# METIS in PArADISE
## *Provenance Management bei der Auswertung von Sensordatenmengen für die Entwicklung von Assistenzsystemen*

Position Statement
Andreas Heuer, Universität Rostock, Institut für Informatik

andreas.heuer@uni-rostock.de

## 1 Einleitung

Wissenschaftliche Experimente, die durch Messungen oder ständige Beobachtungen laufend eine Vielzahl von Daten erheben, müssen durch effiziente Analyseverfahren auf diesen Mess- oder Sensordaten unterstützt werden. Die Mess- und Sensordaten sind so zu verwalten, dass sie im Sinne des Provenance Management rückverfolgbar werden. An der Universität Rostock soll ein langfristiges Forschungsvorhaben im Bereich der Informatik und Elektrotechnik etabliert werden, in dem wissenschaftliche Experimente in der Informatik, der Zellbiologie und der Medizin (neurodegenerative Erkrankungen) auf eine solche Weise unterstützt werden.

Im Bereich der Informatik ist das experimentelle Anwendungsgebiet das der Erforschung und systematischen Entwicklung von Assistenzsystemen. Da in Assistenzsystemen unterstützte Personen durch eine Vielzahl von Sensoren beobachtet werden, müssen auch Privatheitsaspekte bereits während der Phase der Modellbildung berücksichtigt werden, um diese bei der konkreten Konstruktion des Assistenzsystems automatisch in den Systementwurf zu integrieren. Somit gibt es für die Datenbankforscher unter anderem folgende Teilprobleme, die zu lösen sind und die in zwei langfristigen Projektgruppen des Datenbanklehrstuhls zusammengefasst sind:

**PArADISE** (Privacy AwaRe Assistive Distributed Information System Environment): In dieser Projektgruppe werden Techniken zur Auswertung von großen Mengen von Sensordaten entwickelt, die definierte Privatheitsansprüche der späteren Nutzer per Systemkonstruktion erfüllen.

**METIS** (Management, Evolution, Transformation und Integration von Schemata): In dieser Projektgruppe geht es unter anderem um Integrationsverfahren für heterogene Datenbanken, die per se bereits eine Rolle bei der Zusammenführung der unterschiedlichen Datenquellen des Assistenzsystems spielen. Als Seiteneffekt werden die für die Datenintegration entwickelten Techniken der *Global-as-local-view-extension* (GaLVE) auch für

die nötigen inversen Abbildungen zum Provenance Management bei der Verarbeitung der Sensordaten und Ableitung von (Situations-, Aktivitäts-, Intentions-)Modellen benötigt.

Wissenschaftliche Experimente außerhalb der Informatik, etwa zur Auswertung von Sensordaten bei Patienten mit neurodegenerativen Erkrankungen [DEW+12] liegen nicht im Fokus dieses Beitrags.

Im Folgenden werden wir im Abschnitt 2 kurz die Architektur von Assistenzsystemen einführen, wobei wir den Schwerpunkt auf die Phase der Situations-, Aktivitäts- und Intentionserkennung legen. Danach werden wir die Erforschung und Entwicklung von Assistenzsystemen als ein wissenschaftliches Experiment ansehen, in dem Forscher eine große Anzahl von Sensordaten auswerten müssen und aus ihnen Situations-, Aktivitäts- und Intentionsmodelle entwickeln (Abschnitt 3). Zwei der Grundlagenforschungsthemen für die Datenbankforscher, Provenance Management und die Integration von Privatheitsanforderungen, werden wir in Abschnitt 4 einführen.

## 2 Assistenzsysteme

Ähnlich einem menschlichen Assistenten soll ein Assistenzsystem mich unterstützen, im Hintergrund arbeiten (ambient), mich nicht stören, zum richtigen Zeitpunkt eingreifen und Hilfe anbieten (diese in üblichen Fällen auf optischem oder akustischem Wege), vertrauenswürdig und diskret sein und sich bei Bedarf abschalten lassen.

