Editorial

Bernhard Seeger¹ · Birgitta König-Ries² · Theo Härder³

© Springer-Verlag Berlin Heidelberg 2016

1 Schwerpunktthema: Data Management for Bio- and Geosciences

Data-driven Science, eScience, Digital Humanities sind nur einige der Schlagworte, die den dramatischen Wandel in der Art, wie Forschung betrieben wird, in fast allen wissenschaftlichen Disziplinen umreißen. Daten und Analysemethoden haben für eine datengetriebene Wissenschaft in den letzten Jahren in nahezu allen Disziplinen zu neuen Möglichkeiten des Erkenntnisgewinns, aber auch zu neuen Herausforderungen geführt. Dies gilt nicht nur für die Naturwissenschaften, wie z. B. Meteorologie, Genetik und Ökologie, sondern auch für die Geistes- und Sozialwissenschaften. Die Gründe hierfür sind insbesondere die stetig weiterentwickelten Techniken für eine kostengünstige Datenerhebung, die rapid fallenden Kosten für die Datenspeicherung sowie die leistungsfähigeren Netzwerke, wodurch die Nutzer allgegenwärtig auf die Daten zugreifen können. Dadurch haben einerseits Wissenschaftlerinnen und Wissenschaftler vielfältige Chancen, aus diesen Datenmengen

neues Wissen ableiten zu können. Andererseits stellen diese sehr heterogenen und zum Teil auch riesigen Datenmengen die Forschenden vor erhebliche Herausforderungen beim Datenmanagement.

Für die Datenbankforschung ergeben sich durch den Wandel zu einer stärker datengetriebenen Wissenschaft neue interessante Problemstellungen, die erstmals Anfang der Jahrtausendwende in den wegweisenden Arbeiten von Jim Gray aufgegriffen wurden. Die bisherigen Forschungsarbeiten lassen sich grob in zwei Gebiete aufteilen. Zum einen geht es um die effiziente und skalierbare Verarbeitung, Analyse und Visualisierung komplexer wissenschaftlicher Daten, wie z. B. Zeitreihen und Graphen. Zum anderen gibt es vielfältige Arbeiten im Bereich der semantischen Integration heterogener wissenschaftlicher Datenbestände, der semantischen Annotation von erfassten und der davon abgeleiteten Daten, sowie der Sicherstellung der Datenqualität. Interessant dabei sind nicht nur die Methoden, sondern auch der konkrete wissenschaftliche Anwendungskontext.

Dieses Themenheft enthält drei Beiträge, die einen Eindruck von der Vielfalt der Forschung im Bereich des wissenschaftlichen Datenmanagements vermitteln. In dem ersten Beitrag wird das Problem des Auffindens und Integrierens von verteilt vorliegenden, heterogenen Daten in den datenintensiven Wissenschaften betrachtet. Semantische Techniken versprechen hier Unterstützung. Entsprechend wurde beispielsweise in der Biologie und angrenzenden Gebieten in den letzten Jahren großer Aufwand in die Entwicklung von mehr oder weniger stark formalisierten Domänenmodellen gesteckt. Im Ergebnis sind zahlreiche Terminologien und Ontologien entstanden, deren Nutzung jedoch dadurch erschwert wird, dass sie verteilt vorliegen, unterschiedliche (jeweils disziplinspezifische) Sichten auf die Diskursdomäne bieten und über unterschied-

Bernhard Seeger
seeger@mathematik.uni-marburg.de

Birgitta König-Ries
Birgitta.Koenig-Ries@uni-jena.de

✉ Theo Härder
haerder@cs.uni-kl.de

¹ Philipps-Universität Marburg,
Hans-Meerwein-Str. 6, 35032 Marburg,
Deutschland

² Friedrich-Schiller-Universität Jena,
Ernst-Abbe-Platz 2, 07743 Jena, Deutschland

³ AG Datenbanken und Informationssysteme, TU
Kaiserslautern, 67663 Kaiserslautern, Deutschland

liche Schnittstellen und Abfragemechanismen zugreifbar sind. Im Beitrag *A Terminology Service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data* beschreiben Naouel Karam und ihre Koautoren (FU Berlin) eine Plattform, die hier deutliche Verbesserung verspricht: Der im Artikel beschriebene Terminologieservice bietet einheitliche Schnittstellen zum Zugriff auf eine große Zahl unterschiedlicher Vokabularien. Diese werden teils extern eingebunden, teils direkt im System gehalten und werden nach Benutzeranforderungen ausgewählt. Mit den angebotenen Diensten ist es möglich, Use Cases wie etwa die semantische Erweiterung von Anfragen oder die Personalisierung von Suchergebnissen zu unterstützen. Der Dienst wird im Rahmen eines größeren Projektes, GFBio (German Federation for Biological Data) entwickelt und evaluiert.

