Datenbank Spektrum (2012) 12:77–79
DOI 10.1007/s13222-012-0094-6

EDITORIAL

Editorial

Wolf-Tilo Balke

Online publiziert: 23. Mai 2012

1 Schwerpunktthema: Information Extraction

Strukturiertes Wissen aus unstrukturierten oder bestenfalls semi-strukturierten Daten abzuleiten, ist eine der zentralen Herausforderungen der heutigen Wissensgesellschaft. Die Extraktion von Fakten in strukturierter Form aus der Vielfalt frei zugänglicher Information zum Beispiel im World Wide Web beschäftigt die Datenbank-Community deshalb schon eine ganze Weile. Angefangen bei den überschaubaren, meist manuell erstellten Wissensbasen für Expertensysteme in den 80-igern über die massive Indexierung von Dokumenten im Web für die Nutzung durch Suchmaschinen haben die letzten 10 Jahre ungeahnte Durchbrüche im Bereich der intelligenten Algorithmen zur Generierung von Wissen hervorgebracht. Eine Kombination von Technologien für Natural Language Processing (NLP) und Machine-Learning-Algorithmen sowie einfache Inferenzregeln und der Abgleich mit ontologischer Information führen dabei zu relativ hochqualitativen Datenbanken, welche oft mehrere Millionen Fakten für eine Vielzahl von möglichen intelligenten Anwendungen bereitstellen. Aber auch breite Themenfelder wie Business Intelligence oder Content Management, die heutige Unternehmen in zunehmendem Maße durchziehen, wären ohne Information-Extraction-Techniken heute kaum zu beherrschen.

Allerdings sind die Probleme der Informationsextrahierung noch nicht vollständig gelöst. Während die reine Erkennung von Entitäten schon recht erfolgreich automatisiert werden kann, stellt die Erkennung sinnvoller und semantisch aussagekräftiger Relationen zwischen Entitäten noch immer ein großes Problem dar. Zudem zeigt sich, dass eine direkte Einbeziehung menschlicher Kreativität oft schwierige Extraktionsprobleme lösen kann. Der heutigen Stand der Technik und zukünftige Herausforderungen werden deshalb in einem kurzen, einleitenden Übersichtsartikel *Introduction to Information Extraction: Basic Notions and Current Trends* zusammengefasst.

Der erste Beitrag *Fact-Aware Document Retrieval for Information Extraction* beschreibt die Architektur des BlueFact-Systems, bei dem Dokumente, welche strukturierte Information in direkt extrahierbarer Form enthalten, mit entsprechenden Metadaten versehen werden sollen. Idee ist es, damit Ad-hoc-Anfragen zu ermöglichen, bei denen in strukturierter Form Information angefragt werden kann, die dann direkt zum Anfragezeitpunkt aus den Dokumenten mit den vielversprechendsten Metadaten extrahiert wird.

Ein System, das Extraktionsmechanismen zusammen mit deklarativen Anfragemöglichkeiten auf heterogene Datenquellen im Web anwendet, präsentiert der zweite Beitrag *Sequoia – An Approach to Declarative Information Retrieval*. Bisher konnte unstrukturierte Information, z. B. über Suchmaschinen, in der Regel nur mit Schlagworten abgefragt werden. Sequoia erlaubt es, zusätzliche strukturelle Information aus verschiedenen Quellen zu extrahieren, welche es ermöglicht, einfache Operatoren wie z. B. Joins über diesen Quellen auszuführen. Eine Anwendung ist beispielsweise die dynamische Einbindung von Hintergrundinformation aus Nachrichtensammlungen in Twitter Live Streams.

Die direkte Einbindung menschlicher Intelligenz in die Informationsextraktion ist zentrales Thema des dritten Beitrags *Information Extraction meets Crowdsourcing: A Promising Couple*. Hierbei stehen speziell der Aspekt der Datenqualität und Extraktionskosten im Mittelpunkt. Eine eingehende Betrachtung und Klassifizierung verschiedener Nutzungsszenarien zeigt dabei, dass mit Hilfe hybrider Al-

W.-T. Balke (✉)
Institut für Informationssysteme, TU Braunschweig,
Braunschweig, Deutschland
e-mail: balke@ifis.cs.tu-bs.de

gorithmen, die Machine Learning mit direkter menschlicher Interaktion verbinden, eine deutliche Effizienz- und Qualitätssteigerung realisiert werden kann.

Der vierte Beitrag *OPEN – Enabling Non-Expert Users to Extract, Integrate, and Analyze Open Data* beschäftigt sich damit, die Integration und Analyse sogenannter Offener Daten (Open Data) möglichst endnutzerfreundlich zu gestalten. Dazu wird das sogenannte DrillBeyond-Konzept entwickelt, welches Anfragen an eine Datenbank zu stellen erlaubt, die nicht durch das gegebene Schema bzw. die vorhandenen Daten abgedeckt sind, sondern stattdessen automatisch auf offen zugängliche Datensätze im Web abgebildet werden.

