Explorative Suche in Zeitbasierten Primärdaten[1]

Jürgen Bernard[2]

Abstract:

Die Ära des Big Data birgt gewaltige Potenziale für die datenzentrierte Forschung, denen Herausforderungen wie die Größe, die Qualität oder temporale Aspekte der Daten gegenüberstehen. Für die explorative Suche nach unerforschtem Wissen in komplexen Daten benötigen Domänenexperten effektive Analysetechniken und -systeme. Im Design dieser Systeme lassen sich die Kompetenzen von Data Scientists mit denen der Domänenexperten vereinen. Am Beispiel von zeitbasierten Primärdaten präsentiere ich in meiner Dissertation Konzepte, Richtlinien, Techniken und Systeme für die explorative Suche zur Unterstützung der datenzentrierten Forschung. Dabei verfolge ich in einem Visual-Analytics-Ansatz die strikte Kopplung von visuell-interaktiven Benutzerschnittstellen mit algorithmischen Modellen zur Datenanalyse. Beim Design von explorativen Suchsystemen ermögliche ich den Vergleich und die Auswahl von Modellen, unter Einbezug von Domänenexperten.

1 Einleitung

Die Menschheitsgeschichte war stets geprägt von wissenschaftlichen Paradigmen. Nach der experimentellen Wissenschaft in der Antike, der theoretischen Wissenschaft im Mittelalter, und der simulationsbasierten Wissenschft ab dem Computerzeitalter, erleben wir heute das Paradigma der *datenzentrierten Forschung* [HTT09]. Die Suche nach interessanten Strukturen in großen Datenmengen wird zur wissenschaftlichen Praxis. Damit sieht sich die datenzentrierte Forschung mit Herausforderungen aus dem Big-Data-Bereich, wie z.B. der Informationsüberlastung (engl. Information Overload), konfrontiert. Dringend bedarf es neuer, intelligenter Lösungen für die Analyse und Exploration komplexer Daten.

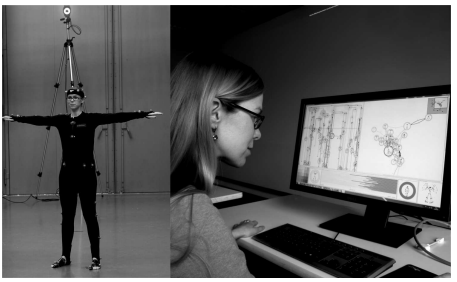
In meiner Dissertation fokussiere ich mich auf *zeitbasierte Primärdaten*, einem Datentyp zur Erfassung von komplexen, temporalen Phänomenen (a.k.a. *Zeitserien, Zeitreihen*), zwei Anwendungsdomänen sind in Abbildung 1 dargestellt. Die temporale Eigenschaft der Daten ermöglicht spezielle Analysetasks, wie etwa die Identifikation von Trends, periodischen Mustern, oder temporalen Anomalien. Primärdaten beschreiben Phänomene in ihrer ursprünglichen Form und unterliegen damit keiner Veränderung oder Manipulation. So birgen zeitbasierte Primärdaten unerforschtes Wissen, welches insbesondere für die datenzentrierte Forschung von großem Interesse ist. Um Erkenntnisse aus den Primärdaten zu ziehen und diese zu verifizieren, bedarf es geeigneter Werkzeuge aus der explorativen und konfirmativen Datenanalyse. Neben der Größe und der Heterogenität komplexer Daten, sind die Qualität und der Zeitbezug spezielle datenseitige Problemstellungen. Zusätzlich zum Dateninhalt (engl. Content), stellen Metadaten (Daten über Daten) eine weitere Komplexität dar. Die Suche nach Zusammenhängen zwischen dem Dateninhalt und Metadaten (z.B. Variablen die temporale Veränderungen in Klimamessungen erklären könnten) ist in vielen Forschungsbereichen höchst relevant, und nicht selten zeitaufwändig.

Eine Vision in der datenzentrierten Forschung ist die Sicherstellung der *Wiederverwendbarkeit* von erhobenen Primärdaten, insbesondere für zeitbasierte, unwiederbringliche Da-

[1] Englischer Titel der Dissertation: “Exploratory Search in Time-Oriented Primary Data”

[2] Fraunhofer-Institut für Graphische Datenverarbeitung IGD, Darmstadt, juergen.bernard@igd.fraunhofer.de

(a) Datenzentrierte Forschung in der Klimaforschung. An der Neumayer Station in der Antarktis werden Wetterphänomene mit Sensoren gemessen. Seit über 30 Jahren werden diese zeitbasierten Primärdaten aus der ganzen Welt zusammengetragen, und stehen der Forschung zur Wiederverwendung bereit.



(b) Links: Tracking menschlicher Bewegungsdaten mit Markern. Domänenexperten in Sportwissenschaften, Medizin, oder Biomechanik interessieren sich für Variationen in Bewegungsabläufen. Rechts: Evaluierung des MotionExplorer Systems zur explorativen Suche in Bewegungsdaten.

Abb. 1: Forschungsgebiete in denen zeitbasierte Primärdaten gemessen, verarbeitet, und anschließend für die wissenschaftliche Wiederverwendung persistiert werden.

ten. Diverse Domänenexperten aus der Forschungslandschaft können so an den selben Daten forschen. Zu weiten Teilen ungelöst ist jedoch das Problem des intuitiven und effektiven Zugangs zu großen, komplexen Datenkollektionen. Die Unterstützung der Suche nach relevanten Daten *ohne* Vorwissen gilt hierbei als besonders schwierig, und bedarf neuartiger, explorativer Datenanalysetechniken. Digitale Bibliotheken und ähnliche Infrastrukturen können hier in Zukunft eine noch zentralere Rolle spielen.

