cdots a_n} = \overline{a_n} \cdots \overline{a_1}$, wobei $a_1, \ldots, a_n$ einzelne Basen (bzw. Buchstaben) sind. Hierdurch wird die chemische Eigenschaft abgebildet, dass sich der DNA Strang $a_1 \cdots a_n$ an den komplementären DNA Strang $\overline{a_1 \cdots a_n}$ anlagern kann. Man beachte hierbei, dass

$$\overline{5'\text{-}a_1 \cdots a_n\text{-}3'} = 3'\text{-}\overline{a_1} \cdots \overline{a_n}\text{-}5' = 5'\text{-}\overline{a_n} \cdots \overline{a_1}\text{-}3'.$$

Eine Technik, die häufig verwendet wird um DNA Stränge mit bestimmten Eigenschaften und ihre Komplemente exponentiell zu vervielfältigen, ist die *Polymerase-Kettenreaktion* (engl. polymerase chain reaction, PCR). Die PCR wiederholt drei biochemische Prozesse mehrmals hintereinander, siehe Abb. 2. Angenommen ein langer DNA-Strang τ, das *Template*, soll vervielfältigt werden. Hierzu werden kurze DNA Stränge, sogenannte *Primer*, welche komplementär zu einem Suffix des Templates sind, zur Lösung hinzugefügt. (Ein Suffix ist eine Basen-Sequenz, die dem $3'$-Ende vorangeht.) Falls $\tau = \gamma\alpha$, wobei α verhältnismäßig kurz ist, dann ist $\overline{\alpha}$ ein geeigneter Primer. Während der *Hybridisierungsphase* lagert sich einer der Primer an das Template an. Durch freie Nukleobasen wird Base nach Base des Templates komplementiert, beginnend am $3'$-Ende des Primers. Dieser Prozess wird *Elongation* genannt. Nachdem das Template vollständig komplementiert wurde, wird der neu entstandene Doppelstrang *denaturiert* (aufgetrennt) und wir erhalten die Einzelstränge τ und $\overline{\tau}$. Wiederholt man diese drei Schritte und fügt Primer hinzu, welche zu einem Suffix von $\overline{\tau}$ komplementär sind (d. h. ein Präfix von τ sind), verdoppelt sich die Anzahl von Templates und ihrer Komplemente nach jedem Zyklus.

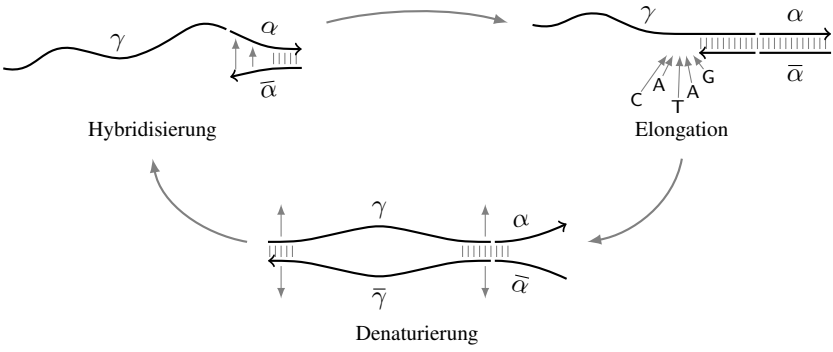


Abbildung 2: Polymerase-Kettenreaktion

Die Haarnadel-Vervollständigung ist die Modifikation eines DNA Strangs, die während der PCR entstehen kann, siehe Abb. 3. Falls ein DNA Strang die Form $\gamma\alpha\beta\overline{\alpha}$ besitzt, so kann der Suffix $\overline{\alpha}$ als Primer auf den DNA-Strang selbst wirken und sich während der Hybridisierung an die DNA Sequenz α anlagern. Eine solche intramolekulare Basenpaarung

wird *Haarnadelstruktur*, oder einfach *Haarnadel*, genannt. Durch Elongation wird der zuvor ungebundene Teil γ des DNA Strangs komplementiert und nach der Denaturierung erhalten wir einen neuen DNA Strang $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$, welcher als *Haarnadel-Vervollständigung* des DNA Strangs $\gamma\alpha\beta\overline{\alpha}$ bezeichnet wird.

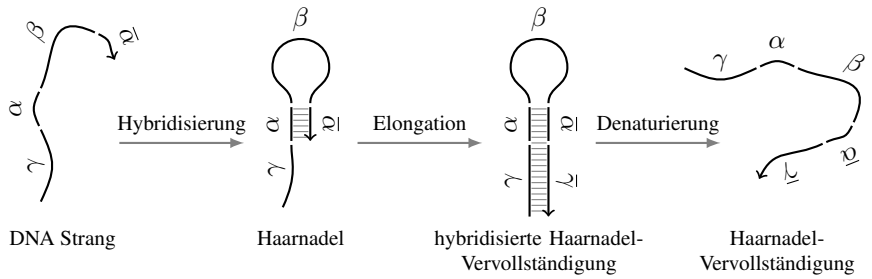


Abbildung 3: Haarnadel-Vervollständigung eines DNA Strangs

Häufig werden Haarnadel-Vervollständigungen als störendes Nebenprodukt von DNA-gestützten Berechnungen gesehen, weshalb in vielen Arbeiten daran geforscht wurde, DNA Bibliotheken (Mengen von DNA Strängen) zu entwerfen, welche keine Haarnadeln oder andere ungewollte Strukturen ausbilden, siehe u. a. [DMG+98, KKL+05, KMT07]. Andererseits wurden auch DNA Algorithmen entworfen, welche Haarnadeln oder Haarnadel-Vervollständigungen nutzen um Berechnungsschritte durchzuführen. Ein Beispiel hierfür ist die *Whiplash-PCR*. Bei dieser Form der PCR werden Haarnadel-Vervollständigungen, welche durch Stopper-Sequenzen kontrolliert sind, verwendet, um zufällige Pfade auf gerichteten Graphen abzulaufen. Da sehr viele dieser zufälligen Pfade parallel berechnet werden können, ist es möglich mithilfe der Whiplash-PCR NP-vollständige Probleme wie das HAMILTON PFAD PROBLEM zu lösen [HAK+97, Win98].

1.2 Haarnadelstrukturen in Formalen Sprachen

In der restlichen Arbeit betrachten wir Haarnadel-Vervollständigungen als Operation auf formalen Sprachen. Der Leser, der nicht mit den Grundlagen der formalen Sprachen vertraut ist, sei auf [HU79] verwiesen.

Wir betrachten Wörter und Sprachen über einem fest gewählten Alphabet Σ. Die Menge aller Wörter wird mit Σ^* bezeichnet und das leere Wort mit 1. Das Alphabet Σ sei mit einer Involution $\overline{\ }$ ausgestattet, d. h. $\overline{\overline{a}} = a$ für alle $a \in \Sigma$. Wie zuvor erweitern wir diese Notation auf Wörter $\overline{a_1 \cdots a_n} = \overline{a_n} \cdots \overline{a_1}$, wobei $a_1, \ldots, a_n \in \Sigma$. Für $w \in \Sigma^*$, bezeichnet $|w|$ die Länge des Wortes. Lässt sich $w = xyz$ schreiben, mit $x, y, z \in \Sigma^*$, so werden x, y, z als *Präfix*, *Faktor*, bzw. *Suffix* von w bezeichnet.

Sei $w = \gamma\alpha\beta\overline{\alpha}$ ein Wort, so wird $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$ eine *(rechte) Haarnadel-Vervollständigung* von w genannt. Da Haarnadeln in der Biochemie nur stabil sind, wenn die Bindung zwischen α und $\overline{\alpha}$ stark genug ist, definieren wir eine Konstante k und fordern $|\alpha| = k$. (Wohlgemerkt änderte sich die Definition nicht, würden wir $|\alpha| \geq k$ fordern.) Sei $w = \alpha\beta\overline{\alpha}\overline{\gamma}$, mit

$|\alpha| = k$, so bezeichnen wir $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$ als eine *(linke) Haarnadel-Vervollständigung* von w.

Seien nun L_1 und L_2 Sprachen, so ist

$$\mathcal{H}_k(L_1, L_2) = \{\gamma\alpha\beta\overline{\alpha}\overline{\gamma} \mid |\alpha| = k \wedge (\gamma\alpha\beta\overline{\alpha} \in L_1 \vee \alpha\beta\overline{\alpha}\overline{\gamma} \in L_2)\}$$

die *Haarnadel-Vervollständigung* von L_1 und L_2, d. h. wir vereinigen alle rechten Haarnadel-Vervollständigungen von Wörtern aus L_1 mit den linken Haarnadel-Vervollständigungen von Wörtern aus L_2. Häufig wird in der Literatur zwischen der rechten, linken und beidseitigen Haarnadel-Vervollständigung unterschieden. Wir erhalten die rechte Haarnadel-Vervollständigung, falls $L_2 = \emptyset$, die linke Haarnadel-Vervollständigung, falls $L_1 = \emptyset$ und die beidseitige Haarnadel-Vervollständigung, falls $L_1 = L_2$. Unsere Definition erlaubt es alle drei Varianten gleichzeitig zu untersuchen.

Da die PCR in der Biochemie selten nach einem Zyklus gestoppt wird, stellt sich die Frage, wie sich DNA Stränge, die durch Haarnadel-Vervollständigungen entstanden sind, weiter entwickeln. Dies inspiriert die Untersuchung der *iterierten Haarnadel-Vervollständigung*. Im iterierten Fall, ergibt es keinen Sinn mit zwei Sprachen zu starten. Es sei L eine formale Sprache. Die iterierte Haarnadel-Vervollständigung von L ist der reflexive und transitive Abschluss der beidseitigen Haarnadel-Vervollständigung

$$\mathcal{H}_k^*(L) = \bigcup_{i \geq 0} \mathcal{H}_k^i(L),$$

wobei

$$\mathcal{H}_k(L) = \mathcal{H}_k(L, L), \qquad \mathcal{H}_k^0(L) = L, \qquad \mathcal{H}_k^{i+1}(L) = \mathcal{H}_k(\mathcal{H}_k^i(L)) \quad \text{für } i \geq 0.$$

Die Haarnadel-Vervollständigung wurde erstmals in einer Arbeit von Cheptea, Martín-Vide und Mitrana 2006 definiert [CMVM06]. In den darauffolgenden Jahren wurde die Haarnadel-Vervollständigung aus formalsprachlichen und algorithmischen Gesichtspunkten untersucht, siehe u. a. [CMVM06, MMY08, MMY09, MMVM09]. In dieser Arbeit betrachten wir neben der Haarnadel-Vervollständigung zwei weitere, verwandte Operationen, die begrenzte Haarnadel-Vervollständigung und die Haarnadel-Verlängerung. Bei der begrenzten Variante ist die Länge des γ-Faktors durch eine Konstante m begrenzt. Diese Variante der Haarnadel-Vervollständigung wurde in [ILM09,ILMM11] untersucht. Bei der Haarnadel-Verlängerung erlauben wir, dass nur ein Teil des ungebunden γ-Faktors komplementiert wird, siehe [MMVM10]. Formale Definitionen beider Operationen werden wir später angeben.

Die folgenden Kapitel fassen die Ergebnisse zusammen, die im Rahmen meiner Dissertation [Kop11] erarbeitet wurden. Aufgrund der Seitenbegrenzung wurde allerdings auf Beweise verzichtet.

2 Haarnadel-Vervollständigungen regulärer Sprachen

In diesem Kapitel seinen L_1 und L_2 reguläre Sprachen. Wir untersuchen ihre Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$. In Kapitel 2.1 beschäftigen wir uns mit der Klasse der

Sprachen, in der $\mathcal{H}_k(L_1, L_2)$ liegt und wir zeigen, dass es Entscheidbar ist, ob $\mathcal{H}_k(L_1, L_2)$ regulär ist. Kapitel 2.2 befasst sich mit Haarnadel-Vervollständigungen von Sprachen in bestimmten Varietäten, d. h. Unterklassen der regulären Sprachen.

2.1 Eindeutigkeit und Regularität

In [CMVM06] wurden die Abschlusseigenschaften verschiedener Sprachklassen unter Haarnadel-Vervollständigung untersucht. Es wurde gezeigt, dass weder die regulären, noch die kontextfreien Sprachen unter Haarnadel-Vervollständigung abgeschlossen sind. Hingegen sind die kontextsensitiven Sprachen unter dieser Operation abgeschlossen. Weiterhin, ist bekannt, dass Haarnadel-Vervollständigung regulärer Sprachen eine linear kontextfreie Sprache ist.

Beispiel 2.1. Sei $\Sigma = \{a, \overline{a}\}$ und $L_1 = a^*\overline{a}^k$. Die (rechte) Haarnadel-Vervollständigung von L_1 ist

$$\mathcal{H}_k(L_1, \emptyset) = \left\{ a^i\overline{a}^j \;\middle|\; i \geq j \geq k \right\}$$

und somit nicht regulär. Wählen wir allerdings $L_2 = \overline{L_1} = a^k\overline{a}^*$, so ist die Haarnadel-Vervollständigung

$$\mathcal{H}_k(L_1, L_2) = \left\{ a^i\overline{a}^j \;\middle|\; i, j \geq k \right\}$$

wieder eine reguläre Sprache.

In diesem Kapitel betrachten wir das Entscheidungsproblem, ob zwei gegebene reguläre Sprachen eine reguläre Haarnadel-Vervollständigung besitzen. In [CMVM06] wurde dies als offenes Problem gestellt. Da unentscheidbar ist, ob eine gegebene lineare Grammatik eine reguläre Sprache erzeugt [Gre68], kann kein allgemeiner Ansatz gewählt werden, um das Problem zu lösen. Eine erste Lösung für das Problem haben wir in [DKM09] gegeben, wo wir zeigten, dass das Problem in polynomieller Zeit entscheidbar ist. Der Grad des Polynoms wurde allerdings nur abgeschätzt auf 20. In dieser Arbeit wird zum einen ein verbesserter Entscheidungsalgorithmus präsentiert, und zum anderen werden wir die Sprachklasse, in der Haarnadel-Vervollständigungen regulärer Sprachen liegen, weiter einschränken.

Theorem 2.2. *Es seien L_1 und L_2 regulär, dann ist die Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$ eindeutig linear kontextfrei.*

Eindeutig bedeutet hier, dass eine lineare Grammatik existiert, welche für jedes Wort in der Sprache genau einen Ableitungspfad besitzt. Dieses neue Resultat ermöglicht es, die Wachstumsfunktion

$$g_{\mathcal{H}_k(L_1,L_2)}(n) = \left|\Sigma^{\leq n} \cap \mathcal{H}_k(L_1, L_2)\right|$$

der Haarnadel-Vervollständigung zu berechnen und sie mit den Wachstumsfunktionen der zugrundeliegenden Sprachen L_1 und L_2 zu vergleichen. Somit lässt sich zum Beispiel berechnen, wie groß die erwartete Anzahl fehlerhafter DNA Stränge ist, die durch Haarnadel-Vervollständigungen während eines PCR Zyklus erzeugt werden (angenommen, dass die Haarnadelbildung ungewollt ist).

Für das Entscheidungsproblem ob $\mathcal{H}_k(L_1, L_2)$ regulär ist, gehen wir davon aus, dass die Sprachen L_1 und $\overline{L_2}$ als deterministische endliche Automaten (DEAs) gegeben sind. Folgende Komplexitäten konnten wir beweisen.

Theorem 2.3. *Es seien L_1 und $\overline{L_2}$ gegeben als DEAs deren Zustandszahl durch n beschränkt ist. Das Problem, ob $\mathcal{H}_k(L_1, L_2)$ regulär ist, ist entscheidbar in*

i.) $\mathcal{O}(n^2)$ *Zeit, falls $L_1 = \emptyset$ oder $L_2 = \emptyset$.*

ii.) $\mathcal{O}(n^6)$ *Zeit, falls $L_1 = \overline{L_2}$.*

iii.) $\mathcal{O}(n^8)$ *Zeit, im Allgemeinen.*

Theorem 2.4. *Es seien L_1 und $\overline{L_2}$ gegeben als DEAs. Das Problem, ob $\mathcal{H}_k(L_1, L_2)$ regulär ist, ist* NL*-vollständig.*

Die Sprachklasse NL enthält diejenigen Probleme, die von einer nicht-deterministischen Turingmaschine in logarithmischem Platz entschieden werden können. Hierbei sei angemerkt, dass die Zugehörigkeit zu NL impliziert, dass das Problem in *Nick's Class* NC_2 liegt und damit effizient parallelisierbar ist, siehe z. B. [Pap94].

2.2 Varietäten

Varietäten sind Unterklassen der regulären Sprachen, welche durch Eigenschaften ihrer syntaktischen Monoide definiert sind. Eine Varietät über endlichen Monoiden ist eine Menge von Monoiden, welche unter Division und direktem Produkt abgeschlossen ist. Für eine ausführliche Einführung in Varietäten über formalen Sprachen sei auf [Pin86] verwiesen. Einige Varietäten besitzen weitere algebraische, kombinatorische und logische Charakterisierungen. So entspricht die Varietät der *aperiodischen* Sprachen $\mathbf{A}$ unter anderem den *Stern-freien* Sprachen und der Klasse von Sprachen, die durch Sätze in *Logik erster Stufe* (first order logic) $\mathrm{FO}[<]$ spezifiziert werden können. Stern-frei bedeutet, dass eine Sprache durch einen regulären Ausdruck ohne Kleene-Stern, dafür aber mit mengentheoretischem Komplement $L^c = \Sigma^* \setminus L$, beschrieben werden kann. Die zweite Varietät, die wir betrachten werden, ist die Varietät $\mathbf{LDA}$, welche der Klasse von Sprachen entspricht, die durch Sätze in *Logik erster Stufe mit zwei Variablen und Nachfolger-Prädikat* $\mathrm{FO}^2[<, +1]$ spezifiziert werden können.

Man beachte, dass die Sprache L_1 in Beispiel 2.1 in der Varietät $\mathbf{LDA} \subsetneq \mathbf{A}$ liegt und dass $\mathcal{H}_k(L_1, \emptyset)$ nicht regulär ist. Sofern die Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$ von zwei Sprachen allerdings regulär ist, bleibt die Zugehörigkeit zu den Varietäten $\mathbf{A}$ und $\mathbf{LDA}$ erhalten.

Theorem 2.5. *Seien L_1 und L_2 Sprachen in $\mathbf{A}$ (bzw. $\mathbf{LDA}$). Die Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$ ist entweder nicht regulär oder sie gehört ebenfalls zur Varietät $\mathbf{A}$ (bzw. $\mathbf{LDA}$).*

3 Haarnadel-Verlängerung

Die Haarnadel-Verlängerung kann während eines PCR-Schrittes entstehen, wenn der Elongationsprozess abgebrochen wird, bevor der ungebundene γ-Teil des des DNA Strangs vollständig komplementiert ist. Sei $w = \gamma_1\alpha\beta\overline{\alpha}$ ein Wort mit $|\alpha| = k$, dann ist $\gamma_1\alpha\beta\overline{\alpha}\overline{\gamma_2}$ eine *(rechte) Haarnadel-Verlängerung* von w, falls γ_2 ein Suffix von γ_1 ist. Des weiteren ist $\gamma_1\alpha\beta\overline{\alpha}\overline{\gamma_2}$ eine *(linke) Haarnadel-Verlängerung* von $\alpha\beta\overline{\alpha}\overline{\gamma_2}$ falls $|\alpha| = k$ und $\overline{\gamma_1}$ ein Präfix von $\overline{\gamma_2}$ (bzw. γ_1 ein Suffix von γ_2) ist, siehe Abb. 4.

rechte Haarnadel-Verlängerung linke Haarnadel-Verlängerung

Abbildung 4: Haarnadel-Verlängerung

Analog zur Haarnadel-Vervollständigung, definieren wir für Sprachen L_1 und L_2 die Haarnadel-Verlängerung

$$\mathcal{HL}_k(L_1, L_2) = \{\gamma_1\alpha\beta\overline{\alpha}\overline{\gamma_2} \mid |\alpha| = k \wedge ((\gamma_1\alpha\beta\overline{\alpha} \in L_1 \wedge \gamma_2 \text{ ist Suffix von } \gamma_1) \vee (\alpha\beta\overline{\alpha}\overline{\gamma_2} \in L_2 \wedge \gamma_1 \text{ ist Suffix von } \gamma_2))\}.$$

Es ist bekannt, dass die Haarnadel-Verlängerung regulärer Sprachen linear kontextfrei und nicht zwingend regulär ist [MMVM10]. Auf den ersten Blick scheint es, als würden sich die Haarnadel-Vervollständigung und die Haarnadel-Verlängerung, angewandt auf reguläre Sprachen, sehr ähnlich verhalten. Betrachten wir erneut die Frage, ob entscheidbar ist, ob reguläre Sprachen eine reguläre Haarnadel-Verlängerung besitzen, so können wir zumindest im einseitigen Fall analoge Resultate beweisen.

Theorem 3.1. *Sei L eine reguläre Sprache. Das Problem, ob die rechte Haarnadel-Verlängerung $\mathcal{HL}_k(L, \emptyset)$ (bzw. linke Haarnadel-Verlängerung $\mathcal{HL}_k(\emptyset, L)$) regulär ist, ist*

i.) NL*-vollständig,*

ii.) entscheidbar in $\mathcal{O}(n^2)$ Zeit,

falls L (bzw. $\overline{L}$) als DEA mit n Zuständen gegeben ist.

Allerdings war es uns nicht möglich ein analoges Resultat für den beidseitigen oder allgemeinen Fall zu beweisen. Das Problem, ob die Regularität einer Haarnadel-Verlängerung regulärer Sprachen entscheidbar ist, bleibt weiterhin offen. Wir können allerdings ein Indiz nennen, warum dieses Problem schwieriger zu lösen sein könnte, als das Regularitätsproblem der Haarnadel-Vervollständigung. Während im Beweis der Theoreme 2.3 und 2.4 ausnutzt wird, dass eine eindeutige lineare Darstellung der Haarnadel-Vervollständigung konstruiert werden kann (vgl. Theorem 2.2), können wir zeigen, dass keine eindeutig lineare Grammatik für die rechte oder beidseitige Haarnadel-Verlängerung der

Sprache $L = (b^+\alpha)^+\overline{\alpha}$ existiert, wobei $\Sigma = \{a, \bar{a}, b, \bar{b}\}$ und $\alpha = a^k$. Die Sprachen $\mathcal{HL}_k(L, \emptyset)$ und $\mathcal{HL}_k(L, L)$ sind also *inhärent mehrdeutig*. Wir können sogar zeigen, dass für jedes $m \in \mathbb{N}$ ein Wort $w \in \mathcal{HL}_k(L, \emptyset)$ (bzw. $w \in \mathcal{HL}_k(L, L)$) existiert, sodass in jeder Grammatik, die $\mathcal{HL}_k(L, \emptyset)$ (bzw. $\mathcal{HL}_k(L, L)$) generiert, das Wort w mindestens m verschiedene Ableitungen besitzt, d. h. jede Grammatik, die $\mathcal{HL}_k(L, \emptyset)$ oder $\mathcal{HL}_k(L, L)$ erzeugt, besitzt einen *unbeschränkten Grad der Mehrdeutigkeit*.

Theorem 3.2. *Die Haarnadel-Verlängerung $\mathcal{HL}_k(L_1, L_2)$ zweier regulärer Sprachen L_1 und L_2 kann inhärent Mehrdeutig sein, selbst wenn $L_1 = \emptyset$ oder $L_2 = \emptyset$.*

4 Iterierte begrenzte Haarnadel-Vervollständigung

Die begrenzte Haarnadel-Vervollständigung ist eine Variante der Haarnadel-Vervollständigung, bei der wir fordern, dass die Länge des γ-Faktors einer Haarnadel-Vervollständigung durch eine Konstante m begrenzt ist. Sei L eine Sprache und $m \geq 1$, so definieren wir die *begrenzte Haarnadel-Vervollständigung* als

$$\mathcal{H}_{k,m}(L) = \{\gamma\alpha\beta\overline{\alpha}\overline{\gamma} \mid |\alpha| = k \wedge |\gamma| \leq m \wedge (\gamma\alpha\beta\overline{\alpha} \in L \vee \alpha\beta\overline{\alpha}\overline{\gamma} \in L)\}.$$

Im Gegensatz zur unbeschränkten Variante sind alle Klassen der Chomsky Hierarchie unter begrenzter Haarnadel-Vervollständigung abgeschlossen [ILM09, ILMM11]. Weiterhin wurde gezeigt, dass die Klassen der kontextfreien, kontextsensitiven und rekursiv aufzählbaren Sprachen unter *iterierte begrenzte Haarnadel-Vervollständigung* $\mathcal{H}^*_{k,m}(L) = \bigcup_{i\geq 0} \mathcal{H}^i_{k,m}(L)$ abgeschlossen sind. Allerdings blieb in den Arbeiten unbeantwortet, ob die regulären Sprachen unter iterierter begrenzter Haarnadel-Vervollständigung abgeschlossen ist. Wir zeigen, dass diese tatsächlich unter iterierter begrenzter Haarnadel-Vervollständigung abgeschlossen sind. Unser Ergebnis ist sogar allgemeiner und lässt sich auf allen Klassen der Chomsky Hierarchie sowie auf alle „klassischen“ Komplexitätsklassen anwenden.

Theorem 4.1. *Sei $\mathcal{C}$ eine Sprachklasse, welche (effektiv) unter Vereinigung, Durchschnitt mit regulären Sprachen und Konkatenation mit regulären Sprachen abgeschlossen ist, dann ist $\mathcal{C}$ auch (effektiv) abgeschlossen unter iterierter begrenzter Haarnadel-Vervollständigung.*

Insbesondere ist die Klasse der regulären Sprachen effektiv unter iterierter begrenzter Haarnadel-Vervollständigung abgeschlossen.

Im Beweis des Theorems wird die Sprache der iterierten begrenzten Haarnadel-Vervollständigung konstruiert. Dies erlaubt die Untersuchung, wie groß ein nicht-deterministischer endlicher Automat (NEA) ist, der die iterierte Haarnadel-Vervollständigung einer regulären Sprache akzeptiert. Wir geben eine untere und obere Schranke für einen solchen NEA an, welche beide exponentiell in der Konstante m sind.

Theorem 4.2.

i.) Es existiert eine reguläre Sprache L, sodass für alle $m \geq 1$ weder $\mathcal{H}_{k,m}(L)$ noch $\mathcal{H}^*_{k,m}(L)$ durch einen NEA mit weniger als 2^m Zuständen akzeptiert werden kann.

ii.) Sei L eine reguläre Sprache, welche durch einen NEA mit n Zuständen akzeptiert wird und $m \geq 1$. Es existiert ein NEA mit $2^{\mathcal{O}(m^2)}n$ Zuständen, welcher die iterierte begrenzte Haarnadel-Vervollständigung $\mathcal{H}_{k,m}(L)$ akzeptiert.

5 Iterierte Haarnadel-Vervollständigungen einelementiger Sprachen

Die Klasse der Sprachen, die durch iterierte Haarnadel-Vervollständigung einelementiger Sprachen (oder Wörter) erzeugt werden, ist gegeben durch

$$\mathrm{HCS}_k = \{\mathcal{H}_k^*(\{w\}) \mid w \in \Sigma^*\}.$$

Diese Sprachklasse wurde erstmals in [MMY08] untersucht. Da die Klasse NL unter iterierter Haarnadel-Vervollständigung abgeschlossen ist [CMVM06], ist HCS_k eine Teilmenge von NL und somit in den kontextsensitiven Sprachen enthalten. Dennoch wurde die Frage, ob HCS_k nicht-reguläre oder nicht-kontextfreie Sprachen enthält, nicht beantwortet und als offenes Problem in [MMY08] gestellt. Wir lösen dieses Problem, indem wir zeigen, dass die iterierte Haarnadel-Vervollständigung $\mathcal{H}_k^*(\{\alpha b \alpha \overline{\alpha} \alpha c \overline{\alpha}\})$ nicht kontextfrei ist, wobei $\Sigma = \left\{a, \bar{a}, b, \bar{b}, c, \bar{c}\right\}$ und $\alpha = a^k$.

Theorem 5.1. *Die iterierte Haarnadel-Vervollständigung einer einelementigen Sprache ist nicht im Allgemeinen kontextfrei.*

Literatur

[CMVM06] Daniela Cheptea, Carlos Martín-Vide und Victor Mitrana. A new operation on words suggested by DNA biochemistry: Hairpin completion. *Transgressive Computing*, Seiten 216–228, 2006.

[DKM09] Volker Diekert, Steffen Kopecki und Victor Mitrana. On the Hairpin Completion of Regular Languages. In Martin Leucker und Carroll Morgan, Hrsg., *ICTAC*, Jgg. 5684 of *LNCS*, Seiten 170–184. Springer, 2009.

[DMG+98] R. Deaton, R. Murphy, M. Garzon, D.R. Franceschetti und S.E. Stevens. Good encodings for DNA-based solutions to combinatorial problems. *Proc. of DNA-based computers DIMACS Series*, 44:247–258, 1998.

[Gre68] Sheila A. Greibach. A Note on Undecidable Properties of Formal Languages. *Mathematical Systems Theory*, 2(1):1–6, 1968.

[HAK+97] Masami Hagiya, Masanori Arita, Daisuke Kiga, Kensaku Sakamoto und Shigeyuki Yokoyama. Towards Parallel Evaluation and Learning of Boolean μ-Formulas with Molecules. In *Second Annual Genetic Programming Conf.*, Seiten 105–114, 1997.

[HU79] J. E. Hopcroft und J. D. Ulman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 1979.

[ILM09] Masami Ito, Peter Leupold und Victor Mitrana. Bounded Hairpin Completion. In *LATA '09: Proceedings of the 3rd International Conference on Language and Automata Theory and Applications*, Seiten 434–445, Berlin, Heidelberg, 2009. Springer-Verlag.

[ILMM11] Masami Ito, Peter Leupold, Florin Manea und Victor Mitrana. Bounded hairpin completion. *Inf. Comput.*, 209:471–485, March 2011.

[KKL+05] Lila Kari, Stavros Konstantinidis, Elena Losseva, Petr Sosík und Gabriel Thierrin. Hairpin Structures in DNA Words. In Alessandra Carbone und Niles A. Pierce, Hrsg., *DNA*, Jgg. 3892 of *LNCS*, Seiten 158–170. Springer, 2005.

[KMT07] Lila Kari, Kalpana Mahalingam und Gabriel Thierrin. The syntactic monoid of hairpin-free languages. *Acta Inf.*, 44(3-4):153–166, 2007.

[Kop11] Steffen Kopecki. *Formal language theory of hairpin formations*. Dissertation, University of Stuttgart, 2011.
`http://elib.uni-stuttgart.de/opus/volltexte/2011/63780`

[MMVM09] Florin Manea, Carlos Martín-Vide und Victor Mitrana. On some algorithmic problems regarding the hairpin completion. *Discrete Applied Mathematics*, 157(9):2143–2152, 2009.

[MMVM10] Florin Manea, Carlos Martín-Vide und Victor Mitrana. Hairpin Lengthening. In Fernando Ferreira, Benedikt Löwe, Elvira Mayordomo und Luís Mendes Gomes, Hrsg., *CiE*, Jgg. 6158 of *LNCS*, Seiten 296–306. Springer, 2010.

[MMY08] Florin Manea, Victor Mitrana und Takashi Yokomori. Some Remarks on the Hairpin Completion. In Erzsebet Csuhaj-Varju und Zoltan Esik, Hrsg., *12th International Conference AFL 2008 Proceedings*, Seiten 302–312, 2008.

[MMY09] Florin Manea, Victor Mitrana und Takashi Yokomori. Two complementary operations inspired by the DNA hairpin formation: Completion and reduction. *Theor. Comput. Sci.*, 410(4-5):417–425, 2009.

[Pap94] Christos H. Papadimitriou. *Computatational Complexity*. Addison Wesley, 1994.

[Pin86] Jean-Éric Pin. *Varieties of Formal Languages*. North Oxford Academic, London, 1986.

[Win98] Erik Winfree. Whiplash PCR for O(1) Computing. In *University of Pennsylvania*, Seiten 175–188, 1998.

Steffen Kopecki wurde am 15. Januar 1983 in Stuttgart geboren. Bis zu seinem 12. Lebensjahr wohnte er in Stuttgart Kaltental, wo er die Grundschule Kaltental und bis zur sechsten Klasse das Fanny-Leicht-Gymnasium Stuttgart besuchte. 1995 zog er mit seiner Familie nach Nürtingen und beendete dort seine Schulausbildung am Hölderlin-Gymnasium Nürtingen im Jahr 2002 mit Abitur. Bevor er zurück nach Stuttgart zum Studieren zog, leistete er im Kreiskrankenhaus Nürtingen Zivildienst als OP-Pfleger. An der Universität Stuttgart begann er den Diplomstudiengang Informatik mit Nebenfach Physik, welchen er im Juni 2009 erfolgreich abschloss. Im gleichen Jahr begann er als Doktorand an der Universität Stuttgart im Institut für formale Methoden der Informatik unter seinem Doktorvater Volker Diekert. Zwei Jahre später, im Juni 2011, verteidigte er seine Dissertation erfolgreich und bekam für die Arbeit die Gesamtnote „mit Auszeichnung bestanden“. Im September 2011 zog Steffen Kopecki nach London in Kanada, wo er heute als Post-Doktorand and der Univeristy of Western Ontario in der Arbeitsgruppe von Lila Kari arbeitet.

Ontologiebasierte Applikationsintegration auf Nutzerschnittstellenebene

Heiko Paulheim
FG Knowledge Engineering
Technische Universität Darmstadt
Hochschulstraße 10
64289 Darmstadt
paulheim@ke.tu-darmstadt.de

Abstract: Typische IT-Landschaften bestehen aus verschiedenen Anwendungen, die parallel genutzt werden. Die Kombination solcher Anwendungen zu integrierten Systemen ist nicht trivial, insbesondere dann nicht, wenn die bestehenden Nutzerschnittstellen weiter verwendet werden sollen. Existierende Ansätze wie Portal- oder Mashup-Lösungen sind vielfach nicht flexibel genug, um eine nahtlose Integration unter Wahrung des Paradigmas der losen Kopplung zu implementieren. Diese Dissertation stellt einen Ansatz vor, der Ontologien und Reasoning nutzt, um Applikationen auf Nutzerschnittstellenebene zu integrieren. Die Arbeit zeigt auf, wie sowohl technologische als auch konzeptionelle Heterogenitäten zwischen den integrierten Applikationen durch den Einsatz von Ontologien überwunden werden können, und diskutiert eine effiziente Implementierung der semantik- und regelgestützten Verarbeitung von Nachrichten zwischen Applikationen. Darüber hinaus wird aufgezeigt, wie der Nutzer bei der Interaktion mit derart integrierten System zusätzlich unterstützt werden kann. Die gesamte Arbeit wird in einer begleitenden Fallstudie aus dem Bereich des Katastrophenmanagements auf ihre praktische Umsetzbarkeit hin geprüft.

1 Einleitung

Applikationsintegration bezeichnet das Zusammenfügen bestehender Anwendungen oder Teile davon zu einem neuen Gesamtsystem, das über die Möglichkeiten der Einzelsysteme hinaus neue Funktionalitäten bereitstellt. Folgt man der klassischen Dreiteilung von Softwaresystemen in Datenhaltung, Applikationslogik und Nutzerschnittstelle [Fow03], so existieren prinzipiell drei Möglichkeiten, eine solche Integration umzusetzen: Integration auf der Ebene der Datenhaltung, auf der Ebene der Applikationslogik, und auf der Ebene der Nutzerschnittstelle [DYB+07].

Wird ein Ansatz auf der Ebene der Datenhaltung oder der Applikationslogik gewählt, so wird stets zumindest eine neue Nutzerschnittstelle implementiert (siehe Abb. 1). Daher haben Ansätze auf der Ebene der Nutzerschnittstelle sowohl für den Entwickler als auch für den Endbenutzer des integrierten Systems. Auf die Nutzerschnittstelle eines interaktiven Systems entfallen ca. 50% des gesamten Entwicklungsaufwandes [MR92], der Wiederverwendungsgrad ist also vielfach höher, wenn ein Integrationsansatz auf der Ebene der

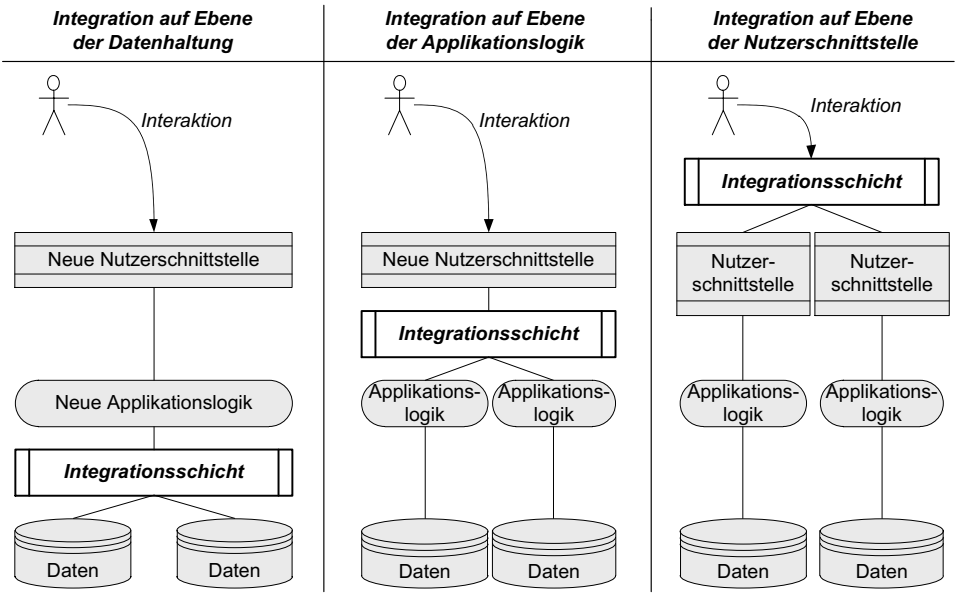


Abbildung 1: Ansätze für Applikationsintegration auf unterschiedlichen Schichten [PP10], angelehnt an [DYB+07]

Nutzerschnittstelle gewählt wird. Außerdem können Nutzer bei derart integrierten Systemen mit bereits bekannten Nutzerschnittstellen interagieren, was die Einarbeitungszeit erheblich verkürzt.

Derzeit verfügbare Ansätze zur Applikationsintegration auf Nutzerschnittstellenebene sind *Portal*- und *Mashup*- Lösungen. Beide erlauben die Wiederverwendung kompletter Applikationen inklusive ihrer Nutzerschnittstellen und die Kombination zu neuen Systemen. Während Portal-Lösungen eher an professionelle Entwickler gerichtet sind und zusätzliche Funktionalitäten wie z.B. Single-Sign-On bieten, sind Mashup-Lösungen eher auf technisch versierte Endbenutzer und semiprofessionelle Programmierer ausgelegt. Untersucht man bestehende Produkte und Ansätze aus beiden Bereichen, so findet man eine gemeinsame Menge an Unzulänglichkeiten, die einer benutzerfreundlichen und gleichzeitig gut wartbaren Lösung entgegen stehen:

- In den meisten Fällen wird Detailwissen über die Implementierung der integrierten Applikationen benötigt, eine Abstraktionsschicht fehlt in der Regel. Dies macht die Integration zeitaufwändig und führt zu einer *engen Kopplung*, die das Gesamtsystem bei Änderungen der integrierten Applikationen potentiell instabil macht.
- Ein gemeinsames Eventmodell fehlt ebenso wie Möglichkeiten der Konvertierung zwischen Datenmodellen der integrierten Applikationen, so dass der Nachrichtenaustausch sowie der *Abgleich konzeptionell heterogener Datenmodelle* größtenteils manuell implementiert werden muss.

- *Technologisch heterogene Applikationen* lassen sich in der Regel nicht beliebig integrieren.

Besondere Probleme bereiten hierbei Interaktionen zwischen Applikationen: soll etwa eine Applikation auf Auswahlaktionen in einer anderen Applikation durch Anzeige von Details reagieren, oder Ziehen und Fallenlassen von Objekten aus einer Applikation in eine andere ermöglicht werden, so treten diese Nachteile besonders auffällig in Erscheinung. In der Regel müssen solche Interaktionen manuell unter Bezugnahme auf Wissen über die Implementierung der Applikationen codiert werden, was Abhängigkeiten zwischen den Applikationen schafft und zu dem Paradigma der losen Kopplung im Widerspruch steht.

Vielen dieser Unzulänglichkeiten lässt sich durch eine zusätzliche Abstraktionsschicht begegnen. In dieser Arbeit werden Ontologien genutzt, um eine solche Abstraktionsschicht zu definieren.

2 Ontologien in der Applikationsintegration

Ontologien sind „formale Spezifikationen einer Konzeptualisierung“ [Gru95], also ein formales Modell der Begriffe, die eine Domäne beschreiben, und ihrer Zusammenhänge. Da sie das Wissen über eine Domäne formal codieren, können auch komplexe Abfragen, die die Kombination verschiedener Fakten zur Beantwortung benötigen, automatisch mit Hilfe von automatischem Schlussfolgern, dem sogenannten *Reasoning*, beantworten werden. Der Nutzen von Ontologien für die Systemintegration wurde bereits früh erkannt; so wurde bereits in einem häufig zitierten Artikel von Uschold und Grüninger aus dem Jahr 1996 die potentielle Verwendung von Ontologien als *inter-lingua* zwischen Systemen zur Herstellung von Interoperabilität vorgeschlagen [UG96].

Die ersten Arbeiten zur Integration auf Datenhaltungsebene mit Ontologien gehen auf die 1980er Jahre zurück. In solchen Arbeiten werden Ontologien genutzt, um eine Abstraktionsschicht über verschiedenen Datenbanken zu bilden. Komplexe Anfragen können mit Hilfe von Begriffen der Ontologie gestellt werden, und ein Integrationssystem übersetzt die Anfragen an die zugrundeliegenden Datenbanksysteme, kombiniert die zurückgelieferten Datensätze, leitet ggf. weitere Fakten daraus ab und liefert eine konsolidierte Antwort zurück [DH05].

Auf der Ebene der Applikationslogik werden Ontologien hauptsächlich im Bereich der *Semantic Web Services* eingesetzt. Hier werden klassische Web-Service-Beschreibungen um formale semantische Beschreibungen angereichert. Damit wird im Idealfall ein komplette, maschinenverständliche Beschreibung eines Dienstes inklusive seiner Vor- und Nachbedingungen sowie seiner Ausführungssemantik und seiner nichtfunktionalen Eigenschaften zur Verfügung gestellt, um so eine automatische Kombination und Ersetzung von Diensten zu ermöglichen [SGA07].

Im Gegensatz zu diesen Ansätzen existieren bislang keine Ansätze, die Ontologien für die Applikationsintegration auf Nutzerschnittstellenebene vorsehen.

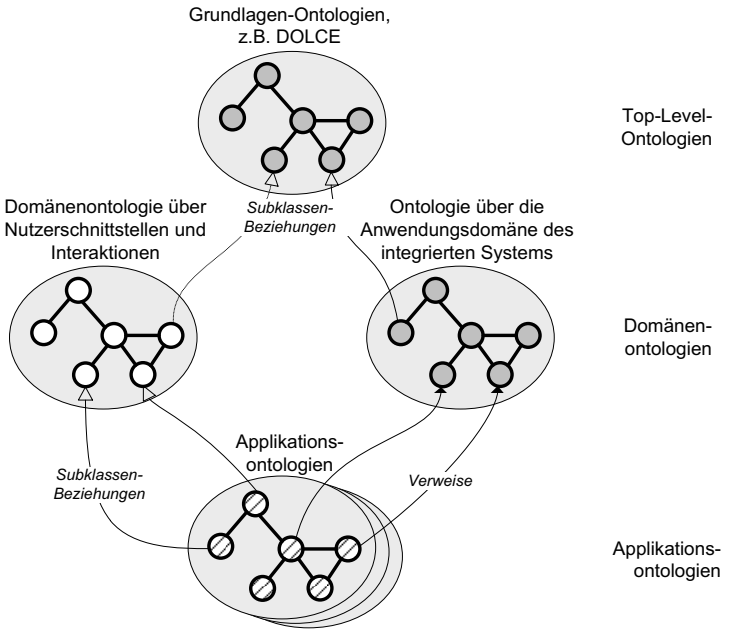


Abbildung 2: Eingesetzte Ontologien, dem Schichtenmodell von Guarino [Gua98] folgend

3 Ansatz

Um eine möglichst *lose Kopplung* der integrierten Applikationen zu erreichen, werden Ontologien überall dort eingesetzt, wo ein Informationsaustausch zwischen den Applikationen stattfinden muss. Sämtliche zwischen den Applikationen ausgetauschten Nachrichten werden komplett mit Hilfe von Ontologien beschrieben, was auch die involvierten Geschäftsobjekte einschließt. Auch Statusinformationen der einzelnen Applikationen, die für andere Applikationen relevant sein können, werden ausschließlich mit Hilfe von Ontologien codiert. Insgesamt kommen drei verschiedene Arten von Ontologien zum Einsatz, wie in Abb. 2 gezeigt:

- Jede einzelne Applikation wird in ihrer eigenen Applikationsontologie beschrieben. Diese formalisiert sämtliche Interaktionsmöglichkeiten mit dieser Applikation. Die Applikationsontologie baut auf zwei verschiedenen Ontologien auf.
- Eine Domänenontologie über Nutzerschnittstellen und Interaktionen beschreibt auf abstrakter Ebene, aus welchen Arten von Komponenten und Interaktionen eine Nutzerschnittstelle bestehen kann.
- Eine Domänenontologie der Geschäftsdomäne beschreibt die konkrete Anwendungsdomäne des integrieten Systems, z.B. das Bankenwesen.

Alle Ontologien werden nutzen die gemeinsame Top-Level-Ontologie DOLCE [MBG+03], wobei dies für die Domänenontologie der Geschäftsdomäne fakultativ ist. Letztere ist nicht Teil des in der Arbeit entwickelten Ansatzes, der domänenunabängig ist.

Jede integrierte Applikation wird in einen sogenannten *Container* verpackt, der die Funktionalität der Anwendung nach außen kapselt (siehe Abb. 3). Der Container stellt die Daten der Anwendungen als *Linked Data* [BHBL09] zur Verfügung, so dass ein Reasoner darauf zugreifen kann. Hierzu werden die Daten der Anwendung mit Hilfe von *Transformationsregeln* in ein ontologiebasiertes Format überführt. Weiterhin übernimmt der Container die Kommunikation mit anderen Anwendungen über Events. Hierzu gehört insbesondere die Koordination von applikationsübergreifenden Interaktionen, z.B. Drag-and-Drop-Aktionen.

Jede Nachricht und jedes Datenobjekt, das zwischen Applikationen ausgetauscht wird, ist komplett mit Ontologien beschrieben. So kann in einer Kombination der verschiedenen Ontologien zum Beispiel ausgedrückt werden, dass der Nutzer eine Aktion vom Typ *Auswählen* mit einem Objekt vom Typ *Kunde* durchgeführt hat, wobei dieses Objekt wiederum weitere Eigenschaften haben kann. Die Ontologie als *inter-lingua* ermöglicht jeder Applikation, Nachrichten von anderen Applikationen zu verstehen, ohne die zugrunde liegende technische Implementierung zu kennen. Damit verlässt sich jede Anwendung lediglich darauf, Nachrichten mit bestimmten Annotationen zu erhalten, direkte Abhängigkeiten zwischen Applikationen werden so eliminiert.

Um die Entkopplung zwischen Applikationen noch stärker voranzutreiben und damit die Modularität und Wartbarkeit des resultierenden Systems noch weiter zu erhöhen, reagieren Container nicht direkt auf Events aus anderen Containern. Die Koordination von anwendungsübergreifenden Interaktionen wird mit Hilfe von *Interaktionsregeln* implementiert, die von einem Reasoner zentral verarbeitet werden. Dieser Reasoner hat Zugriff auf sämtliche Daten und Applikationsontologien. Da die Interaktionsregeln nur auf Konzepten der Ontologien auf Domänenebene, aber nicht auf Konzepten der einzelnen Applikationsontologien formuliert werden, müssen sie in der Regel nicht angepasst werden, wenn sich die Konfiguration des integrierten Gesamtsystems ändert, so dass eine größtmögliche Entkopplung gewährleistet ist.

Der Reasoner kann außerdem auch komplexe Sachverhalte, die in den Domänenontologien modelliert sind, berücksichtigen. So kann beispielsweise folgende Interaktionsregel definiert werden: *Wenn in der Stammdatenanwendung ein Kunde mit Premiumstatus ausgewählt wurde, markiere in der Produktübersicht alle Produkte für Premiumkunden*. Welche Fakten und Axiome dazu führen, dass ein Kunde Premiumstatus bekommt, kann dabei in der Domänenontologie modelliert sein – es lässt sich also zusätzlich zur Entkopplung bei der Integration auch Domänenwissen aus den Anwendungen herausfaktorisieren und anwendungsübergreifend nutzbar machen.

Der Einsatz von Reasoning zur Verarbeitung der Interaktionsregeln ermöglicht zudem eine sehr abstrakte Form der Interaktionsbeschreibung. So ist es zum Beispiel möglich, eine Landkarten-Anwendung wie *Google Maps* mit einer wie folgt formulierten Interaktionsregel zu integrieren: *Wenn der Nutzer ein Objekt, das eine Position hat, auswählt, wird diese Position auf der Karte markiert*. Die Bestimmung, ob ein konkretes Objekt in diese

Kategorie gehört, übernimmt der Reasoner, und es muss nicht a priori bekannt sein, *welche* Arten von Objekten mit Positionen später vom System verarbeitet werden. Damit können Applikationen zum integrierten System hinzugefügt werden und interagieren automatisch mit bereits existierenden Applikationen, ohne dass weitere Anpassungen nötig werden.

Auch globale Interaktionsregeln können definiert werden, um zum Beispiel ein *Linked-Views-Paradigma* zu implementieren: *Wenn der Nutzer ein Objekt selektiert, das auch in anderen Anwendungen bekannt ist, dann hebe dieses Objekt in allen Anwendungen hervor*. Diese Regel kommt ohne Referenz auf die Domänenontologie aus und beschreibt daher eine domänenübergreifende Interaktion, die automatisch greift, sobald zwei Applikationen dieselbe Art von Objekten verarbeiten und die entsprechenden Annotationen an ihre Nachrichten anfügen. Mit Hilfe solcher globaler Interaktionsregeln kann eine grundlegende ad-hoc-Interoperabilität zwischen Anwendungen ermöglicht werden, die greift, ohne dass spezielle Regeln für einzelne Anwendungen definiert werden müssen.

4 Zentrale wissenschaftliche Beiträge

Ein zentrales Artefakt der Arbeit ist die formale Ontologie über Nutzerschnittstellen und Interaktionen [PP11]. Sie stellt ein abstraktes Modell von Nutzerschnittstellen dar, das eine große Bandbreite von Interaktionsmodalitäten abdeckt und zudem durch Nutzung der Top-Level-Ontologie DOLCE stark formalisiert ist. Mit rund 200 Klassen und 500 formalen Axiomen ist sie die derzeit ausdrucksstärkste und umfassendste Ontologie dieser Domäne. Gemeinsam mit formalen Regeln fungiert diese Ontologie als Abstraktionsschicht zwischen integrierten Applikationen und entkoppelt diese voneinander. Da nur noch jede integrierte Applikation auf die Ontologie abgebildet werden muss, sinkt die Komplexität der Integration von $O(n^2)$ auf $O(n)$, und die schwache Kopplung gewährleistet die Wartbarkeit der Anwendung. Eine solche Ontologie kann auch in anderen Bereichen, wie zum Beispiel dem Austausch und der Konvertierung von Nutzerschnittstellenmodellen in der modellgetriebenen Entwicklung, sinnvoll eingesetzt werden.

Durch die Einführung eines abstrakten Modells ist es möglich, sowohl konzeptionell als auch technologisch heterogene Applikationen miteinander zu koppeln. Obgleich es bereits Ansätze gibt, Daten aus Applikationen in ontologiebasierter Form darzustellen, verlassen sich diese in der Regel darauf, dass die Datenmodelle der Applikationen strukturell ähnlich zu der Ontologie sind – eine Annahme, die in der Praxis kaum eintrifft. In dieser Arbeit wurde diesen Ansätzen ein weitaus flexiblerer, regelbasierter Ansatz gegenübergestellt, der in der Lage ist, auch konzeptionell unterschiedliche Klassenmodelle auf beliebige Ontologien abzubilden [POPP12]. Dieser Ansatz ermöglicht nicht nur den Objektaustausch, sondern generell die Entwicklung semantischer Programmiermodelle in weitaus flexiblerer Form, als das bisherige Ansätze leisten.

Neben der konzeptionellen Heterogenität unterstützen die verfügbaren Ansätze zur Integration auf Nutzerschnittstellenebene die Überbrückung technologischer Heterogenität nur unzulänglich. Insbesondere die Implementierung nahtloser Interaktionen, wie z.B. Drag and Drop zwischen Anwendungen, ist bei technologisch verschiedenen Nutzerschnitt-

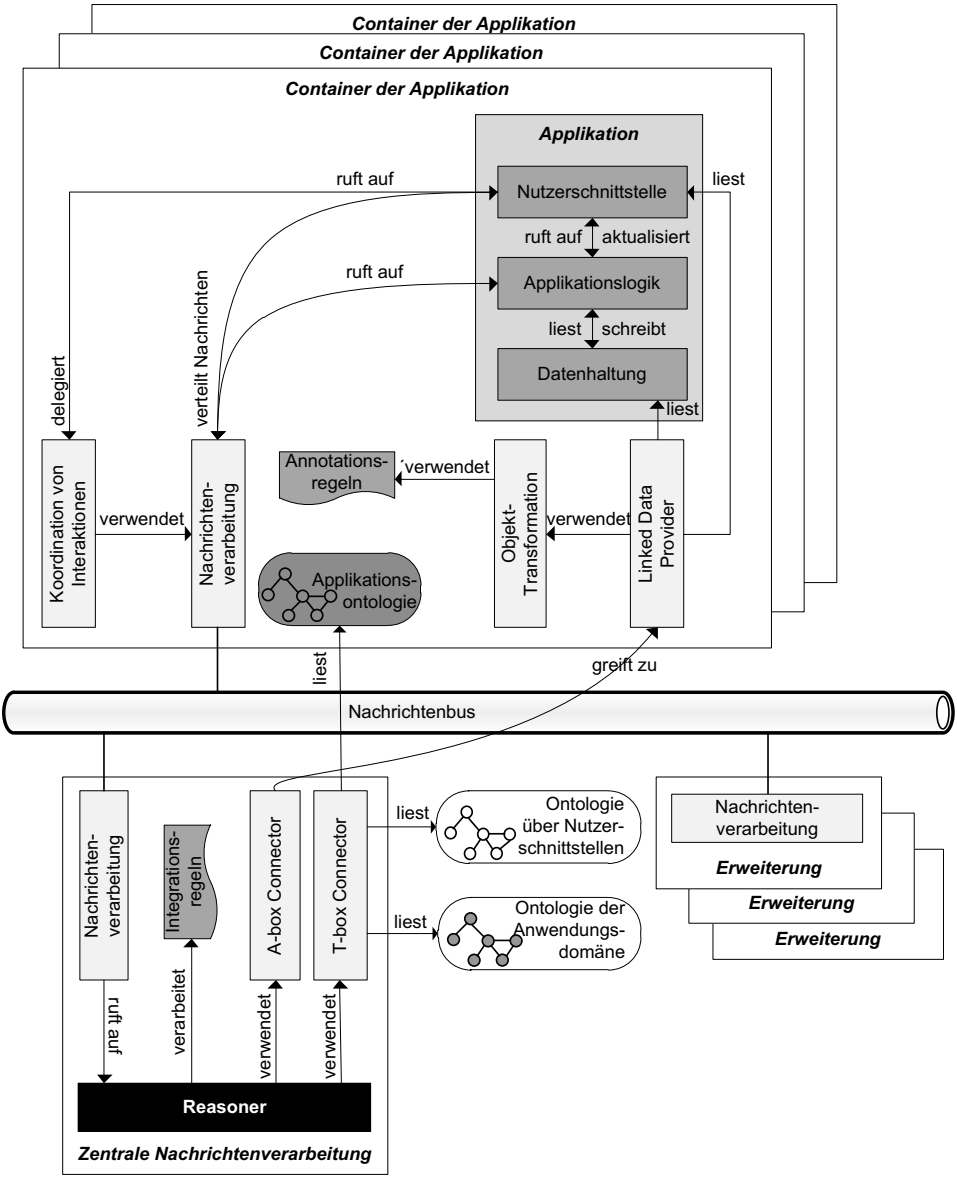


Abbildung 3: Überblick über die Architektur [Pau11b]

stellen, wenn überhaupt, nur sehr aufwändig machbar. Im Rahmen dieser Arbeit wurde am Beispiel von Flex- und Java-Komponenten gezeigt, dass die Einführung einer abstrakten Zwischenschicht auch solche Interaktionen direkt unterstützt [PE10].

Im Umgang mit Nutzerschnittstellen ist die Reaktivität des Systems entscheidend. Gerade bei größeren Faktenmengen stellt die semantische Nachrichtenverarbeitung, wie sie in dieser Arbeit eingesetzt wird, oftmals ein Problem dar, da die Reaktionszeiten für interaktive Systeme deutlich zu lang werden. Im Rahmen dieser Arbeit wurden verschiedene Architekturansätze gegenübergestellt und analysiert. Mit der gewählten Referenzimplementierung ist es möglich, die Reaktionszeiten auch für eine große Anzahl von Applikationen und eine hohe Nachrichtenfrequenz noch deutlich unter der Grenze von einer halben Sekunde zu halten, um eine nutzerfreundliche Interaktion zu gewährleisten [Pau10]. Da die Reaktionszeit in der Praxis oftmals ein Hinderungsgrund für den Einsatz von semantischen Technologien ist, zeigt die Referenzimplementierung Wege auf, wie ontologiegestützte Anwendungen in performanter Form entwickelt werden können, und ebnet damit den Weg für bisher aus Performancegründen nicht umsetzbare Applikationen.

Die Integration mit Hilfe von semantisch annotierten Daten ermöglicht auch neue Interaktionsformen. Da bei dem in der Anwendung gewählten Ansatz sämtliche Daten als Linked Data zur Verfügung stehen, ist es möglich, auf diese neben den existierenden Nutzerschnittstellen parallel mit einem Linked Data Browser zuzugreifen und die Daten zusätzlich als semantisches Netz zu visualisieren. Im Rahmen der Arbeit konnte in einer Nutzerstudie in der Domäne des Katastrophenschutzes gezeigt werden, dass in dieser Domäne typischerweise anfallende Aufgaben mit einer solchen Visualisierung schneller und mit größerer Nutzerzufriedenheit gelöst werden können [Pau11a].

Das in dieser Arbeit entwickelte Framework zur Applikationsintegration auf Ebene der Nutzerschnittstellen wurde im Projekt *SoKNOS* eingesetzt, um ein IT-Unterstützungssystem für Krisenstäbe im Katastrophenschutz zu entwickeln [PDTS+09]. Das System integriert unterschiedlichste Module von der Ressourcenverwaltung über die Nachrichtenverarbeitung bis hin zu interaktiven Lagekarten. Mit Hilfe des Frameworks konnten hier insgesamt 20 verschiedene Anwendungen mit ca. 180 verschiedenen applikationsübergreifenden Interaktionsformen integriert werden. Das System *SoKNOS*, das auf Basis des in dieser Dissertationen entwickelten Frameworks implementiert wurde, wurde bei verschiedenen Messen vorgeführt und mit Endanwendern aus der Katastrpohenschutzdomäne getestet. Die Erfahrungen aus dem Projekt *SoKNOS* zeigen, dass das in dieser Arbeit entwickelte Framework damit auch Anforderungen realer, komplexer IT-Systeme genügt.

5 Zusammenfassung und offene Forschungsfragen

Diese Arbeit stellt einen neuartigen Ansatz vor, um Anwendungsintegration auf Ebene der Nutzerschnittstellen zu ermöglichen. Mit Hilfe von Ontologien werden applikationsübergreifende Interaktionen ermöglicht, ohne das Paradigma der schwachen Kopplung aufzugeben. Der Ansatz erlaubt die Überbrückung von technologischen und konzeptionellen Heterogenitäten und gewährleistet die schnelle Verarbeitung semantisch annotierter Nach-

richten, um den Interaktionsfluss nicht zu verlangsamen. Zudem konnten in einer Nutzerstudie Verbesserungen in der Bedienbarkeit integrierter Systeme durch Visualisierung semantisch annotierter Daten nachgewiesen werden. Zahlreiche Teilergebnisse dieser Arbeit haben Einfluss auf verwandte Forschungsgebiete, in denen Ontologien und semantische Technologien zum Einsatz kommen.

Der in dieser Arbeit entwickelte Ansatz eröffnet weitere relevante Forschungsfragen. So konzentriert sich die Arbeit allein auf Interaktionen mit einem Gerät und einem Endbenutzer; Interaktionen mit multi-modalen Nutzerschnittstellen wurden ebenso ausgeklammert wie Mehrbenutzerinteraktionen. Während die entwickelte Ontologie diese Interaktionsformen prinzipiell abbilden kann, ist die praktische Umsetzung ebenso wie die Benutzbarkeit solcher integrierten Systeme zu untersuchen.

Die in der Arbeit entwickelten Ontologien bilden die integrierten Anwendungen und ihre Interaktionen formal ab. Dies würde es prinzipiell ermöglichen, diese formalen Beschreibungen nicht nur der Maschine, sondern in geeigneter Form auch dem Menschen zugänglich zu machen, um eine explorative Erfahrung des integrierten Systems zu ermöglichen. Geeignete Interaktionsparadigmen und Visualisierungsformen wären hier zu erforschen.

Literatur

[BHBL09] Christian Bizer, Tom Heath und Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[DH05] AnHai Doan und Alon Y. Halevy. Semantic Integration Research in the Database Community: A Brief Survey. *AI Magazine*, 26(1):83–94, 2005.

[DYB+07] Florian Daniel, Jin Yu, Boualem Benatallah, Fabio Casati, Maristella Matera und Regis Saint-Paul. Understanding UI Integration: A Survey of Problems, Technologies, and Opportunities. *IEEE Internet Computing*, 11(3):59–66, 2007.

[Fow03] Martin Fowler. *Patterns of Enterprise Application Architecture*. Addison Wesley, 2003.

[Gru95] Thomas R. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43(5-6):907–928, 1995.

[Gua98] Nicola Guarino, Hrsg. *Formal Ontology and Information Systems*. IOS Press, 1998.

[MBG+03] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino und Alessandro Oltramari. WonderWeb Deliverable D18 – Ontology Library (final). `http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf`, 2003. Letzter Zugriff: 17.04.2012.

[MR92] Brad A. Myers und Mary Beth Rosson. Survey on user interface programming. In *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*, Seiten 195–202. ACM, 1992.

[Pau10] Heiko Paulheim. Efficient Semantic Event Processing: Lessons Learned in User Interface Integration. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral und Tania Tudorache, Hrsg., *The Semantic Web: Research and Applications (ESWC 2010), Part II*, number 6089 in LNCS, Seiten 60–74. Springer, 2010.

[Pau11a] Heiko Paulheim. Improving the Usability of Integrated Applications by Using Interactive Visualizations of Linked Data. *International Journal on Computer Science and Applications (IJCSA)*, 8(2), 2011.

[Pau11b] Heiko Paulheim. *Ontology-based Application Integration*. Springer, 2011.

[PDTS+09] Heiko Paulheim, Sebastian Döweling, Karen Tso-Sutter, Florian Probst und Thomas Ziegert. Improving Usability of Integrated Emergency Response Systems: The SoKNOS Approach. In *Proceedings der 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI) – Informatik 2009*, number 154 in LNI, Seiten 1435–1449, 2009.

[PE10] Heiko Paulheim und Atila Erdogan. Seamless Integration of Heterogeneous UI Components. In Noi Sukaviriya, Jean Vanderdonckt und Michael Harrison, Hrsg., *Proceedings of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS 2010)*, Seiten 303–308. ACM, 2010.

[POPP12] Heiko Paulheim, Daniel Oberle, Roland Plendl und Florian Probst. An Architecture for Information Exchange Based on Reference Models. In *4th International Conference on Software Language Engineering (SLE)*, number 6940 in LNCS. Springer, 2012.

[PP10] Heiko Paulheim und Florian Probst. Application Integration on the User Interface Level: an Ontology-Based Approach. *Data & Knowledge Engineering Journal*, 69(11):1103–1116, 2010.

[PP11] Heiko Paulheim und Florian Probst. A Formal Ontology on User Interfaces – Yet Another User Interface Description Language? In Tim Hussein, Stephan Lukosch, Heiko Paulheim, Jürgen Ziegler und Gaelle Calvary, Hrsg., *Proceedings of the Second Workshop on Semantic Models for Adaptive Interactive Systems (SEMAIS)*, 2011.

[SGA07] Rudi Studer, Stephan Grimm und Andreas Abecker, Hrsg. *Semantic Web Services - Concepts, Technologies and Applications*. Springer, 2007.

[UG96] Mike Uschold und Michael Grüninger. Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 11:93–136, 1996.

Heiko Paulheim wurde am 1.6.1977 in Kassel geboren. Er studierte Mathematik und Anglistik an der Universität Kassel sowie Informatik an der Hochschule Darmstadt und an der Technischen Universität Darmstadt. Er forschte an der Hochschule Darmstadt über den Einsatz von Ontologien im elektronischen Geschäftsverkehr und promovierte bei SAP Research und an der Technischen Universität Darmstadt über die ontologie-basierte Applikationsintegration auf Nutzerschnittstellenebene. Seit 2011 lehrt und forscht er an der Technischen Universität Darmstadt im Fachgebiet Knowledge Engineering, seine aktuellen Forschungsfelder umfassen u.a. die Anwendung von Machine-Learning-Verfahren auf Linked Open Data, Ontology Learning und Ontology Matching. Heiko Paulheim ist Mitglied in zahlreichen Programm-Komitees im Bereich Semantic Web und Künstliche Intelligenz sowie Mitausrichter der Workshopreihen *SEMAIS* (Semantic Models for Adaptive Interactive Systems) und *Know@LOD* (Knowledge Discovery and Data Mining Meets Linked Open Data).

Technologien zur Wiederverwendung von Texten aus dem Web

Martin Potthast

Bauhaus-Universität Weimar
martin.potthast@uni-weimar.de

Abstract: Texte aus dem Web können einzeln oder in großen Mengen wiederverwendet werden. Ersteres wird Textwiederverwendung und letzteres Sprachwiederverwendung genannt. Zunächst geben wir einen Überblick darüber, auf welche Weise Text und Sprache wiederverwendet und wie Technologien des Information Retrieval in diesem Zusammenhang angewendet werden können. In der übrigen Arbeit werden dann eine Reihe spezifischer Retrievalaufgaben betrachtet, darunter die automatische Erkennung von Textwiederverwendungen und Plagiaten, der Vergleich von Texten über Sprachen hinweg, sowie die Wiederverwendung des Webs zur Verbesserung von Suchergebnissen und zur Unterstützung des Schreibens von fremdsprachigen Texten.

1 Einleitung

Etwas wiederzuverwenden bedeutet, es nach seiner ersten Verwendung einem neuen Zweck zuzuführen. Die Wiederverwendung uns umgebender Dinge ist ein alltäglicher Vorgang. Dennoch wird nur selten davon gesprochen, einen Text wiederzuverwenden. Stattdessen spricht man von Zitaten, Übersetzungen, Paraphrasen, Metaphrasen, Zusammenfassungen, Textbausteinen und nicht zuletzt Plagiaten. Sie alle können mit dem Begriff der „Textwiederverwendung" umschrieben und darunter angeordnet werden (siehe Abbildung 1). Einen Text ein zweites Mal zu verwenden ist nichts ungewöhnliches, sondern fester Bestandteil des Schreibens in vielen Genres. Es ist jedoch noch weitgehend unbekannt, wie weit verbreitet die Wiederverwendung von Text heute ist. Das liegt vor allem daran, dass die nötigen Werkzeuge fehlen, dieses Phänomen im großen Stil zu betrachten.

Im Web stehen große Mengen Text zur freien Verfügung. Abgesehen davon, sie einzeln wiederzuverwenden, besteht eine weitere Möglichkeit darin, sie insgesamt wiederzuverwenden, um eine bestimmte Aufgabe (automatisch) zu erledigen. Man spricht in diesem Zusammenhang auch von Sprachwiederverwendung. Da Texte im Web auf unzählige Weisen mit anderen Objekten in Verbindung stehen, liegt hierin großes Potenzial, neue Aufgaben zu finden, die durch geschickte Sprachwiederverwendung besser gelöst werden können. Aus Sicht der Informatik befassen wir uns daher mit folgenden Forschungsfragen:

- Wie und in welchem Umfang können Textwiederverwendungen erkannt werden?
- Welche Aufgaben können durch Sprachwiederverwendung unterstützt werden?

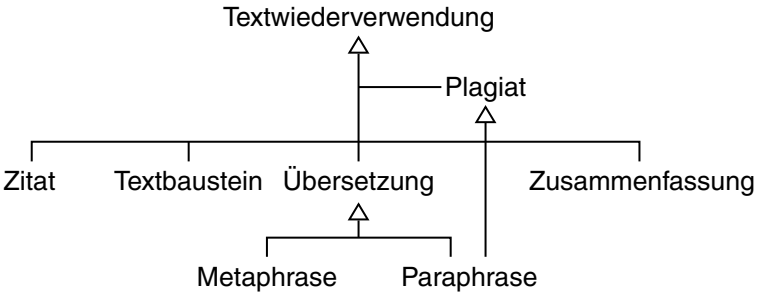


Abbildung 1: Taxonomie bekannter Formen der Textwiederverwendung

1.1 Beiträge

Der vorliegende Text ist eine zusammenfassende, paraphrasierte Übersetzung der in Englisch verfassten Dissertation „Technologies for Reusing Text from the Web" [Pot11]. Die Dissertation ist in zwei Teile gegliedert. Im ersten Teil präsentieren wir Technologien zur Erkennung von Textwiederverwendungen und leisten folgende Beiträge: (1.) Ein einheitlicher Überblick über projektionsbasierte- und einbettungsbasierte Fingerprintingverfahren zur Erkennung fast identischer Texte, sowie die erstmalige Evaluierung einer Reihe dieser Verfahren auf den Revisionshistorien der Wikipedia. (2.) Ein neues Modell zum sprachübergreifenden, inhaltlichen Vergleich von Texten. Das Modell kommt ohne Wörterbücher oder Übersetzungsverfahren aus, sondern benötigt nur eine Menge von Pärchen themenverwandter Texte. Wir vergleichen das Modell in mehreren Sprachen mit herkömmlichen Modellen. (3.) Die erste standardisierte Evaluierungsumgebung für Algorithmen zur Plagiatserkennung. Sie besteht aus Maßen, die die Erkennungsleistung eines Algorithmus' quantifizieren und einem großen Korpus von Plagiaten. Die Plagiate wurden automatisch generiert sowie manuell, mit Hilfe von Crowdsourcing, erstellt. Darüber hinaus haben wir drei internationale Wettbewerbe veranstaltet, in denen insgesamt 32 Forschergruppen ihre Erkennungsansätze gegeneinander antreten ließen.

Im zweiten Teil präsentieren wir auf Sprachwiederverwendung basierende Technologien für drei verschiedene Retrievalaufgaben: (4.) Ein neues Modell zum medienübergreifenden, inhaltlichen Vergleich von Objekten aus dem Web. Das Modell basiert auf der Auswertung der zu einem Objekt vorliegenden Kommentare. In diesem Zusammenhang identifizieren wir Webkommentare als eine in der Forschung bislang vernachlässigte Informationsquelle und stellen die Grundlagen des Kommentarretrievals vor. (5.) Zwei neue Algorithmen zur Segmentierung von Websuchanfragen. Die Algorithmen nutzen Web n-Gramme sowie Wikipedia, um die Intention des Suchenden in einer Suchanfrage festzustellen. Darüber hinaus haben wir mittels Crowdsourcing ein neues Evaluierungskorpus erstellt, das zwei Größenordnungen größer ist als bisherige Korpora. (6.) Eine neuartige Suchmaschine, genannt NETSPEAK, die die Suche nach geläufigen Formulierungen ermöglicht. NETSPEAK indiziert das Web als Quelle für geläufige Sprache in der Form von n-Grammen und implementiert eine Wildcardsuche darauf. Im Folgenden werden die Beiträge genauer beschrieben und eine Auswahl an Ergebnissen präsentiert.

2 Erkennung von Textwiederverwendungen

Für ein Dokument, dessen Originalität in Frage steht, bestehe die Aufgabe darin, alle aus anderen Dokumenten wiederverwendeten Passagen zu identifizieren. Dazu gibt es drei Ansätze: (1.) Die Suche nach Originaldokumenten. (2.) Die Prüfung, ob das Dokument vom angeblichen Autor geschrieben wurde. (3.) Die Prüfung, ob alle Passagen des Dokuments vom gleichen Autor geschrieben wurden. Mit dem ersten Ansatz werden die Schritte, die der Autor des verdächtigen Dokuments zum Auffinden von Texten zur Wiederverwendung gehen musste, nachvollzogen. Die beiden anderen Ansätze basieren darauf, Autoren anhand ihres Schreibstils auseinander zu halten. Im Grunde lassen sich jedoch alle drei Ansätze darauf reduzieren, das verdächtige Dokument (passagenweise) mit anderen zu vergleichen, wobei nach einer „überraschenden" Gleichförmigkeit in Syntax oder Semantik gesucht wird. Bestimmte syntaktische Ähnlichkeiten zeigen gleiche Autoren an, wohingegen semantische Ähnlichkeiten ein mögliches Original entlarven. Im Idealfall würde das verdächtige Dokument mit allen anderen verfügbaren Dokument auf diese Weise verglichen, in der Praxis zwingt der nötige Aufwand aber zur Einschränkung auf wenige Kandidaten. Deshalb müssen diese Kandidaten mit Bedacht gewählt werden, um die Chance auf einen Treffer zu maximieren, sofern es etwas zu treffen gibt.

Mit Hilfe maßgeschneiderter Technologien ist es möglich, den Umfang solcher Untersuchungen erheblich zu steigern und sie zu beschleunigen. In [SMP07] haben wir zu diesem Zweck einen allgemeinen Retrievalprozess vorgeschlagen (siehe Abbildung 2), der die beiden oben diskutierten Schritte (Kandidatenretrieval und detaillierter Vergleich) um einen dritten ergänzt. Die wissensbasierte Nachverarbeitung dient dazu, falsch positive Erkennungen zu vermeiden, korrekte Zitate zu erkennen und Textmodifikationen, die durch den Autor des verdächtigen Dokuments eventuell gemacht wurden, zu visualisieren. All das soll die Bearbeitung solcher Fälle so einfach wie möglich gestalten.

Dieser Erkennungsprozess funktioniert ähnlich für alle in Abbildung 1 gezeigten Formen der Textwiederverwendung, aber es gibt keine allumfassende Lösung. Daher konzentrieren wir uns hier auf Zitate, Textbausteine und Übersetzungen. Ein weiterer Schwerpunkt ist die Evaluierung von Implementierungen dieses Prozesses und die hierfür nötigen Werkzeuge.

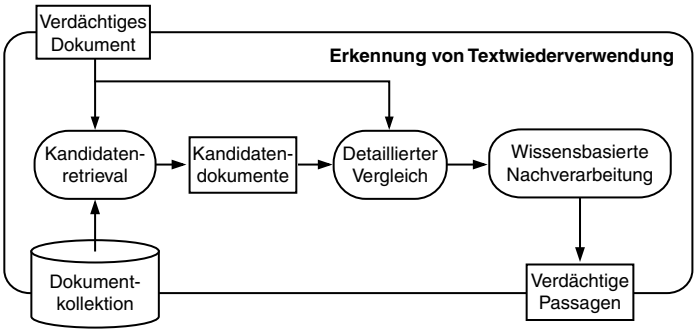


Abbildung 2: Retrievalprozess zur Erkennung von Textwiederverwendungen.

2.1 Fingerprinting zur Erkennung Fast Identischer Texte

Wörtliche Zitate und Textbausteine haben gemeinsam, dass der wiederverwendete Text sehr ähnlich zum jeweiligen Original ist, es aber dennoch Unterschiede geben kann. Beispielsweise werden so wiederverwendete Texte üblicherweise umformatiert, in Zitaten Kommentare eingefügt oder Auslassungen gemacht, in Textbausteinen variable Teile angepasst, und wenn die Wiederverwendung mit der Absicht zu plagiieren geschieht, kleine Modifikationen am Text vorgenommen, um diese Tatsache zu verschleiern. In der Literatur werden diese Formen der Textwiederverwendung daher auch mit dem Begriff „fast identische Texte" umschrieben. Algorithmen zur Erkennung dieser Formen der Wiederverwendung müssen daher robust gegenüber solchen Unterschieden sein.

Eine Klasse von Verfahren, die diese Eigenschaft mitbringt und gleichzeitig sublineare Retrievalzeit ermöglicht, heißt Fingerprinting. Fingerprinting basiert auf Hashing und berechnet für alle Dokumente einer Kollektion einen Fingerprint bestehend aus einer kleinen Zahl von Hashwerten. Anders als mit traditionellen Hashfunktionen werden die Hashwerte so kodiert, dass ähnliche Texte denselben Hashwert erhalten. Im Rahmen unserer Forschung haben wir erstmals das gemeinsame Schema herausgearbeitet, nach dem alle Fingerprintingverfahren arbeiten (siehe Abbildung 3). Kern dieser Verfahren ist die Einbettung hochdimensionaler Dokumentrepräsentationen in niedrigdimensionale Räume. Die anschließende Berechnung von Hashwerten ist gleichzusetzen mit einer Raumpartitionierung, die ähnlichen Dokumenten gleiche Raumabschnitte zuordnet. Zur Dimensionsreduktion in Linearzeit der Dokumentlänge werden Projektion und Einbettung eingesetzt.

Wir haben fünf Fingerprintingverfahren evaluiert und festgestellt, dass das projektionsbasierte Supershingling am besten abschneidet. Ein Problem bei der Evaluierung ist das Fehlen eines Referenzkorpus. Wir schlagen hierfür die Revisionshistorien von Wikipedia-Artikeln vor, die zahlreiche sehr ähnliche Texte aufweisen. Ein überraschendes Ergebnis ist die Tatsache, dass die niedrigdimensionalen Vektoren, die das Fuzzy-Fingerprinting-Verfahren durch Einbettung erzeugt, in Standardretrievalexperimenten ähnlich gut abschneiden wie hochdimensionale Vektorraummodelle: Unabhängig vom Fingerprinting erlaubt dieses Verfahren also die Erzeugung sehr kompakter Dokumentrepräsentationen.

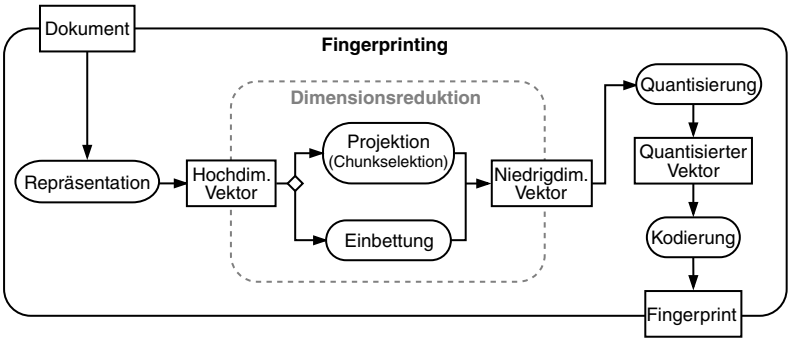


Abbildung 3: Prozess zur Fingerprinterzeugung, der allen Fingerprintingverfahren zugrunde liegt.

2.2 Erkennung von Sprachübergreifender Textwiederverwendung

Übersetzungen sind eine große Herausforderung für die automatische Erkennung von Textwiederverwendung. Das gilt insbesondere für den detaillierten Vergleich eines verdächtigen- mit einem Kandidatendokument des in Abbildung 2 dargestellten Erkennungsprozesses. Anders als innerhalb einer Sprache, kann man sich hier nicht auf syntaktische Überlappungen verlassen. Die Semantik eines Textes zu modellieren und sie automatisch von einer in eine andere Sprache zu überführen, erfordert die Zusammenstellung von Übersetzungswörterbüchern oder parallelen Korpora von Übersetzungen, um damit einen maschinellen Übersetzer zu trainieren. Solche Ressourcen sind schwer zu beschaffen und maschinelles Übersetzen für sich ist ein noch ungelöstes Forschungsproblem.

Wir schlagen das Modell CL-ESA zum sprachübergreifenden Textvergleich vor. Es kommt ohne maschinelle Übersetzung aus und beruht einzig auf *vergleichbaren* Korpora. Das sind Sammlungen von Dokumenten, so dass zu einem Thema in zwei oder mehr Sprachen ein Dokument vorliegt. Die Dokumente können unabhängig voneinander entstanden sein, was ihre Beschaffung bedeutend erleichtert. Die Wikipedia ist zum Beispiel ein vergleichbares Korpus. Mit Hilfe des Modells können zwei verschiedensprachige Dokumente wie folgt verglichen werden: Jedes Dokument wird zunächst mit den Dokumenten des Korpus verglichen, die in seiner Sprache vorliegen, und die Ähnlichkeitswerte aufgezeichnet. Wenn die Ähnlichkeitswerte des einen Dokuments mit denen des anderen übereinstimmen, so sind sich beide sehr ähnlich. Der Grad der Übereinstimmung über alle vergleichbaren Dokumente des Korpus erlaubt eine stufenlose sprachübergreifende Ähnlichkeitsmessung.

CL-ESA wurde mit zwei herkömmlichen Modellen auf Paarungen der Sprachen Englisch, Deutsch, Spanisch, Französisch, Niederländisch und Polnisch verglichen. Mehr als 100 Mio. Vergleiche wurden berechnet und CL-ESA hat sich dabei als konkurrenzfähig erwiesen. Abbildung 4 zeigt das Verhalten von CL-ESA abhängig von seiner Dimensionalität (Zahl vergleichbarer Dokumente). Beim Ranking vergleichbarer Dokumente kann CL-ESA nahezu perfekten Recall erreichen. Außerdem können auch mehrere syntaktisch unabhängige Sprachen gleichzeitig repräsentiert und untereinander verglichen werden.

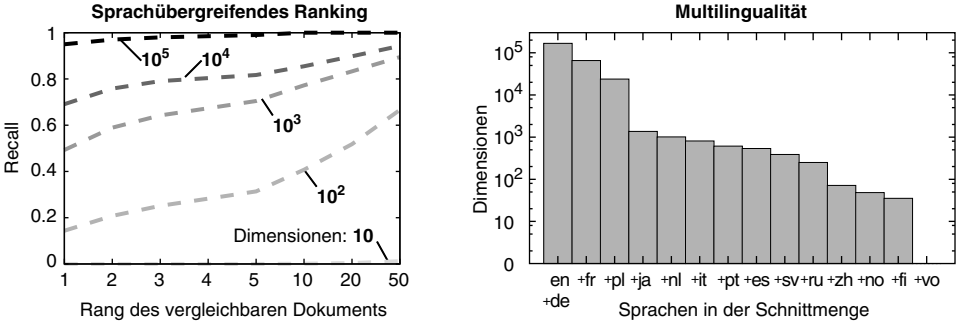


Abbildung 4: Links: CL-ESAs dimensionsabhängige Effektivität im sprachübergreifenden Ranking; Rechts: CL-ESAs Dimensionalität, je mehr Sprachen der Wikipedia gleichzeitig verwendet werden.

2.3 Evaluierung von Verfahren zur Erkennung von Plagiaten

Das Fehlen einer standardisierten Evaluierungsumgebung ist ein schwerwiegendes Problem in empirischer Forschung, da Ergebnisse nicht über Papiere hinweg verglichen oder reproduziert werden können. Wir haben die verfügbare Literatur (205 Papiere) zur Erkennung von Textwiederverwendung und zur Plagiatserkennung in dieser Hinsicht untersucht und festgestellt dass 46% keine Vergleichsverfahren heranziehen, 80% kein freies Korpus verwenden, 77% mit weniger als 1000 Dokumenten evaluieren und insgesamt wenig aussagekräftige Erfolgsmaße eingesetzt werden.