Diese Schwerpunktbeiträge werden ergänzt durch einen Fachbeitrag *Verfahren zur funktionalen Ähnlichkeitssuche technischer Bauteile in 3D-Datenbanken* von Moritz Maier, Jan Schulz und Klaus-Dieter Thoben. Weiterhin finden Sie unter der Rubrik "Datenbankgruppen vorgestellt" einen Beitrag von Wolfgang Lehner zu *Die Datenbankforschungsgruppe der Technischen Universität Dresden stellt sich vor*. Die Rubrik "Dissertationen" ist wiederum recht umfangreich; sie enthält in diesem Heft sieben Kurzfassungen von Dissertationen. Weiterhin erscheinen in der Rubrik "Community" aktuelle Berichte und insbesondere Informationen zur BTW 2013 in Magdeburg.

2 Künftige Schwerpunktthemen

2.1 Scientific Data Management

The past decade has witnessed a dramatic increase in scientific data being generated in the physical, earth, and life sciences. This development is primarily a result of major advancements in sensor technology, surveying techniques, computer-based simulations, and instrumentation of experiments. In a special issue of the "Datenbank-Spektrum", we want to publish original work on different aspects related to the management and analysis of scientific data. The objective of this special issue is to exchange ideas between academia and industry and to discuss recent developments, challenges, and future directions in scientific data management.

Topics of interest include (but are not limited to)

- Modeling and representation of data, metadata, ontologies, and processes for scientific application domains
- Integration and exchange of scientific data
- Design, implementation, and optimization of scientific workflows
- Architectures and components for scientific computing and eScience, including Web portals, repositories, and digital libraries
- Annotation and provenance of scientific data
- Mining and analysis of large-scale scientific datasets
- Case studies and applications related to scientific data management in all domains, with a particular focus on biology, physics, chemistry, medicine, and geography

Guest editors:
Michael Gertz, Heidelberg University,
gertz@informatik.uni-heidelberg.de
Wolfgang Müller, HITS gGmbH,
wolfgang.mueller@h-its.org

2.2 MapReduce Programming Model

MapReduce (MR) is a programming model which facilitates parallel processing of large, distributed, and even heterogeneous data sets. To accelerate the development of specific MR applications, an MR implementation provides a framework dealing with data distribution and scheduling of parallel tasks. The user only has to complement this framework by specifying a *map* function—processing key/value pairs to generate intermediate key/value pairs—and a *reduce* function which groups all records with the same intermediate key and merges all values of such groups.

Using this approach, programs written in such a functional style can automatically exploit large degrees of parallelism and thereby perfectly scale. As a consequence, the MR model had tremendous success in recent years covering many areas of *Big Data* processing. For this reason, the "Datenbank-Spektrum" wants to publish research contributions—especially of the German database community—providing an overview over ongoing work in this particular area.

Submissions covering topics from the following non-exclusive list are encouraged:

- Applications of the MR paradigm
- Optimization of the MR framework and its applications
- MR-conform compilation of DB languages
- Schema flexibility (key/value stores) and MapReduce
- Comparison of applications running under MapReduce/Hadoop and parallel DBMSs
- Cooperation of NoSQL and SQL when processing XXL data

Paper format: 8–10 pages, double column
Notice of intent for a contribution: July 15th, 2012

Guest editor:
Theo Härder, University of Kaiserslautern,
haerder@cs.uni-kl.de
Deadline for submissions: October 1st, 2012

2.3 RDF Data Management

Nowadays, more and more data is modeled and managed by means of the W3C Resource Description Framework (RDF)

and queried by the W3C SPARQL Protocol and RDF Query Language (SPARQL). RDF is commonly known as a conceptual data model for structured information that was standardized to become a key enabler of the Semantic Web to express metadata on the Web. It supports relationships between resources as first-class citizens, provides modeling flexibility towards any kind of schema, and is even usable without a schema at all. Furthermore, RDF allows to collect data starting with very little schema information and refining the schema later, as required. This flexibility led to a wide adoption in many other application domains including life sciences, multifaceted data integration, as well as community-based data collection, and large knowledge bases like DBpedia.

This special issue of the "Datenbank-Spektrum" aims to provide an overview of recent developments, challenges, and future directions in the field of RDF technologies and applications.

Topics of interest include (but are not limited to)

- RDF data management
- RDF access over the Web
- Querying and query optimization over RDF data—especialy when accessed over the Web
- Applications and usage scenarios
- Case studies and experience reports

Paper format: 8–10 pages, double column
Notice of intent for a contribution: Nov. 15th, 2012

Guest editors:
Johann-Christoph Freytag, Humboldt-Universität zu Berlin,
freytag@dbis.informatik.hu-berlin.de
Bernhard Mitschang, University of Stuttgart,
Bernhard.Mitschang@ipvs.uni-stuttgart.de

Deadline for submissions: February 1st, 2013