Ziel meiner Dissertation ist die Unterstützung der datenzentrierten Forschung bei der Wiederverwendung und der Analyse von zeitbasierten Primärdaten. Dazu setze ich das Konzept der explorativen Suche [Ma06, WR09] erstmals für zeitbasierte Primärdaten in die Praxis um. Grundsätzlich repräsentiert die explorative Suche die Idee, verschiedene Informationsbedürfnisse des Nutzers in einem System vereint zu unterstützen. Dabei sollen Aktivitäten vom Abrufen von Faktenwissen (Suche) bis hin zur Erkundung völlig neuer Such- und Informationsräume (Exploration) unterstützt werden. Um die explorative Suche erstmals für zeitbasierte Primärdaten umzusetzen, bediene ich mich der Techniken der Informationsvisualisierung und der Visual Analytics. Die Informationsvisualisierung ist die Lehre der visuell-interaktiven Repräsentierung von abstrakten Daten [CMS99], Visual Analytics erforscht das geeignete Zusammenspiel zwischen automatischer Datenanalyse und visueller Datenexploration [Ke10].

Eine Recherche verwandter Arbeiten ergab insbesondere folgende ungelöste Probleme. Zunächst existierte die explorative Suche weitestgehend nur als Konzept, mit der Ausnahme von Systemen für Textdaten. Es fehlte an Strategien, um das Design geeigneter Systeme auch methodisch zu unterstüzen. Der inhaltsbasierte Zugang zu zeitbasierten Primärdaten stellte ein zentrales technisches Problem dar. So war die Suche bisher nur über Metadaten (Daten über Daten) möglich. Zur Unterstützung der explorativen Datenanalyse lag eine Schwierigkeit darin, einen Überblick über große Mengen an zeitbasierten Primärdaten in einem visuellen Suchsystem anzubieten. Des Weiteren bestand ein Defizit in Suchsystemen darin, dass die Identifikation von Zusammenhängen zwischen Zeitseriendaten (dem Dateninhalt) und Metadaten nicht Teil des analytischen Repertoires war.

In dieser Dissertation beschäftige ich mich mit diesen Herausforderungen und entwickle Methoden, Techniken, und Systeme für die explorative Suche in zeitbasierten Primärdaten.

Es werden Methoden für das Design von explorativen Suchsystemen aufgezeigt (Kapitel 3). Darauf aufbauend stellen die Kapitel 4, 5, und 6 die technischen Schwerpunkte der Dissertation dar. Zunächst löst das erste Visual Analytics System für das visuell-interaktive Preprocessing von Zeitseriendaten das Problem des inhaltsbasierten Zugangs zu zeitbasierten Primärdaten. Ein weiteres Kapitel stellt Richtlinien und Techniken für das Design von Überblicksvisualisierungen für Zeitseriendaten vor. Schließlich werden drei neuartige Techniken für die kombinierte Analyse von Dateninhalt und Metadaten vorgestellt. Die technischen Beiträge dieser Dissertation berücksichtigen explizit die Herausforderung, geeignete algorithmische Modelle in der richtigen Reihenfolge und mit den richtigen Parametern zu wählen. Des Weiteren wird für alle Techniken beschrieben, wie Nutzer in das Design involviert werden können. In Kapitel 7 validiere ich die Methoden und Techniken anhand zweier explorativer Suchsysteme für zeitbasierte Primärdaten.

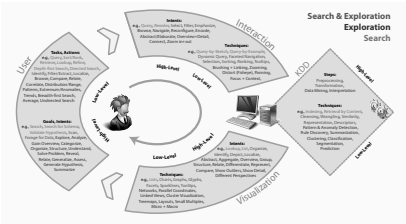
Mit den Ergebnissen dieser Dissertation [Be15c] leiste ich einen Beitrag zur Wiederverwendung von zeitbasierten Primärdaten, insbesondere zur Unterstützung der datenzentrierten Forschung. Nutzer können durch die Definition von visuell-interaktiven Suchanfragen (query-by-sketch, query-by-example) direkt im Dateninhalt suchen. Mit visuell-interaktiven Überblicksdarstellungen sind Nutzer zudem in der Lage unbekannte Zusammenhänge im Suchraum zu explorieren und diese für die Wissenserweiterung zu nutzen. Durch die Öffnung des Designprozesses für den Nutzer und die strikt visuelle Art der Datenrepräsentierung leistet diese Dissertation zudem einen Beitrag zum User-centered Design, sowie zur Kommunikation von Information und Wissen aus zeitbasierten Primärdaten.