Der zweite Beitrag *Supporting Situation Awareness in Spatio-Temporal Databases* von Andreas Behrend, Philip Schmiegel und Andreas Dohr (Uni Bonn) greift ein Thema zur Wissensextraktion und Wissensrepräsentation in Zeitreihen auf. Häufig sind die beobachteten Messergebnisse einer Zeitreihe für den Benutzer nur im Kontext anderer Zeitreihen interessant und können dann entsprechend interpretiert werden. Eine Situation beschreibt dabei ein für den Anwender relevantes Konzept, für das gewisse Eigenschaften der Zeitreihen in einem Zeitintervall erfüllt sein müssen. Diese Situationen werden dann genutzt, um z. B. Anomalien zu erkennen und entsprechend mit Aktionen darauf zu reagieren. Hierfür schlagen die Autoren einen Transitionsgraph vor, in dem die erlaubten und nicht erlaubten Übergänge von Situationen repräsentiert sind. Aus Datenbanksicht sind insbesondere zwei Ansätze von Interesse. Zum einen stellen die Autoren ein System vor, das eine flexible Definition von Situationen unter Verwendung von Datenbanken-Views und deren Optimierung mit Magic Sets unterstützt. Zum anderen erweitern die Autoren bestehende Datenbanken durch Tabellenfunktionen, um die temporale Logik auf Intervallmengen geeignet abzubilden. Zudem geben die Autoren einen Überblick von herausfordernden Problemen bei der Indexierung und Verwaltung von Situationen.

Im dritten Beitrag *Toward a Visual Analytics Approach to Support Multi-Sensor Analysis in Remote Sensing Science* von Mike Sips, Patrick Köthür und Daniel Eggert (Deutsches Geoforschungszentrum Potsdam und Humboldt-Universität zu Berlin) wird konkret auf eine Anwendung im Bereich der Fernerkundung eingegangen, bei der die zeitliche Entwicklung der in Europa von Wald bedeckten Flächen verfolgt und analysiert wird. Dabei werden Fernerkundungsbilder verschiedener Satelliten mit unterschiedlicher räumlicher und zeitlicher Auflösung miteinander kombiniert. Zudem sollen Muster in Zeitreihen identifiziert wer-

den, die den zeitlichen Verlauf des Zustands einzelner Pixels repräsentieren. Um eine hohe Datenqualität zu erreichen, fokussieren sich die Autoren im vorliegenden Beitrag auf die Konzipierung einer auf Metadaten gestützten Auswahl von Sensordaten. Die Arbeit ist Teil des Berliner Projekts GeoMultiSens, in dem *Big Data*-Technologien zur Verarbeitung von Fernerkundungsdaten in angepasster Form verwendet werden. Interessant in diesem Artikel ist insbesondere die Vorgehensweise, wie eine konkrete visuelle Analyseanwendung auf Grundlage von Nutzerrückmeldung erstellt wurde. Um auf veränderte Anforderungen von Nutzern reagieren zu können, wird die Infrastruktur Scientific Data Explorer entwickelt, mit der ähnlich zu einer modellbasierten Softwareentwicklung mit Regeln komplette visuelle Analyseanwendungen generiert werden können.

