Editorial

Matthias Hagen · Benno Stein · Theo Härder

Online publiziert: 5. Februar 2016
© Springer-Verlag Berlin Heidelberg 2016

Schwerpunktthema: Big Data & IR

Unter dem Begriff „Big Data“ werden Daten und Strategien zur Datenverarbeitung gefasst, die einerseits aufgrund ihres Umfangs die Verarbeitungs- und Speicherkapazität eines mittleren bis großen Rechnerclusters voraussetzen und die andererseits aus der immer weiter in den Alltag vordringenden Computertechnik und Sensorik gewonnen werden. Für die Wissenschaft interessant sind deshalb Fragen der Organisation und des Managements von Big Data – aber auch Forschungsfragen, die erst mit Big Data beantwortbar werden, bzw. Problemstellungen, die sich durch die Analyse und das Sammeln von Big Data besser verstehen lassen. Im Zusammenhang mit Information Retrieval als der zielgerichteten Suche nach relevanten Inhalten ergeben sich damit zwei Arten von Herausforderungen: 1) Retrieval *in* Big Data und 2) Verbesserung von Retrieval *durch* Big Data.

Retrieval in Big Data befasst sich zum einen mit der Organisation, dem Management und dem schnellen Zugriff auf Big Data, adressiert aber auch den kreativen Prozess der Suche nach interessanten Forschungsfragen, die sich erst mit Big Data verstehen und beantworten lassen. Neben der Entwicklung leistungsfähiger Frameworks zur Verwaltung von Text-, Multimedia-, Sensor-, Simulations- oder sonstiger Da-

ten geht es folglich auch um die Frage, welche Einblicke durch Big Data heute und in der Zukunft möglich werden.

Die zweite Richtung im Umfeld von Big Data & IR befasst sich mit der Verbesserung von Retrievalverfahren durch den Einsatz von Big Data. Dazu gehören beispielsweise die schon fast klassische Frage der Verbesserung von Websuche oder eCommerce durch maschinelles Lernen auf Daten zum Nutzerverhalten, das Einbeziehen des Nutzerkontexts beim Retrieval oder auch das Ausschöpfen semantischer Datenquellen wie Linked Open Data oder Wissensgraphen.

Dieses Themenheft bietet eine Auswahl von Beiträgen zu den oben genannten Themen:

Der erste Beitrag von Johannes Schildgen und Stefan Deßloch (TU Kaiserslautern) beschreibt die Datentransformationssprache NotaQL als eine mögliche Herangehensweise, die Heterogenität unterschiedlicher Datenquellen aufzulösen. Ziel ist der flexible Zugriff auf Daten und Metadaten sowie auch die Integration schemaloser NoSQL-Systeme.

Ivan Giangreco und Heiko Schuldt (Universität Basel) stellen im zweiten Beitrag mit ADAMpro ein Datenbank- und Retrievalsystem für große Multimedia-Datensätze vor. Speziell die Suche nach ähnlichen Inhalten wird durch geeignete Datenstrukturen und Ergebnispräsentationen unterstützt.

Im dritten Beitrag werden von Fabian Reinartz, Koninika Pal und Sebastian Michel (TU Kaiserslautern) automatische Verfahren zur Erzeugung von Entity-Rankings aus gegebenen Datensätzen beschrieben. Die adressierte Fragestellung ist die automatische Entscheidung, welche Kriterien innerhalb eines Datensatzes für interessante Rankings als besonders vielversprechend erscheinen.

Mit HySpirit behandelt der vierte Beitrag ein System für probabilistisches Datalog (PDatalog). Dabei legen die Autoren Ingo Frommholz (University of Bedfordshire, UK)

M. Hagen (✉) · B. Stein
Faculty of Media, Bauhaus-Universität Weimar, Bauhausstraße 11,
99423 Weimar, Deutschland
E-Mail: matthias.hagen@uni-weimar.de

B. Stein
E-Mail: benno.stein@uni-weimar.de

T. Härder
AG Datenbanken und Informationssysteme, TU Kaiserslautern,
67663 Kaiserslautern, Deutschland
E-Mail: haerder@cs.uni-kl.de

und Thomas Roelleke (Queen Mary University of London, UK) den Fokus auf Wahrscheinlichkeitsabschätzungen in PDataLog, Fuzzy-Prädikaten und verteilter Verarbeitung. In experimentellen Beispielen wird die reale Anwendbarkeit von PDataLog in HySpirit aufgezeigt.