2 Zielsetzung und Problemdefinition

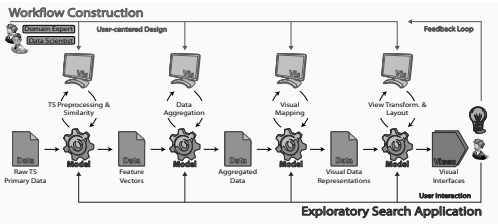
Die übergeordnete Zielsetzung dieser Dissertation lässt sich wie folgt beschreiben: Wie kann, für eine spezifische Forschungsfrage, ein Analysesystem zur Suche und Exploration in großen Datenmengen entwickelt werden, das (a) auf die charakteristischen Eigenschaften der Daten Bezug nimmt, (b) die spezifischen Nutzeranforderungen berücksichtigt, und (c) dabei einen möglichst umfassenden Pool an Analysealgorithmen, Visualisierungs- und Interaktionstechniken bereitstellt. Zudem soll, neben den Lösungen für spezifische datenzentrierte Forschungsprobleme, ein allgemeiner Ansatz hervorgehen, der es Data Scientists und Domänenexperten ermöglicht, gemeinsam explorative Suchsysteme in wesentlich kürzerer Zeit zu entwickeln, als dies derzeit mit dem Stand der Technik möglich ist. Auf Basis einer sorgfältigen Literaturrecherce beschreibe ich im Folgenden die sechs größten Herausforderungen $\mathbf{C}_x$ dieser Zielsetzung.

$\mathbf{C}_{\text{MES}}$ Fehlende Methodologie Data Scientists sind mit einer Reihe von Freiheitsgraden (engl. Design Space) konfrontiert, unter Anderem aufgespannt durch komplexe Daten, individuelle Nutzerwünsche, und spezifische Analysetasks im explorativen Suchkontext. In der Informationsvisualisierung und der Visual Analytics wurden eine Reihe von Konzepten und Techniken erforscht, die auch für das Design von explorativen Suchsystemen von Nutzen wären. Hier bedarf es jedoch zunächst einer genauen Betrachtung von Anknüpfungspunkten und Synergieeffekten.

$\mathbf{C}_{\text{CBA}}$ Inhaltsbasierter Zugang zu Zeitbasierten Primärdaten Die Verwendung des Dateninhalts ist ein vielversprechender Ansatz für effektive Such- und Explorationsmechanismen. Der inhaltsbasierte Zugang zu textuellem Dateninhalt spielt hier eine Pionierrolle. Jedoch gibt es z.B. kaum Digitale Bibliothekssysteme, deren Funktionsumfang die inhalts-



(a) Überblick über Analysetasks aus der Informationsvisualisierung und der Visual Analytics mit einer Relevanz für Such- und Explorationsaktivität. Die Abbildung kondensiert die Information vieler existierender Task Taxonomien.



(b) Referenzworkflow für das Design und die Anwendung von explorativen Suchsystemen. In vier Schritten bestimmen Data Scientists und Domänenexperten algorithmische Modelle und Parameter. Das Resultat ist ein Analyseprozess, integriert in eine visuell-interaktive Benutzerschnittstelle.

Abb. 2: Schematische Darstellungen der zwei wesentlichen konzeptionellen Beiträge.

basierte Suche in Zeitseriendaten unterstützt. Die unsichere Datenqualität und der Zeitbezug sind spezifische Probleme zeitbasierter Primärdaten für den inhaltsbasierten Zugang.

$\mathbf{C}_{\text{CBO}}$ Visuelle Repräsentierung des Inhalts großer Datenmengen Eine wichtige Eigenschaft effektiver Analysesysteme ist die Unterstützung der Nutzer bei der Identifikation von struktureller Information großer Datenmengen. Eine zentrale Herausforderung beim Design solcher Überblickstechniken ist die Aggregation der Daten, welche zudem in eine visuelle Form gebracht, und in das explorative Suchsystem integriert werden müssen. Neben dem Visualisierungsdesign spielt das Interaktionsdesign eine entscheidende Rolle.

$\mathbf{C}_{\text{C+M}}$ Zusammenhänge zwischen Dateninhalt und Metadaten Die Integration von Metadaten in den Analyseprozess ist ein mächtiges, und zu gleich schwieriges analytisches Konzept. Die Charakterisierung von Zusammenhängen (Korrelationen, Assoziationen, etc.), und die Bewertung von deren Interessantheit ist abhängig von der Forschungsfrage. Schließlich bedarf es neuartiger Visualisierungs- und Interaktionstechniken für deren Exploration.

$\mathbf{C}_{\text{MPC}}$, $\mathbf{C}_{\text{UCD}}$ Übergeordnete Problemstellungen Schließlich stellen (a) die Wahl von geeigneten algorithmischen Modellen und Parametern $\mathbf{C}_{\text{MPC}}$, sowie (b) der Einbezug der Nutzergruppe in das Design $\mathbf{C}_{\text{UCD}}$ eigene Problemstellungen dar. Diese haben Einfluss auf die Umsetzung aller technischer Beiträge dieser Dissertation. Das Design von explorativen Suchsystemen führt zu einer Reihe von technischen Freiheitsgraden die sich aus der Kombination von Daten, Nutzern und Tasks ergeben. Die Wahl geeigneter algorithmischer Modelle in geeigneter Reihenfolge, mit geeigneten Parametern $\mathbf{C}_{\text{MPC}}$ ist grundsätzlich schwierig. Bereits kleine Änderungen im Analyseworkflow haben oft gravierende Auswirkungen auf das Analyseergebnis. Den technischen Freiheitsgraden gegenüber steht die Notwendigkeit des Einbezugs der Nutzer in das Design $\mathbf{C}_{\text{UCD}}$, insbesondere bei der Unterstützung von Domänenexperten in der datenzentrierten Forschung. Grundsätzlich sollte der Designprozess iterativ ablaufen. Wichtige Designentscheidungen sollten nutzerbestimmt sein.