1.1 Community-Beiträge in diesem Heft

Die wissenschaftlichen Beiträge werden durch den Fachbeitrag *Speeding up Privacy Preserving Record Linkage for Metric Space Similarity Measures* von Zihad Sehili und Erhard Rahm (Uni Leipzig) ergänzt. Bei der Analyse von personenbezogenen Daten ist es schwierig, nützliche Ergebnisse zu finden und trotzdem einen hohen Grad an Privatsphäre zu bewahren – insbesondere auch dann, wenn personenbezogene Daten von verschiedenen Quellen integriert und analysiert werden müssen. Privacy Preserving Record Linkage (PPRL) widmet sich diesem Problem durch die Kodierung sensibler Attributwerte, so dass die Identifikation von Personen verhindert wird, die Fragestellung der Analyse aber trotzdem bearbeitet werden kann. Die im Beitrag beschriebenen Methoden zielen auf eine Verbesserung von Effizienz und Skalierbarkeit bei PPRL-Anwendungen ab.

Die Rubrik „Datenbankgruppen vorgestellt“ enthält den Beitrag *Die Arbeitsgruppen für Datenbanken und Informationssysteme an der TU Kaiserslautern* von Stefan Deßloch, Theo Härder und Sebastian Michel. In diesem Artikel werden die Ergebnisse der wichtigsten Forschungsprojekte, die in den letzten zehn Jahren abgeschlossen wurden, skizziert und die aktuellen Forschungsthemen vorgestellt. Außerdem werden die Datenbank-bezogenen Lehraufgaben für das Bachelor- und Master-Studium, die im Lehrgebiet Informationssysteme angeboten werden, beschrieben.

In der Rubrik „Kurz erklärt“ erscheint der Artikel *Skyline Queries*, in dem Katja Hose (Aalborg University, Denmark) eine kurze Einführung in die Problemstellung gibt. Sie erklärt die wichtigsten Charakteristika von Skyline Queries und diskutiert die grundsätzlichen Möglichkeiten ihrer Verarbeitung durch den Skyline Operator. Die Probleme der multi-kriteriellen Entscheidungsfindung, die hierbei zu lösen sind, sind auch unter den Namen „Maximum Vector Problem“ oder „Pareto Optimum“ bekannt.

Die Rubrik „Dissertationen“ enthält in diesem Heft 2 Kurzfassungen von Dissertationen aus der deutschsprachigen DBIS-Community.

Schließlich berichtet diese Rubrik „Community“ unter *News* über Ehrungen für Gerhard Weikum, über den 27. GI-Workshop „Grundlagen von Datenbanken“ (GvDB 2016) und weitere aktuelle Informationen, welche die DBIS-Gemeinde betreffen.

Ein Nachruf auf Hans-Jürgen Appelrath, der am 5. August 2016 verstorben ist, erinnert an seine großartigen Leistungen für das OFFIS – Institut für Informatik, die Oldenburger Universität und die deutsche DBIS-Community.

2 Künftige Schwerpunktthemen

2.1 Innovative vertikale Suchlösungen: Konzeption, Umsetzung und Einsatz

In der öffentlichen Wahrnehmung dominieren die großen Web-Suchmaschinen das Bild. Dass es daneben eine Vielfalt spezieller Suchlösungen für ein breites Spektrum unterschiedlichster Nutzungsszenarien gibt, gerät fast in den Hintergrund, obwohl auch diese Lösungen häufig genutzt werden.

Der Begriff der vertikalen Suchlösungen soll dabei breit interpretiert werden. Spezielle Zielgruppen, spezielle Domänen, spezielle Suchgegenstände umreißen nur einige Beispiele. Auch Aspekte des Web Mining, der Unternehmenssuche oder der Shopsuche sind relevant. Gleiches gilt für die Suche nach anderen Medientypen (Bild, Video, ...). Dabei soll es aber nicht um eine Darstellung neuer Verfahren zur Bildanalyse etc. gehen. Vielmehr steht die aus der Anwendungssituation motivierte, zielgerichtet entwickelte oder angepasste Suchlösung im Vordergrund. Die behandelten Systeme sollen im praktischen Einsatz sein und im Beitrag zum Schwerpunktthema als Gesamtsystem betrachtet werden. In Summe möchte das Heft damit einen Überblick über die Vielfalt vertikaler Suchlösungen geben und darstellen, wie Suchlösungen für konkrete Aufgabenstellungen konzipiert und umgesetzt werden können.