Um seine Assistenzaufgaben zu erfüllen, besteht ein Assistenzsystem üblicherweise aus fünf Schichten (siehe Abbildung 1). Dabei deutet die Pyramidenform an, dass in der untersten Schicht dauernd viele Daten (etwa von Sensoren) erzeugt werden, in der obersten Schicht aber nur im Bedarfsfall (also eher selten) ein akustischer oder optischer Hinweis, also eine geringe Datenmenge, ausgegeben wird.

**Sensoren** in der Umgebung der Person sollen Situation und Tätigkeit der Person erfassen, um ihr assistieren zu können. **Ortungskomponenten** sollen die genaue Position der Person bestimmen, etwa zur Detektion dementer Patienten mit Weglauftendenzen. Sensoren und Ortungskomponenten befinden sich in der Umgebung, in benutzten Geräten oder am Körper der Person (Armband, Brille, ...).

Damit ein Assistenzsystem seine Aufgabe erfüllen kann, müssen verschiedene (heterogene) Geräte in der Umgebung der Person **vernetzt** und zur Erreichung des Assistenzziels **spontan gekoppelt** werden.

Sensordaten müssen gefiltert, erfasst, ausgewertet, verdichtet und teilweise langfristig verwaltet werden. Aufgrund der extrem großen Datenmenge (Big Data) muss die **Verarbeitung verteilt** erfolgen: teilweise eine Filterung und Verdichtung schon im Sensor, im nächsterreichbaren Prozessor (etwa im Fernseher oder im Smart Meter in der Wohnung) und im Notfall über das Internet in der Cloud. Neben Daten des Assistenzsystems müssen auch fremde Daten etwa über das Internet berücksichtigt werden.

Nach Auswertung der Sensordaten erfolgt die **Situations-, Handlungs- und Intentionserkennung** (siehe den folgenden Abschnitt 3 für die Entwicklung diesbezüglicher Modelle).


Abbildung 1: Pyramidenarchitektur von Assistenzsystemen

## 3 Big Data Analytics für die Entwicklung von Assistenzsystemen

Aufgrund der Sensor- und Ortungsdaten sowie der weiteren über das Internet erhältlichen Daten muss das Assistenzsystem eine Situations- und Handlungserkennung vornehmen sowie eine Handlungsvorhersage (Intentionserkennung), um proaktiv eingreifen zu können.

Die Situation ist dabei die aktuelle Umgebungsinformation, die Handlung das, was die Person, der assistiert wird, gerade durchführt. Die Intentionserkennung oder Handlungsvorhersage muss voraussagen, was die die Person in Kürze tun wird.

Auch die Handlungs- und Intentionserkennung ist ein aktueller Forschungsgegenstand der Informatiker an der Universität Rostock [KNY+14]. Dabei erheben die Forscher in langen Versuchsreihen eine extrem hohe Anzahl von Sensordaten, aus denen sie mit diversen Analyseverfahren die entsprechenden Modelle ableiten. Diese Analyseverfahren sind — wenn sie ohne Datenbankunterstützung auf Dateisystemen mit Analysewerkzeugen wie R ausgeführt werden — mehrwöchige Prozesse. Ziel der Forscher ist neben der Modellbildungen für Handlung und Intention die Erkenntnis, wie die große Anzahl von Sensoren im Versuch für den praktischen, späteren Einsatz des Assistenzsystems drastisch eingeschränkt werden kann, ohne die Vorhersagequalität zu mindern. Für die Ableitung dieser Informationen müssen unter anderem alle Analysefunktionen invertiert werden, um die für die Modellbildung entscheidenden Anteile der Originaldaten zu finden. Letzteres ist auch ein Problem im Provenance Management, das die Experimentverläufe mit den Ergebnisableitungen begleiten soll. Eine Einschränkung sowohl der Anzahl der Sensoren als auch der Menge und Granularität der erfassten Daten ist auch aus einem anderen Grund wichtig: sie kann die Privatheitsanforderungen der Nutzer des Assistenzsystems realisieren helfen.