3 Konzeptueller Beitrag

Der konzeptionelle Beitrag dieser Dissertation löst das Problem der fehlenden Methodologie für das Design von explorativen Suchsystemen $\mathbf{C}_{\text{MES}}$ in zwei wesentlichen Teilaspekten.

Zunächst wird ein *Überblick über analytische Tasks* gegeben, die für das Design von explorativen Suchsystemen relevant sind. Um Data Scientists das Design von explorativen

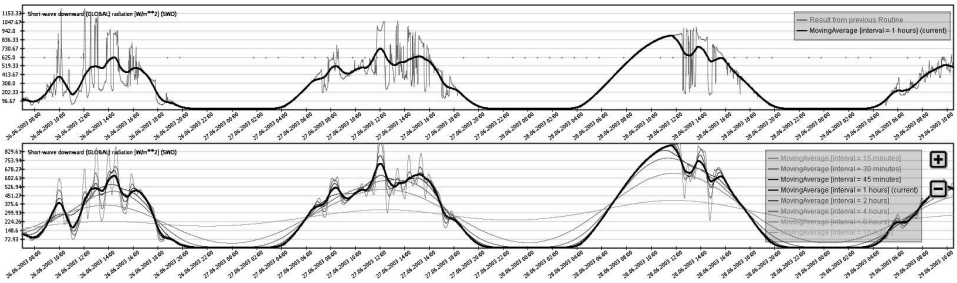


Abb. 3: Visuell-interaktives Preprocessing von Zeitserien. Oben: Nutzer vergleichen eine geglättete Zeitserie (schwarz) mit der ursprünglichen, verrauschten Zeitserie (grau). Unten: Vergleich zwischen der aktuellen Modell-Parameterisierung (schwarz) und 7 Alternativvorschlägen vom System (blau bis braun). Beide Visualisierungen ermöglichen die Optimierung von Modell und Parameterwahl.

Suchsystemen zu erleichtern, sind alle analytischen Tasks in einer einzigen schematischen Darstellung vereint (siehe Abbildung 2a). Diese Darstellung ist das Ergebnis eines Surveys über Tasktaxonomien und Methodologien aus der Informationsvisualisierung, Visual Analytics, sowie den verwandten Domänen Human-Computer Interaction (HCI), Data Mining (DM), und Knowledge Discovery in Databases (KDD).

Der zweite Beitrag beschreibt einen *Referenzworkflow* für das Design und die Anwendung von explorativen Suchsystemen (siehe Abbildung 2b). In vier Schritten wird der Designprozess für visuell-interaktive Benutzerschnittstellen beschrieben. Dabei stellt jeder der vier Schritte eine Instanzierung des Visual Analytics Referenzmodells [Ke10] dar, welches (1) Daten, (2) algorithmische Modelle, (3) visuell-interaktive Nutzerschnittstellen, und (4) resultierende Erkenntnisse in Bezug setzt. Der Referenzworkflow erleichtert das Design von visuell-interaktiven Benutzerschnittstellen, und hilft bei der Optimierung der Wahl, Parameterisierung, und Verschaltung von algorithmischen Modellen.

Zusammenfassend stellt der konzeptuelle Beitrag der Dissertation den Stand der Technik aus der Informationsvisualisierung und der Visual Analytics zusammen, und baut darauf einen allgemeinen Ansatz für das Design von explorativen Suchsystemen auf. Data Scientists haben nun die Möglichkeit, in Kollaboration mit Domänenexperten, spezifische Forschungsfragen gezielt durch neuartige explorative Suchsysteme zu unterstützen.

4 Visuell-Interaktives Preprocessing von Zeitbasierten Primärdaten

Kapitel 4 der Dissertation addressiert das Problem des inhaltsbasierten Zugangs zu zeitbasierten Primärdaten $\mathbf{C}_{\text{CBA}}$. Präsentiert wird der erste Visual Analytics Ansatz für das visuell-interaktive Preprocessing von Zeitserien [Be12a]. So ist es nun möglich die *Qualität* von zeitbasierten Primärdaten visuell-interaktiv zu analysieren und zu optimieren. Hierzu stehen eine Reihe von algorithmischen Modellen aus dem Bereich des Time Series Data Mining zur Verfügung, welche Data Scientists gemeinsam mit Domänenexperten zu einem Preprocessing-Workflow verschalten können. Abbildung 3 zeigt ein solches algorithmisches Modell. Ein Nutzer glättet eine verrauschte Zeitserie durch eine Moving Average Routine. Das System gibt visuelles Feedback durch den Input-Output Vergleich (oben), und durch Parameter-Guidance (unten) $\mathbf{C}_{\text{MPC}}$. Fehler in der Konstruktion von Workflows, wie zum Beispiel Kaskadeneffekte, lassen sich somit direkt erkennen und beseitigen. Um in explorativen Suchsystemen effektive und effiziente Retrieval- und Analysealgorithmen ausführen zu können, stellt der Ansatz zudem algorithmische Modelle zur Verfügung, mit denen Zeitserien in den *Feature Space* transformiert werden können.