Daher haben wir eine von Grund auf neue Evaluierungsumgebung erschaffen, die aus großen Korpora von Plagiaten und neu entwickelten Erfolgsmaßen besteht. Da reale Plagiate schwer zu bekommen sind,[1] wurden Plagiate sowohl künstlich als auch manuell erzeugt. Das Korpus wurde in drei Versionen erstellt,[2] die jede mehr als 25 000 Dokumente mit jeweils zwischen 60- und 90 000 künstlich generierten Plagiaten enthalten. Es wurden eine Reihe von Parametern variiert und Heuristiken eingesetzt, die das Vorgehen eines Plagiators nachbilden und Textmodifikationen vornehmen, die die automatische Erkennung erschweren. Es wurden auch Übersetzungsplagiate von deutschen und spanischen Texten ins Englische generiert. Zusätzlich wurden erstmals mit Hilfe von Amazons Mechanical Turk über 4000 Plagiatsfälle von mehr als 900 Teilnehmern manuell erzeugt.

Weiterhin haben wir vier Erfolgsmaße für Verfahren zur Plagiatserkennung erforscht und entwickelt, die bisher unberücksichtigte Erfolgsaspekte messen: anstatt auf Dokumentebene messen sie auf Passagenebene und berücksichtigen sowohl das verdächtige als auch das Originaldokument sowie die Eindeutigkeit der Erkennung. Gemessen werden Precision, Recall und Granularität der Erkennung plagiierter Passagen. Letzteres ist die durchschnittliche Häufigkeit mit der ein- und derselbe Fall erkannt wird. Das Maß Plagdet kombiniert diese drei, um die Bildung einer Rangfolge von Verfahren zu ermöglichen.

Sowohl das Korpus als auch die Maße wurden erfolgreich im Rahmen dreier internationaler Wettbewerbe zur Plagiatserkennung eingesetzt.[3] Insgesamt sind 32 Forschergruppen aus aller Welt mit ihren Algorithmen in den Disziplinen externe und intrinsische Plagiatserkennung angetreten. Externe Erkennung meint die Suche nach Originalen für ein gegebenes verdächtiges Dokument, intrinsische Erkennung die oben erwähnte Möglichkeit, Plagiate anhand von Schreibstiländerungen im verdächtigen Dokument zu identifizieren. Abbildung 5 zeigt die in den Wettbewerben erzielten Ergebnisse. Keiner der Algorithmen hat alle in den Korpora versteckten Plagiate erkannt und nur wenige erreichen unter allen Maßen gute Bewertungen. Precision erscheint weniger schwer zu erreichen als Recall. Die Werte sind über die Jahre hinweg nicht direkt vergleichbar, da die Korpora zunehmend schwerer konfiguriert wurden. In 2010 wurden externe und intrinsische Plagiatserkennung als gemeinsame Aufgabe abgehalten. Viele neue Ideen wurden im Rahmen der Wettbewerbe getestet und eine deutliche Entwicklung war zu beobachten. Dennoch stecken Algorithmen zur Erkennung von Textwiederverwendung immernoch in den Kinderschuhen.

[1] Es gibt außerdem rechtliche und ethische Probleme, reale Plagiate als Teil von Korpora zu veröffentlichen. Die kürzlich bekannt gewordenen Fälle unter deutschen Politikern stellen keinen repräsentativen Ausschnitt dar.

[2] Das „PAN Plagiarism Corpus" ist frei verfügbar unter: `http://www.webis.de/research/corpora`

[3] Die Wettbewerbe haben im Rahmen unserer Workshopreihe PAN stattgefunden: `http://pan.webis.de`

Externe Plagiatserkennung

	gro	kas	bas	pal	zec	sch	per	val	mal	all
PAN 2009	0.70	0.61	0.60	0.30	0.19	0.14	0.06	0.03	0.02	0.01
	0.74	0.56	0.67	0.67	0.61	0.75	0.66	0.01	0.03	0.37
	0.66	0.70	0.63	0.44	0.37	0.53	0.10	0.46	0.60	0.01
	1.00	1.02	1.11	2.33	4.44	19.43	5.40	1.01	6.78	2.83

	kas	zou	muh	gro	obe	rod	per	pal	sob	got	mic	cos	naw	gup	van	sua	alz	ift
PAN 2010	0.80	0.71	0.69	0.62	0.61	0.59	0.52	0.51	0.44	0.26	0.22	0.21	0.21	0.20	0.14	0.06	0.02	0.00
	0.94	0.91	0.84	0.91	0.85	0.85	0.73	0.78	0.96	0.51	0.93	0.18	0.40	0.50	0.91	0.13	0.35	0.60
	0.69	0.63	0.71	0.48	0.48	0.45	0.41	0.39	0.29	0.32	0.24	0.30	0.17	0.14	0.26	0.07	0.05	0.00
	1.00	1.07	1.15	1.02	1.01	1.00	1.00	1.02	1.01	1.87	2.23	1.07	1.21	1.15	6.78	2.24	17.31	8.68

	grm	gro	obe	coo	rod	rao	pal	naw	gho
PAN 2011	0.56	0.42	0.35	0.25	0.23	0.20	0.19	0.08	0.00
	0.94	0.81	0.91	0.71	0.85	0.45	0.44	0.28	0.01
	0.40	0.34	0.23	0.15	0.16	0.16	0.14	0.09	0.00
	1.00	1.22	1.06	1.01	1.23	1.29	1.17	2.18	2.00

Intrinsische Plagiatserkennung

	sta	hag	zec	sea
Plagdet	0.25	0.20	0.18	0.12
Precision	0.23	0.11	0.20	0.10
Recall	0.46	0.94	0.27	0.56
Granularität	1.38	1.00	1.45	1.70

	obe	sta	kes	aki	rao
Plagdet	0.33	0.19	0.17	0.08	0.07
Precision	0.31	0.14	0.11	0.07	0.08
Recall	0.34	0.41	0.43	0.13	0.11
Granularität	1.00	1.21	1.03	1.05	1.48

Abbildung 5: Ergebnisse dreier Wettbewerbe zur Erkennung von Plagiaten (PAN 2009-2011). Pro Tabelle entspricht jede Spalte einem Teilnehmer. Die Spalten sind nach dem erzielten Plagdet-Wert sortiert. Die Zellenschattierung visualisiert die erzielten Erfolge: Je dunkler, desto besser.

3 Retrievalaufgaben Mittels Sprachwiederverwendung Lösen

Textwiederverwendung bezeichnet die Wiederverwendung einzelner Texte, wohingegen die Wiederverwendung großer Mengen von Texten als Sprachwiederverwendung bezeichnet wird. Ein Übersetzer verwendet Texte beispielsweise einzeln wieder, da sie zumeist unabhängig von anderen übersetzt werden. Ein Linguist hingegen verwendet viele Texte wieder, um anhand der Vorkommen eines Wortes all seine Bedeutungen zu erfassen. Die Wiederverwendung eines Textes geschieht also linear und unabhängig von anderen Texten. Die Wiederverwendung von Sprache dagegen geschieht durch das Ausnutzen bestimmter Texteigenschaften, die viele Texte miteinander teilen, um anhand ihrer Ausprägungen eine Aufgabe zu erfüllen. Es können jedoch nicht bloß linguistische Aufgaben durch die (automatische) Wiederverwendung von Sprache gelöst werden, sondern ungezählte andere: Prominente Beispiele im Information Retrieval sind Wikipedia und Web-n-Gramme. Die Wikipedia ist inzwischen eine weit verbreitete Informationsquelle, nicht nur für Menschen, sondern auch zur Informierung wissensbasierter Modelle und Algorithmen (ein umfassender Überblick ist in [MMLW09] zu finden). Ähnlich erfolgreich werden Web-n-Gramme (Wortfolgen der Länge n und ihre Häufigkeit im Web) eingesetzt. Die Zahl der Aufgaben, die durch Sprachwiederverwendung ganz oder teilweise gelöst werden können ist nicht ersichtlich, da Texte im Web in unzähligen Relationen mit anderen Dingen stehen. Im Rahmen unserer Forschung haben wir zwei neue Ansätze erforscht, um mit Hilfe von Sprachwiederverwendung Aufgaben des Information Retrieval zu lösen: Es handelt sich zum Einen um ein Modell zum inhaltlichen Vergleich von Texten und Objekten aller Mediengattungen und zum Anderen um Algorithmen zum Einfügen intendierter Anführungszeichen in Websuchanfragen. Weiterhin haben wir einen Webdienst entwickelt, der es erlaubt, alle Texte im Web als Formulierungshilfe zu verwenden.

3.1 Ein Modell zum Medienübergreifenden Vergleich von Objekten

Das Web besteht nicht bloß aus Texten, sondern aus Medienobjekten aller Art und selbstverständlich werden auch diese von den Nutzern des Webs gesucht. Suchmaschinen stehen daher vor dem Problem, textuelle Suchanfragen mit Objekten anderer Mediengattungen zu vergleichen. Traditionelle Ansätze hierfür basieren darauf, Korpora bestehend aus Multimediaobjekten auszuwerten, die von Hand mit Metainformationen über ihren Inhalt ausgezeichnet wurden. Mit diesen Daten werden maschinelle Lernverfahren trainiert, die eine Abbildung der Inhaltsangaben auf medienspezifische Charakteristiken erlernen sollen, um so Suchanfragen beantworten zu können. Korpora dieser Machart sind gegenwärtig nur in kleinem Rahmen verfügbar, was die Forschung an dieser Aufgabe behindert.

Wir haben Webkommentare zu Multimediaobjekten als eine weitgehend vernachlässigte Quelle von Informationen identifiziert. Darauf aufbauend schlagen wir ein neues Modell zum medienübergreifenden Vergleich von Objekten aller Mediengattungen vor, das auf Webkommentaren beruht. Das Modell verwendet alle Kommentare zu einem Objekt als Ersatz für eine inhaltliche Beschreibung wieder, um so mit textbasierten Standardmodellen die Kommentare zweier Objekte insgesamt miteinander zu vergleichen. Wird eine inhaltliche Übereinstimmung der Kommentare gemessen, so liegt mit hoher Wahrscheinlichkeit auch eine inhaltliche Übereinstimmung der kommentierten Objekte vor.

Das Modell wurde evaluiert, indem 6000 YouTube-Videos auf diese Weise mit rund 18 000 Artikeln der Nachrichtenseite Slashdot verglichen wurden. Aus den sich so ergebenen Paarungen wurden die 100 manuell ausgewertet, deren gemessene Ähnlichkeit am höchsten war und festgestellt, dass 91 mindestens ein verwandtes und 36 dasselbe Thema aufwiesen (siehe Tabelle 1). Außerdem wurden stichprobenartig weitere 150 Paarungen mit geringeren Ähnlichkeiten ausgewertet und festgestellt, dass thematische Übereinstimmungen ab einer gemessenen Kommentarähnlichkeit von 0.4 im Vektorraummodell gehäuft auftreten. Diese Ergebnisse lassen den Schluss zu, dass die Annahme, auf der unser Modell fußt, zutrifft und dass es zum medienübergreifenden Vergleich von Objekten eingesetzt werden kann. Die einzige Einschränkung dabei ist, dass mindestens rund 100 Kommentare vorhanden sein müssen.

Tabelle 1: Themenvergleich der 100 als am ähnlichsten erkannten Video-Artikel Paare.

Themenvergleich	**Anteil**	**Ähnlichkeit**				**∅ Zahl der Kommentare**	
		min.	∅	max.	σ	Slashdot	YouTube
gleich	36%	0.71	0.78	0.91	0.06	53	927
verwandt	55%	0.71	0.76	0.91	0.04	81	683
ungleich	9%	0.72	0.78	0.87	0.05	104	872
Σ	100%	0.71	0.77	0.91	0.05	74	790

3.2 Algorithmen zur Segmentierung von Suchanfragen

Die vorrangige Art von Suchanfragen an Websuchmaschinen sind Schlüsselwortanfragen. Obwohl die meisten Suchmaschinen auch fortgeschrittenere Anfrageoperatoren und -facetten anbieten, mit denen ein Nutzer das Gesuchte klarer umschreiben kann, werden diese kaum benutzt: Nur 1.12% der Suchanfragen enthalten solche Operatoren [WM07]. Eine der Möglichkeiten besteht darin, Anführungszeichen in Anfragen einzufügen, um so Phrasen als unteilbar zu markieren. Die Suchmaschine kann mit dieser Information die Precision der Suchergebnisse erhöhen, da Dokumente, die die Phrasen nicht enthalten, verworfen werden können. Da die überwältigende Mehrheit der Nutzer einer Suchmaschine, diese Option nicht nutzt, wird ein erhebliches Potential verschenkt.

Wir stellen einen neuen Algorithmus vor, der automatisch Anfragen segmentiert (also Anführungszeichen an geeignete Stellen einer Suchanfrage einfügt). Der Algorithmus basiert auf der Annahme, dass Phrasen, die hinreichend häufig im Web vorkommen, wichtige Konzepte sind, die es lohnt in Anführungszeichen zu setzen. Es werden zunächst alle Segmentierungen der Anfrage aufgezählt und dann jede Segmentierung gewichtet: Das Gewicht errechnet sich aus der Länge aller enthaltenen Phrasen und ihrer Häufigkeit im Web. Damit längere Phrasen eine Chance gegenüber den tendenziell häufiger vorkommenden, kürzeren haben, werden die Gewichte geeignet normalisiert. Am Ende wird die Segmentierung gewählt, deren Gewicht am höchsten ist. Um die Häufigkeit einer Phrase im Web effizient zu ermitteln, verwenden wir das Google-n-Gramm-Korpus, das die Häufigkeit aller im Jahr 2006 im Web vorkommenden Phrasen der Länge $n \leq 5$ Wörter enthält [BF06].

In einer groß angelegten Evaluierung haben wir unseren Algorithmus mit acht weiteren aus der Literatur verglichen. Zu diesem Zweck haben wir ein bislang häufig verwendetes Korpus zur Anfragesegmentierung verwendet, das aus 500 Anfragen besteht. Da dieses Korpus einige Konstruktionsschwächen aufweist und nicht repräsentativ ist, haben wir ein neues Korpus mit mehr als 50 000 Anfragen erstellt, das die Längen- und Häufigkeitsverteilung echter Anfragelogdateien repräsentiert. Zu jeder Anfrage wurden via Amazons Mechanical Turk zehn Personen befragt, an welchen Stellen sie Anführungszeichen einsetzen würden. Auf beiden Korpora übertrifft unser Algorithmus die anderen: Er fügt am ehesten Anführungszeichen dort ein, wo auch Menschen es tun würden, und ist gleichzeitig bedeutend einfacher zu realisieren als die bisherigen Verfahren.

3.3 Ein Werkzeug zur Schreibunterstützung

Die meiste Zeit beim Schreiben verbringt man damit, herauszufinden, wie man etwas schreiben möchte, nicht was. Gute Formulierungen für einen Sachverhalt zu finden, entscheidet darüber, wie gut die Zielgruppe eines Textes ihn versteht. Gerade deutsche Wissenschaftler stehen diesbezüglich vor dem Problem, dass der Diskurs vieler Disziplinen in Englisch stattfindet. Die meisten Deutschen verfügen aber nicht über das Vokabular oder das Sprachgefühl eines Englisch-Muttersprachlers. Die Suche nach Worten und Formulierungen wurde allerdings bis jetzt nicht hinreichend unterstützt.

Wir haben NETSPEAK entwickelt, eine Suchmaschine für geläufige englische Formulierungen.[4] Netspeak indiziert das Web in Form von n-Grammen und ermöglicht eine Wildcardsuche darauf. Suchanfragen bestehen aus kurzen Formulierungen, in die der Nutzer Wildcards dort eingefügt, wo Unsicherheit über die üblicherweise verwendeten Wörter besteht. Die zur Anfrage passenden n-Gramme werden gesucht und die Ergebnisse nach ihrer Häufigkeit im Web sortiert. Auf diese Weise können geläufige von ungebräuchlichen Formulierungen unterschieden werden. Die NETSPEAK zu Grunde liegende Hypothese ist die, dass neben der Korrektheit eines Textes auch die Verwendung geläufiger Formulierungen wichtig ist. Das bringt den Vorteil leichter Verständlichkeit mit sich und schränkt das Fehlerpotenzial ein wenig ein, gerade beim Schreiben in einer fremden Sprache.

4 Ausblick

Wie sähe die Welt aus, wenn alle Textwiederverwendungen im Web offen zutage lägen? Plagiarismus, der „böse Zwilling" der Textwiederverwendung, wäre sinnlos. Doch darüber hinaus würde ein Netzwerk zwischen Webdokumenten ersichtlich, das, anders als das Hyperlinknetzwerk, den Einfluss eines Textes auf andere sichtbar machen würde. Ein solches Netzwerk könnte als weiteres Relevanzsignal in der Websuche dienen, aber auch dazu, Reputation und Honorare zum Urheber eines Textes weiterzuleiten. Im gesamten Web wird dies vermutlich kaum realisierbar sein, in spezifischen Genres, wie der Wissenschaft, ist das aber durchaus denkbar.

Literatur

[BF06] Thorsten Brants und Alex Franz. Web 1T 5-gram Version 1. Linguistic Data Consortium LDC2006T13, 2006.

[MMLW09] Olena Medelyan, David Milne, Catherine Legg und Ian H. Witten. Mining Meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.*, 67:716–754, 2009.

[Pot11] Martin Potthast. *Technologies for Reusing Text from the Web*. Dissertation, Fakultät Medien, Bauhaus-Universität Weimar, 2011.

[SMP07] Benno Stein, Sven Meyer zu Eißen und Martin Potthast. Strategies for Retrieving Plagiarized Documents. In *Proceedings of SIGIR 2007*.

[WM07] Ryen W. White und Dan Morris. Investigating the Querying and Browsing Behavior of Advanced Search Engine Users. In *Proceedings of SIGIR 2007*

Martin Potthast wurde im April 1981 in Steinheim geboren und vollendete seine schulische Laufbahn im Jahr 2000 am Gymnasium St. Xaver in Bad Driburg. Nach dem Zivildienst nahm er 2001 das Studium der Informtik an der Universität Paderborn auf und erhielt Anfang 2005 den Bachelortitel. Mitte 2006 vollendete er das Studium als Diplom-Infomatiker. Seitdem hat er am Lehrstuhl für Content Managment und Web-Technologien der Bauhaus-Universität Weimar promoviert und seine Dissertation Ende 2011 verteidigt.

[4] NETSPEAK ist frei verfügbar unter `http://www.netspeak.org`.

Lernen mit wenigen Beispielen für die visuelle Objekterkennung

Erik Rodner
Lehrstuhl Digitale Bildverarbeitung
Friedrich-Schiller Universität Jena
Erik.Rodner@uni-jena.de

Abstract: Das maschinelle Lernen aus wenigen Beispielen ist ein wichtiges und entscheidendes Problem bei vielen visuellen Erkennungsaufgaben, besonders in industriellen Anwendungen. Im Gegensatz zum Menschen benötigen viele aktuelle Verfahren meistens Hunderte von beschrifteten Beispielbildern. Die Dissertation "*Learning with Few Examples for Visual Recognition Problems*" beschäftigt sich mit diesem Problem und stellt Lösungsmöglichkeiten vor, welche sich auf die Verwendung zweier Konzepte stützen: *Lerntransfer* und *Ein-Klassen-Klassifikation*. Das folgende Dokument bietet eine Zusammenfassung der Ergebnisse der Dissertation.

1 Einleitung

Die Dissertation beschäftigt sich mit Verfahren der visuellen Objekterkennung, welche das Ziel verfolgen, automatisch semantische Informationen aus Bildern zu extrahieren. Dabei sollen zum Beispiel Objekte bekannter Kategorien in einem Bild erkannt und lokalisiert werden. Weiterhin soll die Maschine die Erscheinungsformen einer Objektkategorie selbstständig aus beschrifteten Beispielbildern lernen. Die Fähigkeit, diese Art der automatischen Bildanalyse durchzuführen, ist sowohl in der Robotik als auch bei zahlreichen Anwendungen zwingend notwendig. In den letzten Jahren lässt sich ein drastischer Anstieg an komplexen industriellen Problemstellungen verzeichnen, welche ohne Verfahren des maschinellen Lernens nicht realisierbar sind. Als prägnantes Beispiel sei hier die Fußgängerdetektion [DWSP11] und zahlreiche andere Fahrerassistenzsysteme aufgeführt. Ein Hauptproblem ist die Verfügbarkeit von repräsentativen Lernbeispielen, da die Beschriftung bei vielen Anwendungen zeit- und kostenintensiv ist. *Ziel der Dissertation ist es daher die Anzahl der notwendigen Lernbeispiele durch spezielle Verfahren des maschinellen Lernens zu reduzieren.*

In Abbildung 1 sind die drei Hauptabstraktionsebenen der visuellen Objekterkennung dargestellt. Diese richten sich nach der Art der gewünschten Ausgabe des Systems und des Detaillierungsgrades. Während der Dissertation wurden alle drei Bereiche betrachtet [RD10, FRD10]. Die entwickelten Verfahren sind allgemein für viele Aufgaben des maschinellen Lernens geeignet.

Die vorliegende Arbeit ist wie folgt aufgebaut: zunächst wird allgemein auf die Schwierigkeiten der visuellen Objekterkennung und des Lernens aus wenigen Beispielen einge-

Abbildung 1: Unterschiedliche Aufgabenstellungen der visuellen Objekterkennung: Bildkategorisierung [RD10] (Beschriftung des gesamten Bildes), Objektlokalisation (Beschriftete umschreibende Rechtecke) und Semantische Segmentierung [FRD10] (Beschriftung jedes einzelnen Bildpixels).

gangen. Danach wird das Konzept des Lerntransfers und die in der Dissertation entwickelten Verfahren skizziert. Abschnitt 5 liefert einen Überblick über die Algorithmen der Ein-Klassen-Klassifikation, welche den zweiten Schwerpunkt der Dissertation darstellen. Realisierte Anwendungen werden in Abschnitt 6 kurz zusammengefasst. Abschließend folgt eine Zusammenfassung der Resultate.

Die Zusammenfassung beschreibt nur die Aspekte der Klassifikation. Für eine Beschreibung der Merkmalsauswahl für einzelne Anwendungen sei auf die Dissertation verwiesen.

2 Herausforderungen bei der visuellen Objekterkennung

Bei den Beispielbildern in Abbildung 1 lassen sich die Schwierigkeiten und die Komplexität des automatischen Lernens von Objektkategorien gut erkennen und folgendermaßen zusammenfassen:

1. Die Erscheinungen einzelner Objektkategorien variieren sehr stark durch unterschiedliche Rotationen, Skalierungen, andere Perspektiven, nicht-starre Deformationen, farbliche Gestaltung, Unterkategorien anderer Ausprägung (z.B. verschiedene Arten von Vegetation in den Bildern auf der rechten Seite).

2. Bestimmte Kategorien sind ähnlich zueinander und lassen sich schwierig voneinander trennen (z.B. Kategorie Fenster und Tür auf der rechten Seite).

3. Objekte können sich gegenseitig verdecken (zu sehen in der Straßenszene unten links) und 3D-Informationen sind bei Einzelaufnahmen nicht direkt verfügbar.

4. Die Darstellung von Kategorien und Bildelementen, welche nicht erkannt werden sollen, erschwert die Erkennung zusätzlich (z.B. die Fahrradfahrer im Bild in der Mitte).

Vor allem die große Variabilität der Objekte einer Kategorie lässt sich meist nur durch die Angabe vieler beschrifteter und repräsentativer Beispiele erlernen. So werden zum Beispiel zum Anlernen von Fußgängerdetektoren oft mehrere tausende Beispiele benötigt [DWSP11].

Das Problem bei wenigen Lerndaten manifestiert sich auch als schlecht gestelltes Optimierungsproblem, welches im Folgenden näher erläutert werden soll: Das Lernen von visuellen Aufgaben kann mathematisch als das Schätzen einer Abbildung $f : \mathcal{X} \rightarrow \mathcal{Y}$ von der Menge $\mathcal{X}$ aller Bilder in die Menge $\mathcal{Y}$ aller möglichen Beschriftungen angesehen werden. Die Funktion f wird im Falle der Klassifikation ($\mathcal{Y}$ ist diskret, z.B. $\mathcal{Y} \in \{-1, 1\}$) als Entscheidungsfunktion bezeichnet. Die Schätzung oder das Lernen basiert dabei auf einem gegebenen Lerndatensatz $\mathcal{D}$, welcher n Bilder $\boldsymbol{x}_i \in \mathcal{X}$ und deren Beschriftungen $y_i \in \mathcal{Y}$ enthält. Wird die Aufgabe als reines Schätzproblem betrachtet, führt dies unmittelbar zum entscheidenden Dilemma der Objekterkennung: auf der einen Seite ist die Menge aller möglichen Funktionen f und der Eingaberaum $\mathcal{X}$ selbst hochdimensional, auf der anderen Seite existieren nur wenige gegebene Datenpunkte. Ohne weitere Zusatzinformationen ist diese Situation vergleichbar mit der Regression einer komplizierten Funktion (z.B. Polynom hohen Grades) mit einer geringen Anzahl von Abtastwerten. Genau die Einbindung von zusätzlichem Wissen durch einen sogenannten Lerntransfer (englisch: *knowledge transfer* oder *transfer learning*) ist das Schlüsselkonzept, welches das Lernen aus wenigen Beispielen ermöglicht.

3 Lerntransfer

Betrachtet man die menschlichen Erkennungsleistungen, so ist es anscheinend trotz der beschriebenen Schwierigkeiten beim Lernen visueller Aufgaben als Mensch möglich, neue Objektkategorien oft mit nur einem Beispiel robust zu erlernen [Bie87]. Welche Zusatzinformationen werden aber vom menschlichen Erkennungssystem ausgenutzt um dies zu ermöglichen? Als häufiger Punkt wird die automatische Verwendung von Vorwissen ähnlicher Aufgabenstellungen für das Erlernen einer neuen Aufgabe angeführt. Intuitiv veranschaulicht dies der Lerntransfer beim Erlernen von Sprachen: der Aufwand, eine neue Sprache zu erlernen, ist erheblich geringer, wenn schon verwandte und ähnliche Sprachen bekannt sind, z.B. Französisch und Spanisch. In Abbildung 2 wird dies für die visuelle Objekterkennung anhand der Bildkategorisierung mit Tierklassen illustriert. Bei diesen Tierkategorien existieren viele Gemeinsamkeiten, wie etwa ähnliche Texturmerkmale (Zebra) oder eine ähnliche Konstellation von Objektteilen (Nashorn, Zebra). Die Variation dieser visuellen Komponenten kann daher robust von den verwandten Klassifikationsaufgaben erlernt werden.

In der Dissertation werden mehrere Verfahren vorgestellt, die entwickelt wurden um dieses Konzept beim maschinellen Lernen umzusetzen. Dabei werden unterschiedliche Wissensrepräsentationen von einer Klassifikationsaufgabe (*Unterstützungsaufgabe*) auf eine neue Aufgabe (*Zielaufgabe*) übertragen.

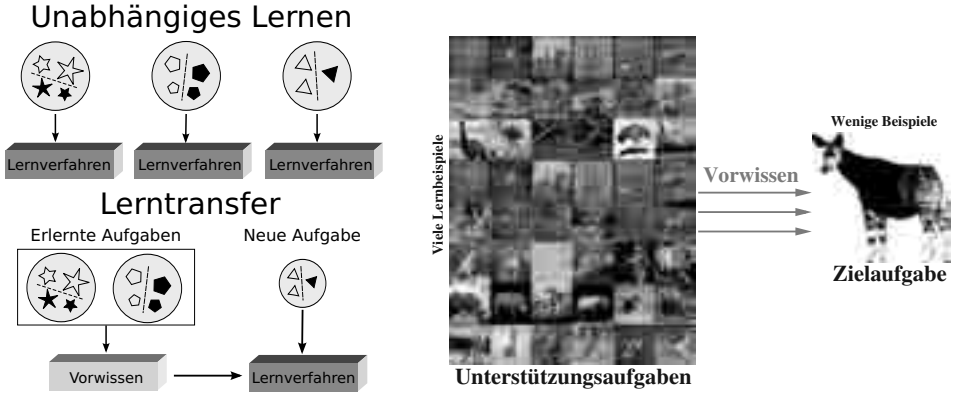


Abbildung 2: (Links) Schematischer Vergleich zwischen unabhängigen Lernen und *Lerntransfer*, (Rechts) Darstellung des Konzeptes des Lerntransfers bei der visuellen Objekterkennung: Durch gezielte Ausnutzung der Ähnlichkeit der neuen Kategorie *Okapi* zu bekannten Objektkategorien mit vielen Lernbeispielen ist eine Reduzierung der Anzahl der notwendigen Lernbeispiele möglich.

4 Adaptiver Lerntransfer mit Gauß-Prozess-Klassifikatoren

Im Rahmen der Dissertation wurde ein Verfahren entwickelt welches auf Kernfunktionen basiert und daher einen nicht-parametrischen Wissenstransfer ermöglicht. Ein besonderer Vorteil dieser Methode ist es, Klassifikationsaufgaben, von denen Wissen transferiert werden soll, automatisch auszuwählen und den Einfluss des Transfers zu adaptieren. Dies wird durch eine effiziente Modellselektion und der Verwendung von semantischen Ähnlichkeiten zwischen Kategoriebegriffen ermöglicht (siehe Übersicht in Abbildung 3). Zunächst wird ein kurzer Überblick über die Gauß-Prozess-Regression und Klassifikation gegeben, da diese ein methodisches Kernelement späterer Algorithmen ist.

Gauß-Prozess-Regression und Klassifikation Viele Klassifikationsverfahren basieren auf einer Parametrisierung $f(\boldsymbol{x};\boldsymbol{\theta})$ der Entscheidungsfunktion. Ausgehend von den Lerndaten $\mathcal{D}$ wird ein Parameter $\boldsymbol{\theta}$ bestimmt, welcher die A-posteriori-Wahrscheinlichkeit $p(\boldsymbol{\theta} \mid \mathcal{D})$ maximiert (vgl. MAP-Schätzung). Die Gauß-Prozess-Regression und Klassifikation kann hingegen anders motiviert werden. Grundidee ist die Betrachtung der Funktion f direkt als Zufallsvariable und die Annahme, dass f gemäß eines Gauß-Prozesses (GP) verteilt ist, d.h. $f \sim \mathcal{GP}(0, \mathcal{K})$. Die Funktion $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ ist die Kovarianzfunktion des Gauß-Prozesses und modelliert die Korrelation $\mathbb{E}\left[f(\boldsymbol{x})f(\boldsymbol{x}')\right]$ von zwei Ausgaben anhand der Ähnlichkeit von $\boldsymbol{x}$ und $\boldsymbol{x}'$.

Im Kontext des maschinellen Lernens wird $\mathcal{K}$ oft als Kern(el)funktion bezeichnet und es lassen sich etliche Formen dieser Funktion zur Modellierung heranziehen. Beispielhaft sei hier die Gaußkernfunktion angeführt:

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right) \quad . \tag{1}$$

An dieser Funktion lässt sich gut erkennen, dass Beispiele mit einer geringen Distanz im Eingaberaum auch zu einer hohen Korrelation der entsprechenden Funktionswerte führen.

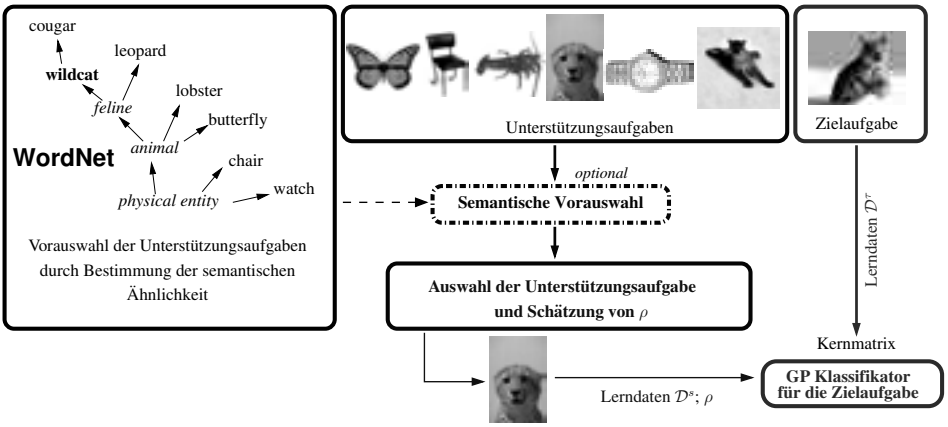


Abbildung 3: Schema des adaptiven Lerntransfers mit Gauß-Prozessen: Aus einer Menge von Klassifikationsaufgaben wird durch einen zweistufigen Prozess eine Unterstützungsaufgabe ausgewählt um das Lernen einer Zielaufgabe anzureichern. Die Auswahl erfolgt auf der Basis von semantischen Ähnlichkeiten und visuellen Informationen.

Tatsächlich ist dies eine der notwendigen Hauptannahmen des maschinellen Lernens: ähnliche Eingaben sollten zu ähnlichen Ausgaben führen. Mit weiteren Annahmen kann die A-posteriori-Verteilung $p(y_* \mid \boldsymbol{x}_*, \mathcal{D})$ der Ausgabe y_* eines neuen Beispiels $\boldsymbol{x}_*$ hergeleitet werden [RW05]. Im Rahmen dieser Zusammenfassung soll auf mathematische Details verzichtet und nur die Gleichung für den Schätzwert angegeben werden:

$$\mathbb{E}(y_* \mid \boldsymbol{x}_*, \mathcal{D}) = \boldsymbol{k}_*^T \left(\mathbf{K} + \sigma_\varepsilon^2 \cdot \mathbf{I}\right)^{-1} \boldsymbol{y} \ . \tag{2}$$

Bei diesem Modell wurde angenommen, dass die gegebenen Ausgaben zusätzlich mit additivem, normalverteilten Rauschen $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ gestört sind. Die Ausgaben y_i des Lerndatensatzes sind im Vektor $\boldsymbol{y} \in \mathcal{Y}^n$ zusammengefasst, $\mathbf{K} \in \mathbb{R}^{n \times n}$ bezeichnet die Kernmatrix, welche die paarweisen Werte der Kernfunktion von den Lerndaten beinhaltet, und im Vektor $\boldsymbol{k}_* \in \mathbb{R}^n$ sind die Werte der Kernfunktion der Lerndaten mit dem neuen Beispiel $\boldsymbol{x}_*$ gespeichert. Die Annahme von normalverteilten Rauschen ist natürlich eine sehr restriktive Annahme, gerade bei der Klassifikation mit diskreten Beschriftungen $y \in \{-1, 1\}$. Andere Rauschmodelle führen hingegen zu Schätzgleichungen, welche nicht in geschlossener Form und nur approximativ ermittelt werden können. In den Untersuchungen der Dissertation zeigte sich, dass eine Anwendung der GP Regression auch auf Klassifikationsaufgaben sinnvoll ist und bei vielen Anwendungen zu besseren Ergebnissen als reine Klassifikationsmodelle führt.

Abhängige Gauß-Prozesse und Lerntransfer Eine entscheidende Frage ist, wie das Konzept des Lerntransfers bei der GP Klassifikation verwendet werden kann. Es sei im Folgenden davon ausgegangen, dass genau zwei binäre Klassifikationsaufgaben gegeben sind, eine Unterstützungsaufgabe s mit Lerndaten $\mathcal{D}^s$ und eine Zielaufgabe τ mit Lerndaten $\mathcal{D}^\tau$. Anstatt die Klassifikatoren für diese Aufgaben jeweils unabhängig voneinander zu lernen ist es das Ziel ein gemeinsames Lernen zu ermöglichen. Durch diesen Schritt ist ein Transfer von Informationen zwischen den Aufgabenstellungen realisierbar.

Ein entscheidendes Konzept sind sogenannte abhängige Gauß-Prozesse [BCW08] (englisch: *dependent Gaussian processes*). Jeder der Klassifikationsaufgaben ist eine Funktion zugeordnet. Diese seien mit f^s für die Unterstützungsaufgabe und f^τ für die Zielaufgabe bezeichnet. Als grundlegende Annahme des Lerntransfers wurden die Ähnlichkeiten der Klassifikationsaufgaben zueinander vorausgesetzt. Diese Annahme kann nun unmittelbar als Korrelation zwischen den Funktionen modelliert werden und es lässt sich folgende gemeinsame A-priori-Annahme aufstellen mit $j, j' \in \{s, \tau\}$:

$$\mathbb{E}(f^j(\boldsymbol{x})f^{j'}(\boldsymbol{x}')) = \mathcal{K}((j,\boldsymbol{x}),(j',\boldsymbol{x}')) = \begin{cases} \mathcal{K}^{\mathcal{X}}(\boldsymbol{x},\boldsymbol{x}') & \text{wenn } j = j' \\ \rho \cdot \mathcal{K}^{\mathcal{X}}(\boldsymbol{x},\boldsymbol{x}') & \text{sonst} \end{cases} . \quad (3)$$

Der Parameter ρ gibt die Korrelation der Klassifikationsaufgaben an. Gleichung (3) kann als Erweiterung der Kernfunktion betrachtet werden und erlaubt es daher den Erwartungswert der A-posteriori-Verteilung von y_* direkt aus Gleichung (2) abzuleiten:

$$\begin{aligned} \mathbb{E}(y_* \mid \boldsymbol{x}_*, \mathcal{D}^s, \mathcal{D}^\tau) &= \boldsymbol{k}_*(\rho)^T(\mathbf{K}(\rho) + \sigma_\varepsilon^2 \cdot \mathbf{I})^{-1}\boldsymbol{y} \\ &= \begin{bmatrix} \boldsymbol{k}_{\tau *} \\ \rho \boldsymbol{k}_{s*} \end{bmatrix}^T \left(\begin{pmatrix} \mathbf{K}_{\tau\tau} & \rho\mathbf{K}_{\tau s} \\ \rho\mathbf{K}_{\tau s}^T & \mathbf{K}_{ss} \end{pmatrix} + \sigma_\varepsilon^2 \cdot \mathbf{I} \right)^{-1} \begin{bmatrix} \boldsymbol{y}_\tau \\ \boldsymbol{y}_s \end{bmatrix} . \quad (4) \end{aligned}$$

Die Indizierung mit s und τ dient der Zuordnung der Werte in den Vektoren und Matrizen zu Lernbeispielen der Zielaufgabe τ oder der Unterstützungsaufgabe s. Bei einer Wahl des Parameters durch $\rho = 0$ erhalten wir das ursprüngliche unabhängige Lernen der Klassifikatoren und bei $\rho = 1$ werden alle Lernbeispiele von s direkt für τ verwendet. Der Parameter erlaubt daher einen adaptiven Lerntransfer.

Auswahl von Unterstützungsklassen Ein Kernelement des entwickelten Verfahrens ist die automatische Schätzung des Parameters ρ und die Auswahl einer Unterstützungsaufgabe aus mehreren Klassifikationsaufgaben mit vielen Lernbeispielen. Dafür wurde eine effiziente Modellselektion mit Leave-one-out Schätzungen entwickelt [RD10], welche die Unterstützungsaufgabe mit der größten zu erwartenden Klassifikationsleistung auswählt.

Eine Auswahl, die nur auf den Bildinformationen basiert, kann natürlich bei der Verwendung von wenigen Lernbeispielen für die Zielaufgabe auch nachteilig sein und so einem sogenannten *negativem Transfer* führen. Daher ist es ratsam, noch zusätzliche Informationsquellen mit einzubeziehen. So ist es zum Beispiel möglich, linguistische semantische Datenbanken, wie etwa WordNet [DDS$^+$09], zu verwenden, um die Ähnlichkeit der Klassifikationsaufgaben auch anhand der Kategoriebezeichnungen durchzuführen. Ein optionaler Schritt des neuen Verfahrens zum adaptiven Lerntransfer ist daher eine Vorauswahl aufgrund von semantischen Ähnlichkeiten, d.h. es werden K Objektkategorien ausgewählt, welche zu der neuen Kategorie am ähnlichsten sind. In den Experimenten konnte gezeigt werden, dass eine Kombination von visuellen und semantischen Informationen vorteilhaft ist, da diese sich ergänzen.

Quantitative Auswertung Alle Methoden wurden quantitativ im Rahmen der Bildkategorisierung ausgewertet. Die Ergebnisse zeigen einen signifikante Steigerung der Erkennungsleistung im Vergleich zu aktuellen Methoden des Lerntransfers und Verfahren, welche keine zusätzlichen Lerndaten anderer Klassifikationsaufgaben verwenden. Abbildung 4 enthält einen Teil der durchgeführten Auswertungen, bei denen der Vorteil des

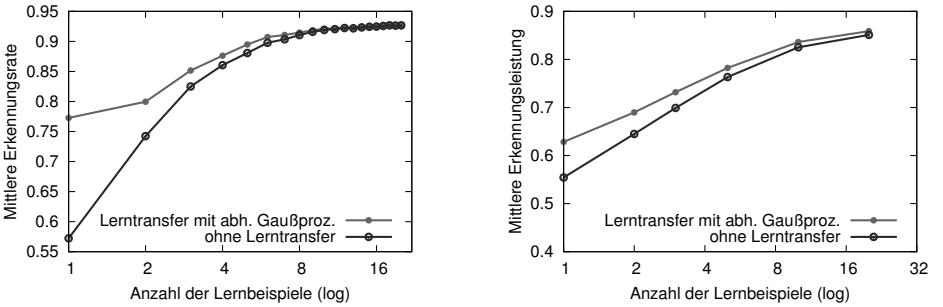


Abbildung 4: Beispielergebnisse der Auswertung des adaptiven Lerntransfers bei der Bildkategorisierung: (Links) mittlere Erkennungsrate bei drei Aufgabenstellungen der Caltech-256 Datenbank, (Rechts) mittlere Erkennungsleistung (average precision) von Aufgabenstellungen der Caltech-101 Datenbank.

Lerntransfers bei wenigen Lernbeispielen gut erkennbar ist. Weiterhin ist zu sehen, dass ab einer gewissen Anzahl von Beispielen, die Informationen im Lerndatensatz der Zielaufgabe genügen und das unabhängige Lernen gleich gute Ergebnisse erzielt wie das Verfahren des Lerntransfers. Für eine ausführliche Beschreibung der Experimente sei auf die Dissertation verwiesen.

Zusammenfassung weiterer Verfahren Eine weitere in der Dissertation vorgestellte Methode erweitert Entscheidungsbaumklassifikatoren um die Möglichkeit, Vorwissen von bereits erlernten Entscheidungsbäumen anderer Aufgaben zu verwenden [RD11]. Zusätzlich wurde eine Ansatz vorgestellt, welcher Informationen über die Merkmalsrelevanz transferiert, um den Lernprozess von randomisierten Entscheidungswäldern anzureichern. Für eine eine detaillierte Beschreibung sei auf die Dissertation verwiesen.

5 Ein-Klassen-Klassifikation

Eine weitere wichtige Art von Aufgabenstellungen mit wenigen Lernbeispielen sind solche, bei denen nur Lerndaten für eine einzige Klasse vorhanden sind. Dieses Szenario ist besonders häufig bei der Defekt- oder Anomaliedetektion zu finden. So sind zum Beispiel viele Bilder eines fehlerfreien Werkstücks vorhanden, jedoch gibt es nur wenige Aufnahmen von fehlerhaften Elementen. Idee vieler Verfahren der Ein-Klassen-Klassifikation oder Ausreißerdetektion ist es, die Verteilung der fehlerfreien Beispiele zu modellieren (z.B. mit Normalverteilungen). Eine Einschätzung eines neuen Beispiels kann dann aufgrund der Likelihood dieser Verteilung oder allgemein eines Neuheitsmaßes erfolgen. Diese Idee ist im linken Teil von Abbildung 5 noch einmal veranschaulicht.

Zur Lösung von Ein-Klassen-Problemen wurden neue Ansätze in der Dissertation entwickelt und vorgestellt, welche direkt vom Konzept der Regression und Klassifikation mit Gauß-Prozessen abgeleitet wurden. So kann unter Annahme eines mittelwertfreien Gauß-Prozesses als A-priori-Verteilung für die latente Funktion f direkt die Gauß-Prozess-

Abbildung 5: (Links) Veranschaulichung der Ein-Klassen-Klassifikation (Rechts) Anwendung der Ein-Klassen-Klassifikation bei der Schätzung von Attributen.

Regression auf die Ausgaben $y_i = 1$ angewendet werden und es ergibt sich:

$$\mathbb{E}(y_* \mid \boldsymbol{x}_*, \mathcal{D}) = \boldsymbol{k}_*^T \left(\mathbf{K} + \sigma_\varepsilon^2 \cdot \mathbf{I}\right)^{-1} \mathbf{1} \ . \tag{5}$$

Dieser Erwartungswert ist direkt als Neuheitsmaß einsetzbar. Weitere Maße ergeben sich unter Einbeziehung der Standardabweichung der Schätzung und bei Verwendung von approximativen Inferenzmethoden, auf die an dieser Stelle aber nicht näher eingegangen werden soll.

Die entwickelten Verfahren weisen viele Gemeinsamkeiten zu bekannten Ansätze wie etwa *support vector data description* [TD04] auf und es lässt sich sogar zeigen, dass Standardverfahren, wie etwa Parzen-Dichteschätzung oder Normalverteilungsklassifikatoren, durch die neuen Verfahren verallgemeinert werden. Ein großer Vorteil ist, dass die Algorithmen zur Klasse der nicht-parametrischen Verfahren gehören, d.h. alle Lerndaten werden bei der Klassifikation eines Beispiels direkt verwendet. In Experimenten wurde in der Dissertation gezeigt, dass die neuen Verfahren zu ähnlichen und sogar oft zu besseren Erkennungsraten als bisherige Methoden führen [KRD10]. Ein weiterer wichtiger Vorteil ist die einfache Implementierung der Algorithmen trotz ihrer theoretischen Komplexität.

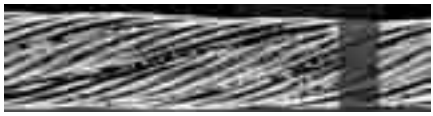
Auf der rechten Seite von Abbildung 5 ist die Anwendung der Verfahren für die Schätzung von Attributen zu sehen. Der Klassifikator wurde mit einer speziellen Art von Stuhl (Kategorie: *windsor chair*) angelernt. In der Erkennungsphase ist es dann möglich eine Menge von Bildern nach der vorhandenen Stärke dieses Attributes zu sortieren.

6 Weitere Anwendungen

Im Folgenden werden weitere Anwendungen der Verfahren vorgestellt, welche im Rahmen der Dissertation studiert wurden.

Defektlokalisierung Die Nützlichkeit der Verfahren der Ein-Klassen-Klassifikation wurde anhand der schwierigen Aufgabenstellung der Defektlokalisierung bei Drahtseilen demonstriert. Die Ergebnisse der Experimente zeigen deutlich, dass die vorgestellten Methoden in der Lage sind, bessere Erkennungsergebnisse als bisherige Standardverfahren (z.B. GMM) zu erzielen und Hinweise auf mögliche Defekte zu liefern. Ein Beispielergebnis ist in Abbildung 6 dargestellt.

Objekterkennung mit
einer mehreren Sensoren



Defektlokalisierung bei Drahtseilen

Aktionserkennung

Abbildung 6: Übersicht über die weiteren untersuchten Anwendungen der entwickelten Methoden: Objekterkennung mit Farb- und Time-of-Flight Kameras, Aktionserkennung und -detektion, Defektlokalisierung. Im rechten oberen Bild ist die Erkennung eines Drahtbruches mit dem entwickelten Verfahren zu erkennen. Die automatische Markierung ist in magenta dargestellt und die rote Markierung am oberen Rand zeigt die manuelle Annotation eines Experten.

Aktionserkennung Das Ziel der visuellen Aktionserkennung ist die Erkennung von Aktionen in Videoaufnahmen. Verfahren der Ein-Klassen-Klassifikation können bei dieser Anwendung vorteilhaft sein, da keine Modellierung von Negativbeispielen, d.h. Sequenzen ohne eine Aktion der Kategorie, notwendig ist. Bei den quantitativen Untersuchungen zeigte sich, dass die Ergebnisse stark von der Wahl des Hyperparameters der Kernfunktion abhängen, aber grundsätzlich eine Detektion von Aktionen möglich ist.

Generische Objekterkennung mit mehreren Sensoren Ein zusätzlicher Aspekt, welcher in der Dissertation untersucht wurde, ist die Entwicklung eines Systems zur generischen Objekterkennung, welches die Sensorinformationen einer Farb- und einer Time-of-Flight-Kamera kombiniert. Eine Time-of-Flight Kamera liefert, ähnlich zur aktuell üblichen Kinect-Kamera der Firma PrimeSense, Tiefendaten in Echtzeit. Dadurch können 3D-Informationen gewonnen werden, die bei manchen Erkennungsaufgaben entscheidend sind. In der Dissertation wurde untersucht, wie eine optimale Fusion der Sensordaten für eine bestimmte Klassifikationsaufgabe erfolgen kann. Dabei wurde erneut ein GP-Klassifikator eingesetzt, welcher es ermöglicht mehrere Kernfunktionen linear gewichtet zu kombinieren. In Experimenten zeigte sich, dass diese Kombination besonders bei wenigen Lerndaten vorteilhaft ist. Ingesamt konnte eine Steigerung der Erkennungsrate von 78.4% auf 88.1% im Vergleich zu bisherigen Verfahren erreicht werden.

7 Zusammenfassung

Ziel der in der Dissertation entwickelten Verfahren ist die Reduzierung der Anzahl von notwendigen Lernbeispielen bei der visuellen Objekterkennung. Dabei wurden mehrere Verfahren entwickelt, welche das Konzept des Lerntransfers beim maschinellen Lernen umsetzen. Grundgedanke ist hierbei die Ausnutzung von Lerndaten bereits bekannter Objektkategorien. Weiterhin wurden neue Methoden der Ein-Klassen-Klassifikation vorgestellt, welche bei der Defektlokalisierung, Aktionserkennung und Bildkategorisierung zum Einsatz kommen.

Die beschriebenen Verfahren sind notwendig, um ein effizientes, kontinuierliches und inkrementelles Lernen zu ermöglichen. Dieser Bereich wird aufgrund der immer höheren Anforderungen an automatisch bestimmte, semantische Information zunehmend an Bedeutung gewinnen.

Literatur

[BCW08] Edwin Bonilla, Kian Ming Chai und Chris Williams. Multi-task Gaussian Process Prediction. In *Advances in Neural Information Processing Systems*, Seiten 153–160. MIT Press, 2008.

[Bie87] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychol Rev*, 94(2):115–147, Apr 1987.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li und L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, Seiten 248 – 255, 2009.

[DWSP11] P. Dollar, C. Wojek, B. Schiele und P. Perona. Pedestrian Detection: An Evaluation of the State of the Art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6(99):1030–1037, 2011.

[FRD10] Björn Fröhlich, Erik Rodner und Joachim Denzler. A Fast Approach for Pixelwise Labeling of Facade Images. In *Proceedings of the 2010 International Conference on Pattern Recognition (ICPR'10)*, Jgg. 7, Seiten 3029–3032, 2010.

[KRD10] Michael Kemmler, Erik Rodner und Joachim Denzler. One-Class Classification with Gaussian Processes. In *Proceedings of the Asian Conference on Computer Vision*, Jgg. 2, Seiten 489–500, 2010.

[RD10] Erik Rodner und Joachim Denzler. One-Shot Learning of Object Categories using Dependent Gaussian Processes. In *Proceedings of the 2010 Annual Symposium of the German Association for Pattern Recognition (DAGM'10)*, Seiten 232–241, 2010.

[RD11] Erik Rodner und Joachim Denzler. Learning with Few Examples for Binary and Multiclass Classification Using Regularization of Randomized Trees. *Pattern Recognition Letters*, 32(2):244–251, 2011.

[RW05] Carl Edward Rasmussen und Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[TD04] David M. J. Tax und Robert P. W. Duin. Support Vector Data Description. *Machine Learning*, 54(1):45–66, 2004.

Erik Rodner studierte Informatik mit Nebenfach Mathematik an der Friedrich-Schiller Universität Jena und erwarb sein Diplom im Jahr 2007 mit Auszeichnung. Im Rahmen seiner Promotion am Lehrstuhl für digitale Bildverarbeitung bei Prof. Joachim Denzler, studierte er die Problemstellung des Lernens mit wenigen Beispielen, welche im vorliegenden Paper kurz zusammengefasst wurde. Seine Dissertation wurde im Jahr 2011 einheitlich mit summa cum laude bewertet und erhielt den Promotionspreis der Universität Jena. Aktuell ist er als PostDoc beschäftigt und erforscht Verfahren der visuellen Objekterkennung.

The Many Faces of Planarity
– Matching, Augmentation, and Embedding Algorithms for Planar Graphs –

Ignaz Rutter

Karlsruher Institut für Technologie (KIT)
Fakultät für Informatik
Am Fasanengarten 5
76131 Karlsruhe
rutter@kit.edu

Abstract: Ein Graph ist *planar*, wenn er sich kreuzungsfrei in die Ebene zeichnen lässt. Planarität ist eine zentrale Eigenschaft, nicht nur im Graphenzeichnen, sondern in der gesamten Graphentheorie. Oftmals lassen sich für planare Graphen stärkere theoretische Aussagen beweisen und effizientere Algorithmen angeben als für allgemeine Graphen. Andererseits tritt Planarität oft auch als Nebenbedingung auf und macht Probleme dadurch schwieriger. Eine besondere Rolle spielen planare Graphen in der Visualisierung, da Kreuzungen die Lesbarkeit von Zeichnungen verschlechtern.

In der vorliegenden Dissertation [Rut11] untersuche ich eine Reihe von Problemen, in denen Planarität auf unterschiedliche Weise auftritt. Im Bereich der kombinatorischen Optimierung wird Planarität als Nebenbedingung für Graphaugmentierungsprobleme sowie als Eingaberestriktion für Matching-Probleme betrachtet und beleuchtet inwiefern dies die Komplexität der jeweiligen Probleme verändert. Der zweite Teil der Arbeit befasst sich mit der Visualisierung planarer Graphen. Bisherige Verfahren zur planaren Visualisierung legen häufig zunächst eine kombinatorische Einbettung fest und optimieren dann im Rahmen dieser Einbettung weitere ästhetische Kriterien. Die Einschränkung auf eine einzige anfangs gewählte Einbettung erweist sich dabei häufig als nachteilig. Ich stelle Verfahren vor, die es ermöglichen über alle Einbettungen eines planaren Graphen zu optimieren und unter allen Einbettungen eine zu finden, die für die Visualisierung am besten geeignet ist.

1 Einleitung

Die Graphentheorie ist ohne Zweifel eine der großen Erfolgsgeschichten der diskreten Mathematik, und insbesondere hinsichtlich der automatisierten Verarbeitung auch der Informatik. Graphen und Netzwerke sind ubiquitär; sie werden auch in Bereichen weit jenseits dieser beiden Urpsrungsgebiete eingesetzt, um Relationen zwischen unterschiedlichsten Entitäten zu modellieren, zu studieren und zu verstehen, beispielsweise in der Physik, Biologie, den Sozialwissenschaften, aber auch um IT-Infrastrukturnetzwerke oder Prozessmodelle zu beschreiben. Menschen sind von Natur aus sehr visuell orientiert. Daher geht die Verwendung von Graphen um komplexe Relationen zu beschreiben häufig einher mit einer entsprechenden Visualisierung der Graphen. Bei zunehmender Anzahl von

Kantenkreuzungen in einer Zeichnung reduziert sich die Lesbarkeit drastisch. Daher ist es intuitiv, Kreuzungen gänzlich zu vermeiden; dies führt zur Definition der Klasse der planaren Graphen.

Heutzutage ist Planarität ein zentrales Konzept, nicht nur im Graphenzeichnen, sondern in der gesamten Graphentheorie. Die Charakterisierung planarer Graphen durch Kuratowski [Kur30] im Jahr 1930, kann als Geburtsstunde der modernen Graphentheorie angesehen werden. Die Charakterisierung über verbotene Substrukturen, nämlich K_5 (den vollständigen Graphen mit fünf Knoten) und $K_{3,3}$ (bestehend aus zwei Gruppen zu je drei Knoten, bei dem jeder Knoten mit allen Knoten der anderen Gruppe verbunden ist), zeigt, dass Planarität ein "endliches" Problem ist und führte zu den ersten polynomiellen Erkennungsalgorithmen. Den ersten Linearzeitalgorithmus zur Erkennung von planaren Graphen veröffentlichten Hopcroft und Tarjan 1974 [HT74], inzwischen ist eine Reihe von linearen Planaritätstests bekannt.

Die planaren Graphen bilden wahrscheinlich eine der am besten untersuchten Graphklassen. Eine Fülle von Literatur zeigt das immense Interesse an ihren Eigenschaften, Zeichenalgorithmen und Optimierungsalgorithmen, die speziell auf planare Graphen zugeschnitten sind. Beispielsweise besitzen planare Graphen gute Zerlegungseigenschaften, lassen sich mit wenigen Farben färben und viele Lösungen für Standardprobleme, die häufig als Subroutine in anderen Algorithmen eingesetzt werden, lassen sich auf planaren Graphen besonders effizient implementieren. Hierzu zählen beispielsweise Matching- und Flussalgorithmen. Zudem dienen planare Graphen oft als Sprungbrett für die Entwicklung effizienter Algorithmen auf allgemeineren Graphklassen, etwa Graphen mit beschränktem Genus oder den sogenannten H-minorenfreien Graphen.

Planarität ist eine Eigenschaft mit vielen unterschiedlichen Aspekten, da sie beispielsweise im Rahmen von kombinatorischen Optimierungsproblemen verschiedene Rollen einnehmen kann, beispielsweise die einer zusätzlichen Nebenbedingung aber auch die einer Eingaberestriktion. Die Nützlichkeit von Planarität und den damit einhergehenden Grapheigenschaften variiert stark mit dem betrachteten Problem und insbesondere mit der Rolle, die Planarität in dem Problem spielt. Um dem Rechnung zu tragen, ist die Arbeit in zwei Teile gegliedert. Im ersten Teil stehen Fragestellungen der kombinatorischen Optimierung im Vordergrund. Dort treten zwei Facetten von Planarität zutage: Einerseits wird Planarität als zusätzliche, hilfreiche Eigenschaft der Eingabe ausgenutzt, andererseits tritt Planarität auch als Nebenbedingung auf, deren Einhaltung durch die Problemstellung gefordert wird und die Probleme häufig schwieriger macht. Der zweite Teil befasst sich mit dem Zeichnen von planaren Graphen. Dabei spielt die Wahl der Einbettung eines planaren Graphen eine wesentliche Rolle für die Qualität der Darstellung. Ich stelle eine Reihe von Verfahren vor, die möglichst gute Einbettungen von planaren Graphen für verschiedene Zeichenstile berechnen.

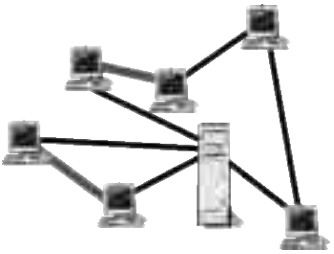


Abbildung 1: Durch Einfügen der dicken Kanten wird das Netzwerk gegen den Ausfall einer einzelnen Kante abgesichert.

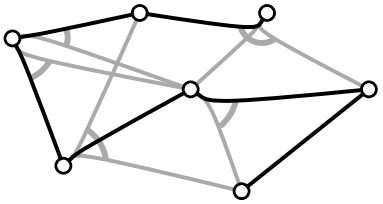


Abbildung 2: Ein Schaltergraph, Kanten desselben Schalters sind gemeinsamen Fußpunkt mit einem Bogen verbunden.

2 Kombinatorische Optimierung auf planaren Graphen

Der erste Teil der Arbeit beschäftigt sich mit verschiedenen kombinatorischen Optimierungsproblemen auf planaren Graphen, dabei tritt Planarität entweder als Restriktion der Eingabe oder als zusätzliche Nebenbedingung auf. Es werden drei unterschiedliche Fragestellungen behandelt.

2.1 Graphaugmentierung

Robustheit ist ein grundlegendes Kriterium beim Aufbau von Infrastruktur-Netzen, wie etwa Computernetzwerken. Ein häufig angewandtes Kriterium ist zum Beispiel zu fordern, dass der Graph zumindest *zweifach kantenzusammenhängend* ist, also nicht durch Ausfall einer einzelnen Kante in mehrere Zusammenhangskomponenten zerfällt. Andererseits ist man daran interessiert, die Kosten möglichst gering zu halten, sodass ein Netzwerk, in dem alle Knoten paarweise miteinander verbunden sind, zu teuer ist. Zudem werden Infrastruktur-Netzwerke häufig nicht vollständig neu aufgebaut, sondern ein bestehendes Netzwerk soll durch zusätzliche Komponenten möglichst kostengünstig erweitert werden, um neuen Anforderungen gerecht zu werden. Man ist also daran interessiert einen gegebenen Graphen, etwa durch Hinzufügen möglichst weniger Kanten, so zu modifizieren, dass er gewisse zusätzliche Eigenschaften erhält. Es wurde bereits eine Reihe solcher *Graphaugmentierungs-Probleme* im Zusammenhang mit Robustheit betrachtet. Dabei soll insbesondere der Zusammenhangsgrad eines gegebenen Graphen durch Hinzufügen möglichst weniger Kanten erhöht werden. Abbildung 1 zeigt ein Beispielnetzwerk, welches durch Einfügen der dicken Kanten zweifach kantenzusammenhängend wird. Hierbei handelt es sich um eine klassische und gut untersuchte Problemstellung. In meiner Arbeit fordere ich nun zusätzlich, dass der augmentierte Graph planar bleibt. Dieses Problem tritt beispielsweise im Graphenzeichnen auf, da viele Zeichenalgorithmen für planare Graphen zweifachen Zusammenhang voraussetzen oder zumindest für diese

Art von Graphen besondere Qualitätsgarantien angeben. Die Forderung möglichst wenige Kanten hinzuzufügen sorgt dafür, dass diese Qualitätsgarantien möglichst gut erhalten bleiben.

Ich untersuche den Komplexitätsstatus einer Reihe von planaren Augmentierungsproblemen. Inbesondere zeige ich, dass das Problem einen planaren Graphen durch Hinzufügen einer minimalen Anzahl von Kanten zweifach kantenzusammenhängend zu machen NP-schwer ist. Dies beantwortet eine offene Frage von Kant, der eine analoge Aussage für zweifachen Knotenzusammenhang gezeigt hat [KB91]. Weiter wird das analoge Problem betrachtet, bei dem der Graph *geometrisch eingebettet* ist, also jeder Knoten bereits eine fest zugewiesene Position hat und die Kanten geradlinig gezeichnet werden müssen. In der Arbeit wird gezeigt, dass einen geometrischen Graphen planar zweifach zusammenhängend zu machen selbst für Bäume NP-schwer ist, und auch das erhöhen zu c-fachem Zusammenhang für $c \geq 3$ schwer ist. Fordert man jedoch zusätzlich, dass die Knoten sich in konvexer Lage befinden, so ist das Problem effizient lösbar. In diesem Fall lässt sich nicht nur die Anzahl der zusätzlich nötigen Kanten minimieren, sondern auch das allgemeinere Kostenminimierungsproblem in $O(n^2)$ Zeit lösen, bei dem jeder möglichen zusätzlichen Kante ein positiver Kostenwert zugeordnet wird. Dieses Problem selbst ohne die Planaritätsbedingung im allgemeinen NP-schwer, und zwar sogar dann, wenn die Gewichte auf die Menge $\{1, 2\}$ beschränkt sind.

2.2 Schaltergraphen

Schaltergraphen bieten erweiterte Modellierungsmöglichkeiten gegenüber gewöhnlichen Graphen. Ein *Schalter* besteht aus einer Menge von Kanten, die sich einen gemeinsamen Knoten teilen. Eine *Konfiguration* wählt aus jedem Schalter eine Kante aus. Ein Schaltergraph beschreibt also eine Familie von Graphen und eine Konfiguration beschreibt ein konkretes Mitglied dieser Familie. Abbildung 2 zeigt einen Schaltergraphen, wobei Kanten, die zum selben Schalter gehören, durch einen Bogen am gemeinsamen Knoten miteinander verbunden sind. Die hervorgehobenen Kanten bilden eine Konfiguration dieses Schaltergraphen, deren resultierender Graph zusammenhängend ist. So lassen sich beispielsweise Computer- und Telefonnetzwerke oder auch Eisenbahnnetze mit Weichen auf sehr natürliche Weise als Schaltergraphen modellieren.

Aufgrund ihres Aufbaus eignen sich Schaltergraphen gut, um graphentheoretische Probleme zu modellieren, die strukturelle Entscheidungen beinhalten. Hat man einen Schaltergraphen vorliegen, ist man daran interessiert herauszufinden, ob seine Familie einen Graphen enthält, der eine gegebene Grapheigenschaft besitzt. Beispielsweise ist man daran interessiert, ob man die Verbindungen der Schalter so auswählen kann, dass zwei gegebene Teilnehmer eines Telefonnetzwerks miteinander verbunden sind. Eine weitere Problemstellung ergibt sich beispielsweise daraus, alle Teilnehmer einer Telefonkonferenz, oder gar alle Teilnehmer des Netzwerks, zusammenzuschalten. Schaltergraphen wurden von Groote und Ploeger eingeführt [GP08] und die vorliegende Arbeit beantwortet eine Reihe von offenen Fragen aus ihrer Arbeit.

Ich gebe effiziente Algorithmen an, mit denen überprüft werden kann, ob eine Verbindung zwischen zwei Knoten hergestellt werden kann, und auch, ob sich das gesamte Netzwerk zusammenschalten lässt; für beliebige Teilmengen hingegen ist das Problem NP-schwer. Zudem wird eine Reihe anderer Eigenschaften hinsichtlich ihrer Komplexität untersucht. So ist es beispielsweise NP-schwer zu entscheiden, ob ein Schaltergraph eine planare Konfiguration besitzt und ob man den Graphen so konfigurieren kann, dass er eulersch ist. Letzteres ist vor allem deswegen erstaunlich, da ein Graph genau dann eulersch ist, wenn er zusammenhängend ist und alle Knoten geraden Grad haben. Konfigurationen für die letztere beiden Eigenschaften können mit Algorithmen aus der Arbeit jedoch separat effizient gefunden werden.

2.3 Große Matchings in planaren Graphen

Ein *Matching* ist eine Teilmenge der Kanten eines Graphen, bei der jeder Knoten zu höchstens einer Kante dieser Menge inzident ist. Das Finden von möglichst großen Matchings ist ein gut untersuchtes Problem aus der kombinatorischen Optimierung, welches häufig als Subroutine in anderen Algorithmen zum Einsatz kommt. Der beste bekannte Algorithmus für dieses Problem hat eine Laufzeit von $O(\sqrt{n}m)$ [MV80]. Erst seit kurzem sind bessere Algorithmen, beispielsweise für planare Graphen, bekannt [MS06]. Allerdings basieren diese Verfahren auf schneller Matrix-Multiplikation, einem nicht sehr praxis-tauglichen Werkzeug. In der Praxis werden daher fast ausschließlich einfachere, langsamere Verfahren mit einer Laufzeit von $O(nm)$ eingesetzt.

Bei der Verwendung eines Matching-Algorithmus als Subroutine ist es häufig nicht zwingend nötig ein größtes Matching zu finden, sondern es genügt ein großes Matching (mit garantierter Mindestgröße) zu finden. Nishizeki und Baybars [NB79] zeigten, dass in planaren Graphen mit festem Minimalgrad stets Matchings einer bestimmten Mindestgröße existieren. Keines der bisherigen Verfahren ist jedoch in der Lage ein solches Matching, von dem man ja weiß, dass es existiert, schneller zu finden als ein größtes Matching. Ich stelle ein Verfahren vor, das ein solches Matching in linearer Zeit berechnet. Zunächst wird ein Verfahren angegeben, das auf allgemeinen Graphen mit Minimalgrad 3 arbeitet und für planare Graphen eine Mindestqualität garantiert, allerdings gelang es nicht mit einem solchen generischen Algorithmus die scharfe Schranke von Nishizeki und Baybars zu erreichen. Erst mit einem modifizierten Verfahren, das Planarität auch auf einer grundlegenderen Ebene ausnutzt und das Verfahren durch die inhärente in einer kombinatorischen Einbettung kodierten Informationen steuert, gelang es in linearer Zeit die scharfe Schranke von Nishizeki und Baybars erreicht. Das Verfahren lässt sich zudem auf größere Minimalgrade verallgemeinern und liefert für diese bessere Schranken.

3 Einbettungen von planaren Graphen

Der zweite Teil der Arbeit beschäftigt sich mit der Problemstellung, Einbettungen von planaren Graphen zu finden, die möglichst gut für bestimmte Visualisierungsarten geeignet sind. Diese Art von Problemen ist inhärent schwierig, da planare Graphen im Allgemeinen exponentiell viele planare Einbettungen besitzen. Es werden Einbettungsprobleme für unterschiedliche Zeichenstile untersucht, sowohl für topologische Zeichnungen, bei denen Kanten als beliebige Kurven gezeichnet werden dürfen, als auch für orthogonale Zeichnungen, bei denen Kanten nur aus horizontalen und vertikalen Streckensegmenten zusammengesetzt werden.

3.1 Planarität partiell eingebetteter Graphen

Durch die Abwesenheit von Kreuzungen sind planare Zeichnungen besonders gut lesbar. Zudem gibt es zahlreiche Algorithmen, die es erlauben zu einem planaren Graphen eine planare Einbettung sowie eine zugehörige Zeichnung zu berechnen. Diese Verfahren erlauben dem Anwender jedoch keine Kontrolle über das resultierende Layout. Wenn sich Netzwerke über die Zeit verändern, ist es wichtig, dass sich die Visualisierung der stabilen Teile eines Netzwerks möglichst wenig verändern, um dem Benutzer eine gute Orientierung zu ermöglichen. In meiner Arbeit betrachte ich daher die grundlegende Frage, ob eine gegebene planare Zeichnung eines Teils eines Netzwerkes sich auf planare Art und Weise zu einer Zeichnung des gesamten Netzwerkes erweitern lässt. Dabei darf der bereits vorgegebene Teil der Zeichnung nicht verändert werden. Die Komplexität dieses Problems hängt stark vom verwendeten Zeichenstil ab. Eine frühere Arbeit zeigte bereits, dass dieses Zeichnungs-Erweiterungsproblem für geradlinige Zeichnungen NP-schwer ist [Pat06]. Ich betrachte das entsprechende Problem für *topologischen Zeichnungen*, bei denen die Knoten eines Graphen durch Punkte und seine Kanten durch beliebige Jordankurven zwischen ihren Endpunkten repräsentiert werden. Abbildung 3 zeigt ein Beispiel eines bereits gezeichneten Teilgraphen mit der Aufgabe, einige weitere Kanten auf planare Weise in die Zeichnung einzufügen (die Instanz ist lösbar, aber nicht ganz offensichtlich; versuchen Sie es!). Versucht man zusätzlich noch die Kante 1–8 hinzuzufügen, so wird die Instanz unlösbar, obwohl der zugrundeliegende Graph selbst noch planar ist. Das partielle Einbettungsproblem unterscheidet sich also in diesem Punkt vom gewöhnlichen Planaritätstest.

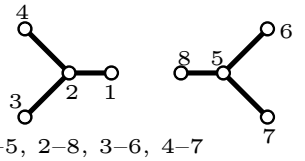


Abbildung 3: Ein partiell eingebetteter Graph. Die Aufgabe besteht darin, die gegebenen Kanten auf planare Weise zusätzlich in die Zeichnung einzufügen.

Für den Fall topologischer Zeichnungen zeige ich, dass das Problem äquivalent ist zu einem entsprechenden Erweiterungsproblem für kombinatorische Einbettungen. Ein *partiell eingebetteter Graph* (PEG) lässt sich als Tripel $(G, H, \mathcal{H})$ beschreiben, dabei ist G ein Graph und $H \subseteq G$ ein Teilgraph mit einer vorgegebenen planaren Einbettung $\mathcal{H}$. Die Fra-

ge besteht nun darin, ob eine planare Einbettung $\mathcal{G}$ von G existiert, deren Einschränkung auf H mit $\mathcal{H}$ übereinstimmt im Sinne folgender beider Eigenschaften: 1) Um jeden Knoten v von H ist die zirkuläre Reihenfolge seiner Nachbarn in H im Uhrzeigersinn um v dieselbe wie in $\mathcal{H}$, und 2) Für jeden (gerichteten) Kreis C in H liegen links bzw. rechts von C in $\mathcal{G}$ dieselben Knoten aus H wie in $\mathcal{H}$. In diesem Fall sagen wir, der PEG sei *planar*.

Überraschenderweise lässt sich das Problem zu testen ob ein PEG planar ist, anders als viele andere Lösungserweiterungsprobleme, effizient lösen, und zwar sogar in linearer Zeit. Das Verfahren hierzu verwendet zunächst eine Zerlegung in (zweifache) Zusammenhangskomponenten und anschließend den SPQR-Baum, eine Datenstruktur zur Repräsentation aller planaren Einbettungen von zweifachen zusammenhängenden planaren Graphen [DT96]. Die oben angegebenen Bedingungen lassen sich dann in diese Zusammenhangskomponenten und auch in den SPQR-Baum "projizieren", wo die Existenz einer Einbettung dann durch lokale Betrachtungen entschieden werden kann. Hierdurch ergibt sich ein relativ einfacher polynomieller Algorithmus, der sich mit Hilfe weiterer algorithmischer Techniken auf lineare Laufzeit beschleunigen lässt.

In einem weiteren Schritt werden zudem, nach dem Vorbild von Kuratowski, der zeigte, dass ein Graph genau dann planar ist, wenn er weder den Graph $K_{3,3}$ noch K_5 enthält, die planaren partiell eingebetteten Graphen durch verbotene Substrukturen charakterisiert. Hierzu werden zunächst die grundlegenden Minor-Operationen auf partiell eingebettete Graphen erweitert. Die neuen PEG-Minor-Operationen definieren eine Ordnung auf der Menge der PEGs, mit der Eigenschaft, dass alle Elemente, die kleiner sind als ein planarer PEG, ebenfalls planar sind. Es muss daher eine Menge von minimalen nicht-planaren PEGs geben, mit der Eigenschaft, dass jeder nicht-planare PEG mindestens eine dieser *verbotenen Substrukturen* enthält. Zusätzlich zu den beiden Graph $K_{3,3}$ und K_5, die sich aus der Forderung der Planarität ergeben, identifiziere ich sieben weitere solche verbotenen Substrukturen (siehe Abbildung 4) und zeige, dass dies genau die minimalen nicht-planaren PEGs sind, die oben angesprochene Menge also endlich und in der Tat recht klein ist. Neben einer genauen kombinatorischen Charakterisierung der planaren PEGs über verbotene Substrukturen ergibt sich hieraus auch ein effizienter *zertifizierender* Planaritätstest für PEGs, der zu einer gegebenen Eingabe entweder eine gültige Einbettungserweiterung berechnet, oder eine verbotene Substruktur extrahiert und so belegt, dass eine gültige Erweiterung nicht existiert. Gerade für komplexe und daher fehleranfällige Algorithmen wie Planaritätstests hat sich ein solches Vorgehen bei der Implementierung von Algorithmen bewährt.

3.2 Simultane Einbettungen

Liegen zwei (oder mehr) Graphen auf derselben Knotenmenge vor, so ist man häufig daran interessiert, diese Graphen miteinander zu vergleichen, beispielsweise durch Angabe einer Zeichnung, die die Ähnlichkeiten möglichst gut hervorhebt. Selbst für Paare planarer Graphen ist der Vereinigungsgraph im Allgemeinen nicht planar. Daher sucht man nach einer sogenannten *simultanen Einbettung mit festen Kanten*, das heißt die Knoten der beiden

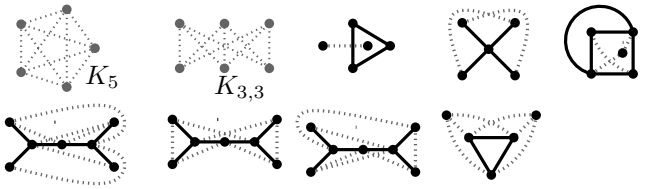


Abbildung 4: Die minimalen nicht-planaren PEGs. Der fest eingebettete Teilgraph H ist schwarz gezeichnet, die zusätzlichen Knoten und Kanten, deren Einbettung noch entschieden werden muss sind hell und gepunktet.

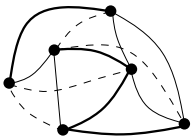


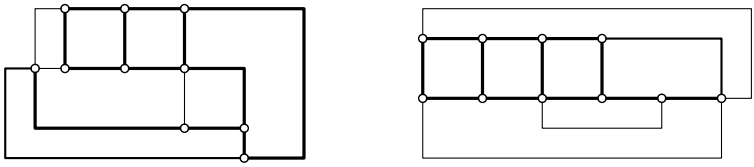
Abbildung 5: Simultane Einbettung zweier Graphen G_1 (durchgezogen) und G_2 (gestrichelt) gemeinsame Kanten sind fett gezeichnet.

Graphen werden an dieselben Positionen gezeichnet, gemeinsame Kanten werden durch dieselbe Kurve dargestellt und jede Zeichnung für sich genommen ist planar, siehe Abbildung 5. Für Tripel von planaren Graphen ist das entsprechende simultane Einbettungs-Problem NP-schwer [GJP$^+$06]. Obwohl sich in den letzten Jahren viele Wissenschaftler mit diesem Problem beschäftigt haben, ist der Komplexitätsstatus des Problems für Paare allgemeiner planarer Graphen noch ungeklärt. Bis dato existieren nur für sehr eingeschränkte Graphklassen effiziente Algorithmen. So lässt sich die Existenz einer simultanen Einbettung effizient überprüfen, wenn beide Graphen außenplanar sind oder wenn einer der beiden Graphen höchstens einen Kreis enthält.

Ich zeige, dass sich das Problem auch dann effizient, und zwar sogar in linearer Zeit, entscheiden lässt, wenn der Durchschnitt beider Graphen zweifach zusammenhängend ist. Dies ist einer der ersten Algorithmen, der zeigt, dass das Problem auch auf einer größeren Klasse von Graphen effizient lösbar ist. Zudem führe ich den Fall, dass der Durchschnitt zusammenhängend ist, zurück auf den Fall, dass der Durchschnitt ein Baum ist und alle weiteren Kanten zwischen Blättern dieses Baumes verlaufen. Dies zeigt neue Richtungen für algorithmische Ansätze für den allgemeinen Fall auf und grenzt die möglichen Bereiche für eine Suche nach potenziell schwierigen Instanzen weiter ein.

3.3 Orthogonale Zeichnungen mit Flexibilitäts-Bedingungen

Der letzte Abschnitt der Arbeit befasst sich mit dem *orthogonalen Zeichnen* von planaren Graphen, wobei Knoten auf dem Gitter platziert werden sollen und die Kanten aus horizontalen und vertikalen Streckensegmenten zusammengesetzt werden. Der Übergang von einem horizontalen auf ein vertikales Segment bzw. umgekehrt wird dabei als Knick



(a) (b)

Abbildung 6: Zwei unterschiedliche orthogonale Zeichnungen eines planaren Graphen.

bezeichnet. Da Kanten mit vielen Knicken die Lesbarkeit der Zeichnungen verschlechtern, ist man bestrebt die Anzahl der Knicke in der Zeichnung möglichst klein zu halten. Traditionell versucht man daher entweder die Gesamtanzahl an Knicken oder auch die maximale Anzahl an Knicken pro Kante zu minimieren. Beide Probleme sind NP-schwer, da es NP-schwer ist zu entscheiden, ob sich ein Graph gänzlich ohne Knicke zeichnen lässt [GT01]. Es ist aber bekannt, dass man mit höchstens zwei Knicken pro Kante stets auskommt [BK94]. Dagegen war bisher unklar, ob man effizient entscheiden kann, ob sich ein Graph mit nur einem Knick pro Kante orthogonal zeichnen lässt.

In dieser Arbeit wird ein neues Problem dieser Art eingeführt, bei dem jeder Kante eine individuelle Maximalzahl von Knicken zugeordnet wird, ihre *Flexibilität*. Dies umfasst das Problem der Zeichenbarkeit mit einem Knick pro Kante, ist aber noch wesentlich allgemeiner, da es dem Benutzer die genaue Kontrolle überlässt, welche Kanten wieviele Knicke haben dürfen. In Abbildung 6 ist ein Beispielgraph mit zwei verschiedenen orthogonalen Zeichnungen gegeben, die Wichtigkeit der Kanten ist durch ihre Dicke angegeben. Obwohl Zeichnung 6a hinsichtlich beider Qualitätsmaße besser ist als Zeichnung 6b, ist letztere übersichtlicher, da wichtige Kanten weniger Knicke haben. Erlaubt man starre Kanten, also solche mit Flexibilität 0, so ist das Problem offensichtlich NP-schwer. Mein Verfahren erlaubt es die Existenz einer Zeichnung im Rahmen der gegebenen Flexibilitäten effizient zu entscheiden, sofern man jeder Kante mindestens einen Knick erlaubt. Dies umfasst das Problem der Zeichenbarkeit mit einem Knick pro Kante und schließt damit die bisherige Lücke zwischen dem Schwere-Resultat für 0-Knick-Zeichenbarkeit und dem Algorithmus für das Zeichnen mit zwei Knicken pro Kante. Zudem liefert es aber einen wesentlich allgemeineren Ansatz zur Berechnung orthogonaler Zeichnungen, der dem Benutzer eine stärkere Kontrolle über die engültige Zeichnung gewährt.

Literatur

[BK94] Therese Biedl und Goos Kant. A better heuristic for orthogonal graph drawings. In *Proc. 2nd Europ. Symp. Algorithms (ESA'94)*, Jgg. 855 of *Lecture Notes Comput. Sci.*, Seiten 24–35. Springer-Verlag, 1994.

[DT96] Giuseppe DiBattista und Roberto Tamassia. On-Line Maintenance of Triconnected Com-

ponents with SPQR-Trees. *Algorithmica*, 15:302–318, 1996.

[GJP+06] Elisabeth Gassner, Michael Jünger, Merijam Percan, Marcus Schaefer und Michael Schulz. Simultaneous Graph Embeddings with Fixed Edges. In F. V. Fomin, Hrsg., *WG '06*, Jgg. 4271 of *Lecture Notes Comput. Sci.*, Seiten 325–335, 2006.

[GP08] Jan F. Groote und Bas Ploeger. Switching graphs. In *Proceedings of the 2nd Workshop on Reachability Problems (RP'2008)*, ENTCS, Seiten 119–135, 2008.

[GT01] Ashim Garg und Roberto Tamassia. On the Computational Complexity of Upward and Rectilinear Planarity Testing. *SIAM J. Comput.*, 31(2):601–625, 2001.

[HT74] John E. Hopcroft und Robert E. Tarjan. Efficient Planarity Testing. *J. ACM*, 21(4):549–568, 1974.

[KB91] Goos Kant und Hans L. Bodlaender. Planar Graph Augmentation Problems. In Frank Dehne, Jörg-Rüdiger Sack und Nicola Santoro, Hrsg., *Proc. 2nd Workshop Algorithms and Data Structures (WADS'91)*, Jgg. 519 of *Lecture Notes Comput. Sci.*, Seiten 286–298. Springer-Verlag, 1991.

[Kur30] Kazimierz Kuratowski. Sur le problème des courbes gauches en topologie. *Fund. Math.*, 15:271–283, 1930.

[MS06] Marcin Mucha und Piotr Sankowski. Maximum Matchings in Planar Graphs via Gaussian Elimination. *Algorithmica*, 45(1):3–20, 2006.

[MV80] Silvio Micali und Vijay V. Vazirani. An $O(\sqrt{|V|} \cdot |E|)$ algorithm for finding maximum matchings in general graphs. In *Proc. 21st Annu. IEEE Sympos. Found. Comput. Sci. (FOCS'80)*, Seiten 17–27, 1980.

[NB79] Takao Nishizeki und Ilker Baybars. Lower bounds on the cardinality of the maximum matchings of planar graphs. *Discrete Math.*, 28(3):255–267, 1979.

[Pat06] Maurizio Patrignani. On Extending a Partial Straight-Line Drawing. *Found. Comput. Sci.*, 17(5):1061–1069, 2006.

[Rut11] Ignaz Rutter. *The Many Faces of Planarity – Matching, Augmentation, and Embedding Algorithms for Planar Graphs –*. Dissertation, Fakultät für Informatik, Karlsruher Institut für Technologie (KIT), July 2011.