Dabei sollen zu jeder Suchlösung insbesondere auch die folgenden Aspekte adressiert werden:

- Beschreibung der konkreten Problemstellung
- Konzeption der umgesetzten Suchlösung
- Unterschiede zu anderen Suchlösungen in derselben Domäne
- Erfahrungen mit der Suchlösung/Evaluation

Mögliche Themen für dieses Heft sind damit beispielsweise:

- Suchmaschinen für spezielle Zielgruppen (Kinder, Senioren, ...)
- Suchmaschinen für spezielle Domänen (Kunst, Medizin, ...)
- Hybridsuchmaschinen, die Web-Inhalte mit Datenbankinhalten kombinieren
- Produktsuche, Shopsuche, Expertensuche, ...

Die Suchsysteme können dabei natürlich auf bestehende Frameworks zurückgreifen, die sie in der konkreten Aufgabenstellung innovativ einsetzen.

Beitragsformat: 8–10 Seiten, zweispaltig, Beiträge sind in Deutsch oder Englisch willkommen.

Ankündigung einer Beitragseinreichung bis zum 15. August 2016

Einreichung der Beiträge für DASP-1-2017 bis zum 1. Oktober 2016

Erscheinen des Themenheftes: März 2017

Gastherausgeber:

Andreas Henrich, Otto-Friedrich-Universität Bamberg
andreas.henrich@uni-bamberg.de

Dirk Lewandowski, Hochschule für Angewandte Wissenschaften Hamburg
dirk.lewandowski@haw-hamburg.de

2.2 Big Graph Data Management

A graph is an intuitive mathematical abstraction to capture how things are connected. In the past decade, the focal point in many data management applications has shifted from individual entities and aggregations thereof toward the connection between entities. Hence today, the graph abstraction is appealing as a natural data model foundation for an increasing range of use cases in interactive as well as analytical graph data management scenarios. Graph-specific use cases can be found in various domains, such as social network analysis, product recommendations, and knowledge graphs. Graph-oriented scenarios also emerge in more traditional enterprise scenarios, such as supply chain management or business process analysis. Therefore, the database community reacts to this newly sparked interest in graph data management with a vast number of projects in research as well as in industry.

Graph management use cases pose novel and unique challenges to data management systems. On the operation-

nal side, typical interactive queries involve transitive closure computation along paths. Common analytical measures, such as page rank and other vertex centrality measures are also significantly more complex than traditional group-by/aggregate queries. From a data structure perspective, the irregular and skewed structure of graphs makes it challenging to achieve a good distribution over non-uniform memory access or cluster nodes for efficient parallelization – particularly, if the graph is large and changing over time. Further challenges among others are declarative graph analytics abstractions for static as well as for dynamic graphs, graph-query-aware optimization strategies, topology indexing, temporal topology indexing, topology estimation, materialized view usage, and maintenance for graph analytical measures.

Graph data management is an exciting research field, now and for the years to come. This special issue aims at exhibiting our community’s current work in the field. We therefore welcome contributions from research and industry that provide original research on the problems mentioned above or that are generally related to big graph data management and processing. We also welcome case studies that showcase the challenges of graph management and graph query processing from a practical perspective, point out particular research questions, and potentially outline novel research directions.

We are looking for contributions from researchers and practitioners in the above described context, which may be submitted in German or in English.

Important dates:

- Notice of intent for a contribution: December 15th, 2016
- Deadline for submissions: February 1st, 2017
- Issue delivery: DASP-2-2017 (July 2017)

Paper format: 8–10 pages, double column (cf. the author guidelines at www.datenbank-spektrum.de).

Guest editors:

Hannes Voigt, TU Dresden
hannes.voigt@tu-dresden.de
Marcus Paradies, SAP
m.paradies@sap.com

2.3 Best Workshop Papers of BTW 2017

This special issue of the “Datenbank-Spektrum” is dedicated to the Best Papers of the Workshops running at the BTW 2017 at the University of Stuttgart. The selected Workshop contributions should be extended to match the format of regular DASP papers.

Paper format: 8–10 pages, double column

Selection of the Best Papers by the Workshop chairs and the guest editor: April 15th, 2017

Guest editor:

Theo Härder, University of Kaiserslautern
haerder@cs.uni-kl.de

Deadline for submissions: June 1st, 2017

Issue delivery: DASP-3-2017 (November 2017)