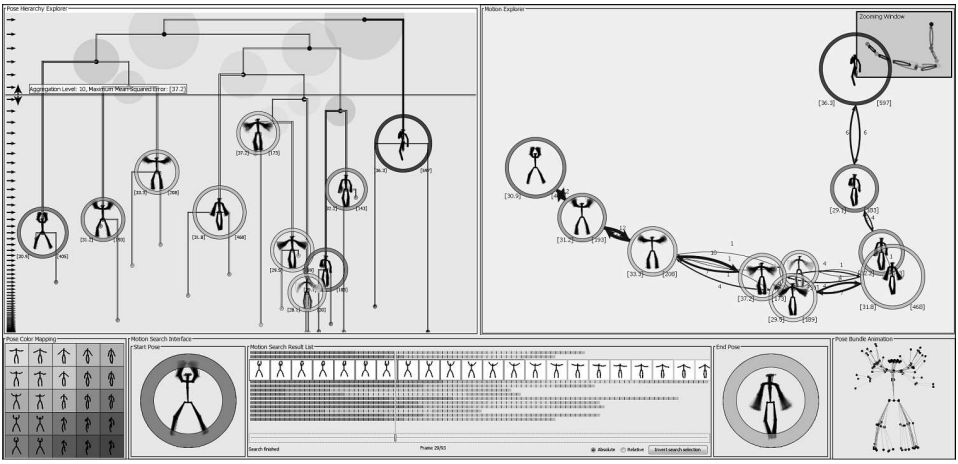


Abb. 4: Explorative Suche in menschlichen Bewegungsdaten. In zwei inhaltsbasierten Überblicksvisualisierungen werden tausende menschlicher Posen (oben links) und Bewegungen zwischen Posen (oben rechts) auf intuitive Weise dargestellt. Mit nur 5 Klicks kann ein Domänenexperte 2 Posen auswählen (hier: rote Pose und grüne Pose), worauf die Suche nach Subsequenzen gestartet wird (siehe unten). Im Beispiel visualisiert das System 12 gefundene Hampelmann-Bewegungen.

Diese Modelle werden häufig als (Zeitserien-) *Deskriptoren* bezeichnet, welche nun auch visuell-interaktiv in den Workflow integrierbar sind. Schließlich schließt dieser Ansatz die visuell-interaktive Definition von Ähnlichkeitsmaßen für Zeitserien mit ein.

Zusammenfassend stellt der präsentierte wissenschaftliche Beitrag zur Konstruktion von Preprocessing Workflows einen vollständigen Umstieg auf einen visuell-interaktiven Ansatz dar. Auf effiziente Weise können Data Scientists, gemeinsam mit Domänenexperten, effektive Preprocessing Workflows für zeitbasierte Primärdaten konstruieren $\mathbf{C}_{\text{UCD}}$. Nach Abschluss eines Workflows kann dieser dann voll-automatisch Zeitserien prozessieren. Eine der naheliegenden Anwendungen ist der nutzerzentrierte, inhaltsbasierte Zugang zu zeitbasierten Primärdaten mit der anschließenden Integration in explorative Suchsysteme.

5 Visueller Überblick über den Dateninhalt

Kapitel 5 der Dissertation beschreibt Lösungen zur visuellen Repräsentierung großer Datenmengen $\mathbf{C}_{\text{CBO}}$ in drei technischen Aspekten. Zunächst werden visuell-interaktive Techniken für den Clusteranalyseprozess präsentiert [Be11]. Hier werden unter Anderem neuartige Techniken für die halb-überwachte (engl. semi-supervised) Clusteranalyse, und für die Qualitätsbewertung von Clusterergebnissen vorgestellt. Der zweite Beitrag zeigt auf wie aggregierte Daten, als Produkt des Clusteringprozesses, visuell repräsentiert werden können. Hierbei liegt besonderes Augenmerk auf der kombinierten Visualisierung von Clusters und deren Datenpunkten, sowie auf dem wissenschftlich korrekten Umgang mit *Farbe*, als ähnlichkeitserhaltende, visuelle Variable zum visuellen Vergleich von Clusters [Be15b]. Der dritte Beitrag zeigt Möglichkeiten auf, wie aggregierte Daten auf sinnvolle Weise in einem 2D Layout arrangiert werden können. Von besonderer Wichtigkeit ist hier der sinnvolle Einsatz von Algorithmen zur Dimensionsreduktion (Projektion) zur Beibehaltung struktureller Information hochdimensionaler Daten.

In allen drei Beiträgen wird Wert auf den Einbezug von Nutzerwünschen gelegt $\mathbf{C}_{\text{UCD}}$, sowie auf die richtige Wahl von algorithmischen Modellen und Modellparametern $\mathbf{C}_{\text{MPC}}$. So

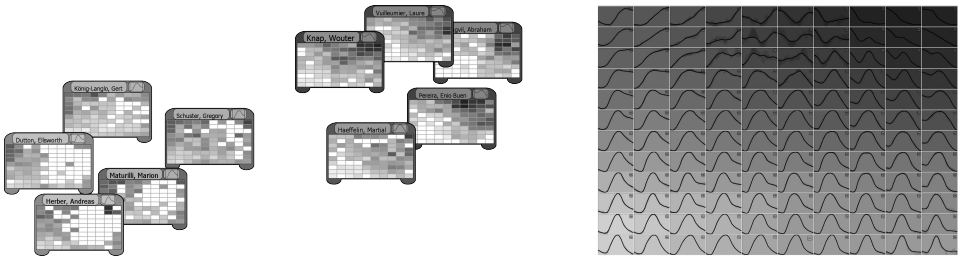


Abb. 5: Analyse von Zusammenhängen zwischen zehn Forschern (links) und den von ihnen gemessenen und publizierten Primärdaten im Überblick (rechts). Ein Layout arrangiert die zehn Forscher anhand der Ähnlichkeit ihner Daten. Interessanterweise bilden sich zwei Gruppen von Forschern deutlich heraus. Gespräche mit dem Domänenexperten ergaben, dass die Forschergruppe 'braun' überwiegend im Antarktischen misst. Eine Mosaikmetapher zeigt die Daten als visuelle Signaturen.