Ignaz Rutter, geboren 1981, studierte Informatik an der Universität Karlsruhe. Er schloss 2007 als Diplom-Informatiker ab und wurde mit dem Absolventenpreis für den besten Studienabschluss im akademischen Jahr 2006/2007 ausgezeichnet. Seit 2007 arbeitet er als wissenschaftlicher Mitarbeiter am Lehrstuhl Prof. Wagner am jetzigen Karlsruher Institut für Technologie (KIT). Die Promotion "summa cum laude" zum vorliegenden Thema erfolgte im Juli 2011. Im Januar 2012 wurde er mit dem Hermann-Billing-Preis für eine herausragende Dissertation an der Fakultät für Informatik des KIT im Jahr 2011 ausgezeichnet. Seine Interessensgebiete umfassen Algorithmen, Graphentheorie, Graphenzeichnen und kombinatorische Optimierung.

Die Struktur dominierender Mengen in Graphen

Oliver Schaudt
Institut für Informatik
Universität zu Köln
Weyertal 80, 50931 Cologne, Germany
schaudt@zpr.uni-koeln.de

Abstract: Ein zentrales Konzept in der Graphentheorie ist das der Dominierung. Eine *dominierende Menge* eines Graphen G ist eine Teilmenge X der Knoten, für die jeder Knoten aus $V(G) \setminus X$ einen Nachbarn in X besitzt. Anschaulich formuliert, eine dominierende Menge in einem Netzwerk ist ein Komitee, bei dem gilt, dass jedes Nicht-Mitglied ein Mitglied kennt. Dominierende Mengen, ihre Anwendungen und Varianten sind in der Forschungsliteratur sehr gut untersucht.

In dieser Arbeit tragen wir grundlegende Untersuchungen zu den strukturellen und algorithmischen Eigenschaften dominierender Mengen bei. Im Mittelpunkt stehen dabei spezielle dominierende Mengen, deren induzierte Teilgraphen zusätzlichen strukturellen Bedingungen genügen.

1 Einleitung

Ein zentrales Konzept in der Graphentheorie ist das der Dominierung. Eine *dominierende Menge* in einem gegebenen Graphen $G = (V, E)$ ist eine Teilmenge X der Knoten, sodass jeder Knoten aus $V \setminus X$ einen Nachbarn in X besitzt. Anschaulich formuliert, eine dominierende Menge in einem Netzwerk ist ein Komitee, so dass jedes Nicht-Mitglied ein Mitglied des Komitees kennt. Dominierende Mengen, ihre Anwendungen und Varianten sind in der Forschungsliteratur sehr gut untersucht. Das Konzept der Dominierung bezieht seine Motivation aus den Bereichen des Netzwerkdesing, der Standortprobleme und der Theorie sozialer Netzwerke (s. [HHS98]). Ein Beispiel für den Einsatz dominierender Mengen in der Praxis ist das Routen von Nachrichten in mobilen ad-hoc Netzwerken und Sensornetzwerken [BDTC04].

In der Literatur wurden zahlreiche Varianten des Dominierungsbegriffs vorgestellt und in mehreren tausend wissenschaftlichen Artikeln untersucht [HHS98]. Unter den wichtigsten und am besten untersuchten Konzepten sind die folgenden.

Eine *zusammenhängende dominierende Menge* eines zusammenhängenden Graphen G ist eine dominierende Menge X deren induzierter Teilgraph, im folgenden mit $G[X]$ bezeichnet, zusammenhängend ist. Der Zusammenhang motiviert sich aus der Anforderung an die dominierenden Knoten, untereinander kommunizieren zu können. Die minimale Kardinalität einer zusammenhängenden dominierenden Menge wird als $\gamma_c(G)$ bezeichnet. Offensichtlich besitzt jeder zusammenhängende Graph eine zusammenhängende dominierende Menge – bspw. die ganze Knotenmenge.

Eine *total dominierende Menge* X ist eine dominerende Menge ohne isolierte Knoten, d.h. jeder Knoten des Graphen hat einen Nachbarn in X. Diese Bedingung motiviert sich aus Anwendungen, in denen die dominierenden Knoten eine Kontrollfunktion haben, etwa zur Detektion von Fehlfunktionen im Netzwerk. Die minimale Kardinalität einer total dominierenden Menge wird mit γ_t bezeichnet.

Eine gut untersuchte Variante total dominierender Mengen sind die *Paar-dominierenden* Mengen. Hier wird gefordert, dass der durch die Paar-dominierende Menge induzierte Teilgraph ein perfektes Matching besitzt. Die dominerende Menge kann also in Paare adjazenter Knoten partitioniert werden. Insbesondere ist eine Paar-dominerende Menge stets total dominierend. Die minimale Kardinalität einer Paar-dominierenden Menge wird mit γ_p bezeichnet. Jeder Graph ohne isolierte Knoten besitzt eine Paar-dominierende Menge und demnach auch eine total dominierende Menge. Auf der anderen Seite kann ein Graph mit isolierten Knoten keine total dominierende Menge besitzen. Daher beschränken wir uns im folgenden auf Graphen ohne isolierte Knoten.

Alle oben genannten Parameter sind NP-vollständig, weswegen man an möglichst scharfen oberen (und unteren) Schranken interessiert ist. Neben der Untersuchung dieser quantitativen Parameter ist eine weitere zentrale Frage im Bereich der Dominierung die Bestimmung der strukturellen Eigenschaften der dominierenden Mengen. Dabei ist man an der Beschreibung der Struktur des von der dominierende Menge induzierten Teilgraphen interessiert. Beispiele hierfür sind das Problem der dominierenden Clique [CK90, CK88, BT90, Dra94]. Hier sucht man zusammenhängenden dominierenden Mengen, welche einen vollständigen Teilgraphen induzieren. Das Problem motiviert sich aus der Analyse sozialer Netzwerke [CK88]. Ein weiteres Beispiel ist das Problem der dominierenden induzierten Bäume [CSM04, Rau07, YWD09]. Hier sucht man zusammenhängende dominierende Mengen mit azyklischem induzierten Teilgraphen. Motiviert ist dieses Problem durch das Design von *virtual backbones* in mobilen ad-hoc Netzwerken [YWD09]. Weitere Beispiele motivieren sich aus dem Kontext der Dominierung selber.

Die folgenden Resultate stammen aus dem Spannungsfeld zwischen Graphentheorie und kombinatorischer Optimierung. Wir beschäftigen uns aus struktureller und komplexitätstheoretischer Sicht mit Dominierung. Alle im folgenden genannten Resultate sind, sofern nicht anders gekennzeichnet, der Dissertation entnommen. Die Resultate sind als Teilpublikationen [SS11, Sch12a, Sch12b, Sch12f, Sch12d, Sch12e, Sch12c, Sch11] erschienen.

1.1 Notation

Mit P_n bezeichnen wir den induzierten Pfad auf n Knoten. Ebenso ist C_n der induzierte Kreis auf n Knoten. Den vollständigen Graphen auf n Knoten notieren wir als K_n. $K_{n,m}$ ist der vollständige bipartite Graph auf n und m Knoten.

Es seien G und H zwei Graphen. Ist kein induzierter Teilgraph von G isomorph zu H, so sagen wir G ist *H-frei*.

2 Strukturelle Eigenschaften dominierender Mengen

2.1 Die NP-Vollständigkeit struktureller Eigenschaften zusammenhängender und total dominierender Mengen

Alle im folgenden angeführten Graphenklassen werden ausführlich definiert in [BD99]. Zu einer vorgegebenen Graphenklasse $\mathcal{G}$ betrachten wir das Problem, ob ein gegebener Graph G eine zusammenhängende bzw. total dominierende Menge X besitzt, sodass $G[X] \in \mathcal{G}$. Darüberhinaus betrachten wir das Problem, wie klein eine solche Menge X gewählt werden kann. Es ist bekannt, dass das Existenz- und Minimierungsproblem der dominierenden Cliquen [CK90] und das der dominierenden induzierten Bäume NP-vollständig sind [CSM04, Rau07].

Durch einen gemeinsamen Reduktionsansatz verallgemeinern wir diese einzelnen Resultate wie folgt. Das Existenzproblem einer zusammenhängenden oder total dominierenden Menge, deren induzierter Teilgraph zu einer der folgenden Graphenklassen gehört, ist NP-hart zu entscheiden:

- perfekte Graphen; Meyniel Graphen; (p, q)-chordale Graphen, festes $p \geq 4, q \geq 1$; schwach chordale Graphen.
- Paritätsgraphen; Distanz-vererbende Graphen; Ptolemäische Graphen.
- Bipartite Graphen; planare bipartite Graphen; chordale bipartite Graphen; azyklische Graphen.
- asteroidal-triple-freie Graphen; co-comparability Graphen; trapezoide Graphen; Permutationsgraphen; bipartite Permutationsgraphen; streng chordale Graphen.
- Intervallgraphen; Einheitsintervallgraphen.
- circular-arc-Graphen; unit-circular-arc-Graphen.
- Graphen mit unizyklischen Zusammenhangskomponenten; Kaktusgraphen.
- $K_{1,r}$-freie Graphen, festes $r \geq 3$.
- r-partite Graphen, $r \geq 2$.
- dreiecksfreie Graphen; vollständige bipartite Graphen; Sterne $(\{K_{1,n} : n \in \mathbb{N}\})$.

Diese Liste ist bei weitem nicht vollständig.

Darüberhinaus ist das Problem, zusammenhängende oder total dominierende Mengen von vorgegener Struktur zu minimieren, ebenso schwierig. Es sei $\mathcal{G}$ eine Klasse aus obiger Liste. Zu einem gegebenem Graphen G und $k \in \mathbb{N}$ ist es NP-hart zu entscheiden, ob G eine zusammenhängende bzw. total dominierende Menge X besitzt, für die $G[X] \in \mathcal{G}$ und $|X| \leq k$. Hierbei kann für fast alle Klassen $\mathcal{G}$ die Instanz G stärker eingeschränkt werden, bspw. auf bipartite Graphen vom Maximalgrad 4.

2.2 Strukturelle Eigenschaften zusammenhängender dominierender Mengen in Distanz-verberbenden Graphen

Die algorithmischen Eigenschaften zusammenhängender dominierende Mengen von Distanz-vererbenden Graphen sind sehr gut untersucht [DM88, BD98, Dra94, CY01]. Beispielsweise können kardinalitätsminimale zusammenhängende dominierende Mengen in linearer Zeit bestimmt werden. Unser Resultat hierzu ist die erste vollständige strukturelle Beschreibung zusammenhängender dominierender Mengen in Distanz-vererbenden Graphen.

Es sei G ein zusammenhängender Distanz-vererbender Graph mit $\gamma_c(G) \geq 2$. Sind X und Y zwei beliebige inklusionsminimale zusammenhängende dominierende Mengen, so sind die jeweils induzierten Teilgraphen $G[X]$ und $G[Y]$ isomorph. Insbesondere haben je zwei inklusionsminimale zusammenhängende dominierende Mengen die gleiche Größe. D.h. es gibt keine Unterschied zwischen inklusionsminimalen und kardinalitätsminimalen zusammenhängenden dominierenden Mengen.

Als Konsequenz ergibt sich, dass, im Gegensatz zum allgemeinen Fall (s. Abschnitt 2.1), die strukturellen Eigenschaften zusammenhängender dominierender Mengen in Distanz-vererbenden Graphen algorithmisch einfach erfasst werden können.

Darüberhinaus gilt die folgende stärkere Aussage: Es sei G ein zusammenhängender Distanz-vererbender Graph mit $\gamma_c(G) \geq 2$ und X eine inklusionsminimale zusammenhängende dominierende Menge von G. H sei ein zusammenhängender induzierter Teilgraph von G und Y eine inklusionsminimale zusammenhängende dominierende Menge von H. Dann gilt, dass $H[Y]$ ein induzierter Teilgraph von $G[X]$ ist.

2.3 Die strukturellen Eigenschaften total dominierender Mengen

In einer Reihe von Artikeln entwickelten Bacsó und Tuza die Theorie der sogenannten *strukturellen Dominierung* (s. [Bac09, Tuz08]). Die Idee der strukturellen Dominierung ist es, hinreichende Bedingungen für die Existenz zusammenhängender dominierender Mengen mit vorgeschriebender Struktur in Form verbotener induzierter Teilgraphen anzugeben. Wir übertragen diese Idee auf den Fall totaler Dominierung.

Die *Corona* eines Graphen G entsteht aus G durch das Ankleben eines Blattes an jeden Knoten von G (s. Abb. 1). Wir verwenden die Notation $Cor(G)$ für die Corona von G.

Ein Beispiel für Resultate in diesem Bereich ist das folgende Resultat. Es sei G ein Graph und $\mathcal{F}$ eine Klasse von Graphen, welche keinen induzierten Pfad enthält. Ist G $Cor(F)$-frei für alle $F \in \mathcal{F}$, dann hat G einen total dominierende Menge X für die $G[X]$ ein $\mathcal{F}$-freier Graph ist. Dabei sind die Coronas der Graphen aus $\mathcal{F}$ notwendigerweise verboten, da keine total dominierende Menge von $Cor(F)$ einen F-freien induzierten Teilgraphen hat.

Als Beispiel sei $\mathcal{F} = \{K_{1,r}\}$ für $r \geq 3$. Dann sagt obiger Satz, dass jeder $Cor(K_{1,r})$-freie Graph eine total dominierende Menge besitzt, deren induzierter Teilgraph $K_{1,r}$-frei

ist. Eine Anwendung dieses Beispiels ist die Schranke (3) an γ_p/Γ_t (s. Abschnitt 3.1). $K_{1,3}$ und $Cor(K_{1,3})$ sind dargestellt in Abb. 1.

3 Abhängigkeiten unter Dominierungsparametern

Die Berechnung aller in der Einleitung und im folgenden vorgestellten Dominerungsparameter ist NP-vollständig [HHS98]. Daher ist man an Schranken an die Parameter und deren Verhältnisse interessiert. Die Resultate in diesem Abschnitt arbeiten solche Schranken unter strukturellen Bedingungen an die betrachteten Graphen heraus.

3.1 Paar-Dominierung gegenüber totaler Dominierung

Wie in der Einleitung beschrieben ist jede Paar-dominierende Menge insbesondere total dominierend. In diesem Abschnitt beschäftigen wir uns mit der Frage, um welchen Faktor Paar-dominierende Mengen größer sind als total dominierende Mengen. Dabei geben wir strukturelle Bedingungen an, um diesen Faktor zu beschränken. Neben γ_t und γ_p untersuchen wir ausserdem die Parameter Γ_t und Γ_p, definiert als maximale Kardinalität einer inklusionsminimalen total dominierenden Menge bzw. Paar-dominierenden Menge.

Für jedes $r \geq 3$ bezeichnen wir mit T_r den Graphen, welcher aus $K_{1,r}$ durch einfache Unterteilung aller Kanten entsteht. Die Graphen $K_{1,3}$, T_3 und die Corona von $K_{1,3}$ sind in Abb. 1 dargestellt.

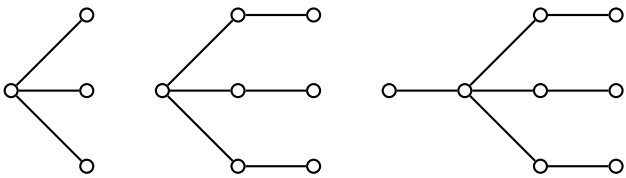


Abbildung 1: $K_{1,3}$, T_3 und die Corona von $K_{1,3}$.

Für jedes $r \geq 3$ gilt die Schranke

$$\gamma_p/\gamma_t \leq 2 - 2/r \tag{1}$$

für $K_{1,r}$-freie Graphen. In der Literatur war bis jetzt nur der Spezialfall $k = 3$ bekannt [DB98]. Darüberhinaus gilt (1) auch für jeden (C_5, T_r)-freien Graphen.

Aus (1) ergibt sich direkt die Schranke

$$\gamma_p/\gamma_t \leq 2 - 2/(\Delta + 1). \tag{2}$$

Dabei bezeichnet Δ den Maximalgrad des Graphen. Die Schranken (1) und (2) werden angenommen für jeden Wert $r \geq 3$ bzw. $\Delta \geq 2$. Überraschender Weise gelten die Schranken

(1) und (2) unter den gleichen Voraussetzungen ebenso für das Verhältnis Γ_p/Γ_t.

Für das Verhältnis γ_p/Γ_t gelten die folgenden Schranken. Für jedes $r \geq 3$ ist

$$\gamma_p/\Gamma_t \leq 2 - 2/r \tag{3}$$

für alle Graphen, welche die Corona von $K_{1,r}$ nicht als induzierten Teilgraphen enthalten. Diese Voraussetzung ist deutlich schwächer als die für (1) geforderte Bedingung, $K_{1,r}$-frei zu sein. Darüberhinaus verletzt die Corona von $K_{1,r}$, $r \geq 3$, die Schranke (3). Demnach ist sie notwendigerweise verboten – in diesem Sinne ist (3) optimal. Als Konsequenz leitet sich wiederum

$$\gamma_p/\Gamma_t \leq 2 - 2/\Delta \tag{4}$$

ab. Beide Schranken, (3) und (4), werden angenommen für jeden Wert $r \geq 3$ bzw. $\Delta \geq 2$.

3.2 Totale Dominierung und Paar-Dominierung gegenüber zusammenhängender Dominierung

Jede zusammenhängende dominierende Menge mit mehr als einem Knoten ist insbesondere total dominierend. Daher gilt $\gamma_t \leq \gamma_c$ für alle Graphen mit $\gamma_c \geq 2$. Es stellt sich sie Frage, welche Bedingungen hinreichend sind für $\gamma_c \leq \gamma_t$. Das folgende Resultat gibt die vollständige Liste der verbotenen induzierten Teilgraphen für $\gamma_c \leq \gamma_t$ an. Überraschender Weise sind diese induzierten Teilgraphen ebenso verboten für $\gamma_c \leq \Gamma_t$, $\gamma_c \leq \gamma_p$ und $\gamma_c \leq \Gamma_p$.

Sei $\phi \in \{\gamma_t, \Gamma_t, \gamma_p, \Gamma_p\}$ beliebig. Die folgenden Bedingungen sind äquivalent:

- Jeder zusammenhängende induzierte Teilgraph von G erfüllt $\gamma_c \leq \phi$.
- G ist (P_7, C_7, F_1, F_2)-frei (s. Abb. 2).

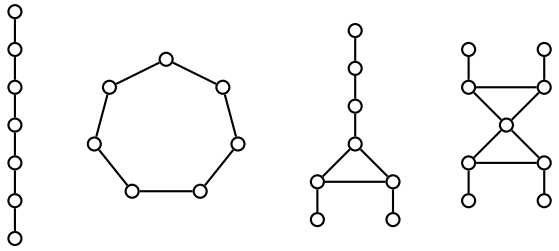


Abbildung 2: Die Graphen P_7, C_7, F_1 und F_2.

3.3 Induzierte Paar-dominierende Mengen

Eine *induzierte Paar-dominierende* Menge ist eine dominierende Menge, deren induzierter Teilgraph 1-regulär ist. Anders formuliert handelt es sich um eine Paar-dominierende Menge, deren Knoten durch ein induziertes Matching perfekt gematcht werden können. Nicht jeder Graph besitzt eine induzierte Paar-dominierende Menge, das entsprechende Entscheidungsproblem ist NP-vollständig.

Es gilt die Äquivalenz folgender Aussagen für jeden Graphen G:

1. Jeder induzierte Teilgraph von G erfüllt $\gamma_p/\Gamma_t \leq 1$.
2. Jeder induzierte Teilgraph von G hat eine induzierte Paar-dominierende Menge.
3. G enthält weder die Corona von P_3, noch die Corona von K_3, noch C_5 als induzierten Teilgraphen (s. Abb. 3).

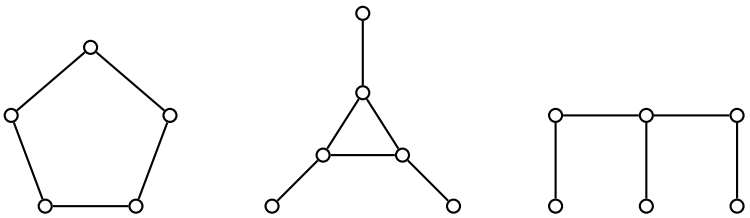


Abbildung 3: C_5, die Corona von K_3 und die Corona von P_3.

Die Klasse der $(C_5, Cor(P_3), Cor(K_3))$-freien Graphen hat weitere interessante Eigenschaften in Bezug auf Paar-dominierende Mengen.

Für jeden $(Cor(P_3), Cor(K_3))$-freien Graphen G auf n Knoten gilt

$$\gamma_p(G) \leq 2 \left\lceil \frac{n}{4} \right\rceil . \tag{5}$$

Dies bedeutet, dass eine kardinalitätsminimale Paar-dominierende Menge eines $(Cor(P_3), Cor(K_3))$-freien Graphen wenig mehr als die Hälfte der Knoten ausmacht. Die Schranke (5) ist optimal, da sie für jeden induzierten Pfad P_n, $n \geq 2$, scharf ist [HS98].

Haynes, Lawson und Studer [SHL00] führen den Parameter γ_{ip} ein, der für jeden Graphen definiert ist, welcher eine induzierte Paar-dominierende Menge besitzt. Er misst die Größe einer kardinalitätsminimalen induzierten Paar-dominierenden Menge. Ist zusätzlich zu $Cor(P_3)$ und $Cor(K_3)$ auch C_5 verboten, so gilt $\gamma_{ip} = \gamma_p$, d.h. es existiert eine induzierte Paar-dominierende Menge welche eine kardinalitätsminimale Paar-dominierende Menge ist. Daraus leitet sich mit (5) die Schranke

$$\gamma_{ip}(G) \leq 2 \left\lceil \frac{n}{4} \right\rceil$$

ab, welche für jeden $(C_5, Cor(P_3), Cor(K_3))$-freien Graphen gilt.

Wir notieren die minimale Größe eines inklusionsmaximalen induzierten Matchings mit im_-. Wie oben beschrieben, gibt es zu jeder induzierten Paar-dominierenden Menge X ein induziertes Matching, dessen gematchte Knoten genau X sind. Demnach gilt $2im_- \leq \gamma_{ip}$. Überraschender Weise gilt aber auch $\gamma_{ip} \leq 4im_-$ für alle $(C_5, Cor(P_3), Cor(K_3))$-freien Graphen. Insgesamt gilt also

$$2im_- \leq \gamma_{ip} \leq 4im_-.$$

4 Effizient totale Dominierung in Graphen

Eine *effizient total dominierende Menge* in einem Graphen $G = (V, E)$ ist eine Menge $X \subseteq V$ sodass jeder Knoten von G genau einen Nachbarn in X besitzt. D.h. $|N(v) \cap X| = 1$ für alle $v \in V$. Effizient total dominierende Mengen sind somit insbesondere total dominierend, aber die Rückrichtung ist nicht korrekt. Zu entscheiden, ob ein gegebener Graph eine effizient total dominierende Menge besitzt, ist ein NP-vollständiges Problem.

Effizient total dominierende Mengen motivieren sich aus folgenden Umformulierungen. Eine effizient total dominierende Menge eines Graphen $G = (V, E)$ entspricht genau einem perfekten Matching des *Nachbarschaftshypergraphen* $(V, \{N(v) : v \in V\})$. Perfekte Matchings in Nachbarschaftshypergraphen sind wiederum schlicht Exact Cover der Adjazenzmatrix. Diese wiederum können als ganzzahlige Punkte im Partitionierungspolytop der Adjazenzmatrix verstanden werden.

In der Arbeit wurden folgende komplexitätstheoretische Fragen bearbeitet:

- Auf welchen Graphenklassen kann die Existenz effizient total dominierender Mengen in polynomieller Zeit entschieden werden? Auf welchen bleibt das Problem NP-vollständig?
- Auf welchen Graphenklassen können unter der Vorgabe von Knotengewichten gewichtsminimale effizient total dominierende Mengen gefunden werden?

Zu diesen Fragen haben wir die folgenden Resultate. Das gewichtete Problem kann auf folgenden Instanzen effizient gelöst werden:

- odd-sun-freie chordale Graphen (in $\mathcal{O}(n^3)$ Schritten und linearer Programmierung),
- Balancierte Graphen (mit linearer Programmierung)
- Kantengraphen (in $\mathcal{O}(n^3)$ Schritten),
- Klauenfreie Graphen (in $\mathcal{O}(n^3)$ Schritten).

Darüberhinaus kann das Entscheidungsproblem kann auf folgenden Instanzen effizient gelöst werden:

- $star_{3,3,3}$-freie chordale Graphen (in $\mathcal{O}(n^3)$ Schritten),

- Kantengraphen (in $\mathcal{O}(n^2)$ Schritten).

Das Entscheidungsproblem bleibt dagegen NP-vollständig, wenn die Instanzen auf planare bipartite Graphen vom Maximalgrad 3 eingeschränkt werden.

Literatur

[Bac09] Gabor Bacsó. Complete description of forbidden subgraphs in the structural domination problem. *Discrete Mathematics*, 309:2466–2472, 2009.

[BD98] Andreas Brandstädt und Feodor F. Dragan. A linear-time algorithm for connected r-domination and Steiner tree on distance-hereditary graphs. *Networks*, 31:177–182, 1998.

[BD99] Andreas Brandstädt und Feodor F. Dragan. *Graph classes: a survey*. SIAM, 1999.

[BDTC04] Jeremy Blum, Min Ding, Andrew Thaeler und Xiuzhen Cheng. Connected Dominating Set in Sensor Networks and MANETs. In *Handbook of Combinatorial Optimization*, Seiten 329–369. Kluwer Academic Publishers, 2004.

[BT90] Gabor Bacsó und Zsolt Tuza. Dominating cliques in P_5-free graphs. *Periodica Mathematica Hungarica*, 21:303–308, 1990.

[CK88] Margaret B. Cozzens und Laura L. Kelleher. Dominating sets in social network graphs. *Mathematical Social Sciences*, 16:267–279, 1988.

[CK90] Margaret B. Cozzens und Laura L. Kelleher. Dominating cliques in graphs. *Annals of Discrete Mathematics*, 86:101–116, 1990.

[CSM04] Xue-Gang Chen, Liang Sun und Alice A. McRae. Tree domination in graphs. *Ars Combinatoria*, 73:193–203, 2004.

[CY01] Gerard J. Chang und Hong-Gwa Yeh. Weighted connected k-domination and weighted k-dominating clique in distance-hereditary graphs. *Theoretical Computer Science*, 263:3–8, 2001.

[DB98] Ronald D. Dutton und Robert C. Brigham. Domination in claw-free graphs. *Congressus Numerantium*, 132:69–75, 1998.

[DM88] Alessandro D'Atri und Marina Moscarini. Distance-Hereditary Graphs, Steiner Trees, and Connected Domination. *SIAM Journal of Computation*, 17:521–538, 1988.

[Dra94] Feodor F. Dragan. Dominating cliques in distance-hereditary graphs. *Lecture Notes in Computer Science*, 824:370–381, 1994.

[HHS98] Theresa W. Haynes, Stephen T. Hedetniemi und Peter J. Slater. *Fundamentals of Domination in Graphs*. Marcel Dekker, Inc., 1998.

[HS98] Teresa W. Haynes und Peter J. Slater. Paired-domination in graphs. *Networks*, 32:199–206, 1998.

[Rau07] Dieter Rautenbach. Dominating and large induced trees in regular graphs. *Discrete Mathematics*, 307:3177–3186, 2007.

[Sch11] Oliver Schaudt. On the existence of total dominating subgraphs with a prescribed additive hereditary property. *Discrete Mathematics*, 311:2095–2101, 2011.

[Sch12a] Oliver Schaudt. Efficient total domination in digraphs. *Journal of Discrete Algorithms*, 15:32–42, 2012.

[Sch12b] Oliver Schaudt. A note on connected domination in distance-hereditary graphs. *Discrete Applied Mathematics*, 160:1394–1398, 2012.

[Sch12c] Oliver Schaudt. On weighted efficient total domination. *Journal of Discrete Algorithms*, 10:61–69, 2012.

[Sch12d] Oliver Schaudt. Paired- and induced paired-domination in (E,net)-free graphs. *Discussiones Mathematicae Graph Theory*, 32:473–485, 2012.

[Sch12e] Oliver Schaudt. Total domination versus paired domination. *Discussiones Mathematicae Graph Theory*, 32:435–447, 2012.

[Sch12f] Oliver Schaudt. When the connected domination number is at most the total domination number. *Discrete Applied Mathematics*, 160:1281–1284, 2012.

[SHL00] Daniel S. Studer, Teresa W. Haynes und Linda M. Lawson. Dominating cliques in distance-hereditary graphs. *Ars Combinatoria*, 57:111–128, 2000.

[SS11] Oliver Schaudt und Rainer Schrader. The complexity of connected dominating sets and total dominating sets with specified induced graphs. erscheint in Information Processing Letters, 2011.

[Tuz08] Zsolt Tuza. Hereditary domination in graphs: Characterization with forbidden induced subgraphs. *SIAM Journal of Discrete Mathematics*, 22:849–853, 2008.

[YWD09] Ruiyun Yu, Xingwei Wang und Sajal K. Das. EEDTC: Energy-efficient dominating tree construction in multi-hop wireless networks. *Pervasive and Mobile Computing*, 5:318–333, 2009.

Oliver Schaudt

Oliver Schaudt wurde am 03. Februar 1986 in Bergisch Gladbach geboren. Er ist verheiratet und hat eine Tochter. Nach seinem Abitur im Juni 2005 begann er ein Studium der Mathematik (Nebenfach Informatik) an der Universität zu Köln. Im Dezember 2009 schloss er sein Studium als Diplom-Mathematiker ab. Sein Diplom wurde mit der Note *mit Auszeichnung* bewertet. Die anschliessende Promotion am Institut für Informatik der Universität zu Köln schloss er im Oktober 2011 ab. Die Promotion wurde mit der Note *summa cum laude* bewertet. Während der Promotionszeit arbeitete er als wissenschaftlicher Mitarbeiter am Institut für Informatik.

Aktuell arbeitet er an der Université Pierre et Marie Curie (Paris 6) als Postdoc.

Atomic Basic Blocks
Eine Abstraktion für die gezielte Manipulation der Echtzeitsystemarchitektur

Fabian Scheler

Friedrich-Alexander-Universität Erlangen-Nürnberg
Department Informatik
Lehrstuhl für Informatik 4 Verteilte Systeme und Betriebssysteme
fs@cs.fau.de

Abstract: Je nach Anwendungsfall ist das ereignis- oder das zeitgesteuerte Paradigma für die Realisierung eines Echtzeitsystems zu bevorzugen. Diese Paradigmen nehmen aber auch entscheidenden Einfluss auf die interne Struktur eines Echtzeitsystems, etwa die Koordinierung gleichzeitiger Ereignisbehandlungen bei Erzeuger-Verbraucher-Beziehungen oder kritischen Abschnitten, was eine spätere Anpassung des gewählten Paradigmas schwierig gestaltet. Um trotz dieser strukturellen Unterschiede eine werkzeuggestützte Anpassung des Echtzeitparadigmas zu ermöglichen, wurde im Rahmen dieser Dissertation die Abstraktion *Atomic Basic Block* und darauf basierend der *Real-Time Systems Compiler* entwickelt, ein Transformationswerkzeug, das einen automatisierten Übergang von ereignis- zu zeitgesteuerten Systemen ermöglicht.

1 Einleitung

Echtzeitrechensysteme sind oft in eine physikalische Umwelt eingebettet, die ihr Verhalten entscheidend prägt. Aus der Umwelt hervorgehende Stimuli initiieren Berechnungen im Echtzeitrechensystem, die sowohl funktional korrekt ablaufen müssen, als auch zeitlichen Beschränkungen unterworfen sind. Allein korrekte Ergebnisse zu liefern reicht also nicht aus, Berechnungsergebnisse müssen auch bis zu einem bestimmten Termin vorliegen. Wird ein Termin verpasst und das Berechnungsergebnis zu spät geliefert, kann dies ebenso ernsthafte Konsequenzen haben wie fehlerhaft berechnete Werte.

Die Abläufe innerhalb eines Echtzeitrechensystems, vom Eintreten des Stimulus bis hin zur Bereitstellung des Ergebnisses, unterscheiden sich je nach der verwendeten *Echtzeitsystemarchitektur*. Sie entscheidet, wie Ereignisbehandlungen aktiviert und wie Abhängigkeiten zwischen verschiedenen Ereignisbehandlungen implementiert werden. Die bekanntesten Ausprägungen solcher Echtzeitsystemarchitekturen sind das *ereignisgesteuerte* und das *zeitgesteuerte Paradigma*. Ersteres koppelt Ereignisbehandlungen durch Unterbrechungen direkt an externe Ereignisse und implementiert Abhängigkeiten zwischen Ereignisbehandlungen explizit mithilfe von Synchronisationsmechanismen wie Semaphoren oder Schlossvariablen. Zeitgesteuerte Systeme hingegen ordnen Ereignisbehandlungen in einer

vorab berechneten Ablauftabelle an, die zyklisch abgearbeitet wird und Abhängigkeiten zwischen verschiedenen Ereignisbehandlungen bereits berücksichtigt.

Je nach Anwendungsszenario bieten beide Paradigmen unterschiedliche Vorteile. So punkten ereignisgesteuerte Systeme durch Flexibilität, wenn sich die Ereigniszeitpunkte in der physikalischen Umwelt nicht exakt vorhersagen lassen, während zeitgesteuerte Systeme bei strikt periodischen Anwendungen ihre Stärken ausspielen können. Bei der Entwicklung des Space Shuttle entschied man sich beispielsweise für einen ereignisgesteuerten Ansatz, um flexibel auf sich ändernde Anforderungen in dem lang andauernden Entwicklungsvorhaben eingehen zu können [Car84]. Die Vorhersagbarkeit zeitgesteuerter Systeme hingegen kommt vor allem bei der Entwicklung und Verifikation fehlertoleranter Systeme zum Tragen, weshalb die Flugsteuerung des F-18-Kampfflugzeugs aufwändig von einer ereignis- auf eine zeitgesteuerte Architektur übertragen wurde [SG90].

Kann man die Wahl zwischen dem ereignis- und dem zeitgesteuerten Paradigma zu Beginn eines Entwicklungsvorhaben noch relativ frei treffen, ist eine spätere Migration mit hohem Aufwand verbunden. Die verwendete Echtzeitsystemarchitektur beeinflusst nämlich in großem Maße die Implementierung eines Echtzeitsystems. Ereignisgesteuerte Implementierungen sind häufig mit Aufrufen von Synchronisationsmechanismen durchsetzt, um Abhängigkeiten zwischen verschiedenen Ereignisbehandlungen explizit herzustellen, während diese Abhängigkeiten in zeitgesteuerten Systemen kaum noch sichtbar sind. Eine Migration zwischen diesen beiden Welten erfordert also zunächst eine mühsame Aufarbeitung dieser Abhängigkeiten, um sie korrekt auf die Zielarchitektur zu übertragen.

Im Rahmen dieser Dissertation wurde nun eine Möglichkeit entwickelt, diese Migration mithilfe eines Transformationswerkzeugs zu automatisieren. Die Abhängigkeiten zwischen verschiedenen Ereignisbehandlungen werden hierfür durch globale Abhängigkeitsgraphen unabhängig von der Echtzeitsystemarchitektur beschrieben. Diese Abhängigkeitsgraphen bestehen aus *Atomic Basic Blocks* (ABBs) und basieren auf interprozeduralen Kontrollflussgraphen, wie sie im Übersetzerbau verwendet werden. Im Gegensatz zu Kontrollflussgraphen können aber auch Abhängigkeiten zwischen verschiedenen Ereignisbehandlungen ausgedrückt werden. Die zeitlichen Eigenschaften der physikalischen Umwelt, etwa die Periodizität der einzelnen Stimuli oder der Termin, bis zu dem eine Ereignisbehandlung abgeschlossen sein muss, werden in einem *Systemmodell* gekapselt und mit den durch ABBs beschriebenen Ereignisbehandlungen verknüpft. ABBs und das Systemmodell sind die Grundlage des im Rahmen dieser Dissertation entwickelten *Real-Time Systems Compilers* (RTSC), eines Quelltexttransformationswerkzeugs, um die Echtzeitsystemarchitektur eines Echtzeitsystems gezielt zu manipulieren.

Im folgenden Abschnitt 2 werden kurz die charakteristischen Merkmale einer Echtzeitsystemarchitektur dargelegt. Anschließend beschreibt Abschnitt 3 die Abstraktion *Atomic Basic Block*. Zusammen mit dem in Abschnitt 4 vorgestellten *Systemmodell* bildet sie die grundlegenden Abstraktion für den *Real-Time Systems Compiler*, der in Abschnitt 5 präsentiert wird. Abschnitt 6 fasst die wesentlichen Beiträge dieser Dissertation noch einmal zusammen und gibt Ausblick auf zukünftige Arbeiten.

```
Message *serialMsg;

ISR(SerialByte) {
  unsigned char rcv = getByte();
  msg_addTo(serialMsg,rcv);

  if(msg_complete(serialMsg)) {
    buffer_insert(buf,serialMsg);
    SetEvent(MsgHandler,MsgRcv);
  }
}
```

```
TASK(MsgHandler) {
  Message *currentMsg = 0;
  Initialisation();

  WaitEvent(MsgRcv);
  ClearEvent(MsgRcv);
  currentMsg = buffer_get(buf);
  handler(currentMsg);

  TerminateTask();
}
```

Abbildung 1: Byte-weiser Empfang einer Nachricht über die serielle Schnittstelle, implementiert durch die Unterbrechungsbehandlung **ISR**(SerialByte) und den Faden **TASK**(MsgHandler)

.

2 Echtzeitsystemarchitekturen

Eine Echtzeitsystemarchitektur stellt Mechanismen zur Verfügung, um Ereignisbehandlungen als Reaktion auf regelmäßig und unregelmäßig auftretende Stimuli (*periodische* und *nicht-periodische Ereignisse*) zu aktivieren und um gerichtete Abhängigkeiten (Sequentialisierung von Vorgänger und Nachfolger) und ungerichtete Abhängigkeiten (Koordinierung kritischer Abschnitte) zu implementieren. Außerdem bietet sie eine deterministische Ablaufplanung, um mehrere Ereignisbehandlungen abwechselnd auf demselben Prozessor auszuführen.

Anhand des Beispiels in Abbildung 1 soll ihr Einfluss auf die Struktur einer Echtzeitanwendung erläutert werden. Das Beispiel verwendet das Betriebssystem AUTOSAR OS [AUT09], das eine ereignisgesteuerte Echtzeitsystemarchitektur implementiert, und stellt den Empfang einer Nachricht über die serielle Schnittstelle Byte für Byte durch die Unterbrechungsbehandlung **ISR**(SerialByte) und den Faden **TASK**(MsgHandler) dar.

Die Unterbrechungsbehandlung **ISR**(SerialByte) wird durch das Eintreffen eines Bytes an der seriellen Schnittstelle aktiviert, wodurch sie auch direkt an diesen Stimulus gebunden wird. Die Unterbrechungsbehandlung holt das empfangene Byte von der seriellen Schnittstelle ab und fügt es zur aktuellen Nachricht hinzu. Vollständige Nachrichten werden anschließend in einem Puffer zwischengespeichert. Die gerichtete Abhängigkeit zwischen dem Erzeuger **ISR**(SerialByte) der Nachricht und deren Verbraucher **TASK**(MsgHandler) wird im vorliegenden Beispiel explizit durch die Systemaufrufe SetEvent und WaitEvent hergestellt. Sobald die explizite Wartebedingung dieser gerichteten Abhängigkeit erfüllt ist, kann **TASK**(MsgHandler) die Nachricht aus dem Puffer entnehmen und weiter verarbeiten. Zusätzlich existiert hier noch eine ungerichtete Abhängigkeit, die jedoch nicht abgebildet ist. Das Einfügen der neuen Nachricht und ihre Entnahme aus dem Puffer stellen kritische Abschnitte dar, die geeignet abgesichert werden müssen. Im Falle von AUTOSAR OS könnte dies beispielsweise durch Schlossvariablen geschehen, die eine überlappende Ausführung dieser kritischen Abschnitte verhindern.

Ereignisbehandlung	Startzeitpunkt
ISR(SerialByte)	50 μs
...	...
ISR(SerialByte)	100 μs
...	...
ISR(SerialByte)	450 μs
...	...
TASK(MsgHandler)	500 μs

Tabelle 1: Ausschnitt aus einer möglichen statischen Ablauftabelle für das Beispiel aus Abbildung 1.

Solche Abhängigkeiten werden in zeitgesteuerten Echtzeitsystemarchitekturen, wie sie etwa OSEKtime [OSE01] anbietet, implizit ohne Zuhilfenahme von Systemaufrufen wie `SetEvent` und `WaitEvent` umgesetzt. Hierfür werden die Ereignisbehandlungen in einer Ablauftabelle zeitlich so angeordnet, dass diese Abhängigkeiten gewahrt bleiben. Tabelle 1 skizziert auszugsweise eine solche Ablauftabelle. Sie platziert **ISR**(SerialByte) vor **TASK**(MsgHandler), um so implizit die gerichtete Abhängigkeit sicherzustellen und eine überlappungsfreie Ausführung der kritischen Abschnitte zu erreichen. Die konkreten Startzeitpunkte der einzelnen Ereignisbehandlungen in der Ablauftabelle ergeben sich dabei aus einer Analyse der physikalischen Umgebung. So können die Startzeitpunkte von **ISR**(SerialByte) beispielsweise aus der Datenrate der seriellen Schnittstelle abgeleitet werden. **TASK**(MsgHandler) wird im Vergleich zu **ISR**(SerialByte) wesentlich seltener aktiviert, weil eine komplette Nachricht meist aus mehreren Bytes besteht, die erst von der Unterbrechungsbehandlung geliefert werden müssen.

Dieses Beispiel demonstriert bereits anschaulich, wie unterschiedlich verschiedene Echtzeitsystemarchitekturen mit gerichteten und ungerichteten Abhängigkeiten umgehen, und welchen Einfluss dies auf die Implementierung einer Anwendung hat. Die Abhängigkeiten zwischen **ISR**(SerialByte) und **TASK**(MsgHandler) existieren auch in der zeitgesteuerten Umsetzung, nur sind sie nicht mehr explizit sichtbar, weil sie sich einzig in der Reihenfolge der Ablauftabelle niederschlagen, wohingegen sie in der ereignisgesteuerten Implementierung direkt an den entsprechenden Systemaufrufen abgelesen werden können. Eine generische und architekturneutrale Beschreibung dieser Abhängigkeitsbeziehungen ist also die Voraussetzung, um die Abbildung eines Echtzeitsystems auf verschiedene Echtzeitsystemarchitekturen zu ermöglichen. Dies leisten ABBs, die im folgenden Abschnitt vorgestellt werden.

3 Atomic Basic Blocks

ABBs beschreiben gerichtete und ungerichtete Abhängigkeiten zwischen verschiedenen Ereignisbehandlungen unabhängig von der verwendeten Echtzeitsystemarchitektur. Hierfür erweitern ABBs die aus dem Übersetzerbau bekannten *Grundblöcke*. Grundblöcke und auch interprozedurale Kontrollflussgraphen alleine reichen nämlich nicht aus, um auch kontrollflussübergreifende Abhängigkeiten zwischen verschiedenen Aktivitätsträgern dar-

zustellen. ABBs fassen daher mehrere aneinander hängende Grundblöcke zusammen und erzeugen so eine Vergröberung des lokalen Kontrollflussgraphen einer Funktion. Auf Funktionsebene zeichnen sie den Kontrollflussgraphen der Funktion nach und bilden so *lokale ABB-Graphen*. ABBs stellen aber auch Ansatzpunkte für kontrollflussübergreifende gerichtete und ungerichtete Abhängigkeiten dar. ABBs enden daher immer an *ABB-Endpunkten*, die genau diese Ansatzpunkte im Kontrollflussgraphen markieren. ABBs lassen sich mithilfe folgender Regeln aus einem Kontrollflussgraphen extrahieren:

1. ABBs umfassen einen oder mehrere Grundblöcke einer Funktion, die einen zusammenhängenden Teilgraphen des Kontrollflussgraphen dieser Funktion bilden.
2. Jeder ABB besitzt einen eindeutigen Grundblock $Entry(ABB)$, über den dieser ABB betreten wird, seinen *Eingang*. Ebenso existiert für jeden ABB höchstens ein Grundblock $Exit(ABB)$, über den der ABB verlassen wird, der *Ausgang*. Sie sind die einzigen Grundblöcke eines ABB, die Vorgänger bzw. Nachfolger im Kontrollflussgraphen der Funktion besitzen können, die nicht Bestandteil des ABB sind.
3. ABBs reichen vom Ende des vorhergehenden ABB bis zu einem ABB-Endpunkt.
4. Liegt ein ABB-Endpunkt innerhalb eines Grundblocks, wird dieser Grundblock entsprechend in zwei Grundblöcke geteilt.

ABB-Endpunkte sind *Quellen* oder *Ziele* kontrollflussübergreifender Abhängigkeiten oder *künstliche ABB-Endpunkte*. Letztere dienen lediglich dazu, die oben genannte Regel 2 zu erfüllen und eine eindeutige Zerlegung eines Kontrollflussgraphen in ABBs zu gewährleisten. Quellen und Ziele werden durch Systemaufrufe markiert, die Abhängigkeiten zwischen gleichzeitigen Kontrollflüssen herstellen, beispielsweise die Systemaufrufe `SetEvent` und `WaitEvent` des in Abschnitt 2 betrachteten Beispiels.

Gerichtete kontrollflussübergreifende Abhängigkeiten zwischen *vereinbaren ABB-Endpunkten* verknüpfen schließlich lokale ABB-Graphen einzelner Funktionen zu einem Wald *globaler ABB-Graphen*, die sämtliche Abhängigkeitsbeziehungen in einem Echtzeitsystem beschreiben. ABB-Endpunkte heißen vereinbar, wenn es sich bei ihnen um kompatible Quellen und Ziele handelt (z. B. die bereits genannten Systemaufrufe `SetEvent` und `WaitEvent`) und sie sich auf entsprechend zusammenhängende Systemobjekte beziehen (z. B. dasselbe Ereignis). Kritische Abschnitte werden ihrerseits durch eine Menge von zusammenhängenden ABBs eines lokalen ABB-Graphen beschrieben. Zwischen zwei kritischen Abschnitten besteht eine ungerichtete Abhängigkeit, wenn ihnen dasselbe Betriebsmittel zugeordnet wurde. Sie bilden also einen *ungerichteten globalen ABB-Graphen*.

Abbildung 2 zeigt den zu Abbildung 1 gehörenden globalen ABB-Graphen. Grundblöcke sind dabei als eckige Kästen und ABBs als Kästen mit abgerundeten Ecken dargestellt. Die gerichteten Kanten innerhalb der Ereignisbehandlungen **`ISR`**`(SerialByte)` und **`TASK`**`(MsgHandler)` spiegeln den ursprünglichen Kontrollflussgraphen wider. Wegen der zu den Systemaufrufen `SetEvent` und `WaitEvent` gehörenden ABB-Endpunkten wurden die Grundblöcke $BB2$ und $BB4$ jeweils in die Grundblöcke $BB2a$ und $BB2b$ beziehungsweise $BB4a$ und $BB4b$ aufgeteilt. Nachdem diese ABB-Endpunkte auch vereinbar sind, werden die beiden lokalen ABB-Graphen an dieser Stelle durch eine gerichtete

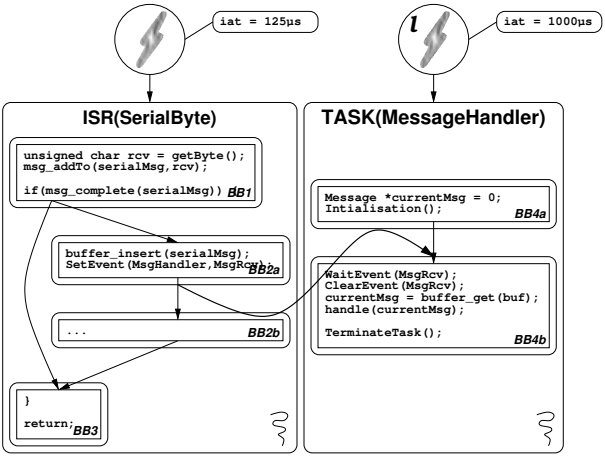


Abbildung 2: Globaler ABB-Graph des Beispiels aus Abbildung 1

Abhängigkeit zu einem globalen ABB-Graphen verknüpft. `SetEvent`, die Quelle der Abhängigkeit, wurde dem ersten Fragment $BB2a$ des ehemaligen Grundblocks $BB2$ zugeordnet, während das Ziel `WaitEvent` nun Bestandteil des zweiten Fragments $BB4b$ des früheren Grundblocks $BB4$ ist. Die übrigen ABBs in diesem Beispiel werden durch künstliche ABB-Endpunkte abgeschlossen. Auf die Darstellung der ungerichteten, in den Operationen `buffer_insert` und `buffer_get` enthaltenen Abhängigkeit wurde zugunsten der besseren Übersicht verzichtet.

4 Systemmodell

Neben der internen, durch ABB-Graphen beschrieben Struktur einer Echtzeitanwendung, müssen in Echtzeitsystemen zeitliche Rahmenbedingungen befolgt werden, die das kontrollierte physikalische Objekt bestimmt. Diese Informationen werden für die Manipulation der Echtzeitsystemarchitektur benötigt. Die hier angestrebte Migration ereignisgesteuerter Systeme auf zeitgesteuerte Echtzeitsystemarchitekturen verwendet dieses Wissen beispielsweise für die Berechnung von Ablauftabellen. Demzufolge werden diese Eigenschaften in einem *Systemmodell* notiert und dem im nachfolgenden Abschnitt beschriebenen Transformationswerkzeug zur Verfügung gestellt.

Diese Informationen umfassen *Ereignisse*, die angeben, wie häufig bestimmte *Aufgaben* und die damit verbundenen Ereignisbehandlungen im Echtzeitsystem ausgelöst werden, und *Termine*, die den spätesten möglichen Fertigstellungszeitpunkt einer Ereignisbehandlung markieren. Weiterhin unterscheidet man *periodische Ereignisse*, von denen *Periode*, *Phase* und *Jitter* bekannt sind, und *nicht-periodische Ereignisse*, von denen man nur die *minimale Zwischenankunftszeit* kennt, sowie *physikalische* und *logische Ereignisse*. Physikalische Ereignisse werden durch das kontrollierte Objekt erzeugt, während logische Ereignisse

Zustandsübergänge innerhalb der Anwendung kenntlich machen, die gesonderte Behandlung erfordern. Die Unterscheidung physikalischer und logischer Ereignisse erlaubt eine wesentlich präzisere Angabe der zeitlichen Eigenschaften des Echtzeitsystems. Während sich die zeitlichen Eigenschaften physikalischer Ereignisse alleine durch die Analyse des zu kontrollierenden Objekts erschließen, werden die eines logischen Ereignisses nämlich auch durch die Anwendung selbst bestimmt. Zusätzliches Anwendungswissen kann so in die zeitliche Beschreibung des Systems eingebracht werden. Die Verknüpfung zwischen globalen ABB-Graphen und dem Systemmodell erfolgt, indem Ereignisse und Termine immer einer entsprechenden Ereignisbehandlung zugeordnet werden, die wiederum durch einen ABB-Graphen dargestellt wird.

In Abbildung 2 sind zwei Ereignisse als in Kreise gefasste Blitze abgebildet. Ein physikalisches Ereignis aktiviert dabei die Aufgabe **`ISR`**`(SerialByte)` mit einer minimalen Zwischenankunftszeit von 125 μs, die durch die Übertragungsrate der seriellen Schnittstelle bestimmt wird. **`TASK`**`(MsgHandler)` hingegen wird von einem logischen Ereignis ausgelöst, dessen größere minimale Zwischenankunftszeit von 1000 μs auch von der Größe einer Nachricht abhängt. Könnte man dieses Anwendungswissen nicht durch ein logisches Ereignis nutzbar machen, müsste man für beide Aufgaben die kürzere Zwischenankunftszeit von 125 μs annehmen, was zu einer deutlich pessimistischeren Betrachtung führen würde.

5 Der Real-Time Systems Compiler

Der *Real-Time System Compiler* (RTSC) ist ein Quelltexttransformationswerkzeug, um die Echtzeitsystemarchitektur eines Echtzeitsystems gezielt zu beeinflussen. Der RTSC verarbeitet hierfür ein als Quelltext vorliegendes *Quellsystem* und das zugehörige Systemmodell, das die behandelten Ereignisse und Termine mit den Ereignisbehandlungen im Quellsystem verknüpft und in Form einer *Aufgabendatenbank* bereitgestellt wird. Eine weitere Eingabe des RTSC ist die Aufgabendatenbank des Zielsystems. Sie beschreibt, durch welche im Quellsystem implementierten Ereignisbehandlungen die Ereignisse des Zielsystems behandelt werden sollen und welchen zeitlichen Beschränkungen sie unterliegen. Hier erfordert beispielsweise der Übergang von nicht-periodischen auf entsprechende abfragende Ereignisse eine Anpassung ihrer zeitlichen Eigenschaften. Auf eine algorithmische Übertragung der Aufgabendatenbanken vom Quell- ins Zielsystem wurde aber bewusst verzichtet, um gezielt auf abweichende zeitliche Eigenschaften des Zielsystems eingehen zu können. Ausgabe des RTSC ist die Implementierung des Zielsystems, das die transformierten Ereignisbehandlungen des Quellsystems enthält.

Abbildung 3 stellt die an einen Übersetzer angelehnte Struktur des RTSC schematisch dar. Das Front-End und das Back-End hängen jeweils von der Echtzeitsystemarchitektur des Quell- beziehungsweise des Zielsystems ab. Die Implementierung des Middle-Ends hingegen ist architekturunabhängig, es wird aber durch das Zielsystem beeinflusst, was in der Abbildung durch entsprechende Schattierungen verdeutlicht wird. Die gerichteten, durchgezogenen Pfeile stehen für Ein- und Ausgabebeziehungen zwischen dem RTSC und dem Quell- beziehungsweise Zielsystem, die gepunkteten Pfeile zwischen den Implementierungen des Quell- und des Zielsystems und den Aufgabendatenbanken deuten die Verknüpfung der Ereignisse in der Aufgabendatenbank und den Ereignisbehandlungen

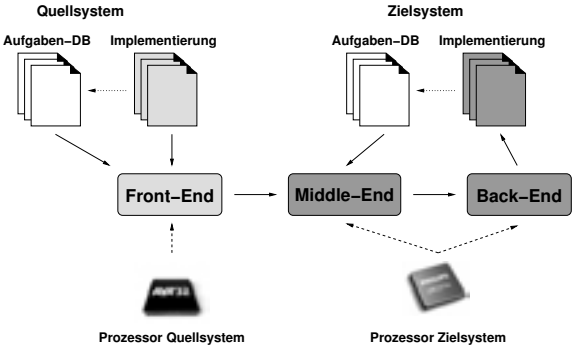


Abbildung 3: Grobgliederung des RTSC in Front-End, Middle-End und Back-End

in der Implementierung an. Die gestrichelten, von den Prozessoren des Quell- und des Zielsystems ausgehenden Pfeile kennzeichnen prozessorabhängige Schritte im RTSC. Die Prozessorabhängigkeit schlägt sich vor allem in der Analyse der maximalen Ausführungszeit nieder, die in verschiedene Analysen und Transformationsschritte eingebracht wird.

Grundsätzlich ist der RTSC als generisches Transformationswerkzeug zwischen verschiedenen Echtzeitsystemarchitekturen konzipiert, diese Dissertation konzentriert sich aber exemplarisch auf die Migration ereignisgesteuerter auf AUTOSAR OS basierender Echtzeitsysteme auf die zeitgesteuerte Architektur von OSEKtime. Ferner unterstützt der RTSC derzeit noch keine verschiedenen Prozessortypen im Quell- und Zielsystem, wie es in Abbildung 3 angedeutet ist, auch wenn dies prinzipiell möglich wäre.

5.1 Front-End

Das Front-End extrahiert einen globalen ABB-Graphen aus dem Quellsystem, der gerichtete und ungerichtete Abhängigkeiten innerhalb des Quellsystems unabhängig von dessen Echtzeitsystemarchitektur darstellt. Hierzu werden zunächst die Elemente des Quellsystems identifiziert, die Ereignisbehandlungen implementieren. Anschließend wird in der kompletten Implementierung nach ABB-Endpunkten gesucht, und es werden lokale ABB-Graphen erzeugt, die an den ABB-Endpunkten zu einem globalen ABB-Graphen verknüpft werden.

5.2 Middle-End

Im Middle-End wird der globale ABB-Graph mithilfe des statischen Ablaufplanungsalgorithmus von Adelzaher und Shin [AS99] in einer statischen Ablauftabelle zeitlich geordnet. Dazu werden zunächst Abhängigkeitsmuster aus dem ABB-Graphen entfernt, die der gewählte Algorithmus nicht direkt verarbeiten kann. Zu diesen Mustern zählen beispielsweise auch Funktionsaufrufe. Mehrfach aufgerufene Funktionen führen zu ABBs, die von mehreren Vorgängern aktiviert werden. Der Algorithmus geht aber davon aus, dass immer alle Vorgänger eingeplant werden müssen, bevor ein gemeinsamer Nachfolger ausführungsbereit wird. Bei mehrfachen Funktionsaufrufen wird der lokale ABB-Graph

der aufgerufenen Funktion beispielsweise entsprechend oft vervielfältigt, bis die Einstiegs-ABBs dieser Funktionen nur noch einen einzigen Vorgänger besitzen. Durch eine statische Analyse wird außerdem die maximale Ausführungszeit der einzelnen ABB bestimmt, ohne die eine statische Ablaufplanung nicht möglich wäre.[1]

Die hier angewandten Transformationen selbst sind unabhängig von der jeweiligen Echtzeitsystemarchitektur, es hängt aber von ihr ab, ob sie benötigt werden. In diesem Sinne wird das Middle-End zwar von der Echtzeitsystemarchitektur des Zielsystems beeinflusst, auf die Implementierung der verwendeten Transformationen wirkt sie sich aber nicht aus.

5.3 Back-End

Im letzten Schritt wird das aufbereitete Quellsystem auf die Elemente der Echtzeitsystemarchitektur des Zielsystems abgebildet. Dabei wird ein Anwendungsrumpf erstellt, der die transformierten Ereignisbehandlungen aktiviert, und die im Middle-End berechnete statische Ablauftabelle wird als Konfiguration für das OSEKtime-Betriebssystem ausgegeben. Weiterhin findet im Back-End auch die Übersetzung der transformierten Echtzeitanwendung in ein für den Prozessor des Zielsystems geeignetes Maschinenprogramm statt.

5.4 Implementierung

Als Grundlage für den RTSC wurde das LLVM-Projekt [LA04] ausgewählt, das ein Baukastensystem für die Entwicklung eigener Übersetzer anbietet und sämtliche Anforderungen des RTSC erfüllt. Die LLVM stellt eine Fülle von Analysen und Transformationen aus dem Bereich des Übersetzerbaus zur Verfügung und gestattet auf einfache Weise den Zugriff und die Manipulation der Kontrollflussgraphen der vorliegenden Echtzeitanwendung, was für die Erzeugung und die Transformation des ABB-Graphen notwendig ist. Die Abstraktion ABB und die zugehörigen ABB-Graphen konnten mithilfe der LLVM direkt als Aggregation einzelner, von der LLVM angebotener Grundblöcke implementiert und als zusätzliche Abstraktionsebene über den Kontrollflussgraphen gelegt werden.

6 Zusammenfassung und Ausblick

Die hier zusammengefasste Dissertation ebnet den Weg für eine gezielte Manipulation der Echtzeitsystemarchitektur eines Echtzeitsystems, um es so werkzeuggestützt auf seinen Einsatzzweck zuzuschneiden. Ermöglicht wird dies durch die Abstraktion *Atomic Basic Block*, die mit den Mitteln der Echtzeitsystemarchitektur hergestellte gerichtete und ungerichtete Abhängigkeiten architekturunabhängig darstellt. Ein *Systemmodell* kapselt die zeitlichen Eigenschaften der physikalischen Umwelt und verknüpft sie mit der durch ABB-Graphen beschriebenen Echtzeitanwendung. Darauf aufbauend wurde mit dem *Real-Time Systems Compiler* erstmalig ein Werkzeug geschaffen, um die Manipulation der Echtzeitsystemarchitektur automatisiert und ohne manuelle Eingriffe vorzunehmen.

[1] Die Analyse erfolgt derzeit mit Hilfe des Werkzeugs Absint aiT (`http://www.absint.de/ait`).

Im Rahmen des durch die DFG geförderten Projekts „Aspektorientierte Echtzeitsystemarchitekturen“ (AORTA) werden die Arbeiten an den Atomic Basic Blocks und dem RTSC fortgeführt. Dabei soll zum einen die Manipulation der Echtzeitsystemarchitektur mit der Umkehrung der hier entwickelten Migration, nämlich dem Übergang von zeit- auf ereignisgesteuerte Systeme vervollständigt werden. Zum anderen soll der Begriff „Echtzeitsystemarchitektur“ um verteilte Systeme sowie Mehrkernsysteme erweitert werden. Schließlich beeinflussen auch die Interaktionsmöglichkeiten über gemeinsamen Speicher oder ein dediziertes Kommunikationssystem die interne Struktur einer Echtzeitanwendung und stellen somit eine Facette der Echtzeitsystemarchitektur dar.

Literatur

[AS99] Tarek F. Abdelzaher und Kang G. Shin. Combined Task and Message Scheduling in Distributed Real-Time Systems. *IEEE TPDS*, 10(11):1179–1191, 1999.

[AUT09] AUTOSAR. Specification of Operating System (Version 4.0.0). Bericht, Automotive Open System Architecture GbR, Dezember 2009.

[Car84] Gene D. Carlow. Architecture of the space shuttle primary avionics software system. *CACM*, 27(9):926–936, 1984.

[LA04] Chris Lattner und Vikram Adve. LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation. In *Proceedings of the 2004 International Symposium on Code Generation and Optimization (CGO'04)*, Mar 2004.

[OSE01] OSEK/VDX Group. Time Triggered Operating System Specification 1.0. Bericht, OSEK/VDX Group, Juli 2001. `http://portal.osek-vdx.org/files/pdf/specs/ttos10.pdf`.

[SG90] T. Shepard und M. Gagne. A model of the F18 mission computer software for pre-run-time scheduling. In *10th Int. Conf. on Dist. Comp. Sys. (ICDCS '90)*, Seiten 62–69, Washington, DC, USA, Mai 1990. IEEE.

Fabian Scheler geboren 1981, begann im Jahr 2000 ein Informatik-Studium in Erlangen, welches er 2005 mit Auszeichnung abschloss. Im Anschluss daran trat er dort eine Stelle als wissenschaftlicher Mitarbeiter bei Prof. Dr.-Ing. Wolfgang Schröder-Preikschat am Lehrstuhl für Verteilte Systeme und Betriebssysteme an. Seine Tätigkeit verband er mit einer Lehrtätigkeit und intensiven Forschungsarbeiten auf den Gebieten Echtzeitsysteme, eingebettete Systeme und Betriebssysteme, die schließlich in seiner Dissertation mündeten. Im April 2011 promovierte Fabian Scheler schließlich mit Auszeichnung an der Technischen Fakultät der Friedrich-Alexander-Universität. Nach einem kurzen Aufenthalt bei der AREVA NP GmbH kehrte er im August 2011 an den Lehrstuhl für Verteilte Systeme und Betriebssysteme zurück, um die während seiner Doktorarbeit begonnene Forschung an *Atomic Basic Blocks* und dem *Real-Time Systems Compiler* im Rahmen des von der DFG geförderten AORTA-Projekts fortzuführen.

Ground Station Networks for Efficient Operation of Distributed Small Satellite Systems

Marco Schmidt

Institut für Informatik
Universität Würzburg
marco.w.schmidt@googlemail.com

Abstract: Satellitenformationen und Konstellationen rücken immer mehr in den Fokus aktueller Weltraumforschung, ausgelöst durch die jüngsten Fortschritte in der Kleinsatelliten-Entwicklung. Der Einsatz von verteilten Weltraumsystemen ermöglicht die Realisierung von innovativen Anwendungen auf Basis von hoher zeitlicher und räumlicher Auflösung in Observationsszenarien. Allerdings bringt dieses neue Paradigma der Raumfahrttechnik auch Herausforderungen in verschiedenen Forschungsfeldern mit sich. In dieser Arbeit werden neue Netzwerk-Konzepte für Raumfahrtmissionen unter Einsatz von Bodenstationsnetzwerken vorgestellt. Die präsentierten Verfahren koordinieren verfügbare Bodenstationsressourcen um einen robusten und effizienten Kommunikationslink zu ermöglichen.

1 Einführung

In den letzten 10 Jahren konnte in der Weltraumforschung eine Entwicklung zu immer kleineren Satelliten beobachtet werden. Dies stellt auch spannende Herausforderungen an die Informatik, da die Defizite durch die Miniaturisierung (z.B. erhöhte Störanfälligkeit) durch die Borddatenverarbeitung zu kompensieren sind. Insbesondere stehen hier Informatikmethoden für einen effizienten Betrieb des Satelliten durch vernetzte Bodenstationen im Mittelpunkt dieser Arbeit.

Der Trend zu immer kleineren Satelliten entstand ursprünglich aus dem akademischen Umfeld, welches die miniaturisierten Satelliten vor allem zur Ausbildung und Technologiedemonstration nutzten. Mittlerweile findet das Kleinsatellitenkonzept auch bei kommerziellen Forschungsinstituten und Raumfahrtagenturen immer mehr Anklang. Für Kleinsatelliten haben sich inzwischen die Begriffe Picosatellit (ein Satellit bis 1 kg Gesamtmasse) und Nanosatellit (bis 10 kg Gesamtmasse) durchgesetzt. Die Anzahl der in den letzten Jahren in den Orbit gebrachten Pico- und Nanosatelliten ist enorm gestiegen. Es findet also ein Paradigmenwechsel von konventionellen, großen, multifunktionellen Satelliten zu kleinen Satelliten mit geringer Masse für Einzelaufgaben statt.
Die Universität Würzburg, Lehrstuhl für Informatik 7, trägt mit dem UWE Programm (Universität Würzburg Experimentalsatellit) seit Beginn maßgeblich zu der Standardisierung und Entwicklung von Kleinsatelliten bei. So konnte schon mit UWE-1 (siehe Bild 1)

demonstriert werden, wie Kleinsatelliten als Experimentierplattform effizient genutzt werden können [SZ06]. Die zu dem Themenbereich ÏP in Space"durchgeführten Experimente wurden mehrfach international ausgezeichnet. UWE-1 war der erste deutsche Picosatellit der erfolgreich in einen niedrigen Erdorbit gebracht wurde. Das Ziel des UWE Projektes ist in naher Zukunft Satelliten-Schwärme und Formationen im Orbit zu etablieren, allerdings gibt es in dem Gebiet der verteilten Satellitennetze noch eine Reihe an offenen Forschungsfragen.

Abbildung 1: Der UWE-1 Satellit der Universität Würzburg

Diese Arbeit beschäftigt sich daher mit der effizienten Nutzung von Ressourcen für den Betrieb von verteilten Kleinsatellitenmissionen. Dabei werden vor allem zwei Problemstellungen näher untersucht und Verfahren zur Lösung vorgestellt. Der erste Schwerpunkt der Arbeit beschäftigt sich mit Satellite Scheduling, d.h. der Zuordnung von Satelliten zu Bodenstationen. Das Problem des Satellite Range Scheduling ist np vollständig [BWH04]. Bisherige Ansätze bezogen sich allerdings hauptsächlich auf die Optimierung der Auslastung in einem Bodenstationsnetz. Im Falle von Kleinsatelliten kommt es aber weniger auf die optimale Auslastung an, es ist eher wichtig flexibel auf Ausfälle reagieren zu können. Aus diesem Grund wird ein neuartiger Ansatz vorgestellt, das sogenannte Redundant Satellite Scheduling (RRS) [SS09], welcher besser die Bedürfnisse von Kleinsatellitenmissionen erfüllt.
Der zweite Schwerpunkt der Arbeit beschäftigt sich mit Datenmanagement in Bodenstationsnetzen. Das Empfangen eines Satelliten mit mehreren Bodenstationen gleichzeitig wurde erst durch das Kleinsatellitenkonzept intensiv umgesetzt. Daher gibt es bisher keine Ansätze welche die Synchronisation von Satellitendaten auf mehreren, lose gekoppelten Bodenstationen behandelt. Im Rahmen der Arbeit wurde ein System entwickelt, welches die Satellitendaten im Netz nicht nur selbstständig synchronisiert, sondern dies auch für die Erkennung und Korrektur von Übertragungsfehlern nutzt. Gerade für die oft stark gestörten Funkkanäle von Kleinsatelliten ist dies ein idealer Ansatz um die Linkqualität zu verbessern.

2 Eigenschaften und Besonderheiten von Kleinsatelliten und Bodenstationsnetzwerken

Die Vorteile des Kleinsatellitenkonzepts sind vielfältig, vor allem sind die geringen Startkosten ein Grund für den großen Erfolg. Denn die Startkosten eines Satelliten korrelieren direkt mit der Gesamtmasse des Satelliten, auf diese Weise ist es möglich dass Universitäten auch mit geringem Budget regelmäßig eigene Kleinsatelliten ins Weltall schicken. Ein wichtiger Aspekt, der das Kleinsatellitenkonzept für die Forschung besonders interessant macht, ist die Möglichkeit erstmals ganze Satelliten-Schwärme oder Formationen effizient in einen niedrigen Erdorbit zu platzieren. Auf Grund der geringen Masse der einzelnen Satelliten kann eine große Anzahl von Satelliten mit einer einzigen Rakete ins Weltall gebracht werden. Die Implementierung solcher Multi-Satelliten Systeme ermöglicht die Umsetzung neuartiger Anwendungen: So können mehrere Satelliten in einer Formation zu einem virtuellen Instrument, z.B. zu einem virtuellen Teleskop, zusammengeschaltet werden. Sensornetzwerke bestehend aus Satelliten können zur hochauflösenden Messung von physikalischen Größen, z.B. Atmosphärenparameter für die Weltraumwetterforschung, eingesetzt werden. Die Einsatzmöglichkeiten von verteilten Satellitensystemen für Raumfahrtanwendungen sind vielfältig, viele Forscher betrachten den Schritt zu verteilten System im Weltraum als die nächste Generation der Raumfahrtmissionen.
Allerdings bergen verteilte Satelliten-Systeme auch einige Herausforderungen. Nehmen wir beispielhaft einen Schwarm mit 10 Satelliten und ein verfügbares Netz mit 30 Bodenstationen an. In konventionellen Missionen wird ein einzelner Satellit von einer einzelnen Bodenstation gesteuert. Im verteilten Fall treten allerdings ganz neue Problemstellungen auf: Wie können Daten effizient von einem Satellitennetz im Weltraum zu einem Netz von Bodenstationen übertragen werden? Welcher Satellit sollte zu welcher Zeit mit welcher Bodenstation Kontakt aufnehmen? Diese Fragen scheinen auf den ersten Blick trivial zu sein, denn in terrestrischen Anwendungen gibt es bereits eine Reihe an Lösungen. Bei genauerem Hinsehen treten allerdings sowohl theoretische, als auch praktische Probleme auf, z.B. die hohe Dynamik der Satelliten oder die beschränkten Ressourcen von Kleinsatelliten. Es gilt daher Wege zu finden, wie vorhandene Flaschenhälse umgangen werden können. Der in dieser Arbeit verfolgte Ansatz bezieht sich dabei auf die Verwendung von vernetzten Bodenstationen.

Im Rahmen von universitären Kleinsatellitenprojekten wurden meist auch Bodenstationen zum Empfang der Satelliten eingerichtet. Diese Bodenstationen sind aus COTS Komponenten aufgebaut und entsprechen im Wesentlichen einer einfachen Funkstation. Im Gegensatz zu den Anlagen der großen Raumfahrtbehörden zeichnen sich diese akademischen Bodenstationen durch eine simple Architektur aus. Zudem gibt es keine klare Unterscheidung mehr in Missions- und Kontrollzentrum. Solch eine Bodenstation ist auch an der Universität Würzburg vorhanden (siehe Bild 2), diese kann direkt für den Kontakt mit

den Satelliten der UWE Reihe benutzt werden. Da die Kleinsatelliten meist in niedrige Erdorbits gebracht werden, ist deren Kommunikationszeit mit den Bodenstationen auf wenige Minuten am Tag beschränkt. Um die Kontaktzeit zu erhöhen, werden Bodenstationen in diversen Projekten (z.B. GENSO [SK07] oder GSN [NN06]) miteinander vernetzt. Die Nutzung mehrerer, vernetzter Bodenstationen stellt eine einfache Möglichkeit dar die Kommunikationszeit mit den einzelnen Satelliten zu erhöhen.

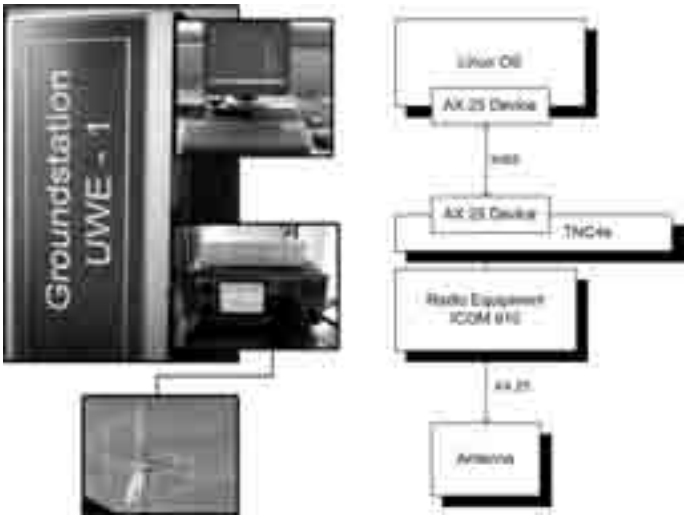


Abbildung 2: Bodenstation der Uni Würzburg

Bodenstationsnetzwerke wie GENSO entstanden aus den Funkstationen von Kleinsatellitenprojekten und besitzen einige Besonderheiten, diese Netze können nicht direkt mit den großen, klassischen Bodenstationsnetzwerken der Raumfahrtagenturen verglichen werden. Bereits die Topologie unterscheidet sich sehr voneinander: Klassische Bodenstationsnetzwerke werden aus sehr unterschiedlichen Empfangsanlagen zusammengeschaltet, die meistens nur für einen Missionstyp benutzt werden können. Diese Empfangsanlagen werden zentral gesteuert und gewartet. Der Ansatz aus dem Kleinsatellitenumfeld hingegen besteht aus einer großen Anzahl von sehr ähnlichen Bodenstationen, welche nur sehr lose miteinander gekoppelt sind. Die Bodenstationen können typischerweise über das Internet miteinander kommunizieren und Daten austauschen, jedoch liegt hier eher eine dezentrale Kontrolle vor. Jede Bodenstation ist mit einem Forschungsinstitut assoziiert, welches für die entsprechende Bodenstation zuständig ist. Es gibt daher in Netzwerken wie GENSO keine zentrale Instanz, die eine kollektive Wartung oder den Ausfall einer Bodenstation ausgleichen kann. Allerdings kann durch die große Anzahl der gleichartigen Bodenstationen durch intelligente Vernetzung dieser Nachteil ausgeglichen werden.

Die Motivation der Arbeit war die zugrunde liegende Infrastruktur an Bodenstationen zu nutzen um den Betrieb von Kleinsatelliten effizient zu gestalten. Gerade die Struktur sol-

cher Netze, d.h. viele, lose gekoppelte Bodenstationen, lässt sich nutzen um einen Mehrwert für Kleinsatellitenmissionen zu generieren. Da die Stationen untereinander kompatibel sind, ist es möglich durch eine intelligente Vernetzung die verteilte Struktur effizient einzusetzen. Die Arbeit behandelt dabei vor allem zwei Aspekte, die im nächsten Abschnitt ausführlich beschrieben werden.

3 Beiträge der Arbeit

Die Arbeit umfasst zwei Schwerpunkte: Der erste Teil beschäftigt sich mit dem Themenfeld des Scheduling, d.h. der Zuordnung von Satelliten, bzw. Kontaktfenstern zu Bodenstationen. Im zweiten Teil wird ein Datenmanagement System vorgestellt, mit welchem die Kommunikationslinks von Kleinsatelliten effizienter genutzt werden können.

3.1 Scheduling

Die Kontaktfenster zwischen Satellit und Bodenstation sind bei niedrigen Erdorbits auf maximal 15 Minuten begrenzt, diese treten ungefähr zwei bis dreimal pro Tag auf. Daher ist eine Bodenstation mit einem Low Earth Orbit (LEO) Satelliten nur zu etwa 2 % am Tag im Einsatz. Mit mehreren Satelliten kann natürlich eine höhere Auslastung erreicht werden, allerdings ist das auftretende Planungsproblem sehr komplex. Die Zuordnung von Kontaktfenstern zu Bodenstationen wird als Satellite Range Scheduling (SRS) Problem bezeichnet (siehe Bild 3) und ist np vollständig.

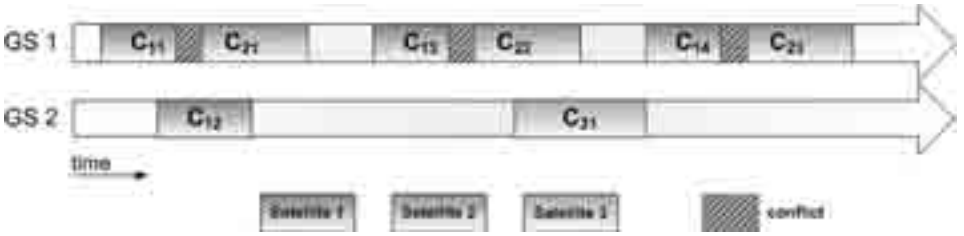


Abbildung 3: Zuordnung von Kontaktfenstern zu zwei Bodenstationen

Die großen Raumfahrtagenturen haben sich schon ausführlich mit diesem Problem beschäftigt, vor allem um die Auslastung ihrer Bodenstationsnetze zu steigern. Die hier erforschten Verfahren lassen sich allerdings nicht einfach auf Kleinsatelliten übertragen. Dies liegt an sehr unterschiedlichen Anforderungen an das Scheduling: Während die Raumfahrtagenturen über große Zeiträume planen müssen, wird bei Kleinsatelliten eher ein kurzer Zeithorizont bevorzugt. Zudem muss bei Kleinsatelliten jederzeit mit Ausfällen gerechnet werden, das Scheduling Verfahren muss also sehr flexibel sein. In dieser Arbeit

wurde eine ausführliche Analyse der Anforderungen des klassischen Ansatzes der Raumfahrtagenturen mit den neuen Ansatz aus dem Kleinsatellitenbereich durchgeführt. Ein besonders wichtiger Unterschied ist, dass bei den großen Raumfahrtagenturen die Bodenstationsnetze an Kunden vermietet werden, d.h. auf Anfrage werden einzelne Bodenstationen für andere Forscher reserviert. Im Gegensatz hierzu werden im Kleinsatellitenbereich die Bodenstationsressourcen "freiwillig"geteilt, d.h. mehrere Institute nutzen Bodenstationen gemeinsam um den Kontakt zu den Satelliten auszubauen. Dies geschieht ohne Mietkosten, da beide Parteien von den gemeinsam genutzten Ressourcen profitieren. Jedoch hat diese Art der Ressourcennutzung einen wichtigen Einfluss auf das Scheduling: Während bei den Raumfahrtagenturen die Anfragen auf Kontaktzeiten sehr genau und spezifisch definiert sind, wird im Kleinsatellitenbereich versucht möglichst viele freie Ressourcen für den eigenen Satelliten zu erhalten. Es ist also durchaus möglich dass mehrere Bodenstationen gleichzeitig einen Satelliten verfolgen, wenn diese Bodenstationen nicht anderweitig im Einsatz sind. Dies führt dazu, dass die Anfragen auf Kontaktzeiten (Scheduling Requests) nicht genau definiert sind, sondern nach einer variablen Anzahl an Ressourcen fragen. Dieses Problem wurde in der Arbeit als „Redundant Scheduling“ eingeführt und näher untersucht.

Da bisherige Ansätze aus dem Bereich des Satellite Range Scheduling nicht für die speziellen Anforderungen von Kleinsatelliten geeignet sind, wurde ein neues Verfahren entwickelt. Dieser Ansatz besteht im Wesentlichen aus einer Kostenfunktion, die es ermöglicht den Aspekt des redundanten Scheduling zu integrieren. So kann der Grad der Redundanz in den erzeugten Plänen gesteuert werden. Es wurde mathematisch gezeigt, wie die Kostenfunktion sich in unterschiedlichen Szenarien verhält. Für die Suche nach einem geeigneten Zeitplan wurden zwei Suchverfahren implementiert. Zum Einen eine Hillclimber Suche, zum Anderen eine Tiefensuche in Kombination mit dem Branch and Bound Verfahren. Das vorgestellte Verfahren wurde in Software implementiert und mit einer Reihe von Problemtypen und -größen evaluiert. Dabei wurde darauf geachtet möglichst realitätsnahe Probleme zu erstellen, indem die Orbitdaten von bereits gestarteten Satelliten und die Lage von existierenden Bodenstationen verwendet wurden. Die Größe der Probleme wurde durch die Anzahl der Satelliten und Bodenstationen variiert. Untersucht wurde die durchschnittlich benötigte Zeit, die für die Erstellung des Schedule benötigt wurde, sowie die Anzahl der nicht eingeplanten Anfragen (Unsatisfied requests).

In den Experimenten konnte gezeigt werden, dass die Hillclimber Suche für fast alle Problemgrößen sehr gut funktioniert. So ist es immer möglich mit dieser Suche in weniger als 5 Sekunden einen neuen Schedule zu erzeugen, das System garantiert dadurch ein hohes Maß an Flexibilität bei Ausfällen einzelner Bodenstation. Der Einzige Schwachpunkt des Hillclimbing Algorithmus sind Problemtypen in welchen eine wesentlich größere Anzahl an Satelliten als Bodenstationen untergebracht werden müssen. Hier kommt zusammen, dass die Ressourcen durch die geringe Anzahl an Bodenstationen sehr begrenzt sind, außerdem hängt der Hillclimber Algorithmus leicht in lokalen Minima fest. Daher sollte bei stark überzeichneten Problemen (Oversubscribed Scheduling) eher das Branch and Bound Verfahren verwendet werden. Ein wichtiges Ergebnis ist, dass das Problem der „Unsatis-

fied requests“ elegant umgangen werden kann. Während andere Arbeiten aus dem Bereich des Satellite Range Scheduling meist versuchen die Anzahl der „Unsatisfied requests“ zu minimieren, kann dies hier durch eine Vergrößerung des Zeithorizonts mit einfachen Mitteln erreicht werden. Dies ist allerdings nur möglich, da zum Einen der Zeithorizont bei Kleinsatelliten sehr flexibel ist, und zum Anderen da innerhalb weniger Sekunden ein neuer (suboptimaler) Schedule erzeugt werden kann. Auf diese Weise kann dafür gesorgt werden dass nicht zu viele Anfragen aus dem finalen Plan fallen.
Das entwickelte System ist besonders gut geeignet für kommende, hochverteilte Raumfahrtmissionen. Gerade in Anbetracht der Tatsache dass zukünftige Missionen eine große Anzahl von Kleinsatelliten einsetzen werden (Missionen mit 50 Satelliten in einer Rakete sind bereits genehmigt), verdeutlicht die Notwendigkeit von geeigneten Scheduling Verfahren.