## 4 Nötige Grundlagenforschung: Privatheit und Provenance

Die Entwickler der Analysewerkzeuge müssen ihr Assistenzziel und die notwendigen Sensordaten zur Erreichung des Ziels und zur grundlegenden Situations-, Handlungs- und Intentionserkennung formulieren. Diese Zielformulierung wird dann in Anfragen auf Datenbanken transformiert. Weiterhin können die **Privatheitsansprüche** des Nutzers vordefiniert oder von jedem Nutzer selbst individuell verschärft werden. Auch diese Privatheitsansprüche werden in Anfragen (Sichten) auf Datenbanken umgesetzt. Durch Abgleich des Informationsbedarfs des Assistenzsystems und der Privatheitsansprüche des Nutzers kann dann die Datenbankkomponente des Assistenzsystems entscheiden, wie die Menge an Sensordaten selektiert, reduziert, komprimiert oder aggregiert werden muss, um beiden Parteien im System gerecht zu werden[Gru14].

Ein entscheidendes Kriterium für die Vertrauenswürdigkeit eines Assistenzsystems ist noch die Frage, wie nah am Sensor die Daten bereits reduziert und verdichtet werden können: Wenn der Sensor so intelligent ist, dass er bestimmte Filtermechanismen von Datenbanksystemen beherrscht, so kann dieser bereits eine Vorfilterung vornehmen. Nur die für das Assistenzziel unabdingbaren Daten, die die Privatheit des Nutzers nicht verletzen, können dann im Rahmen des Cloud Data Management des Anbieters der Assistenzfunktionalität entfernt und verteilt gespeichert werden.

Ein Teil der Datenbankforschung im Gebiet **Provenance** konzentriert sich auf einzelne Modelle, Operatoren und Annotationen für die Rekonstruktion der Rohdaten [CAB+14]. Eine durchgängige und automatisierte Vorgehensweise fehlt jedoch. Daneben wurden Provenance-Eigenschaften charakterisiert und formalisiert [BKT01], meist für das Tracing der Daten vorwärts [GKT07] und nicht für die automatische Rekonstruktion rückwärts. Für Letzteres werden Charakterisierungen inverser Analyseoperatoren und inverser Schemaabbildungen benötigt. Inverse Schemaabbildungen und ihre Eigenschaften werden zwar untersucht [Fag07], derzeit aber noch nicht im Bereich Provenance angewandt, sondern eher bei der Schema-Evolution [CMDZ10].

In [FHLM96] wurde die Idee der **inversen Schemaabbildungen** bereits für die Integration heterogener Datenbanken eingeführt. Diese Idee wurde in [SBLH14] auf neuere Entwicklungen in der Theorie inverser Schemaabbildungen angepasst. Dabei wurden die klassischen Schemaabbildungen verallgemeinert auf den Fall, dass nicht alle relevanten Daten im (relationalen) Schema repräsentiert werden können, sondern zusätzlich auf Instanzebene erfasst werden müssen. Diese zusätzlichen Annotationen helfen bei der inversen Abbildung verdichteter Daten.

Im Gegensatz zur traditionellen Datenintegration sollen Anfragen im integrierten System bei diesem Ansatz nicht an das globale Schema gestellt werden, sondern weiterhin an die lokalen Schemata, wobei diese um Daten aus anderen Quellen erweitert werden. Dieser Ansatz wird als *Global-as-local-view-extension* (GaLVE) bezeichnet.

**Inverse Schema-Instanz-Abbildungen** werden nun auch für die Invertierung, also Rückverfolgung, von Analyseprozessen, also allgemeineren Datenbankanfragen, benötigt. Bisher wurden allerdings für die Datenbankintegration nur einfache Anfragen, bestehend aus Selektion, Projektion und Verbund, berücksichtigt. Analyseprozesse erfordern eine Erwei-

terung der bestehenden Technik auf statistische Funktionen (skalare Funktionen, Aggregatfunktionen, OLAP) und allgemeine Workflows. Hier muss ermittelt werden, welche Zusatzinformationen zur Gewährleistung der Rückverfolgbarkeit erfasst werden müssen.

