Das PArADISE-Projekt

Big-Data-Analysen für die Entwicklung von Assistenzsystemen (Extended Abstract)*

Andreas Heuer
Lehrstuhl DBIS, Institut für Informatik
Universität Rostock
18051 Rostock, Deutschland
heuer@informatik.uni-rostock.de

Holger Meyer
Lehrstuhl DBIS, Institut für Informatik
Universität Rostock
18051 Rostock, Deutschland
hme@informatik.uni-rostock.de

ZUSAMMENFASSUNG

Bei der Erforschung und systematischen Entwicklung von Assistenzsystemen fallen eine große Menge von Sensordaten an, aus denen Situationen, Handlungen und Intentionen der vom Assistenzsystem unterstützten Personen abgeschätzt (modelliert) werden müssen. Neben Privatheitsaspekten, die bereits während der Phase der Modellbildung berücksichtigt werden müssen, sind die *Performance* des Analysesystems sowie die *Provenance* (Rückverfolgbarkeit von Modellierungsentscheidungen) und die *Preservation* (die langfristige Aufbewahrung der Forschungsdaten) Ziele unserer Projekte in diesem Bereich. Speziell sollen im Projekt PArADISE die Privatheitsaspekte und die Performance des Systems berücksichtigt werden. In einem studentischen Projekt wurde innerhalb einer neuen *experimentellen* Lehrveranstaltung im reformierten Bachelor- und Master-Studiengang Informatik an der Universität Rostock eine Systemplattform für eigene Entwicklungen geschaffen, die auf Basis von klassischen zeilenorientierten Datenbanksystemen, aber auch spaltenorientierten und hauptspeicheroptimierten Systemen die Analyse der Sensordaten vornimmt und für eine effiziente, parallelisierte Verarbeitung vorbereitet. Ziel dieses Beitrages ist es, die Ergebnisse dieser studentischen Projektgruppe vorzustellen, insbesondere die Erfahrungen mit den gewählten Plattformen PostgreSQL, DB2 BLU, MonetDB sowie R (als Analysesystem) zu präsentieren.

1. EINLEITUNG

Ein Forschungsschwerpunkt am Institut für Informatik der Universität Rostock ist die Erforschung und systematische Entwicklung von Assistenzsystemen, etwa im DFG-Graduiertenkolleg MuSAMA. Da in Assistenzsystemen unterstützte Personen durch eine Vielzahl von Sensoren beobachtet werden, müssen bei der datengetriebenen Modellierung von Situationen, Handlungen und Intentionen der Personen aus großen Datenmengen mittels Machine-Learning-Methoden entsprechende Modelle abgeleitet werden: ein Performance-Problem bei einer Big-Data-Analytics-Fragestellung.

Da Personen *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 die Teilprobleme der performanten Berechnung der Modelle als auch der Wahrung der Privatheitsansprüche des Nutzers, die zu lösen sind und die in einer langfristigen Projektgruppe des Datenbanklehrstuhls angegangen werden: im Projekt **PArADISE** (Privacy AwaRe Assistive Distributed Information System Environment) werden effiizionte Techniken zur Auswertung von großen Mengen von Sensordaten entwickelt, die definierte Privatheitsansprüche der späteren Nutzer per Systemkonstruktion erfüllen.

Während wir in [Heu15] ausführlicher auf die Verknüpfung der Aspekte *Privatheit* (Projekt PArADISE) und *Provenance* (Projekt METIS) eingegangen sind, werden wir uns in diesem Beitrag auf die beiden Schwerpunkte des PArADISE-Projektes konzentrieren, das ist neben der Privatheit die *Performance* durch Parallelität und Verteilung.