3.2 Data Management

Momentan setzen Kleinsatelliten fast ausschließlich auf Funkfrequenzen aus dem Amateurfunkbereich. Diese werden meistens in Verbindung mit einfachen Dipol Antennen verwendet, welche eine fast omnidirektionale Abstrahlcharakteristik besitzen. Die Empfangskegel auf der Erdoberfläche sind daher relativ groß, ein Picosatellit über Deutschland kann in ganz Europa empfangen werden. Bodenstationsnetzwerke wie GENSO umfassen sehr viele Stationen in Europa, die zu einem Netz zusammengeschaltet sind. Daher werden oft mehrere Stationen gleichzeitig verwendet um einen Satelliten zu verfolgen (bzw. mit der Antenne zu tracken). Dies bedeutet, dass mehrere Empfangsstationen die Daten des Satelliten gleichzeitig empfangen. Der parallele Empfang hat den Vorteil dass typischerweise mehr Daten dekodiert werden können. Die Funklinks von Kleinsatelliten sind fehleranfällig und das redundante Tracken mit mehreren Stationen ermöglicht daher einen besseren Datendurchsatz. Allerdings liegen dem Operator momentan lediglich die Datenströme der einzelnen Bodenstationen vor, d.h. ein Datenpaket kann parallel empfangen in jedem Datenstrom vorliegen, es kann aber nicht erkannt werden ob es sich um ein identisches Datenpaket handelt, oder um verschiedene Datenpakete. Der Grund hierfür ist, dass die Empfangszeit der Datenpakete voneinander abweicht und der Dateninhalt durch Übertragungsfehler verfälscht sein kann. Zudem werden Housekeeping-Daten meist ohne eindeutige ID verschickt.
Ziel der Arbeit war daher ein System zu entwickelt, welches automatisch die Datenströme synchronisiert und sortiert, so dass der Operator aus den parallel empfangenen Datenströmen einen virtuellen Datenstrom erhält mit den empfangenen Datenpaketen aus den unterschiedlichen Bodenstationen. Anders ausgedrückt kann man sagen, dass Datenpakete rein an ihrer Empfangszeit zugeordnet werden, um mehrere, parallel empfangene Datenpakete miteinander vergleichen zu können.

Die erste Problemstellung ist die Synchronisation der Daten in einem Bodenstationsnetzwerk. Zwar sind bereits eine Reihe an Techniken zur Zeitsynchronisation (z.B. NTP) weit

verbreitet, allerdings kann für eine entfernte Bodenstation nicht garantiert werden, dass die Zeit synchronisiert ist. Daher wurde die Anforderung an das System gestellt, dass es selbstständig für die Zeitsynchronisation der Bodenstation sorgen muss. Um nun auch die Datenpakete (bzw. deren Empfangszeit) zu synchronisieren müssen mehrere Probleme gelöst werden: In Bild 4 ist dargestellt welche Einflüsse auf die Empfangszeit eines Datenpakets von einem Satelliten wirken. Zum Einen ist die Signallaufzeit abhängig von der geographischen Lage der Bodenstation, hinzu kommt eine Systemverzögerung durch die Bodenstation selbst. Werden nun die Daten von einem Server gesammelt und verglichen, so muss auch noch die Verzögerung durch das Internet berücksichtigt werden. Daher ist es nicht möglich die Empfangszeiten der einzelnen Pakete miteinander zu vergleichen, die Inhalte der Datenpakete können durch Übertragungsfehler verfälscht sein.
Um die Datenpakete synchronisieren zu können, wurde folgender Ansatz entwickelt: Die Zeitsynchronisation der Bodenstationen untereinander wurde mit Hilfe eines modifizierten SNTP Protokolls durchgeführt. Die Systemverzögerung wurde durch eine verteilte Messung und einem Mittelwertfilter bestimmt. Die Verzögerung durch das Internet wurde mit Zeitstempeln umgangen. Mit Hardware-in-the-loop Experimenten wurde gezeigt, dass mit diesem Ansatz mehr als 99 % der Pakete zuverlässig synchronisiert werden konnten.

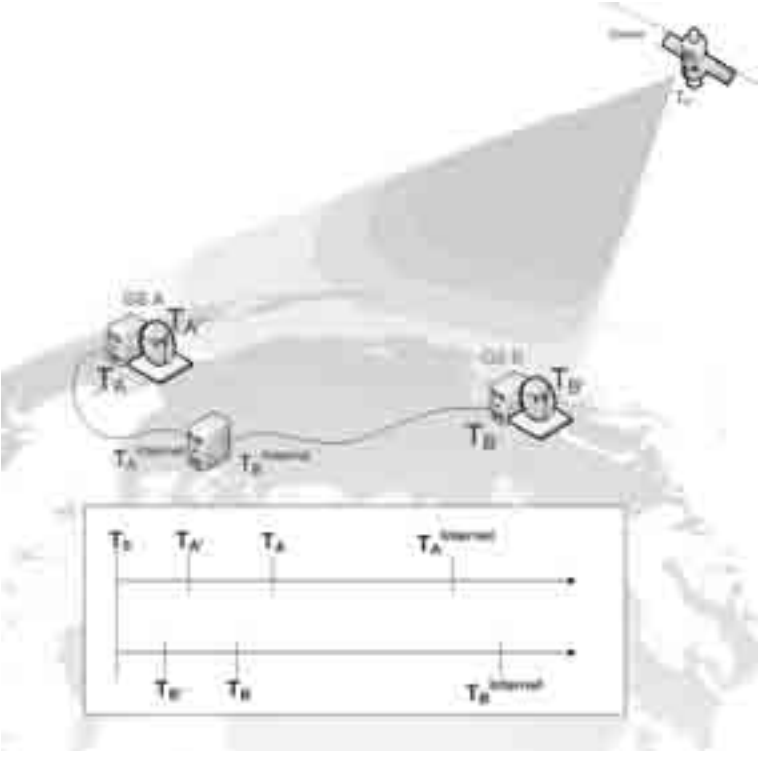


Abbildung 4: Paralleler Datenempfang von einem Satelliten

Auf Basis der synchronisierten Datenpakete sollen nun Übertragungsfehler erkannt und korrigiert werden. Das in Kleinsatelliten hauptsächlich eingesetzte AX.25 Protokoll unterstützt keine Vorwärtsfehlerkorrektur, sondern lediglich eine Fehlererkennung über ein CRC Feld. Es wurden nun verschiedene Verfahren verglichen um redundant empfangene Daten zur Fehlererkennung und Korrektur nutzen zu können. Besonders performant arbeitet der im Rahmen der Arbeit entwickelte Ground Station Majority Voting (GSMV) Algorithmus, der die Daten auf Bitebene vergleicht und Abweichungen im Mehrheitsentscheid korrigiert. Das CRC Feld wird anschließend genutzt um zu überprüfen ob die Korrektur erfolgreich war. In Experimenten konnte gezeigt werden, dass sich mit diesem System die Bitfehlerrate schon mit kleinen Bodenstationsnetzen signifikant reduzieren lässt. Gerade im Hinblick auf die beschränkten Funkkanäle heutiger Kleinsatelliten kann mit diesem Verfahren die Kommunikation beträchtlich optimiert werden.
Weiterer Vorteil des Systems ist, dass der Einsatz auf jeglichen Bodenstationsnetzwerken anwendbar ist ohne die Hardware zu modifizieren. So kann bestehende Infrastruktur verwendet werden um den Funkkontakt zu bereits im Orbit befindlichen Satelliten effizienter zu nutzen. Mit diesen neuen Ansatz kann durch die intelligente Vernetzung von terrestrischen Stationen eine Mission zusätzlich unterstützt werden.

4 Zusammenfassung

Im Rahmen der Arbeit wurden zunächst die Herausforderungen und Anforderungen von verteilten Kleinsatellitensystemen abgeleitet. Der Fokus lag dabei auf dem Satellitenbetrieb verteilter Raumfahrtmissionen, speziell die optimale Ressourcennutzung mit Hilfe von Bodenstationsnetzwerken. Die Unterschiede zum klassischen Ansatz (große Einzelsatelliten in Verbindung mit dedizierten Bodenstationen) wurden ausführlich dargestellt.
Der erste Teil der Arbeit beschäftigt sich mit der Zuteilung von Kontaktfenstern zu Bodenstationen (Satellite Range Scheduling). Da bisherige Ansätze sich auf die Optimierung der Auslastung fokussieren, wurde ein neues Verfahren entwickelt welches besser die Anforderungen von verteilten Kleinsatellitensystemen erfüllt. Diese erfordern ein hohes Maß an Flexibilität, da es regelmäßig zu Ausfällen von einzelnen Stationen kommt. Dies wurde erreicht in dem auf eine optimale Lösung des Problems verzichtet wird, dadurch ist es aber möglich bei Bedarf innerhalb weniger Sekunden einen neuen Plan zu erstellen. Des Weiteren wurde der Aspekt des redundanten Scheduling vorgestellt, welcher es ermöglicht die Ressourcen von Bodenstationsnetzwerken besser zu nutzen. Das vorgestellte Verfahren wurde als das Redundant Request Scheduling Problem (RRSS) veröffentlicht und auf internationalen Konferenzen vorgestellt. Das implementierte System wurde ausführlich mit verschiedenen Problemgrößen und -typen evaluiert. Es konnte gezeigt werden, wie das entwickelte System für zukünftige, hochverteilte Kleinsatellitenmissionen eingesetzt werden kann.
Im zweiten Teil der Arbeit wurde der Bereich des Datenmanagements behandelt. Hier wurde ein System entwickelt, welches Datenpakete von Satelliten selbstständig in einem Bodenstationsnetz synchronisiert. Auf Basis dieser Synchronisierung können Übertragungs-

fehler erkannt und korrigiert werden. Für die Synchronisation der Satellitenlinks wurde ein Algorithmus vorgestellt, der für die Anforderungen in einem lose gekoppelten Bodenstationsnetz zugeschnitten ist. Ein mathematisches Modell für die erzielbare Korrekturrate wurde abgeleitet. Mit Hardware-in-the-loop Experimenten wurde das System evaluiert. Es konnte gezeigt werden, wie mit den momentan bereits verfügbaren Ressourcen die Bitfehlerrate im Übertragungskanal von Kleinsatelliten signifikant gesenkt werden kann.
Ob in Zukunft die Miniaturisierung im Bereich der Raumfahrt in diesem Tempo weiter fortschreiten wird ist schwer vorherzusagen, jedoch lässt sich an der großen Menge an geplanten Raumfahrtmissionen ganz klar ein Trend zu verteilten Anwendungen im Weltall erkennen. Die in dieser Arbeit vorgestellten Konzepte und Verfahren stellen einen wichtigen Schritt in die Richtung intelligent vernetzter Raumfahrtmissionen dar.

Literatur

[BWH04] L. Barbulescu, D. Whitley und A. How. Leap Before You Look: An Effective Strategy in an Oversubscribed Scheduling Problem. In *International Workshop on Scheduling a Scheduling Competition AAAI 2004*, Seiten 143–148, 2004.

[NN06] Y. Nakamura und S. Nakasuka. Ground Station Networks to Improve Operations Efficiency of Small Satellites and its Operation Scheduling method. In *IAC*, 2006.

[SK07] G. Shirville und B. Klofas. GENSO: A Global Ground Station Network. In *AMSAT Symposium*, 2007.

[SS09] M. Schmidt und K. Schilling. A Scheduling System with redundant scheduling capabilities. In *IWPSS*, 2009.

[SZ06] M. Schmidt und F. Zeiger. Design and Implementation of In Orbit Experiments for the Pico Satellite UWE-1. In *International Astronautical Congress*, number IAC-06-E2.1.07. 2006, 2006.

Marco Schmidt studierte Informatik an der Universität Würzburg, mit seiner Diplomarbeit zum ersten deutschen Picosatelliten UWE-1 schloss er das Studium erfolgreich ab. Er begann 2006 sein Promotionsstudium an der Universität Würzburg und arbeitete zugleich als wissenschaftlicher Mitarbeiter am Lehrstuhl für Robotik und Telematik. Nach seiner Promotion 2011 leitet er die Raumfahrtgruppe des Lehrstuhls. Marco Schmidt engagiert sich zudem besonders für die Förderung von Studenten als aktives Mitglied der Deutschen Gesellschaft für Luft- und Raumfahrt - Lilienthal-Oberth e.V (DGLR) und des Space Education and Outreach Committee (SEOC) der International Astronautical Federation (IAF). Seine Arbeiten im Bereich der Kleinsatelliten-Forschung wurden mehrfach national und international ausgezeichnet.

Konstruktion selbst-organisierender Softwaresysteme

Hella Seebach

Institut für Software und Systems Engineering, Universität Augsburg
seebach@informatik.uni-augsburg.de

Abstract: Selbst-Organisation ist die Fähigkeit eines Systems, ohne externe Eingriffe die eigene Struktur zu verändern. Sie ermöglicht es somit, dass ein System selbständig sowohl auf interne Veränderungen als auch auf Umweltveränderungen reagieren kann. Dieses Phänomen wird in vielen Disziplinen wie der Physik, der Philosophie und der Biologie untersucht. In der Informatik wurde die Forschung in Richtung dieser Eigenschaft ebenfalls intensiviert. Es wird analysiert, wie Selbst-Organisationsmechanismen, die in der Natur beobachtbar sind, auf Informatiksysteme übertragen werden können. Das Ziel ist es, die steigende Komplexität der Computersysteme in den Griff zu bekommen und die Systeme robuster gegenüber Veränderungen der Umwelt zu machen. In dieser Dissertation wurde ein Verfahren entwickelt, mit dem solche selbst-organisierenden Softwaresysteme top-down, standardisiert und reproduzierbar konstruiert werden können. Eine der größten Herausforderungen war dabei, mit dem nicht vorhersagbaren Verhalten der Systeme umzugehen.

1 Einführung

Die Anforderungen an Softwaresysteme der Zukunft nehmen rasant zu. Vor allem Verteiltheit, Flexibilität und Wiederverwendbarkeit werden gefordert. Zudem sind kurze Entwicklungszeiten wünschenswert. All diese Forderungen sind notwendig, um mit den ständigen Technologieentwicklungen und Anwenderanforderungen, die im Bereich der Informatik aufkommen, mithalten zu können. Die Leistungsfähigkeit und Vernetzung der Rechner steigt stetig und sie können immer komplexere Dienste übernehmen. Es sind somit die technischen Voraussetzungen für flexible, verteilte, vielseitige Systeme vorhanden. Es ist an der Zeit, dass die Softwaretechnik diese Entwicklungen annimmt und Techniken zur Verfügung stellt, mit dieser Komplexität umzugehen und um in diesem Bereich zuverlässige Software standardisiert zu entwickeln.

Die Nutzung der beschriebenen Möglichkeiten macht die Softwaresysteme jedoch noch komplexer und für den Benutzer schwer verwalt- und nachvollziehbar. Um diesen Problemen zu begegnen, entstanden die Forschungsgebiete *Autonomic Computing* [KC03] und *Organic Computing* [Sch05]. In den Bereich des Organic Computing ist diese Dissertation einzuordnen. Organic Computing geht im Vergleich zu Autonomic Computing über Serverarchitekturen hinaus und befasst sich generell mit technischen Systemen, die anhand lokaler Regeln zu einem selbst-organisierenden Gesamtverhalten führen, das „lebensähnlich“ (organisch) wirkt [MS04]. Das Verhalten eines selbst-organisierenden Systems zeigt oft sehr gute Eigenschaften bezüglich Skalierbarkeit und Robustheit gegenüber

Störeinflüssen oder Parameteränderungen, weshalb sich selbst-organisierende Systeme gut als Paradigma für zukünftige, komplexe technische Systeme eignen. Das Verhalten des Gesamtsystem ist im Allgemeinen nicht direkt aus den lokalen Regeln ableitbar. Problematisch ist dabei, dass lokale Regeln nicht unmittelbar nur zu positivem globalem Verhalten führen. Deswegen werden nicht nur Techniken zum Bau solcher Systeme benötigt, sondern auch Techniken, um „schlechtes“ globales Verhalten zu vermeiden.

Im Schwerpunktprogramm „Organic Computing“ der Deutschen Forschungsgemeinschaft [For10] wurden unter anderem Observer-/Controller-Architekturen [RMB+06], und Kooperationsmechanismen [PMGT08] für selbst-organisierende Systeme entwickelt, die ein allgemeines Vorgehen bei der Entwicklung selbst-organisierender Systeme vorschlagen. Trotz dieser generellen Ideen für Organic Computing Systeme befasst sich bisher kaum jemand mit der Fragestellung, wie solche hoch dynamischen Systeme durchgängig von Konzeptmodell bis hin zu Code modelliert und letztendlich konstruiert werden können. Genau diese Vorgehensweise ist jedoch notwendig, um trotz Selbst-Organisation und den damit unvorhersagbaren Verhaltensänderungen eine Akzeptanz der Systeme zu erreichen. Traditionelle Softwaretechnik kommt bei diesen komplexen selbst-organisierenden Systemen an ihre Grenzen [DMSGK06]. So werden beispielsweise Techniken benötigt, um die Flexibilität in die Systeme zu integrieren oder um den Systemen zur Laufzeit Entscheidungshilfen zu geben, damit diese sich „positiv/gewünscht“ verhalten.

Das Ziel dieser Arbeit war es, einen Ansatz für eine ganze Klasse von Systemen zu entwickeln, mit dem durchgängig und „top-down“ solche selbst-organisierenden, zuverlässigen Systeme entwickelt werden können. Mit so einem Vorgehen ist es möglich, die Systeme auf einer formalen Grundlage aufzusetzen und Verhaltensgarantien abzugeben obwohl Selbst-Organisation integriert ist. Leider kann es keinen allumfassenden Ansatz für jede Art von selbst-organisierenden Systemen geben, da diese inhärent zu unterschiedlich sind. Jedoch können die Systeme anhand ihrer Struktur und Eigenschaften klassifiziert werden und anschließend für diese Klassen von Systemen Techniken und Modelle entwickelt werden. Die Dissertation stellt Techniken für die Klasse der Ressourcenflusssysteme vor und gibt einen Ausblick auf weitere Systemklassen, die mit den entwickelten Techniken selbst-organisierend realisiert werden können. Die entwickelte Methodik umfasst mehrere Elemente, die in einer Software Engineering Guideline zusammengefasst werden. Die Guideline gibt dem Entwickler eine Anleitung an die Hand, wann welche Techniken wie anzuwenden sind. Die Techniken reichen von Softwaretechnik und Algorithmen bis hin zu Formalen Methoden, um erwünschtes globales Verhalten zu erreichen und unerwünschtes zu vermeiden.

2 Selbst-Organisation kontrollieren durch Verhaltenskorridore

Selbst-organisierende Systeme sind in der Lage, sich selber neu zu konfigurieren und sich im Falle von Änderungen ihres Zustandes (z.B. Ausfall einer Fähigkeit) bzw. der Umgebung an diese anzupassen. Diese Fähigkeiten der Systeme machen sie im Vergleich zu traditionellen, robuster, zuverlässiger und auch leichter zu warten, da kein menschliches Eingreifen mehr erforderlich ist. Jedoch stellt sich die Frage, wie sichergestellt werden

kann, dass diese Systeme auch nur das tun, wozu sie entworfen wurden. Der entwickelte *Restore Invariant Approach (RIA)* stellt eine Technik dar, die eine Grundlage für die Konstruktion solcher Systeme mit sogenannten Selbst-X-Eigenschaften (Selbst-Heilung, Selbst-Optimierung, etc.) schafft. Der RIA ist die Schnittstelle, die zentrale Idee, um Softwaretechniken und formale Analyse für selbst-organisierende Systeme auf eine gemeinsame Basis zu stellen und eine einheitliche Zielvorstellung zu entwickeln.

Die Grundidee ist es, dem System einen Verhaltenskorridor vorzugeben, in dem es sich völlig frei bewegen kann. Sobald das System durch Einflüsse wie Ausfälle, neue Komponenten oder neue Aufgaben diesen Korridor verlässt, greift ein Rekonfigurationsmechanismus ein, der das System wieder in den Korridor zurückbringt. Wie auch traditionelle Systeme sind selbst-organisierende Systeme prinzipiell nicht vor Fehlern oder neu auftretenden Ereignissen geschützt. Die Zustände, die ein selbst-organisierendes System durchlaufen kann, können somit unterschieden werden in funktionale Zustände $\mathcal{S}_{func}$, in denen das System seine Aufgabe erfüllt und Zustände $\mathcal{S}_{reconf}$, in denen das System gerade rekonfiguriert, um nach dem Auftreten einer Veränderung, wieder in einen funktionalen Zustand zurückzukehren. Für den Zustandsraum $\mathcal{S}$ des Systems gilt die folgende Bedingung:

$$\mathcal{S} := \mathcal{S}_{func} \cup \mathcal{S}_{reconf} \text{ , mit } \mathcal{S}_{func} \cap \mathcal{S}_{reconf} = \emptyset.$$

Das heißt, ein System ist immer entweder funktionsfähig oder wird gerade rekonfiguriert. Der Korridor zieht somit die Linie zwischen den funktionalen Systemzuständen und den nicht erwünschten Zuständen, in denen rekonfiguriert werden muss.

Formal kann das Aufsplitten der Zustände als prädikatenlogische Formel beschrieben werden, die zu *wahr* ausgewertet wird für die Zustände $\mathcal{S}_{func}$ und zu *falsch* für die Zustände $\mathcal{S}_{reconf}$. Nachdem die Formel für alle funktionalen Zustände wahr ist, wird von einer *Invariante*, die für das System gelten soll, gesprochen. Der Verhaltenskorridor für selbst-organisierende Systeme wird über Bedingungen (Constraints) spezifiziert, d.h. es wird auf den Systemkomponenten definiert, was zulässig ist und was nicht. Die Konjunktion all dieser Constraints bildet die Invariante. Um diese Bedingungen spezifizieren zu können, wird ein Modell des Systems benötigt (s. Kapitel 4). Auf diesem Modell werden die Bedingungen mit Hilfe der Object Constraint Language (OCL) spezifiziert. Diese müssen zur initialen Konfiguration und zur Laufzeit permanent ausgewertet werden. Bei der Verletzung eines Constraints und somit der Invariante, wird der Korridor verlassen und es muss eine neue Belegung für die Zustandsvariablen gefunden werden, so dass die Invariante wieder Gültigkeit hat (restore). Dazu muss die Invariante über Zustandsvariablen sprechen, denen neue Werte zugewiesen werden können. Dies ist der Fall, wenn das Produktivsystem Freiheitsgrade, sprich Redundanz enthält. Der Rekonfigurationsmechanismus muss also für eine Belegung sorgen, die einem Zustand des Produktivsystems entspricht, der innerhalb des Korridors liegt.

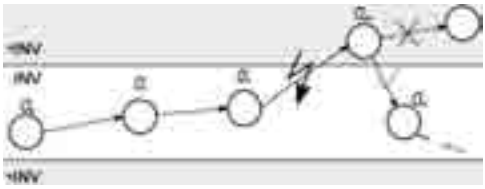


Abbildung 1: Verhaltenskorridor

Abbildung 1 zeigt den Zusammenhang zwischen Systemzuständen, Invariante (INV) und dem Fall, dass in dem Schritt von σ_2 zu σ_{err} das System einen Fehlerzustand betritt, der

außerhalb des Korridors liegt. Durch die Rekonfiguration betritt das System im nächsten Schritt σ_4 wieder den Korridor (mit $\sigma_0, \sigma_1, \sigma_2, \sigma_4 \in \mathcal{S}_{func}$ und $\sigma_{err} \in \mathcal{S}_{reconf}$).

Liegt ein solcher Verhaltenskorridor und eine Systemspezifikation für ein selbst-organisierendes System vor, können Aussagen über das Verhalten des Systems gemacht werden, bzw. über Eigenschaften, die das System innerhalb des Korridors aufweist.

Der RIA ermöglicht es einem Systementwickler, den funktionalen Teil des Systems herkömmlich zu beschreiben und zu modellieren und den Rekonfigurationsteil völlig separat zu betrachten. Der Teil, der für die Rekonfiguration zuständig ist, muss lediglich die Aufgabe erfüllen, Invarianten wiederherzustellen. Die Berechnung korrekter Variablenbelegungen kann als „Constraint Satisfaction Problem“ beschrieben werden. Wie dies konkret gelöst wird, bleibt dem Rekonfigurationsmechanismus überlassen. Beispiele für Rekonfigurationsmechanismen (s. Kapitel 5) sind Constraint Solver oder genetische Algorithmen. Ebenso gibt die Theorie des RIA nicht vor, dass diese Invariante systemweit ausgewertet werden muss. In den meisten Fällen können die Bedingungen lokal überwacht und ausgewertet werden. Dies erlaubt lokale, also dezentrale Rekonfiguration.

Wie bereits erwähnt ist es nicht möglich, diesen Verhaltenskorridor für jegliche selbstorganisierende Systeme einmalig zu spezifizieren. Vielmehr ist es die Aufgabe Systemklassen zu finden, die modelliert werden und auf deren Modellen dieser Korridor definiert werden kann. In der Dissertation wurde die Systemklasse der Ressourcenflusssysteme eingehend untersucht und modelliert. Im Folgenden wird eine Beispielanwendung (s. Kapitel 3) aus dieser Systemklasse vorgestellt und anschließend das Organic Design Pattern (s. Kapitel 4), welches alle notwendigen Komponenten und deren Verhalten beschreibt. Auf diesen Komponenten wurden alle Constraints definiert, die notwendig sind, um den Verhaltenskorridor für Ressourcenflusssysteme im Generellen zu spezifizieren.

3 Fallstudie: adaptive Produktionszelle

Inspiriert von den Problemstellungen die herkömmliche Produktionsstraßen aufweisen (starr, unflexibel, hoch spezialisiert [HNS+10]) und den Bestrebungen, diese flexibler zu gestalten [BS07], ist die Anwendung der adaptiven Produktionszelle entstanden. Selbst-Organisation ist ein Mittel um diese Systeme einerseits dezentral und ausreichend flexibel, andererseits beweisbar zuverlässig zu konstruieren. Die adaptive Produktionszelle beinhaltet Roboter und mobile Plattformen (Carts), die sich an wechselnde Bedingungen anpassen können. Das heißt, dass Roboter verschiedene Fähigkeiten haben (Werkzeugwechsel) und Carts so mobil sind, dass sie Ressourcen zwischen beliebigen Robotern transportie-

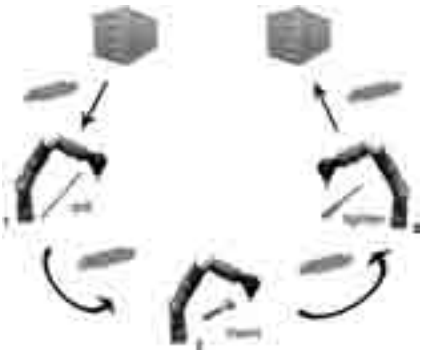


Abbildung 2: Adaptive Produktionszelle

ren können. Eine initiale Konfiguration, wie sie in Abbildung 2 gezeigt wird, ist eine Möglichkeit so eine Produktionszelle aufzubauen. Jeder Roboter verfügt über drei Fähigkeiten (drill, insert, tighten). Die Idee der adaptiven Produktionszelle ist es, die eingeführten Freiheitsgrade zu nutzen, um auf Ausfälle und neue Aufgaben, die dem System gestellt werden, reagieren zu können. Gerade für Kleinserien ist die adaptive Produktionszelle gut geeignet. Durch die Flexibilität in den Transportwegen, können die Roboter in beinahe beliebiger Reihenfolge angesteuert werden und somit kann ohne große Umbauarbeiten eine neue Aufgabe erfüllt werden. Des Weiteren stellte ein fixes Förderband bisher einen *Single Point of Failure* dar, d.h. wenn das Band stehen blieb, musste die gesamte Produktionsstraße angehalten werden.

4 Definition selbst-organisierender Ressourcenflusssysteme

Das Organic Design Pattern (ODP) [SOR07] ist ein Werkzeug, das die Konstruktion von selbst-organisierenden Systemen erleichtert und formale Aussagen über Systemeigenschaften ermöglicht. Das Herzstück des Patterns ist ein erprobtes Rollenkonzept, das es erlaubt, Systeme zu modellieren, die sich selbständig an neue Aufgaben anpassen und verschiedene Ressourcen mit verschiedenen Aufgaben gleichzeitig bearbeiten. Die im ODP vordefinierten Constraints spezifizieren den angesprochenen Verhaltenskorridor und auch dessen Überwachung zur Laufzeit. Des Weiteren gibt das Pattern das Verhalten der funktionalen Teile des Systems bereits vor. In [NSS+10] wird gezeigt, wie dieses Verhalten verifiziert werden kann. Unter der Annahme, dass der Rekonfigurationsmechanismus ausschließlich die Konfiguration beeinflusst und nicht in das Verhalten der funktionalen Komponenten eingreift, ist somit verifiziert, dass das selbst-organisierende System funktional korrekt arbeitet. Durch einen verifizierten Result Checker [FNSR11] wird sichergestellt, dass die Konfigurationen, die von dem Rekonfigurationsmechanismus berechnet wurden, korrekt sind bzgl. der Einhaltung aller Constraints.

Die statischen Elemente eines ODP-Systems und deren Zusammenhänge sind in Abbildung 3 als UML Klassendiagramm aufgezeigt. Das wichtigste Konzept ist der *Agent*, der entsprechend einer Aufgabe (*Task*) die *Resource* mit einer oder mehreren Fähigkeiten (*Capabilities*) bearbeitet. Der Agent kann zu jedem Zeitpunkt maximal eine Ressource bearbeiten. Der Task beschreibt eine Abfolge von Capabilities (ordered, nonunique), die auf eine Ressource angewendet werden sollen. Er ändert sich während der Verarbeitung einer Ressource nicht. Der Zustand (*state*) der Ressource ist immer ein Präfix des Tasks. Agenten, die eine Ressource produzieren, stellen den Startpunkt des Ressourcenflusses dar, Agenten, die eine Ressource konsumieren, entsprechend den Endpunkt. Jeder Agent zeichnet sich durch die Fähigkeiten, die er besitzt (*availableCapabilities*) und seine Interaktionsmöglichkeiten aus. Letztere geben an, welchen anderen Agenten der jeweilige Agent Ressourcen geben (*outputs*) kann und von welchen anderen Agenten er Ressourcen annehmen (*inputs*) kann. Die Agenten verfügen nur über ihr lokales Wissen, d.h. sie kennen im Allgemeinen nur die Agenten, die ihnen über genau diese „Input-/Output-Relationen“ bekannt sind. Die Kommunikation in ODP-Systemen ist nicht eingeschränkt, d.h. sobald ein Agent einen anderen Agenten „kennt“, kann er mit diesem kommunizieren.

Ein Agent kann mehrere *Rollen* haben. Die Zuweisung von Rollen zu Agenten wird Rollenallokation genannt (*allocatedRoles*). Selbst-Organisation wird in dieser Systemklasse als ein Rollenallokationsproblem angesehen: Welcher Agent muss zu welchem Zeitpunkt, welche Rolle ausführen, damit das Gesamtsystem korrekt arbeitet? Die momentan vom Agenten gewählte Rolle bestimmt, welche Fähigkeiten der Agent auf die vorliegende Ressource anwendet. Eine Rolle setzt sich aus drei Teilen zusammen, einer Vorbedingung (*precondition*), einer Sequenz von Fähigkeiten und einer Nachbedingung (*postcondition*). Die Vorbedingung gibt an, von welchem Agenten (*port*) der Agent die Ressource bekommt, welchen Task die Ressource hat und in welchem Bearbeitungszustand (*state*) sie momentan ist. Die Sequenz von Fähigkeiten (*capabilitiesToApply*), die in der Rolle angegeben sind, bestimmen, was der Agent konkret mit der Ressource zu tun hat. Das kann von „nur durchreichen“ (leere Sequenz) bis „alles erledigen“ (alle Capabilities des Tasks) reichen. Die Nachbedingung gibt wiederum an, mit welchem Zustand die Ressource den Agent verlässt, welche Aufgabe die Ressource hat und zu welchem Agent (*port*) die Ressource weitergegeben werden soll. In dem Beispiel der adaptiven Produktionszelle (s. Abbildung 2) nimmt Roboter 1 die Ressource von Cart 1 in dem Zustand [], mit dem Task [$drill$, $insert$, $tighten$], führt $drill$ aus und gibt die Ressource weiter an Cart 2 mit dem Zustand [$drill$].

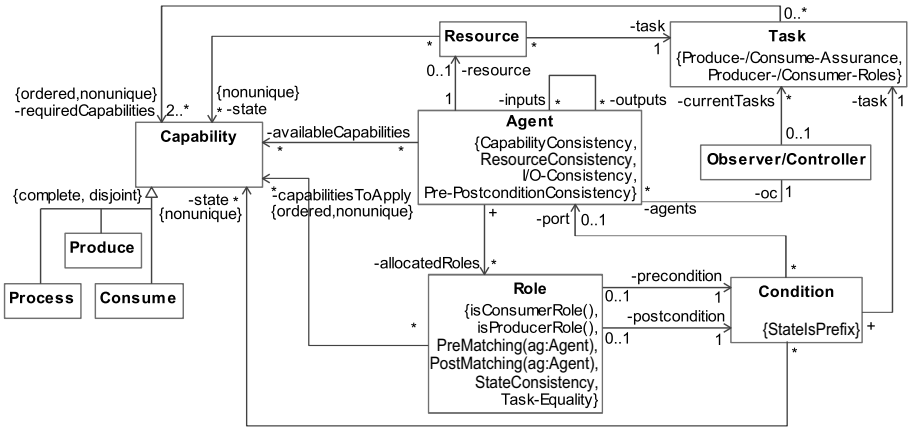


Abbildung 3: Organic Design Pattern - Konstruktionsmodell

Wie in Kapitel 2 beschrieben, werden eine Reihe von Constraints benötigt, die auf den Komponenten des Zielsystems definiert werden können, um Verhaltensgarantien abzugeben. Die hier vorgestellte statische Struktur bildet nun die Grundlage, um mit OCL-Constraints [OMG06] einen Verhaltenskorridor zu spezifizieren. Die Constraints sind in abgekürzter Schreibweise in den Komponenten des ODP zu finden. An dieser Stelle wird exemplarisch ein Constraint angegeben. Der *CapabilityConsistency*-Constraint (vgl. Listing 1) eines Agenten besagt, dass ein Agent nur Rollen zugewiesen bekommen darf, in denen er Capabilities anwenden muss, die er in seiner *availableCapabilities*-Relation hat.

Diese Constraints werden auf der einen Seite benötigt, um eine korrekte Rollenallokation zu berechnen und auf der anderen Seite, um durch Beobachtung eine Verletzung der

Invariante festzustellen. Im gegebenen Beispiel können die Agenten (Roboter und Carts) direkt feststellen, ob ihre Rollenallokation noch korrekt ist und ggf. eine Verletzung des Constraints und somit der Invariante melden bzw. lokal verarbeiten. Fällt nun bei einem Roboter der Bohrer *drill* aus muss er mit einem anderen Roboter die Rolle tauschen. Die Carts rekonfigurieren ihre Transportrollen entsprechend und stellen somit wieder einen korrekten Ressourcenfluss her. Für diese Rollenrekonfiguration muss genügend Redundanz im System vorhanden sein, sowohl in den Capabilities als auch in den Input-Output Relationen.

```
1 context Agent inv: availableCapabilities ->
2   includesAll(allocatedRoles.capabilitiesToApply)
```

Listing 1: CapabilityConsistency

Der konkrete Rekonfigurationsmechanismus, der das System nach Verlassen des Korridors wieder in diesen zurück bringt, ist gekapselt in der Observer/Controller-Komponente (O/C). Das Ergebnis des Rekonfigurationsalgorithmus ist eine neue Allokation von Rollen zu Agenten, die das Gesamtsystem in einen Zustand bringt, in dem es konform zu den Constraints arbeitet und wieder in der Lage ist, das Systemziel korrekt zu erfüllen. Die Constraints und statischen Komponenten des ODP bestimmen die Ergebnisform des Rekonfigurationsalgorithmus. Die Klassen, Relationen und Kardinalitäten gekoppelt mit den Constraints spezifizieren korrekte Resultate und somit welche Rollenallokationen für Systeme, die mit dem ODP modelliert wurden, gültig sind. Ein Zustand eines ODP-Systems ist eine konkrete Instanz des ODP. Die Zustände σ_0-σ_4 aus Abbildung 1 entsprechen jeweils einer korrekt konfigurierten Instanz des ODP. Der Schritt zwischen diesen Zuständen entspricht einer Änderung der ausgehenden Instanz, z.B. durch neue Objekte oder veränderte Assoziationen. $\sigma_{err} \in \mathcal{S}_{reconf}$ entspricht einer Instanz, die im Moment nicht korrekt konfiguriert ist, da z.B. in dem Schritt, der zu diesem Zustand führte, ein Agent die Fähigkeit, die er für seine Rolle benötigt, verloren hat. Die statischen und dynamischen Konzepte des ODP sind in der Referenzimplementierung *ODP Runtime Environment (ORE)* komplett umgesetzt. Die Implementierung setzt auf Jadex, einem Multi-Agenten-System auf und unterstützt eine einfache Konstruktion selbst-organisierender Ressourcenflusssysteme.

5 Rekonfiguration selbst-organisierender Ressourcenflusssysteme

Detektiert nun ein Agent lokal eine Constraintverletzung, muss ein Teil des Systems rekonfiguriert werden. Um diese Rekonfiguration durchzuführen bietet die vorliegende Arbeit prinzipiell zwei verschiedene Möglichkeiten an. Zum einen kann das gesamte System durch eine zentrale Instanz rekonfiguriert werden, die das Wissen über das Gesamtsystem hält. Die zweite Möglichkeit ist eine dezentrale Rekonfiguration die nur einen kleinen Teil des Systems mit neuen Rollen ausstattet. Eine zentrale Instanz zieht immer einige Nachteile mit sich, die hier nicht im Detail erläutert werden sollen. Vielmehr wird im Folgenden ein kurzer Einblick in die dezentrale Rekonfiguration gegeben, die minimale Eingriffe in das System ermöglicht und Entscheidungen auf lokalem Wissen fällt. Somit stellt nach [DMSGK05] die dezentrale Rekonfiguration einen stark selbst-organisierenden Mechanismus dar.

Die Grundidee der dezentralen Rekonfiguration [ASN+11] ist die Bildung von Agentengruppen, *Koalitionen* genannt, die in der Lage sind, verletzte Constraints wieder zu erfüllen. Fehler, die durch den Ausfall von Capabilities, Input- oder Output-Ports sowie vollständiger Agenten entstehen, werden kompensiert, indem für einen Teil des Systems neue Rollen berechnet werden. Die anderen Bereiche des Systems, die an der Rekonfiguration nicht beteiligt sind, können für den betroffenen Task weiterhin die alloziierten Rollen ausführen. Dabei existieren die Koalitionen ausschließlich in Phasen der Rekonfiguration. Sind diese abgeschlossen, so lösen sich die Koalitionen wieder auf. Dieser Mechanismus profitiert von der festgelegten Systemstruktur durch das ODP und kann somit anhand lokaler Regeln über die Gültigkeit der Invariante Entscheidungen treffen.

Diese dezentrale Kontrolle bietet die Möglichkeit nur sehr selektiv in den laufenden Betrieb der Systeme einzugreifen. Es werden nur die Agenten angehalten, bei denen eine Constraintverletzung vorliegt, sowie angrenzende Agenten, die bei der Wiederherstellung einer korrekten Rollenallokation helfen können. Das Ziel ist es, effizient und effektiv den Ressourcenfluss wiederherzustellen. In Systemen, die mit dem ODP konstruiert wurden und die den Koalitionsbildungsmechanismus zur Rekonfiguration nutzen, ist jeder einzelne Agent in der Lage, eine Koalition zu bilden, um das System wieder in einen Zustand innerhalb des Korridors zurückzuführen. Dieser Mechanismus greift somit nicht in die Eigenständigkeit der Agenten ein, was im Sinne der Selbst-Organisation natürlich wünschenswert ist. Wenn die Entscheidung für die Rekonfiguration auf einen dezentralen Mechanismus fällt, muss jedoch in Kauf genommen werden, dass dieser Mechanismus nicht immer die global optimale Lösung finden kann, speziell wenn er auf rein lokalem Wissen arbeitet. Ein Vorteil des in dieser Dissertation präsentierten Gesamtansatzes ist es, dass im ORE verschiedene Rekonfigurationsmechanismen für die Realisierung starker (dezentraler) sowie schwacher (zentraler) selbst-organisierender Systeme bereitgestellt werden. Der Entwickler eines selbst-organisierenden Ressourcenflusssystems ist damit in der Lage, die für seine Anwendung optimale Rekonfigurationsstrategie mittels eines Plugin-Mechanismus einzusetzen.

6 Innovationen und Ausblick

Motiviert von der zunehmenden Komplexität zukünftiger Softwaresysteme ist es in dieser Dissertation gelungen, einen Ansatz zu entwickeln, der das Design und die Konstruktion selbst-organisierender Systeme erleichtert, mit denen die steigende Komplexität handhabbar wird. Wesentliche Ergebnisse dieser Arbeit sind der *Restore Invariant Approach*, der die Definition eines Korridor gewünschten Verhaltens fordert und die getrennte Betrachtung des funktionalen Systems und den Mechanismen der Selbst-Organisation ermöglicht; das *Organic Design Pattern*, welches die Klasse der selbst-organisierenden Ressourcenflusssysteme beschreibt; verschiedenste Rekonfigurationsmechanismen, sowohl zentraler als auch dezentraler Natur; das *ODP Runtime Environment*, das alle entwickelten Mechanismen in einem Framework vereint und die *Software Engineering Guideline* [SNSR10], die letztendlich eine Anleitung zur einfachen, reproduzierbaren Entwicklung selbst-organisierender Systeme darstellt. Auf einige dieser Punkte konnte leider im Rahmen die-

ses Beitrags nicht eingegangen werden, ausführliche Beschreibungen finden sie jedoch in [See11].

Aus den Ergebnissen der Dissertation ergeben sich einige neue Möglichkeiten und interessante Fragen. Die Spezifikation der korrekten Systemkonfiguration basiert bereits auf OCL, eine weiterführende spannende Frage ist, was mit dieser Spezifikation noch auf Modellebene ausgedrückt werden kann. Interessant ist dabei die Analyse eines konkreten Systems (d.h. einer Instanz des ODP), bzgl. der Redundanzverteilung. Es könnte getestet werden, ob das System optimal konfiguriert ist, damit der Selbst-Organisationsmechanismus die Zuverlässigkeit und Robustheit des Systems wesentlich verbessern kann. Untersucht werden kann diese Instanz unter anderem auch auf das Vorkommen von *Single Point of Failures*. Solche Schwachstellen bereits auf Modellebene zu finden ist eine erstrebenswerte Erweiterung des bisherigen Ansatzes. Eine weitere Herausforderung ist es, die Erfahrung mit dem Design selbst-organisierender Systeme zu nutzen, um das ODP auch für andere Systemklassen nutzbar zu machen. Interessant sind dabei die folgenden Fragestellungen: welche Konzepte des ODP werden für welche Systemklasse benötigt; werden manche Teile eventuell nicht benötigt; müssen neue Konzepte und Constraints definiert werden; wie gliedern sich diese Konzepte ein; welche Abhängigkeiten bestehen unter den einzelnen Konzepten und den Constraints, die über diese Konzepte sprechen. Die Vision ist, dass nach einer ausgiebigen Analyse weiterer Systemklassen eine Art Baukasten für selbst-organisierende Systeme entsteht.

Literatur

[ASN+11] G. Anders, H. Seebach, F. Nafz, J.-P. Steghöfer und W. Reif. Decentralized Reconfiguration for Self-Organizing Resource-Flow Systems Based on Local Knowledge. In *Proceedings of EASe 2011*, Las Vegas, Nevada, USA, April 2011.

[BS07] B. Bickel und M. Schuster. *Trends, Methoden und Grundsätze moderner Fabrik- und Produktionsplanung*. GRIN Verlag, 2007.

[DMSGK05] G. Di Marzo Serugendo, M.-P. Gleizes und A. Karageorgos. Self-organization in multi-agent systems. *The Knowledge Engineering Review*, 20(02):165–189, 2005.

[DMSGK06] G. Di Marzo Serugendo, M.P. Gleizes und A. Karageorgos. Self-Organisation and Emergence in MAS: An Overview. *Informatica*, 30(1):45–54, 2006.

[FNSR11] P. Fischer, F. Nafz, H. Seebach und W. Reif. Ensuring Correct Self-Reconfiguration in Safety-Critical Applications by Verified Result Checking. In *Workshop Organic Computing as part of ICAC 2011*, Karlsruhe, Germany, June 2011.

[For10] Deutsche Forschungsgemeinschaft. http://www.organic-computing.de/SPP, 2010. DFG SPP 1183 - Organic Computing Initiative.

[HNS+10] A. Hoffmann, F. Nafz, H. Seebach, A. Schierl und W. Reif. Developing Self-Organizing Robotic Cells using Organic Computing Principles. In *Workshop on Bio-Inspired Self-Organizing Robotic Systems, 2010 IEEE International Conference on Robotics and Automation (ICRA 2010)*. Springer, 2010.

[KC03] J.O. Kephart und D.M. Chess. The Vision of Autonomic Computing. *Computer*, 36(1):41–50, 2003.

[MS04] C. Müller-Schloer. Organic Computing - On the Feasibility of Controlled Emergence. In *Proceedings of the 2nd IEEE/ACM/IFIP international conference on Hardware/-software codesign and system synthesis*, Seiten 2–5. ACM, 2004.

[NSS+10] F. Nafz, H. Seebach, J.-P. Steghöfer, S. Bäumler und W. Reif. A Formal Framework for Compositional Verification of Organic Computing Systems. In *Proceedings of the 7th International Conference on Autonomic and Trusted Computing*. Springer, 2010.

[OMG06] OMG. Object Constraint Language, OMG Available Specification, 2006. Version 2.0.

[PMGT08] H. Parzyjegla, Jaeger M., Mühl G. und Weis T. Model-driven Development and Adaptation of Autonomous Control Applications. *IEEE Distributed Systems Online (DSOnline)*, 9(11), November 2008.

[RMB+06] U. Richter, M. Mnif, J. Branke, C. Müller-Schloer und H. Schmeck. Towards a generic observer/controller architecture for Organic Computing. *INFORMATIK 2006 – Informatik für Menschen!*, P-93:112 – 119, 2006.

[Sch05] H. Schmeck. Organic Computing - A New Vision for Distributed Embedded Systems. In *Proceedings of the Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing*, ISORC '05, Seiten 201–203, Washington, DC, USA, 2005. IEEE Computer Society.

[See11] H. Seebach. *Konstruktion selbst-organisierender Softwaresysteme*. Dissertation, University of Augsburg, Augsburg, Germany, 2011. (in German).

[SNSR10] H. Seebach, F. Nafz, J.-P. Steghöfer und W. Reif. A Software Engineering Guideline for Self-organizing Resource-Flow Systems. In *Proceedings of the 4th IEEE International Conference on Self-Adaptive and Self-Organizing Systems*, 2010.

[SOR07] H. Seebach, F. Ortmeier und W. Reif. Design and Construction of Organic Computing Systems. In *Proceedings of the IEEE Congress on Evolutionary Computation*, 2007.

Hella Seebach wurde 1978 in Oldenburg geboren und hat Informatik an der Universität Augsburg studiert. Sie promovierte am Lehrstuhl „Softwaretechnik und Programmiersprachen“ bei Professor Reif. Zur Zeit koordiniert Frau Seebach dort den Bereich Organic Computing, der das DFG geförderte Projekte SAVE ORCA und die DFG-Forschergruppe OC-Trust umfasst. In beiden Projekten stehen selbst-organisierende Systeme im Mittelpunkt. SAVE ORCA befasst sich mit den generellen Konstruktionsparadigmen und formaler Korrektheit der Systeme, während in OC-Trust der Fokus auf der Entwicklung von Trustmechanismen für selbst-organisierende Systeme liegt.

Adaptive Verfahren zur nutzerzentrierten Organisation von Musiksammlungen

Sebastian Stober
Data & Knowledge Engineering Gruppe
Fakultät für Informatik, Otto-von-Guericke-Universität, Magdeburg
Sebastian.Stober@ovgu.de

Abstract: Music Information Retrieval (MIR) Systeme müssen fazettenreiche Informationen verarbeiten und gleichzeitig mit heterogenen Nutzern umgehen können. Insbesondere wenn es darum geht, eine Musiksammlung zu organisieren, stellen die verschiedenen Sichtweisen der Nutzer, verursacht durch deren unterschiedliche Kompetenz, musikalischen Hintergrund und Geschmack, eine große Herausforderung dar. Diese Herausforderung wird hier adressiert, indem adaptive Verfahren für verschiedene Elemente von MIR Systemen vorgeschlagen werden: Datenadaptive Techniken zur Merkmalsextraktion werden beschrieben, welche zum Ziel haben, die Qualität und Robustheit der aus Audioaufnahmen extrahierten Informationen zu verbessern. Das klassische Problem der Genreklassifikation wird aus einer neuen nutzerzentrierten Sichtweise behandelt – anknüpfend an die Idee idiosynkratischer Genres, welche die persönlichen Hörgewohnheiten eines Nutzer besser widerspiegeln. Eine adaptive Visualisierungstechnik zur Exploration und Organisation von Musiksammlungen wird entwickelt, die insbesondere Darstellungsfehler adressiert, welche ein weit verbreitetes und unumgängliche Problem von Techniken zur Dimensionsreduktion sind. Darüber hinaus wird umrissen, wie diese Technik eingesetzt werden kann, um die Interessantheit von Musikempfehlungen zu verbessern, und neue blickbasierte Interaktionstechniken ermöglicht. Schließlich wird ein allgemeiner Ansatz für adaptive Musikähnlichkeit vorgestellt, welcher als Kern für eine Vielzahl adaptiver MIR Anwendungen dient. Die Einsatzmöglichkeiten der beschriebenen Verfahren werden an verschiedenen Anwendungsprototypen gezeigt.

Dank immer fortgeschrittenerer Analysemethoden aus dem Music Information Retrieval (MIR) werden zukünftige Generationen von Programmen zur Verwaltung von Musiksammlungen immer besser den Inhalt der Musikstücke verstehen können. Dadurch wird es möglich, Musikstücke inhaltlich miteinander zu vergleichen und beispielsweise ähnliche Stücke zu gruppieren, anstatt einfach nur nach Genre, Künstler und Album zu sortieren. Dabei ergeben sich jedoch auch neue Herausforderungen und Problemstellungen. Einerseits müssen die MIR Systeme auf der Datenseite mit einer Vielzahl verschiedener Facetten von Musik umgehen (wie z.B. Melodie, Harmonie, Rhythmus, Dynamik, Instrumentierung, Text) und mit Merkmalen, welche die Musikstücke auf ganz unterschiedlichen Ebenen beschreiben und sich dabei auf verschiedene Facetten beziehen. Gleichzeitig muss aber auch mit einer starken Varianz unter den Benutzern eines MIR Systems gerechnet werden mit Unterschieden im Hörverhalten, Musikgeschmack und nicht zuletzt im musikalischen Hintergrund. Dies führt beispielsweise dazu, dass im Allgemeinen nicht alle Nutzer eines Systems Musikstücke auf die gleiche (objektive) Art und Weise vergleichen. So mag ein Musiker dazu neigen, stärker auf Strukturen, Instrumentierung oder

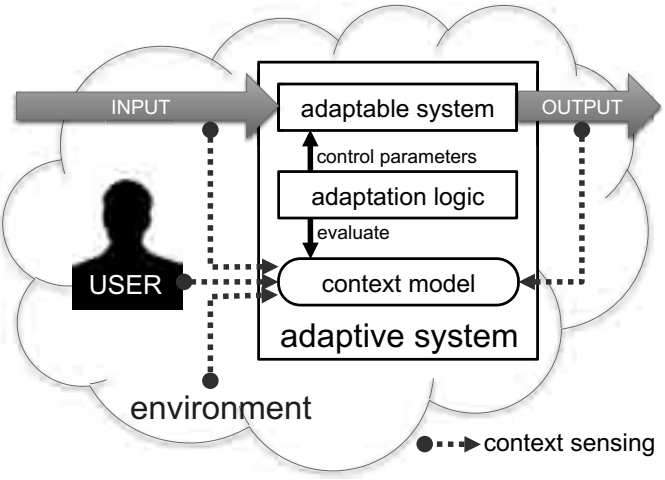


Abbildung 1: Allgemeines Modell eines adaptiven Systems.

Harmonien zu achten und dabei vielleicht seinem Instrument eine besondere Gewichtung geben. Nicht-Musiker werden sich beim Vergleichen möglicherweise eher auf die Klangfarbe oder allgemeine Stimmung eines Stückes stützen. Andere wiederum achten, sofern sie der jeweiligen Sprache mächtig sind, auf den Liedtext. Um mit der großen Diversität in den Musikinformationen und bei den Benutzern von MIR Systemen umzugehen, bietet sich als eine Lösungsmöglichkeit die Verwendung adaptiver Methoden an.

Als Grundlage für die Entwicklung adaptiver MIR-Methoden wurden zunächst die folgenden Definitionen sowie ein allgemeines Modell für adaptive Systeme erarbeitet. Ein System wird als *adaptierbar* bezeichnet, wenn sein Verhalten mittels von außen zugänglicher Parameter manuell angepasst werden kann. Ein *adaptives* System nimmt hingegen diese Anpassung selbständig vor, wobei die Änderung im Verhalten vom Kontext abhängt und zielorientiert sein muss, so dass das Systemverhalten bezüglich eines gegebenen Maßes optimiert wird. Der Kontext ist dabei im weitest möglichen Sinn zu verstehen und umfasst die allgemeine Arbeitsumgebung, den Benutzer und die Ein- und Ausgabedaten (z.B. deren Historie oder statistische Charakteristika). Für die automatische Anpassung muss das System den Kontext (zumindest teilweise) wahrnehmen, ein Kontextmodell als interne Repräsentation konstruieren und schließlich daraus mittels einer Adaptionslogik geeignete Systemänderungen ableiten. Der Kern des Systems kann dabei wiederum als einfaches adaptierbares System betrachtet werden wie in Abbildung 1 dargestellt. Diese Sichtweise ermöglichte eine systematische Analyse und den Vergleich verschiedenster MIR Verfahren mit Blick auf deren Adaptivität [SN12], welche in Umfang und Systematik bisher einmalig ist. Dieser Überblick weißt auch auf besonders vielversprechende Anwendungsfelder für adaptive Techniken im MIR hin, welche in der Arbeit behandelt wurden und in den folgenden Abschnitten zusammengefasst sind.

1 Adaptive Merkmalsextraktion aus Musikaufnahmen

Bei der Analyse von Musikaufnahmen, dem ersten Schritt des generellen Retrievalprozess, kann mit Hilfe adaptiver Methoden eine höhere Qualität und Robustheit der extrahierten Merkmale erreicht werden. Hier wurden zwei Ansätze entwickelt, die im Rahmen von Diplomarbeiten betreuter Studenten umgesetzt wurden. Der erste Ansatz beschäftigt sich mit dem Problem, wie die Melodie aus einer Stereo-Musikaufnahme extrahiert werden kann [DNS07]. Die Idee dabei ist, zunächst mittels eines weit verbreiteten Karaoke-Filters ein Signal für die Hintergrundmusik zu extrahieren. Dieses Signal wird dann zum Einstellen eines Störsignalfilters verwendet, der versucht, möglichst viel der Hintergrundmusik aus der Aufnahme zu entfernen, während die Melodie erhalten bleiben sollte. Für einen solchen Anwendungsfall ist der Störsignalfilter jedoch an sich nicht vorgesehen, da ein weitestgehend konstantes Störsignal vorausgesetzt wird, während die Hintergrundmusik hingegen oft starke zeitliche Veränderlichkeit aufweist. Abhilfe schafft hier, den Störfilter als adaptierbaren Kern in einem adaptiven System einzubetten, welches die Filterparameter dynamisch bezüglich eines kleinen Zeitfensters wählt. Somit wird zu jedem Zeitpunkt nur ein kleiner zeitlich begrenzter Ausschnitt der Hintergrundmusik zur Anpassung des Filters verwendet, was der Annahme eines konstanten Störsignals deutlich näher kommt und schließlich zu einem besseren Filterergebnis führt.

Der zweite Ansatz beschäftigt sich mit der Fehlerkorrektur bei der Erkennung von Akkorden in Musikaufnahmen [RSN08]. Die harmonische Akkordfolge ist ein wichtiges Merkmal zur Indexierung und Analyse westlicher Musik. Daher beschäftigt sich eine Vielzahl von Arbeiten im MIR mit dieser Problematik. Trotz großer Fortschritte sind Klassifikationsfehler jedoch keine Seltenheit. Dies liegt unter anderem an der Verwendung von Musikinstrumenten und Effekten, die nicht nur harmonische Signalanteile erzeugen. Basierend auf einer umfangreichen Studie existierender Ansätze zur Akkorderkennung, wurde hier eine Unterteilung in drei Phasen vorgenommen: Merkmalsextraktion, Akkordklassifikation und Nachverarbeitung. Anschließend wurde ein adaptives Verfahren für die Nachverarbeitung vorgeschlagen, welches unabhängig von den ersten zwei Arbeitsschritten angewendet werden kann. Dabei werden (ungenaue) Informationen über die Akkorde in der (zeitlichen) Nachbarschaft als Kontext verwendet, um mögliche Fehlklassifikationen zu korrigieren. Dazu bildet ein probabilistischer Klassifikator den adaptierbaren Kern des adaptiven Systems. Als Systemparameter werden die Marginalwahrscheinlichkeiten der einzelnen Akkorde dynamisch bezüglich der Häufigkeit in der Nachbarschaft angepasst und so der Klassifikationsprozess beeinflusst. Dies führte zu einer deutlichen Verbesserung der Klassifikationsgenauigkeit in Experimenten mit drei verschiedenen Basis-Klassifikationsverfahren.

2 Genreklassifikation mit nutzerspezifischen Genrekategorien

Die Einordnung von Musikstücken in Genres ist ein häufig verwendeter Ansatz zur Sortierung von Musiksammlungen – insbesondere für große Kataloge im Handel. Eine solche Sortierung ist jedoch oftmals nur begrenzt hilfreich, da einerseits generische Genres wie

"Rock" oder "Pop" zu undifferenziert sind und es andererseits aber auch schwierig ist, für sehr spezifische Genres wie "Scottish Lo-Fi Post-Rock" einen Konsens unter verschiedenen Nutzern zu finden. Es wurde daher ein alternativer Ansatz untersucht: Anstatt Musik in mehr oder weniger künstliche Schubladen zu pressen, könnte ein nutzeradaptives MIR System Genrekategorien lernen, die auf den individuellen Nutzer zugeschnitten und somit sinnvoll und intuitiv verständlich sind. Verschiedene frühe Studien aus dem MIR deuten zumindest darauf hin, dass sich sinnvolle nutzerbezogene Genrekategorien aus dem individuellen Nutzungsverhalten ableiten lassen wie z.B. "Musik zum Autofahren". In einer Vorstudie im Rahmen der Dissertation wurde daher zunächst ein Prototyp zur Aufzeichnung von einfachen Wetterinformationen als Hörkontext entworfen und mit einer kleinen Nutzergruppe getestet. Außerdem wurde eine Reihe von weitergehenden Möglichkeiten zur automatischen Aufzeichnung von Kontextinformationen vorgeschlagen, die mit einfachen technischen Mitteln realisierbar wären. Da jedoch einige dieser Möglichkeiten stark in die Privatsphäre der Nutzer eingreifen, wurde zunächst eine umfangreiche Studie zur Akzeptanz der Aufzeichnungstechniken durchgeführt [SSN09]. Im Rahmen der CeBIT 2009 und mit Hilfe eines Online-Fragebogens wurden insgesamt 461 Personen befragt. Die Umfrageergebnisse zeigen deutlich, dass die potentiellen Nutzer der Aufzeichnung von Hörkontextinformation sehr kritisch gegenüber stehen und ihre Privatsphäre Vorrang hat. Als generelle Richtlinie für die zukünftige Entwicklung personalisierter MIR Anwendungen kann aus den Antworten außerdem gefolgert werden, dass die Benutzer jederzeit die volle Kontrolle haben müssen – sowohl über die aufgezeichneten Kontextinformationen als auch darüber, ob und inwieweit diese zur Adaption verwendet werden.

3 Fokusadaptive Visualisierung von Musiksammlungen

Die Visualisierung von Musiksammlungen stellt den ersten der beiden Hauptschwerpunkte dieser Arbeit dar. Viele Ansätze zur Visualisierung einer Musiksammlung basieren auf Techniken, bei denen Objekte (Musikstücke, Alben oder Künstler) aus einem hochdimensionalen Merkmalsraum für die Darstellung in den zwei- oder dreidimensionalen Raum abgebildet werden. Ziel dabei ist es, die Objekte in einer Art Karte der Sammlung so anzuordnen, dass benachbarte Objekte einander sehr ähnlich sind und die Ähnlichkeit mit wachsendem Abstand in der Karte abnimmt. Dabei kommt es durch die Dimensionsreduktion zwangsläufig zu Verzerrungen der Abstände. Als Folge kann es vorkommen, dass benachbarte Objekte sich gar nicht so sehr ähneln, wie es die Darstellung in der Karte vermuten lässt, oder weit von einander entfernte Objekte sehr ähnlich sind. Im letzteren Fall sind beide Regionen durch eine Art "Wurmloch" (über den hochdimensionalen Merkmalsraum) miteinander verbunden und gehören ursprünglich im Merkmalsraum zur gleichen Nachbarschaft. Im Rahmen der Arbeit wurde daher die fokusadaptive SpringLens entwickelt, eine interaktive Visualisierungstechnik, die eine globale Sicht auf eine Musiksammlung ermöglicht und mit adaptiven Filterfunktionen und multifokalem Zoom die beschriebenen Verzerrungsprobleme gezielt adressiert. (Der Name der Technik leitet sich aus dem ihr zugrundeliegenden Verfahren zur nichtlinearen Verzerrung von Bildern [GGSS06] ab.) Dabei wird speziell auf das Phänomen der Wurmlöcher eingegangen, welches bei der

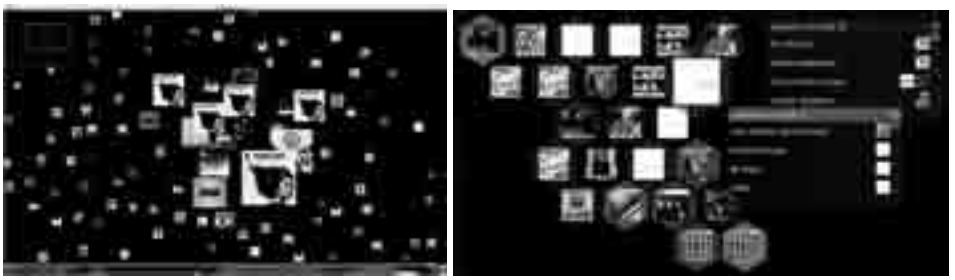


Abbildung 2: Links: MusicGalaxy Prototyp zur Exploration von Musiksammlungen mit Hilfe der fokusadaptiven SpringLens Visualisierung. Rechts: BeatlesExplorer Prototyp zur interaktiven Strukturierung der Musik der Beatles mit einer wachsenden selbstorganisierenden Karte. Demo-Videos zu den Prototypen sind unter `http://www.dke-research.de/aucoma/thesis` verfügbar.

Exploration besonders problematisch ist, da Nutzer im Falle verzerrter Nachbarschaften schnell relevante Objekte übersehen können. Zur Visualisierung der Wurmlöcher in der Kartendarstellung wird eine adaptive multifokale Fischaugenlinse verwendet. Diese besteht aus einem Primärfokus, welcher vom Nutzer gesteuert wird, und einem datengetriebenen Sekundärfokus. Beim Primärfokus handelt es sich um eine gewöhnliche Fischaugenlinse. Mit dieser kann der Nutzer in eine Region hineinzoomen, die ihn interessiert. Die Verzerrung der Linse führt dazu, dass mehr Platz für die Darstellung von Details in dieser Region geschaffen wird, indem die anderen (weniger interessanten) Regionen nach außen gedrückt und komprimiert dargestellt werden. So kann sich der Nutzer die Region von Interesse genauer anschauen, ohne deren Einordnung in den Kontext der gesamten Sammlung zu verlieren. Der Sekundärfokus umfasst mehrere solcher Fischaugenlinsen, die jedoch kleiner sind und nicht direkt vom Nutzer sondern vom System dynamisch in Abhängigkeit vom Primärfokus angepasst werden: Ändert sich der Primärfokus, wird im Hintergrund eine Nächste-Nachbar-Suche im ursprünglichen Merkmalsraum initiiert. Werden nächste Nachbarn zu den Objekten im Primärfokus zurückgeliefert, die sich nicht in direkter Nachbarschaft befinden, werden an den entsprechenden Stellen Sekundärlinsen eingefügt. In der resultierenden verzerrten Darstellung rücken dadurch die entfernten nächsten Nachbarn mit in den Fokus und näher an die Region von Interesse. Basierend auf dieser fokusadaptiven Visualisierungstechnik wurde der Prototyp "MusicGalaxy" [SN10] entwickelt, welcher in Abbildung 2 (links) gezeigt wird. Dabei werden die Musikstücke einer Sammlung als Sterne einer Galaxie visualisiert. Das zugrundeliegende Ähnlichkeitsmaß für die Abstandsberechnung im Merkmalsraum unterstützt eine flexible Anzahl von Facetten, deren Gewichtung adaptierbar ist. Dank spezieller Datenstrukturen können bei einer Änderung des Maßes die Positionen der Musikstücke in der Galaxiekarte in Echtzeit aktualisiert werden wodurch eine interaktive Exploration ermöglicht wird. Zentraler Aspekt ist hier jedoch zunächst die fokusadaptive Visualisierung während sich der folgende Abschnitt dann im Detail mit den Ähnlichkeitsmaßen beschäftigt. Die Entwicklung von MusicGalaxy erfolgte in einem nutzerzentrierten Designprozess. Auf der CeBIT 2010 wurden Meinungen von 112 Besuchern zum ersten interaktiven Prototyp gesammelt. Die nächste überarbeitete Version wurde mit drei Personen ausgiebig getestet

und im Anschluss weiter verbessert. Abschließend wurde eine vergleichende Studie mit Hilfe eines Eyetrackers durchgeführt, an der 30 Personen teilnahmen. Die Ergebnisse der Studie belegen die Nützlichkeit des Sekundärfokus zur Exploration. Im Vergleich mit der herkömmlichen "Pan & Zoom" Technik, die als Standard für kartenbasierte Nutzerschnittstellen betrachtet werden kann, wurden von der fokusadaptiven SpringLens durchweg bessere Werte für Nützlichkeit, Benutzbarkeit und Intuitivität erreicht. Weiterhin konnten mit Hilfe des Eyetrackers verschiedene Navigationsstrategien der Benutzer identifiziert werden. Wie ein weiterer Prototyp zur Exploration von Fotosammlungen beweist, beschränkt sich der Einsatzbereich der entwickelten Visualisierungstechnik nicht nur auf Musikdaten. Verschiedene weiterführende Nutzungsszenarien werden in Abschnitt 5 diskutiert.

4 Kontextadaptive Musikähnlichkeit

Musikähnlichkeit ist der Schlüssel für eine Vielzahl von MIR Anwendungen und stellt daher den zweiten Schwerpunkt dieser Arbeit dar. Ähnlichkeitsmaße werden beispielsweise benötigt, um die Ergebnisliste einer Suche zu sortieren, um bei der Organisation ähnliche Musikstücke zu gruppieren oder um ähnliche Stücke empfehlen zu können. Wie eingangs motiviert, hat Musik jedoch viele Facetten, die nicht zwangsläufig für alle Nutzer gleich wichtig sein müssen. Vielmehr kann die Wichtigkeit einzelner Facetten beispielsweise in Abhängigkeit vom musikalischen Hintergrund des Nutzers oder dessen Retrievalaufgabe stehen. Im Rahmen der Arbeit wurde daher zunächst ein Modell vorgeschlagen, welches subjektive Musikähnlichkeit durch ein parametrisierbares Abstandsmaß mit einer beliebigen Anzahl von Facetten umsetzt. Um den *subjektiven* Abstand zweier Musikstücke zu bestimmen, wird dabei pro Facette ein *objektiver* Abstand berechnet und mit einem subjektiven Faktor zwischen 0 und 1 gewichtet zum Gesamtabstand aufaddiert. Durch Wahl geeigneter Facettengewichte ist dieses Abstandsmaß adaptierbar nach den Vorstellungen des Nutzers. Dank des einfachen linearen Modells ist es zudem auch leicht verständlich. Jedoch kann nicht immer davon ausgegangen werden, dass der Nutzer sich der von ihm beim Vergleich zweier Musikstücke vorgenommenen Facettengewichtung auch bewusst ist. Daher wurde auch ein allgemeiner Ansatz entwickelt, der es einem MIR System erlaubt, die Gewichte aus der Interaktion mit dem Nutzer zu lernen [Sto11]. Der Lernprozess wird dabei als Optimierungsproblem (unter Bedingungen) oder alternativ als binäres Klassifikationsproblem modelliert. In beiden Fällen wird das Verfahren durch relative Abstandsbedingungen gesteuert, die als atomare Informationseinheiten das Kontextmodell bilden. Drei Beispielanwendungen veranschaulichen, wie solche relativen Abstandsbedingungen aus verschiedenen realen Interaktionsszenarien abgeleitet werden können. Bei der ersten Anwendung wurden in Zusammenarbeit mit Forschern des Meertens Instituts in Utrecht Abstandsmaße zur Klassifikation von Volksliedern gelernt. Dazu wurden Annotationen der Experten ausgewertet. Bei der zweiten Anwendung, dem "BeatlesExplorer" (Abbildung 2, rechts), handelt es sich um einen weiteren Prototyp, der im Rahmen dieser Dissertation entwickelt wurde. Er dient zur Exploration der Musik der Beatles und berücksichtigt mehr als 20 Facetten wie z.B. Text, Harmonie, Rhythmus, Instrumentierung und Produzenten. Mit Hilfe einer wachsenden selbstorganisierenden Karte (Growing Self-

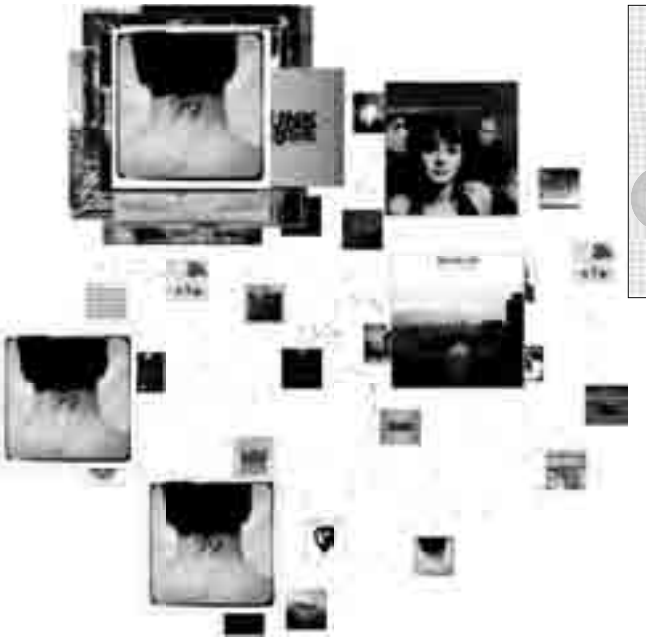
Organizing Map; GSOM), werden ähnliche Musikstücke in hexagonale Zellen zusammengefasst, wobei benachbarte Zellen wiederum ähnlich zueinander sind. Durch Drag&Drop kann der Nutzer interaktiv nach seinen Vorstellungen Musikstücke in andere Zellen verschieben. Daraus lassen sich relative Abstandsbedingungen ableiten und das Abstandsmaß anpassen. Die dritte Beispielanwendung ist der bereits vorgestellte Prototyp MusicGalaxy. Hier können Nutzer Objekte durch Taggen in Gruppen einordnen, woraus sich ebenfalls relative Abstandsbedingungen ableiten lassen. Zur Adaption der Facettengewichte bezüglich des Kontextmodells werden schließlich verschiedene Verfahren beschrieben mit leicht unterschiedlichen Zielfunktionen vorgeschlagen, die passend zum Anwendungsszenario ausgewählt werden können. In einer abschließenden Evaluierung mit dem MagnaTagATune Datensatz[1] werden die Verfahren miteinander verglichen. Damit wird von der Modellierung bis hin zur Anwendung und Evaluierung ein umfassender Rahmen für die Entwicklung von MIR Systemen basierend auf einem adaptiven Ähnlichkeitsmaß beschrieben. Darüber hinaus können die entwickelten Techniken auch problemlos auf andere Medien wie z.B. Bilder angewendet werden, solange geeignete Abstandsmaße für die entsprechenden Facetten vorhanden sind.

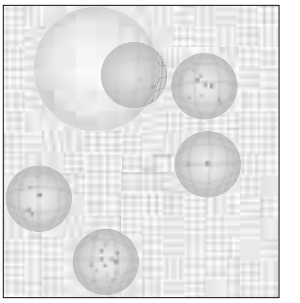
5 Weiterführende Arbeiten

5.1 Bisoziative Exploration von Musiksammlungen

In Zusammenarbeit mit Stefan Haun wurde im Rahmen des EU-Projektes "BISON"[2] ein Ansatz zur Entdeckung von Bisoziationen in großen Informationsräumen entwickelt [SHN11]. Der Begriff der Bisoziation geht auf den Künstler Arthur Köstler zurück [Kös64] und bezeichnet eine Assoziation, bei der Domaingrenzen überschritten werden und somit eine nicht offensichtliche Verknüpfung hergestellt wird. Durch Zweckentfremdung des Sekundärfokus der fokusadaptiven SpringLens können Bisoziationen zwischen Objekten in einer Sammlung hervorgehoben werden. Am Beispiel von MusicGalaxy wird gezeigt, wie sich dadurch interessante und überraschende Musikempfehlungen finden lassen. Die Technik lässt sich jedoch auch ohne Weiteres auf andere Anwendungsszenarien übertragen. Die grundlegende Idee dabei ist, zwei grundverschiedene Sichten auf die Sammlung in der Visualisierung miteinander zu kombinieren. Die primäre Sicht wird dazu direkt durch die Galaxiekarte dargestellt und kann durch die Anpassung des zugrundeliegenden Abstandsmaßes verändert werden. Die sekundäre Sicht wird hingegen indirekt durch den Sekundärfokus der SpringLens visualisiert, welcher nun nächste Nachbarn bezüglich der sekundären Sicht hervorhebt. Dabei kann die sekundäre Sicht verschiedene Formen haben. Zum einen ist es möglich, ein alternatives Abstandsmaß zu verwenden – beispielsweise eines mit Facettengewichten, die orthogonal zum Abstandsmaß der primären Sicht gewählt sind (Vgl. Abbildung 3). So könnte z.B. ein Abstandsmaß, welches klangliche Eigenschaften der Musikstücke berücksichtigt, für die Berechnung der Karte mit einem sekundären

[1] `http://tagatune.org/Magnatagatune.html`
[2] Bisociation Network for Creative Information Discovery, `http://www.bisonet.eu/`





projection weights

dynamics	0.0
rhythm	1.0
timbre	0.0

distortion weights

dynamics	1.0
rhythm	0.0
timbre	1.0

Abbildung 3: Kombination von zwei Abstandsmaßen mit orthogonalen Facettengewichten. Links: MusicGalaxy (invertiertes Farbschema). Rechts oben: Entsprechende Linsenverzerrung bestehend aus vom Benutzer kontrollierten Primärfokus (rot) und adaptiven Sekundärfokus mit mehreren kleinen Linsen (blau). Rechts unten: Facettengewichte zur Berechnung der Galaxiekarte (projection weights) und zur Erkennung nächster Nachbarn für die Verzerrung (distortion weights).

Abstandsmaß, welches die Songtexte vergleicht, kombiniert werden. Lenkt der Benutzer nun den Primärfokus auf ein Musikstück, wird der Sekundärfokus textlich ähnliche Stücke hervorheben, wobei in der Karte weit entfernte Stücke besonders interessant sind, da sie ein ähnliches Thema anders musikalisch umsetzen. Durch Navigation über den Sekundärfokus kann der Benutzer somit schrittweise an sowohl musikalisch als auch textlich ansprechende Musikstücke herangeführt werden. Alternativ ist es auch möglich, für die sekundäre Sicht eine Graphstruktur zu verwenden, in der Beziehungen zwischen Objekten explizit in Form von Kanten vorliegen. Nächste Nachbarn können hier direkt durch Traversieren des Graphen gefunden werden. Diese Konstellation ist besonders interessant, weil sie die Zusammenführung von karten- und graphbasierten Ansätzen ermöglicht, die sonst nur getrennt betrachtet werden, weil sie zwei völlig unterschiedlichen Explorationsstrategien entsprechen. MusicGalaxy konstruiert beispielsweise aus den Informationen der Musicbrainz Datenbank[3] einen Graph für die Musikstücke in der Sammlung, der zur Steuerung des Sekundärfokus verwendet werden kann. Dabei werden unter anderem auch Beziehungen zu Künstlern, Alben, Plattenlabeln mit einbezogen.

[3] http://musicbrainz.org/

5.2 Blickgesteuerter Adaptiver Fokus

In Zusammenarbeit mit Sophie Stellmach wurde auf Basis der fokusadaptiven SpringLens schließlich eine blickgestützte Benutzerschnittstelle zur Exploration von Mediensammlungen umgesetzt. Die grundlegende Idee ist hier, dass der Primärfokus statt mit der Maus direkt mit dem Blick gesteuert werden kann. Dabei kommt ein Eyetracker zum Einsatz, der bereits zur Evaluierung der fokusadaptiven SpringLens Visualisierung genutzt wurde. Blickinformationen bieten sich als natürliche Eingabemethode an, da der Blick oft einer manuellen Aktion vorausgeht. Die Verwendung von Fischaugenlinsen in der fokusadaptiven SpringLens hat zudem den Vorteil, dass dadurch Objekte von Interesse lokal vergrößert werden und somit durch den Blick auch leichter ausgewählt werden können. Bei rein blickgesteuerten Benutzerschnittstellen tritt jedoch das sogenannte "Midas Touch" Problem [Jac90] auf, welches nach König Midas aus der griechischen Mythologie benannt ist, der alles, was er anfasste, in Gold verwandelte – ob er wollte oder nicht. Hier bezieht sich dies auf die Schwierigkeit, gewollte von ungewollten Aktionen zu unterscheiden. Um diesem Problem entgegenzuwirken, wurde hier eine zweite Eingabemodalität hinzugenommen. Untersucht wurde dabei die Kombination mit einer Tastatur (repräsentativ für Tasteneingabegeräte wie z.B. Fernbedienungen) und mit einem Smartphone (mit Lagesensor und Touch-Eingabe). Durch eine frühe Einbindung von Benutzern in einen nutzerzentrierten Designprozess konnten entsprechend intuitive und natürliche Interaktionstechniken herausgearbeitet werden. Diese wurden schließlich im "GazeGalaxy" Prototyp umgesetzt, der eine Erweiterung von MusicGalaxy darstellt. Die Ergebnisse einer ersten Nutzerstudie mit diesem Prototyp belegen, dass der Blick tatsächlich als natürlicher Eingabekanal dienen kann und die Verwendung zur Steuerung der Fischaugenlinse als intuitiv betrachtet wird, solange bestimmte grundlegende Designrichtlinien berücksichtigt werden: Erstens sind Blickdaten von Natur aus ungenau und daher sollte die Interaktion nicht von genauen Positionen abhängen. Zweitens sollten Benutzer ihre Aktionen durch zusätzliche explizite Befehle bestätigen können, um ungewollte Aktionen zu vermeiden. Der resultierende Konferenzbeitrag [SSDN11] erhielt auf der NGCA 2011 den Best Paper Award. An weiteren Verbesserungen der Interaktionstechniken wird derzeit gearbeitet.

Danksagung

Die hier zusammengefasste Dissertation fand im Rahmen des DFG-finanzierten Projektes "AUCOMA - Adaptive User-Centered Organisation of Music Archives" statt und wurde zusätzlich durch ein Promotionsstipendium der Studienstiftung des deutschen Volkes gefördert. Die Entwicklung der bisoziativen SpringLens wurde durch die Europäische Kommission unterstützt (FP7-ICT-2007- C FET-Open, contract no. BISON-211898).

Literatur

[DNS07] Alexander Duda, Andreas Nürnberger und Sebastian Stober. Towards Query by Singing/Humming on Audio Databases. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*, 2007.

[GGSS06] Tobias Germer, Timo Götzelmann, Martin Spindler und Thomas Strothotte. SpringLens: Distributed Nonlinear Magnifications. In *Eurographics 2006 - Short Papers*, 2006.

[Jac90] Robert J. K. Jacob. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people*, 1990.

[Kös64] Arthur Köstler. *The Act of Creation*. Macmillan, 1964.

[RSN08] Johannes Reinhard, Sebastian Stober und Andreas Nürnberger. Enhancing Chord Classification through Neighbourhood Histograms. In *Proceedings of the 6th International Workshop on Content-Based Multimedia Indexing (CBMI'08)*, 2008.

[SHN11] Sebastian Stober, Stefan Haun und Andreas Nürnberger. Creating an Environment for Bisociative Music Discovery and Recommendation. In *Proceedings of Audio Mostly 2011 – 6th Conference on Interaction with Sound*, 2011.

[SN10] Sebastian Stober und Andreas Nürnberger. A Multi-Focus Zoomable Interface for Multi-Facet Exploration of Music Collections. In *Extended Proceedings of 7th International Symposium on Computer Music Modeling and Retrieval (CMMR'10)*, 2010.

[SN12] Sebastian Stober und Andreas Nürnberger. Adaptive Music Retrieval - A State of the Art. *Multimedia Tools and Applications*, 2012. (Online First Article).

[SSDN11] Sophie Stellmach, Sebastian Stober, Raimund Dachselt und Andreas Nürnberger. Designing Gaze-supported Multimodal Interactions for the Exploration of Large Image Collections. In *Proceedings of 1st International Conference on Novel Gaze-Controlled Applications (NGCA'11)*, 2011.

[SSN09] Sebastian Stober, Matthias Steinbrecher und Andreas Nürnberger. A Survey on the Acceptance of Listening Context Logging for MIR Applications. In *Proceedings of the 3rd Workshop on Learning the Semantics of Audio Signals (LSAS'09)*, 2009.

[Sto11] Sebastian Stober. Adaptive Distance Measures for Exploration and Structuring of Music Collections. In *Proceedings of AES 42nd Conference on Semantic Audio*, 2011.

Sebastian Stober, geboren am 10. Dezember 1980 in Halberstadt, studierte von 2000 bis 2005 Informatik mit Nebenfach Mathematik an der *Otto-von-Guericke-Universität* in Magdeburg mit einem siebenmonatigen Forschungsaufenthalt an der *University of Melbourne* in Australien. Seit 2006 ist er wissenschaftlicher Mitarbeiter in der *Data & Knowledge Engineering Gruppe* mit den Forschungsschwerpunkten Maschinelles Lernen, Musik Information Retrieval und Adaptive Systeme. Er organisierte 2006 den *1st International Workshop on Learning Semantics of Audio Signals (LSAS)* und zwei Folgeveranstaltungen 2007 und 2009. Seit 2007 ist er Mitorganisator des *International Workshop on Adaptive Multimedia Retrieval (AMR)*. Seine Arbeit als Doktorand beendete er 2011 erfolgreich mit seiner Dissertation zum Thema "Adaptive Methods for User-Centered Organization of Music Collections."

Multioperator Weighted Monadic Datalog

Torsten Stüber

Institut Theoretische Informatik
Fakultät Informatik
Technische Universität Dresden
Nöthnitzer Str. 46
01062 Dresden
torsten.stueber@tu-dresden.de

Abstract: In dieser Arbeit stellen wir ein formales Modell zur Verarbeitung von baumstrukturierten Daten vor. Dieses vereint und generalisiert Konzepte von Baumautomaten, Attributgrammatiken, Monadic Datalog und MSO-Logik. Unser Modell verarbeitet Bäume unter Einsatz von *Multioperator-Monoiden* und erhält dadurch eine hohe Ausdrucksstärke und vielseitige Einsetzbarkeit. In unserer Arbeit beschreiben und vergleichen wir vier verschiedene Semantiken sowie zahlreiche Normalformen. Einige der Semantiken sind nur dann einsetzbar, wenn das Modell kein zirkuläres Verhalten aufweist; daher entwickeln wir des Weiteren einen Zirkularitätstests. Wir betrachten beispielhaft einige Instanzen unseres Modells und stellen Bezüge zu existierenden Berechnungsmodellen her.
Diese Kurzfassung ist eine Zusammenfassung der Kapitel der Dissertation.

1 Einführung

Einer der Kernaspekte der Informatik ist die Untersuchung und Entwicklung von Berechnungsmodellen. Berechnung in diesem Sinne ist die Verarbeitung von Eingabe- zu Ausgabedaten. Gewöhnlich sind die zu verarbeitenden Daten nach einem gewissen Schema organisiert, zum Beispiel in einer Baumstruktur. Baumstrukturierte Daten haben viele Anwendungen in der Informatik, beispielsweise im Bereich der semistrukturierten Datenbanken, der formalen Sprachtheorie oder der Übersetzung natürlicher Sprachen.