können Data Scientists nun z.B. in den Clusteranalyseprozess eingreifen, bzw. auf effektive Weise geeignete Clusteralgorithmen und Parameter identifizieren. Abbildung 4 zeigt ein Anwendungsbeispiel das von allen drei technischen Aspekten profitiert.

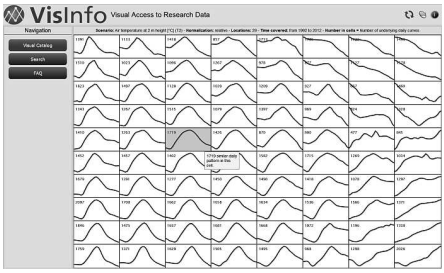
6 Kombinierte Analyse von Dateninhalt und Metadaten

Kapitel 6 der Dissertation löst Probleme bei der Analyse von Zusammenhängen zwischen dem Dateninhalt (Zeitserien) und Metadaten $\mathbf{C}_{\text{C+M}}$. Es werden drei Techniken präsentiert, die es Nutzern ermöglichen auf visuell-interaktive Weise nach interessanten Zusammenhängen zwischen zeitbasierten Primärdaten und Metadaten zu suchen. Die Ansätze unterscheiden sich in der Festlegung der abhängigen Variable, und damit in der Art der unterstützten wissenschaftlichen Fragestellung. Im ersten Ansatz wird die Zielvariable im Dateninhalt festgelegt [Be12c]. Das System präsentiert interessante Metadaten, die mit dem Dateninhalt in Zusammenhang stehen (z.B. "Verläufe hoher Luftfeuchtigkeit treten verstärkt in tropischen Regionen auf"). Der Zweite Ansatz unterstützt das inverse Informationsbedürfnis: für ein festgelegtes Metadatenattribut werden interessante Zeitverläufe exploriert (Beispiel Finanzbranche: "die Internetbranche erlebte Ende der 90er Jahre einen starken Aufwärstrend") [Be12b]. Schließlich wird ein Explorationssystem vorgestellt, das gänzlich ohne die Festlegung von Zielvariablen arbeitet [Be14]. Auf einen Blick werden die interessantesten Zusammenhänge im gesamten Datensatz aufgezeigt. Das System nimmt damit Domänenexperten nicht nur das Testen von Hypothesen ab, sondern auch deren Entdeckung und Formulierung.

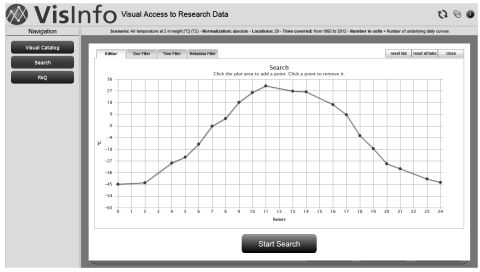
Allen drei Ansätzen gemein ist die Definition von Interessantheit. Domänenexperten legen fest welches algorithmische Modell Interessantheiten berechnet (z.B. Korrelationsmaße, Signifikanztests, etc.), und die visuell-interaktiven Techniken präsentieren automatisch die interessantesten Zusammenhänge. Abbildung 5 zeigt ein Beispiel für den zweiten Ansatz.

7 Designstudien - Explorative Suchsysteme

In Kapitel 7 der Dissertation werden die neuartigen Konzepte, Guidelines, und Techniken in die Praxis umgesetzt. In zwei Anwendungsbeispielen wird das Konzept der explorativen Suche erstmals für zeitbasierte Primärdaten umgesetzt. Beide Systeme wurden unter Einbezug von Domänenexperten geplant, entwickelt, verfeinert, und evaluiert $\mathbf{C}_{\text{UCD}}$ $\mathbf{C}_{\text{MPC}}$.



(a) Visueller Katalog: Überblick über zeitbasierte Primärdaten. Hier: 200000 Temperaturverläufe.



(b) Sketchbasierte, visuelle Benutzerschnittstelle zur inhaltsbasierten Suche in zeitbasierten Primärdaten.

Abb. 6: VisInfo: Explorative Suche in großen Mengen an Temperaturverläufen, zur Klimaforschung.

VisInfo [Be15a] ist ein Digitales Bibliothekssystem, das den visuellen Zugang zu zeitbasierten Primärdaten, und damit deren Wiederverwendung ermöglicht. In einer Überblicksvisualisierung können Domänexperten in Tageverläufen explorieren (siehe Abbildung 6a). Die Zeitserien stammen aus einem frei zugängigen Klimadaten-Repository (‘Open Data’). Zur inhaltsbasierten Suche in den Daten können interessante Zeitverläufe direkt ausgewählt werden (‘query-by-example’). Alternativ untersützt VisInfo das Skizzieren von Suchtermen (‘query-by-sketch’), illustriert in Abbildung 6b. Suchergebnisse werden in VisInfo visuell-interaktiv repräsentiert, so können Domänenexperten die gefundenen zeitbasierten Primärdaten direkt einsehen und darin browsen. Facetten ermöglichen den Einbezug von Metadaten bei der Treffereschließung. Zudem lassen sich die Suchtreffer im Kontext einer Kalender- und einer Geobasierten Visualisierung interpretieren.