2. ASSISTENZSYSTEM-ENTWICKLUNG ALS BIG-DATA-PROBLEM

Um seine Assistenzaufgaben zu erfüllen, besteht ein Assistenzsystem üblicherweise aus fünf Schichten [Heu15]. In der untersten Schicht werden ständig viele Daten (etwa von Sensoren) erzeugt, in der obersten Schicht wird aber nur im Bedarfsfall (also eher selten) ein akustischer oder optischer Hinweis, also eine geringe Datenmenge, ausgegeben.

In der mittleren der fünf Schichten müssen Sensordaten 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, beispielsweise Wartungspläne beim Auto oder die elektronische Patientenakte beim Patienten. Allgemein können hier natürlich auch die Daten sozialer Netz-

*Eine Langfassung dieses Artikels ist erhältlich als [HM15] unter http://www.ls-dbis.de/digbib/dbis-tr-cs-04-15.pdf

27^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 26.05.2015 - 29.05.2015, Magdeburg, Germany.

werke, Kalenderdaten der Nutzer oder Wettervorhersage-Daten ausgewertet werden, falls sie für das Assistenzziel eine Rolle spielen.

Eine Kernaufgabe bei der Erforschung und Entwicklung ist die datengetriebene Modellierung von Situationen, Handlungen und Intentionen, die eine Fragestellung im Forschungsgebiet Big Data Analytics sind. Big Data [Mar15] ist ein derzeitiges Hype-Thema nicht nur in der Informatik, das in seiner technischen Ausprägung auf vielfältige Forschungsprobleme führt. Technisch gesehen sind Big-Data-Probleme mit den vier *V* (Volume, Velocity, Variety, Veracity) charakterisiert. *Big Data Analytics* ist nun das Problem komplexer Analysen auf diesen Daten. In Datenbankbegriffen sind diese komplexen Analysen iterative Anfrageprozesse.

3. DIE VIER P ZU DEN VIER V

Die Forschungsschwerpunkte der Rostocker Datenbankgruppe lassen sich in diesem Zusammenhang mit vier *P* charakterisieren, die im Folgenden näher erläutert werden sollen.

Forschung und Entwicklung: In der Forschungs- und Entwicklungsphase eines Assistenzsystems ist das vorrangige Ziel, eine effiziente Modellbildung auf großen Datenmengen zu unterstützen. Dabei sollte möglichst automatisch eine Selektion der Daten (Filterung wichtiger Sensordaten nach einfachen Merkmalen) und eine Projektion der Daten (die Beschränkung der großen Sensormenge auf wenige, besonders aussagekräftige Sensoren) vorgenommen werden. Die nötige Effizienz in dieser Phase führt auf unser Forschungsthema **P3: Performance**. Da während der Entwicklung bei fehlerhafter Erkennung von Handlungen und Intentionen die dafür zuständigen Versuchsdaten ermittelt werden müssen, führt die Rückverfolgbarkeit der Analyseprozesse in der Entwicklung auf unsere Forschungsthemen **P2: Provenance Management** und **P4: Preservation** (Langfristarchivierung von Forschungsdaten).

Einsatz: In der Einsatzphase eines Assistenzsystems sind dagegen Privatheitsansprüche vorherrschend, die im Gesamtsystem durch stufenweise Datensparsamkeit erreicht werden können (unser Forschungsthema **P1: Privatheit**). Eine weitere Verdichtung (auch Reduktion und Aggregation) der live ausgewerteten Daten unterstützen aber nicht nur die Privatheit, sondern auch die Performance.

Die vier *P* behandeln wir in drei langfristigen Forschungsprojekten (METIS, PArADISE, HyDRA), in diesem Beitrag konzentrieren wir uns auf den Aspekt **P3 (Performance)** des PArADISE-Projektes.

4. DAS PARADISE-PROJEKT

Im Projekt **PArADISE** (Privacy AwaRe Assistive Distributed Information System Environment) arbeiten wir derzeit an Techniken zur Auswertung von großen Mengen von Sensordaten, die definierte Privatheitsansprüche der späteren Nutzer per Systemkonstruktion erfüllen.