Anwendungen, bei denen Bäume als Eingabe verwendet werden, lassen sich gemäß der Art der Ausgabe ihrer Verarbeitung einteilen. Die wichtigsten Vertreter einer solchen Einteilung sind:

- *Prüfen einer Eigenschaft*: beispielsweise die Überprüfung darauf, ob der Eingabebaum ein Binärbaum ist,
- *Berechnung eines Zahlenwertes*: beispielsweise das Zählen der Blattknoten eines Baumes,
- *Transformation des Baumes*: beispielsweise die Übersetzung des Ableitungsbaumes

eines englischen Satzes in einen entsprechenden Ableitungsbaum eines deutschen Satzes im Anwendungsbereich der Syntax-basierten maschinellen Übersetzung,

- *Transformation des Baumes bei gleichzeitiger Berechnung eines Zahlenwertes*: beispielsweise die probabilistische Übersetzung eines Ableitungsbaumes im Anwendungsbereich der statistischen maschinellen Übersetzung.

Die Informatik hat eine große Anzahl an Berechnungsmodellen hervorgebracht und untersucht. Gängige Beispiele sind die Turingmaschine, Programmiersprachen, neuronale Netze oder endliche Automaten, also Modelle mit begrenztem Speicher. Forschung im Bereich der Automatentheorie hat zu vielen fruchtbaren Entdeckungen und Anwendungen geführt, da endliche Automaten meist eine einfache, reguläre Struktur besitzen, effizient verarbeiten können, robuste Berechnungsklassen definieren (d.h. vergleichbare Modelle weisen die gleiche Ausdrucksmächtigkeit auf) und oftmals befriedigende Berechnungsmächtigkeiten aufweisen.

Im Folgenden geben wir einen Überblick über endliche Modelle zur Verarbeitung von baumstrukturierten Daten. Wir konzentrieren uns dabei auf zwei Entwicklungszweige solcher Modelle.

Endliche Automaten und Transducer Die Theorie der Baumautomaten [GS97] begann in der Mitte der 1960er Jahre und verallgemeinerte das Konzept der endlichen Zeichenreihenautomaten. Die Semantik eines FTA ist eine Baumsprache; sie besteht aus allen Bäumen, die durch den FTA akzeptiert werden.

Das Konzept des FTA kann auf natürliche Art und Weise durch eine Gewichtung der Transitionen erweitert werden. Das Rechnen mit diesen Gewichten erfordert den Einsatz einer algebraische Struktur; gewöhnlich sind Halbringe [HW98] für diese Anwendung besonders geeignet. Diese Erweiterung führt zu dem Konzept der gewichteten Baumautomaten (*weighted tree automaton* – WTA) [BR82]. Ein gegebener WTA definiert eine Abbildung von der Menge der Eingabebäume in die Trägermenge des Halbrings. Gewichtete Baumautomaten haben viele Anwendungen in der Informatik, zum Beispiel für die Befehlsauswahl in Kompilern, Baummusterabgleich, und die natürliche Sprachverarbeitung [KM09].

WTA können als eine Erweiterung von FTA angesehen werden. Weitere Modelle, die von FTA durch ähnliche Erweiterungen abgeleitet wurden, sind tiefgründig untersucht und fruchtbar angewandt wurden. Wir nennen hier beispielhaft das Modell der Baumübersetzer (*tree transducer* – TT) [Eng75]. Ein TT definiert eine *Baumübersetzung*, also ein Abbildung von Eingabebäumen in Mengen von Ausgabebäumen. Auch TT können durch einen Halbring gewichtet werden. Das führt zu dem Modell der gewichteten Baumübersetzer (*weighted tree transducer* – WTT).

Da FTA, WTA, TT und WTT eine ähnliche Struktur aufweisen, wurde ein vereinheitlichendes Modell entwickelt, welches alle ebengenannten Modelle subsumiert. Dieses Modell, bezeichnet als gewichteter Multioperator-Baumautomat (*weighted multioperator tree automaton* – WMTA), wurde erstmalig von Kuich [Kui99] untersucht und nutzt zur Berechnung eine Algebra namens Multioperator-Monoid (kurz *M-Monoid*). Durch gezielten Einsatz verschiedener M-Monoide kann ein WMTA die Modelle FTA, WTA, TT und WTT

simulieren. Das heißt, dass Untersuchungen und Resultate für WMTA entsprechende Untersuchungen und Resultate für die vier letztgenannten Modelle automatisch beinhalten.

Monadic Datalog *Monadic Datalog* [GK02], ein Fragment von Datalog, ist ein Mittel zum formalen Beschreiben von Baumsprachen und zur Knotenauswahl von Bäumen. Ein Monadic-Datalog-Programm (kurz: MD) besteht im Wesentlichen aus einer Menge von Regeln.

Monadic Datalog kann auf natürliche Art und Weise für Rangbäume als auch für rangfreie Bäume verwendet werden. Damit hat es Anwendungen im Bereich der semistrukturierten Datenbanken und XML. Das große praktische Potential von Monadic Datalog ist wie folgt begründet: (i) die Benutzung von Regeln zur Spezifikation von Eigenschaften von Baumsprachen ist intuitiv und natürlich, (ii) die Zeitkomplexität zur Auswertung eines MD ist linear im Bezug auf die Größe des Eingabebaumes und des betrachteten MD und (iii) die Klasse der Baumsprachen, die mit Monadic Datalog spezifiziert werden können, ist die Klasse der durch FTA erkennbaren Baumsprachen (vgl. [GK02]).

Gewichtetes Monadic Datalog (*weighted monadic datalog* – WMD) ist die durch einen Halbring gewichtete Variante von Monadic Datalog. Es wurde gezeigt, dass WMD strikt ausdrucksstärker als WTA sind und dass es ähnlich wie Monadic Datalog die Eigenschaft hat, in Linearzeit auswertbar zu sein (sowohl in Abhängigkeit von der Größe des Eingabebaums als auch des WMD) (vgl. [SV08]).

Eine weitere Abwandlung von Monadic Datalog, die Monadic-Datalog-Baumübersetzer (kurz MDTT), erlauben die Spezifikation von endlichen Baumübersetzungen (d.h., die Menge der Ausgabebäume für jeden Eingabebaum ist endlich) als auch von unendlichen Baumübersetzungen (d.h., die Menge der Ausgabebäume für einen gegebenen Eingabebaum kann unendlich sein). Es wurde gezeigt, dass MDTT mindestens die Ausdrucksstärke von attributierten Baumübersetzern aufweisen [Fül81] (siehe [BS09]).

Zielstellung dieser Arbeit In dieser Arbeit untersuchen wir eine Generalisierung von MD, WMD und MDTT, die darüber hinaus die Modelle FTA, WTA, TT und WMTA subsumiert. Dieses Modell heißt *Multioperator-gewichtetes Monadic Datalog* (*multioperator weighted monadic datalog* – MWMD). In Tabelle 2 wird dargestellt, wie sich dieses Modell in die Landschaft der wichtigsten obengenannten Modelle eingliedert.

	Baumsprache	Baumreihe	Baumübersetzung	Allgemeine Berechnung
Baumautomaten	FTA	WTA	TT	WMTA
Monadic Datalog	MD	WMD	MDTT	MWMD

Tabelle 1: Eine Übersicht über die wichtigsten genannten Formalismen.

MWMD stellen also ein verallgemeinertes Konzept zum Verarbeiten baumstrukturierter Daten dar. Sie erben dabei die positiven Eigenschaften sowohl von der auf Baumautomaten als auch von der auf Monadic Datalog basierenden Modelle: Auswertbarkeit bei ge-

ringer Zeitkomplexität, einfache Spezifizierbarkeit von Baumsprachen, Baumreihen oder Baumübersetzungen, große Ausdrucksmächtigkeit und ein großes Gebiet von Anwendungen durch die Verwendung von M-Monoiden.

Der universell einsetzbare Formalismus MWMD ermöglicht darüber hinaus eine unifizierte Sichtweise auf die zahlreichen Anwendungen und Eigenschaften konventioneller Modelle zum Verarbeiten von Bäumen und führt diese zu einem Ganzen zusammen. In der Dissertation werden im Speziellen Resultate präsentiert, die belegen, dass eine einheitliche Betrachtung baumbasierter Modelle gewinnbringend ist.

2 Grundlagen

Da MWMD eine Vielzahl an Verarbeitungsmodellen vereinen, basieren sie auf einer großen Zahl von Konzepten der Mathematik und Theoretischen Informatik. In diesem Kapitel besprechen wir die wichtigsten Begriffe, die für das grundlegende Verständnis der nachfolgenden Kapitel notwendig sind. Dabei gehen wir im Speziellen auf Operationen mit unendlicher Stelligkeit ein. Das Kapitel schließt mit einer Einführung in das Gebiet der Hypergraphen und Hyperpfade ab. Dabei richten wir besonderes Augenmerk auf Hyperpfad-Segmente, Abhängigkeitsbeziehungen und Dekomposition sowie Komposition von Hyperpfaden.

3 M-Monoide

Dieses Kapitel behandelt die zentrale algebraische Struktur unseres Formalismus, sogenannte Multioperator-Monoide (kurz *M-Monoide*). Unsere Definition der M-Monoide basiert auf dem Begriff der distributiven M-Monoide (oder distributiven Ω-Monoide), die auf Kuich [Kui99] zurückgehen.

Ein M-Monoid besteht zum einen aus einem kommutativen Monoid und zum anderen aus einer Δ-Algebra (für eine Signatur Δ); das Monoid und die Δ-Algebra sind dabei auf der gleichen Trägermenge definiert. Formal ist ein M-Monoid also ein Tupel $\mathcal{A} = (A, +, \mathbf{0}, \theta)$, wobei $(A, +, \mathbf{0})$ ein kommutatives Monoid und (A, θ) eine Δ-Algebra ist. M-Monoide sind ein essenzieller Bestandteil für die Definition der Semantik eines MWMD-Programms, da deren Auswertung die Monoid-Operation und die Operationen der Δ-Algebra des M-Monoids benutzt.

Es zeigt sich, dass die Operationen eines M-Monoids im Allgemeinen nicht ausreichend sind, um die Semantik eines beliebigen MWMD zu berechnen. Das tritt immer dann ein, wenn der betrachtete MWMD ein zirkuläres Verhalten aufweist. Diesem Problem begegnen wir dadurch, dass wir zwei Erweiterungen von M-Monoiden einführen, mit denen es möglich ist, wohldefinierte Ausgabewerte für zirkuläre MWMD zu berechnen. Diese Erweiterungen heißen *ω-vollständige* und *ω-stetige M-Monoide*. Des Weiteren arbeiten wir Beziehungen zwischen ω-vollständigen und ω-stetigen M-Monoiden heraus.

4 M-gewichtete Monadic-Datalog-Programme

In diesem Kapitel präsentieren wir das Kernmodell der Arbeit und beschreiben dabei detailliert die Syntax und Semantik von MWMD. Die syntaktische Struktur eines MWMD orientiert sich an der Syntax von MDTT [BS09] und WMD [SV08]. Wir beschreiben diese nun im Ansatz.

Sei Σ ein Rangalphabet und Δ eine Signatur. Ein MWMD-Programm *über* Σ *und* Δ ist ein Tripel (P, R, q), wobei P ein Rangalphabet ist, in dem jedes Symbol ein- oder nullstellig ist, R eine endliche Menge und q ein einstelliges Symbol aus P ist. Die Elemente von P heißen *nutzerdefinierte Prädikate*, die Elemente von R *Regeln*, und q ist das *Abfrageprädikat*. Jede Regel r aus R ist von der Form

$$\textit{Kopf} \leftarrow \textit{Rumpf}\,;\,\textit{Guard}\ .$$

Der Kopf der Regel ist dabei ein prädikatenlogisches Atom über einem Prädikat aus P, also von der Form $p()$ für ein nullstelliges p oder $p(x)$ für ein einstelliges p; x ist eine Variable und wird einem vorgegebenen Vorrat an Variablen entnommen. Der Rumpf ist ein Baum mit Symbolen aus Δ, an dessen Blättern zusätzlich prädikatenlogische Atome über P stehen dürfen; diese Atome sind auf gleiche Art und Weise aufgebaut wie der Regelkopf. Der Guard ist eine endliche Menge von Atomen über der Prädikatenmenge sp_Σ:

$$\mathrm{sp}_\Sigma = \{\mathrm{root}^{(1)}, \mathrm{leaf}^{(1)}\} \cup \{\mathrm{label}_\sigma^{(1)} \mid \sigma \in \Sigma\} \cup \{\mathrm{child}_i^{(2)} \mid i \in [\mathrm{maxrk}(\Sigma)]\}\ .$$

Die Definition der Semantik von MWMD ist komplex und stellt einen Schwerpunkt der Dissertation dar. Um eine reichhaltige Theorie der MWMD zu entwickeln, führen wir zwei verschiedene Arten von Semantiken ein, die als *Fixpunktsemantik* und *Hypergraphsemantik* bezeichnet sind. Die Fixpunksemantik hat Ähnlichkeit zur Initial-Algebra-Semantik eines WTA [BR82]; dagegen weist die Hypergraphsemantik Parallelen zur Lauf-Semantik von gewichteten Baumautomaten auf. Die Fixpunktsemantik basiert auf der Semantik von MDTT, WMD und MD, die Hypergraphsemantik ist dagegen ein neuartiges Konzept.

Jede dieser beiden Semantikarten benötigt drei Eingaben: ein MWMD-Programm, einen Eingabebaum und ein M-Monoid. Die Semantiken sind so definiert, dass sie den Eingabebaum gesteuert durch das MWMD-Programm auswerten und dabei die Operationen des M-Monoids anwenden. Die Ausgabe ist schließlich ein Element des M-Monoids. Betrachtet man also ein festes MWMD-Programm und ein festes M-Monoid, dann sind die Semantiken Abbildungen von Eingabebäumen in die Trägermenge des M-Monoids; eine solche Abbildung heißt Baumreihe.

Die Fixpunktsemantik basiert auf der Anwendung eines *Konsequenz-Operators* und geht auf die Definition der Semantik der Logikprogrammierung bzw. Hornklauseln und Monadic Datalog [GK02] zurück. Wir beschreiben nun die grobe Idee der Fixpunktsemantik. Zunächst werden die Variablen jeder Regel r in R durch Knoten des Eingabebaumes auf all solche Weisen instanziiert, dass der Guard von r erfüllt ist; lautet der Guard beispielsweise $\{\mathrm{label}_\sigma(x), \mathrm{child}_2(x, y)\}$, dann darf x nur durch einen mit σ beschrifteten Knoten

und y durch den zweiten Kindknoten von x instanziiert werden. Eine *Interpretation* ordnet jeder Instanz der in den Regeln vorkommenden Atomen ein Element des betrachteten M-Monoids zu. Ausgehend von einer festen Startinterpretation wird durch wiederholtes Anwenden des Konsequenz-Operators schrittweise eine Zielinterpretation berechnet. Der Konsequenz-Operator nimmt dabei Bezug auf die instanziierten Regeln: beispielsweise bedeutet die Regelinstanz

$$a \leftarrow \delta\big(b, c\big)\,;\{\dots\}\,,$$

dass die Konsequenz-Interpretation der Atominstanz a den Wert $\delta\big(I(b), I(c)\big)$ zuordnet, wobei I die aktuelle Interpretation ist und die Operation δ durch die Δ-Algebra des M-Monoids interpretiert wird. Das Ergebnis der Semantik ist dann der Wert, den die Zielinterpretation der Atominstanz zuordnet, die aus dem Abfrageprädikat und dem Wurzelknoten des Eingabebaums besteht.

Die Hypergraphsemantik ordnet einem MWMD-Programm und einem Eingabebaum einen Hypergraphen, den *Abhängigkeitshypergraphen*, zu. Aus diesem Hypergraphen wird eine Menge von Termen über der Signatur Δ extrahiert. Jeder dieser Terme wird anschließend in der Δ-Algebra des betrachteten M-Monoids ausgewertet. Die Ergebnisse der Auswertung für jeden Term werden dann durch das kommutative Monoid verknüpft. Das entstehende Element des M-Monoids ist schließlich das Ergebnis der Hypergraphsemantik.

Für jede der beiden Semantiken definieren wir eine *endliche* und eine *unendliche* Variante. Die endlichen Varianten der Semantiken sind dabei nur auf einer Teilklasse der MWMD-Programme definiert, den *schwach nichtzirkulären* MWMD*-Programmen*. Außerhalb dieser Klasse liegende MWMD-Programme weisen beim Berechnen der Semantik ein zirkuläres Verhalten auf und sind nur mit den unendlichen Varianten der Semantiken anwendbar; diese erfordern dabei allerdings ein ω-stetiges M-Monoid (bei der Fixpunktsemantik) oder ein ω-vollständiges M-Monoid (bei der Hypergraphsemantik) als Eingabe. Insgesamt untersuchen wir also vier verschiedene Semantiken.

Die folgende Tabelle gibt eine Übersicht darüber, welche der vier Semantiken unter welchen Umständen einsetzbar ist.

	Fixpunktsemantik	Hypergraphsemantik
endlich	schwach nichtzirkuläre MWMD beliebige M-Monoide	schwach nichtzirkuläre MWMD beliebige M-Monoide
unendlich	beliebige MWMD ω-stetige M-Monoide	beliebige MWMD ω-vollständige M-Monoide

Tabelle 2: Eine Übersicht über die Semantiken von MWMD-Programmen.

Wir schließen das Kapitel mit einem Vergleich der vier Semantiken ab und erarbeiten hinreichende Bedingungen für ihre Äquivalenz (siehe Theorem 4.53).

5 Normalformen

In diesem Kapitel betrachten wir vier syntaktische Teilklassen der MWMD, sogenannte *eingeschränkte*, *zusammenhängende*, *lokale* und *echte* MWMD. Wir beschreiben diese Klassen nun informell.

Eingeschränkt: die Position der in den Regeln vorkommenden Variablen muss bestimmten strukturellen Eigenschaften genügen.

Zusammenhängend: die Variablen in den Regeln müssen logisch zusammenhängen.

Lokal: die Regeln müssen so definiert sein, dass die Variablen bei der Bildung von Regelinstanzen nur mit direkt benachbarten Knoten des Eingabebaumes instanziiert werden; die Struktur der Regeln ähnelt dabei den Regeln von Attributgrammatiken [Cou84].

Echt: alle nutzerdefinierten Prädikate sind einstellig.

Die Teilklasse der zusammenhängenden MWMD wurde in dieser Form erstmalig von Gottlob und Koch [GK02, Theorem 4.2] für MD eingeführt; sie wurde weiterhin in [SV08, BS09] untersucht. Die anderen drei Klassen wurden in [BS09] für MDTT eingeführt.

Wir untersuchen hinreichende Bedingungen, die eine Übereinstimmung zwischen diesen syntaktischen Klassen garantieren. Es handelt sich dabei also um Bedingungen, unter denen diese Teilklassen (und Schnitte der Teilklassen) Normalformen von MWMD bilden (siehe Theorem 5.8). Dazu analysieren wir, wann ein gegebenes MWMD-Programm in ein semantisch äquivalentes MWMD-Programm transformiert werden kann, der zu einer der genannten Teilklassen gehört. Hierbei ist zu klären, was wir unter „semantischer Äquivalenz“ verstehen. In der Tat gebrauchen wir die stärkste Definition von semantischer Äquivalenz, die in diesem Kontext möglich ist. Genauer gesagt präsentieren wir Konstruktionen, die alle vier im vorherigen Kapitel eingeführten Semantiken erhalten. Dies stellt eine besondere Herausforderung dar und ist im Detail sehr technisch. Die Konstruktionen, die wir in diesem Kapitel vorstellen, basieren auf Konstruktionen in [BS09].

6 Zirkularitätstest

In diesem Kapitel beweisen wir, dass es ein effektives Verfahren gibt, mit dem man entscheiden kann, ob ein MWMD-Programm schwach nichtzirkulär ist (siehe Theorem 6.1). Dieses Resultat ist wichtig, da nur so entschieden werden kann, welche der vier Semantiken für ein MWMD-Programm in einer konkreten Situation anwendbar sind.

Die Definition von schwacher Nichtzirkularität beruht auf dem Begriff der Nichtzirkularität von Attributgrammatiken [Cou84] und WMD [SV08]. Ein Entscheidungsverfahren für die Nichtzirkularität von Attributgrammatiken, bezeichnet als Zirkularitätstest, wurde erstmals durch Knuth [Knu68] untersucht. Es basiert auf einer rekursiven Konstruktion von endlichen Mengen von Graphen, sogenannten IS-Graphen, die auf Zyklen überprüft werden.

In dieser Dissertation verfolgen wir nicht den Ansatz, einen Zirkularitätstest zu entwickeln,

der auf IS-Graphen basiert, da er sich als zu schwierig und komplex erweist. Stattdessen wenden wir die folgende Idee an. Sei M ein MWMD-Programm und L_M die Menge der Eingabebäume, für die M ein zirkuläres Verhalten aufweist. Dann ist M schwach nichtzirkulär genau dann, wenn L_M leer ist. Wir zeigen, dass effektiv eine MSO-Formel [TW68] konstruiert werden kann, die L_M definiert. Das impliziert, dass L_M eine erkennbare Baumsprache ist. Dann folgt die Entscheidbarkeit der schwachen Nichtzirkularität aus der Tatsache, dass das Leerheitsproblem für erkennbare Baumsprachen entscheidbar ist.

7 Gewichtetes Monadic Datalog

In diesem Kapitel zeigen wir, dass MWMD das Konzept der WMD subsumiert. Dazu untersuchen wir die Semantik von MWMD für die Teilklasse von M-Monoiden, die das algebraische Verhalten von Halbringen [HW98] simulieren. Um möglichst starke Resultate zu erhalten, betrachten wir sogar M-Monoide, die starke Bimonoide [DSV10] simulieren. Dieses Kapitel ist eine überarbeitete und erweiterte Version der Arbeit [SV08]; wir merken an, dass in jener Arbeit WMD über Halbringen und rangfreien Bäumen betrachtet wurde; dagegen behandelt diese Dissertation WMD über starken Bimonoiden und Rangbäumen.

Kern unserer Untersuchungen sind die Ausdrucksstärke und die Effizienz der Auswertung von WMD. Im Speziellen vergleichen wir die endliche und unendliche Variante der Semantik von WMD (Lemma 7.15) und zeigen, dass WMD unter Benutzung des Booleschen Halbrings MD simulieren können und dass WMD streng ausdrucksstärker sind als WTA (Theorem 7.18). Wir schließen die Betrachtungen durch einen Beweis dafür ab, dass WMD effizient ausgewertet werden können (Theorem 7.21).

8 Monadic-Datalog-Baumübersetzer

Dieses Kapitel behandelt das Konzept der MDTT. Wir zeigen, dass die Klasse der MDTT in der Klasse der MWMD enthalten ist. Das erreichen wir durch den Einsatz von speziellen M-Monoiden. Diese M-Monoide verhalten sich ähnlich einer Termalgebra und sorgen so zum Beispiel bei der Hypergraphsemantik dafür, dass die Auswertung eines Termes den Term selbst erzeugt. Dies macht deutlich, dass MDTT nichts anderes als MWMD sind, bei denen von einer konkreten semantischen Domäne abstrahiert wurde. Dieses Kapitel ist eine überarbeitete und erweiterte Version der Arbeit [BS09], in der MDTT erstmalig untersucht wurden.

Wir zeigen, dass es eine scharfe Abgrenzung zwischen solchen MDTT gibt, für die die Semantik in linear beschränkter Zeit vollständig berechnet werden kann (diese MDTT sind demnach für praktische Zwecke einsetzbar) und solchen MDTT, für die die Semantik nicht in endlicher Zeit berechnet werden kann. Erstere Sorte von MDTT bezeichnen wir als *ausführbar* und zeigen, dass Ausführbarkeit von MDTT entscheidbar ist (siehe Lemma 8.12).

MDTT und (nichtdeterministische) attributierte Baumübersetzer [Fül81] sind konzeptuell eng verwandt. Wir beweisen, dass attributierte Baumübersetzer und eingeschränkte MDTT die gleiche Ausdrucksmächtigkeit haben (siehe Theorem 8.21).

9 Gewichtete Multioperator-Baumautomaten

Dieses Kapitel behandelt das Konzept der WMTA. Wir zeigen, dass MWMD insbesondere WMTA simulieren können. Dazu definieren wir eine syntaktische Teilklasse von MWMD, so dass sich die MWMD in dieser Klasse exakt wie WMTA verhalten. Dieses Kapitel ist eine überarbeitete Version der wichtigsten Resultate der Arbeiten [SVF09, FSV10]. Wir beschränken uns auf einen Beweis der folgenden beiden Hauptresultate:

1. Wir betrachten M-Monoide, welche gewisse zusätzliche Eigenschaften erfüllen. Wir zeigen, dass die Klasse der Baumreihen, welche durch WMTA über einem solchen M-Monoid erkannt werden, dekomponiert werden kann in (1) die Klasse der Relabelings, gefolgt von (2) der Klasse der charakteristischen Baumtransformationen von erkennbaren Baumsprachen, gefolgt von (3) der Klasse der Baumreihen, die durch Homomorphismus-WMTA erkannt werden. Ein Homomorphismus-WMTA ist ein WMTA mit genau einem Zustand (siehe Theorem 9.17).
2. Wir präsentieren eine alternative Charakterisierung der Klasse der Baumreihen, die durch WMTA erkannt werden. Diese Charakterisierung basiert auf sogenannten *M-Ausdrücken*, einer neuen Art von gewichteter MSO-Logik. Diese Charakterisierung ist ein Büchi-artiges Resultat [Büc60] für die Klasse der Baumreihen, die durch WMTAerkannt werden (siehe Theorem 9.26).

Literatur

[BR82] J. Berstel und C. Reutenauer. Recognizable formal power series on trees. *Theoret. Comput. Sci.*, 18(2):115–148, 1982.

[BS09] M. Büchse und T. Stüber. Monadic Datalog Tree Transducers. In A. H. Dediu, A.-M. Ionescu und C. Martín-Vide, Hrsg., *LATA*, Jgg. 5457 of *Lecture Notes in Computer Science*, Seiten 267–278. Springer, 2009.

[Büc60] J. R. Büchi. Weak Second–order arithmetic and finite automata. *Zeitschr. für math. Logik und Grundl. der Mathem.*, 6:66–92, 1960.

[Cou84] B. Courcelle. Attribute grammars: definitions, analysis of dependencies, proof methods. In B. Lorho, Hrsg., *Methods and tools for compiler construction*, Seiten 81–102. Cambridge University Press, 1984.

[DSV10] M. Droste, T. Stüber und H. Vogler. Weighted automata over strong bimonoids. *Inform. Sci.*, 180:156–166, 2010.

[Eng75] J. Engelfriet. Bottom-up and top-down tree transformations - a comparison. *Math. Systems Theory*, 9(3):198–231, 1975.

[FSV10] Z. Fülöp, T. Stüber und H. Vogler. A Büchi-Like Theorem for Weighted Tree Automata over Multioperator Monoids. *Theory of Computing Systems*, Seiten 1–38, 2010.

[Fül81] Z. Fülöp. On attributed tree transducers. *Acta Cybernet.*, 5:261–279, 1981.

[GK02] G. Gottlob und C. Koch. Monadic Queries over Tree-Structured Data. In *LICS '02: Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science*, Seiten 189–202, Washington, DC, USA, 2002. IEEE Computer Society.

[GS97] F. Gécseg und M. Steinby. Tree Languages. In G. Rozenberg und A. Salomaa, Hrsg., *Handbook of Formal Languages*, Jgg. 3, Kapitel 1, Seiten 1–68. Springer-Verlag, 1997.

[HW98] U. Hebisch und H.J. Weinert. *Semirings - Algebraic Theory and Applications in Computer Science*. World Scientific, Singapore, 1998.

[KM09] K. Knight und J. May. Applications of Weighted Automata in Natural Language Processing. In M. Droste, W. Kuich und H. Vogler, Hrsg., *Handbook of Weighted Automata*, Kapitel 14. Springer-Verlag, 2009.

[Knu68] D.E. Knuth. Semantics of context–free languages. *Math. Systems Theory*, 2:127–145, 1968.

[Kui99] W. Kuich. Linear systems of equations and automata on distributive multioperator monoids. In *Contributions to General Algebra 12 - Proceedings of the 58th Workshop on General Algebra "58. Arbeitstagung Allgemeine Algebra", Vienna University of Technology. June 3-6, 1999*, Seiten 1–10. Verlag Johannes Heyn, 1999.

[SV08] T. Stüber und H. Vogler. Weighted monadic datalog. *Theor. Comput. Sci.*, 403(2-3):221–238, 2008.

[SVF09] T. Stüber, H. Vogler und Z. Fülöp. Decomposition of weighted multioperator tree automata. *Int. J. Foundations of Computer Sci.*, 20(2):221–245, 2009.

[TW68] J.W. Thatcher und J.B. Wright. Generalized finite automata theory with an application to a decision problem of second-order logic. *Math. Syst. Theory*, 2(1):57–81, 1968.

Torsten Stüber

Von 2001 bis 2006 studierte Torsten Stüber an der Technischen Universität Dresden zunächst Informatik und anschließend im internationalen Masterstudiengang *Computational Logic*. 2005 führte er ein Auslandssemester an der University of Auckland, Neuseeland, durch. Als bester Absolvent seines Jahrgangs an der Fakultät Informatik der TU Dresden wurde er 2007 mit der Lohrmann-Medaille ausgezeichnet.

Seit Ende 2006 ist er wissenschaftlicher Mitarbeiter am Lehrstuhl *Grundlagen der Programmierung* der TU Dresden. In dieser Funktion hat er sich vorrangig mit der Theorie der Baumautomaten und -Logiken beschäftigt und über dieses Thema im Februar 2011 promoviert. Im Verlauf seiner wissenschaftlichen Arbeit wurde sein Interesse für die Bereiche der *natürlichen Sprachverarbeitung* und des *maschinellen Lernens* geweckt und er hat seine Tätigkeit nach seiner Promotion verstärkt auf diese Gebiete konzentriert.

Suche und Lernen im Immunsystem: Modelle der T-Zell-Zirkulation und der Negativauslese

Johannes Textor

Theoretical Biology & Bioinformatics
Universität Utrecht
Padualaan 8
3584 CH Utrecht, Niederlande
johannes.textor@gmx.de

Abstract: Das Immunsystem ist ein hochkomplexes verteiltes System, das größtenteils aus multifunktionalen und stets mobilen Zellen besteht. Viele wichtige immunologische Vorgänge sind bis heute nur unzureichend verstanden. In den letzten Jahren wurden durch die Entwicklung neuer Technologien wie der Multiphotonenmikroskopie neuartige und reichhaltige Daten gewonnen. Um aus diesen Daten fundierte Erkenntnisse zu gewinnen, werden in der modernen Immunologie zunehmend Computersimulationen und mathematische Modelle eingesetzt. Diese Arbeit befasst sich zum Einen mit der Überwachung des Organismus durch zirkulierende T-Zellen, die als stochastischer Suchprozess modelliert wird, und zum Anderen mit der Generierung von T-Zell-Rezeptoren durch die sogenannte Negativauslese, die als stochastischer Lernprozess modelliert wird. Beide Modelle werden anhand formaler Methoden aus der theoretischen Informatik untersucht, um ein tieferes Verständnis informationsverarbeitender und stochastischer Aspekte der modellierten Prozesse zu erlangen. Zudem werden aus den Modellen quantitative Vorhersagen generiert, die anhand experimenteller Daten validiert werden. Der Erkenntnisgewinn dieser formalen und quantitativen Analysen wird im Kontext aktueller immunologischer Forschung diskutiert und bewertet. Unter Anderem gelingt es, durch das stochastische Zirkulationsmodell eine Vielzahl bisher isolierter experimenteller Daten quantitativ in Beziehung zu setzen, während das algorithmische Modell der Negativauslese die aktuell besten Vorhersagen zur Erkennung von HIV-Peptiden durch sogenannte CD8-T-Zellen liefert.

1 Einleitung

Die Arbeit besteht aus zwei weitgehend unabhängigen Teilen, die allerdings einer gemeinsamen Struktur und einem gemeinsamen wissenschaftlichen Ansatz folgen, nämlich dem der mathematischen Biologie. Ähnlich der theoretischen Physik wird hierbei zunächst ein gutes *Problemverständnis* angestrebt, um dann das zu untersuchende Problem geeignet als mathematisches Modell zu *formalisieren*. Das erhaltene Modell wird dann im Hinblick auf analytische sowie quantitative *Lösungen* untersucht, die dann, und das ist der wichtigste Schritt, in biologisch sinnvolle *Erkenntnisse und Vorhersagen* übersetzt werden müssen. Die Vorhersagen dienen sowohl der Falsifizierbarkeit des Modells als auch der besseren Planung zukünftiger Experimente. Als Besonderheit verwenden wir in der

Dissertation Techniken aus der Wahrscheinlichkeitstheorie, der Lerntheorie und der Algorithmik, wohingegen die "traditionelle" mathematische Biologie vor allem kontinuierliche Modellierungstechniken wie Differentialgleichungen einsetzt. Die Resultate der Arbeit demonstrieren, dass solche Techniken tatsächlich zum Erkenntnisgewinn für die moderne Immunologie von Nutzen sein können.

2 T-Zell-Zirkulation

Die T-Zellen sind neben den B-Zellen eine der zwei Zellarten des Immunsystems, die dem Körper völlig unbekannte Pathogene erkennen können. Jede T-Zelle ist auf die Erkennung einiger weniger körperfremder Muster spezialisiert *(Spezifität)*, und kaum eine T-Zelle gleicht der Anderen *(Diversität)*. Durch die hohe Anzahl von T-Zellen (ca. 10^{11} beim Menschen, 10^8 bei einer Maus) wird so insgesamt ein umfassender Schutz erreicht. Dies bedeutet aber, dass die Anzahl von T-Zellen, die auf eine bislang unbekannte Infektion reagieren können, extrem klein ist – in einer Maus sind dies unter Umständen nur 20 Zellen [MCP+07]! Diese würden natürlich nicht ausreichen, um eine Infektion wirksam zu bekämpfen. Erkennt eine T-Zelle ein Antigen in einem Lymphknoten oder der Milz, beginnt sie sich deshalb zu teilen, wodurch sich die Population spezifischer Zellen innerhalb weniger Tage millionenfach vergrößert. Je früher ein eingedrungenes Antigen durch die wenigen spezifischen T-Zellen gefunden und erkannt wird, desto schneller kann diese Vermehrung beginnen. Daher befinden sich T-Zellen ständig auf Patrouille durch Lymphknoten, Milz und andere lymphatische Organe (Abbildung 1).

Die Wege der T-Zell-Zirkulation werden seit den 1960er Jahren erforscht. Bereits früh wurde bekannt, dass T-Zellen etwa einmal pro Tag einen Lymphknoten besuchen, wo sie mehrere Stunden verweilen. Über die Wege der Zellen *innerhalb* des Lymphgewebes konnte man allerdings lange Zeit nur spekulieren. Man nahm an, dass lymphatische Organe auf eine gerichtete, synchronisierte Art durchwandert werden. Im Jahr 2002 gelang es dann mittels der Zweiphotonenmikroskopie erstmals, einzelne T-Zellen beim Durchwandern von Lymphknoten im lebenden Tier zu beobachten [MWPC02]. Das Resultat war für viele Immunologen eine große Überraschung: Die Migration der T-Zellen im Gewebe gleicht einer zufälligen Bewegung, einem *random walk*. Die Pioniere der Zweiphotonenmikroskopie beschrieben daher die T-Zell-Zirkulation erstmals als einen stochastischen Suchprozess [WPMC03], was einen Paradigmenwechsel in der Immunologie auslöste.

In der Arbeit verfolgten wir den naheliegenden Ansatz, ein diskretes mathematisches Modell der T-Zell-Zirkulation aufzustellen und dieses Modell mit Techniken zu untersuchen, die man z.B. auch bei der Analyse randomisierter Suchheuristiken anwendet. Durch die kleine Zahl der spezifischen T-Zellen ist ein diskreter Ansatz sinnvoller als ein kontinuierliches Differentialgleichungsmodell. Beispielsweise kann in einer Maus die Zahl der spezifischen T-Zellen kleiner sein als die Zahl der Lymphknoten (etwa 35), so dass viele Lymphknoten zum Zeitpunkt einer Infektion keine spezifischen T-Zellen enthalten. Hieraus sieht man unmittelbar die Notwendigkeit der Zirkulation. Bei kontinuierlichen Modellen dagegen wird oft davon ausgegangen, dass die Population der spezifischen T-Zellen gleichmäßig auf alle Organe verteilt ist [SPN97], so dass der beschriebene Fall in

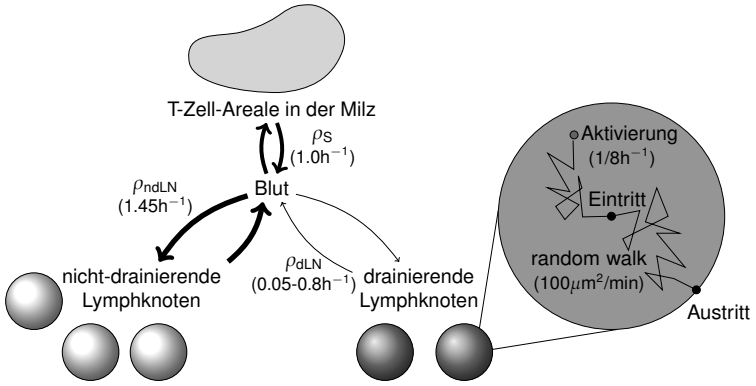


Abbildung 1: Zirkulation von T-Zellen zwischen Blut, Lymphknoten und Milz. Bei systemischen Infektionen wie *listeria*, die sich über das Blut verbreiten, wird die Immunantwort vorwiegend in der Milz gebildet, während bei lokalen Infektionen wie *influenza* oder *herpes simplex* die lokal drainierenden Lymphknoten – dies sind oft nur etwa 5% aller Lymphknoten – die wichtigste Rolle bei der Abwehr spielen. Vom Blut aus wandern T-Zellen nach dem Zufallsprinzip in die Milz oder einen Lymphknoten ein, wo sie mehrere Stunden nach Antigen suchen und dann wieder in das Blut zurückkehren. Innerhalb eines Lymphknotens bewegen sich T-Zellen im Wesentlichen zufällig [WPMC03] und werden aktiviert, wenn sie dabei auf Antigen stoßen.

solchen Modellen nicht berücksichtigt werden kann.

Der grundlegende Aufbau des vorgeschlagenen Modells wird in Abbildung 1 gezeigt. T-Zellen migrieren zwischen dem Blut, der Milz, sowie (bei lokalen Infektionen) drainierenden und nicht-drainierenden Lymphknoten hin und her. Die Verweilzeit der Zellen im Blut wird durch eine Exponentialverteilung modelliert, während für die lymphatischen Organe sowohl konstante Verweilzeiten als auch die sich durch einen *random walk* ergebende Verweilzeitverteilung untersucht wurden. Die Parameter des Modells werden anhand der Literatur festgelegt. Zur grundlegenden Validierung werden quantitative Vorhersagen generiert und mit experimentellen Daten abzugleichen. Einige Resultate dieses Vergleichs werden in Abbildung 2 (A,B) gezeigt. Das Modell nähert experimentell ermittelte Kinetiken der T-Zell-Rekrutierung für verschiedene Infektionstypen (*listeria*, *influenza*, und *herpes simplex*) gut an. Aus den Ergebnissen kann geschlossen werden, dass bei lokalen Infektionen mit wenigen drainierenden Lymphknoten wie *herpes simplex* ohne eine erhöhte Zufuhr von Lymphozyten zu drainierenden Lymphknoten eine effektive Immunabwehr nicht möglich wäre (Abbildung 2B).

Nach der quantitativen Validierung des Modells erfolgt eine tiefer gehende formale Analyse der grundlegenden Modelleigenschaften. Aus der Perspektive stochastischer Suche interessiert uns dabei besonders folgendes Optimierungsproblem: Wie lange sollte eine T-Zelle idealerweise in jedem Lymphknoten bleiben? Wird jeder Lymphknoten nur kurz durchsucht, besteht das Risiko, ein vorhandenes Antigen nicht zu finden. Wird umgekehrt zu viel Zeit auf die Durchsuchung verwendet, so könnte sich unterdessen anderswo eine Infektion ausbreiten, die dann mitunter erst zu spät erkannt wird. Dieses Problem ist

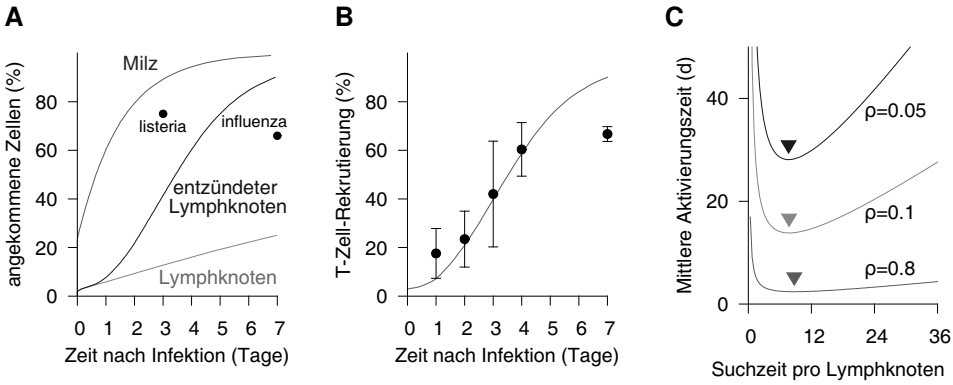


Abbildung 2: Quantitative Vorhersagen des Zirkulationsmodells im Vergleich mit experimentellen Daten in Mäusen. (A) Ankunftsgeschwindigkeit der zirkulierenden T-Zellen in der Milz (rote Linie), in einem Lymphknoten (grüne Linie) und in einem entzündeten, anschwellenden Lymphknoten (blaue Linie) verglichen mit T-Zell-Rekrutierungsdaten für *listeria* (Abwehr in der Milz) und *influenza* (Abwehr in Lymphknoten) [vHGS$^+$09]. (B) Verschwinden spezifischer T-Zellen aus der Zirkulation durch Aktivierung in den drainierenden Lymphknoten für *herpes simplex* (Punkte und Fehlerbalken: experimentelle Daten [SJHC11]; Linie: Modellvorhersage). (C) Die erwartete Zeit der Aktivierung einer T-Zelle als Funktion der Suchzeit pro Lymphknoten (Linien) und die sich daraus ergebende optimale Suchzeit (Dreiecke) für verschiedene Einflussraten in drainierende Lymphknoten, wie sie bei einer typischen Entzündung auftreten [SPS$^+$05], und einer Aktivierungsrate von 8 Stunden [MHvA04].

eng verwandt mit der Frage des *optimalen Restarts* bei Las-Vegas-Algorithmen mit bekannter Laufzeitverteilung. Für diese wurde bereits gezeigt, dass keine Strategie besser sein kann als das Neustarten des Algorithmus in festen Zeitintervallen [LSZ93]. Da auch das Erkennen von Antigen durch zirkulierende T-Zellen als Las-Vegas-Algorithmus aufgefasst werden kann, lässt sich diese Erkenntnis unmittelbar übertragen. In der Arbeit untersuchen wir darauf aufbauend die konkrete "Laufzeitverteilung", die sich durch das T-Zell-Zirkulationsszenario ergibt.

Diese "Laufzeitverteilung" wird durch folgende Parameter definiert: Sei ρ die Wahrscheinlichkeit, in einen drainierenden Lymphknoten einzuwandern; T die mittlere Zeit, die zwischen zwei Lymphknotenbesuchen im Blut, der Milz und anderswo verbracht wird; sowie α der Parameter einer Exponentialverteilung, die die Aktivierungszeit in einem drainierenden Lymphknoten beschreibt[1]. Sei H der Zeitpunkt der Aktivierung der T-Zelle in einem drainierenden Lymphknoten, gerechnet vom Beginn der Infektion. Welche Verweilzeit R pro Lymphknoten minimiert den Erwartungswert $E[H]$? Man kann nun zeigen, dass gilt:

$$R = \frac{1}{\ln(1-\alpha)} \left(W_{-1} \left(-\frac{(1-\alpha)^{\frac{T}{1-\rho}}}{e} \right) + 1 \right) - \frac{T}{1-\rho}$$

Hier bezeichnet W_{-1} den Nebenast der Lambertschen W-Funktion. Die durch diese Glei-

[1]Die Verwendung der Exponentialverteilung ergibt sich anhand wohlbekannter mathematischer Eigenschaften des dreidimensionalen random walks, und approximiert den random walk erstaunlich genau.

chung vorhergesagten Optima (Abbildung 2C) befinden sich in einem Bereich von etwa 9 Stunden, was zu experimentellen Daten über die tatsächliche Verweilzeit [SPN97] konsistent ist. Mittels asymptotischer Analysen lässt sich außerdem zeigen, dass das Optimum recht stabil gegenüber Parameterschwankungen ist, wie sie im biologischen Kontext zu erwarten sind. Als weiteres, recht überraschendes Ergebnis zeigt sich, dass der Parameter ρ im biologisch sinnvollen Bereich ($\rho \ll 1$) kaum Einfluss auf das Optimum hat. Die entzündungsvermittelte Erhöhung der Rekrutierungsrate und die Verweilzeit im Lymphknoten tragen also weitgehend unabhängig voneinander zur Effizienz der stochastischen T-Zell-Suche bei.

Zusammenfassend kann gesagt werden, dass das vorgeschlagene Modell trotz seiner Einfachheit in der Lage ist, experimentelle Daten aus verschiedensten Techniken wie Lymphdrainage, Histologie, Durchflusszytometrie und Zweiphotonenmikroskopie quantitativ in Verbindung zu bringen. Die Simulationen und die formale Analyse des stochastischen Modells lieferten quantitative, qualitative und funktionale Erkenntnisse, die durch experimentelle Ansätze nur schwer oder gar nicht hätten gewonnen werden können.

3 Negativauslese

Der zweite Teil der Arbeit befasst sich mit der Generierung und Negativauslese von T-Zellen. Diese Prozesse stellen sicher, dass immunokompetente T-Zellen normale Bestandteile des Organismus (Selbst) tolerieren und nur fremdartige Substanzen (Nichtselbst) angreifen. Im Wesentlichen wird dies erreicht, indem T-Zellen durch stochastische DNA-Rearrangements "zufällig generiert" werden, und dann normalen Proteinen aus dem Selbst ausgesetzt werden. Reagieren die neu erzeugten T-Zellen auf solche harmlosen Substanzen, werden sie wieder getötet. Die Arbeit greift ein etabliertes *algorithmisches* Modell [FPAC94] der Negativauslese auf. Der grundlegende Aufbau des Algorithmus wird in Abbildung 3 verdeutlicht: Als Grundlage definiert man eine Pattern-Matching-Funktion, die jedem Element einer Musterklasse Π eine Menge von Elementen eines Universums $\mathcal{U}$ zuordnet (z.b. $\Pi = \mathcal{U} = \mathbb{R}^2$ mit dem euklidischen Abstand als Matching-Funktion). Der Algorithmus generiert zunächst per Zufall eine Pattern-Menge P und gleicht diese dann mit der Eingabemenge S ab. Alle Pattern, die ein Element von S matchen, werden gelöscht. Im Idealfall beschreiben die dann übrigen Pattern ungefähr die dem Algorithmus unbekannte Bipartitionierung von $\mathcal{U}$, anhand derer die Beispiele generiert wurden.

Die Arbeit betrachtet zunächst grundlegende Eigenschaften dieses Algorithmus, die von der verwendeten Pattern-Matching-Funktion weitgehend unabhängig sind, aus der Perspektive der algorithmischen Lerntheorie. Dieser Ansatz zeigt Verbindungen zum konsistenten Lernen und dem Versionsraumlernen auf. Darauf aufbauend untersuchen wir verschiedene Mechanismen der Verallgemeinerung, die eine wesentliche Eigenschaft lernender Systeme ist. Verallgemeinerung kann zum Beispiel durch die Verwendung sogenannter *restringierter Pattern* erfolgen, bei denen ein zusätzlicher Parameter die Anzahl der gematchten Elemente pro Pattern kontrolliert (z.B. könnte man bei den Kreis-Pattern aus Abbildung 3 die Verallgemeinerung über den Radius einstellen). Weiterhin wird eine Verallgemeinerung des Negativauslese-Algorithmus vorgeschlagen, bei der nicht nur

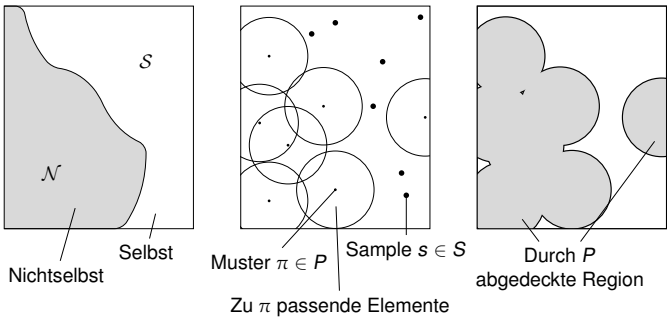


Abbildung 3: Illustration des Negativauslese-Algorithmus (siehe Text).

betrachtet wird *ob* ein zu klassifizierendes Element x durch ein Pattern aus P gematcht wird, sondern *wie viele* Pattern aus P das Element x matchen. Diesem Wert wird der Name *Sampling-Distanz* gegeben. Über einen Schwellwert für die Sampling-Distanz kann ebenfalls der Grad der Verallgemeinerung eingestellt werden.

In der Literatur wurden verschiedene konkrete Pattern-Matching-Funktionen zur Modellierung der Antigenerkennung durch T-Zellen vorgeschlagen. Diese basieren meist auf der Tatsache, dass T-Zellen nicht mit kompletten Proteinen interagieren, sondern nur mit kleinen Abbauprodukten von Proteinen (Peptiden). Abbildung 4 zeigt dieses Prinzip für die CD8-T-Zellen, die der Abwehr intrazellulärer Pathogene wie Viren dienen, und in deren Fall die Peptide typischerweise lineare Ketten von 9 Aminosäuren *(nonamere)* sind. Diese Beobachtung motivierte eine Klasse stringbasierter Modelle von T-Zell-Rezeptoren, bei denen T-Zellen als Strings und die Peptiderkennung als String-Matching formalisiert werden. Stringbasierte Modelle haben in der immunologischen Literatur bereits eine recht lange Tradition, und haben in jüngerer Zeit vielbeachtete qualitative Erkenntnisse geliefert [KRQ$^+$10]. Allerdings ist die Nutzbarkeit dieser Modelle durch das Problem beschränkt, dass eine direkte Implementierung des Negativauslese-Algorithmus meist exponentielle Laufzeit aufweist, da die Pattern-Menge P exponentielle Größe hat [Sti09].

Angesichts dieses Problems stellen wir die Frage, unter welchen Bedingungen die Negativauslese effizient durch einen Ein-Ausgabe-äquivalenten Algorithmus simuliert werden kann. Wir zeigen zunächst allgemein, wie die Komplexität dieser Aufgabe anhand formaler Entscheidungs- und Zählprobleme charakterisiert werden kann, und analysieren dann nach dieser Methodik mehrere konkrete Pattern-Matching-Funktionen. Für einige davon erhalten wir tatsächlich effiziente Simulationsalgorithmen. Dies gelingt vor allem durch Anwendung von effizienten Datenstrukturen wie Präfix- und Suffixbäumen (die allerdings teilweise technisch komplex ist). Für andere Pattern-Matching-Funktionen führen wir dagegen Reduktionsbeweise durch, die zeigen, dass eine effiziente Simulation unter der Annahme P$\neq$NP nicht möglich ist.

Glücklicherweise ist unser effizienter Simulationsansatz für einige wichtige stringbasierte Modelle durchführbar, unter anderem für das sogenannte r*-contiguous-Modell* [PPP93], für das bereits mehrere Jahre nach einem Polynomialzeitverfahren gesucht worden war

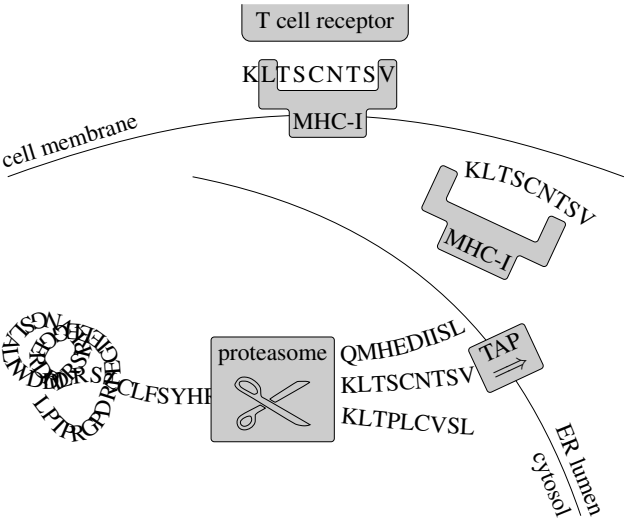


Abbildung 4: Schema der Antigenerkennung durch sogenannte CD8-T-Zellen (oder auch T-Killerzellen). Innerhalb jeder Zelle des menschlichen Organismus werden Stoffwechselprodukte kontinuierlich durch das Proteasom abgebaut und dabei in kleine Stücke aus wenigen Aminosäuren (Peptide) zerlegt. Diese Peptide werden durch das Molekül TAP in das endoplasmatische Retikulum transportiert, wo einige Peptide (meistens der Länge 9) an MHC-Klasse-I-Moleküle binden. Die Komplexe aus MHC und Peptid werden dann an die Zelloberfläche transportiert, wo sie den CD8-T-Zellen präsentiert werden. Durch das Prinzip der Negativauslese sind diese Zellen so generiert, dass sie normale Stoffwechselprodukte der Körpers nicht erkennen können, und daher im Normalfall nicht auf die präsentierten Peptide reagieren. Werden aber nun in der Zelle anormale Proteine synthetisiert, z.B. durch einen Virusbefall, so kann dies durch eine CD8-T-Zelle mit der richtigen Spezifität erkannt werden. Dies führt dann zum programmierten Zelltod der infizierten Zelle.

[Sti09]. Hierbei wird ein T-Zell-Rezeptor als Tupel aus einem Strings s der Länge ℓ und einer natürlichen Zahl $r \leq \ell$ modelliert, z.B. $(\texttt{KLTSCNTSV}, 3)$. Eine solches Pattern matcht alle Strings der Länge ℓ, die zu s in mindestens r aufeinanderfolgenden Positionen identisch sind. Durch unsere algorithmischen Verbesserungen ist es nun möglich, quantitative Vorhersagen aus diesem Modell auch für große Eingabedatensätze zu generieren.

Zum Abschluss der Arbeit wenden wir die neu entwickelten algorithmischen Verfahren zur Untersuchung der Immunogenität von HIV-Peptiden an. Wie bei den meisten Viren können bei HIV nicht alle viralen Peptide, die den T-Zellen gezeigt werden, auch durch diese erkannt werden. Konkret werden von HIV-infizierten Zellen etwa 90 verschiedene Peptide[2] präsentiert, wovon nur etwa die Hälfte von T-Zellen erkannt wird. Die andere Hälfte unterteilt sich wiederum in Peptide, die nur selten erkannt werden (kryptisch), solche, die manchmal erkannt werden (subdominant), und einige wenige, die meistens erkannt werden (dominant). Es ist bislang kaum verstanden, welche Eigenschaften der Peptide die Immunogenität bestimmen, und dies ist eines der wichtigsten ungelösten Pro-

[2]Die genaue Zahl hängt vom entsprechenden Mutanten des Virus ab.

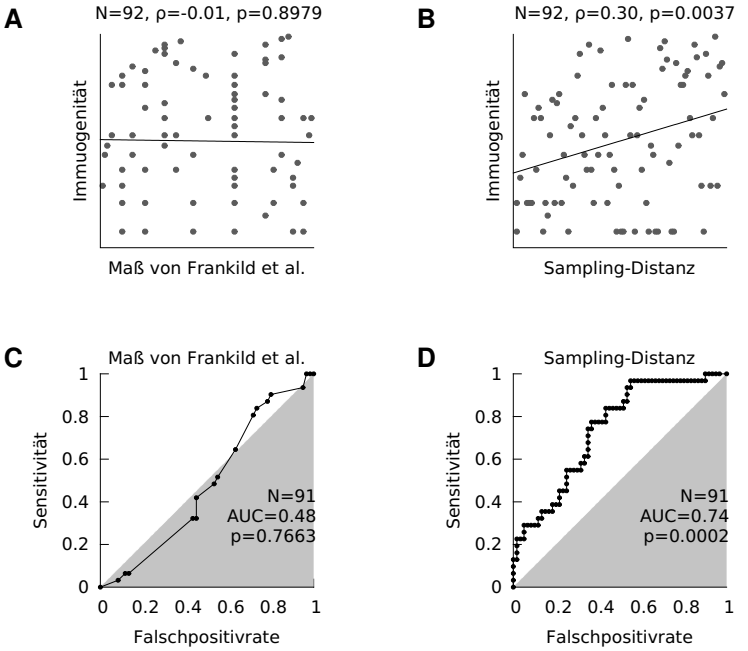


Abbildung 5: Quantitative Vorhersagen eines Modells aus der Literatur (A,C; [FdBL$^+$08]) und des Negativauslesemodells aus der Dissertation (B,D) im Vergleich mit experimentellen Daten zur HIV-Infektion. (A,B) Für 92 Peptide der Länge 9 wurde anhand veröffentlichter Daten von 120 HIV-Patienten [FKA$^+$04] die Erkennungshäufigkeit jedes Peptids bestimmt. Die Abbildungen zeigen die Korrelation (ρ) zwischen diesem Wert und den Vorhersagen des Literaturmodells (A) und des Negativauslesemodells (B). (C,D) Anhand der HIV-Datenbank des Los Alamos National Laboratory [YKB$^+$09] wurde für 91 Peptide der Länge 9 bestimmt, welche davon prinzipiell immunogen sind. Die Abbildung zeigen die sich ergebenden ROC-Kurven, wenn man die Immunogenität der 91 Peptide mit der Literaturmethode (C) und dem Negativauslesemodell (D) vorhersagt. Zum Vergleich: Zufälliges Raten (graue Fläche) entspricht einer diagonalen ROC-Kurve mit einer Fläche (AUC) von 0.5, während ein perfekter Klassifikator eine AUC von 1.0 erreicht. Die p-Werte geben die Wahrscheinlichkeit an, dass entsprechende Ergebnisse durch zufälliges Raten zu Stande kommen.

bleme der modernen Immunologie. In der Arbeit stellen wir die Hypothese auf, dass einige Peptide unter anderem deswegen nicht oder nur schwer erkennbar sein könnten, weil die meisten dafür spezifischen T-Zellen die Negativauslese nicht überleben.

Um diese Hypothese zu testen, simulieren wir die Generierung und Negativauslese eines Repertoires künstlicher T-Zellen nach dem r-contiguous-Modell (mit einigen biochemisch motivierten Erweiterungen). Als Eingabemenge werden alle 140,000 bekannten Nonamere des menschlichen Proteoms verwendet. Tatsächlich korrelieren die Vorhersagen des Modells signifikant mit der Immunogenität der untersuchten Peptide (Abbildung 5), und erklären diese sogar weitaus besser als das bislang beste bekannte Modell aus der Literatur [FdBL$^+$08]. Somit kann erstmals gezeigt werden, dass das r-contiguous-Modell trotz seiner Einfachheit tatsächlich quantitative Vorhersagekraft aufweist. Gleichzeitig wird eine neuartige Erklärung für die Immunogenität von HIV-Peptiden gefunden.

4 Ausblick

Wie zu Beginn betont wurde, lag der Hauptfokus der Arbeit auf der Generierung neuer Erkenntnisse für die Immunologie mit Methoden der (vor allem theoretischen) Informatik. Jedoch zeigt die Arbeit auch Anknüpfungspunkte zu aktuellen Fragestellungen der Informatik selbst auf. Diese ergeben sich vor allem zu bioinspirierten randomisierten Suchheuristiken (im ersten Teil) sowie zu sogenannten „künstlichen Immunsystemen" für die Computersicherheit (zweiter Teil). Besonders hervorzuheben ist, dass der Algorithmus zur Negativauslese ursprünglich als Grundlage für Intrusion-Detection-Systeme konzipiert wurde [FPAC94]. Nach anfangs vielversprechenden Resultaten wurde die Forschung in diese Richtung nach etwa 10 Jahren weitgehend aufgegeben, was vor allem an der unzureichenden Effizienz des Algorithmus lag [Sti09]. Da die Arbeit dieses Effizienzproblem löst, erscheint weitere Forschung in diese Richtung nun wieder sinnvoll. Von allgemeinem Interesse dürfte auch die Grundlage für eine lerntheoretische Betrachtung des Immunsystems sein, die im zweiten Teil geschaffen wird. Das Immunsystem ist neben dem zentralen Nervensystem eines der beiden wichtigen kognitiven Systeme des menschlichen Körpers, wurde aber bisher kaum lerntheoretisch untersucht. Eine Vertiefung dieser Forschung birgt das Potential, gemeinsame Mechanismen des Lernens und der Gedächtnisbildung in beiden Systemen zu identifizieren.

Literatur

[FdBL+08] Sune Frankild, Rob J. de Boer, Ole Lund, Morten Nielsen und Can Keşmir. Amino Acid Similarity Accounts for T Cell Cross-Reactivity and for "Holes" in the T Cell Repertoire. *PLoS one*, 3(3):e1831, 2008.

[FKA+04] N. Frahm, B. T. Korber, C. M. Adams et al. Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities. *Journal of Virology*, 78:2187–2200, 2004.

[FPAC94] Stephanie Forrest, Alan S. Perelson, Lawrence Allen und Rajesh Cherukuri. Self-Nonself Discrimination in a Computer. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Seiten 202–212. IEEE Computer Society Press, 1994.

[KRQ+10] Andrej Košmrlj, Elizabeth L. Read, Ying Qi et al. Effects of thymic selection of the T-cell repertoire on HLA class I–associated control of HIV infection. *Nature*, 465:350–354, 2010.

[LSZ93] Michael Luby, Alistair Sinclair und David Zuckerman. Optimal speedup of Las Vegas algorithms. *Information Processing Letters*, 47:173–180, 1993.

[MCP+07] J. J. Moon, H. H. Chu, M. Pepper et al. Naive CD4+ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity*, 27:203–213, 2007.

[MHvA04] Thorsten R. Mempel, Sarah E. Henrickson und Ulrich H. von Andrian. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature*, 427:154–159, 2004.

[MWPC02] M. J. Miller, S. H. Wei, I. Parker und M. D. Cahalan. Two-photon imaging of lymphocyte motility and antigen response in intact lymph node. *Science*, 296:1869–1873, 2002.

[PPP93] Jerome K. Percus, Ora E. Percus und Alan S. Perelson. Predicting the size of the T-cell receptor and antibody combining region from consideration of efficient self-nonself discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, 90(5):1691–1695, 1993.

[SJHC11] A. T. Stock, C. M. Jones, W. R. Heath und F. R. Carbone. Rapid recruitment and activation of CD8+ T cells after herpes simplex virus type 1 skin infection. *Immunology and Cell Biology*, 89:143–148, 2011.

[SPN97] Dov J. Stekel, Claire E. Parker und Martin A. Nowak. A model of lymphocyte recirculation. *Immunology Today*, 18(5):217–221, 1997.

[SPS+05] Kelly A. Soderberg, Geoffrey W. Payne, Ayuko Sato et al. Innate control of adaptive immunity via remodeling of lymph node feed arteriole. *Proceedings of the National Academy of Sciences of the USA*, 102(45):16315–16320, 2005.

[Sti09] Thomas Stibor. Foundations of R-Contiguous Matching in Negative Selection for Anomaly Detection. *Natural Computing*, 8:613–641, 2009.

[vHGS+09] J. W. van Heijst, C. Gerlach, E. Swart et al. Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient. *Science*, 325:1265–1269, 2009.

[WPMC03] Sindy H. Wei, Ian Parker, Mark J. Miller und Michael D. Cahalan. A stochastic view of lymphocyte motility and trafficking within the lymph node. *Immunological Reviews*, 195:136–159, 2003.

[YKB+09] Karina Yusim, Bette T. M. Korber, Christian Brander et al. HIV Molecular Immunology 2009. Bericht LA-UR 09-05941, Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico, 2009.

Johannes Textor, geboren 1979 in Göttingen, studierte von 1999 bis 2006 Informatik an der Universität zu Lübeck und promovierte dort von 2006 bis 2011 am Institut für Theoretische Informatik. Seit September 2011 ist er als Postdoktorand im Institut für Theoretische Biologie und Bioinformatik der Universität Utrecht in den Niederlanden tätig. Dort forscht er in verschiedenen Kooperationen mit dem niederländischen Krebsforschungsinstitut, der Harvard Medical School, den National Institutes of Health, und der Washington University St. Louis weiter schwerpunktmäßig zur Migration von T-Zellen.

Neue Ansätze zum computergestützten Entwurf epitopbasierter Impfstoffe

Nora C. Toussaint

Immunology Program & Computational Biology Program
Lucille Castori Center for Microbes, Inflammation & Cancer
Sloan-Kettering Institute
Memorial Sloan-Kettering Cancer Center
1275 York Ave, New York, NY 10065, U.S.A.
toussain@mskcc.org

Abstract: Trotz zahlreicher Erfolge des traditionellen Impfstoffentwurfs, gibt es immer noch Krankheiten, gegen die bisher kein geeigneter Impfstoff entwickelt werden konnte. Die wohl bekanntesten Beispiele sind HIV-Infektion und Krebs. Hier sind neue, rational entworfene Impfstoffe, wie zum Beispiel epitopbasierte Impfstoffe, eine vielversprechende Alternative. Der gezielte Einsatz computergestützter Methoden im epitopbasierten Impfstoffentwurf verspricht darüber hinaus verbesserte Impfstoffe bei kürzerer Entwicklungszeit und geringeren Kosten.

Im Rahmen der hier zusammengefassten Dissertation wurden relevante Probleme des epitopbasierten Impfstoffentwurfs formalisiert und mit Hilfe von Methoden der kombinatorischen Optimierung und des maschinellen Lernens gelöst. Die Anwendung der vorgestellten Methoden in realistischen Impfstoffentwurfsstudien lieferte vielversprechende Ergebnisse, die das große Potenzial des computergestützten rationalen Impfstoffentwurfs verdeutlichen.

1 Einführung

Die Entwicklung von Impfstoffen gehört zu den bedeutendsten Fortschritten in der Geschichte der modernen Medizin. Die Grundidee von Impfungen ist es, Immunität zu generieren, ohne die eigentliche Krankheit hervorzurufen. Dazu machen sich Impfstoffe das Gedächtnis des adaptiven Teils des Immunsystems zu Nutze. Ziel der Immunoinformatik ist es, über die Modellierung des Immunsystems zu einem besseren Verständnis immunologischer Prozesse beizutragen. Mit Hilfe dieser Modelle soll eine gezielte Veränderung des immunologischen Gedächtnisses und damit eine optimale Behandlung möglich werden.

Das adaptive Immunsystem wird aktiv, wenn es ein immunogenes Proteinfragment erkennt. Im Inneren einer Wirtszelle werden die Proteine eines Pathogens – ebenso wie alle anderen Proteine – in kleine Proteinfragmente, die *Peptide*, zerlegt. Einige dieser Peptide binden an Moleküle des Haupthistokompatibilitätskomplexes (MHC - *major histocompatibility complex*), die sie an der Zelloberfläche präsentieren. Erkennt eine T-Zelle einen

solchen MHC-Peptid-Komplex wird eine Immunantwort hervorgerufen (Abbildung 1). Peptide, die eine Immunantwort hervorrufen können, werden *immunogene Peptide* oder auch *Epitope* genannt. Die Proteine, aus denen sie hervorgegangen sind, heißen *Antigene*.

Epitope stellen die kleinsten immunogenen Teile eines Antigens dar. Die Verwendung von Epitopen als Komponenten eines Impfstoffs bietet viele Vorteile. Epitopbasierte Impfstoffe (EBIs) können gezielt hochimmunogene Regionen von targetspezifischen Antigenen ins Visier nehmen. Darüber hinaus können EBIs auf die MHC-Ausprägungen einer spezifischen Person zugeschnitten werden und sind somit im Bereich der personalisierten Medizin einsetzbar. EBIs haben noch viele weitere gute Eigenschaften, die sie für die Pharmaindustrie besonders interessant machen. Es ist daher kaum verwunderlich, dass sie in den letzten Jahren sehr viel Aufmerksamkeit erweckt haben.

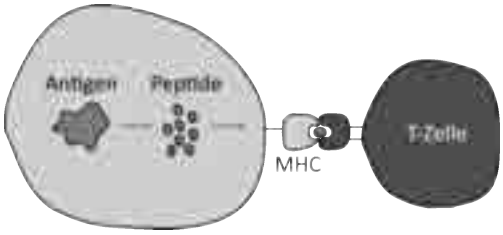


Abbildung 1: Im Inneren der Wirtszelle wird ein Antigen in kleinere Stücke, die Peptide, zerlegt. MHC-Moleküle binden solche Peptide und präsentieren sie an der Zelloberfläche. Wird ein MHC-Peptid-Komplex von einer T-Zelle erkannt, wird eine Immunantwort hervorgerufen.

Der Entwurf eines EBIs kann grob in drei Schritte unterteilt werden: Epitopbestimmung, Epitopselektion und Epitopassemblierung (Abbildung 2). In jedem dieser Schritte können computergestützte Methoden eingesetzt werden, um die Arbeit von Immunologen zu vereinfachen und sie in ihren Entscheidungen zu unterstützen. Die Anwendung solcher *in-silico*-Methoden im EBI-Entwurf verspricht verbesserte Impfstoffe bei einer kürzeren Entwicklungszeit und geringeren Kosten.

Das noch recht junge interdisziplinäre Feld der Immunoinformatik ist geprägt von proprietären Daten – nicht zuletzt aufgrund der Bedeutung immunologischer Daten für die pharmazeutische Industrie – und von einer Grundskepsis der Immunologen gegenüber Versuchen, das Immunsystem durch Gleichungen zu beschreiben. Die hier zusammengefasste Dissertation [Tou11] versucht zur Etablierung informatischer Methoden in der

Abbildung 2: Epitopbasierter Impfstoffentwurf. Ausgehend von einer Menge von targetspezifischen Antigenen werden Kandidatenepitope bezüglich einer Zielpopulation bestimmt (*Epitopbestimmung*). Aus der resultierenden Menge von Kandidatenepitopen muss die für die Verwendung in einem EBI am besten geeignete Teilmenge ausgewählt werden (*Epitopselektion*). Die ausgewählten Epitope werden zum EBI zusammengesetzt (*Epitopassemblierung*). (Die Abbildung basiert auf [TK09b].)

Immunologie beizutragen, indem sie immunologische Probleme formalisiert und mit theoretisch fundierten Methoden reproduzierbar und, soweit möglich, optimal löst. Um neue und verbesserte *in-silico*-Methoden zur Epitopbestimmung zu ermöglichen, wurden Methoden des maschinellen Lernens verwendet und weiterentwickelt. Darüber hinaus wurden die Probleme der Epitopselektion und -assemblierung formalisiert und Methoden der kombinatorischen Optimierung verwendet, um diese Probleme erstmalig optimal zu lösen.

2 Epitopbestimmung

Ausgehend von einer Menge von Antigenen werden im ersten Schritt des EBI-Entwurfs Kandidatenepitope bestimmt und experimentell validiert. Kandidatenepitope sind diejenigen targetspezifischen Peptide, die eine adaptive Immunantwort in der Zielpopulation hervorrufen können. Die Einbindung von *in-silico*-Methoden in diesen Schritt kann die Anzahl der durchzuführenden Experimente drastisch reduzieren.

Aufgrund der komplexen Abhängigkeit der T-Zell-Reaktivität vom Immunsystem des Wirts sowie des unvollständigen Verständnisses der zugrundeliegenden Prozesse ist die Vorhersage von T-Zell-Epitopen ein äußerst schwieriges Problem. Da T-Zellen immunogene Peptide nur erkennen können, wenn diese von MHC-Molekülen präsentiert werden, ist die Bindung von Peptiden an MHC-Moleküle notwendig – allerdings nicht hinreichend – für die Auslösung einer Immunantwort. Die Prozesse, die für die Bindung eines Peptids an ein MHC-Molekül zuständig sind, sind gut verstanden. Da gezeigt wurde, dass das Potenzial eines Peptids, eine T-Zell-basierte Immunantwort hervorzurufen, also die *Immunogenität* eines Peptids, gut mit der MHC-Bindeaffinität des Peptids korreliert, wird das Epitopbestimmungsproblem üblicherweise auf das Problem der MHC-Bindevorhersage reduziert.

In der hier zusammengefassten Dissertation werden drei Ansätze zur Epitopvorhersage vorgestellt: zwei Ansätze zur MHC-Bindevorhersage und ein Ansatz zur Verbesserung der Vorhersage von Immunogenität.

2.1 MHC-Bindevorhersage

Das Hauptproblem bei der Vorhersage von Peptiden, die an ein bestimmtes MHC-Molekül binden, ist der Mangel an (frei) verfügbaren experimentellen Bindedaten. Mehrere tausend genetische MHC-Varianten, die sogenannten *Allele*, sind bekannt. Unterschiedliche MHC-Varianten binden unterschiedliche Peptidrepertoires und unterschiedliche Individuen tragen unterschiedliche MHC-Varianten. Daraus folgt, dass in jedem Menschen unterschiedliche Peptide von MHC-Molekülen präsentiert werden. Da T-Zellen Epitope nur im Komplex mit MHC-Molekülen erkennen, kann ein Peptid, das in einem Individuum eine T-Zell-Antwort hervorruft, in einem anderen Individuum vom Immunsystem unbeachtet bleiben. Darüber hinaus sind bestimmte MHC-Allele in einer Population häufiger als andere und die Verteilung von MHC-Allelen variiert zwischen unterschiedlichen Populationen. Die Berücksichtigung der MHC-Verteilung in der Zielpopulation ist für den Entwurf eines

effektiven EBIs daher unumgänglich.

Klassische Ansätze zur MHC-Bindevorhersage basieren auf allelspezifischen Modellen und benötigen eine gewisse Menge an experimentellen Bindedaten für das jeweilige Allel [PBC94, BLW+03]. Für einen Großteil der bekannten MHC-Allele sind jedoch nicht genügend Bindedaten verfügbar. Im Rahmen der hier zusammengefassten Dissertation werden zwei Ansätze zur Überwindung dieses Problems vorgestellt.

Neue Kernfunktionen

Ein Ansatz zur Umgehung des Problems der mangelnden Bindedaten für ein bestimmtes MHC-Allel ist die Einbindung zusätzlichen biologischen Wissens. MHC-Moleküle binden Peptide in gestreckter Form (Abbildung 3). Innerhalb des Komplexes interagieren die Peptidbestandteile, d.h. die Aminosäuren, sowohl untereinander als auch mit den angrenzenden Aminosäuren des MHC-Moleküls. Jede der Aminosäuren des Peptids trägt somit zur Bindeaffinität zwischen Peptid und MHC-Molekül bei. Der jeweilige Beitrag hängt von den physikochemischen Eigenschaften der Aminosäure (z.B. Ladung, Größe, Hydrophobizität), von den physikochemischen Eigenschaften der Aminosäuren in der näheren Umgebung und auch von der Position der Aminosäure innerhalb des Peptids ab.

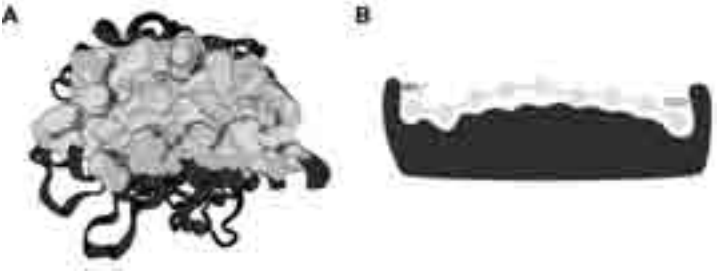


Abbildung 3: MHC-Peptid-Bindung. A) Kristallstruktur eines MHC-Peptid-Komplex. B) Schematische Darstellung der gestreckten Bindung eines Peptids (grün) an ein MHC-Molekül (rot).

Das Wissen über physikochemische Eigenschaften von Aminosäuren wurde erfolgreich verwendet, um die MHC-Bindevorhersage zu verbessern (z.B. [NLW+03]). Die Anordnung der Aminosäuren im Peptid wurde jedoch nicht berücksichtigt. Die hier zusammengefasste Dissertation stellt einen Ansatz basierend auf Supportvektormaschinen (SVMs) [Vap95] vor, der erstmalig das Wissen über physikochemische Aminosäureeigenschaften mit Informationen über die Anordnung der Aminosäuren im Peptid kombiniert [TWKR10, WTA+10].

Ausgehend von einer Menge von N Trainingsbeispielen – in unserem Fall Peptide $\mathbf{x_i}$ mit ihrem Label $y_i \in \{\pm 1\}$ – lernen SVMs eine Entscheidungsfunktion, mit der Vorhersagen über ungesehene Peptide getroffen werden können:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x_i}, \mathbf{x}) + b\right).$$

Die Trainingsbeispielgewichte $\{\alpha_i\}_{i=1}^N$ und der Bias b werden durch Lösen eines quadratischen Optimierungsproblems bestimmt. Die Funktion $k(\mathbf{x}, \mathbf{x}')$ wird als *Kern* bezeichnet. Sie stellt ein Ähnlichkeitsmaß zwischen zwei Peptidsequenzen dar und hat einen wesentlichen Einfluss auf die Vorhersagequalität der SVM.
Ein Kern, der sich besonders gut eignet, die gestreckte Anordnung der Peptide im MHC-Peptid-Komplex zu berücksichtigen, ist der sogenannte *Weighted-Degree*-Kern (WD-Kern) [RS04]. Der WD-Kern vergleicht zwei Sequenzen gleicher Länge L basierend auf den Substrings, aus denen sie sich zusammensetzen. Der WD-Kern von Grad d ist folgendermaßen definiert:

$$k_d^{wd}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=1}^{d} \beta_\ell \sum_{i=1}^{L-\ell-1} \mathbf{I}\big(\mathbf{x}_{[i:i+\ell]}, \mathbf{x'}_{[i:i+\ell]}\big),$$

wobei I die Identitätsfunktion ist, $\mathbf{x}_{[i:i+\ell]}$ der Substring der Länge ℓ an Position i und β_ℓ ein Gewicht, das Übereinstimmungen in längeren Substrings niedriger gewichtet. Ein wesentlicher Nachteil des WD-Kerns bei der Vorhersage von MHC-Bindung ist jedoch, dass für ihn alle Aminosäuren gleich ähnlich sind. Eine Berücksichtigung physikochemischer Eigenschaften von Aminosäuren ist nicht ohne Weiteres möglich.

In der Dissertation wird eine Gruppe von Kernfunktionen vorgestellt, die die Vorteile des WD-Kerns mit denen von physikochemischen Deskriptoren für Aminosäuren verbindet. Ausführliche empirische Untersuchungen im Rahmen der Dissertation haben gezeigt, dass sich der *WD-RBF-Kern* besonders gut für die MHC-Bindevorhersage eignet. Dieser Kern ist wie folgt definiert

$$k_{d,\sigma}^{wd,\Psi}(\mathbf{x}, \mathbf{x}') \quad = \quad \sum_{\ell=1}^{d} \beta_\ell \sum_{i=1}^{L-\ell+1} \exp\left(-\frac{\sum_{j=1}^{\ell} \|\Psi(\mathbf{x}_j) - \Psi(\mathbf{x}'_j)\|^2}{2\sigma^2} \right),$$

wobei $\mathbf{x}_j$ die Aminosäure an Position j im Peptid $\mathbf{x}$ ist. Die Abbildung $\Psi : \Sigma \to \mathbb{R}^m$ bildet eine Aminosäure $x \in \Sigma$ auf m physikochemische Deskriptoren ab. Effiziente Implementierungen der in der Dissertation vorgestellten Kerne sind als Teil der Open-Source-Machine-Learning-Toolbox Shogun (http://www.shogun-toolbox.org) [SRH+10] öffentlich verfügbar gemacht worden.

Ein statistischer Vergleich mit dem WD-Kern auf 35 unterschiedlichen MHC-Allelen zeigt, dass die Verwendung des WD-RBF-Kerns zu einer Verbesserung der Vorhersagequalität führt: der WD-RBF-Kern liefert in 23 Fällen bessere Ergebnisse als der WD-Kern, der nur in acht Fällen die besseren Vorhersageergebnisse liefert. Die Verbesserung durch den WD-RBF-Kern ist besonders deutlich, wenn nur wenig Trainingsdaten zur Verfügung stehen.

Neue Ansätze des Transferlernens

Der zweite in der Dissertation vorgestellte Ansatz zur MHC-Bindevorhersage kommt aus dem Bereich des sogenannten Transferlernens und nutzt zur Umgehung des Problems der mangelnden Daten strukturelle Ähnlichkeiten zwischen unterschiedlichen MHC-Varianten aus. Diese strukturellen Ähnlichkeiten erlauben es, experimentelle Bindedaten einer MHC-Variante für die Vorhersage der Bindeeigenschaften einer anderen zu verwenden.

Der Unterschied zwischen zwei MHC-Varianten zeigt sich in der Bindetasche, also in dem Bereich, in dem das MHC-Molekül das Peptid bindet. Ausgehend von einer Menge von 3D-Strukturen von MHC-Peptid-Komplexen konnten wir feststellen, welche Positionen in der Sequenz eines MHC-Moleküls mit dem gebundenen Peptid interagieren. Die Aminosäuren an den jeweiligen Positionen bilden das Profil der Bindetasche – eine Art Fingerabdruck, anhand dessen MHC-Varianten unterschieden werden können. Je ähnlicher das Profil zweier MHC-Varianten, umso ähnlicher sind auch die Peptide, die sie binden.

Während klassische Ansätze zur MHC-Bindevorhersage ein Modell pro MHC-Variante trainieren, trainieren wir ein einziges SVM-Modell für alle MHC-Varianten gemeinsam. Als Eingabe dient ein MHC-Peptid-Paar, repräsentiert durch einen Vektor, der sich aus physikochemischen Deskriptoren für das Profil der Bindetasche der MHC-Variante und für die Aminosäuren des Peptids zusammensetzt. Dieser Ansatz ermöglichte erstmalig Bindevorhersagen für alle bekannten MHC-Varianten, unabhängig von der Menge der für das jeweilige Allel zur Verfügung stehenden Bindedaten. Die Vorhersagegenauigkeit der Methode ist sowohl für Allele mit Bindedaten als auch für Allele ohne Bindedaten vergleichbar oder sogar besser als die von allelspezifischen Methoden.

2.2 Immunogenitätsvorhersage

Die Bindung eines Peptids an ein MHC-Molekül ist lediglich eine notwendige Voraussetzung für die Induktion einer T-Zell-Antwort, d.h. für die Immunogenität eines Peptids. Die Vorhersage von MHC-bindenden Peptiden löst das Epitopbestimmungsproblem im EBI-Entwurf demnach nicht vollständig.

Das Hauptproblem bei der Vorhersage von immunogenen Peptiden ist die komplexe Abhängigkeit der T-Zell-Reaktivität vom Immunsystem des Patienten. Bereits existierende Methoden zur Vorhersage von Immunogenität [BR04, TH07] berücksichtigen diese Abhängigkeit nicht, sondern beziehen lediglich die Aminosäuresequenz des Peptids in die Vorhersage mit ein. Diese sehr starke Vereinfachung des Problems spiegelt sich in einer geringen Vorhersagegenauigkeit wieder.

In der Dissertation wird ein Ansatz vorgestellt, der erstmalig zusätzlich zur Peptidsequenz Wissen über eine relevante Eigenschaft des Immunsystems einbezieht: das adaptive Immunsystem ist *selbsttolerant*, d.h. körpereigene Proteine und Peptide rufen im Allgemeinen keine Immunantwort hervor. Daraus lässt sich folgern, dass körperfremde Peptide, die eine hohe Ähnlichkeit zu Selbstpeptiden haben, wahrscheinlich keine T-Zell-gesteuerte Immunantwort hervorrufen.