MotionExplorer [Be13] ist ein exploratives Suchsystem zur Analyse menschlicher Bewegungsdaten, Abbildung 4 verleiht einen visuellen Eindruck des Systems. Im Anwendungsbeispiel werden verschiedene menschliche Bewegungsabläufe einer großen Motion-Capturing Datenbank analysiert. Einige davon beschreiben Bewegungen eines “Hampelmanns”. In zwei Überblicksvisualisierungen können Domänenexperten große Mengen an Dateninhalt explorieren. Dabei unterscheiden sich die beiden visuellen Benutzerschnittstellen durch ihre Fokussierung auf menschliche Posen (oben links) und Bewegungen (oben rechts). Interessante Posen können für die inhaltsbasierte Suche verwendet werden (‘query-by-example’). Auf die Definition einer Start- und Endpose hin sucht ein Retrievalalgorithmus automatisch nach entsprechenden Bewegungsabläufen. Suchergebnisse sind in MotionExplorer visuell-interaktiv explorierbar (unten). So werden alle gefundenen Bewegungsabläufe in einer Liste dargestellt, ein Sliderwerkzeug ermöglicht das Browsen in der temporalen Information der Suchtreffer, ähnlich wie bei der Navigation eines Videos. Ein Animationsfenster ermöglicht den visuellen Vergleich aller Suchtreffer (unten rechts).

VisInfo und MotionExplorer stellen zwei der ersten explorativen Suchsysteme für zeitbasierte Primärdaten dar. Beide Designstudien belegen die Anwendbarkeit der Konzepte, Richtlinien, und Techniken dieser Dissertation. Beide explorativen Suchsysteme enthalten konkrete Lösungen für die sechs vorrangigen Problemstellungen (vgl. Kapitel 2).

8 Zusammenfassung

In dieser Dissertation habe ich Konzepte, Richtlinien, Techniken, und Systeme für die explorative Suche in zeitbasierten Primärdaten vorgestellt. Als vorrangige Nutzergruppe

galten Domänenexperten in der datenzentrierten Forschung. Die wissenschaftlichen Beiträge erstrecken sich über den gesamten Workflow der Zeitserienanalyse, beginnend mit "rohen" zeitbasierten Primärdaten, bis hin zu Systemen mit visuell-interaktiven Benutzerschnittstellen für die Zeitserienanalyse. Die Suche in zeitbasierten Primärdaten wird durch visuelle Benutzerschnittstellen unterstützt ('query-by-example', 'query-by-sketch'). Überblicksvisualisierungen über den Dateninhalt und neuartige visuell-interaktive Techniken zur Analyse von Zusammenhängen zwischen dem Dateninhalt und Metadaten, stellen die wissenschaftlichen Beiträge zur Exploration dar. In dieser Dissertation wurde Visual Analytics als ein vielversprechender Weg für das Design von effektiven und effizienten visuell-interaktiven Analysesystemen vorgestellt. Mit der Möglichkeit zur Visualisierung von Zwischenergebnissen des Workflows ermögliche ich den Einbezug von Nutzern in das Design. Durch diese Herangehensweise entstehen robuste und zugleich generalisierbare Workflows für die Zeitserienanalyse, die im Anschluss vollautomatisch ausgeführt werden können. In Kombination mit dem nutzerzentrierten Ansatz, ermöglicht ein hoher Grad an Automatisierung auch die Simplifizierung von visuellen Benutzerschnittstellen, was zu einfachen und intuitiven explorativen Suchsystemen führt. Dies wird auch in den beiden Anwendungsbeispielen deutlich, in denen explorative Suchsysteme mit einfachen und intuitiven Interaktionsdesigns demonstriert wurden. Unterstüztend wirkte in diesem Zusammenhang auch die iterative Herangehensweise bei der Entwicklung, sowie die fortlaufende Evaluierung der Systeme mit der jeweiligen Zielgruppe. Beide Systeme zeigen auf, wie die datenzentrierte Forschung durch die explorative Suche unterstützt werden kann. Durch die vorgestellten Analysetechniken trägt diese Dissertation auch zum Prozess bei, Hypothesen künftig effizienter und effektiver testen und formulieren zu können. So wurden eine Reihe von Techniken vorgestellt, die, in Kombination mit der Definition von Interessantheitsmaßen durch den Nutzer, ähnliche Muster und interessante Zusammenhänge vollkommen automatisch erkennen und visuell repräsentieren.

An die Ergebnisse dieser Dissertation knüpfen eine Reihe von Forschungsfragen an. Naheliegend ist der Einbezug neuer Daten, Nutzer, Analysetasks. Darüber hinaus eignet sich die entstandene Basis durch die visuell-interaktive Herangehensweise für kollaborative Forschungsunternehmungen. Weiter lässt sich der Ansatz der Interessantheitsdefinition durch den Nutzer, und die damit verbundene Automatisierung des Forschungsprozesses, auf weitere Szenarien übertragen. Schließlich eröffnet die Vision einer nutzerbasierten Definition von Ähnlichkeit hochdimensionaler Datenobjekte neuartige Forschungsfragen für datenanalytische Forschungsgebiete, wie etwa Informationsvisualisierung, Visual Analytics, Visual Data Mining, Machine Learning, oder Active Learning.