Christoph Koch (Friedrich-Schiller-Universität Jena und DATEV eG Nürnberg) widmet sich im fünften Beitrag der Generierung synthetischer Datenbankstatistiken, die für die Abschätzung der Performance von SQL-Statements benötigt werden. Im Beitrag werden dazu Explain-Mechanismen durch künstlich erzeugte Statistiken von Performance-Indikatoren vorgeschlagen.

Im sechsten Beitrag wird die Wichtigkeit von Zusammenfassungen für die Suche und Visualisierung im Kontext von Big Data thematisiert. Daniel Blank, Andreas Henrich und Stefan Kufer (Universität Bamberg) beschreiben am Beispiel von räumlichen Daten und allgemeinen metrischen Räumen die Nützlichkeit von Zusammenfassungen in verteilten Szenarien für die Ressourcenauswahl und -visualisierung.

Community-Beiträge in diesem Heft

Die Rubrik „Datenbankgruppen vorgestellt“ enthält den Beitrag *Die Abteilung Datenbanken und Informationssysteme am Max-Planck-Institut für Informatik*, in dem Gerhard Weikum zunächst die seit 2005 verfolgte wissenschaftliche Vision der Abteilung skizziert, bevor er die aktuellen Forschungsrichtungen in vier großen Themenfeldern beschreibt. Außerdem gibt er einen Hinweis, welche der in der Abteilung entwickelte Open-Source Software öffentlich verfügbar ist und welche Datenressourcen für die Forschung verwendet werden können.

In der Rubrik „Kurz erklärt“ erscheint in diesen Heft ein Beitrag von Uta Störl (Hochschule Darmstadt), Meike Klette (Universität Rostock) und Stefanie Scherzinger (OTH Regensburg) zum aktuellen Stichwort *Objekt-NoSQL-Mapping*. Objekt-Relationale-Mapper (OR-Mapper) werden schon seit vielen Jahren bei der Persistierung von Objekten in relationalen Datenbanken eingesetzt. Mittlerweile werden aber auch für NoSQL-Datenbanken dedizierte Systeme, sogenannte Objekt-NoSQL-Mapper (ON-Mapper), angeboten, die teilweise als Erweiterung von existierenden OR-Mappern, teilweise aber auch als eigenständige Produkte entwickelt wurden.

Erfreulich umfangreich ist in diesem Heft die Rubrik „Dissertationen“, in der 13 Kurzfassungen von Dissertationen aus der deutschsprachigen DBIS-Community vorgestellt werden.

Die Rubrik „Community“ enthält schließlich unter *News* weitere aktuelle Informationen – beispielsweise Ergebnisse zur Wahl der DFG-Fachkollegiaten –, welche die DBIS-Gemeinde betreffen.

Künftige Schwerpunktthemen

1 Schutz der Privatsphäre in einer ubiquitären Welt

Mit immer mehr mobilen Geräten und Sensoren werden u. a. große Mengen an persönlichen Daten gesammelt, verarbeitet und transformiert. Solche Sammlungen personenbezogener Daten sind auf der einen Seite notwendig, um personenspezifische Angebote machen zu können, die dem Empfänger örtlich und zeitlich von Nutzen sind, oder um Trends zu erkennen und somit Planungen in unterschiedlichen Bereichen genauer und effizienter ausführen zu können. Auf der anderen Seite dienen sie häufig dazu, individuelle Personenprofile zu erstellen, die zum Vorteil oder Nachteil der beschriebenen Person genutzt werden können.