Unter Verwendung einer effizienten Trie-Struktur und einer geeigneten Ähnlichkeitsfunktion ermittelt unser Ansatz für jedes betrachtet Peptid die Ähnlichkeit zu den Peptiden des menschlichen Proteoms. Einbeziehung dieser Information in die Immunogenitätsvorhersage führt zu einer deutlichen Verbesserung der Vorhersagegenauigkeit [TFZ$^+$11].

3 Epitopselektion

Nachdem im ersten Schritt des EBI-Entwurfs eine Menge von Kandidatenepitopen bestimmt wurde, wird im zweiten Schritt aus dieser Menge die Teilmenge ausgewählt, die die beste Immunantwort in der Zielpopulation verspricht. Aufgrund von regulatorischen, ökonomischen und praktischen Überlegungen kann nur eine kleine Menge der im vorhergehenden Schritt bestimmten Kandidatenepitope in den EBI einbezogen werden. Da der Erfolg des Impfstoffs von den ausgewählten Epitopen abhängt, ist es äußerst wichtig, die optimale Kombination von Peptiden zu identifizieren, also die Menge von Peptiden, die die bestmögliche Immunantwort in der Zielpopulation hervorruft.

Weder die übliche manuelle Selektion durch Experten noch die bisher publizierten computergestützten Methoden [GMB+05, VSRL07] können garantieren, dass die getroffene Auswahl und somit der resultierende Impfstoff optimal sind. Wir stellen einen auf ganzzahliger linearer Programmierung (ILP - *integer linear programming*) basierenden Ansatz vor, der eine beweisbar optimale Lösung liefert. Unser Ansatz erlaubt eine elegante und flexible Formulierung unterschiedlicher Anforderungen an die auszuwählenden Epitope. Für typische Problemgrößen beträgt die Laufzeit nur wenige Sekunden [TDK08].

Ein guter Impfstoff zeigt eine hohe Gesamtimmunogenität, d.h. er hat das Potenzial starke Immunität in einem Großteil der Zielpopulation hervorzurufen. Die Zielfunktion maximiert also die Gesamtimmunogenität der ausgewählten Epitope. Zusätzlich werden je nach Fall unterschiedliche Anforderungen an die auszuwählenden Epitope gestellt. Diese Anforderungen werden im ILP als Bedingungen formuliert. Ein Beispiel ist in ILP 1 gegeben. Das ILP wählt k Epitope aus (1a) – aus jedem der n Antigene mindestens eins (1b) – die insgesamt im Kontext von mindestens m MHC-Varianten immunogen sind (1c, 1d).

ILP 1 Ausschnitt aus einem ILP zur Epitopselektion

$$
\begin{array}{lll}
\text{Maximiere} & \sum_{e\in E} \mathsf{x_e} \sum_{a\in A} p(a)\, i(e,a)\,, & \\
\text{so dass} & \sum_{e\in E} \mathsf{x_e} = k & \text{(1a)}\\
\forall i \in \{1,\ldots,n\}: & \sum_{e\in E_i \cap I} \mathsf{x_e} \geq 1 & \text{(1b)}\\
\forall a \in A: & \sum_{e\in I_a} \mathsf{x_e} \geq \mathsf{y_a} & \text{(1c)}\\
 & \sum_{a\in A} \mathsf{y_a} \geq m & \text{(1d)}
\end{array}
$$

DEFINITIONEN

A	Menge der MHC-Allele der Zielpopulation
E_i	Menge der Kandidatenepitope von Antigen i
E	Menge aller Kandidatenepitope ($E = E_1 \cup \ldots \cup E_n$)
I_a	Menge von Epitopen, die im Kontext von MHC-Variante a immunogen sind
I	Menge aller bezüglich A immunogenen Epitope ($I = \bigcup_{a\in A} I_a$)

PARAMETER

$i(e,a)$	Immunogenität von Epitop e im Kontext von MHC-Variante a
k	Anzahl der auszuwählenden Epitope
$p(a)$	Wahrscheinlichkeit, dass die MHC-Variante a in der Zielpopulation vorkommt
m	Mindestanzahl von MHC-Varianten, die berücksichtigt werden sollen

VARIABLEN

$\mathsf{x_e} = 1$	falls Epitop e zur optimalen Menge gehört, sonst $\mathsf{x_e} = 0$
$\mathsf{y_a} = 1$	falls MHC-Variante a von der Epitopmenge berücksichtigt wird, sonst $\mathsf{y_a} = 0$

Die Formalisierung des Epitopselektionsproblems und die Formulierung als ILP führen zu deutlich besseren Ergebnissen und damit – zumindest theoretisch – zu deutlich besseren Impfstoffen als bereits existierende Lösungsansätze (Abbildung 4). Um diese Methoden Immunologen zugänglich zu machen, wurde ein öffentlich verfügbarer Webservice entwickelt (http://www.epitoolkit.org/optitope) [TK09a].

4 Epitopassemblierung

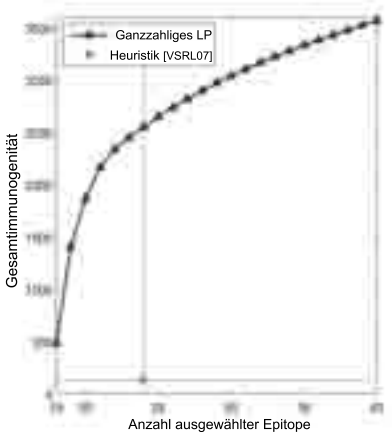


Abbildung 4: Epitopselektion. Vergleich des hier vorgestellten optimalen Ansatz mit der in [VSRL07] präsentierten Heuristik.

Der letzte Schritt des EBI-Entwurfs befasst sich mit der Verabreichung des Impfstoffs. Dazu werden die Peptide üblicherweise zu einem einzelnen langen Polypeptid zusammengefügt. Da eine ungünstig gewählte Epitopanordnung zum Abbau der gewünschten Epitope im Wirt führen kann, ist die optimale Anordnung von wesentlicher Bedeutung für den Erfolg des Impfstoffs. Wir präsentieren den ersten computergestützten Ansatz zur Lösung dieses komplexen Problems. Unsere graphentheoretische Formulierung ermöglicht es, für realistische Problemgrößen die optimale Anordnung der Epitope in vertretbarer Zeit zu bestimmen.

Sei $G = (V, E, w)$ ein vollständiger, gerichteter und gewichteter Graph mit Knoten V, Kanten E und Gewichten w. Jeder Knoten repräsentiert ein Epitop und jede Kante eine mögliche Konkatenation des jeweiligen Epitoppaares. Das Gewicht w_{ab} der Kante (a, b) entspricht dem Logarithmus der Wahrscheinlichkeit, dass die Epitope a und b im Konstrukt $a - b$ abgebaut werden. In dieser Formulierung entspricht die optimale Epitopanordnung dem minimalen Hamiltonpfad, d.h. dem Pfad, der jeden Knoten genau einmal besucht und minimales Gewicht hat. Durch Hinzufügen eines weiteren Knotens, der mit allen anderen Knoten über jeweils eine Hin- und eine Rückkante verbunden wird, wird das Problem in die Suche nach dem minimalen Hamiltonkreis umgewandelt. Die Bestimmung des minimalen Hamiltonkreises entspricht dem wohlbekannten und ausgiebig studierten *Traveling Salesman Problem* (TSP). Verwendung einer bereits publizierten ILP-Formulierung des TSP [Pat03] erlaubt es, das Epitopassemblierungsproblem für realistische Epitopzahlen in vertretbarer Zeit optimal zu lösen.

In einer theoretischen Impfstoffentwurfsstudie wurde gezeigt, dass bei einer optimierten Anordnung deutlich mehr der ausgewählten Epitope von MHC-Molekülen an der Zelloberfläche präsentiert werden können als bei einer zufälligen Anordnung [TMKL11].

5 Zusammenfassung und Ausblick

EBIs haben das Potenzial, eine wichtige Rolle bei der Bekämpfung von Krebs aber auch von neu aufkommenden Infektionskrankheiten wie zum Beispiel der Schweinegrippe zu spielen. Beim EBI-Entwurf müssen zahlreiche Probleme gelöst und Entscheidungen getroffen werden. Hier kann die Informatik einen wesentlichen Beitrag leisten.

Die hier zusammengefasste Dissertation [Tou11] befasst sich mit der Formalisierung wesentlicher Probleme des EBI-Entwurfs und der Entwicklung theoretisch fundierter Methoden, um diese Probleme reproduzierbar und optimal zu lösen. Die hier vorgestellten Methoden zur Epitopbestimmung ermöglichen deutlich verbesserte Bindevorhersagen für MHC-Allele mit wenigen oder keinen Bindedaten, sowie eine verbesserte Immunogenitätsvorhersage. Die Verwendung von ILP-basierten Ansätzen zur Lösung des Epitopselektions- und des Epitopassemblierungsproblems liefert erstmalig garantiert optimale EBI-Konstrukte.

Langfristig wird die Verwendung von *in-silico*-Methoden zur Epitopbestimmung, -selektion und -assemblierung den Impfstoffentwurfsprozess drastisch verändern. Die Verwendung standardisierter Ansätze für den Entwurf von EBIs verspricht insbesondere kürzere Entwicklungszeiten, was essenziell sein kann, wenn es zu neu aufkommenden Infektionskrankheiten kommt. Darüber hinaus ist eine schnelle und auch kostengünstige Impfstoffentwicklung unverzichtbar für vollständig personalisierte Impfstoffe, insbesondere für die Immuntherapie von Krebs.

Literatur

[BLW+03] S. Buus, S. L. Lauemøller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm und S. Brunak. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, 62(5):378–384, 2003.

[BR04] M. Bhasin und G. P. S. Raghava. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine*, 22(23-24):3195–3204, 2004.

[GMB+05] A. S. De Groot, L. Marcon, E. A. Bishop, D. Rivera, M. Kutzler, D. B. Weiner und W. Martin. HIV vaccine development by computer assisted design: the GAIA vaccine. *Vaccine*, 23(17-18):2136–2148, 2005.

[NLW+03] M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemøller, K. Lamberth, S. Buus, S. Brunak und O. Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, 12(5):1007–1017, 2003.

[Pat03] G. Pataki. Teaching integer programming formulations using the traveling salesman problem. *SIAM review*, 45(1):116–123, 2003.

[PBC94] K. C. Parker, M. A. Bednarek und J. E. Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 152(1):163–175, 1994.

[RS04] G. Rätsch und S. Sonnenburg. Accurate Splice Site Detection for *Caenorhabditis elegans*. In B. Schölkopf, K. Tsuda und J.-P. Vert, Hrsg., *Kernel Methods in Computational Biology*, Seiten 277–298. MIT Press, 2004.

[SRH+10] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl und V. Franc. The SHOGUN Machine Learning Toolbox. *J Mach Learn Res*, 11:1799–1802, 2010.

[TDK08] N. C. Toussaint, P. Dönnes und O. Kohlbacher. A mathematical framework for the selection of an optimal set of peptides for epitope-based vaccines. *PLoS Comput Biol*, 4(12):e1000246, 2008.

[TFZ+11] N. C. Toussaint, M. Feldhahn, M. Ziehm, S. Stevanović und O. Kohlbacher. T-cell epitope prediction based on self-tolerance. In *Proceedings of the Second Immunoinformatics and Computational Immunology Workshop*, 2011.

[TH07] C.-W. Tung und S.-Y. Ho. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*, 23(8):942–949, 2007.

[TK09a] N. C. Toussaint und O. Kohlbacher. OptiTope – a web server for the selection of an optimal set of peptides for epitope-based vaccines. *Nucleic Acids Res*, 37(Web Server issue):W617–22, 2009.

[TK09b] N. C. Toussaint und O. Kohlbacher. Towards in silico design of epitope-based vaccines. *Expert Opin Drug Discovery*, 4(10):1047–1060, 2009.

[TMKL11] N. C. Toussaint, Y. Maman, O. Kohlbacher und Y. Louzoun. Universal peptide vaccines – optimal peptide vaccine design based on viral sequence conservation. *Vaccine*, 29(47):8745–8753, 2011.

[Tou11] N.C. Toussaint. *New approaches to in silico design of epitope-based vaccines*. Dissertation, Universität Tübingen, 2011.

[TWKR10] N. C. Toussaint, C. Widmer, O. Kohlbacher und G. Rätsch. Exploiting physicochemical properties in string kernels. *BMC Bioinf*, 11 Suppl 8:S7, 2010.

[Vap95] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[VSRL07] T. Vider-Shalit, S. Raffaeli und Y. Louzoun. Virus-epitope vaccine design: informatic matching the HLA-I polymorphism to the virus genome. *Mol Immunol*, 44(6):1253–1261, 2007.

[WTA+10] C. Widmer, N. C. Toussaint, Y. Altun, O. Kohlbacher und G. Rätsch. Novel machine learning methods for MHC class I binding prediction. In T. Dijkstra, E. Tsivtsivadze, E. Marchiori und T. Heskes, Hrsg., *Pattern Recognition in Bioinformatics*, Jgg. 6282 of *Lecture Notes in Computer Science*, Seiten 98–109. Springer Berlin / Heidelberg, 2010.

Nora C. Toussaint studierte Diplom-Informatik mit Nebenfach Biologie an der Humboldt-Universität zu Berlin. Bis Ende 2011 war sie als wissenschaftliche Mitarbeiterin an der Universität Tübingen tätig und promovierte im Bereich Immunoinformatik bei Prof. Dr. Oliver Kohlbacher. Während ihrer Promotion hat sie Forschungsaufenthalte am Centrum Wiskunde & Informatica (CWI) in Amsterdam und an der Bar-Ilan Universität in Israel verbracht. Derzeit arbeitet sie als *Postdoc* in den Forschungsabteilungen Immunologie und Bioinformatik des Memorial Sloan-Kettering Cancer Centers in New York.

Ultraschall-Mosaicing und Bewegungsmodellierung: Anwendungen der medizinischen Bildregistrierung

Christian Wachinger

Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
wachinge@mit.edu

Abstract: Ultraschall ist eine der interessantesten klinischen Bildgebungsmodalitäten. Vor allem die Kosteneffizienz und die Sicherheit für den Patienten haben zu einer weiten Verbreitung geführt. Die Anwendungen von Ultraschall sind aber bisher aufgrund hohen Bildrauschens, spezifischer Artefakte und Abhängigkeit vom Aufnahmewinkel limitiert. So ist die Erstellung von Atmungsmodellen zur Unterstützung klinischer Eingriffe bisher auf Aufnahmen der Kernspin- oder Computertomographie beschränkt. Diese modernen Modellierungsverfahren werden aber auch in der Zukunft, angesichts hoher Kosten, nur einem beschränkten Patientenanteil zur Verfügung stehen. Damit eine breite Patientenmasse von diesen Fortschritten profitieren kann, müssen preiswertere Alternativen gefunden werden. Ultraschall stellt derzeit die aussichtsreichste Alternative dar, allerdings müssen neue Algorithmen entwickelt werden, die sich den Herausforderungen der Ultraschallbildgebung stellen. In meiner Dissertation habe ich mich genau mit der Entwicklung dieser Verfahren beschäftigt, mit der Fokussierung auf Ultraschall-Mosaicing und 4D Atmungsmodellierung. Um gute Ergbnisse zu erzielen, haben wir die gesamte Verarbeitungskette, beginnend bei der Demodulation der Rohdaten über die Erstellung von 4D Zeitserien bis zur korrekten Registrierung, analysiert und neue Beiträge geleistet. Wir haben diese auf zahlreichen internationalen Konferenzen vorgestellt und in top-tier Journalen publiziert. Besonders hervorzuheben ist, dass es uns als Erste gelungen ist ein Bewegungsmodell der Leber während der Atmung anhand von Ultraschalldaten zu erstellen. Dies war erst möglich nach der Entwicklung eines neues Systems zur Erstellung von 4D Ultraschalldaten anhand von Manifold Learning and eines neuen Registrierverfahrens, das die räumlichen als auch die zeitliche Komponente gleichzeitig berücksichtigt und damit die Erzeugung eines korrekten Atmumgsmodells ermöglicht.

1 Einführung

Die letzten 20 Jahre waren revolutionär für die medizinische Bildgebung. Die Entwicklung und Einführung neuer bildgebender Geräte in den klinischen Alltag ermöglicht heutzutage die Darstellung relevanter Strukturen in 3D, über Zeit in 4D, in verschiedenen Kontrasten, sowie deren anatomische als auch funktionelle Charakterisierung. Mittlerweile wird es oft als selbstverständlich angesehen, dass die Daten in digitaler Form vorliegen. Dies beschleunigte jedoch den Einzug der Informatik in eine Domäne die bisher vornehmlich von der Zusammenarbeit zwischen Physikern und Medizinern geprägt war. Für die Weiterent-

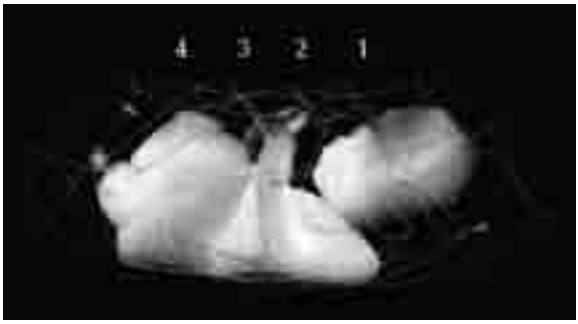


Abbildung 1: Mosaic eines Babyphantoms anhand von vier 3D Aufnahmen.

wicklung in dieser interdisziplinären Forschungsrichtung sind nun aber auch Informatikkompetenzen essentiell, um die Verarbeitung von großen Datenmengen, die Überlagerung von Bildern verschiedener Modalitäten, die Navigation innerhalb des Patienten, und die Analyse medizinischer Bilddaten zu verbessern. In der Informatik wird die Arbeit in dieser Richtung meist als medizinische Bildverarbeitung betitelt, wobei die Bezeichnung natürlich nur einen Teilbereich der damit assoziierten Techniken widerspiegelt. Die vorliegende Dissertation wurde zum einen von dem Wunsch getrieben medizinische Prozeduren zu verbessern um Patienten in Zukunft eine bessere Behandlung anbieten zu können und zum anderen von der Notwendigkeit der Entwicklung theoretisch fundierter Verfahren.

1.1 Ultraschall

Die bildgebende Modalität auf die sich die Dissertation konzentriert ist Ultraschall. Im Gegensatz zu anderen Modalitäten ist Ultraschall kostengünstig, ohne Belastung für den Patienten, und in Echtzeit. Des Weiteren kann Ultraschall direkt im Krankenbett aufgenommen werden. Diese Vorteile haben zu einer weiten Verbreitung von Ultraschallgeräten im klinischen Alltag geführt. Besonders hervorzuheben ist hierbei dass sich die Verbreitung nicht nur auf finanziell gut ausgestattete Kliniken beschränkt, sondern aufgrund der Kosteneffizienz auch auf Entwicklungsländern ausdehnt. Der Nachteil von Ultraschall ist jedoch, dass die Interpretation und Aufnahme von Ultraschallbildern ein spezielles Training erfordert. Vor allem das inhärente Speckle-Rauschen, die Bildartefakte, und die Richtungsabhängigkeit unterscheiden es von anderen Modalitäten und erschweren das sofortige Verständnis. Diese Besonderheiten von Ultraschall führen auch dazu, dass Standardalgorithmen der Bildverarbeitung meist zu schlechten Ergebnissen auf Ultraschallbildern führen. Das Schlüsselelement um gute Ergebnisse zu erhalten besteht in der richtigen Modellierung. Hierfür wird ein gutes Verständnis der physikalischen Prinzipien von Ultraschall benötigt sowie die genaue Kenntnis der theoretischen Grundlagen existierender Methoden. Erst die Erfahrung in beiden Bereichen ermöglicht die Entwicklung neuer Methoden für Ultraschall die in klinischen Anwendungen eingesetzt werden können.

1.2 Mosaicing und Bewegungsmodellierung

Die klinischen Anwendungen mit denen sich die Dissertation beschäftigt sind die Erstellung von Mosaiken und die Bewegungsmodellierung mit Hilfe von 4D Ultraschalldaten. Die Motivation für die Erstellung von Ultraschallmosaiken ist ähnlich zu der für die Erstellung von Fotomosaiken oder Panoramabildern. Zum einen ermöglicht es die Erstellung qualitativ hochwertigerer Bilder, da die Informationen von mehreren Bildern kombiniert werden können, und zum anderen die Präsentation eines größeren Bildbereichs. Ein Beispiel eines Mosaics aus vier 3D Ultraschallaufnahmen ist in Abbildung 1 zu sehen. In mehreren Studien wurden hierfür die klinischen Vorteile ausgeführt. Erstens, ermöglicht Ultraschall-Mosaicing die räumliche Beziehung zwischen anatomischen Strukturen besser zu verstehen, die zu groß für eine Aufnahme sind [KCK+03]. Zweitens, haben Ultraschalldiagnostiker die Möglichkeit Strukturen aus verschiedenen Blickwinkeln darzustellen [LRJ+05]. Drittens, ermöglicht es Größen- und Distanzmessungen von großen Organen [KCK+03]. Viertens, können individuelle Strukturen im größeren räumlichen Kontext identifiziert werden [DIG+02]. Schließlich ermöglicht es der größere Darstellungsbereich auch Experten die nicht an die Ultraschallbildgebung gewöhnt sind die räumlichen Verhältnisse besser zu verstehen [HSK+03], und somit die Lücke zwischen den Modalitäten zu verringern und Diagnosen in Ultraschall an andere Experten zu übermitteln.

Die zweite klinische Anwendung von Interesse ist die Bewegungsmodellierung, wobei wir uns vor allem auf die Bewegung aufgrund von Atmung konzentrieren. Die Atmung ist ein zyklischer, irregulärer Prozess der zur Deformation der abdominalen und thorakalen Regionen führt. Das Atmungssignal stellt hierbei den aktuellen Atmungszustand des Patienten dar. Für eine Vielzahl von Anwendungen ist es erforderlich dem aufgenommenen Ultraschallbild den korrespondierenden Atmungszustand zuzuweisen. Unter anderem haben wir eine Technik basierend auf manifold learning entwickelt, die es ermöglicht den Atmungszustand des Patienten nur anhand der aufgenommenen Bilddaten zu erkennen, ohne die Verwendung externer Trackingsysteme. Diese Methode haben wir für die Erstellung von 4D Ultraschalldaten verwendet. Abbildung 2 illustriert ein extrahiertes Signal aus einer Ultraschallsequenz zusammen mit dem Referenzsignal eines Trackingsystems. Zeitlich aufgelöstee Bilddaten werden extensiv eingesetzt um Herz [vS08] und Leberbewegung [vS08] zu analysieren. Mit Hilfe der erstellten 4D Daten ist es möglich ein Bewegungsmodell des Organs zu berechnen. Wir haben hierfür ein neues Registrierverfahren entwickelt, dass sowohl die räumliche als auch die zeitliche Komponente gleichzeitig berücksichtigt. Das extrahierte Bewegungsmodell führt zum Beispiel zu Vorteilen in der Lokalisierung für die Strahlentherapie, da die Bestrahlung gesunden Gewebes minimiert werden kann [CMM+08]. Zudem ermöglicht es eine exaktere Erhitzung und Zerstörung pathogenen Gewebes mit fokussiertem Ultraschall [TSM+03].

1.3 Registrierung

Die beiden vorgestellten Anwendungen basieren auf der gleichen zugrundeliegenden Technik, der Bildregistrierung oder kurz Registrierung. In der Registrierung werden zwei Bilder

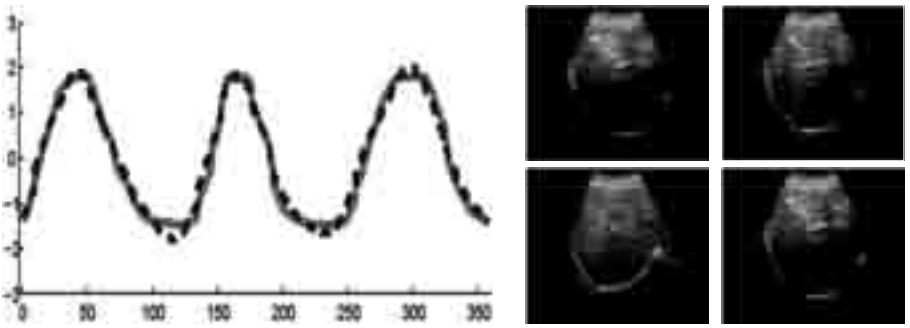


Abbildung 2: Links: Atmungssignal extrahiert aus einer Sequenz von Ultrschallbildern mit Manifold Learning (rot). Referenzsignal anhand eines externen Trackingsystems (blau). Hohe Korrelation von 95% der beiden Singale. Rechts: Auszüge aus der Ultraschallsequenz.

die dasselbe Objekt darstellen korrekt überlagert. Sie ermöglicht es die Bilder innerhalb eines gemeinsamen Koordinatensystems anzuzeigen und damit komplementäre Informationen zwischen den Bildern zu propagieren. Es wird hierbei zwischen der rigiden (Verschiebung und Rotation) und nicht-rigiden (auch Deformation) Registrierung unterschieden. Während für die Erstellung von Mosaiken vornehmlich rigide Methoden eingesetzt werden benötigt man für die Bewegungsmodellierung nicht-rigide Methoden. Eine weitere Gemeinsamkeit ist, dass in beiden Fällen nicht nur zwei Bilder registriert werden müssen, sondern eine Gruppe von Bildern. Hierfür wurden spezielle, grupppenbasierte Registrierungsmethoden entwickelt, die wir durch eigene Beiträge ergänzt haben.

Von besonderer Bedeutung ist die mathematische Modellierung der Registrierung. Hierbei werden vor allem Konzepte aus dem Teilgebiet der Stochastik angewendet. Eine übliche Annahme ist hierbei die Unabhängigkeit der Pixel untereinander, um die Ableitung einfacher zu gestalten. Wir die Unabhängigkeit durch die Markov-Bedingung ersetzt. Dies ermöglichte es uns neue theoretische Einsichten zu erhalten, Verfahren zu standardisieren, und neue Registrierkonzepte vorzuschlagen [WN12a]. Darüberhinaus hat dieses Modell ein neues multi-modales Registrierverfahren motiviert [WN12b].

1.4 Gliederung

Im Folgenden werden wir kurz das Prinzip der Ultraschallbildgebung skizzieren und eigene Beträge im Bereich der Demodulation vorstellen. Darauf aufbauend entwickeln wir Registrierverfahren die an Ultraschall angepasst sind und zu besseren Ergebnissen als alternative Verfahren führen. Wir konzentrieren uns hierbei auf ein neues Ähnlichkeitsmaß für Mosaicing, dass auf der Nakagami-Verteilung basiert. Abschließend diskutieren wir die Akquise von 4D Ultraschalldaten mit einer Wobblersonde und die korrekte Registrierung dieser Daten mit einem neuen Verfahren, das die räumliche als auch zeitliche Dimension berücksichtigt.

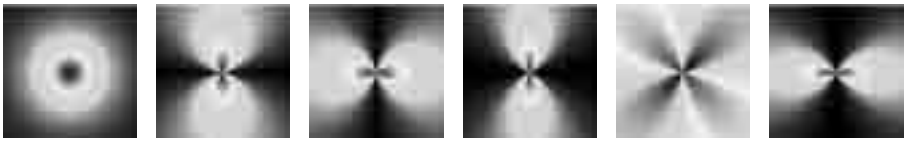


Abbildung 3: Magnituden der 2D Hilberttransformationen mit log-Gabor Kernel im Frequenzraum.

2 Ultraschall Demodulation

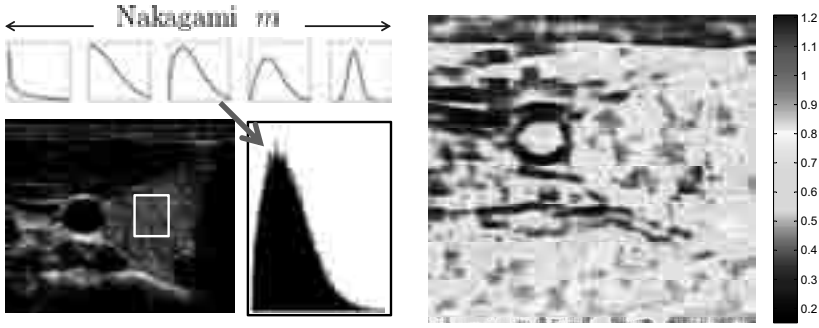
Für die Ultraschallbildgebung werden elektrische Impulse in mechanische umgewandelt unter Ausnutzung des piezoelektrischen Effekts. Die vibrierenden Piezoelemente erzeugen eine Schallwelle im umgebenden Gewebe, die an Übergängen mit unterschiedlichem akustischem Widerstand zu einem gewissen Teil reflektiert wird. Die reflektierte Welle wird in ein elektrisches Signal umgewandelt, dass noch von den hochfrequenten Anteilen der Trägerwelle gekennzeichnet ist. Das Radiofrequenzsignal (RF Signal) muss demoduliert werden um das informationstragende Signal zu extrahieren. Genauer wird hierfür ein Hüllenkurven-Detektor eingesetzt. Dieser basiert auf der Berechnung des analytischen Signals. Das analytische Signal ist ein komplexwertiges Signal, dass das ursprüngliche Signal g als Realteil und das Hilbert-transformierte Signal $\mathcal{H}\{g\}$ als Imaginärteil enthält. Um die Amplitude A des Signals zu erhalten wird der Betrag des analytischen Signals berechnet:

$$A = \sqrt{g^2 + \mathcal{H}\{g\}^2}. \tag{1}$$

Diese Demodulation wird bisher in 1D, separat für jeden Ultraschallstrahl durchgeführt. In unserer Arbeit haben wir die Erweiterung des analytischen Signals auf 2D [WSF09] für die Demodulation von RF Daten verwendet [WKN11]. Die Daten werden dafür mit 2D Hilbert Transformationen der ersten und zweiten Ordnung gefiltert. Die Filtermagnituden im Frequenzraum sind in Abbildung 3 dargestellt. In der Dissertation haben wir anhand einer Vielzahl von Ultraschallbildern die verbesserte Qualität der 2D Demodulation illustriert. Neben der qualitativen Evaluierung haben wir auch einen Weg gefunden die Ergebnisse quantitativ mit Hilfe der Nakagami Verteilung zu untermauern. Anhand eines theoretischen Modells von Ultraschall wurde die Nakagami Verteilung als akkurates Modell für die Verteilung von demodulierten Ultraschalldaten vorgestellt [SDR$^+$02]. Die Nakagami-Verteilung mit Form- m und Skalenparameter ω ist folgedermaßen definiert:

$$p(x \mid m, \omega) = \frac{2m^m x^{2m-1}}{\Gamma(m)\omega^m} \exp\left(-\frac{m}{\omega}x^2\right), \forall x \in \mathbb{R}_+. \tag{2}$$

mit Gammafunktion Γ. Wir haben mit Hilfe von goodness-of-fit Tests evaluiert wie gut die empirische Verteilung der demodulierten Daten dem theoretischen Modell entspricht. Dies ist graphisch in der Abbildung 4(a) illustriert. In einem Fenster wird die empirische Verteilung berechnet. Anschließend werden die Parameter m und ω geschätzt. Dargestellt ist nur der wichtigere Formparamter m. Durch Bestimmung der P-Werte wird anschließend die Übereinstimmung der empirischen Verteilung (Histogramm) mit der geschätzten



(a) Nakagami Fit (b) m Parameter

Abbildung 4: Links: Schätzung des m Parameters der Nakagami-Verteilung anhand empirischer Bilddaten. Die Nakagami-Verteilung ist für verschiedene m Werte dargestellt (blaue Kurven). Die Nakagami-Verteilung die am Besten den Daten entspricht ist als rote Kurve eingezeichnet. Rechts: m Werte über das ganze Bild berechnet. Auffällig ist die hohe Variation.

Nakagami-Verteilung (rote Kurve) quantifiziert. Die Ergebnisse zeigen, dass die Demodulation mit Hilfe des 2D analytischen Signals zu eindeutig besseren Ergebnissen führt. Dieses Ergebnis hat direkte Konsequenzen für die Segmentierung, Klassifikation und Registrierung in Ultraschall die auf dem Nakagami-Modell aufbauen. Des Weiteren haben wir herausgefunden, dass das 2D analytische Signal besser dafür geeignet ist Merkmale in Bildern zu erkennen. Eine Anwendung dafür ist die Nadelerkennung. Die Arbeit wurde vom Patentamt der TU München zum Patent angemeldet.

3 Ultraschall-Registrierung

Nach der Demodulation der Daten ist der nächste Schritt die korrekte Registrierung. Die zwei wichtigen Komponenten der Registrierung mit denen wir uns beschäftigen sind die Optimierung und die Ähnlichkeitsmaße. Für die Optimierung haben wir den efficient second-order minimization (ESM) Algorithmus für die simultane Registrierung von Ultraschalldaten hergeleitet [WN09]. Simultane Registrierung bezieht sich hierbei auf die simultane Optimierung von mehreren Transformationmatrizen wie es bei der gruppenbasierten Registrierung für Mosaicing benötigt wird. Unsere Experimente zeigen die schnellere Konvergenz von ESM im Vergleich zum Gauß-Newton Verfahren.

Unsere Beiträge bezüglich der Ähnlichkeitsmaße beinhalten die mathematische Herleitung eines neuen Rahmenwerks für multivariate Maße für die simultane Registrierung. Die Herleitung geschieht ausgehend von einer wahrscheinlichkeitstheoretischen Modellierung der Registrierung. Die neue Klasse von Ähnlichkeitsmaßen ist besonders für Ultraschall-Moasaicing relevant, da die variierende Anzahl überlappender Bilder kein Problem darstellt. Wir haben das neue Rahmenwerk in mathematischer Beziehung zu bereits existierende Ansätzen gestellt. Dies wurde durch Variation der initialen Annahmen und durch

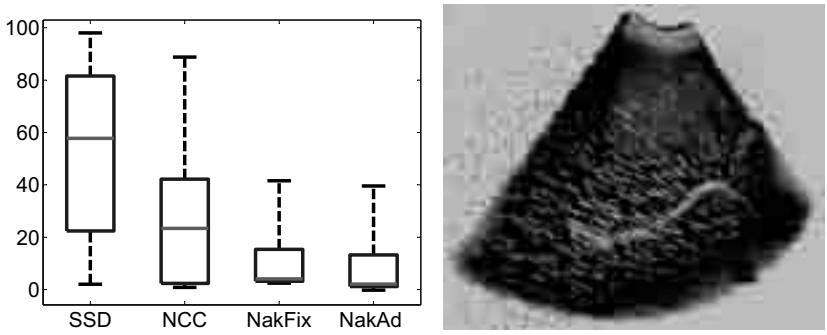


Abbildung 5: Links: Boxplot der Ergebnisse der Registrierexperimente. NakFix: Nakagami-Verteilung mit globalen Parameterwerten. NakAd: Lokale Anpassung der Parameterwerte. Y-Achse: Registrierfehler. Rechts: Volume Rendering eines Ultrschallvolumens mit dem zugehörigen Deformationsfeld.

Anwendung probabilistischer Gesetze ermöglicht.

In einem weiteren Beitrag haben wir ultraschall-spezifische Ähnlichkeitsmaße hergeleitet, in dem wir die Gauß-Rauschverteilung durch die Nakagami-Verteilung ersetzt haben [WKN12]. Die Annahme einer Nakagami-Verteilung ist adäquater für Ultraschalldaten, wie bereits früher dargestellt wurde. Interessanterweise profitieren wir hier direkt von der Demodulation der Daten mit dem 2D analytischen Signal, da dies zu Daten führt die besser dem theoretischen Modell entsprechen. Für die Verwendung der Nakagami-Verteilung müssen die zugehörigen Parameter spezifiziert werden. Anstatt diese global heuristisch zu setzen, schätzen wir sie direkt von den Daten. Die Notwendigkeit für eine lokale Schätzung ist in Abbildung 4(b) illustriert. Der m Parameter variiert sehr stark innerhalb des Ultraschallbildes. Würden wir mit einem einzigen globalen Parameter arbeiten, wie es in früheren Arbeiten der Fall ist, so wäre dieser immer nur für einen bestimmten Bereich optimal. Durch die lokale Adaptation des Ähnlichkeitsmaßes erhalten wir für jeden Bereich die am besten passende Metrik. Unsere Ergebnisse für Ultraschall-Mosaicing mit Ähnlichkeitsmaßen basierend auf der Nakagami-Verteilung zeigen eine klare Verbesserung der Registrierung im Vergleich zu Standardmetriken wie sum of squared differences (SSD) und normalized cross correlation (NCC), siehe Abbildung 5.

4 Ultraschall Bewegungsmodellierung

Voraussetzung für die Bewegungsmodellierung sind 4D Ultraschalldaten. Wir haben hierfür eine neue Technik entwickelt, basierend auf der bereits erwähnten Extraktion des Atmungssignals mit manifold learning [WYRN12]. Wir nehmen Ultraschalldaten mit einer Wobblersonde über mehrere Atmungszyklen auf, siehe Abbildung 6. Die direkte Verwendung der Daten ist nicht möglich, da sich während einer 3D Aufnahme die sich aus mehreren 2D Aufnahmen über verschiedene Winkel zusammensetzt, der Atemzustand des

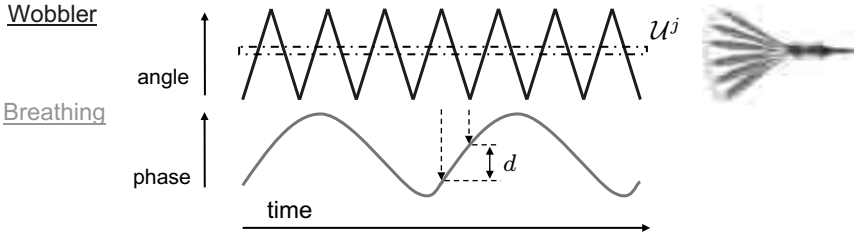


Abbildung 6: Wobblerwinkel (blau) und Atmugssignal (grau) über Zeit. Gestrichelte Linie zeigt die Veränderung des Atemzustands d während eines Durchlaufs an. Strichpunkt Linie zeigt Aufnahmen mit gleichem Winkel über verschiedene Zyklen hinweg an.

Patienten ändert. Die Veränderung ist durch d in der Abbildung dargestellt. Mit der entwickelten Technik weisen wir jedem 2D Ultraschallbild den korrespondierenden Atmungszustand zu. Somit können wir die Daten retrospektiv umsortieren und konsistente 4D Daten erzeugen. Neben Ultraschall haben wir die entwickelte Technik auch auf Magnetresonanzdaten angewandt.

Für die Nachfolgende deformierbare Registrierung haben wir die bereits vorher erwähnten Registrierfahren auf ein B-Spline basiertes Deformationsmodell erweitert. Die bedeutendste Neuerung der neuen Methode ist die Berücksichtigung der zeitlichen Dimension während der Registrierung. Bereits existierende Verfahren in der Literatur verwenden einen zweistufigen Prozess. Zuerst werden die Daten der Zeitreihe registriert und im zweiten Schritt wird das zeitlich zusammengesetzte Deformationsfeld regularisiert. Wir betten die Daten in einen um eine Dimension (Zeit) erweiterten Raum ein. Für die Registrierung von 3D Daten arbeiten wir mit 4D Deformationsfeldern. Wir garantieren somit, dass zu jedem Zeitpunkt in der Registrierung ein glattes Deformation in räumlicher als auch zeitlicher Dimension besteht. Dies in Kombination mit dem bereits vorher beschriebenen, simultanen Registrieransatz charakterisiert die vorgeschlagene Methode zur Bewegungsmodellierung.

In Abbildungen 7 und 8 zeigen wir die Registrierung einer synthetischen Ringsequenz. Wir zeigen die ursprüngliche Sequenz, die deformierten Ringe und die zugehörigen Deformationsfelder. Außerdem zeigen wir die Einbettung in den um eine Dimension augmentierten Raum, in diesem Fall 3D. Im Schnitt entlang der zeitlichen Richtung können wir sehr gut die Veränderung in den Bilddaten über Zeit einschätzen. Diese vorteilhaften Visualisierungsmöglichkeiten ergeben sich ganz natürlich aus dem verwendeten Modellierungsansatz. In Abbildung 5 zeigen wir eine Ausschnitt aus der Bewegungsmodellierung von Leberdaten in 4D. Illustriert ist ein Ultraschallvolumen mit Volume Rendering zusammen mit dem Deformationsfeld bezüglich der ersten Aufnahme in der Sequenz.

5 Schlussfolgerung

In der Dissertation haben wir Verfahren entwickelt um das Ultraschall-Mosaicing und die Bewegungsmodellierung zu verbessern. Dies beginnt bei der Demodulation der Ultraschall

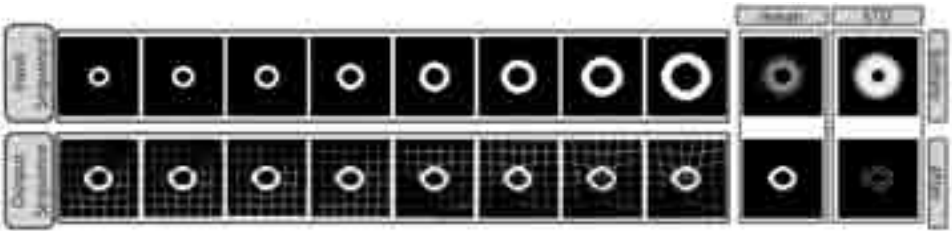
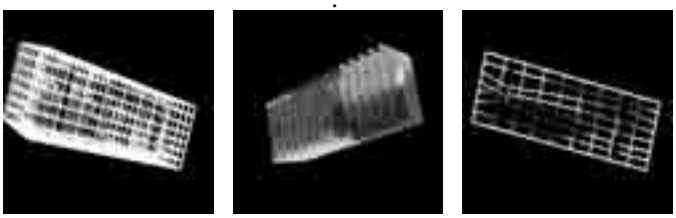


Abbildung 7: Eingabe- und Ausgabesequenz für das Ringexperiment.



(a) (b) (c)

Abbildung 8: (a) $(N+1)$D wireframe Mesh. (b) ND Meshes in $(N+1)$D. (c) Schnitt des $(N+1)$D Mesh entlang der zeitlichen Richtung.

RF Daten, setzt sich fort bei der Herleitung neuer Ähnlichkeitsmaße und Optimierungsverfahren für die Registrierung, und endet in der Bewegungsmodellierung in 4D. Besonders die vorgestellte Modellierung der Leberbewegung anhand Ultraschall wurde in dieser Art zum ersten Mal durchgeführt. Bisher wurde dies nur mit Hilfe von Daten der Computer- oder Kernspintomographie versucht. Wie bereits erwähnt stellen die speziellen Charakteristika von Ultraschall besondere Herausforderungen dar. Es besteht aber auch ein enormes Potenzial darin akkurate Bewegungsvorhersagen mit der weitaus kostengünstigeren Ultraschallvariante zu erstellen und somit einer breiteren Patientenmasse anzubieten.

Literatur

[CMM+08] R. Colgan, J. McClelland, D. McQuaid, PM Evans, D. Hawkes, J. Brock, D. Landau und S. Webb. Planning lung radiotherapy using 4D CT data and a motion model. *Physics in Medicine and Biology*, 53:5815, 2008.

[DIG+02] C.F. Dietrich, A. Ignee, M. Gebel, B. Braden und G. Schuessler. Imaging of the Abdomen. *Z Gastroenterol*, 40:965–970, 2002.

[HSK+03] Wolfgang Henrich, Annette Schmider, Siri Kjos, Boris Tutschek und Joachim W. Dudenhausen. Advantages of and applications for extended field-of-view ultrasound in obstetrics. *Archives of Gynecology and Obstetrics*, V268:121–127, Jun 2003.

[KCK+03] Se Hyung Kim, Byung Ihn Choi, Kyoung Won Kim, Kyoung Ho Lee und Joon Koo Han. Extended Field-of-View Sonography: Advantages in Abdominal Applications. *Journal of Ultrasound in Medicine*, 22(4):385–394, 2003.

[LRJ+05] Y.L. Leung, A.L. Roshier, S. Johnson, R. Kerslake und D.S. McNally. Demonstration of the appearance of the paraspinal musculoligamentous structures of the cervical spine using ultrasound. *Clin Anat*, 18(2):96–103, 2005.

[SDR+02] PM Shankar, VA Dumane, JM Reid, V Genis, F Forsberg, CW Piccoli und BB Goldberg. Classification of ultrasonic B-mode images of breast masses using Nakagami distribution. *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions on*, 48(2):569–580, 2002.

[TSM+03] C. Tempany, E.A. Stewart, N. McDannold, B.J. Quade, F.A. Jolesz und K. Hynynen. MR Imaging–guided Focused Ultrasound Surgery of Uterine Leiomyomas: A Feasibility Study1. *Radiology*, 226(3):897, 2003.

[vS08] Martin von Siebenthal. *Analysis and Modelling of Respiratory Liver Motion using 4DMRI*. Dissertation, Eidgenössische Technische Hochschule ETH Zürich, 2008.

[WKN11] Christian Wachinger, Tassilo Klein und Nassir Navab. The 2D Analytic Signal on RF and B-mode Ultrasound Images. In *Information Processing in Medical Imaging (IPMI)*, 2011.

[WKN12] Christian Wachinger, Tassilo Klein und Nassir Navab. Locally adaptive Nakagami-based ultrasound similarity measures. *Ultrasonics*, 52(4):547 – 554, 2012.

[WN09] Christian Wachinger und Nassir Navab. Similarity Metrics and Efficient Optimization for Simultaneous Registration. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[WN12a] Christian Wachinger und Nassir Navab. A Contextual Maximum Likelihood Framework for Modeling Image Registration. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[WN12b] Christian Wachinger und Nassir Navab. Entropy and Laplacian images: Structural representations for multi-modal registration. *Medical Image Analysis*, 16(1):1 – 17, 2012.

[WSF09] L. Wietzke, G. Sommer und O. Fleischmann. The Geometry of 2D Image Signals. In *CVPR*, Seiten 1690–1697, 2009.

[WYRN12] Christian Wachinger, Mehmet Yigitsoy, Eric Rijkhorst und Nassir Navab. Manifold Learning for Image-Based Breathing Gating in Ultrasound and MRI. In *Medical Image Analysis*, 2012.

Christian Wachinger, geboren am 10. Februar 1982, hat 2007 sein Diplom in Informatik von der TU München erhalten. Während dieser Zeit absolvierte er ein Zusatzstudium am CDTM im Breich Technology Management, studierte ein Jahr an der Telecom ParisTech und Ecole Centrale Paris, France, und verbrachte sechs Monate in Princeton, USA. In 2011 hat er seine Doktorarbeit mit dem Titel *Ultrasound Mosaicing and Motion Modeling: Applications in Medical Image Registration* an der TU München erfolgreich verteidigt. Derzeit ist er als Post-Doc im Computer Science and Artificial Intelliance Laboratory (CSAIL) am Massachusetts Institute of Technology (MIT), Cambridge mit einer Zweitanstellung im Department für Neurologie an der Harvard Medical School.

Verhaltensprofile – Ein Relationaler Ansatz zur Verhaltenskonsistenzanalyse

Matthias Weidlich

Technion – Israel Institute of Technology
32000 Haifa, Israel
weidlich@tx.technion.ac.il

Abstract: Die Entwicklung von Informationssystemen im Unternehmensumfeld wird oft durch Geschäftsprozessmodelle unterstützt. Unterschiedliche Modellierungsziele resultieren allerdings in unterschiedlichen Modellen desselben Prozesses. Nichtsdestotrotz sollten die entsprechenden Modelle konsistent, d.h. frei von Widersprüchen sein. Die Striktheit des Konsistenzbegriffes steht hierbei in Konflikt mit der Eignung der Prozessmodelle für einen bestimmten Zweck. Dieser Beitrag stellt einen Ansatz für die Analyse von Verhaltenskonsistenz vor, welcher sich fundamental von existierenden Arbeiten unterscheidet. Grundlage des Ansatzes ist eine Verhaltensabstraktion, das Verhaltensprofil eines Prozessmodells, welches für bestimmte Modellklassen effizient berechenbar ist. Auf Basis von Verhaltensprofilen werden Konsistenzbegriffe und Konsistenzmaße, sowie ergänzende Analysetechniken vorgestellt.

1 Einführung

Das Erstellen einer Systemspezifikation auf Basis von Anforderungen ist entscheidend für den Erfolg von Softwareentwicklungsprojekten. Im Unternehmensumfeld hat sich die Modellierung von Geschäftsprozessen bewährt um die Kluft zwischen Geschäfts- und Softwareentwicklung zu überbrücken. Prozessmodelle unterstützen u.a. den Abgleich von geschäftlichen Anforderungen mit der Funktionalität von Informationssystemen [LPB99], sowie die Entwicklung von Prozessorientierten Informationssystemen [Kin09].

Prozessmodelle beschreiben den Ablauf zur Erreichung eines Geschäftszieles [Wes07]. Wie alle konzeptionellen Modelle, stellen sie eine reduzierte Abbildung eines Originals dar [Küh06]. Auch zeichnen sie sich durch ein pragmatisches Merkmal aus. Der *Zweck* der Modellierung gibt vor, welche Eigenschaften des Originals in das Modell übernommen werden und welche Art von Reduktion durchgeführt wird. Differenzen zwischen Prozessmodellen desselben Prozesses sind somit oft unterschiedlichen Modellierungszielen geschuldet. Die oft zitierte Problematik des 'Business-IT-Gap' [RP00] ist ein Beispiel. Prozessmodelle welche geschäftliche Anforderungen darstellen, setzen den Fokus darauf, *was* im Rahmen eines Geschäftsprozesses ausgeführt wird. Implementierungsnahe Prozessmodelle hingegen zeigen, *wie* der Prozess unter Berücksichtigung eines konkreten Systemumfelds realisiert wird. Unterschiede zwischen diesen Modellen entsprechen der Regel, ihre Vermeidung würde die Eignung der Modelle für einen gewissen Zweck beeinträchtigen. Nichtsdesto-

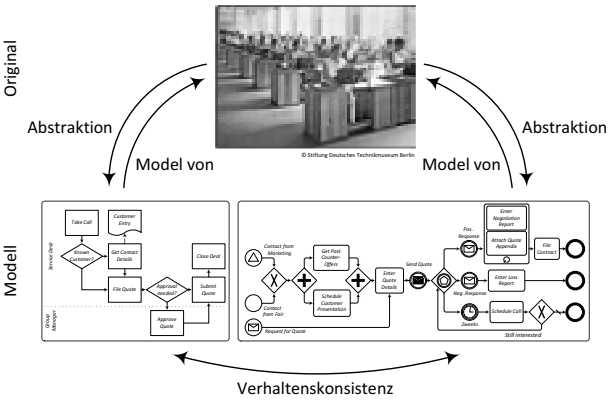


Abbildung 1: Die Frage der Verhaltenskonsistenz: Prozessmodelle stellen verschiedene reduzierte Abbildungen eines Prozesses (eines Originals) dar, sollen jedoch widerspruchsfrei sein.

trotz sollten die entsprechenden Modelle konsistent, d.h. frei von Widersprüchen sein. Da Prozessmodelle zuvorderst Verhalten beschreiben, kommt der Kontrollflussperspektive eine besondere Bedeutung zu. Diese Frage der Verhaltenskonsistenz ist in Abbildung 1 dargestellt und lässt sich wie folgt zusammenfassen:

Wie analysiert man Verhaltenskonsistenz für Prozessmodelle desselben Prozesses?

Verhaltenskonsistenz wird oft im Sinne von Widerspruchsfreiheit definiert, siehe [Zel95], und bezieht sich auf eine grundsätzliche Fragestellung der Informatik. Für Verhaltensmodelle im Allgemeinen und Prozessmodelle im Besonderen gibt es eine Vielzahl von Ansätzen zur Analyse von Verhaltenskonsistenz. Jene basieren auf Verhaltensäquivalenzen und nehmen an, dass Prozessmodelle in einer hierarchischen Verfeinerungsrelation stehen. Folglich weisen sie eine hohe Berechnungskomplexität auf und erlauben es nicht, den Konsistenzbegriff graduell für einen bestimmten Anwendungsfall anzupassen.

Dieser Beitrag stellt eine Analyse von Verhaltenskonsistenz vor, welche sich fundamental von existierenden Arbeiten unterscheidet. Kern des Ansatzes ist das Verhaltensprofil eines Prozessmodells. Das Verhaltensprofil wird als Menge von Relationen definiert, welche Verhaltenscharakteristika von Prozessmodellen beschreiben. Es ist eine Abstraktion des Verhaltens, die für bestimmte Modellklassen effizient berechenbar ist. Auf Basis von Verhaltensprofilen werden Konsistenzbegriffe und Konsistenzmaße für Prozessmodelle vorgestellt. Weiter definiert der Ansatz ergänzende Analysetechniken: eine Algebra für Verhaltensprofile und eine Modellsynthese. Der vorliegende Beitrag stellt eine Zusammenfassung des Ansatzes dar. Eine vollständige Beschreibung ist in [Wei11] zu finden.

Der Beitrag gliedert sich wie folgt. Der nächste Abschnitt führt ein Beispiel ein. Abschnitt 3 stellt Verhaltensprofile vor. Die Anwendung von Verhaltensprofilen für die Analyse von Verhaltenskonsistenz wird in Abschnitt 4 gezeigt. Abschnitt 5 diskutiert verwandte Arbeiten. Der Beitrag schließt mit einer Zusammenfassung und einem Ausblick.

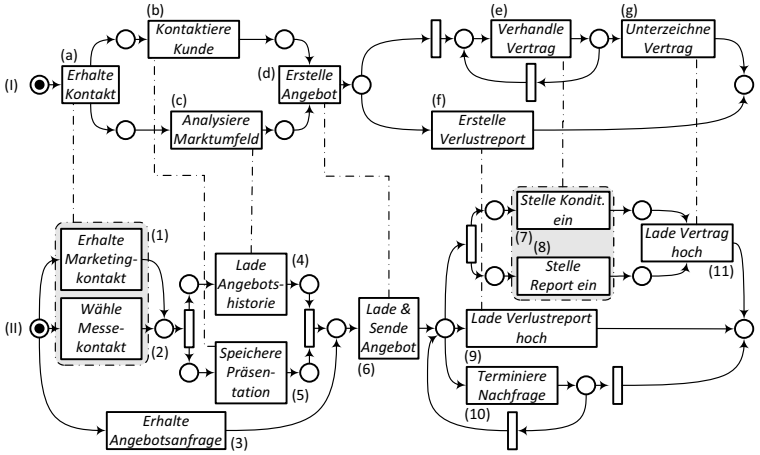


Abbildung 2: Zwei Prozessmodelle, gegeben als markierte Petri-Netze, welche den gleichen Prozess darstellen, aber für unterschiedliche Zwecke erstellt worden sind.

2 Ein Beispiel

Abbildung 2 illustriert das Problem der Verhaltenskonsistenzanalyse mit zwei Prozessmodellen, welche als markierte Petri-Netze vorliegen. Trotz Unterschieden in Syntax und Semantik, sind die meisten Prozessmodellierungssprachen vom Petri-Netz-Formalismus inspiriert. Dies legt die Verwendung von Petri-Netzen für die Konsistenzanalyse nahe.

Die beiden Netze in Abbildung 2 stellen einen Angebotserstellprozess dar. Ein potentieller Kunde wird kontaktiert und es kommt zu einer Angebotserstellung. Bei einer positiven Antwort wird ein Vertrag ausgehandelt, eine negative Antwort führt zu einer Kontaktverlustmeldung. Modell I beschreibt den Prozess aus der Geschäftssicht (Was wird getan?), während Modell II eine implementierungsnahe Sicht einnimmt (Wie wird es umgesetzt?). Obwohl beide Modelle den gleichen Prozess darstellen, unterscheiden sie sich in etlichen Details. Zum Beispiel sieht ausschließlich Modell II eine Nachfrage vor, sofern keine Antwort auf das Angebot empfangen wurde. Nichtsdestotrotz gibt es viele korrespondierende (nicht zwangsläufig semantisch äquivalente) Paare von Transitionen. So entspricht das Unterzeichnen des Vertrages in Modell I dem Hochladen des Vertrages in das System in Modell II, vor dem Hintergrund der unterschiedlichen Modellierungszwecke. Abbildung 2 illustriert diese Korrespondenzen zwischen den Transitionen. Das Finden von Korrespondenzen kann mittels Techniken des Ontology Matching [ES07] unterstützt werden.

Das Beispiel illustriert die besonderen Herausforderungen der Verhaltenskonsistenzanalyse. Die Modelle unterscheiden sich sowohl in der Frage, welche Eigenschaften des Prozesses in das Modell übernommen werden, als auch in der Frage, wie die Eigenschaften reduziert werden. Die Modelle zeigen unterschiedliche Granularität, welche in komplexen 1:n oder n:m Korrespondenzen resultiert. Auch gibt es Unterschiede in der Abdeckung des Prozesses, so dass die Modelle nicht in einer hierarchischen Verfeinerungsrelation stehen.

3 Das Verhaltensprofil

Im Folgenden wird das Konzept des Verhaltensprofils vorgestellt. Des Weiteren wird die Ermittlung von Verhaltensprofilen für eine bestimmte Klasse von Petri-Netzen diskutiert.

Definition. Das Verhaltensprofil eines Petri-Netzes basiert auf seiner Trace-Semantik, die das Verhalten als, potentiell unendlich große Menge von Ausführungsfolgen definiert. Sei $N = (P, T, F)$ ein Petri-Netz. Dann ist $\sigma : \{1, \ldots, n\} \mapsto T$ eine Schaltfolge der Länge $n \in \mathbb{N}$. Die Trace-Semantik ist durch alle Schaltfolgen gegeben, welche in der initialen Markierung M_i des Netzes aktiviert sind und wird mit $\mathcal{T}(N, M_i)$ bezeichnet.

Auf Grundlage der Ausführungsfolgen, definiert das Verhaltensprofil eine Menge von Verhaltensrelationen über Paare von Transitionen. Die Relationen basieren wiederum auf einer Basisrelation, genannt schwache Ordnung[1]. Ein Transitionspaar $(t_1, t_2) \in T \times T$ ist Teil der *schwachen Ordnung* $\succ \subseteq T \times T$, genau dann, wenn es eine Ausführungsfolge $\sigma \in \mathcal{T}(N, M_i)$ der Länge n gibt, in welcher die erste Transition vor der zweiten Transition auftritt, d.h. $\sigma(k) = t_1$ und $\sigma(l) = t_2$ für $1 \leq k < l \leq n$.

Weitere Verhaltensrelationen ergeben sich aus den unterschiedlichen Kombinationen, in welchen zwei Transitionen in schwacher Ordnung stehen können.

- Ein Paar $(t_1, t_2) \in T \times T$ ist Teil der *Exklusivitätsrelation* $+ \subseteq T \times T$, genau dann, wenn $(t_1, t_2) \notin \succ$ und $(t_2, t_1) \notin \succ$. Exklusive Transitionen treten demnach niemals gemeinsam in einer Ausführungsfolge auf.
- Ein Paar $(t_1, t_2) \in T \times T$ ist Teil der *strikten Ordnungsrelation* $\rightsquigarrow \subseteq T \times T$, genau dann, wenn $(t_1, t_2) \in \succ$ und $(t_2, t_1) \notin \succ$. Sofern eine Ausführungsfolge beide Transitionen enthält, tritt t_1 vor t_2 auf.
- Ein Paar $(t_1, t_2) \in T \times T$ ist Teil der *Interleavingrelation* $\| \subseteq T \times T$, genau dann, wenn $(t_1, t_2) \in \succ$ und $(t_2, t_1) \in \succ$. Das heißt, in mindestens einer Ausführungsfolge tritt t_1 vor t_2 auf, sowie umgekehrt.

Als Verhaltensprofil wird die Menge der drei Relationen bezeichnet. Die Definition der Relationen impliziert, dass Exklusivität und Interleaving symmetrisch sind, während die strikte Ordnungsrelation irreflexiv und antisymmetrisch ($(t_1, t_2) \in \rightsquigarrow \wedge\, t_1 \neq t_2 \Rightarrow (t_2, t_1) \notin \rightsquigarrow$) ist. Die Relationen sind paarweise disjunkt und partitionieren zusammen mit der inversen strikten Ordnung $\reflectbox{$\rightsquigarrow$}$ das Kreuzprodukt der Transitionen. Es gilt $t + t$, wenn t höchstens einmal in einer Ausführungsfolge auftreten kann und $t \| t$, sofern t potentiell mehrmals auftritt. Das Verhaltensprofil kann als Matrix dargestellt werden, nebenstehend illustriert für Modell I in Abbildung 2. Es gilt $e + f$ und $f + f$ da die Transitionen nicht gemeinsam in einer Ausführungsfolge auftreten. Da b vor f auftreten kann, umgekehrt jedoch nicht, gilt $b \rightsquigarrow f$. Es gilt $b \| c$ aufgrund der nebenläufigen Aktivierung. Interleaving kann jedoch auch in Schleifen begründet sein (z.B. $e \| e$).

Tabelle 1: Verhaltensprofil für Modell I

	a	b	c	d	e	f	g
a	+	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$
b	$\reflectbox{$\rightsquigarrow$}$	+	$\|$	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$
c	$\reflectbox{$\rightsquigarrow$}$	$\|$	+	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$
d	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	+	$\rightsquigarrow$	$\rightsquigarrow$	$\rightsquigarrow$
e	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\|$	+	$\rightsquigarrow$
f	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	+	+	+
g	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	$\reflectbox{$\rightsquigarrow$}$	+	+

[1] Der Begriff der Ordnung ist informell zu verstehen. Die in diesem Abschnitt eingeführten Relationen erfüllen die Anforderungen an eine Quasiordnung oder Halbordnung nur für bestimmte Klassen von Netzen.

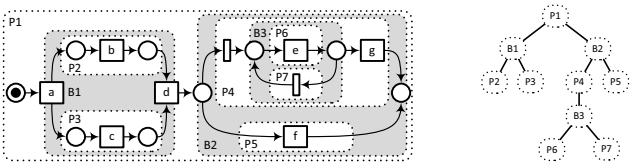


Abbildung 3: Ermittlung des Verhaltensprofils durch strukturelle Dekomposition eines Netzes: 3(a) dreifach-zusammenhängende Graphfragmente, 3(b) die Fragmente als Zerlegungsbaum.

Die Relationen werden als Konsequenz der Existenz von bestimmten Ausführungsfolgen gebildet. Kausalitäten zwischen Transitionen, bzw. ihrem Auftreten, werden nicht im Verhaltensprofil erfasst. Für die Anwendung im Rahmen der Verhaltenskonsistenzanalyse wird später diskutiert, dass dies eine gewünschte Eigenschaft ist. In anderen Anwendungsfällen kann die Abstraktion von Kausalitäten jedoch unerwünscht sein. Für diese Fälle wird in [Wei11] ein kausales Verhaltensprofil definiert. Jenes erweitert den hier vorgestellten Begriff um eine Relation, welche eine kausale Kopplung des Auftretens zweier Transitionen erfasst.

Ermittlung von Verhaltensprofilen. Das Verhaltensprofil stellt eine Abstraktion der Trace-Semantik eines Petri-Netzes dar. Für bestimmte Klassen von Netzen ist eine Betrachtung der Ausführungsfolgen jedoch nicht notwendig um auf das Verhaltensprofil zu schließen. Stattdessen können die Relationen des Verhaltensprofils direkt von der Struktur des Netzes abgeleitet werden. Das Problem der Zustandsraumexplosion durch Nebenläufigkeit, welche mit einer exponentiellen Anzahl von Ausführungsfolgen einhergeht, wird somit umgangen. Ohne auf die formalen Details einzugehen, wird im Folgenden einen Ansatz zur Ermittlung des Verhaltensprofils für Petri-Netze zusammengefasst, welche die Free-Choice, Workflow und Soundness Eigenschaften zeigen. Diese Eigenschaften sind sowohl syntaktischer (beispielsweise gibt es eine initiale Stelle ohne Vorgänger und eine finale Stelle ohne Nachfolger) als auch semantischer Natur (so gibt es keine Verklemmungen). Für die Prozessmodellierung hat diese Netzklasse eine große Bedeutung. Die Grundkonzepte der meisten Prozessmodellierungssprachen können auf entsprechende Netze zurückgeführt werden [LVD09].

Der Ansatz zur Ermittlung des Verhaltensprofils wendet eine Graphzerlegung an, welche Fragmente anhand ihres Zusammenhangs identifiziert. Die dreifach-zusammenhängende Zerlegung bestimmt Fragmente, welche je zwei Randknoten haben, die das Fragment mit dem Rest des Graph verbinden. Weiter sind die Fragmente frei von Überschneidungen, so dass ein Zerlegungsbaum die Hierarchie des Enthaltenseins abbildet.

Die Anwendung dieser Zerlegungstechnik für den Graphen eines Petri-Netzes ist in Abbildung 3 am Beispiel des vorab diskutierten Netzes illustriert. Fragmente werden durch Teilnetze gebildet, welche Stellen oder Transitionen als Randknoten haben. Die Teilnetze lassen sich anhand ihrer strukturellen Eigenschaften klassifizieren. So ist Teilnetz B2 eine Stellen-begrenzte (die Randknoten sind Stellen) azyklische Bond-Struktur (mehrere unabhängige Pfade zwischen den Randknoten). In [Wei11] wurde gezeigt, wie sich die

Relation des Verhaltensprofils für zwei Transitionen aus dem Zerlegungsbaum und der Klassifikation der Fragmente ableiten lässt. Als Beispiel seien hier Transitionen e und f gewählt. Beide Transitionen lassen sich in dem Zerlegungsbaum lokalisieren (Teilnetz P5 bzw. P6). Nun wird der niedrigste gemeinsame Vorfahre im Zerlegungsbaum bestimmt (Teilnetz B2), sowie der Pfad von der Wurzel des Baums zu diesem Vorfahren untersucht. Da dieser Pfad frei von Stellen-begrenzten zyklischen Bond-Fragmenten ist und der Vorfahre ein Stellen-begrenztes azyklisches Bond-Fragment ist, gilt $e + f$, d.h. die Transitionen sind exklusiv zueinander. Mit dieser Methode lässt sich die Verhaltensrelation für ein Paar von Transitionen in linearer Zeit zu der Größe des Petri-Netzes bestimmen, sofern alle Schleifen im Netz wohlstrukturiert sind. Sofern dies nicht der Fall ist, greift ein komplementärer Ansatz. Dieser basiert auf Resultaten der Petri-Netz Theorie zur Ableitung der Nebenläufigkeitsrelation und erlaubt die Ermittlung des Verhaltensprofils in kubischer Zeit zur Netzgröße. Die Kombination beider Ansätze ermöglicht somit die effiziente Ermittlung des Verhaltensprofils.

Sofern ein Petri-Netz die syntaktischen und semantischen Anforderungen nicht erfüllt, kann das Verhaltensprofil aus dem Unfolding des Netzes abgeleitet werden. Das Unfolding stellt eine kompakte Repräsentation des Zustandsraumes dar, dessen Berechnung jedoch ein NP-vollständiges Problem ist. Die generelle Anwendbarkeit des Ansatzes geht demnach mit einer hohen Berechnungskomplexität einher.

4 Konsistenzanalyse auf Basis von Verhaltensprofilen

Der Ansatz zur Konsistenzanalyse mit Verhaltensprofilen ist schematisch in Abbildung 4 dargestellt und umfasst die folgenden Schritte:

(1) Korrespondenzen zwischen Transitionen zweier Petri-Netze werden konstruiert.
(2) Die Verhaltensprofile der Petri-Netze werden ermittelt.
(3) Mittels der Verhaltensprofile werden Konsistenzkriterien überprüft.
(4) Mittels der Verhaltensprofile werden Konsistenzmaße berechnet.
(5) Eine Algebra und eine Modellsynthese für Verhaltensprofile erlauben das Untersuchen der Gemeinsamkeiten im Verhalten.
(6) Sofern Ausführungsdaten für den Prozess in Form von Logs vorliegen, wird die Übereinstimmung mit dem modellierten Verhalten gemessen.

Im Folgendem werden die Schritte (3) - (5) des Ansatzes erläutert.

4.1 Konsistenzkritieren

Die Grundidee, der auf Verhaltensprofilen basierenden Konsistenzkriterien, lautet wie folgt: Die Verhaltensrelationen zweier Petri-Netze sollten für alle möglichen zwei Paare von korrespondierenden Transitionen übereinstimmen. So wird für das Beispiel in Abbildung 2 verlangt, dass die Exklusivität zwischen den Transitionen e und f in Modell I, auch für die entsprechenden Transitionspaare $(7, 9)$ und $(8, 9)$ in Modell II gilt. Das Beispiel zeigt

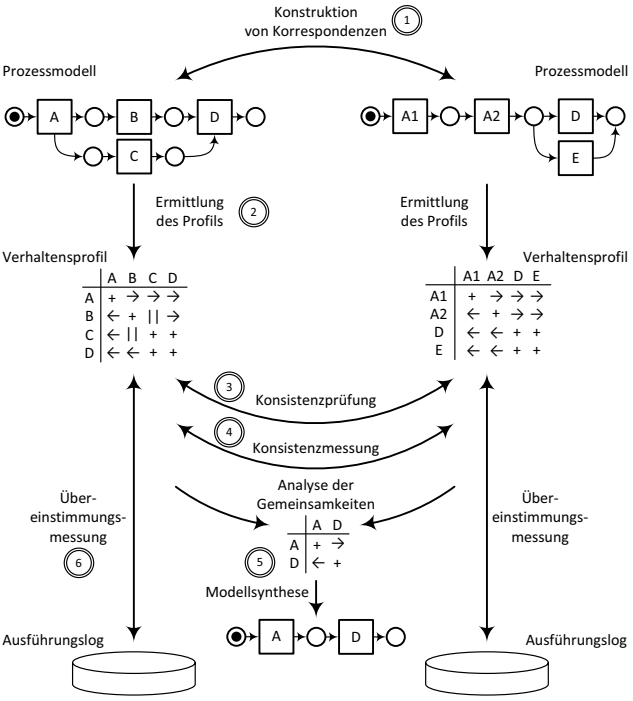


Abbildung 4: Überblick über die Konsistenzanalyse mit Verhaltensprofilen.

auch, dass der Begriff des Verhaltensprofiles besonders für die Analyse geeignet ist, da nur die Existenz von Ausführungsfolgen, nicht aber Kausalitäten betrachtet werden. Die unterschiedliche Abdeckung des Prozesses durch die Modelle führt zu unterschiedlichen Kausalitäten. In Modell I führt das Auftreten von Transition a immer zu einem Auftreten von Transition d. In Modell II hingegen gilt dies nicht für die Transitionspaare $(1, 6)$ und $(2, 6)$, da die Möglichkeit einer direkten Angebotsanfrage berücksichtigt wurde. Diese Unterschiede, welche aus unterschiedlichen Eintritts- oder Austrittspunkten des modellierten Prozesses resultieren, sind oft zwischen Prozessmodellen zu beobachten, welche unterschiedlichen Zwecken dienen. Das Verhaltensprofil erlaubt es davon zu abstrahieren und Konsistenz auf Basis der Ordnung des potentiellen Auftretens zu definieren. So gilt die strikte Ordnung zwischen Transitionen a und d in Modell I auch für die Paare $(1, 6)$ und $(2, 6)$ in Modell II.

In diesem Sinne wurde ein Spektrum von Konsistenzkriterien auf Verhaltensprofilen definiert. Die einzelnen Kriterien unterscheiden sich in der Auswahl der Transitionspaare und Relationen, welche konsistent sein müssen. So können zum Beispiele Unterschiede bzgl. der Wiederholbarkeit eines Auftretens einer Transition toleriert werden, wenn Transitionspaare der Identitätsrelation nicht überprüft werden. Im Rahmen einer empirischen Studie mit Prozessmodellierungsexperten zeigte sich, dass diese Konsistenzdefinition eine gute Approximation des Konsistenzempfindens der Experten aufweist [Wei11].

4.2 Konsistenzmaße

Der Abgleich der Verhaltensrelationen für Paare von korrespondierenden Transitionen ist auch die Grundlage von Konsistenzmaßen. Jene ergeben sich aus dem Verhältnis der Anzahl der Transitionspaare eines Modell, für welche alle korrespondierenden Transitionspaare konsistent sind und der Anzahl der Transitionspaare, für welche dieses nicht gilt. Auf Basis der bereits erwähnten unterschiedlichen Auswahl der betrachteten Transitionspaare und Relationen ergibt sich somit auch ein Spektrum von Konsistenzmaßen.

Die Konsistenzmaße erlauben es die Abweichungen im Verhalten zweier Petri-Netze zu quantifizieren. Für das Beispiel in Abbildung 2 wird nur eine geringe Abweichung gemessen und ein Konsistenzwert von ≈ 0.98 erreicht. Der einzige Unterschied in die Verhaltensprofilen ist, dass Transition *e* in Modell I potentiell mehrfach auftreten kann, was für Transitionen 7 und 8 in Modell II ausgeschlossen ist.

4.3 Analyse von Verhaltensgemeinsamkeiten

Um die mittels der Konsistenzkriterien und Maße erworbenen Ergebnisse interpretieren zu können, ist es hilfreich die Verhaltensgemeinsamkeiten und Unterschiede zu extrahieren. Auf Basis dessen kann entschieden werden, ob gewisse Abweichungen im Verhalten zweier Modelle akzeptabel sind. Im Rahmen des vorgestellten Ansatzes, wird diese Analyse durch eine Mengenalgebra für Verhaltensprofile realisiert. Jene verlangt eine Normalisierung von komplexen Korrespondenzen. Für eine Menge von Transitionen eines Modells, welche Teil einer komplexen Korrespondenz sind, wird die dominierende Verhaltensrelation zu allen anderen Transitionen des Netzes algorithmisch bestimmt.

Grundlage der Algebra ist eine Hierarchie der Relationen des Verhaltensprofils. Die Exklusivitätsrelation gilt als strikteste Relation, da sie das gemeinsame Auftreten zweier Transitionen ausschließt. Die Interleavingrelation hingegen gilt als schwächste Relation, da sie keine Einschränkung bzgl. des gemeinsamen Auftretens zweier Transitionen trifft. Auf Basis dieser Hierarchie definiert die Algebra für Verhaltensprofile die Relationen Äquivalenz, Enthaltensein und Leerheit, sowie die Operationen Komplement, Schnitt und Vereinigung. So ist ein Verhaltensprofil in einem anderen enthalten, genau dann, wenn alle Transitionspaare des ersten Profils Teil einer schwächeren Verhaltensrelation im Vergleich zu den korrespondierenden Transitionspaaren im zweiten Profil sind. Der Schnitt hingegen kombiniert die jeweils strikteste Verhaltensrelation für alle Transitionspaare. Die algebraischen Relationen und Operationen erlauben es zahlreiche Analysefragestellungen zu beantworten. Zum Beispiel können Schnitt und Vereinigung als größter gemeinsamer Teiler bzw. kleinstes gemeinsames Vielfaches des Verhaltens angesehen werden.

Die Analyse und Interpretation der mittels algebraischer Operationen erzeugten Verhaltensprofile wird durch eine Modellsynthese unterstützt. Jene erzeugt zu einem Verhaltensprofil ein entsprechendes Petri-Netz und lässt sich als Umkehroperation der in Abschnitt 3 beschriebenen Ermittlung eines Verhaltensprofils auffassen.

5 Verwandte Arbeiten

Verhaltenskonsistenz ist bereits aus einer Vielzahl von Perspektiven betrachtet worden. Verhaltensvererbung [BvdA01] zeigt, wie die für statische Modelle bekannten Vererbungskonzepte auf Verhaltensmodelle übertragen werden. Die Einhaltung von Konsistenz ist auch die Grundlage von Arbeiten über die Verfeinerung von Verhaltensmodellen [vGG01]. Diese Arbeiten haben gemeinsam, dass sie auf klassischen Verhaltensäquivalenzen basieren, so dass sie eine hohe Berechnungskomplexität aufweisen. Auch wird immer eine hierarchische Verfeinerungsrelation zwischen Prozessmodellen angenommen oder erzeugt. Dies ist als wesentliche Einschränkung zu sehen, wie das Zitat von Knöpfel et al. [KGT05] aufzeigt: '*nonhierarchical transformations are not rare exceptions. In the transition from high-level models of the application domain to the implementation model, we find them anywhere'*.

Verhaltensrelationen werden auch verwendet um Modelle aus Ausführungsdaten zu gewinnen (Process Mining) [vdA11]. Hierbei werden jedoch üblicherweise Relationen eingesetzt, welche die direkte Nachfolgebeziehung von Transitionen erfassen, z.B. im Alpha-Algorithmus [vdA11]. Sie sind somit weniger geeignet um Konsistenz von Modellen zu beurteilen, welche eine unterschiedliche Abdeckung eines Prozesses aufzeigen. Ähnliche Relationen, genannt Causal Footprints, wurden in [vDDM08] zur Ähnlichkeitsmessung verwendet. Causal Footprints haben jedoch den Nachteil, dass für ein Modell keine eindeutige Definition existiert.

6 Zusammenfassung & Ausblick

Dieser Beitrag hat einen Ansatz für die Analyse von Verhaltenskonsistenz vorgestellt. Auf Basis einer relationalen Verhaltensabstraktion, dem Verhaltensprofil, wurden Konsistenzbegriffe, Konsistenzmaße und ergänzende Analysetechniken diskutiert. Der Ansatz ist in Gänze in der Dissertation [Wei11] beschrieben. Jene Arbeit enthält weitere experimentelle Untersuchungen bzgl. der Anwendbarkeit der effizienten Ermittlungsalgorithmen und ihrem Laufzeitverhalten, sowie empirische Untersuchungen zur Eignung der Konsistenzkriterien.

Auch wenn die Definition von Verhaltensprofilen mit dem Ziel der Konsistenzanalyse erfolgte, so hat sich gezeigt, dass das Konzept eine grundsätzliche Bedeutung hat. So wurde es bereits für die indexbasierte Suche in Prozessmodelldatenbanken, die glossarbasierte Generierung von Benennungsvorschlägen für Modellelemente, das Erkennen von Datenanomalien in Prozessmodellen, sowie die Optimierung von komplexen Ereignisabfragen in Systemen der Komplexen Ereignisverarbeitung erfolgreich eingesetzt.

Literatur

[BvdA01] Twan Basten und Wil M. P. van der Aalst. Inheritance of behavior. *J. Log. Algebr. Program.*, 47(2):47–145, 2001.

[ES07] Jérôme Euzenat und Pavel Shvaiko. *Ontology matching*. Springer, 2007.

[JvdA09] Kurt Jensen und Wil M. P. van der Aalst, Hrsg. *Transactions on Petri Nets and Other Models of Concurrency (TOPNOC) II*, Jgg. 5460 of *LNCS*. Springer, 2009.

[KGT05] Andreas Knöpfel, Bernhard Gröne und Peter Tabeling. *Fundamental Modeling Concepts*. Wiley, 2005.

[Kin09] Ekkart Kindler. Model-Based Software Engineering and Process-Aware Information Systems. In *TOPNOC* [JvdA09], Seiten 27–45.

[Küh06] Thomas Kühne. Matters of (Meta-)Modeling. *Software and System Modeling*, 5(4):369–385, 2006.

[LPB99] J. Luftman, R. Papp und T. Brier. Enablers and inhibitors of business-IT alignment. *Communications of the AIS*, 1(3), 1999.

[LVD09] Niels Lohmann, Eric Verbeek und Remco M. Dijkman. Petri Net Transformations for Business Processes - A Survey. In *TOPNOC* [JvdA09], Seiten 46–63.

[RP00] Colette Rolland und Naveen Prakash. Bridging the Gap Between Organisational Needs and ERP Functionality. *Requir. Eng.*, 5(3):180–193, 2000.

[vdA11] Wil van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[vDDM08] Boudewijn F. van Dongen, Remco M. Dijkman und Jan Mendling. Measuring Similarity between Business Process Models. In *CAiSE*, Jgg. 5074 of *LNCS*, Seiten 450–464. Springer, 2008.

[vGG01] Rob J. van Glabbeek und Ursula Goltz. Refinement of actions and equivalence notions for concurrent systems. *Acta Inf.*, 37(4/5):229–327, 2001.

[Wei11] Matthias Weidlich. *Behavioural Profile - A relational approach to behaviour consistency*. Dissertation, Hasso Plattner Institut, Universität Potsdam, 2011.

[Wes07] Mathias Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer, 2007.

[Zel95] Stephan Zelewski. Petrinetzbasierte Modellierung komplexer Produktionssysteme. Band 9: Beurteilung des Petrinetz-Konzepts. Bericht, Universität Leipzig, 1995.

Matthias Weidlich ist Postdoktorand und außerordentlicher Dozent am Technion – Israel Institute of Technology in Haifa, Israel. Von 2003 bis 2008 studierte er Softwaresystemtechnik am Hasso Plattner Institut (HPI) an der Universität Potsdam. Während des Studiums verbrachte er Auslandssemester an der EFREI in Paris, Frankreich, sowie bei SAP Research in Brisbane, Australien. 2008 schloss er das Studium als Master of Science mit Auszeichnung ab. Seit Mai 2008 war er als Wissenschaftlicher Mitarbeiter am HPI an der Universität Potsdam tätig. Dort wurde er im November 2011 mit dem Prädikat ‘summa cum laude’ promoviert. Forschungsschwerpunkte seiner Arbeit sind die formale Analyse von Prozessmodellen und Datenheterogenität von Informationssystemen. Seine Forschungsresultate sind in internationalen Fachzeitschriften erschienen (u.a. IEEE Transactions on Software Engineering, Information Systems, The Computer Journal, Acta Informatica). Auf der 8th International Conference on Service Oriented Computing (ICSOC 2010) wurde er mit dem Best Paper Award ausgezeichnet.

Symbolische Methoden für die probabilistische Verifikation
– Zustandsraumreduktion und Gegenbeispiele –

Ralf Wimmer
Lehrstuhl für Rechnerarchitektur, Technische Fakultät
Albert-Ludwigs-Universität Freiburg
Georges-Köhler-Allee 51, 79110 Freiburg im Breisgau
wimmer@informatik.uni-freiburg.de

Abstract: Ein bekanntes Hindernis für die formale Verifikation von Systemen bildet die potentiell stark anwachsende Größe des Zustandsraums, genannt „Zustandsraumexplosion". Dieses Problem konnte für digitale Schaltungen durch den Einsatz symbolischer Methoden zufriedenstellend gelöst oder zumindest entscheidend entschärft werden. Für probabilistische Systeme, die als Markow-Kette oder Markow-Entscheidungsprozess modelliert sind, brachte die direkte Übertragung dieser symbolischen Methoden bisher keinen Durchbruch.

In dieser Arbeit stellen wir zwei neue Ansätze vor, mit denen Markow-Modelle mit sehr großen Zustandsräumen verifiziert werden können. Die erste Methode ist ein symbolisches Verfahren zur Vorverarbeitung: Zu jedem Markow-Modell berechnen wir mit rein symbolischen Verfahren das kleinste Modell, das in den interessierenden Eigenschaften mit dem Original-Modell übereinstimmt. Die Verifikation kann danach auf dem minimierten Modell durchgeführt werden. Das zweite offene Problem, das in der Dissertation gelöst wird, ist die symbolische Berechnung von Gegenbeispielen, wenn Sicherheitseigenschaften von Markow-Ketten mit diskreter Zeit verletzt sind. Anhand von Experimenten wird gezeigt, dass die neu entwickelten Verfahren den bisher verfügbaren Verfahren hinsichtlich der Laufzeit bzw. der Größe der handhabbaren Systeme deutlich überlegen sind.

1 Einführung

Der Einsatz von computergesteuerten Systemen in sicherheitskritischen Umgebungen (wie beispielsweise in Fahrzeugen oder für medizinische Zwecke) und ihre gleichzeitig immer größer werdende Komplexität machen es unerlässlich, die zuverlässige und korrekte Funktionsweise des jeweiligen Systems sicherzustellen. Allerdings haben praktisch relevante Systeme längst eine Komplexität erreicht, die es unmöglich macht, mehr als einen verschwindend geringen Anteil aller Ablaufmöglichkeiten durch Simulation zu überprüfen. Deshalb wurden formale Methoden, sogenannte Model-Checking-Algorithmen, entwickelt, mit denen man Systeme automatisch auf vorgegebene Eigenschaften prüfen kann. Dabei wird auf clevere Weise der gesamte Zustandsraum eines Systems systematisch durchsucht, so dass damit – im Gegensatz zur Simulation – auch die Fehlerfreiheit bewiesen werden kann. Während erste Verfahren noch auf kleine Systeme mit relativ wenigen Zuständen beschränkt waren, verhalfen symbolische Methoden dem Model Checking für digitale Schaltungen zum Durchbruch [BCM$^+$92]. Symbolische Methoden stellen den Zustandsraum eines Systems nicht durch Aufzählen der Zustände und der Übergänge dazwischen dar, sondern beispielsweise als Lösungen einer Formel oder als Pfade in einem Entschei-

dungsdiagramm (engl. Ordered Binary Decision Diagram, OBDD). Dieser sind in vielen Fällen deutlich kompakter als die Aufzählung aller Zustände und Zustandsübergänge.