Literaturverzeichnis

[Be11] Bernard, Jürgen; von Landesberger, Tatiana; Bremm, Sebastian; Schreck, Tobias: Multiscale visual quality assessment for cluster analysis with Self-Organizing Maps. In: IS&T/SPIE Conference on Visualization and Data Analysis (VDA). SPIE Press, S. 78680N.1 – 78680N.12, 2011.

[Be12a] Bernard, Jürgen; Ruppert, Tobias; Goroll, Oliver; May, Thorsten; Kohlhammer, Jörn: Visual-Interactive Preprocessing of Time Series Data. In (Kerren, Andreas; Seipel, Stefan, Hrsg.): SIGRAD. Jgg. 81 in Linköping Electronic Conference Proceedings. Eurographics, S. 39–48, 2012.

[Be12b] Bernard, Jürgen; Ruppert, Tobias; Scherer, Maximilian; Kohlhammer, Jörn; Schreck, Tobias: Content-based Layouts for Exploratory Metadata Search in Scientific Research Data. In: Joint Conference on Digital Libraries (JCDL). ACM, S. 139–148, 2012.

[Be12c] Bernard, Jürgen; Ruppert, Tobias; Scherer, Maximilian; Schreck, Tobias; Kohlhammer, Jörn: Guided Discovery of Interesting Relationships Between Time Series Clusters and Metadata Properties. In: International Conference on Knowledge Management and Knowledge Technologies (i-KNOW). ACM, New York, NY, USA, S. 22:1–22:8, 2012.

[Be13] Bernard, Jürgen; Wilhelm, Nils; Kruger, Bjorn; May, Thorsten; Schreck, Tobias; Kohlhammer, Jörn: MotionExplorer: Exploratory search in human motion capture data based on hierarchical aggregation. IEEE Transactions on Visualization and Computer Graphics (TVCG), 19(12):2257–2266, 2013.

[Be14] Bernard, Jürgen; Steiger, Martin; Widmer, Sven; Lücke-Tieke, Hendrik; May, Thorsten; Kohlhammer, Jörn: Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. Computer Graphics Forum (CGF), 33(3):291–300, 2014.

[Be15a] Bernard, Jürgen; Daberkow, Debora; Fellner, Dieter; Fischer, Katrin; Koepler, Oliver; Kohlhammer, Jörn; Runnwerth, Mila; Ruppert, Tobias; Schreck, Tobias; Sens, Irina: VisInfo: a digital library system for time series research data based on exploratory search - a user-centered design approach. Internat. Journal on Digital Libraries, 16(1):37–59, 2015.

[Be15b] Bernard, Jürgen; Steiger, Martin; Mittelstädt, Sebastian; Thum, Simon; Keim, Daniel; Kohlhammer, Jörn: A survey and task-based quality assessment of static 2D colormaps. In: SPIE, Visualization and Data Analysis (VDA). Jgg. 9397, 2015.

[Be15c] Bernard, Jürgen: Exploratory search in time-oriented primary data. dissertation, Technische Universität Darmstadt, Graphisch-Interaktive Systeme (GRIS), Darmstadt, Germany, 2015.

[CMS99] Card, Stuart K.; Mackinlay, Jock D.; Shneiderman, Ben, Hrsg. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Publishers, CA, USA, 1999.

[HTT09] Hey, Anthony J. G.; Tansley, Stewart; Tolle, Kristin M.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

[Ke10] Keim, D.; Kohlhammer, Jörn; Ellis, G.; Mansmann, F., Hrsg. Mastering the Information Age: Solving Problems with Visual Analytics. VisMaster, http://www.vismaster.eu/book/, 2010.

[Ma06] Marchionini, Gary: Exploratory Search: From Finding to Understanding. Commun. ACM, 49(4):41–46, 2006.

[WR09] White, Ryen W.; Roth, Resa A.: Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lect. on Information Concepts, Retrieval, and Services, 1:1–98, 2009.

Jürgen Bernard studierte Informatik an der TU Darmstadt, mit den Schwerpunkten Computergraphik und Bionik. Seine Diplomarbeit schrieb er 2009, über Methoden zur visuell-interaktiven Clusteranalyse. Er promovierte an der TU Darmstadt, zunächst am Fachgebiet Graphisch-Interaktive Systeme (GRIS) und schließlich am Fraunhofer Institut für Graphische Datenverarbeitung (IGD). Seine Dissertation mit dem Titel "Exploratory Search in Time-Oriented Primary Data" [Be15c] verteidigte Jürgen Bernard im Herbst 2015. Schwerpunkte seiner wissenschaftlichen Arbeit sind die visuell-interaktive Analyse von multidimensionalen und zeitbasierten Daten. Seine Anwendungsdomänen erstrecken sich von Digitalen Bibliotheken, über Klimaforschung, menschliche Bewegungsanalyse, bis hin zur Analyse von Patientendaten. Als Autor von über 40 Publikationen hat Jürgen Bernard, neben der nutzerzentrierten Projektarbeit, seinen eigenen Forschungsschwerpunkt definiert, den er seit März 2016 als Post-Doc an der TU Darmstadt weiter verfolgt.