Aus den genannten Gründen wird es immer wichtiger, den Datenschutz in einer ubiquitären Welt im Kontext von Big Data nicht nur juristisch abzusichern (Bundesdatenschutzgesetz). Vielmehr wird es immer dringlicher, auch technische Möglichkeiten, Mechanismen und Ansätze zu entwerfen und zu realisieren, die es Personen ermöglichen, die Kontrolle über ihre Daten besser spezifizieren sowie ihre Nutzung und Weitergabe besser kontrollieren und nachvollziehen zu können. Trotz großer Fortschritte im Bereich des Schutzes der Privatsphäre durch unterschiedliche Techniken besteht weiterhin eine große Herausforderung darin, skalierbare Ansätze und Lösungen sowohl für die Nutzung personenbezogener Daten durch Dritte als auch deren Kontrolle durch den „Spender“ zu entwickeln und zu realisieren.

Somit ist es das Ziel des Themenheftes, neben einer Einführung in das Thema skalierbare Ansätze und Lösungen für das Sammeln, Verarbeitung und Analysieren personenbezogener Daten in unterschiedlichen Anwendungsdomänen zu beschreiben. Mögliche Themen für dieses Themenheft sind (nicht ausschließlich):

- Schutz der Privatsphäre im Bereich Big Data generell beim Sammeln, Integrieren, Verarbeiten und Analysieren von Daten
- Technische Umsetzung juristischer (gesetzlicher) Vorgaben entsprechend des deutschen Rechts bzw. des EU-Rechts und Vorgaben in unterschiedlichen Bereichen wie beispielsweise dem Gesundheitsbereich (Arzt, Krankenhaus, Versicherer) oder dem Finanzbereich (Banken)
- Sprachen zur Beschreibung von Privacy-Präferenzen und deren Überprüfung (in skalierbarer Form)
- Datenaustausch unter Berücksichtigung von Privacy-Präferenzen und gesetzlichen Vorgaben
- Anfragebearbeitung in Datenbanksystemen unter Berücksichtigung von Privacy-Präferenzen
- Quantitative Bewertung von Ansätzen zum Schutz der Privatsphäre im Kontext der Nutzung geschützter Daten

- Anforderungsanalysen für den Schutz der Privatsphäre in Anwendungsdomänen – Schutz gegen Genauigkeit der Daten
- Anforderungen des Schutzes der Privatsphäre für räumliche und zeitbezogene Daten in verschiedenen Anwendungsbereichen
- Infrastrukturen zum Schutz der Privatsphäre
- Modelle zum Schutz der Privatsphäre bei Zugriff oder Datennutzung

Gastherausgeber:

Johann-Christoph Freytag, HU Berlin

freytag@informatik.hu-berlin.de

Eric Buchmann, Karlsruher Institut für Technologie

eric.buchmann@kit.edu

2 Data management for bio- and geosciences

Like many other scientific disciplines, research in the bio- and geosciences follow more and more a data-driven approach. Big Data in the classical sense is only one of the issues, but probably more often the everyday problem of scientists is to cope with lots of „small“, heterogeneous data that needs to be integrated to answer complex questions.

This special issue addresses the arising challenges and solutions for data management in these areas. We are interested in both survey papers and papers describing original research dealing with the following or similar topics in the context of the bio- and geosciences:

- Data-intensive science
- Data management
- Data integration
- Spatio-temporal data processing
- Scientific workflows
- Semantic web technologies
- Visualization and visual analytics
- Data stream management
- Case studies and applications

Important dates:

- Notice of intent for a contribution: April 15th, 2016
- Deadline for submissions: June 1st, 2016
- Issue delivery: DASP-3-2016 (November 2016)

Paper format: 8–10 pages, double column

Guest editors:

Bernhard Seeger, Philipps-Universität Marburg

seeger@mathematik.uni-marburg.de

Birgitta König-Ries, Friedrich-Schiller-Universität Jena

Brigitta.Koenig-Ries@uni-jena.de

3 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 Schwerpunkt als Gesamtsystem betrachtet werden. In Summe möchte das Heft damit einen Überblick über die Vielfalt vertikaler Suchlösungen geben und darlegen, 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

4 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 operational 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