Die Herausforderungen der neuesten Zeit ergeben sich durch den Einsatz von Mikroprozessoren in eingebetteten Systemen, die kontinuierliche Größen messen und diese digital verarbeiten (sogenannte hybride Systeme) oder die in unsicheren Umgebungen arbeiten: So werden z. B. Nachrichten über unzuverlässige Kanäle verschickt und können verloren gehen, randomisierte Protokolle werden eingesetzt und Daten kommen mit zeitlichen Schwankungen an. Solche stochastischen Systeme werden oft als Markow-Ketten oder Markow-Entscheidungsprozesse modelliert. Um für derartige Systeme Fragen nach der Zuverlässigkeit oder Verfügbarkeit beantworten zu können, bildete in den letzten zwei Jahrzehnten die Entwicklung von Model-Checking-Algorithmen für stochastische Systeme (siehe z. B. [HJ94, BHHK03]) einen Schwerpunkt in der Verifikationsforschung.

Ein Problem, das die Anwendung dieser Algorithmen auf reale Systeme schwierig macht, ist deren Anzahl an Zuständen, die im Allgemeinen exponentiell in der Zahl der Systemkomponenten wächst. Die symbolischen Methoden, die OBDDs zur Darstellung der Zustandsräume verwenden, wurden zum Teil auf Markow-Modelle verallgemeinert, skalierten jedoch nicht im selben Maße für Markow-Ketten wie für Schaltkreise. Für einige Modellklassen wie beispielsweise interaktive Markow-Ketten (IMCs) oder Markow-Entscheidungsprozesse mit kontinuierlicher Zeit (CTMDPs) sind bis heute keine symbolischen Verfahren verfügbar. Darüber hinaus ist es generell nicht möglich, mit den üblichen Verfahren für Markow-Ketten gleichzeitig zum Ergebnis der Eigenschaftsprüfung ein Gegenbeispiel zu erhalten, wenn eine gewünschte Eigenschaft verletzt ist. Zwar wurden im Wesentlichen seit 2003 verschiedene Verfahren vorgeschlagen, mit denen für Markow-Ketten Gegenbeispiele erzeugt werden können (beispielsweise [HKD09, AL10]), allerdings beruhen all diese Verfahren auf einer expliziten Darstellung des Zustandsraums und sind daher auf verhältnismäßig kleine Systeme beschränkt; gut skalierende symbolische Verfahren waren bis jetzt nicht verfügbar.

Das Ziel der Dissertation war folglich, symbolische Verfahren zu entwickeln, die zum einen die Eigenschaftsprüfung für größere Systeme ermöglichen und es zum anderen gestatten, für Markow-Modelle mit sehr großen Zustandsräumen Gegenbeispiele zu erzeugen. Entsprechend ist die Dissertation in zwei Teile gegliedert: Im ersten Teil wird ein Verfahren zur *symbolischen Minimierung* einer ganzen Reihe von Markow-Modellen vorgestellt. Dabei können verschiedene Klassen von interessierenden Eigenschaften im minimierten System erhalten bleiben: Beispielsweise erhält man unterschiedliche minimale Systeme abhängig davon, ob die Wahrscheinlichkeit, innerhalb einer vorgegebenen Anzahl von Schritten eine Menge von Zuständen zu erreichen, erhalten bleiben soll oder lediglich die Wahrscheinlichkeit, irgendwann eine solche Zustandsmenge zu erreichen (unabhängig von der Anzahl der Schritte). Das formale Mittel für die Minimierung sind Äquivalenzrelationen auf dem Zustandsraum eines Systems, die *Bisimulationen* genannt werden. Dabei werden Zustände, die schrittweise dasselbe beobachtbare Verhalten zeigen, als äquivalent angesehen. Indem man die gröbste Bisimulation berechnet und dann zum Quotientensystem übergeht, dessen Zustände gerade den Äquivalenzklassen der Bisimulation entsprechen, erhält man das kleinste System, das dasselbe beobachtbare Verhalten wie das Originalsystem zeigt. Nach der Minimierung kann deshalb die vorgegebene Eigenschaft auf dem minimierten System überprüft werden, das in vielen Fällen deutlich kleiner als das Originalsystem ist. Der

vorgestellte Algorithmus bildet ein einheitliches Framework, das nicht nur verschiedene Systemklassen – beschriftete Transitionssysteme, Markow-Ketten mit diskreter und kontinuierlicher Zeit sowie interaktive Markow-Ketten – minimieren, sondern auch diverse Minimierungskriterien berücksichtigen kann und leicht auf weitere Kriterien erweiterbar ist. Es wird sowohl die Laufzeit als auch der Speicherplatzbedarf optimiert, und es wird eine umfangreiche Anwendung auf die Analyse sicherheitskritischer Systeme vorgestellt, für die der entwickelte Minimierungsalgorithmus eine erfolgreiche Analyse ermöglicht.

Im zweiten Teil der Arbeit wird gezeigt, wie mit symbolischen Methoden *Gegenbeispiele* für Markow-Ketten mit diskreter Zeit (DTMCs) berechnet werden können. Wir erweitern dazu eine Methode namens *Bounded Model Checking* (BMC) [BCC+03], die bereits industriell sehr erfolgreich zur Fehlersuche in digitalen Schaltungen eingesetzt wird. Dabei wird die Existenz von Pfaden vorgegebener Länge, die eine Sicherheitseigenschaft verletzen, als propositionales Erfüllbarkeitsproblem formuliert. Jede Lösung des Problems entspricht genau einem Systemablauf, der zu einem Fehler führt. Während für digitale Schaltungen in der Regel ein einzelner solcher Ablauf ausreicht, um eine Sicherheitseigenschaft („Es wird nie ein sicherheitskritischer Zustand erreicht“) zu widerlegen, sind bei Sicherheitseigenschaften für DTMCs („Die Wahrscheinlichkeit, einen sicherheitskritischen Zustand zu erreichen, ist höchstens λ“) Mengen von Abläufen nötig, deren gemeinsame Wahrscheinlichkeitsmasse die vorgegebene Schranke λ überschreitet. Es wird beschrieben, wie man BMC auf DTMCs anwenden kann, um effizient ein kompaktes Gegenbeispiel zu erzeugen. Experimente zeigen, dass das resultierende Verfahren in der Lage ist, deutlich größere Systeme zu verarbeiten als die bisherigen Verfahren.

Im Folgenden werden kurz die wichtigsten Verfahren und Ergebnisse, die im Rahmen dieser Dissertation entwickelt wurden, zusammengefasst. Für die Details wird auf die Dissertation und die Konferenz- und Zeitschriftenbeiträge, in denen die Ergebnisse publiziert wurden, verwiesen.

2 Symbolische Zustandsraumreduktion

In diesem Abschnitt stellen wir das neu entwickelte symbolische Minimierungsverfahren vor, das den ersten Teil der Dissertation bildet. Grundlage dafür bildet ein Algorithmus von Blom und Orzan [BO05a, BO05b], der die Minimierung von beschrifteten Transitionssystemen (LTSs) bezüglich starker und branching Bisimulation gestattet. Man beginnt mit einer initialen Partition des Zustandsraums, die entweder durch Zustandsbeschriftungen vorgegeben sein kann oder andernfalls einen einzelnen Block mit allen Zuständen enthält. Der Algorithmus beruht darauf, alle Zustände anhand einer Signatur so zu charakterisieren, dass Zustände mit verschiedenen Signaturen nicht äquivalent sein können. Eine neue Einteilung der Zustände in Klassen erfolgt durch Aufspalten der Blöcke der aktuellen Partition gemäß den Signaturen. Man iteriert diesen Vorgang so lange, bis ein Fixpunkt erreicht ist. Das Ergebnis ist die gröbste starke bzw. branching Bisimulation, welche die Anfangspartition verfeinert. Dieser Algorithmus wurde für den Einsatz in einer verteilten Umgebung entwickelt und verwendet explizite Darstellungen des Zustandsraums.

Zunächst wird in Kapitel 3 der ursprüngliche Algorithmus dahingehend erweitert, dass er nicht nur in der Lage ist, die starke und die branching Bisimulation für LTSs zu ver-

wenden, sondern im Wesentlichen alle Minimierungskriterien, die in der Literatur eine Rolle spielen, nämlich zusätzlich die schwache, orthogonale, safety-, η-, delay- und progressing Bisimulation [WHH+06]. Außerdem kann bei der Minimierung unterschiedliches Divergenzverhalten der Zustände berücksichtigt werden, sofern dies nicht bereits durch die Definition der Bisimulation erfolgt. Die Dissertation liefert einen Beweis für die Korrektheit des Verfahrens für all diese Bisimulationstypen. Das Verfahren kann bei Bedarf leicht um weitere Bisimulationen erweitert werden, indem jeweils eine geeignete Signatur formuliert wird.

In dieser Form verwendet der Minimierungsalgorithmus immer noch explizite Darstellungen der Zustandsräume und ist dadurch auf Systeme beschränkt, die klein genug sind, um in den Hauptspeicher zu passen. Er wird nun so modifiziert, dass er ausschließlich mit symbolischen Datenstrukturen in Form von OBDDs arbeitet. Dazu werden Methoden entwickelt, bei denen die Laufzeit der Berechnungen nicht mehr direkt von der Größe des dargestellten Systems abhängt, sondern nur noch von der Größe der Darstellung. Letztere kann – und ist es auch in vielen praktischen Fällen – sehr viel kleiner sein als eine explizite Darstellung. Den Kern des symbolischen Algorithmus bildet dabei eine geschickte Partitionsdarstellung, die eine effiziente Berechnung der Signaturen und der Verfeinerung der aktuellen Partition gestattet. Eine ganze Reihe von Optimierungen – zum Beispiel auf die Partitionsdarstellung angepasste BDD-Operationen zum Zugriff auf die Blöcke der aktuellen Partition; das Auslassen von Blöcken der Partition bei der Verfeinerung, von denen festgestellt werden kann, dass sie in der aktuellen Iteration nicht aufgespaltet werden können; und die Anpassung der Signaturen für eine effizientere Berechnung – reduzieren die Laufzeit beträchtlich. Dadurch ergibt sich ein flexibles Verfahren zur Minimierung von LTSs, das durch den Einsatz symbolischer Datenstrukturen mit deutlich größeren Systemen umgehen kann als Konkurrenzverfahren und das hinsichtlich der Laufzeit konkurrierenden symbolischen Verfahren wie dem von Bouali und de Simone [Bd92] weit überlegen ist.

Während in Kapitel 3 Verfahren für beschriftete Transitionssysteme entwickelt wurden, die keine stochastischen Transitionen besitzen, folgt in Kapitel 4 die Erweiterung der Techniken auf Markow-Ketten mit diskreter (DTMCs) und mit kontinuierlicher Zeit (CTMCs) sowie interaktive Markow-Ketten (IMCs), welche die stochastischen Übergänge der CTMCs mit den interaktiven Transitionen der LTSs kombinieren. Es werden geeignete Signaturen definiert für alle drei stochastischen Systemklassen, und die symbolische Verfeinerung wird so angepasst, dass sie im Falle von IMCs mit Paaren von Signaturen arbeitet. Schwierigkeiten bereiten dabei numerische Instabilitäten [WB10]. Aufgrund von Rundungsfehlern, die bei Verwendung der üblichen Gleitkommaarithmetik auftreten, ergeben sich in den Signaturen von eigentlich äquivalenten Zuständen kleine Unterschiede, die dazu führen, dass sie in unterschiedliche Äquivalenzklassen gelangen. Als Lösungsmöglichkeiten entwickelten wir neben einer Implementierung, die mit rationaler Arithmetik arbeitet, eine weitere Version, bei der die Übergangsraten zwischen Zuständen als reine Symbole behandelt, d. h. nicht als Zahlen interpretiert werden. Dadurch werden nur noch natürliche Zahlen zum Zählen der Kanten mit gleichem Symbol verwendet, was ohne Rundungsprobleme erfolgen kann. Dadurch erhält man als Ergebnis die kleinste Markow-Kette, deren Verhalten für alle möglichen Wahlen der Ratensymbole mit der ursprünglichen Markow-Kette übereinstimmt. Der große Vorteil dabei ist, dass nach der Minimierung die konkreten Werte der Symbole beliebig angepasst werden können, ohne den Quotienten neu berechnen zu müssen. Benötigt man für die konkreten Werte das minimale System, kann dies durch erneutes Minimieren

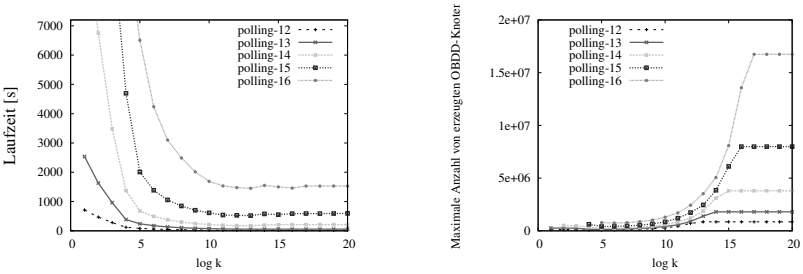


Abbildung 1: Abhängigkeit von Laufzeit (links) und Speicherplatz (rechts) des hybriden Verfahrens von der Zahl k der gleichzeitig verfeinerten Blöcke

des interpretationsunabhängigen Quotientensystems erhalten werden.

Experimente mit unserem Tool SIGREF und verschiedenen Fallstudien zeigten Folgendes: (1) Bei Verwendung von Gleitkommaarithmetik werden in vielen Fällen zu feine Quotientensysteme berechnet und in einem Fall terminierte sogar die Minimierung wegen der numerischen Probleme nicht. (2) Die Verwendung rationaler Arithmetik ist nur unwesentlich teurer als Gleitkommaarithmetik, da der Hauptteil der Rechenzeit nicht für arithmetische Operationen verwendet wird, sondern beispielsweise für Cache-Lookups, auf welche die rationale Arithmetik keinen Einfluss hat. (3) Die Behandlung der Raten als reine Symbole erzeugte in den untersuchten Fällen interpretationsunabhängige Quotienten, die nicht wesentlich größer sind als die mit festen Werten berechneten Systeme. Der Größenunterschied betrug in der Regel weniger als 10 %. Auch die Laufzeit ist mit den Zeiten der beiden anderen Varianten vergleichbar. Damit sind die numerischen Probleme ohne wesentliche Laufzeit- und Speicherplatzeinbußen gelöst.

Kapitel 5 optimiert den Speicherverbrauch bei der Minimierung. Derisavi stellte einen alternativen symbolischen Algorithmus zur Minimierung von CTMCs vor, der eine andere Partitionsdarstellung verwendet [Der07]. Diese besteht aus maximal $\lceil \log_2 n \rceil$ OBDDs, wobei n die Zahl der Zustände der CTMC ist. Experimente haben gezeigt, dass der Speicherverbrauch dieses Algorithmus in vielen Fällen deutlich geringer ist der als von SIGREF; dafür ist SIGREFs Laufzeit um Größenordnungen kleiner als die von Derisavis Verfahren. Aus diesem Grund haben wir in Kooperation mit Derisavi ein hybrides Verfahren entwickelt, das die Vorteile beider Algorithmen vereint [WDH10]. Die Idee ist, die gesamte Partition in der kompakten Darstellung von Derisavi zu halten. Daraus werden Gruppen von k Blöcken – k ist ein Parameter – extrahiert, in SIGREFs effizient zu verfeinerndes Format gebracht, verfeinert und zurück in die kompakte Darstellung gebracht. Die verfeinerten Blöcke werden durch das Ergebnis der Verfeinerung ersetzt. Dies wird wiederholt, bis sich keine Änderungen der Partition mehr ergeben. Die Wahl von k bestimmt, wie viel Speicherplatz und Laufzeit benötigt werden (siehe Abbildung 1). Je kleiner k ist, d. h. je weniger Blöcke in einem Schritt verfeinert werden, desto geringer ist der Platzbedarf und desto größer die Laufzeit. Für $k = 1$ enthält man im Wesentlichen Derisavis Verfahren, für $k = n$ unser signaturbasiertes Verfahren von SIGREF. Damit kann der schon den Speicherplatz reduziert werden, aber das Ziel ist, den Wert von k automatisch so einzustellen, dass die Laufzeit minimal ist, ohne dass der verfügbare Speicherplatz überschritten wird. Dazu wird folgende Strategie verwendet: Man beginnt mit dem SIGREF-Verfeinerungsverfahren, bis der zur

Verfügung stehende Speicherplatz nicht mehr ausreicht. Dann wird die zuletzt erfolgreich verfeinerte Partition gesichert und die aktuelle Iteration abgebrochen. Der Wert von k wird nun so gewählt, dass man zwei Gruppen von Blöcken erhält. Die Verfeinerung wird mit dem hybriden Verfahren und der zuletzt erfolgreich berechneten Partition fortgesetzt. Jedesmal, wenn jetzt zu viel Speicherplatz benötigt wird, wird die aktuelle Iteration abgebrochen, der Wert von k halbiert und die Berechnung mit der zuletzt abgespeicherten Partition fortgesetzt. Falls bei $k = 1$ der verfügbare Speicher nicht ausreicht, kann die Minimierung mit dem zur Verfügung stehenden Speicher nicht durchgeführt werden. Die Ergebnisse zeigen, dass für kleine CTMCs, bei denen der Speicherverbrauch keine Einschränkung darstellt, die Effizienz von SIGREF erreicht wird. Bei großen Modellen jedoch, bei denen die Minimierung mit dem SIGREF-Algorithmus scheitert, ist der hybride Algorithmus in der Lage, die Minimierung durchzuführen. Dabei ist er z. T. um Größenordnungen schneller als der Algorithmus von Derisavi. Damit ist das Ziel erreicht, die Vorteile beider Algorithmen – die Laufzeiteffizienz von SIGREF und die Speichereffizienz von Derisavis Algorithmus – zu kombinieren.

Der erste Teil der Arbeit endet in Kapitel 6 mit einer Anwendung der symbolischen Bisimulationsminimierung, in der für industrielle Statechart-Modelle Wahrscheinlichkeiten für die zeitbeschränkte Erreichbarkeit von sicherheitskritischen Zuständen [BHH$^+$09] berechnet werden. Statecharts sind ein weit verbreiterter Formalismus zur Systemmodellierung und als Teil von UML 2 standardisiert. Da Statecharts an sich keine stochastischen Informationen enthalten, wird es dem Benutzer ermöglicht, Ereignisse zu markieren, die mit einer zeitlichen Verteilung auftreten. Im nächsten Schritt wird aus dem Statechart-Modell ein symbolisch dargestelltes LTS erzeugt, dessen Transitionsbeschriftungen gerade den markierten Ereignissen entsprechen. Die übrigen Ereignisse werden als nicht beobachtbar angesehen. Dieses Transitionssystem wird mit Hilfe unseres symbolischen Algorithmus verkleinert, indem wir den Quotienten der branching Bisimulation berechnen. Nun reichert man das minimierte Transitionssystem um die Zeitinformation an. Das Ergebnis ist eine interaktive Markow-Kette. Im Allgemeinen ist sie zu groß für die Weiterverarbeitung. Deshalb folgt im Anschluss nochmals eine Minimierungsphase – diesmal mit der branching Bisimulation für IMCs. Zum Zeitpunkt, als diese Arbeit entstanden ist, war noch kein Verfahren – weder ein symbolisches noch ein explizites – verfügbar, um direkt auf IMCs Erreichbarkeitswahrscheinlichkeiten zu berechnen (dieses Problem wurde 2010 von Zhang und Neuhäußer [ZN10] gelöst). Deshalb verwenden wir eine Transformation in uniforme Markow-Entscheidungsprozesse mit kontinuierlicher Zeit (CTMDPs), welche die Erreichbarkeitswahrscheinlichkeiten erhält, und berechnen diese anschließend mit dem Algorithmus von Baier et al. [BHKH05]. Durch diesen Ablauf gelingt es, Fragen wie z. B. „Wie groß ist die Wahrscheinlichkeit, innerhalb von 2 Stunden in einen sicherheitskritischen Zustand zu gelangen?“ für industriell relevante Modelle effizient zu beantworten.

3 Gegenbeispiele für Markow-Ketten

Im zweiten Teil der Dissertation wird das Problem gelöst, Gegenbeispiele für DTMCs mit sehr vielen Zuständen mit symbolischen Methoden zu berechnen. Es wird dabei angenommen, dass eine Sicherheitseigenschaft der Form „Die Wahrscheinlichkeit, einen sicherheitskritischen Zustand zu erreichen, ist höchstens λ” verletzt ist. Model Checking für

DTMCs wird in der Regel auf das Lösen eines linearen Gleichungssystems zurückgeführt, das gerade die gesuchte Wahrscheinlichkeit ergibt. Deshalb erhält man bei DTMCs nicht automatisch ein Gegenbeispiel, wenn eine Sicherheitseigenschaft verletzt ist. Gegenbeispiele sind jedoch von zentraler Bedeutung für die Korrektur fehlerhafter Systeme oder für die abstraktionsbasierte Verifikation, bei der eine zu grobe Abstraktion mit Hilfe von Gegenbeispielen an geeigneten Stellen verfeinert wird. Bisherige Verfahren zur Erzeugung von Gegenbeispielen für DTMCs beruhten auf Algorithmen zur Berechnung kürzester Wege in Graphen [HKD09] oder heuristischen Suchverfahren [AL10]. Diese setzen jedoch alle explizit dargestellte Zustandsräume voraus und sind daher auf verhältnismäßig kleine Systeme beschränkt.

Wir entwickeln ein symbolisches Verfahren zur Berechnung von Gegenbeispielen. Grundlage dafür ist das aus der Verifikation von Schaltkreisen bekannte Bounded Model Checking (BMC) [BCC$^+$03]. Dabei wird die Existenz von Pfaden einer festen Länge, die eine Sicherheitseigenschaft verletzen, als logische Formel beschrieben. Deren Erfüllbarkeit wird mit Hilfe eines geeigneten Solvers geprüft. Jede erfüllende Belegung der Formel entspricht genau einem Ablauf, der vom Anfangszustand des Systems zu einem sicherheitskritischen Zustand führt. Während ein einzelner solcher Pfad bei Schaltkreisen als Gegenbeispiel ausreicht, ist bei einer DTMC eine Menge von derartigen Pfaden notwendig, so dass ihre gemeinsame Wahrscheinlichkeitsmasse die Schranke λ überschreitet.

In Kapitel 9 wird gezeigt, wie das klassische BMC-Verfahren für Schaltkreise erweitert werden kann, so dass Pfadmengen als Gegenbeispiele für DTMCs effizient berechnet werden können [WBB09]. Zunächst muss eine Darstellung der Transitionsrelation als Erfüllbarkeitsproblem erzeugt werden. Da die Entscheidung, ob eine solche Formel erfüllbar ist, ein NP-hartes Problem ist, ist es von zentraler Bedeutung, eine möglichst kompakte Formel zu erzeugen. Dazu gehen wir von einer symbolischen Darstellung des Zustandsraums aus, die in Form eines Entscheidungsdiagramms gegeben ist, wie sie beispielsweise der Model Checker PRISM [HKNP06] erzeugt. Um die Darstellung zu verkleinern, werden zunächst die genauen Transitionswahrscheinlichkeiten ignoriert, und es werden Techniken zur Verkleinerung von OBDDs wie Sifting und Don't-Care-Minimierung angewendet. Mit Hilfe der Tseitin-Transformation erhält man aus dem OBDD eine Formel für die Transitionsrelation, deren Länge linear in der Größe des OBDDs ist. Durch k-faches Abrollen des Systems und Erweiterung um Formeln, die den Anfangszustand bzw. die sicherheitskritischen Zustände beschreiben, erzeugen wir eine Formel, deren erfüllende Belegungen genau den Pfaden der Länge k entsprechen, die vom Anfangszustand zu einem sicherheitskritischen Zustand führen.

Ein Gegenbeispiel wird nun folgendermaßen erzeugt: Man beginnt mit Pfadlänge $k = 0$ und wiederholt die folgenden Schritte solange, bis die Wahrscheinlichkeitsmasse der gefundenen Pfade die Grenze λ übersteigt. Man prüft die BMC-Formel auf Erfüllbarkeit; ist sie unerfüllbar, erhöht man die Pfadlänge um eins. Gibt es eine erfüllende Belegung, entspricht diese einem neuen Pfad, den man der Pfadmenge hinzufügt. Man schließt den abgearbeiteten Pfad aus dem Suchraum des Solvers aus und startet den Suchprozess erneut. Falls die Sicherheitseigenschaft verletzt ist, terminiert diese Prozedur nach endlich vielen Schritten. Eine Verbesserungsmöglichkeit ergibt sich aus folgender Beobachtung: Enthält das System Schleifen, können diese beliebig oft durchlaufen werden. Indem Gegenbeispiele nicht als einfache lineare Pfade dargestellt werden, sondern als azyklische Pfade, deren

Zustände annotiert sind mit Schleifen, kann man die Zahl der Pfade, die nötig sind, um genügend Wahrscheinlichkeitsmasse zu erhalten, in vielen Fällen stark verringern und gleichzeitig die Laufzeit verkleinern, da wir Pfade, die durch mehrfaches Abwickeln von Schleifen entstehen, von vorn herein aus dem Suchraum ausschließen können.

In Kapitel 10 evaluieren wir das Tool SBMC, das das beschriebene Verfahren implementiert, anhand einiger Fallstudien und vergleichen es mit dem expliziten Verfahren von Han et al. [HKD09]. Es zeigte sich, dass die Laufzeiten für alle Benchmarks, die beide Verfahren verarbeiten konnten, vergleichbar sind. Jedoch kann SBMC auf deutlich größere Systeme angewendet werden, als dies für das Vergleichsverfahren aufgrund seines Speicherverbrauchs möglich war.

Zum Abschluss der Arbeit folgt in Kapitel 11 eine Übersicht über Ideen, die in Zukunft weiter verfolgt werden und zum Teil inzwischen wurden. Zum einen handelt es sich dabei um weitere Optimierungen des BMC-Verfahrens für DTMCs, zum anderen um Erweiterungen auf allgemeinere Systemtypen wie Markow-Reward-Modelle.

Zu den Optimierungen gehört die Erzeugung von beliebig verschachtelten regulären Ausdrücken für die Pfadmengen, wie sie von Han et al. als Repräsentation von Gegenbeispielen vorgeschlagen wurden [HKD09]. Dadurch kann eine weitere Reduktion der Größe der Gegenbeispiele gegenüber unserem implementierten Ansatz mit azyklischen Pfaden, die mit einfachen Schleifen annotiert sind, erreicht werden.

Bisher haben wir die konkreten Übergangswahrscheinlichkeiten ignoriert und angenommen, dass kürzere Pfade in der Regel für den Benutzer zur Fehlersuche nützlicher sind als längere. Man kann jedoch auch die Wahrscheinlichkeit der Pfade als Optimierungskriterium verwenden, indem man anstelle eines rein propositionalen Erfüllbarkeitsproblems ein sogenanntes SMT-Problem erzeugt, bei dem neben booleschen Atomen lineare Ungleichungen über reellen Variablen in der Formel vorkommen. Damit lässt sich eine Formel konstruieren, die genau für diejenigen Pfade der Länge k erfüllt ist, die zu einem kritischen Zustand führen und eine vorgegebene Mindestwahrscheinlichkeit p haben. Von Systemen, die Pfade zu einem sicherheitskritischen Zustand mit sehr unterschiedlichen Wahrscheinlichkeiten enthalten, erhoffen wir uns durch diese Technik bessere Gegenbeispiele und eine Reduktion der Laufzeit, da insgesamt weniger Pfade benötigt werden.

Weiterhin lässt sich die Erzeugung von Gegenbeispielen mit der Bisimulationsminimierung kombinieren. Dabei wendet man zunächst die symbolische Minimierung mittels starker Bisimulation auf das System an. Dadurch reduziert sich in den meisten Fällen die Anzahl der Zustände deutlich. Für das minimierte System berechnet man dann mittels BMC ein Gegenbeispiel. Ein Pfad im minimierten System entspricht einer Folge von Äquivalenzklassen im Originalsystem. Man kann effizient das Gegenbeispiel für das minimierte System zurückübersetzen in ein Gegenbeispiel für das Originalsystem. In vielen Fällen wird dies für die Fehlerkorrektur nicht nötig sein, sondern es dürfte genügen, von allen schrittweise äquivalenten Pfaden einen Repräsentanten zur Verfügung zu stellen. Dadurch lässt sich das Gegenbeispiel weiter verkleinern.

Mit demselben Ansatz, mit dem man Pfade mit höherer Wahrscheinlichkeit bevorzugen kann, ist man auch in der Lage, Markow-Reward-Modelle zu behandeln. Bei diesen sind Zustände und/oder Transitionen einer DTMC um Kosten bzw. Belohnungen erweitert. Es werden Eigenschaften der Form „Die Wahrscheinlichkeit, dass das Erreichen eines Zustands

Kosten größer als c verursacht, ist höchstens λ“ betrachtet. Gegenüber den DTMCs müssen die gefundenen Pfade dahingehend eingeschränkt werden, dass sie Kosten $> c$ verursachen müssen. Dies lässt sich direkt in das erzeugte SMT-Problem integrieren.

Insgesamt ist das BMC-Verfahren ein mächtiges und flexibles Werkzeug, um effizient Gegenbeispiele für Markow-Ketten mit sehr großem Zustandsraum zu erzeugen. Durch die vorgeschlagenen Verbesserungen und Erweiterungen wird sich seine Laufzeit noch weiter reduzieren und die unterstützte Modellklasse erweitern lassen.

4 Zusammenfassung und Ausblick

In der Dissertation haben wir zwei Probleme bei der Verifikation stochastischer Systeme gelöst: Im ersten Teil haben wir ein symbolisches *Minimierungsverfahren* vorgestellt, das zu einem Markow-Modell das kleinste berechnet, das in den interessierenden Eigenschaften mit dem ursprünglichen übereinstimmt. Im zweiten Teil haben wir gezeigt, wie man mit Hilfe von Bounded Model Checking effizient *Gegenbeispiele* für DTMCs berechnen kann. Dadurch, dass beide Verfahren symbolische Datenstrukturen verwenden und dafür optimiert sind, sind sie insbesondere auch auf sehr große Systeme anwendbar.

Das Tool SIGREF ist dabei, sich zu einem Standardwerkzeug für die symbolische Minimierung zu entwickeln und fand bereits Eingang in mehrere Anwendungen in der probabilistischen Verifikation. Das Werkzeug zur Generierung von Gegenbeispielen wird aktiv weiterentwickelt. Im Moment werden die in Kapitel 11 beschriebenen Optimierungen und Erweiterungen integriert, um das Tool SBMC noch mächtiger zu machen. Es ist geplant, SBMC beispielsweise für die gegenbeispielgesteuerte Abstraktionsverfeinerung (CEGAR) einzusetzen.

Literatur

[AL10] Husain Aljazzar und Stefan Leue. Directed Explicit State-Space Search in the Generation of Counterexamples for Stochastic Model Checking. *IEEE Trans. on Software Engineering*, 36(1):37–60, 2010.

[BCC+03] Armin Biere, Alessandro Cimatti, Edmund M. Clarke, Ofer Strichman und Yunshan Zhu. Bounded Model Checking. *Advances in Computers*, 58:118–149, 2003.

[BCM+92] Jerry R. Burch, Edmund M. Clarke, Kenneth L. McMillan, David L. Dill und L. J. Hwang. Symbolic Model Checking: 10^{20} States and Beyond. *Information and Computation*, 98(2):142–170, 1992.

[Bd92] Amar Bouali und Robert de Simone. Symbolic Bisimulation Minimisation. In *Proc. of CAV*, Band 663 von LNCS, Seiten 96–108. Springer, 1992.

[BHH+09] Eckard Böde, Marc Herbstritt, Holger Hermanns, Sven Johr, Thomas Peikenkamp, Reza Pulungan, Jan Rakow, Ralf Wimmer und Bernd Becker. Compositional Dependability Evaluation for STATEMATE. *IEEE Trans. on Software Engineering*, 35(2):274–292, 2009.

[BHHK03] Christel Baier, Boudewijn Haverkort, Holger Hermanns und Joost-Pieter Katoen. Model-Checking Algorithms for Continuous-Time Markov Chains. *IEEE*

Trans. on Software Engineering, 29(7):1–18, 2003.

[BHKH05] Christel Baier, Holger Hermanns, Joost-Pieter Katoen und Boudewijn R. Haverkort. Efficient Computation of Time-Bounded Reachability Probabilities in Uniform Continuous-Time Markov Decision Processes. *Theoretical Computer Science*, 345(1):2–26, 2005.

[BO05a] Stefan Blom und Simona Orzan. A Distributed Algorithm for Strong Bisimulation Reduction of State Spaces. *Software Tools for Technology Transfer*, 7(1):74–86, 2005.

[BO05b] Stefan Blom und Simona Orzan. Distributed State Space Minimization. *Software Tools for Technology Transfer*, 7(3):280–291, 2005.

[Der07] Salem Derisavi. A Symbolic Algorithm for Optimal Markov Chain Lumping. In *Proc. of TACAS*, Band 4424 von LNCS, Seiten 139–154. Springer, 2007.

[HJ94] Hans Hansson und Bengt Jonsson. A Logic for Reasoning about Time and Reliability. *Formal Aspects of Computing*, 6(5):512–535, 1994.

[HKD09] Tingting Han, Joost-Pieter Katoen und Berteun Damman. Counterexample Generation in Probabilistic Model Checking. *IEEE Trans. on Software Engineering*, 35(2):241–257, 2009.

[HKNP06] Andrew Hinton, Marta Kwiatkowska, Gethin Norman und David Parker. PRISM: A Tool for Automatic Verification of Probabilistic Systems. In *Proc. of TACAS*, Band 3920 von LNCS, Seiten 441–444. Springer, 2006.

[WB10] Ralf Wimmer und Bernd Becker. Correctness Issues of Symbolic Bisimulation Computation for Markov Chains. In *Int'l GI/ITG Conf. on Measurement, Modelling and Evaluation of Computing Systems (MMB)*, Band 5987 von LNCS, Seiten 287–301. Springer, 2010.

[WBB09] Ralf Wimmer, Bettina Braitling und Bernd Becker. Counterexample Generation for Discrete-time Markov Chains using Bounded Model Checking. In *Proc. of VMCAI*, Band 5403 von LNCS, Seiten 366–380. Springer, 2009.

[WDH10] Ralf Wimmer, Salem Derisavi und Holger Hermanns. Symbolic Partition Refinement with Automatic Balancing of Time and Space. *Performance Evaluation*, 67(9):815–835, 2010.

[WHH+06] Ralf Wimmer, Marc Herbstritt, Holger Hermanns, Kelley Strampp und Bernd Becker. Sigref – A Symbolic Bisimulation Tool Box. In *Proc. of ATVA*, Band 4218 von LNCS, Seiten 477–492. Springer, 2006.

[ZN10] Lijun Zhang und Martin R. Neuhäußer. Model Checking Interactive Markov Chains. In *Proc. of TACAS*, Band 6015 von LNCS, Seiten 53–68. Springer, 2010.

Ralf Wimmer studierte an der Albert-Ludwigs-Universität Freiburg Informatik und Mikrosystemtechnik und erhielt 2004 das Diplom mit Auszeichnung in Informatik. Für seine Diplomarbeit wurde er mit dem VDI-Nachwuchsförderpreis ausgezeichnet. 2011 promovierte er mit Auszeichnung an der Albert-Ludwigs-Universität bei Prof. Dr. Bernd Becker im DFG Transregio-Sonderforschungsbereich SFB/TR 14 AVACS zu symbolischen Methoden für die Verifikation probabilistischer Systeme. Derzeit ist Ralf Wimmer Akademischer Rat an der Universität Freiburg. Er ist Autor von über 20 Publikationen.

Entwicklung einer Komplexitätstheorie für randomisierte Suchheuristiken: Black-Box-Modelle*

Carola Winzen

Universität des Saarlandes und Max-Planck-Institut für Informatik
Campus E1 4
66123 Saarbrücken
winzen@mpi-inf.mpg.de

Abstract: Randomisierte Suchheuristiken sind problemunabhängige Algorithmen, die sowohl im wissenschaftlichen als auch im industriellen Kontext zur Optimierung von schwierigen Problemen genutzt werden. Sie sind einfach zu implementieren, lassen sich vielseitig einsetzen und liefern überraschend häufig bereits in kurzer Zeit sehr gute Ergebnisse. Daher sind randomisierte Suchheuristiken weit verbreitet. Ein großes Problem in Anwendung von randomisierten Suchheuristiken ist jedoch die Tatsache, dass sich schwer vorhersagen lässt, ob sich das zu optimierende Problem gut durch eine geeignete Heuristik lösen lässt oder ob andere problemspezifische Verfahren deutliche besser geeignet sind.

Mit meiner Dissertation leisten wir einen Beitrag zur Entwicklung einer Komplexitätstheorie für randomisierte Suchheuristiken. Unser langfristiges Ziel ist die Charakterisierung von Problemklassen in solche, die sich schnell und zuverlässig durch Suchheuristiken optimieren lassen und solche, für die grundsätzlich andere Methoden besser geeignet sind.

1 Einleitung

Trotz immer schnellerer Computer lassen sich viele Probleme unseres täglichen Lebens selbst mit den besten bekannten Methoden nicht effizient lösen. Für andere Probleme mag ein effizienter, auf das Problem zugeschnittener Algorithmus zwar existieren, jedoch würde die Entwicklung eines solchen zu lange dauern oder aber zu viele Kosten verursachen. In beiden Situationen bieten *randomisierte Suchheuristiken* eine geeignete Alternative zu maßgeschneiderten Algorithmen.

Randomisierte Suchheuristiken sind Algorithmen, die, basierend auf dem Prinzip des Zufalls und ohne das konkret zugrunde liegende Problem zu kennen, Lösungsvorschläge erstellen, diese evaluieren und sich durch (in der Regel lokale) Modifikationen nach und nach einer optimalen Lösung annähern. Damit sind randomisierte Suchheuristiken vielseitig einsetzbare Algorithmen, die aufgrund ihrer hohen Flexibilität nicht nur im industriel-

*Originaltitel der Arbeit: Toward a Complexity Theory for Randomized Search Heuristics: Black-Box Models. Die Dissertation ist in englischer Sprache verfasst.

len Kontext weit verbreitet sind, sondern auch im rein wissenschaftlichen Umfeld immer häufiger Anwendung finden.

Viele Suchheuristiken sind durch Phänomene der Natur inspiriert. So gibt es beispielsweise *evolutionäre* Algorithmen, *Ameisenalgorithmen* und *artifizielle Immunsysteme*, die auf in der Natur beobachteten Prinzipien aufbauen. Die beiden klassischen Suchheuristiken *Simulated Annealing* und *Threshold Accepting* hingegen haben Bausteine, die aus der Physik motiviert sind.

Eine der einfachsten Suchheuristiken ist der (1+1) evolutionäre Algorithmus (EA), dessen Pseudocode wir in Algorithmus 1 darstellen.

Algorithmus 1: (1+1) EA zur Maximierung einer Funktion $f\colon \{0,1\}^n \to \mathbb{R}$.

1 **Initialisierung:** Wähle einen zufälligen Suchpunkt $x \in \{0,1\}^n$ und evaluiere $f(x)$;
2 **Optimierung: for** $t = 1, 2, 3, \ldots$ **do**
3 Kopiere $y \leftarrow x$, flippe dann jeden Eintrag in y mit Wahrscheinlichkeit $1/n$ und evaluiere $f(y)$; //Variation
4 **if** $f(y) \geq f(x)$ **then** $x \leftarrow y$; //Selektion

Wie viele Iterationen braucht der (1+1) EA, für eine gegebene Zielfunktion f, bis er zum ersten Mal einen *optimalen* Suchpunkt anfragt? Diese Frage ist Gegenstand der sogenannten *Laufzeitanalyse*. Trotz der vielen erfolgreichen Anwendungsbeispiele von Suchheuristiken in Wissenschaft und Praxis ist die Theorie der Laufzeitanalyse erst seit wenigen Jahrzehnten Betrachtungsgegenstand der Informatik. Insbesondere die Arbeiten von Ingo Wegener haben einen bedeutenden Beitrag zur Entwicklung einer konsistenten Theorie von randomisierten Suchheuristiken geleistet. Während wir jedoch heutzutage spezifische Algorithmen auf spezifischen Problemen untersuchen können, fehlt es uns derzeit an einem guten Verständnis, in welchen Situationen problemunabhängige Heuristiken in kurzer Laufzeit gute Lösungen liefern können. Eine Komplexitätstheorie ähnlich zu der in der klassischen Algorithmik ist daher wünschenswert.

Mit dieser Arbeit tragen wir zur Entwicklung einer solchen Komplexitätstheorie für Suchheuristiken bei. Wir zeigen zunächst anhand verschiedener Beispiele, dass existierende Modelle die Schwierigkeit eines Problems nicht immer zufriedenstellend erfassen. Wir schlagen daher ein weiteres Modell vor. In unserem *ordnungsbasierten Black-Box-Modell* lernen die Algorithmen keine exakten Funktionswerte, sondern bloß die Rangordnung der bislang angefragten Suchpunkte. Dieses Modell gibt für manche Probleme eine bessere Einschätzung der typischen Laufzeit von randomisierten Suchheuristiken. Jedoch gibt es auch im neuen Modell Funktionenklassen, deren Komplexität als zu gering einzuschätzen ist.

Ich hoffe den Leser dieser Kurzzusammenfassung meiner Dissertation davon überzeugen zu können, dass die Entwicklung einer Komplexitätstheorie für randomisierte Suchheuristiken eine spannende Aufgabe ist, dass es erste vielversprechende und interessante Ergebnisse gibt und dass die vorgeschlagenen Black-Box-Modelle viele spannende Fragestellungen zur weiteren Forschungsarbeit aufwerfen.

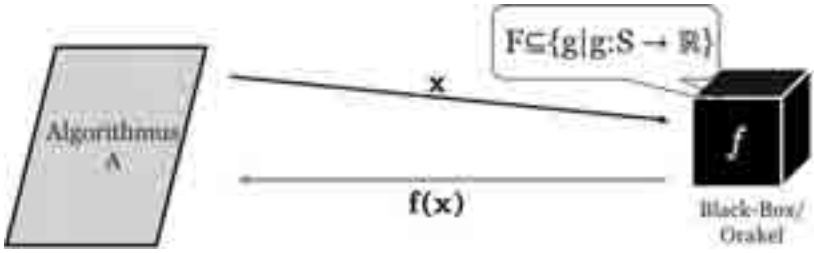


Abbildung 1: Illustration des Grundmodells

2 Die beiden Grundmodelle

Da randomisierte Suchheuristiken (RSH) problemunabhängig sind, können sie über das konkret zu lösende Problem nur Informationen erhalten, indem sie Lösungskandidaten, im Folgenden auch *Suchpunkte* genannt, evaluieren. Dabei gehen wir davon aus, dass diese Evaluierung von einem Orakel durchgeführt. Ist f die zu optimierende Funktion, so gibt das Orakel zu jedem angefragten Suchpunkt x dessen Funktionswert $f(x)$ preis. Die Heuristik „lernt“ somit zwar den konkreten Wert $f(x)$, erhält jedoch keine weitere Information über die Funktion f. Diese Konstellation wird in der Literatur häufig *Black-Box-Optimierung* genannt.

Abbildung 1 illustriert das Grundmodell: das Orakel bzw. die Black-Box-wählt eine Klasse von Problemen $\mathcal{F}$ sowie eine konkrete Probleminstanz $f \in \mathcal{F}$. Wir nehmen an, dass die Probleme $\mathcal{F}$ als Funktionen $g : S \to \mathbb{R}$ modelliert sind. Die Menge S nennen wir *Suchraum,* die Elemente $x \in S$ *Suchpunkte.* Die Funktionenklasse $\mathcal{F}$ ist der Heuristik bekannt, nicht aber die konkrete Funktion f. Das Ziel der Heuristik ist die Maximierung von f. In jedem Schritt fragt die Heuristik einen Suchpunkt x an, woraufhin das Orakel dessen Funktionswert $f(x)$ preisgibt. Die Wahl des Suchpunktes x hängt in der Regel vom bislang angesammelten Wissen über f ab. Da wir uns für *randomisierte* Suchheuristiken interessieren, erlauben wir, dass der Algorithmus *zufällige* Entscheidungen trifft. Das heißt, wir nehmen an, dass der Algorithmus Zugriff auf einen (unbegrenzt langen) String von Zufallszahlen hat.

Ist nun $\mathcal{A}$ eine Klasse von Algorithmen und $A \in \mathcal{A}$ ein konkreter Algorithmus dieser Klasse, so bezeichnen wir mit $T(A, f)$ die erwartete Anzahl von Funktionsevaluationen (Anfragen an das Orakel) bis der Algorithmus A zum ersten Mal einen *optimalen* Suchpunkt $x \in \arg\max f$ anfragt (*first hitting time*). Wir nennen $T(A, f)$ die *Laufzeit* von Algorithmus A für die Funktion f. Die worst-case Laufzeit $\sup_{f\in\mathcal{F}} T(A, f)$ von A auf der Funktionenklasse $\mathcal{F}$ bezeichnen wir mit $T(A, \mathcal{F})$. Schließlich definieren wir, wie in der Komplexitätstheorie üblich, die *Komplexität von $\mathcal{F}$ für $\mathcal{A}$* als die bestmögliche worst-case Laufzeit, $\inf_{A\in\mathcal{A}} T(A, \mathcal{F})$.

Betrachten wir die Klasse $\mathcal{A}$ aller Algorithmen, so nennen wir $T(\mathcal{A}, \mathcal{F})$ die *uneingeschränkte Black-Box-Komplexität von $\mathcal{F}$*. Dies ist die Black-Box-Komplexität, die Droste, Jansen und Wegener in ihrer Arbeit [DJW06] betrachten.

Obwohl die uneingeschränkte Black-Box-Komplexität viele interessante und zum Teil ungelöste Probleme hervorgebracht hat, mussten bereits Droste, Jansen und Wegener feststellen, dass die Betrachtung *aller* Black-Box-Algorithmen für viele Funktionenklasse eine als deutlich zu niedrig anzusehende Einschätzung der Komplexität gibt. Besteht beispielsweise $\mathcal{F}$ aus nur einer Funktion $\mathcal{F} = \{f\}$, so ist die uneingeschränkte Black-Box-Komplexität von $\mathcal{F}$ eins: ein optimaler Algorithmus für dieses Problem berechnet zunächst *offline*, das heißt ohne Interaktion mit dem Orakel, einen optimalen Suchpunkt $x \in \arg\max f$ und fragt diesen in der ersten Iteration an. Ein weiteres Beispiel für eine zu geringe Einschätzung der tatsächlichen Komplexität ist das MAXCLIQUE-Problem. Dieses NP-schwere Problem hat eine uneingeschränkte Black-Box-Komplexität von $O(n^2)$, zertifiziert durch den Algorithmus, der zunächst die Präsenz aller potentiell möglichen Kanten anfragt und dann offline eine maximale Clique berechnet.

2.1 Das unbiased Black-Box-Modell

Nachdem die Ergebnisse von Droste, Jansen und Wegener zunächst wenig Nachfolgearbeiten zur Komplexitätstheorie randomisierter Suchheuristiken hervorbrachten, nahm die Forschung zu Black-Box-Komplexität durch ein neues Modell von Lehre und Witt für pseudo-Bool'sche Funktionen $f : \{0,1\}^n \to \mathbb{R}$ neue Fahrt auf. Lehre und Witt beobachten, dass viele Suchheuristiken in der Auswahl neuer Suchpunkte nicht sehr selektiv sind, sondern sowohl Bitpositionen als auch Bitwerte gleich (*unbiased*) behandeln. Daher betrachten sie nur solche Algorithmen, die diesem Schema folgen. Formal nennen wir eine Familie $(\mathcal{D}(. \mid y^1, \dots, y^k))_{y^1,\dots,y^k \in \{0,1\}^n}$ von Wahrscheinlichkeitsverteilungen auf $\{0,1\}^n$ *k-adisch unbiased,* wenn sie für alle möglichen Inputs $y^1, \dots, y^k$ invariant unter Hamming-Automorphismen ist, das heißt, wenn sie für alle Permutationen σ von $[n] := \{1, \dots, n\}$ und alle Punkte $z \in \{0,1\}^n$ die folgende Bedingung erfüllt:

$$\forall x \in \{0,1\}^n : \mathcal{D}(x \mid y^1, \dots, y^k) = \mathcal{D}(\sigma(x \oplus z) \mid \sigma(y^1 \oplus z), \dots, \sigma(y^k \oplus z)) \,.$$

Die *k-adische unbiased Black-Box-Komplexität von $\mathcal{F}$* ist die Komplexität von $\mathcal{F}$ bezüglich aller k-adisch unbiased Black-Box-Algorithmen (Algorithmen, die dem Schema von Algorithmus 2 folgen).

Algorithmus 2: Modell eines k-adisch unbiased Black-Box-Algorithmus zur Maximierung einer Funktion $f : \{0,1\}^n \to \mathbb{R}$

1 **Initialisierung:** Wähle $x^0 \in \{0,1\}^n$ zufällig uniform. Frage $f(x^0)$ an;
2 **for** $t = 1, 2, 3, \dots$ **do**
3 Abhängig von $\big(f(x^0), \dots, f(x^{t-1})\big)$ wähle k Indizes $i_1, \dots, i_k \in [0..t-1]$ sowie eine k-adische unbiased Familie $(D(. \mid y^1, \dots, y^k))_{y^1,\dots,y^k \in \{0,1\}^n}$ von Wahrscheinlichkeitsverteilungen auf $\{0,1\}^n$;
4 Ziehe x^t gemäß der Verteilung $D(. \mid x^{i_1}, \dots, x^{i_k})$. Frage $f(x^t)$ an;

3 Die ONEMAX-Funktionenklasse und Mastermind

Eine klassische Testfunktion in der Analyse randomisierter Suchheuristiken ist die Funktion ONEMAX, welche jedem Suchpunkt $x \in \{0,1\}^n$ die Anzahl der Einsen in x zuordnet. Die Laufzeiten von Heuristiken auf dieser Funktion wird häufig analysiert um zu verstehen, wie sich die Heuristik in „einfachen“ Bereichen des Optimierungsproblems verhält. Um einen Vergleich von unbiased Black-Box-Komplexität mit der uneingeschränkten Black-Box-Komplexität zu erlauben, betrachten wir in der Regel eine Verallgemeinerung der ONEMAX-Funktion: Für jeden Bitstring $z \in \{0,1\}^n$ definieren wir die Funktion OM_z, die jedem Suchpunkt $x \in \{0,1\}^n$ die Anzahl $|\{j \in [n] | z_j = x_j\}|$ der Übereinstimmungen von x mit z zuordnet. Mit ONEMAX_n bezeichnen wir die Menge $\{\text{OM}_z | z \in \{0,1\}^n\}$ aller solcher Funktionen.

Bereits 1963 haben Erdős und Rényi [ER63] gezeigt, dass die uneingeschränkte Black-Box-Komplexität von ONEMAX_n $\Theta(n/\log n)$ ist. Unwissentlich hat Chvátal [Chv83] dieses Ergebnis verallgemeinert: die uneingeschränkte Black-Box-Komplexität von ONEMAX_n ist $\Theta(n/\log n)$ auch dann, wenn wir statt der beiden „Farben“ $0, 1$ eine beliebe konstante Anzahl $k \in \mathbb{N}$ von Farben betrachten und ONEMAX_n auf natürliche Weise zu der Menge der Funktionen

$$\{\text{OM}_z : [k]^n \to [0..n], x \mapsto \{j \in [n] | x_j = z_j\} | z \in [k]^n\}$$

erweitern. In dieser Definition kürzen wir $[n] := \mathbb{N}_{\leq n} := \{1, \ldots, n\}$ und $[0..n] := [n] \cup \{0\} = \{0, 1, \ldots, n\}$ ab.

Die typische Laufzeit von randomisierten Suchheuristiken auf ONEMAX_n ist mit $\Theta(n \log n)$ Funktionsevaluationen sehr viel größer als die uneingeschränkte Black-Box-Komplexität von ONEMAX_n. Letztere spiegelt somit die Komplexität dieser sehr einfachen Funktionenklasse für randomisierte Suchheuristiken nicht gut wider. Dass Lehre und Witt zeigen können, dass die unäre (1-adische) unbiased Black-Box-Komplexität von ONEMAX_n von der Größenordnung $n \log n$ ist, macht ihr Modell zu einem interessanten Kandidaten für Black-Box-Modelle: Komplexität der Funktionenklasse und Laufzeit der Suchheuristiken stimmen überein.

Die Funktionenklasse ONEMAX_n ist sehr verwandt mit dem Mastermindspiel. Mastermind ist ein Brettspiel für zwei Personen. Der Kodierer, im Folgenden mit Carole bezeichnet, legt mit den bunten Spielstiften ein geheimes Codewort. Paul, der zweite Spieler soll dieses Codewort herausfinden. In jeder Runde gibt er dazu mit den bunten Spielstiften einen Rateversuch ab. Carole beantwortet jeden solchen Versuch mit schwarzen und weien Antwortstiften. Für jede Übereinstimmung des Rateversuchs mit dem geheimen Codewort gibt sie Paul einen schwarzen Antwortstift, für die richtige Farbe am falschen Platz je einen weißen. Pauls Ziel ist es, das Codewort mit möglichst wenig Rateversuchen herauszufinden. Black-Box-Algorithmen zur Optimierung von ONEMAX_n mit k Farben entsprechen genau den möglichen Strategien, die Paul im Mastermindspiel mit k Farben hat, wenn Carole statt mit schwarzen und weißen nur mit den schwarzen Antwortstiften antwortet und ihm somit nur die Anzahl der *genauen Übereinstimmungen* seines Rateversuchs mit ihrem geheimen Codewort bekannt gibt. Interessanterweise verändert die Reduktion auf schwarze Antwortstifte die asymptotische Komplexität des Mastermindspiels nicht.

4 Neue Ergebnisse zu den beiden Grundmodellen

Im ersten Teil der Dissertation beschäftigen wir uns mit den oben eingeführten Grundmodellen: dem uneingeschränkten Black-Box-Modell und den k-adischen unbiased Black-Box-Modellen. Für eine Reihe verschiedener Probleme zeigen wir, dass die Black-Box-Komplexitäten deutlich niedriger sind als die typischen Laufzeiten von Heuristiken.

Ergebnisse für die unbiased Black-Box-Modelle mit Aritäten ≥ 2. Lehre und Witt analysieren in [LW10] das unäre Black-Box-Modell, in dem Algorithmen nur einen einzigen Suchpunkt nutzen dürfen um einen neuen Suchpunkt zu finden. Solche Algorithmen nennen wir *mutationsbasiert*.

Im vierten Kapitel der Dissertation, welches auf der Veröffentlichung [DJK+11] basiert, beschäftigen wir uns mit den unbiased Black-Box-Modellen für Aritäten ≥ 2. In diesen Modellen sind Algorithmen auch in der Lage zwei oder mehr Suchpunkte durch sogenanntes *Crossover* miteinander zu kombinieren. Eine wichtige Fragestellung in der Theorie randomisierter Suchheuristiken ist die nach dem Nutzen von Crossover-Operatoren. Während sie in der Praxis regelmäßig Anwendung finden, gibt es bislang nur wenig Evidenz, dass sie auch beweisbar bessere Ergebnisse liefern.

Wir zeigen, dass für $2 \leq k \leq n$ die k-adische unbiased Black-Box-Komplexität von $\textsc{OneMax}_n$ auf $O(n/\log k)$ fällt. Dieses Ergebnis lässt prinzipiell zwei Interpretationen zu. Zum einen liefert es eine Indikation, dass crossoverbasierte Algorithmen tatsächlich einen Vorteil gegenüber rein mutationsbasierten Algorithmen haben. Zum anderen wirft es die Frage auf, wie gut das unbiased Modell für Aritäten ≥ 2 die Wirklichkeit widerspiegelt. Die Antwort auf diese Frage kennen wir momentan nicht, glauben aber, dass in beiden Aussagen etwas Wahres liegt.

Zudem zeigen wir, dass auch für die Funktionenklasse $\textsc{LeadingOnes}_n$ die Komplexität von $\Theta(n^2)$ im unären Fall auf $O(n \log n)$ im binären (d. h. 2-adischen) Modell fällt. Die Funktion $\textsc{LeadingOnes} : \{0,1\}^n \to [0..n], x \mapsto \max\{j \in [0..n] | \forall i \leq j : x_i = 1\}$ misst die Länge des längsten Prefixes aus Einsen. Diese Funktion ist wie $\textsc{OneMax}$ eine Standardtestfunktion. Sie wird durch die Funktionenklasse $\textsc{LeadingOnes}_n :=$

$$\left\{\textsc{Lo}_{z,\sigma} : \{0,1\}^n \to \mathbb{N}, x \mapsto \max\{j \in [0..n] \mid \forall i \leq j : x_{\sigma(i)} = z_{\sigma(i)}\} \mid z \in \{0,1\}^n, \sigma \in S_n\right\}$$

auf natürliche Weise verallgemeinert.[1]

Diese Schranke von $O(n \log n)$ scheint auf den ersten Blick bestmöglich („scharf“) zu sein und es liegt nahe zu vermuten, dass diese Schranke auch für das uneingeschränkte Modell bestmöglich ist. Überraschenderweiser können wir jedoch zeigen, dass bereits die 3-adische unbiased Black-Box-Komplexität von $\textsc{LeadingOnes}_n$ höchstens $O(n \log n/\log\log n)$ ist. Das ist das fünfte Kapitel der Arbeit (siehe auch [DW11a]). Interessanterweise sind die tatsächlichen Komplexitäten von $\textsc{LeadingOnes}_n$ in den verschiedenen Black-Box-Modellen bis heute nicht bekannt.

Ergebnisse für das unäre unbiased Black-Box-Modell. Die bislang vorgestellten Ergebnisse beschäftigen sich mit dem Vorteil von höheren Aritäten. Im sechsten Kapitel der

[1] Mit S_n bezeichnen wir die Menge aller Permutationen σ der Menge $[n]$.

Dissertation, welches auf der Veröffentlichung [DKW11] basiert, zeigen wir schließlich, dass auch das unäre unbiased Modell für manche Funktionenklassen eine als deutlich zu niedrig anzusehende Komplexität aufweist. Damit meinen wir, dass die unäre unbiased Komplexität deutlich niedrig ist als die typische Laufzeit von randomisierten Suchheuristiken. Unter anderem zeigen wir, dass ein NP-schweres Teilproblem von PARTITION eine unäre unbiased Black-Box-Komplexität von nur $O(n \log n)$ hat.

5 Das Black-Box-Modell mit beschränktem Speicher

Die im ersten Teil der Dissertation vorgestellten Ergebnisse zeigen, dass wir uns mit restriktiveren Modellen beschäftigen müssen, wenn wir über Black-Box-Komplexität eine realistische Einschätzung von Laufzeiten randomisierter Suchheuristiken erzielen wollen. Ein solches Modell, welches tatsächlich schon in der Arbeit [DJW06] vorgeschlagen wurde, beschränkt den Speicher von Suchheuristiken. Die dem Modell zugrunde liegende Beobachtung ist die Tatsache, dass sich viele Heuristiken nicht merken, welche Suchpunkte sie bereits angefragt haben. Stattdessen wird in der Regel sogar nur die beste bislang bekannte Lösung und deren Funktionswert gespeichert. Droste, Jansen und Wegener schlagen daher ein *memory-restricted* Modell vor. In diesem Black-Box-Modell mit beschränktem Speicher werden nur solche Algorithmen berücksichtigt, die zu jedem Zeitpunkt nur einen Suchpunkt und dessen Funktionswert speichern. Die Entscheidung für den nächsten Suchpunkt darf nur auf dem aktuellen Inhalt des Speichers basieren. Insbesondere hat der Algorithmus keinen Zugang zu einem Iterationszähler. Wurde ein zweiter Suchpunkt angefragt und dessen Funktionswert evaluiert, so muss der Algorithmus entscheiden, ob er seinen aktuellen Speicher durch den zuletzt gefragten Suchpunkte und dessen Funktionswert ersetzt oder ob der Speicher nicht aktualisiert und die neu gewonnene Information komplett verworfen wird.

Die Black-Box-Komplexität mit beschränktem Speicher lässt sich gut am Mastermind-Beispiel verdeutlichen. Wir spielen das Spiel mit nur zwei Reihen. Existiert eine leere Reihe, so kann Paul – basierend auf der Information, die auf dem Brett verfügbar ist – eine neue Vermutung abgeben. Carole sagt ihm, wie nah seine Farbkombination an ihrem Codewort ist. Sind beide Reihen des Spielbretts belegt, so muss Paul sich entscheiden, welche Reihe er auf dem Brett belässt und welche er entfernt. Beim Entfernen vergisst er zugleich die Information, die in dieser Reihe kodiert war. Zudem weiß er zu keinem Zeitpunkt, seit wie viele Runden die beiden das Spiel schon spielen.

Droste, Jansen und Wegener vermuten, dass diese Einschränkung des Speichers die Komplexität des Mastermindspiels mit zwei Farben, d. h. die Komplexität von ONEMAX_n, von $\Theta(n/\log n)$ im Grundmodell auf $\Omega(n \log n)$ im Modell mit beschränktem Speicher erhöht. Diese Vermutung widerlegen wir im siebten Kapitel der Dissertation, welches auf der Veröffentlichung [DW12] basiert und wohl als eines der Kernergebnisse dieser Arbeit bezeichnet werden darf. Wir zeigen, dass nicht nur ONEMAX_n eine memory-restricted Black-Box-Komplexität von $O(n/\log n)$ hat, sondern dass dies auch die memory-restricted Komplexität des Mastermindspiels mit einer konstanten Anzahl k von Farben ist. Dieses Ergebnis ist bestmöglich, wie sich leicht durch informationstheoretische Argumente be-

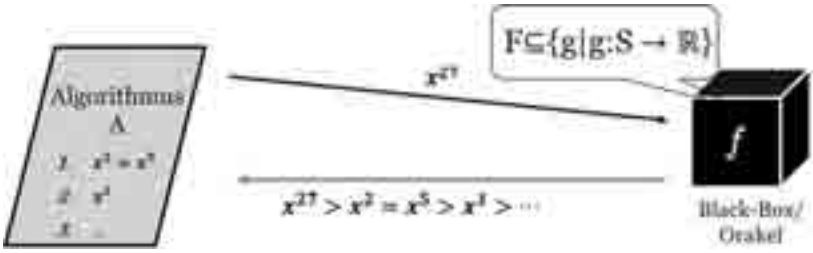


Abbildung 2: Illustration des ordnungsbasierten Modells

weisen lässt.

Eine Kernidee des Beweises ist die blockweise Identifizierung des Codewortes. Auf den verkleinerten Suchraum können wir die Ergebnisse von Erdős und Rényi [ER63] bzw. von Chvátal [Chv83] anwenden. Mit dieser Technik benötigen wir $O(s/\log s)$ Iterationen um einen Block der Länge $s = n^\varepsilon$ zu identifizieren, wobei $\varepsilon > 0$ eine beliebige Konstante ist. Da es $\lceil n/s \rceil$ solcher Blöcke gibt, ist die Gesamtlaufzeit $O(n/\log n)$. Da wir nur einen Suchpunkt (eine „Reihe“) als Speicherplatz haben, speichern wir die Antworten auf unsere Fragen in einem zu diesem Zeitpunkt ungenutzten Teil des Strings. Das wirft die Schwierigkeit auf, dass wir beim Kodieren der Antworten den String selber verändern und wir somit aufpassen müssen, wie wir Caroles Antworten interpretieren. Diese und verwandte Schwierigkeiten lösen wir durch eine aufwendige „Buchhaltung“.

6 Ein ordnungsbasiertes Komplexitätsmodell

Unser Ergebnis zum Modell mit beschränktem Speicher zeigt, dass hier durch aufwendige Techniken sehr viel Information gespeichert werden kann. Das spiegelt das Verhalten von randomisierten Suchheuristiken nicht gut wider. Daher schlagen wir im achten Kapitel der Dissertation (vgl. auch [DW11b]) ein neues ordnungsbasiertes Modell vor. Das Modell basiert auf der Überlegung, dass viele Heuristiken statt der *absoluten* Funktionswerte nur *relative* Funktionswerte bei der Auswahl des nächsten Suchpunktes zugrunde legen.[2] Daher beschränken wir uns in unserem *ordnungsbasierten Black-Box-Modell* auf solche Algorithmen, die demselben Prinzip folgen. Formal können wir dies erreichen, indem die Funktionsevaluation eines Suchpunktes x nicht dessen Funktionswert $f(x)$ preisgibt, sondern nur dessen relative Qualität unter allen bislang angefragten Suchpunkten.

Abbildung 2 illustriert dieses Modell. In diesem Beispiel weiß der Algorithmus aus seinen bisherigen Fragen bereits, dass Suchpunkte x^2 und x^5 besser sind als x^1, welcher wiederum besser ist als alle anderen bislang angefragten Suchpunkte. Basierend auf diesem Wissen , dass $f(x^2) = f(x^5) > f(x^1)$ gilt, fragt er nun Suchpunkt x^{27} an und bekommt als Antwort, dass dieser besser als alle bislang angefragten Suchpunkte ist.

[2] Im (1+1) evolutionären Algorithmus (vgl. Algorithmus 1) wird dies in Zeile 4 deutlich.

Wir zeigen, dass in diesem Modell die Black-Box-Komplexität von der Funktionenklasse BINARYVALUE$_n$, das am besten als gewichtete Version des Mastermindspiels illustriert werden kann, deutlich höher ist als deren uneingeschränkte Black-Box-Komplexität. Während die letztere nur $O(\log n)$ ist, ist die ordnungsbasierte Komplexität linear in n. Interessanterweise und für uns zunächst überraschend können wir für das ordnungsbasierte Black-Box-Modell jedoch auch zeigen, dass sich die Komplexität von ONEMAX$_n$ in diesem Modell nicht verändert: sie ist wie im uneingeschränkten Fall $\Theta(n/\log n)$.

Ein nettes Beispiel für die Tatsache, dass das ordnungsbasierte Modell viele natürliche Fragen aufwirft, ist das folgende Waagenproblem: Gegeben seien n voneinander unterscheidbare Kugeln verschiedenen Gewichts. Wie oft müssen wir die vorhandene Apothekerwaage nutzen bis wir die n Kugeln in zwei Teilmengen gleichen Gewichts unterteilen können?

7 Anwendung auf kombinatorische Probleme

Im dritten Teil der Dissertation wenden wir die verschiedenen Black-Box-Modelle auf die beiden kombinatorischen Probleme „Minimum Spanning Tree“ und „Single-Source Shortest Paths“ an. Während das Problem minimaler Spannbäume auf natürliche Weise als Funktionenklasse $\{g : \{0,1\}^m \to \mathbb{R}\}$ modelliert werden kann, ist die Modellierung des Problems der Berechnung kürzester Wege nicht eindeutig. In der Regel wird eine Modellierung als Klasse von Funktionen $\{g : [n]^n \to \mathbb{R}\}$ bevorzugt. Aus diesem Grund führen wir in diesem Kapitel auch verschiedene Möglichkeiten zur Verallgemeinerung des unbiased Black-Box-Modells ein. Dieses Kapitel basiert auf der Veröffentlichung [DKLW11].

8 Zusammenfassung

Mit dem Ziel eine Komplexitätstheorie für randomisierte Suchheuristiken (RSH) zu entwickeln haben wir die Black-Box-Komplexitäten verschiedener Funktionenklassen in den beiden Grundmodellen sowie im Modell mit beschränktem Speicher analysiert. Zudem haben wir ein ordnungsbasiertes Komplexitätsmodell vorgeschlagen, welches für manche Funktionenklassen eine bessere Einschätzung ihrer Optimierbarkeit durch RSH liefert. Keines der untersuchten Modelle scheint jedoch typische Laufzeiten von randomisierten Suchheuristiken auf *allen* Funktionenklassen widerzuspiegeln.

Die vorgelegte Dissertation wirft eine Reihe interessanter Fragestellungen auf. Zwei Beispiele für vielversprechende Anknüpfungspunkte sind zum einen die Frage nach anderen, potenziell besser geeigneten Komplexitätsmodellen, zum anderen aber auch die Nutzung der Erkenntnisse aus den Black-Box-Modellen zur Entwicklung besserer Suchheuristiken. Zudem gibt es eine Vielzahl von Problemen, deren Black-Box-Komplexität nicht bekannt ist. Insbesondere für die Berechnung von *unteren Schranken* scheinen die vorhandenen Methoden nicht auszureichen. Eine Weiterentwicklung der Beweistechniken in randomisierten Modellen ist daher erstrebenswert.

Literatur

[Chv83] Vasek Chvátal. Mastermind. *Combinatorica*, 3:325–329, 1983.

[DJK+11] Benjamin Doerr, Daniel Johannsen, Timo Kötzing, Per Kristian Lehre, Markus Wagner und Carola Winzen. Faster Black-Box Algorithms Through Higher Arity Operators. *Proc. of the 11th ACM Workshop on Foundations of Genetic Algorithms (FOGA)*, Seiten 163–172, 2011.

[DJW06] Stefan Droste, Thomas Jansen und Ingo Wegener. Upper and Lower Bounds for Randomized Search Heuristics in Black-box Optimization. *Theory of Computing Systems*, 39:525–544, 2006.

[DKLW11] Benjamin Doerr, Timo Kötzing, Johannes Lengler und Carola Winzen. Black-Box Complexities of Combinatorial Problems. *Proc. of the 13th Annual Genetic and Evolutionary Computation Conference (GECCO)*, Seiten 981–988, 2011.

[DKW11] Benjamin Doerr, Timo Kötzing und Carola Winzen. Too Fast Unbiased Black-Box Algorithms. *Proc. of the 13th Annual Genetic and Evolutionary Computation Conference (GECCO)*, Seiten 2043–2050, 2011.

[DW11a] Benjamin Doerr und Carola Winzen. Breaking the $O(n \log n)$ Barrier of Leading-Ones, 2011. Veröffentlichung ausstehend.

[DW11b] Benjamin Doerr und Carola Winzen. Towards a Complexity Theory of Randomized Search Heuristics: Ranking-Based Black-Box Complexity. *Proc. of the 6th International Computer Science Symposium in Russia (CSR)*, Seiten 15–28, 2011.

[DW12] Benjamin Doerr und Carola Winzen. Playing Mastermind With Constant-Size Memory. *Proc. of the Symposium on Theoretical Aspects of Computer Science (STACS)*, Seiten 441–452, 2012.

[ER63] Paul Erdős und Alfréd Rényi. On Two problems of Information Theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 8:229–243, 1963.

[LW10] Per Kristian Lehre und Carsten Witt. Black-box Search by Unbiased Variation. *Proc. of the 12th Annual Genetic and Evolutionary Computation Conference (GECCO)*, Seiten 1441–1448, 2010.

Carola Winzen wurde am 5. März 1984 in Würselen geboren. Nach einem Schüleraustausch in Tobatí, Paraguay, im Schuljahr 2000/01 zog sie nach Konstanz, wo sie 2003 ihr Abitur mit der Note 1,0 ablegte. Das Studium der Mathematik absolvierte Frau Winzen an der Christian-Albrechts-Universität zu Kiel in nur 8 Semestern (Gesamtnote *sehr gut*). Nach Beendigung des Studiums entschied sich Frau Winzen zunächst bei der Unternehmensberatung McKinsey & Company erste Berufserfahrungen zu sammeln. Dort arbeitete sie 2 Jahre lang vor allem an Netzwerkoptimierungs- und Fahrplanproblemen.

Im Januar 2010 nahm sie das Promotionsstudium der Informatik an der Universität des Saarlandes auf. Am Max-Planck-Institut für Informatik arbeitet sie in Gruppe „Algorithmen und Komplexität“ von Prof. Dr. Dr. h.c. mult. Kurt Mehlhorn. Die Promotionsprüfung legte Frau Winzen am 16. Dezember 2011 mit der Gesamtnote *summa cum laude* ab. Ihr Promotionsstudium wurde durch das Google Europe Fellowship in Randomized Algorithms gefördert.

Dynamische Code-Evolution für Java

Thomas Würthinger

Institut für Systemsoftware
Johannes Kepler University Linz
wuerthinger@ssw.jku.at

Abstract: Dynamische Code-Evolution ermöglicht strukturelle Änderungen an laufenden Programmen. Das Programm wird temporär angehalten, der Programmierer verändert den Quelltext und dann wird die Ausführung mit der neuen Programm-Version fortgesetzt.

Diese Arbeit beschreibt einen neuartigen Algorithmus für die unlimitierte dynamische Neudefinition von Java-Klassen in einer virtuellen Maschine. Die unterstützten Änderungen beinhalten das Hinzufügen und Entfernen von Feldern und Methoden sowie Veränderungen der Klassenhierarchie. Der Zeitpunkt der Veränderung ist nicht beschränkt und die aktuell laufenden Ausführungen von alten Versionen einer Methode werden fortgesetzt. Mögliche Verletzungen der Typsicherheit werden erkannt und führen zu einem Abbruch der Neudefinition. Die entwickelten Techniken können die Entwicklung neuer Programme beschleunigen sowie den Versionswechsel ohne Abschaltpause von Server-Anwendungen ermöglichen. Die Arbeit präsentiert auch ein Programmiermodell für sichere dynamische Aktualisierungen und diskutiert nützliche Limitierungen, die es dem Programmierer ermöglichen, über die semantische Korrektheit einer Aktualisierung Schlussfolgerungen zu ziehen.

Alle Algorithmen sind in der Java HotSpot VM implementiert und es wird derzeit seitens Oracle an der Integration in die offizielle Java-Version gearbeitet. Die Evaluierung zeigt, dass die neuen Fähigkeiten weder vor noch nach einer dynamischen Veränderung einen negativen Einfluss auf die Spitzenleistung der virtuellen Maschine haben.

1 Einführung

Der Eingriff in das Verhalten eines laufenden Programms wurde bereits früh in der Geschichte der Informatik bearbeitet [Fab76]. Die Forschung fokussierte dabei auf prozedurale Programmiersprachen: Bei einem Eingriff ersetzte man die Definitionen von Funktionen und es gab eigene Methoden zur Umwandlung der Daten. Mit der Einführung von objekt-orientierten Programmiersprachen wurden Klassendefinitionen und Subtypbeziehungen ein wichtiger Teil eines Programms. Die dynamische Veränderung eines Programms muss in diesem Kontext auch das Layout von existierenden Objekten sowie die Semantik von Methodenaufrufen aufgrund der aktuellen Klassenhierarchie berücksichtigen.

Die Benutzung einer virtuellen Maschine (VM) zur Ausführung von Programmen hilft beim Lösen dieser neuen Herausforderungen: Eine VM erhöht die Möglichkeiten führ dynamische Code-Evolution aufgrund der zusätzlichen Abstraktionsschicht zwischen dem

ausgeführten Programm und der Hardware. Die Hauptaufgaben dieser Zwischenschicht sind automatische Speicherverwaltung, dynamisches Klassenladen und Programmverifikation. Die in dieser Arbeit präsentierten Algorithmen für die dynamische Veränderung von Klassendefinitionen verwenden die existierende Infrastruktur der VM.

Die dynamische Veränderung von Java-Programmen ist derzeit unter dem Namen "hotswapping" bekannt, da sie auf das Austauschen von Methodendefinitionen beschränkt ist. Die Verbesserung dieser Funktionalität um zusätzliche Veränderungsmöglichkeiten ist von hoher Priorität für viele Java-Programmierer. Dies zeigt sich unter anderem auf der Oracle-Webseite für "requests for enhancements" wo diese Verbesserung zu den am meisten unterstützten Anfragen gehöhrt (Bug ID: 4910812) [Ora11].

2 Stufen der Code-Evolution

Aufgrund unserer Erfahrungen bei der Implementierung der Prototyp-VM schlagen wir die Unterscheidung der folgenden vier Stufen der Code-Evolution vor, die sich durch die Komplexität der Implementierung in einer VM unterscheiden:

Austauschen von Methodendefinitionen: Die einfachste mögliche Veränderung ist das Austauschen der Bytecodes einer Methode. Diese Stufe ist bereits in aktuellen Produkt-VMs implementiert und wird "hotswapping" genannt.

Hinzufügen und Entfernen von Methoden: Die VM verwaltet eine Tabelle mit Zeigern auf die virtuellen Methoden für jede Klasse. In einem objekt-orientierten Programm kann das Hinzufügen oder Entfernen einer Methode zu einer Veränderung der Tabelle in der betroffenen Klasse oder in Subklassen führen. Weiters muss Maschinencode mit Referenzen auf betroffene Methoden invalidiert oder neu berechnet werden.

Hinzufügen und Entfernen von Feldern: Bis zu dieser Stufe haben die Veränderungen nur die Metadaten der VM beeinflusst. Jetzt müssen auch die aktuell existierenden Objektinstanzen des laufenden Programms modifiziert werden. Auch Maschinencode der vom Layout von Klassen abhängig ist muss invalidiert werden.

Hinzufügen und Entfernen von Supertypen: Diese Stufe ist die komplexeste mögliche Veränderung für objekt-orientierte Sprachen. Die möglichen Auswirkungen auf die virtuelle Maschinen beeinhalten zusätzlich zu den Auswirkungen aller vorherigen Stufen noch die Möglichkeit, dass das Typsystem verletzt ist.

Die Veränderung eines Java-Programms kann auch danach klassifiziert werden, ob sie binär kompatibel ist [Dmi01]. Die hellgrauen Bereiche in Abbildung 1 stellen binär kompatible Veränderungen dar, die dunkelgrauen Bereiche binär inkompatible Veränderungen. Wir beschreiben Lösungen für das Problem von binär inkompatiblen Veränderungen in Abschnitt 5. Die im Rahmen der Arbeit entwickelte Prototyp-VM ist die erste VM, die alle vorgestellten Stufen der Code-Evolution unterstützt.

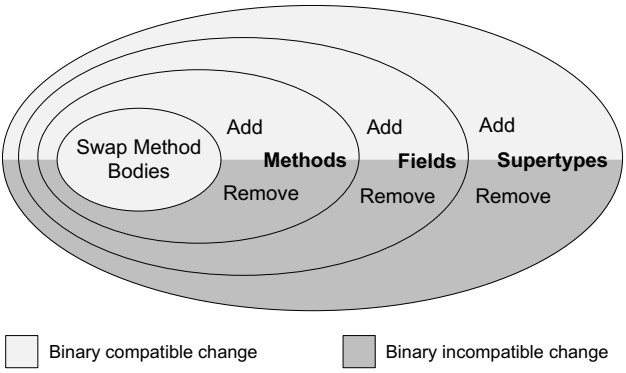


Abbildung 1: Stufen der Code-Evolution.

3 Anwendungsbereiche

Dynamische Code-Evolution kann in verschiedenen Bereichen eingesetzt werden, die jeweils ihre eigenen speziellen Anforderungen mitbringen. Wir unterscheiden vier Hauptanwendungsbereiche:

Beschleunigte Programmentwicklung. Wenn ein Programmierer häufig kleine Veränderungen einer Applikation mit einer langen Startzeit macht, hilft dynamische Code-Evolution die Produktivität bei der Entwicklung zu erhöhen. Anstatt das Programm jedesmal neu zu starten, kann der Programmierer sofort nach der Veränderung das geänderte Verhalten der Anwendung beobachten.

Langlebige Server-Anwendungen. Kritische Server-Anwendungen, die nicht heruntergefahren werden dürfen, können nur mit dynamischer Code-Evolution verändert werden. Für diese Anwendung ist es wichtig, dass das Programm im Normalbetrieb nicht langsamer läuft und ein besonderes Augenmerk liegt auf der Korrektheit einer Veränderung.

Dynamische Sprachen. Dynamische Veränderungen zur Laufzeit sind ein fixer Bestandteil vieler dynamischer Sprachen und die Unterstützung von dynamische Code-Evolution auf VM-Ebene kann die Implementierung dieser dynamischer Sprachen vereinfachen.

Dynamische Aspekt-Orientierte Programmierung. Code-Evolution ist auch relevant für aspekt-orientierte Programmierung (AOP). Es gibt verschiedene dynamische AOP-Werkzeuge deren Limitierungen sich in den beschränkten Möglichkeiten zur Code-Evolution begründen [CST03, VBAM09]. Diese Werkzeuge profitieren sofort von den neuen Code-Evolution-Algorithmen.

Der Fokus unserer Implementierung ist die Unterstützung einer beschleunigten Programmentwicklung, da dies der populärste Anwendungsfall von Code-Evolution ist. Im Rahmen

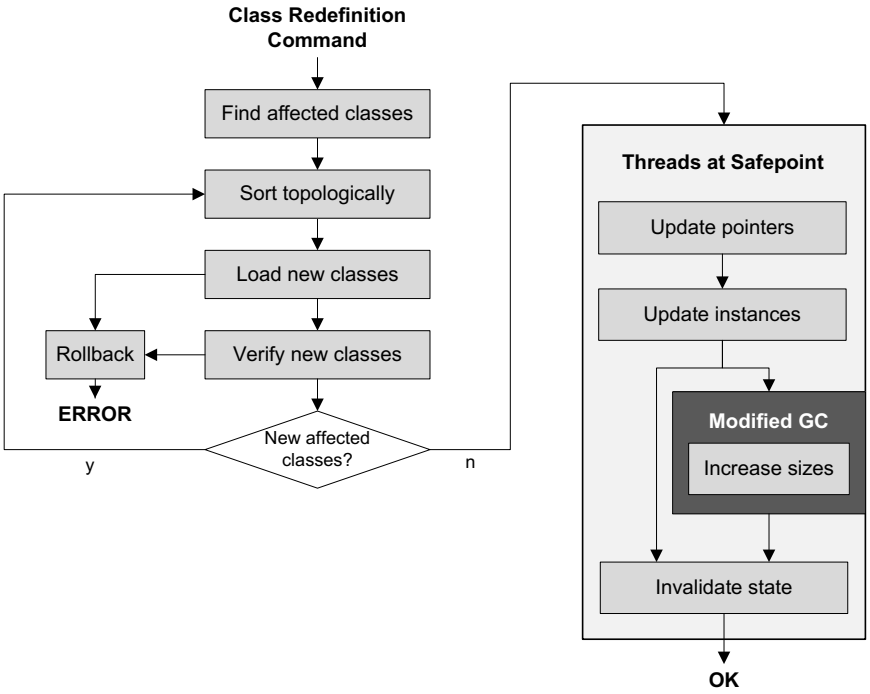


Abbildung 2: Schritte des Algorithmus zur Klassenredefinition.

der Arbeit gehen wir aber insbesondere auch auf die Herausforderungen in bezug auf Korrektheit einer Veränderung ein, die bei langlebigen Server-Anwendungen eine große Rolle spielt (siehe Abschnitt 6).

4 Algorithmus

Unser Algorithmus für unlimitierte Klassenredefinition ist als Modifikation einer Java VM implementiert. Die in dieser Arbeit vorgestellten generellen Konzepte sind jedoch generell auf virtuelle Maschinen anwendbar, die statisch typisierte und objekt-orientierte Programme ausführen. Für unseren Prototyp haben wir eine Produkt-VM anstatt einer Forschungs-VM gewählt, um sicherzugehen, dass die Evaluierungsergebnisse in einer Produkt-Umgebung gültig sind und die Algorithmen auch in einer hochoptimierten VM anwendbar sind.

Abbildung 2 gibt eine Übersicht der Schritte, die von unserem Algorithmus zur Klassenredefinition ausgeführt werden. Als ersten Schritt erzeugt der Algorithmus eine Liste der von der Redefinition betroffenen Klassen, die auch die Subklassen von redefinierten Klassen enthält. Diese Liste wird topologisch auf Basis der Subtyp-Beziehungen sortiert

und definiert im weiteren Verlauf die Reihenfolge in der die Klassen verarbeitet werden. Anschließend werden die neuen Klassen geladen und zum Typ-Universum der VM hinzugefügt. Sie bilden ein Neben-Universum, da die alten Klassen zu diesem Zeitpunkt noch im System verbleiben. Dieser Ansatz ermöglicht es auch, dass die Ausführung des alten Java-Programms zu diesem Zeitpunkt noch parallel zur Klassenredefinition erfolgen kann. Für das Laden und die Verifikation der neuen Klassen werden die normalen in der VM bereits implementierten Mechanismen verwendet. Falls die Verifikation einer Klasse fehlschlägt wird die Redefinition abgebrochen und der ursprüngliche Systemzustand wiederhergestellt. Bevor die Redefinition endgültig durchgeführt wird, muss noch einmal geprüft werden ob neue betroffene Klassen von der parallel laufenden Java-Applikation geladen wurden. In diesem Fall müssen auch sie redefiniert werden.