Ein erster Prototyp ist von einer studentischen Arbeitsgruppe erstellt worden. Derzeit können Analysen zur Modellbildung auf Sensordaten in SQL-92, SQL:2003 oder iterativen Ansätzen über SQL-Anweisungen realisiert und auf die Basissysteme DB2 (zeilenorientiert oder spaltenorientiert: DB2 BLU), PostgreSQL (zeilenorientiert) sowie MonetDB (spaltenorientiert und hauptspeicheroptimiert) abgebildet werden.

Während die grundlegenden Forschungsarbeiten zu PArADISE durch zwei Stipendiaten des Graduiertenkollegs MuSAMA (Hannes Grunert und Dennis Marten) in 2013 und 2014 starteten, wurden die ersten softwaretechnischen Umsetzungen des Projektes durch eine studentische Projektgruppe im Wintersemester 2014/2015 vorgenommen. Hier wurden dann verschiedene SQL-Anfragen und R-Programme zur Lösung der grundlegenden Regressions- und Korrelationsprobleme entwickelt, wobei als Vorgabe (zum Vergleich) folgende fünf Stufen realisiert werden sollten:

1. Umsetzung von Regression und Korrelation in Standard-SQL-92 (also per Hand, da keine Analysefunktionen außer den klassischen Aggregatfunktionen wie COUNT, SUM und AVG vorhanden).
2. Umsetzung in SQL:2003 mit den entsprechenden OLAP-Funktionen.
3. Umsetzung mit rekursivem oder iterativem SQL, sofern in den Systemen möglich.
4. Eine Integration der SQL-Anfrage mit R-Auswertungen.
5. Eine R-Auswertung pur ohne Kopplung an SQL.

Die in MuSAMA bisher verwendete Lösung mit *Plain R* wies dabei die schlechteste Effizienz auf, auch wenn man den Prozess des initialen Ladens der Daten in den Hauptspeicher herausrechnet. Unter den Varianten mit einer Analyse in reinem SQL-92 (Regression per Hand mit Aggregatfunktionen umgesetzt) war die MonetDB-Lösung etwas besser als die DB2-Variante, PostgreSQL fiel stärker ab. Die SQL:2003-Lösung konnte in MonetDB mangels vorhandener OLAP- und Rekursions-Fähigkeiten nicht umgesetzt werden, DB2 war hier wiederum deutlich besser als PostgreSQL. Weiterhin bemerkt man im Vergleich von SQL-92 und SQL:2003, dass der Optimierer von DB2 als auch PostgreSQL die direkte Verwendung der OLAP-Funktionen belohnt. Die beste Performance aller Varianten erreichte jedoch MonetDB mit integrierten R-Funktionen.

5. DANKSAGUNGEN

Wir danken der studentischen Projektgruppe PArADISE im Wintersemester 2014/2015, die im Rahmen einer experimentellen Projekt-Lehrveranstaltung die Basis für die softwaretechnische Umsetzung des PArADISE-Projektes gelegt hat: Pia Wilsdorf, Felix Köppl, Stefan Lüdtke, Steffen Sachse, Jan Svacina, Dennis Weu.

6. LITERATUR

[Heu15] Heuer, A.: METIS in PArADISE: Provenance Management bei der Auswertung von Sensordatenmengen für die Entwicklung von Assistenzsystemen. In: *Lecture Notes in Informatics, Band 242, BTW 2015 Workshop-Band, 131 – 135*, 2015.

[HM15] Heuer, A.; Meyer, H.: Das PArADISE-Projekt: Big-Data-Analysen für die Entwicklung von Assistenzsystemen. Technischer Bericht CS-04-15, Institut für Informatik, Universität Rostock, 2015.

[Mar15] Markl, V.: Gesprengte Ketten - Smart Data, deklarative Datenanalyse, Apache Flink. *Informatik Spektrum*, Band 38, Nr. 1, S. 10–15, 2015.