Da in diesem seitenbeschränkten Beitrag leider viele Aspekte und Literaturhinweise nicht aufgenommen werden konnten, findet sich eine Langfassung dieses Position Papers unter `www.andreas-heuer.de/files/metis-in-paradise-long.pdf`

## Literatur


[BKT01] Peter Buneman, Sanjeev Khanna und Wang Chiew Tan. Why and Where: A Characterization of Data Provenance. In Jan Van den Bussche und Victor Vianu, Hrsg., *ICDT*, Jgg. 1973 of *Lecture Notes in Computer Science*, Seiten 316–330. Springer, 2001.

[CAB+14] Lucian Carata, Sherif Akoush, Nikilesh Balakrishnan, Thomas Bytheway, Ripduman Sohan, Margo Seltzer und Andy Hopper. A Primer on Provenance. *Commun. ACM*, 57(5):52–60, Mai 2014.

[CMDZ10] Carlo Curino, Hyun Jin Moon, Alin Deutsch und Carlo Zaniolo. Update Rewriting and Integrity Constraint Maintenance in a Schema Evolution Support System: PRISM++. *PVLDB*, 4(2):117–128, 2010.

[DEW+12] Martin Dyrba, Michael Ewers, Martin Wegrzyn, Ingo Kilimann, Claudia Plant, Annahita Oswald, Thomas Meindl, Michela Pievani, Arun L. W. Bokde, Andreas Fellgiebel, Massimo Filippi, Harald Hampel, Stefan Klöppel, Karlheinz Hauenstein, Thomas Kirste und Stefan J. Teipel. Combining DTI and MRI for the Automated Detection of Alzheimer's Disease Using a Large European Multicenter Dataset. In Pew-Thian Yap, Tianming Liu, Dinggang Shen, Carl-Fredrik Westin und Li Shen, Hrsg., *Multimodal Brain Image Analysis - Second International Workshop, MBIA 2012, Held in Conjunction with MICCAI 2012, Nice, France, October 1-5, 2012. Proceedings*, Jgg. 7509 of *Lecture Notes in Computer Science*, Seiten 18–28. Springer, 2012.

[Fag07] Ronald Fagin. Inverting schema mappings. *ACM Trans. Database Syst.*, 32(4), 2007.

[FHLM96] Guntram Flach, Andreas Heuer, Uwe Langer und Holger Meyer. Transparente Anfragen in föderativen Datenbanksystemen. In *Proceedings zum Workshop Föderierte Datenbanken*, Seiten 45–49, 1996.

[GKT07] Todd J. Green, Gregory Karvounarakis und Val Tannen. Provenance semirings. In Leonid Libkin, Hrsg., *PODS*, Seiten 31–40. ACM, 2007.

[Gru14] Hannes Grunert. Distributed Denial of Privacy. In Erhard Plödereder, Lars Grunske, Eric Schneider und Dominik Ull, Hrsg., *44. Jahrestagung der Gesellschaft für Informatik, Informatik 2014, Big Data - Komplexität meistern, 22.-26. September 2014 in Stuttgart, Deutschland*, Jgg. 232 of *LNI*, Seiten 2299–2304. GI, 2014.

[KNY+14] Frank Krüger, Martin Nyolt, Kristina Yordanova, Albert Hein und Thomas Kirste. Computational State Space Models for Activity and Intention Recognition. A Feasibility Study. *PLOS ONE*, November 2014. 9(11): e109381. doi:10.1371/journal.pone.0109381.

[SBLH14] Georgi Straube, Ilvio Bruder, Dortje Löper und Andreas Heuer. Data Integration in a Clinical Environment Using the Global-as-Local-View-Extension Technique. In *HIS*, Jgg. 8423 of *Lecture Notes in Computer Science*, Seiten 148–159. Springer, 2014.