Zur Durchführung der Veränderung werden alle laufenden Java-Threads angehalten. Zusätzlich werden globale Locks benutzt um parallele Kompilierung oder paralleles Laden von Klassen zu verhindern. Anschließend werden in einer vollständigen Iteration über den Speicherbereich der VM alle Zeiger auf alte Klassen durch Zeiger auf neue Klassen ersetzt. Im Rahmen dieser Iteration wird auch das Layout von Objekten verändert deren Größe gleich geblieben oder sich verringert hat. Für die Anpassung von vergrößerten Objekten ist ein leicht modifizierter Speicherbereinigungs-Lauf durchgeführt, in dessen Rahmen die Objekte auf die neue Größe angepasst werden. Im nächsten Schritt werden innerhalb der VM in verschiedenen Bereichen Zustände invalidiert, die nicht mehr konsistent mit den neuen Klassendefinitionen sind. Zum Schluss werden alle Locks wieder freigegeben und die Ausführung der Java-Threads setzt mit der neuen Programm-Version fort.

5 Binär Inkompatible Veränderungen

Damit eine Veränderung binär kompatibel ist, muss jedes vorher gültige Klassenelement auch nach der Änderung weiterhin gültig sein. Weiters müssen alle Subtyp-Beziehungen erhalten bleiben. Binär kompatible Veränderungen einer Klasse sind: Das Hinzufügen von Feldern, Methoden oder implementierten Interfaces. Binär inkompatible Veränderungen einer Klasse sind: Das Entfernen von Feldern, Methoden oder Subtyp-Beziehungen. Die VM muss sich um eine sichere Lösung für binär inkompatible Veränderungen kümmern, da alter Code nach einer Versionsveränderung weiterhin ausgeführt werden kann. Dies ist möglich wenn zum Zeitpunkt der Veränderung alte Methoden noch aktiv sind.

Wir unterscheiden vier Lösungen für das Problem von entfernten Klassenelementen:

Statische Überprüfung: Eine statische Erreichbarkeitsanalyse überprüft ob die fortgeführte Programmausführung eine Instruktion erreichen kann, die auf ein entferntes Klassenelement verweist. In diesem Fall wird die Redefinition abgelehnt.

Dynamische Überprüfung: Die Redefinition wird immer durchgeführt und es wird sichergestellt, dass alte Methoden immer im Interpreter ausgeführt werden. Dieser überprüft zur Laufzeit ob gerade eine entsprechende in der neuen System-Version ungültige Instruktion ausgeführt wird und wirft dann eine Ausnahme.

Zugriff auf entfernte Klassenelemente: In dieser Lösung wird statt der Ausnahme auf das in der neuen Programmversion entfernte aber noch im System verfügbare Klassenelement zugegriffen. Diese Möglichkeit steht nicht für Instanzfelder zur Verfügung, da diese unwiederbringlich gelöscht werden.

Zugriff auf alte Klassenelemente: In allen bisherigen Konfigurationen wird ein Methodenaufruf immer auf die neueste Version einer Methode weitergeleitet. Dies kann jedoch zu Problemen führen wenn die alte Methode nicht mit der neuen aufgerufenen Methode kompatibel ist. Als mögliche Lösung unterstützt die DCE VM auch noch die dynamische Suche nach der exakt richtigen Version einer Methode aufgrund der Version der aufrufenden Methode.

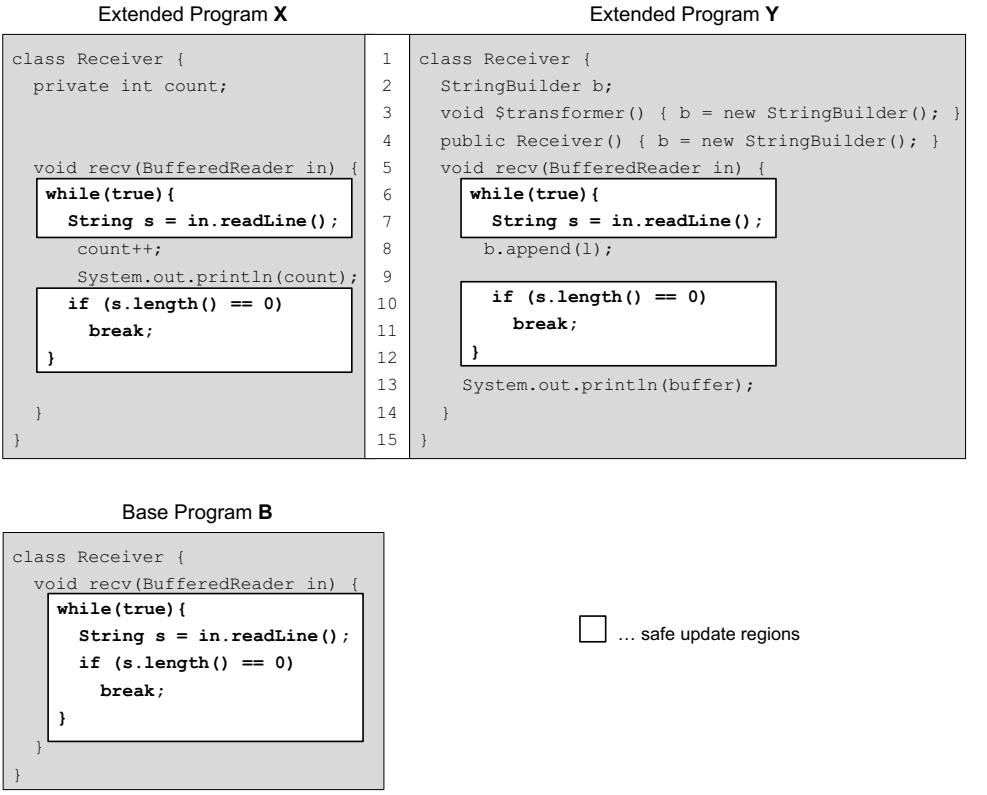
Es ist eine notwendige Invariante einer statisch typisierten VM, dass zu jedem Zeitpunkt der Typ eines Werts ein Subtyp des statisch deklarierten Typs des den Wert beinhaltenden Felds besitzt. Wenn das nicht mehr der Fall ist, kann Typsicherheit nicht mehr garantiert werden, was höchstwahrscheinlich zu einem Absturz der VM führt, wenn der Wert das nächste Mal benutzt wird. Bei der Entfernung einer Subtyp-Beziehung durch eine Klassenredefinition kann ein derartiger Fall auftreten. Um dennoch das Entfernen von Subtyp-Beziehungen nicht vollständig zu verbieten, setzt die DCE VM einen Algorithmus ein, der in einer Iteration über alle Werte die Einhaltung der Invariante überprüft. Sollte die Invariante nicht garantiert sein, wird die Klassenredefinition abgelehnt und der ursprüngliche Systemzustand wiederhergestellt.

6 Sicherer Versionswechsel

Die Techniken zur Klassenredefinition, die in den vorhergehenden Abschnitten beschrieben wurden, definieren die Veränderung eines Programms auf der Ebene von Feldern, Methoden und Supertypen. In diesem Abschnitt gehen wir einen Schritt weiter und betrachten die detailierten Unterschiede zwischen zwei Methodendefinitionen auf Bytecode-Ebene. Wir stellen eine neue Lösung für das Verändern von Java-Methoden, die zum Zeitpunkt der Redefinition aktiv sind, vor. Das ist insbesondere für Methoden mit langlaufenden oder endlosen Schleifen hilfreich, die ansonsten nie verändert werden könnten.

Wir definieren ein Programmiermodell, das uns den Wechsel zwischen einem Basisprogramm und einem erweiterten Programm erlaubt. Das erweiterte Programm muss vom Basisprogramm abgeleitet sein: Es darf sich durch hinzugefügte Klassenelemente, neue Klassen und hinzugefügte Bytecode-Abschnitte vom Basisprogramm unterscheiden. Der Wechsel vom Basisprogramm zum erweiterten Programm kann zu einem beliebigen Zeitpunkt erfolgen. Beim Wechsel zurück zum Basisprogramm stellen wir sicher, dass sich alle ausführenden Threads außerhalb der hinzugefügten Bytecode-Abschnitte befinden.

Die vielen Restriktionen der Unterschiede zwischen dem Basisprogramm und dem erweiterten Programm verhindern viele mögliche Versionswechsel. Es ist jedoch möglich sowohl das aktuell ausführende Programm `X` als auch das neue Programm `Y` als erweiterte Programme eines virtuellen gemeinsamen Basisprogramms `B` zu betrachten. Auf diese




```
class Receiver {
  private int count;

  void recv(BufferedReader in) {
    while(true){
      String s = in.readLine();
      count++;
      System.out.println(count);
      if (s.length() == 0)
        break;
    }
  }
}
```

```
1  class Receiver {
2    StringBuilder b;
3    void $transformer() { b = new StringBuilder(); }
4    public Receiver() { b = new StringBuilder(); }
5    void recv(BufferedReader in) {
6      while(true){
7        String s = in.readLine();
8        b.append(l);
9
10       if (s.length() == 0)
11         break;
12     }
13     System.out.println(buffer);
14   }
15 }
```

```
class Receiver {
  void recv(BufferedReader in) {
    while(true){
      String s = in.readLine();
      if (s.length() == 0)
        break;
    }
  }
}
```

Abbildung 3: Wechsel zwischen zwei Programmversionen.

Weise kann dann die Transformation von `X` nach `Y` als eine Transformation von `X` nach `B` und von `B` nach `Y` angesehen werden.

Abbildung 3 zeigt zwei verschiedene Programme mit einer Schleife, in der Daten empfangen werden. Das Programm `X` auf der linken Seite zählt die Anzahl der Textzeilen und das Programm `Y` auf der rechten Seite speichert die Textzeilen in einem `StringBuilder`-Objekt. Das gemeinsame Basisprogramm ist weiß markiert. Alle notwendigen Restriktionen zwischen den beiden erweiterten Programmen und dem Basisprogramm werden befolgt, deshalb kann auf sichere Art und Weise zwischen den beiden Programmversionen gewechselt werden. Das Basis-Programm wird selbst nie ausgeführt, es dient lediglich als Definition für die sicheren Code-Regionen in denen der Wechsel stattfinden darf.

	DYMOS	Ginseng	Upstare	CLOS	Smalltalk	JDrums	DVM	JVolve	Hotswap	DCE VM
Austauschen von Methoden	X	X	X	X	X	X	X	X	X	X
Methoden Hinzufügen/Entfernen	X	X	X	X	X	X	X	X		X
Felder Hinzufügen/Entfernen	X	X	X	X	X	X	X	X		X
Klassenhierarchie-Veränderungen				X	X					X
Atomare Klassenredefinition								X	X	X
Veränderung Aktiver Methoden			X							X
Kein Performance-Verlust				X	X			X	X	X
Entfernung Virtueller Methoden										X
Literatur-Referenz	[CL83]	[NHSO06]	[MB09]	[Ste90]	[GR83]	[AR00]	[MPG+00]	[SHM09]	[Dmi01]	

Tabelle 1: Feature comparison of systems that allow dynamic changes to running programs.

7 System-Vergleich

Tabelle 1 vergleicht die Funktionalitäten verschiedener dynamischer Code-Evolutions-Systeme mit der DCE VM. Die meisten Systeme erlauben die Veränderung von Methoden und Feldern eines Programms. Im Fall von prozeduralen Systemen wie zum Beispiel DYMOS werden anstelle von Feldern globale Datenbereiche verändert.

Veränderungen der Klassenhierarchie sind nur in Systemen erlaubt, die mit dem Konzept von Metaklassen-Definitionen ausgestattet sind (z.B., CLOS und Smalltalk). Diese Systeme unterstützen jedoch keine atomare Redefinition von mehreren Klassen wie die DCE VM. Die Möglichkeit zum Verändern von Methoden, die gerade aktiv sind, bietet Gingseng. Während das System von Gingseng flexibler in der Art der Veränderung ist, bietet es keine Möglichkeit, Aussagen über die semantische Korrektheit einer Veränderung zu machen und unterstützt auch nicht das Konzept von sicheren Update-Regionen.

Dynamische Code-Evolution kann ohne Performance-Verlust unterstützt werden, wenn ein Programm in einer VM ausgeführt wird (z.B., CLOS, Smalltalk, JVolve, Hotswap und DCE VM). Andere Techniken resultieren in teilweise in signifikanten Performance-Verlusten (z.B., 38.5% für Ginseng). Die Möglichkeit, bereits entfernte Methoden aufgrund der Version der derzeit ausgeführten Methode und des Aufrufadressaten auszuführen, ist nur in der DCE VM verügbar.

8 Zusammenfassung

Die vorliegende Dissertation enthält folgende wissenschaftliche Beiträge:

- Wir beschreiben einen neuen Algorithmus zur Klassenredefinition in einer Java VM (siehe Abschnitt 4).
- Wir erlauben das Hinzufügen und Entfernen von Feldern und Methoden zur Laufzeit und unterstützen auch die Veränderung der Klassenhierarchie.
- Wir diskutieren die möglichen Probleme, die durch binär inkompatible Veränderungen ausgelöst werden.
- Wir beschreiben eine Lösung für das Problem von gelöschten Klassenelementen (siehe Abschnitt 5).
- Wir schlagen einen Algorithmus zur Prüfung der Typsicherheit im Fall von entfernten Supertypen vor (siehe Abschnitt 5).
- Wir präsentieren ein eingeschränktes Programmiermodell für sichere dynamische Updates von Java-Programmen (siehe Abschnitt 6).
- Wir beschreiben drei verschiedene Fallstudien, die mögliche Anwendungsbereiche aufzeigen.
- Wir zeigen, dass unser Ansatz keinen Performance-Verlust für das ausgeführte Java-Programm vor oder nach der Klassenredefinition bedeutet.
- Wir evaluieren die Performance des Algorithmus zur Veränderung von Objektinstanzen an ausgewählten Micro-Benchmarks.

Der Hauptbeitrag dieser Arbeit ist die Dynamic Code Evolution VM (DCE VM). Nach unserem besten Wissen ist die DCE VM die erste VM für eine statisch typisiert objektorientierte Sprache, die unlimitierte Unterstützung von Klassenredefinitionen bietet, ohne die Ausführungsgeschwindigkeit zu beeinträchtigen. Die DCE VM hat signifikantes Interesse unter Java-Entwicklern hervorgerufen. Es gibt auch bereits Pläne, die Modifikationen in den Quelltext der HotSpot VM zu übernehmen. Dies würde die unlimitierte Code-Evolution mehreren Millionen Java-Entwicklern verfügbar machen. Ein Prototyp der DCE VM mit Binärdaten und Quelltext können von `http://ssw.jku.at/dcevm/` heruntergeladen werden.

Literatur

[AR00] Jesper Andersson und Tobias Ritzau. Dynamic code update in JDrums. In *Workshop on Software Engineering for Wearable and Pervasive Computing*, 2000.

[CL83] Robert P. Cook und Insup Lee. DYMOS: A Dynamic Modification System. *SIGSOFT Software Engineering Notes*, 8:201–202, March 1983.

[CST03] Shigeru Chiba, Yoshiki Sato und Michiaki Tatsubori. Using HotSwap for Implementing Dynamic AOP Systems. In *Proceedings of the Workshop on Advancing the State-of-the-Art in Run-time Inspection*, 2003.

[Dmi01] Mikhail Dmitriev. *Safe Class and Data Evolution in Large and Long-Lived Java Applications*. Dissertation, University of Glasgow, 2001.

[Fab76] Robert S. Fabry. How to Design a System in Which Modules Can Be Changed on the Fly. In *Proceedings of the International Conference on Software Engineering*, Seiten 470–476. IEEE Computer Society, 1976.

[GR83] Adele Goldberg und David Robson. *Smalltalk-80: the Language and its Implementation*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1983.

[MB09] Kristis Makris und Rida Bazzi. Immediate Multi-threaded Dynamic Software Updates Using Stack Reconstruction. In *Proceedings of the USENIX Annual Technical Conference*. USENIX Association, 2009.

[MPG+00] Scott Malabarba, Raju Pandey, Jeff Gragg, Earl Barr und J. Fritz Barnes. Runtime Support for Type-Safe Dynamic Java Classes. In *Proceedings of the European Conference on Object-Oriented Programming*, Seiten 337–361. Springer-Verlag, 2000.

[NHSO06] Iulian Neamtiu, Michael Hicks, Gareth Stoyle und Manuel Oriol. Practical Dynamic Software Updating for C. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*. ACM Press, 2006.

[Ora11] Oracle Corporation. *Top 25 RFEs (Requests for Enhancements)*, 2011. `http://bugs.sun.com/top25_rfes.do`.

[SHM09] Suriya Subramanian, Michael Hicks und Kathryn S. McKinley. Dynamic Software Updates: a VM-Centric Approach. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, Seiten 1–12. ACM Press, 2009.

[Ste90] Guy L. Steele, Jr. *Common LISP: the Language*. Digital Press, second. Auflage, 1990.

[VBAM09] Alex Villazón, Walter Binder, Danilo Ansaloni und Philippe Moret. Advanced Runtime Adaptation for Java. In *Proceedings of the International Conference on Generative Programming and Component Engineering*, Seiten 85–94. ACM Press, Oktober 2009.

Thomas Würthinger wurde am 1. September 1986 geboren. Er studierte Informatik an der Johannes Kepler Universität und absolvierte sein Doktoratsstudium 2011. Die Promotion erfolgte unter den Auspizien des österreichischen Bundespräsidenten. Während seines Studiums absolvierte er Auslandspraktika bei Sun Microsystems, Oracle und Google. Seit April 2011 arbeitet er im Oracle-Forschungslabor. Seine vorrangigen Forschungsgebiete sind Übersetzerbau, dynamische Programmanalyse und virtuelle Maschinen. Er ist Leiter des OpenJDK-Projekts "Graal".

Theorie künstlicher Immunsysteme

Christine Zarges

Department of Computer Science, University of Warwick
Coventry CV4 7AL, United Kingdom
zarges@dcs.warwick.ac.uk

Abstract: Künstliche Immunsysteme sind adaptive Systeme, die sich bezüglich ihrer Komponenten und Funktionsweisen an Theorien natürlicher Immunsysteme orientieren und diese nachahmen. Wir analysieren verschiedene Mechanismen, die dabei typischerweise zum Einsatz kommen. Mit dieser Arbeit stellen wir uns damit einer der größten Herausforderungen des Gebietes und leisten einen signifikanten Beitrag zu dessen Weiterentwicklung. Wir untersuchen zwei zentrale Komponenten künstlicher Immunsysteme, Variation und Alterung. Da unsere theoretischen Analysen zum Verständnis der verwendeten Verfahren in der Praxis beitragen sollen, liegt unser Fokus auf der praktischen Relevanz. Einem grundlegenden Aspekt dieser Fragestellung ist abschließend ein eigener Abschnitt gewidmet.

1 Einleitung

Unter künstlichen Immunsystemen versteht man eine Klasse von der Natur inspirierter Algorithmen, die sich an Prozessen in natürlichen Immunsystemen orientieren [dCT02]. Ihre Entwicklung und Untersuchung ist ein noch recht junges Forschungsgebiet, das sich zum Einen der Computational Intelligence und zum Anderen dem Gebiet der randomisierten Suchheuristiken zuordnen lässt. Weitere bekannte Suchheuristiken sind beispielsweise evolutionäre Algorithmen, Ameisen- und allgemein Schwarmsysteme oder auch simulierte Abkühlung und randomisierte lokale Suche. Im Gegensatz zu diesen anderen Verfahren ist das Gebiet der künstlichen Immunsysteme jedoch weitergehender oder enger verzahnt mit der Forschung innerhalb der Immunologie. Aus diesem Grund lassen sich zwei wesentliche Teilaspekte der Erforschung künstlicher Immunsysteme identifizieren: zum Einen die Modellierung von natürlichen Immunsystemen mit Hilfe informatischer Methoden mit dem Ziel, die Arbeitsweise natürlicher Immunsysteme besser zu verstehen; zum Anderen die Entwicklung immun-inspirierter Verfahren zur Problemlösung. Typische Anwendungen künstlicher Immunsysteme sind Lernmethoden, Klassifikation, Anomalie-Erkennung sowie Optimierung. Wir beschäftigen uns mit diesem zweiten Aspekt künstlicher Immunsysteme. Der Fokus liegt auf künstlichen Immunsystemen, die dem Vorbild der klonalen Selektion folgen und in der Optimierung eingesetzt werden. Optimierung ist eine der wichtigsten Anwendungen randomisierter Suchheuristiken.

Ein oft genanntes Problem in diesem Bereich der künstlichen Immunsysteme war das Fehlen einer theoretischen Fundierung. Ein Grund ist insbesondere die Tatsache, dass die

meisten Algorithmen auf der direkten Anwendung einzelner Immunprinzipien auf das vorliegende Problem basieren. Es existierten lediglich einige wenige Konvergenzanalysen. Ergebnisse über erwartete Laufzeiten der Algorithmen oder theoretische Arbeiten, die die Funktionsweise der Algorithmen erklären, fehlten vollständig. Um eine strukturierte Entwicklung dieser Algorithmen und Verfahren weiter voranzutreiben, ist jedoch ein theoretisches Verständnis der grundlegenden Elemente unverzichtbar. Dieses Problem wurde von führenden Wissenschaftlern im Bereich der künstlichen Immunsysteme erkannt und in einem wegweisenden Positionspapier als eine der grundlegenden und wichtigsten Herausforderung für die Zukunft des Gebietes hervorgehoben [THSC08]. Hier liegt der Fokus insbesondere auf dem Verständnis der einzelnen Komponenten von künstlichen Immunsystemen, mit dem Ziel Vorhersagen darüber zu treffen, für welche Problemklassen bestimmte Algorithmen besonders vielversprechend sind.

Wir stellen uns genau dieser Fragestellung und leisten damit einen entscheidenden und grundlegenden Beitrag zur Weiterentwicklung des Forschungsgebietes. Als erster derartiger Beitrag schaffen wir die Grundlage für weitergehende Forschungsarbeiten. Aufgrund der Ähnlichkeit zu anderen von der Natur inspirierten Verfahren ist es wünschenswert, künstliche Immunsysteme mit anderen solchen Verfahren zu vergleichen und damit allgemeine Resultate im Bereich randomisierter Suchheuristiken zu erzielen. Deshalb orientiert sich unsere Methodik an Analysen aus diesem Gebiet und vergleicht Ergebnisse mit vorhanden Ergebnissen zu anderen randomisierten Suchheuristiken.

Um Vergleichbarkeit zu gewährleisten, betrachten wir künstliche Immunsysteme in einem allgemeinen Rahmen, der sich an allgemeinen randomisierten Suchheuristiken orientiert. Dieses allgemeine Schema ist in Abbildung 1 skizziert. Wir beschränken unsere Betrachtungen auf pseudo-boolesche Optimierungsprobleme der Dimension n, d. h. der Optimierung von Funktionen $f\colon \{0,1\}^n \to \mathbb{R}$. Der betrachtete Algorithmus ist rundenbasiert und verwaltet eine in der Regel rein zufällig initialisierte Menge von μ Suchpunkten (Population). In jeder Runde des Algorithmus wird ein Nachkomme erzeugt. Dazu werden einer oder mehrere der Suchpunkte ausgewählt (Elter), variiert und bewertet. Die Variation kann dabei entweder eine Mutation, d. h. eine zufällige Veränderung eines einzelnen Suchpunktes, oder eine Rekombination mehrerer Suchpunkte sein. Anschließend werden aus den ursprünglichen Suchpunkten sowie des Nachkommen basierend auf deren Bewertung (Fitness) μ Suchpunkte für die nächste Runde ausgewählt. Dieses Vorgehen wird solange wiederholt, bis ein vorher definiertes Abbruchkriterium erfüllt ist.

Meist ist man bei der Analyse derartiger Algorithmen insbesondere an der Zeit interessiert, die benötigt wird, um das erste mal einen optimalen Suchpunkt zu finden. Das erlaubt das Abbruchkriterium in der Analyse zu ignorieren. Da angenommen wird, dass die Funktionsauswertung die teuersten Operationen des Algorithmus darstellen, beschränkt man sich auf die Analyse der Anzahl an Funktionsauswertungen bzw. Runden, die benötigt werden, um ein Optimum zu erreichen. Diese Anzahl bezeichnet man als Optimierzeit.

Um bestimmte Eigenschaften eines Algorithmus herauszuarbeiten, betrachtet man meist Beispielfunktionen. Im Rahmen der folgenden Betrachtungen ziehen wir hierzu zum Einen bekannte und häufig betrachtete Beispielfunktionen zu Rate, um vergleichende Ergebnisse zu anderen Algorithmen zu erhalten. Zum Anderen konstruieren wir eigene Funktionen, um Eigenschaften der betrachteten Operatoren exemplarisch herauszuarbeiten.

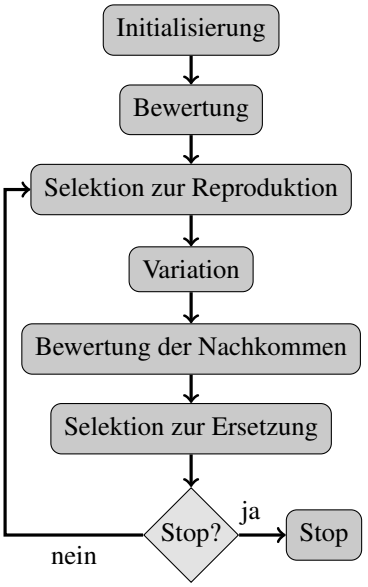


Abbildung 1: Allgemeines Schema einer randomisierten Suchheuristik.

Im Folgenden geben wir einen Überblick über zentrale Ergebnisse und einen Ausblick auf mögliche weitergehende Arbeiten. Wir betrachten zunächst immun-inspirierte Variationsoperatoren, d. h. Mutationsoperatoren, die in realen künstlichen Immunsystemen zum Einsatz kommen (Abschnitt 2). Anschließend befassen wir uns mit dem Konzept des Alterns, einem in vielen künstlichen Immunsystemen wichtigen Diversitätsmechanismus (Abschnitt 3). Ein Schwerpunkt liegt dabei auf dem Verständnis der betrachteten Konzepte und der Anwendbarkeit der theoretischen Ergebnisse in der Praxis. Hierbei gilt es insbesondere verschiedene Parametrisierungen zu vergleichen und so Leitfäden für die weitergehende Anwendung zu erarbeiten. Dieser praktischen Relevanz theoretischer Ergebnisse ist zum Abschluss ein eigener Teil gewidmet (Abschnitt 4). Die Darstellung hier konzentriert sich auf eine verständliche Darstellung der zentralen Ergebnisse und Folgerungen, die aus diesen gezogen werden können. Konkrete (asymptotische) Ergebnisse, vollständige Definitionen und rigoros bewiesene Theoreme können in [Zar11] nachgelesen werden. Die interessanten analytischen und methodischen Probleme bei der Betrachtung künstlicher Immunsysteme werden ebenfalls nur dort diskutiert.

2 Mutation in künstlichen Immunsystemen

Im Gegensatz zu anderen randomisierten Suchheuristiken kommen bei künstlichen Immunsystemen meist große Mutationswahrscheinlichkeiten zum Einsatz. Man spricht von Hypermutation. Basierend auf immunologischen Theorien existiert eine Vielzahl derartiger Hypermutationen. Besonders verbreitet sind sogenannte invers fitness-proportionale

Mutationen. Wir betrachten im Folgenden diese Gruppe der Hypermutationen sowie das relativ neue Konzept der zusammenhängenden Hypermutationen. Abschließend beschäftigen wir uns allgemein mit den Effekten großer Mutationswahrscheinlichkeiten.

Invers fitness-proportionalen Mutationen liegt die Idee zugrunde, dass bereits „gute“ Zellen weniger stark mutiert werden sollten als „schlechte“ Zellen. Zur Umsetzung dieser Idee existieren verschiedene konkrete Implementierungen. Den hier betrachteten Varianten ist gemeinsam, dass sie die Mutationswahrscheinlichkeit, d. h. die Wahrscheinlichkeit mit der ein einzelnes Bit des Suchpunktes kippt, durch eine von der Fitness abhängigen Funktion beschreiben. Dabei stellen die adaptive Mutation, welche die invers fitness-proportionale Idee direkt umsetzt, und eine auf dem Hamming-Abstand basierende Mutation die einfachsten und im Grunde nicht immun-basierten Methoden der Implementierung dar. Im Gegensatz dazu sind CLONALG und opt-aiNet die in den entsprechend benannten Immunalgorithmen verwendeten Mutationsoperatoren. Wir analysieren das Verhalten dieser Operatoren auf einer sehr einfachen und weit verbreiteten Beispielfunktion namens ONEMAX, die die Anzahl der Einsen in einem gegebenen Bit-String maximiert. Diese Beispielfunktion ist häufig der Startpunkt theoretischer Analyse, da sie zum Einen einfach zu optimieren ist und zum Anderen exemplarisch aufzeigt, ob ein betrachteter Algorithmus grundsätzlich in der Lage ist, zum Beispiel eine lokale Suche nachzuahmen.

Wir stellen fest, dass die adaptive Mutationswahrscheinlichkeit keine Probleme mit der Optimierung der betrachteten Funktion hat, wohingegen die auf dem Hamming-Abstand basierende Methode mit hoher Wahrscheinlichkeit nicht in der Lage ist, diese sehr einfache Beispielfunktion effizient zu optimieren. Dies liegt an zu großen Mutationswahrscheinlichkeiten, die lokale Suche nicht mehr simulieren können. Dies stellt insbesondere ein Problem dar, wenn der Algorithmus sich bereits dem Optimum angenähert hat und nur noch kleinere lokale Änderungen nötig sind.

Ähnliche Effekte lassen sich bei den beiden betrachteten immun-inspirierten Operatoren beobachten. Hierbei ist allerdings zu beachten, dass die Funktionsweise der Operatoren erheblich von der gewählten Parametrisierung abhängt. Zum Einen muss ein Anwender über die Größe eines Dämpfungsparameter entscheiden, zum Anderen stellt sich die Frage, auf welche Art und Weise die hier notwendige Normalisierung der Fitness durchgeführt wird. Wir betrachten verschiedene Werte für Dämpfungsparameter sowie zwei unterschiedliche Normalisierungsmethoden. Bei der ersten Methode gehen wir davon aus, dass wir nur eine Population der Größe 1 betrachten und zur Normalisierung den optimalen Funktionswert zur Rate ziehen. Bei der zweiten Methode betrachten wir allgemeine Populationen der Größe μ und verwenden jeweils den aktuell besten bekannten Funktionswert zur Normalisierung. Diese Methodik wird in der Praxis meist verwendet. Wir beweisen, dass die Verwendung einer größeren Population und dieser zweiten Normalisierungsmethode entscheidend für den Erfolg der CLONALG-Mutation ist, wohingegen opt-aiNet für beide Varianten gute Ergebnisse erzielen kann. In beiden Fällen ist eine angemessene Wahl für den Dämpfungsparameter entscheidend.

Für Hypermutationen ist es also entscheidend, dass die Mutationswahrscheinlichkeit zumindest in der Nähe des Optimums nicht zu groß wird. Sonst kann man globale Optima nicht exakt finden. Hypermutationen sind also eher dazu geeignet, „robuste“ Lösungen zu finden, da sie große lokale Optima gegenüber isolierten globalen Optima bevorzugen.

Die praktische Relevanz der theoretischen Ergebnisse ist ein zentraler Aspekt unserer Betrachtungen. Aus diesem Grund werden für die meisten Resultate begleitende Experimente durchgeführt. Im Fall von CLONALG bringen diese experimentellen Betrachtungen wichtige zusätzliche Erkenntnisse. Versuche mit verschiedenen Suchraumdimensionen belegen, dass der Operator mit passendem Dämpfungsparameter auch für die einfache Normalisierungsmethode funktionieren kann. Insbesondere ist bis zu einer Dimension von 10^5 kaum ein Unterschied zum optimalen mutations-basierten evolutionären Algorithmus zu erkennen. Dies erklärt, warum der Operator bislang trotz seiner potenziellen Probleme in der Praxis erfolgreich eingesetzt wurde.

Im Gegensatz zu invers fitness-proportionalen Mutationen bestimmen zusammenhängende Mutationen keine fitness-abhängige Funktion für die Mutationswahrscheinlichkeit. Sie legen einen Teilbereich des betrachteten Suchpunktes, in dem die Mutation stattfindet, fest. Innerhalb dieses Bereiches wird dann jedes Bit mit einer festgelegten Wahrscheinlichkeit p gekippt. Der restliche Teil des Suchpunktes bleibt unverändert. Wir betrachten drei verschiedene Instanziierung dieser Idee. Bei der ersten Variante wählen wir zwei zufällige Positionen im Bit-String und betrachten den durch diese Positionen begrenzten Bereich. Bei der zweiten und dritten Variante wählen wir hingegen zufällig einen Startpunkt und eine Länge für den Bereich, wobei der Bereich bei der zweiten Variante am Ende des Bit-Strings abgeschnitten wird und die dritte Variante zyklisch ist.

Trotz ihrer Ähnlichkeit sind zentrale Unterschiede zwischen den Varianten zu beobachten. Während Variante 1 eine Tendenz hat, Bits in der Mitte des Bit-Strings zu kippen, weist Variante 2 eine Tendenz zu Bits am Ende des Bit-Strings auf. Variante 3 hingegen hat keine derartige Tendenz und kippt jedes Bit mit gleicher Wahrscheinlichkeit. Des Weiteren ist festzuhalten, dass die bei Variante 1 beobachtete Tendenz im Gegensatz zu Variante 2 symmetrisch ist. Variante 2 weist eine höhere Wahrscheinlichkeiten für 1-Bit-Mutationen am Ende des Bit-Strings auf. Da 1-Bit-Mutationen für die Optimierung zentral sein können, ist eine derartige Tendenz ohne zusätzliches Problemwissen nicht wünschenswert. Dies gilt insgesamt, so dass in der Regel die Verwendung von Variante 3 vorzuziehen ist.

Eine weitere Eigenschaft der betrachteten Operatoren ist, dass die Wahl eines Extremwertes für p im Allgemeinen nicht empfehlenswert ist, da beispielsweise für $p = 1$ keine Konvergenz garantiert werden kann. In diesem Fall unterliegt der gesamte ausgewählte Bereich einer Mutation, was zu einer Stagnation des Optimierungsprozesses führen kann.

Wir stellen fest, dass zusammenhängende Mutationen trotz ihrer sehr hohen Mutationswahrscheinlichkeit keine Probleme mit der zuvor betrachten Beispielfunktion ONEMAX haben, weil man mit nicht zu geringer Wahrscheinlichkeit 1-Bit-Mutationen durchführt. Allerdings ist man bei Problemen, bei denen dies entscheidend ist, im Vergleich zu Standardmutation in evolutionären Algorithmen bzw. lokaler Suche langsamer. Andererseits weisen zusammenhängende Mutationen enorme Vorteile bei Problemen auf, bei denen Mehr-Bit-Mutationen zentral sind, so dass solche Operatoren da vorzuziehen sind. Wir erkennen, dass eine Kombination aus Standardmutationen aus evolutionären Algorithmen sowie zusammenhängenden Mutationen zu robusteren Algorithmen führen kann.

Wir widmen uns abschließend allgemeinen Konsequenzen großer Mutationswahrscheinlichkeiten. Wir betrachten die Funktionsklasse der strikt monotonen Funktionen, d. h. Funk-

tionen, bei denen sich der Funktionswert erhöht, wenn ausschließlich Nullen zu Einsen gekippt werden. Werden sowohl Nullen als auch Einsen gekippt, führt dies zu unvergleichbaren Suchpunkten, bei denen die Fitness beliebig festgelegt werden kann. Wir untersuchen Mutationswahrscheinlichkeiten der Form c/n für eine Konstante c und interessieren uns für den Einfluss des Parameters c auf die Optimierzeit.

Der bisher betrachtete Algorithmus mit Populationsgröße 1 hat mit $c \leq 1$ eine polynomielle Optimierzeit für jede monotone Funktion. Die Erhöhung von c lässt die Optimierzeit für einige monotone Funktionen von polynomiell auf exponentiell steigen. Das zeigt erstmalig, dass bereits die Erhöhung der Mutationswahrscheinlichkeit um einen konstanten Faktor drastische Auswirkungen auf die Optimierzeit haben kann. Darum ist bei der Verwendung von Hypermutationen besondere Vorsicht geboten.

3 Alterungsmechanismen

In künstlichen Immunsystemen orientieren sich Alterungsmechanismen an der endlichen Lebensdauer von Immunzellen. Wir betrachten das weit verbreitete Konzept des statischen Alterns, bei dem jeder Suchpunkt ein Alter hat, das in jeder Runde um eins wächst. Ein Parameter bestimmt die maximale Lebensdauer eines Suchpunktes. Neue Suchpunkte erhalten Alter 0, falls sie eine Verbesserung gegenüber ihren Eltern darstellen, sonst erben sie das Alter. Verkleinert sich die Population durch das Entfernen zu alter Zellen, wird sie mit rein zufälligen neuen Suchpunkten mit Alter 0 aufgefüllt. Diesen Mechanismus aus dem Bereich künstlicher Immunsysteme vergleichen wir mit ähnlichen Mechanismen aus anderen randomisierten Suchheuristiken und betrachten einen evolutionären Alterungsmechanismus. Der zentrale Unterschied zum statischen Altern ist, dass ein neuer Suchpunkt in jedem Fall Alter 0 zugewiesen bekommt.

Es ist leicht einzusehen, dass die Wahl der Lebensdauer entscheidend für die Performanz ist. Diese Wahl ist problemabhängig und sehr schwierig. Man kann aber zentrale Eigenschaften für die beiden betrachteten Operatoren festhalten.

Ein Alterungsmechanismus kann nur dann Einfluss auf das Verhalten des Algorithmus haben, wenn die maximale Lebensdauer nicht so groß gewählt wurde, dass kein Punkt der Population sie je erreicht. Auf der anderen Seite muss sie ausreichend groß gewählt werden, damit dem Algorithmus genügend Zeit für die Verbesserung der Suchpunkte bleibt. Dies ist insbesondere für das statische Altern entscheidend, da neue Suchpunkte hier das Alter des Elter erben können. So kann es dazu kommen, dass sämtliche Suchpunkte in der Population das gleiche Alter haben und gemeinsam aussterben. Dies entspricht einem Neustart des Algorithmus. Ein derartiger Neustart ist wünschenswert, falls der Algorithmus in einem lokalen Optimum stecken geblieben ist. Passiert ein derartiger Neustart aber zu früh, gleicht der Algorithmus einer rein zufälligen Suche und kann kaum noch optimieren. Die maximale Lebensdauer für statisches Altern muss daher so gewählt sein, dass mindestens lokale Verbesserungen der Suchpunkte möglich sind.

Beim evolutionären Altern ist die Perspektive eine leicht andere. Hier kann es ausreichend sein, Kopien eines aktuell besten Suchpunktes zu erzeugen, da hier Nachkommen in je-

dem Fall Alter 0 erhalten. Hier muss also das Alter groß genug sein, um aktuell beste Suchpunkte zu kopieren, so lange lokale Verbesserungen möglich sind.

Beim Vergleich der beiden Operatoren in typischen Situationen fällt auf, dass das statische Altern in der Lage ist, Neustarts zu simulieren, da das Ausbleiben von Verbesserungen dazu führt, dass irgendwann alle Suchpunkte gleich alt sind. Beim evolutionären Altern ist dies mit hoher Wahrscheinlichkeit nicht der Fall, da jeder Nachkomme Alter 0 erhält, was eine größere Altersdiversität impliziert. Diese Altersdiversität ist von Vorteil, wenn der Algorithmus während des Optimierungsprozesses auf ein sogenanntes Plateau trifft, einen Bereich im Suchraum, in dem benachbarte Punkte gleiche Fitness haben. Im Gegensatz zum statischen Altern erlaubt das evolutionäre Altern hier einen zufälligen Lauf auf dem Plateau und so sein Durchschreiten. Statisches Altern beobachtet lediglich die Fitness der Punkte und behandelt diese Situation identisch zu einem lokalen Optimum. Bei nicht sehr kleinen Plateaus geschieht also ebenfalls ein Neustart, so dass sie zu unüberwindlichen Hindernissen werden.

Eine zentrale Beobachtung ist, dass die Vorteile beider Varianten des Alterns miteinander kombiniert werden können. Hierzu ist es ausreichend, das statische Altern so zu erweitern, dass ein Nachkomme nur das Alter seines Elter erbt, falls er entweder eine Verschlechterung oder eine Kopie darstellt. So erben unterschiedliche Punkte mit gleicher Fitness –wie auf dem Plateau– ihr Alter nicht. Diese einfache Ergänzung führt zu einem beweisbar besseren Operator, der in der Praxis vorzuziehen ist.

Als wichtigen Vorteil von Alterungsmechanismen haben wir die Fähigkeit, Neustarts zu simulieren, erkannt. Neustarts sind ein weit verbreitetes und oft erfolgreiches Konzept, das sich leichter und effizienter auch direkt implementieren lässt. Die Fähigkeit Neustarts zu simulieren sollte darum nicht der zentrale Vorteil von Alterungsmechanismen sein.

Wir nehmen eine strukturiertere Sichtweise auf Alterungsmechanismen ein und betrachten andere potenzielle Vorteile sowie das Zusammenspiel mit anderen Teilen der Algorithmen. Hierbei ist insbesondere die Selektion zur Ersetzung spannend, da diese entscheidenden Einfluss auf die Altersdiversität hat. Wir betrachten sogenannte partielle Neustarts, d. h. Runden, in denen nur ein Teil der Suchpunkte ausstirbt und durch zufällige neue Suchpunkte ersetzt wird. Diese neuen Punkte sind in der Regel deutlich schlechter als die bereits über einen längeren Zeitraum verbesserten Suchpunkte, so dass sie meist nur für einen kurzen Zeitraum in der Population verweilen und rasch durch Kopien besser Suchpunkte ersetzt werden. Falls dies dazu führt, dass nach einer Weile alle Punkte gleich alt sind, kommt es wieder zu einem vollständigen Neustart. Für einen partiellen Neustart braucht man Altersdiversität. Dann können Alterungsmechanismen zum Beispiel in Algorithmen, die neben Mutation auch Rekombination mehrerer Suchpunkte verwenden, Vorteile bringen. Wir demonstrieren das für Probleme, bei denen die Rekombination eines lokalen Optimums mit einem rein zufälligen Punkt zu signifikanten Verbesserung führen kann und weisen immense Effizienzsteigerungen nach.

Wir untersuchen mehrere Varianten der Selektion zur Ersetzung und ihre Fähigkeiten eine für unser Ziel ausreichende Altersdiversität sicherzustellen. Hierzu betrachten wir einen Algorithmus, der in jeder Runde einen neuen Suchpunkt mittels Mutation oder Rekombination erzeugt. Wir untersuchen Diversitätsmechanismen, die genau dann zum Einsatz

kommen, wenn es mehrere „schlechteste“ Suchpunkte gibt, von denen einer eliminiert werden muss, und zeigen, dass bereits relativ einfache Mechanismen den gewünschten Effekt erzielen können. Hierzu gehören insbesondere die Methode, einen Suchpunkt zu eliminieren, dessen Alter am häufigsten in der aktuellen Population vorkommt. Eine weitere erfolgsversprechende Methode entfernt einen Suchpunkt mit minimaler Altersdifferenz zu dem Suchpunkt, der in der aktuellen Runde erzeugt wurde. Die Eliminierung eines rein zufälligen schlechten Punktes reicht hingegen nicht aus.

In experimentellen Untersuchungen wird deutlich, dass bei derartigen Anwendungen größere Populationsgrößen helfen, da auf diese Weise mehr partielle Neustarts durchgeführt werden. Vorher waren hierfür lediglich Beispielprobleme bekannt, die speziell derart konstruiert wurden, um die Existenz solcher Probleme zu beweisen. Wir halten abschließend fest, dass unsere Untersuchungen zeigen, dass die häufig in der Literatur zu findende Aussage, Alterungsmechanismen an sich tragen zur Diversität der Population bei, so nicht korrekt ist, da andere Elemente der Algorithmen wie die Selektion ebenfalls einen entscheidenden Beitrag leisten müssen.

4 Praktische Relevanz theoretischer Resultate

Die bisher vorgestellten Ergebnisse geben Einblick in die Funktionsweise verschiedener Mechanismen aus dem Bereich randomisierter Suchheuristiken. Diese Algorithmen werden in der Praxis angewendet, wenn kein problem-spezifischer Algorithmus verfügbar ist. Primäres Ziel unserer Analysen ist daher die Entwicklung besserer Heuristiken voranzutreiben. Das Erreichen wir zum Einen durch Erlangung eines tiefergehenden Verständnisses der unterschiedlichen Verfahren und zum Anderen durch die Entwicklung von Leitfäden für Praktiker zur Auswahl und Parametrisierung verschiedener Operatoren. Die praktische Relevanz ist für theoretische Resultate hierbei von entscheidender Bedeutung. Natürlich sind Praktiker an realen Rechenzeiten interessiert. Darum ist es entscheidend, dass das für unsere Analyse zugrunde liegende Kostenmodell Rechenzeiten realistisch abbildet. Es stellt sich also die Frage, ob und wenn ja unter welchen Bedingungen das Zählen von Runden bzw. Funktionsauswertungen ein ausreichend realistisches Maß ist.

Wir erinnern daran, dass bei der Wahl des Kostenmaßes angenommen wird, dass Funktionsauswertungen die teuersten Operationen des Algorithmus sind und sämtliche andere Operationen vernachlässigt werden können, so dass die Anzahl der Funktionsauswertungen eine gute Schätzung der tatsächlichen Laufzeit dargestellt. Es stellt sich die Frage, ob dies bei in der Praxis betrachteten Problemen und Algorithmen der Fall ist. Insbesondere die Verwendung von zusätzlichen Mechanismen, wie Rekombination oder Diversitätsmechanismen erhöhen die Laufzeit einer einzelnen Runde, so dass der Vergleich eines relativ komplizierten mit einem sehr einfachen Algorithmus irreführende Ergebnisse liefern kann. Wir beweisen, dass dieser Effekt bereits bei sehr einfachen Funktionen und Algorithmen auftreten kann, wenn die Funktionsauswertung in Realität nicht viel teurer ist als beispielsweise die Mutation. Dies ist selbstverständlich auch abhängig von der jeweiligen Implementierung der Algorithmen. Des Weiteren kann es sein, dass asymptotische Ergebnisse für praktisch relevante Aussagen zu grob sind. Hier können exakte Analysen weiterhelfen.

Motiviert durch diese Beobachtung entwickeln wir ein weitergehendes Kostenmodell, das ähnlich wie im Algorithm Engineering Implementierungsdetails in die Analyse mit einbezieht und eine genauere Analyse liefert. Wir betrachten die konkrete Implementierung einer einfachen randomisierten Suchheuristik mit Populationsgröße 1, bei der in jeder Runde ein Nachkomme durch Mutation erzeugt und jedes Bit mit Wahrscheinlichkeit c/n gekippt wird. Wir identifizieren zwei unterschiedliche Arten von Runden, zum Einen Runden, in denen kein Bit kippt, d. h. es wird nur eine Kopie des ursprünglichen Suchpunktes erzeugt. In diesem Fall ist keine neue Funktionsauswertung notwendig. Zum Anderen gibt es Runden, in denen Bits gekippt werden, so dass eine teure Funktionsauswertung durchgeführt werden muss. Wir bestimmen experimentell das Kostenverhältnis dieser beiden Rundenarten und führen mit Hilfe dieses neuen Kostenmodells eine exakte Analyse durch, anhand derer wir einen optimalen Wert für c bestimmen können. Bestimmt man solche optimalen Werte für c mit der herkömmlichen Methode, bei der jede Runde Kosten 1 verursacht, erhält man andere Resultate, die im Gegensatz zu unseren Ergebnissen mit realen Laufzeiten nicht konsistent sind. Unsere Art der Kombination von experimenteller und praktischer Analyse ist neu auf dem Gebiet der randomisierten Suchheuristiken und hat das Potenzial die Lücke zwischen Theorie und Praxis zu verkleinern.

5 Zusammenfassung und Ausblick

Die hier vorgestellte Arbeit ist die erste rigorose Laufzeitanalyse im Bereich der künstlichen Immunsysteme, was sie bahnbrechend und wegweisend macht. Ihr Hauptziel ist es, zum Verständnis der Arbeitsweise von künstlichen Immunsystemen beizutragen, Hinweise auf vielversprechende Anwendungsgebiete zu geben und zum Entwurf besserer künstlicher Immunsysteme beizutragen. Wir haben verschiedene Aspekte praktisch eingesetzter künstlicher Immunsysteme theoretisch betrachtet – zum Einen Hypermutationen, zum Anderen Altersmechanismen. Sämtliche Resultate wurden mit bereits vorhanden theoretischen Resultaten für andere randomisierte Suchheuristiken verglichen, um die unterschiedlichen Heuristiken voneinander abzugrenzen. Dabei lag ein besonderer Fokus auf der praktischen Relevanz der theoretischen Resultate. Dies wird unter anderem durch die Durchführung zusätzlicher experimentelle Untersuchungen gewährleistet. Zusätzlich haben wir von unseren theoretischen Ergebnissen Leitfäden zur Parametrisierung und zum Einsatz der betrachteten Operatoren abgeleitet. Die praktische Ausrichtung der Arbeit wird weiterhin durch das Hinterfragen des für die Analyse randomisierter Suchheuristiken verwendeten Kostenmaßes unterstrichen. Hier gibt die Arbeit Hinweise, auf welche Art und Weise praktisch relevante Theorie betrieben werden kann.

Eine der größten Herausforderungen für die Zukunft dieses Gebietes ist es sicherlich, die theoretische Analyse auf weitere Operatoren und ganze Algorithmen auszuweiten. Dazu gehört unter Anderem die Analyse des Zusammenspiels verschiedener Aspekte der Algorithmen, so wie wir es teilweise bereits am Beispiel von Alterungsmechanismen und Selektion durchgeführt haben. Auf lange Sicht ist es wünschenswert, einen vollständigen

Leitfaden zu erstellen, der für bestimmte Anwendungsgebiete die am vielversprechendsten erscheinenden Methoden aus dem Bereich der randomisierten Suchheuristiken identifiziert.

Eine weitere große Herausforderung für das Gebiet der künstlichen Immunsysteme ist die Vereinigung der beiden großen Teilbereiche, d. h. der Immun-Modellierung und der Algorithmenentwicklung. Es stellt sich die Frage, ob die theoretischen Ansätze und Methoden dieser Arbeit auch im Bereich der Immunologie hilfreich sein können, um vorhandene Modelle zu analysieren und damit zum weitergehenden Verständnis von natürlichen Immunsystemen beizutragen.

Literatur

[dCT02] Leandro Nunes de Castro und Jonathan Timmis. *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer, 2002.

[THSC08] Jon Timmis, Andrew Hone, T. Stibor und E. Clark. Theoretical advances in artificial immune systems. *Theoretical Computer Science*, 403(1):11–32, 2008.

[Zar11] Christine Zarges. *Theoretical Foundations of Artificial Immune Systems*. Dissertation, Fakultät für Informatik, TU Dortmund, Deutschland, 2011.

Christine Zarges hat von 2001–2007 an der TU Dortmund Informatik mit Nebenfach Betriebswirtschaftslehre studiert und ihr Diplom mit Auszeichnung abgeschlossen. Anschließend hat sie an der TU Dortmund am Lehrstuhl für Effiziente Algorithmen und Komplexitätstheorie an ihrer Promotion gearbeitet und diese im Juli 2011 mit der Note ausgezeichnet vollendet. Die Dissertation wurde mit dem Dissertationspreis der TU Dortmund ausgezeichnet. Außerdem wurde Christine Zarges im Jahr 2010 als einzige Deutsche mit einem Google Anita Borg Memorial Scholarship gefördert. Aktuell ist sie als Postdoktorandin an der University of Warwick in Großbritannien tätig. Ihr Aufenthalt dort wird vom Deutschen Akademischen Austausch Dienst finanziert.

Ihre Forschungsschwerpunkte liegen im Bereich der künstlichen Immunsysteme und randomisierten Suchheuristiken. Gegenwärtig hat Christine Zarges vier Zeitschriftenartikel sowie 14 begutachtete Konferenzbeiträge veröffentlicht, von denen zwei mit Best Paper Awards ausgezeichnet wurden. Seit 2012 ist sie Mitglied im Editorial Board der Zeitschrift „Evolutionary Computation“. Außerdem ist sie im Jahre 2012 Mitveranstalterin eines Tutorials und eines Workshops zu ihrem Dissertationsthema auf zwei führenden internationalen Konferenzen des Gebietes.

GI-Edition Lecture Notes in Informatics

Dissertations

Vol. D-1: Ausgezeichnete Informatikdissertationen 2000
Vol. D-2: Ausgezeichnete Informatikdissertationen 2001
Vol. D-3: Ausgezeichnete Informatikdissertationen 2002
Vol. D-4: Ausgezeichnete Informatikdissertationen 2003
Vol. D-5: Ausgezeichnete Informatikdissertationen 2004
Vol. D-6: Ausgezeichnete Informatikdissertationen 2005
Vol. D-7: Ausgezeichnete Informatikdissertationen 2006
Vol. D-8: Ausgezeichnete Informatikdissertationen 2007
Vol. D-9: Ausgezeichnete Informatikdissertationen 2008
Vol. D-10: Ausgezeichnete Informatikdissertationen 2009
Vol. D-11: Ausgezeichnete Informatikdissertationen 2010
Vol. D-12: Ausgezeichnete Informatikdissertationen 2011

P-1 Gregor Engels, Andreas Oberweis, Albert Zündorf (Hrsg.): Modellierung 2001.

P-2 Mikhail Godlevsky, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications, ISTA'2001.

P-3 Ana M. Moreno, Reind P. van de Riet (Hrsg.): Applications of Natural Lan-guage to Information Systems, NLDB'2001.

P-4 H. Wörn, J. Mühling, C. Vahl, H.-P. Meinzer (Hrsg.): Rechner- und sensorgestützte Chirurgie; Workshop des SFB 414.

P-5 Andy Schürr (Hg.): OMER – Object-Oriented Modeling of Embedded Real-Time Systems.

P-6 Hans-Jürgen Appelrath, Rolf Beyer, Uwe Marquardt, Heinrich C. Mayr, Claudia Steinberger (Hrsg.): Unternehmen Hochschule, UH'2001.

P-7 Andy Evans, Robert France, Ana Moreira, Bernhard Rumpe (Hrsg.): Practical UML-Based Rigorous Development Methods – Countering or Integrating the extremists, pUML'2001.

P-8 Reinhard Keil-Slawik, Johannes Magenheim (Hrsg.): Informatikunterricht und Medienbildung, INFOS'2001.

P-9 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Innovative Anwendungen in Kommunikationsnetzen, 15. DFN Arbeitstagung.

P-10 Mirjam Minor, Steffen Staab (Hrsg.): 1st German Workshop on Experience Management: Sharing Experiences about the Sharing Experience.

P-11 Michael Weber, Frank Kargl (Hrsg.): Mobile Ad-Hoc Netzwerke, WMAN 2002.

P-12 Martin Glinz, Günther Müller-Luschnat (Hrsg.): Modellierung 2002.

P-13 Jan von Knop, Peter Schirmbacher and Viljan Mahni_ (Hrsg.): The Changing Universities – The Role of Technology.

P-14 Robert Tolksdorf, Rainer Eckstein (Hrsg.): XML-Technologien für das Semantic Web – XSW 2002.

P-15 Hans-Bernd Bludau, Andreas Koop (Hrsg.): Mobile Computing in Medicine.

P-16 J. Felix Hampe, Gerhard Schwabe (Hrsg.): Mobile and Collaborative Business 2002.

P-17 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Zukunft der Netze –Die Verletzbarkeit meistern, 16. DFN Arbeitstagung.

P-18 Elmar J. Sinz, Markus Plaha (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2002.

P-19 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund.

P-20 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund (Ergänzungsband).

P-21 Jörg Desel, Mathias Weske (Hrsg.): Promise 2002: Prozessorientierte Methoden und Werkzeuge für die Entwicklung von Informationssystemen.

P-22 Sigrid Schubert, Johannes Magenheim, Peter Hubwieser, Torsten Brinda (Hrsg.): Forschungsbeiträge zur "Didaktik der Informatik" – Theorie, Praxis, Evaluation.

P-23 Thorsten Spitta, Jens Borchers, Harry M. Sneed (Hrsg.): Software Management 2002 – Fortschritt durch Beständigkeit

P-24 Rainer Eckstein, Robert Tolksdorf (Hrsg.): XMIDX 2003 – XML-Technologien für Middleware – Middleware für XML-Anwendungen

P-25 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Commerce – Anwendungen und Perspektiven – 3. Workshop Mobile Commerce, Universität Augsburg, 04.02.2003

P-26 Gerhard Weikum, Harald Schöning, Erhard Rahm (Hrsg.): BTW 2003: Datenbanksysteme für Business, Technologie und Web

P-27 Michael Kroll, Hans-Gerd Lipinski, Kay Melzer (Hrsg.): Mobiles Computing in der Medizin

P-28 Ulrich Reimer, Andreas Abecker, Steffen Staab, Gerd Stumme (Hrsg.): WM 2003: Professionelles Wissensmanagement – Er-fahrungen und Visionen

P-29 Antje Düsterhöft, Bernhard Thalheim (Eds.): NLDB'2003: Natural Language Processing and Information Systems

P-30 Mikhail Godlevsky, Stephen Liddle, Heinrich C. Mayr (Eds.): Information Systems Technology and its Applications

P-31 Arslan Brömme, Christoph Busch (Eds.): BIOSIG 2003: Biometrics and Electronic Signatures

P-32 Peter Hubwieser (Hrsg.): Informatische Fachkonzepte im Unterricht – INFOS 2003

P-33 Andreas Geyer-Schulz, Alfred Taudes (Hrsg.): Informationswirtschaft: Ein Sektor mit Zukunft

P-34 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 1)

P-35 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 2)

P-36 Rüdiger Grimm, Hubert B. Keller, Kai Rannenberg (Hrsg.): Informatik 2003 – Mit Sicherheit Informatik

P-37 Arndt Bode, Jörg Desel, Sabine Rathmayer, Martin Wessner (Hrsg.): DeLFI 2003: e-Learning Fachtagung Informatik

P-38 E.J. Sinz, M. Plaha, P. Neckel (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2003

P-39 Jens Nedon, Sandra Frings, Oliver Göbel (Hrsg.): IT-Incident Management & IT-Forensics – IMF 2003

P-40 Michael Rebstock (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2004

P-41 Uwe Brinkschulte, Jürgen Becker, Dietmar Fey, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle, Thomas Runkler (Edts.): ARCS 2004 – Organic and Pervasive Computing

P-42 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Economy – Transaktionen und Prozesse, Anwendungen und Dienste

P-43 Birgitta König-Ries, Michael Klein, Philipp Obreiter (Hrsg.): Persistance, Scalability, Transactions – Database Mechanisms for Mobile Applications

P-44 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): Security, E-Learning. E-Services

P-45 Bernhard Rumpe, Wofgang Hesse (Hrsg.): Modellierung 2004

P-46 Ulrich Flegel, Michael Meier (Hrsg.): Detection of Intrusions of Malware & Vulnerability Assessment

P-47 Alexander Prosser, Robert Krimmer (Hrsg.): Electronic Voting in Europe – Technology, Law, Politics and Society

P-48 Anatoly Doroshenko, Terry Halpin, Stephen W. Liddle, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications

P-49 G. Schiefer, P. Wagner, M. Morgenstern, U. Rickert (Hrsg.): Integration und Datensicherheit – Anforderungen, Konflikte und Perspektiven

P-50 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 1) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-51 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 2) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-52 Gregor Engels, Silke Seehusen (Hrsg.): DELFI 2004 – Tagungsband der 2. e-Learning Fachtagung Informatik

P-53 Robert Giegerich, Jens Stoye (Hrsg.): German Conference on Bioinformatics – GCB 2004

P-54 Jens Borchers, Ralf Kneuper (Hrsg.): Softwaremanagement 2004 – Outsourcing und Integration

P-55 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): E-Science und Grid Adhoc-Netze Medienintegration

P-56 Fernand Feltz, Andreas Oberweis, Benoit Otjacques (Hrsg.): EMISA 2004 – Informationssysteme im E-Business und E-Government

P-57 Klaus Turowski (Hrsg.): Architekturen, Komponenten, Anwendungen

P-58 Sami Beydeda, Volker Gruhn, Johannes Mayer, Ralf Reussner, Franz Schweiggert (Hrsg.): Testing of Component-Based Systems and Software Quality

P-59 J. Felix Hampe, Franz Lehner, Key Pousttchi, Kai Ranneberg, Klaus Turowski (Hrsg.): Mobile Business – Processes, Platforms, Payments

P-60 Steffen Friedrich (Hrsg.): Unterrichtskonzepte für inforrmatische Bildung

P-61 Paul Müller, Reinhard Gotzhein, Jens B. Schmitt (Hrsg.): Kommunikation in verteilten Systemen

P-62 Federrath, Hannes (Hrsg.): „Sicherheit 2005“ – Sicherheit – Schutz und Zuverlässigkeit

P-63 Roland Kaschek, Heinrich C. Mayr, Stephen Liddle (Hrsg.): Information Systems – Technology and ist Applications

P-64 Peter Liggesmeyer, Klaus Pohl, Michael Goedicke (Hrsg.): Software Engineering 2005

P-65 Gottfried Vossen, Frank Leymann, Peter Lockemann, Wolffried Stucky (Hrsg.): Datenbanksysteme in Business, Technologie und Web

P-66 Jörg M. Haake, Ulrike Lucke, Djamshid Tavangarian (Hrsg.): DeLFI 2005: 3. deutsche e-Learning Fachtagung Informatik

P-67 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 1)

P-68 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 2)

P-69 Robert Hirschfeld, Ryszard Kowalcyk, Andreas Polze, Matthias Weske (Hrsg.): NODe 2005, GSEM 2005

P-70 Klaus Turowski, Johannes-Maria Zaha (Hrsg.): Component-oriented Enterprise Application (COAE 2005)

P-71 Andrew Torda, Stefan Kurz, Matthias Rarey (Hrsg.): German Conference on Bioinformatics 2005

P-72 Klaus P. Jantke, Klaus-Peter Fähnrich, Wolfgang S. Wittig (Hrsg.): Marktplatz Internet: Von e-Learning bis e-Payment

P-73 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): “Heute schon das Morgen sehen“

P-74 Christopher Wolf, Stefan Lucks, Po-Wah Yau (Hrsg.): WEWoRC 2005 – Western European Workshop on Research in Cryptology

P-75 Jörg Desel, Ulrich Frank (Hrsg.): Enterprise Modelling and Information Systems Architecture

P-76 Thomas Kirste, Birgitta König-Riess, Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Informationssysteme – Potentiale, Hindernisse, Einsatz

P-77 Jana Dittmann (Hrsg.): SICHERHEIT 2006

P-78 K.-O. Wenkel, P. Wagner, M. Morgenstern, K. Luzi, P. Eisermann (Hrsg.): Land- und Ernährungswirtschaft im Wandel

P-79 Bettina Biel, Matthias Book, Volker Gruhn (Hrsg.): Softwareengineering 2006

P-80 Mareike Schoop, Christian Huemer, Michael Rebstock, Martin Bichler (Hrsg.): Service-Oriented Electronic Commerce

P-81 Wolfgang Karl, Jürgen Becker, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle (Hrsg.): ARCS´06

P-82 Heinrich C. Mayr, Ruth Breu (Hrsg.): Modellierung 2006

P-83 Daniel Huson, Oliver Kohlbacher, Andrei Lupas, Kay Nieselt and Andreas Zell (eds.): German Conference on Bioinformatics

P-84 Dimitris Karagiannis, Heinrich C. Mayr, (Hrsg.): Information Systems Technology and its Applications

P-85 Witold Abramowicz, Heinrich C. Mayr, (Hrsg.): Business Information Systems

P-86 Robert Krimmer (Ed.): Electronic Voting 2006

P-87 Max Mühlhäuser, Guido Rößling, Ralf Steinmetz (Hrsg.): DELFI 2006: 4. e-Learning Fachtagung Informatik

P-88 Robert Hirschfeld, Andreas Polze, Ryszard Kowalczyk (Hrsg.): NODe 2006, GSEM 2006

P-90 Joachim Schelp, Robert Winter, Ulrich Frank, Bodo Rieger, Klaus Turowski (Hrsg.): Integration, Informationslogistik und Architektur

P-91 Henrik Stormer, Andreas Meier, Michael Schumacher (Eds.): European Conference on eHealth 2006

P-92 Fernand Feltz, Benoît Otjacques, Andreas Oberweis, Nicolas Poussing (Eds.): AIM 2006

P-93 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 1

P-94 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 2

P-95 Matthias Weske, Markus Nüttgens (Eds.): EMISA 2005: Methoden, Konzepte und Technologien für die Entwicklung von dienstbasierten Informationssystemen

P-96 Saartje Brockmans, Jürgen Jung, York Sure (Eds.): Meta-Modelling and Ontologies

P-97 Oliver Göbel, Dirk Schadt, Sandra Frings, Hardo Hase, Detlef Günther, Jens Nedon (Eds.): IT-Incident Mangament & IT-Forensics – IMF 2006

P-98 Hans Brandt-Pook, Werner Simonsmeier und Thorsten Spitta (Hrsg.): Beratung in der Softwareentwicklung – Modelle, Methoden, Best Practices

P-99 Andreas Schwill, Carsten Schulte, Marco Thomas (Hrsg.): Didaktik der Informatik

P-100 Peter Forbrig, Günter Siegel, Markus Schneider (Hrsg.): HDI 2006: Hochschuldidaktik der Informatik

P-101 Stefan Böttinger, Ludwig Theuvsen, Susanne Rank, Marlies Morgenstern (Hrsg.): Agrarinformatik im Spannungsfeld zwischen Regionalisierung und globalen Wertschöpfungsketten

P-102 Otto Spaniol (Eds.): Mobile Services and Personalized Environments

P-103 Alfons Kemper, Harald Schöning, Thomas Rose, Matthias Jarke, Thomas Seidl, Christoph Quix, Christoph Brochhaus (Hrsg.): Datenbanksysteme in Business, Technologie und Web (BTW 2007)

P-104 Birgitta König-Ries, Franz Lehner, Rainer Malaka, Can Türker (Hrsg.) MMS 2007: Mobilität und mobile Informationssysteme

P-105 Wolf-Gideon Bleek, Jörg Raasch, Heinz Züllighoven (Hrsg.) Software Engineering 2007

P-106 Wolf-Gideon Bleek, Henning Schwentner, Heinz Züllighoven (Hrsg.) Software Engineering 2007 – Beiträge zu den Workshops

P-107 Heinrich C. Mayr, Dimitris Karagiannis (eds.) Information Systems Technology and its Applications

P-108 Arslan Brömme, Christoph Busch, Detlef Hühnlein (eds.) BIOSIG 2007: Biometrics and Electronic Signatures

P-109 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.) INFORMATIK 2007 Informatik trifft Logistik Band 1

P-110 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.) INFORMATIK 2007 Informatik trifft Logistik Band 2

P-111 Christian Eibl, Johannes Magenheim, Sigrid Schubert, Martin Wessner (Hrsg.) DeLFI 2007: 5. e-Learning Fachtagung Informatik

P-112 Sigrid Schubert (Hrsg.) Didaktik der Informatik in Theorie und Praxis

P-113 Sören Auer, Christian Bizer, Claudia Müller, Anna V. Zhdanova (Eds.) The Social Semantic Web 2007 Proceedings of the 1st Conference on Social Semantic Web (CSSW)

P-114 Sandra Frings, Oliver Göbel, Detlef Günther, Hardo G. Hase, Jens Nedon, Dirk Schadt, Arslan Brömme (Eds.) IMF2007 IT-incident management & IT-forensics Proceedings of the 3rd International Conference on IT-Incident Management & IT-Forensics

P-115 Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron and Dirk Walther (Eds.) German conference on bioinformatics GCB 2007

P-116 Witold Abramowicz, Leszek Maciszek (Eds.)
Business Process and Services Computing
1st International Working Conference on Business Process and Services Computing
BPSC 2007

P-117 Ryszard Kowalczyk (Ed.)
Grid service engineering and manegement
The 4th International Conference on Grid Service Engineering and Management
GSEM 2007

P-118 Andreas Hein, Wilfried Thoben, Hans-Jürgen Appelrath, Peter Jensch (Eds.)
European Conference on ehealth 2007

P-119 Manfred Reichert, Stefan Strecker, Klaus Turowski (Eds.)
Enterprise Modelling and Information Systems Architectures
Concepts and Applications

P-120 Adam Pawlak, Kurt Sandkuhl, Wojciech Cholewa,
Leandro Soares Indrusiak (Eds.)
Coordination of Collaborative Engineering - State of the Art and Future Challenges

P-121 Korbinian Herrmann, Bernd Bruegge (Hrsg.)
Software Engineering 2008
Fachtagung des GI-Fachbereichs Softwaretechnik

P-122 Walid Maalej, Bernd Bruegge (Hrsg.)
Software Engineering 2008 - Workshopband
Fachtagung des GI-Fachbereichs Softwaretechnik

P-123 Michael H. Breitner, Martin Breunig, Elgar Fleisch, Ley Pousttchi, Klaus Turowski (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Technologien, Prozesse, Marktfähigkeit
Proceedings zur 3. Konferenz Mobile und Ubiquitäre Informationssysteme
(MMS 2008)

P-124 Wolfgang E. Nagel, Rolf Hoffmann, Andreas Koch (Eds.)
9th Workshop on Parallel Systems and Algorithms (PASA)
Workshop of the GI/ITG Speciel Interest Groups PARS and PARVA

P-125 Rolf A.E. Müller, Hans-H. Sundermeier, Ludwig Theuvsen, Stephanie Schütze, Marlies Morgenstern (Hrsg.)
Unternehmens-IT:
Führungsinstrument oder Verwaltungsbürde
Referate der 28. GIL Jahrestagung

P-126 Rainer Gimnich, Uwe Kaiser, Jochen Quante, Andreas Winter (Hrsg.)
10th Workshop Software Reengineering (WSR 2008)

P-127 Thomas Kühne, Wolfgang Reisig, Friedrich Steimann (Hrsg.)
Modellierung 2008

P-128 Ammar Alkassar, Jörg Siekmann (Hrsg.)
Sicherheit 2008
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 4. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
2.-4. April 2008
Saarbrücken, Germany

P-129 Wolfgang Hesse, Andreas Oberweis (Eds.)
Sigsand-Europe 2008
Proceedings of the Third AIS SIGSAND European Symposium on Analysis, Design, Use and Societal Impact of Information Systems

P-130 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
1. DFN-Forum Kommunikations-technologien Beiträge der Fachtagung

P-131 Robert Krimmer, Rüdiger Grimm (Eds.)
3rd International Conference on Electronic Voting 2008
Co-organized by Council of Europe, Gesellschaft für Informatik and E-Voting. CC

P-132 Silke Seehusen, Ulrike Lucke, Stefan Fischer (Hrsg.)
DeLFI 2008:
Die 6. e-Learning Fachtagung Informatik

P-133 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 1

P-134 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 2

P-135 Torsten Brinda, Michael Fothe, Peter Hubwieser, Kirsten Schlüter (Hrsg.)
Didaktik der Informatik –
Aktuelle Forschungsergebnisse

P-136 Andreas Beyer, Michael Schroeder (Eds.)
German Conference on Bioinformatics
GCB 2008

P-137 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2008: Biometrics and Electronic Signatures

P-138 Barbara Dinter, Robert Winter, Peter Chamoni, Norbert Gronau, Klaus Turowski (Hrsg.)
Synergien durch Integration und Informationslogistik
Proceedings zur DW2008

P-139 Georg Herzwurm, Martin Mikusz (Hrsg.)
Industrialisierung des Software-Managements
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschaftsinformatik

P-140 Oliver Göbel, Sandra Frings, Detlef Günther, Jens Nedon, Dirk Schadt (Eds.)
IMF 2008 - IT Incident Management & IT Forensics

P-141 Peter Loos, Markus Nüttgens, Klaus Turowski, Dirk Werth (Hrsg.)
Modellierung betrieblicher Informationssysteme (MobIS 2008)
Modellierung zwischen SOA und Compliance Management

P-142 R. Bill, P. Korduan, L. Theuvsen, M. Morgenstern (Hrsg.)
Anforderungen an die Agrarinformatik durch Globalisierung und Klimaveränderung

P-143 Peter Liggesmeyer, Gregor Engels, Jürgen Münch, Jörg Dörr, Norman Riegel (Hrsg.)
Software Engineering 2009
Fachtagung des GI-Fachbereichs Softwaretechnik

P-144 Johann-Christoph Freytag, Thomas Ruf, Wolfgang Lehner, Gottfried Vossen (Hrsg.)
Datenbanksysteme in Business, Technologie und Web (BTW)

P-145 Knut Hinkelmann, Holger Wache (Eds.)
WM2009: 5th Conference on Professional Knowledge Management

P-146 Markus Bick, Martin Breunig, Hagen Höpfner (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Entwicklung, Implementierung und Anwendung
4. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2009)

P-147 Witold Abramowicz, Leszek Maciaszek, Ryszard Kowalczyk, Andreas Speck (Eds.)
Business Process, Services Computing and Intelligent Service Management
BPSC 2009 · ISM 2009 · YRW-MBP 2009

P-148 Christian Erfurth, Gerald Eichler, Volkmar Schau (Eds.)
9th International Conference on Innovative Internet Community Systems
I^2CS 2009

P-149 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
2. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-150 Jürgen Münch, Peter Liggesmeyer (Hrsg.)
Software Engineering
2009 - Workshopband

P-151 Armin Heinzl, Peter Dadam, Stefan Kirn, Peter Lockemann (Eds.)
PRIMIUM
Process Innovation for Enterprise Software

P-152 Jan Mendling, Stefanie Rinderle-Ma, Werner Esswein (Eds.)
Enterprise Modelling and Information Systems Architectures
Proceedings of the 3rd Int'l Workshop EMISA 2009

P-153 Andreas Schwill, Nicolas Apostolopoulos (Hrsg.)
Lernen im Digitalen Zeitalter
DeLFI 2009 – Die 7. E-Learning Fachtagung Informatik

P-154 Stefan Fischer, Erik Maehle Rüdiger Reischuk (Hrsg.)
INFORMATIK 2009
Im Focus das Leben

P-155 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2009:
Biometrics and Electronic Signatures
Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-156 Bernhard Koerber (Hrsg.)
Zukunft braucht Herkunft
25 Jahre »INFOS – Informatik und Schule«

P-157 Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, Peter Stadler (Eds.)
German Conference on Bioinformatics 2009

P-158 W. Claupein, L. Theuvsen, A. Kämpf, M. Morgenstern (Hrsg.)
Precision Agriculture Reloaded – Informationsgestützte Landwirtschaft

P-159 Gregor Engels, Markus Luckey, Wilhelm Schäfer (Hrsg.)
Software Engineering 2010

P-160 Gregor Engels, Markus Luckey, Alexander Pretschner, Ralf Reussner (Hrsg.)
Software Engineering 2010 – Workshopband (inkl. Doktorandensymposium)

P-161 Gregor Engels, Dimitris Karagiannis Heinrich C. Mayr (Hrsg.)
Modellierung 2010

P-162 Maria A. Wimmer, Uwe Brinkhoff, Siegfried Kaiser, Dagmar Lück-Schneider, Erich Schweighofer, Andreas Wiebe (Hrsg.)
Vernetzte IT für einen effektiven Staat
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010

P-163 Markus Bick, Stefan Eulgem, Elgar Fleisch, J. Felix Hampe, Birgitta König-Ries, Franz Lehner, Key Pousttchi, Kai Rannenberg (Hrsg.)
Mobile und Ubiquitäre Informationssysteme
Technologien, Anwendungen und Dienste zur Unterstützung von mobiler Kollaboration

P-164 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2010: Biometrics and Electronic Signatures Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-165 Gerald Eichler, Peter Kropf, Ulrike Lechner, Phayung Meesad, Herwig Unger (Eds.)
10th International Conference on Innovative Internet Community Systems (I^2CS) – Jubilee Edition 2010 –

P-166 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
3. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-167 Robert Krimmer, Rüdiger Grimm (Eds.)
4th International Conference on Electronic Voting 2010
co-organized by the Council of Europe, Gesellschaft für Informatik and E-Voting.CC

P-168 Ira Diethelm, Christina Dörge, Claudia Hildebrandt, Carsten Schulte (Hrsg.)
Didaktik der Informatik
Möglichkeiten empirischer Forschungsmethoden und Perspektiven der Fachdidaktik

P-169 Michael Kerres, Nadine Ojstersek Ulrik Schroeder, Ulrich Hoppe (Hrsg.)
DeLFI 2010 - 8. Tagung der Fachgruppe E-Learning der Gesellschaft für Informatik e.V.

P-170 Felix C. Freiling (Hrsg.)
Sicherheit 2010
Sicherheit, Schutz und Zuverlässigkeit

P-171 Werner Esswein, Klaus Turowski, Martin Juhrisch (Hrsg.)
Modellierung betrieblicher Informationssysteme (MobIS 2010)
Modellgestütztes Management

P-172 Stefan Klink, Agnes Koschmider Marco Mevius, Andreas Oberweis (Hrsg.)
EMISA 2010
Einflussfaktoren auf die Entwicklung flexibler, integrierter Informationssysteme
Beiträge des Workshops der GI-Fachgruppe EMISA (Entwicklungsmethoden für Informationssysteme und deren Anwendung)

P-173 Dietmar Schomburg, Andreas Grote (Eds.)
German Conference on Bioinformatics 2010

P-174 Arslan Brömme, Torsten Eymann, Detlef Hühnlein, Heiko Roßnagel, Paul Schmücker (Hrsg.)
perspeGKtive 2010
Workshop „Innovative und sichere Informationstechnologie für das Gesundheitswesen von morgen“

P-175 Klaus-Peter Fähnrich, Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für die Informatik
Band 1

P-176 Klaus-Peter Fähnrich, Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für die Informatik
Band 2

P-177 Witold Abramowicz, Rainer Alt, Klaus-Peter Fähnrich, Bogdan Franczyk, Leszek A. Maciaszek (Eds.)
INFORMATIK 2010
Business Process and Service Science – Proceedings of ISSS and BPSC

P-178 Wolfram Pietsch, Benedikt Krams (Hrsg.)
Vom Projekt zum Produkt
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschafts-informatik (WI-MAW), Aachen, 2010

P-179 Stefan Gruner, Bernhard Rumpe (Eds.)
FM+AM`2010
Second International Workshop on Formal Methods and Agile Methods

P-180 Theo Härder, Wolfgang Lehner, Bernhard Mitschang, Harald Schöning, Holger Schwarz (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW)
14. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme“ (DBIS)

P-181 Michael Clasen, Otto Schätzel, Brigitte Theuvsen (Hrsg.)
Qualität und Effizienz durch informationsgestützte Landwirtschaft, Fokus: Moderne Weinwirtschaft

P-182 Ronald Maier (Hrsg.)
6th Conference on Professional Knowledge Management
From Knowledge to Action

P-183 Ralf Reussner, Matthias Grund, Andreas Oberweis, Walter Tichy (Hrsg.)
Software Engineering 2011
Fachtagung des GI-Fachbereichs Softwaretechnik

P-184 Ralf Reussner, Alexander Pretschner, Stefan Jähnichen (Hrsg.)
Software Engineering 2011
Workshopband
(inkl. Doktorandensymposium)

P-185 Hagen Höpfner, Günther Specht, Thomas Ritz, Christian Bunse (Hrsg.)
MMS 2011: Mobile und ubiquitäre Informationssysteme Proceedings zur 6. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2011)

P-186 Gerald Eichler, Axel Küpper, Volkmar Schau, Hacène Fouchal, Herwig Unger (Eds.)
11th International Conference on Innovative Internet Community Systems (I^2CS)

P-187 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
4. DFN-Forum Kommunikations-technologien, Beiträge der Fachtagung 20. Juni bis 21. Juni 2011 Bonn

P-188 Holger Rohland, Andrea Kienle, Steffen Friedrich (Hrsg.)
DeLFI 2011 – Die 9. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V.
5.–8. September 2011, Dresden

P-189 Thomas, Marco (Hrsg.)
Informatik in Bildung und Beruf
INFOS 2011
14. GI-Fachtagung Informatik und Schule

P-190 Markus Nüttgens, Oliver Thomas, Barbara Weber (Eds.)
Enterprise Modelling and Information Systems Architectures (EMISA 2011)

P-191 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2011
International Conference of the Biometrics Special Interest Group

P-192 Hans-Ulrich Heiß, Peter Pepper, Holger Schlingloff, Jörg Schneider (Hrsg.)
INFORMATIK 2011
Informatik schafft Communities

P-193 Wolfgang Lehner, Gunther Piller (Hrsg.)
IMDM 2011

P-194 M. Clasen, G. Fröhlich, H. Bernhardt, K. Hildebrand, B. Theuvsen (Hrsg.)
Informationstechnologie für eine nachhaltige Landbewirtschaftung
Fokus Forstwirtschaft

P-195 Neeraj Suri, Michael Waidner (Hrsg.)
Sicherheit 2012
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 6. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)

P-196 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2012
Proceedings of the 11th International Conference of the Biometrics Special Interest Group

P-197 Jörn von Lucke, Christian P. Geiger, Siegfried Kaiser, Erich Schweighofer, Maria A. Wimmer (Hrsg.)
Auf dem Weg zu einer offenen, smarten und vernetzten Verwaltungskultur
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2012

P-198 Stefan Jähnichen, Axel Küpper, Sahin Albayrak (Hrsg.)
Software Engineering 2012
Fachtagung des GI-Fachbereichs Softwaretechnik

P-199 Stefan Jähnichen, Bernhard Rumpe, Holger Schlingloff (Hrsg.)
Software Engineering 2012
Workshopband

P-200 Gero Mühl, Jan Richling, Andreas Herkersdorf (Hrsg.)
ARCS 2012 Workshops

P-201 Elmar J. Sinz Andy Schürr (Hrsg.)
Modellierung 2012

P-202 Andrea Back, Markus Bick, Martin Breunig, Key Pousttchi, Frédéric Thiesse (Hrsg.)
MMS 2012:Mobile und Ubiquitäre Informationssysteme

P-203 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
5. DFN-Forum Kommunikations-technologien
Beiträge der Fachtagung

P-204 Gerald Eichler, Leendert W. M. Wienhofen, Anders Kofod-Petersen, Herwig Unger (Eds.)
12th International Conference on Innovative Internet Community Systems (I2CS 2012)

P-205 Manuel J. Kripp, Melanie Volkamer, Rüdiger Grimm (Eds.)
5th International Conference on Electronic Voting 2012 (EVOTE2012)
Co-organized by the Council of Europe, Gesellschaft für Informatik and E-Voting.CC

P-206 Stefanie Rinderle-Ma, Mathias Weske (Hrsg.)
EMISA 2012
Der Mensch im Zentrum der Modellierung

P-207 Jörg Desel, Jörg M. Haake, Christian Spannagel (Hrsg.)
DeLFI 2012: Die 10. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V.
24.–26. September 2012

P-208 Ursula Goltz, Marcus Magnor, Hans-Jürgen Appelrath, Herbert Matthies, Wolf-Tilo Balke, Lars Wolf (Hrsg.)
INFORMATIK 2012

P-209 Hans Brandt-Pook, André Fleer, Thorsten Spitta, Malte Wattenberg (Hrsg.)
Nachhaltiges Software Management

P-210 Erhard Plödereder, Peter Dencker, Herbert Klenk, Hubert B. Keller, Silke Spitzer (Hrsg.)
Automotive – Safety & Security 2012
Sicherheit und Zuverlässigkeit für automobile Informationstechnik

The titles can be purchased at:

Köllen Druck + Verlag GmbH
Ernst-Robert-Curtius-Str. 14 · D-53117 Bonn
Fax: +49 (0)228/9898222
E-Mail: druckverlag@koellen.de