Datenbank Spektrum (2019) 19:1–3
https://doi.org/10.1007/s13222-019-00310-1

EDITORIAL

Editorial

Erhard Rahm[1] · **Theo Härder**[2]

1 Schwerpunktthema: Scalable and intelligent Data Services and Solutions – Results from the Big Data Competence Center ScaDS Dresden/Leipzig

Das Schwerpunktthema dieser Ausgabe widmet sich ausgewählten Forschungsergebnissen des seit 2014 bestehenden BMBF-geförderten Big-Data-Kompetenzzentrums ScaDS (Competence Center for *Sca*lable *D*ata Services and *S*olutions) Dresden/Leipzig. Hierzu beinhaltet das Themenheft einen einleitenden Überblicksartikel sowie vier Artikel zu spezielleren Ergebnissen.

Im ersten Beitrag *Big Data Competence Center ScaDS Dresden/Leipzig: Overview and selected research activities* erläutern die neun Autoren von der Universität Leipzig und der TU Dresden die wesentlichen Forschungs- und Anwendungsgebiete von ScaDS und präsentieren zudem bereits einzelne Ergebnisse zu Themen, die in den nachfolgenden Artikeln nicht im Fokus stehen, u. a. zur Anreicherung sowie der holistischen Integration von Daten. Generell wird im ScaDS Dresden/Leipzig ein breites Themenspektrum in der Big-Data-Forschung und deren Anwendungsgebieten adressiert, wobei mehrere prototypische Systemlösungen entstanden, u. a. zur auf große Datenmengen skalierbaren Datenintegration sowie zur Analyse von Graphdaten (Gradoop-System).

Der nachfolgende Artikel *Large-Scale Time Series Analytics – Novel Approaches for Generation and Prediction* der Autoren M. Hahmann, C. Hartmann, L. Kegel und W. Lehner von der TU Dresden widmet sich der Datenanalyse großer Zeitreihen. Ein Schwerpunkt ist dabei die Gewinnung von Zeitreihen aus bestehenden Daten, z. B. um ein System gezielt evaluieren zu können. Neben der Vorstellung bekannter, meist domänen-spezifischer Ansätze für eine derartige Generierung von Zeitreihen, wird ein domänen-übergreifend nutzbarer Ansatz vorgeschlagen, der auf statistischen Eigenschaften der vorliegenden Daten basiert. Der zweite Teil des Aufsatzes widmet sich der Vorhersage künftiger Zeitreihenereignisse. Hierzu wird der sogenannte CSAR-Ansatz vorgestellt, der auch bei unvollständigen bzw. fehlerhaften Eingabedaten anwendbar ist.

Der Beitrag *ScaDS Research on Scalable Privacy-preserving Record Linkage* der Leipziger Autoren M. Franke, M. Gladbach, Z. Sehili, F. Rohde und E. Rahm stellt Forschungsergebnisse zur Privacy-bewahrenden Datenintegration vor, z. B. um patientenbezogene Informationen aus verschiedenen Datenquellen unter Wahrung des Datenschutzes für eine verbesserte Datenanalyse zu kombinieren. Hierzu erfolgt eine Kodierung personenidentifizierender Attributwerte wie Name und Geburtsdatum durch Bitvektoren, welche durch eine vertrauenswürdige Instanz, einer sogenannten Linkage Unit, für das Erkennen übereinstimmender Personen abgeglichen werden. Die Autoren stellen insbesondere neue Ansätze zur Skalierung dieser Linkage-Ansätze vor, bei denen ein paralleles Linkage auf einem Hadoop-Cluster mit Apache Flink erfolgt und die Anzahl der Vergleiche von Bitvektoren durch Blocking- und Filter-Ansätze stark reduziert wird. Als besonders effektiv stellt sich dabei ein Blocking auf Basis von Locality-sensitive Hashing (LSH) heraus.

Der Big-Data-Anwendungsbereich Digital Humanities steht im Fokus des Artikels *A Big Data Case Study in Digital Humanities: Creating a Performance Benchmark for Canonical Text Services* von G. Heyer und J. Tiepmar von der Universität Leipzig. Die Autoren erläutern die Realisierung des sogenannten CTS-Dienstes (Canonical Text Services), mit dem bestimmte Textbereiche in hierarchisch strukturierten Dokumenten mit einer permanenten Referenz versehen werden können, um damit genaue Annotationen und Querverweise zu ermöglichen. Zudem wird die Leistungsfähigkeit der Implementierung analysiert.

Erhard Rahm
rahm@informatik.uni-leipzig.de

✉ Theo Härder
haerder@cs.uni-kl.de

[1] Fakultät für Mathematik und Informatik, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Deutschland

[2] AG Datenbanken und Informationssysteme, TU Kaiserslautern, 67663 Kaiserslautern, Deutschland

Der letzte Beitrag *BIGGR: Bringing Gradoop to Applications* der Leipziger Autoren M.A. Rostami, M. Kricke, E. Peukert, S. Kühne, M. Wilke, S. Dienst und E. Rahm beschreibt die Ergebnisse des ScaDS-assoziierten BIGGR-Projektes, um das an der Universität Leipzig entwickelte, verteilte Graphanalysesystem Gradoop in die Analyseplattform Knime zu integrieren, um damit eine breitere Nutzung zu ermöglichen. Die realisierte Lösung erlaubt die visuelle Definition von Analyse-Workflows unter Nutzung der Gradoop-Operatoren zur Transformation und Analyse von Graphdaten (z. B. sozialen oder bibliographischen Netzen). Für die Auswertung großer Datenmengen können die Gradoop-Operatoren verteilt auf einem Cluster mit Apache Flink ausgeführt werden. Zudem werden unterschiedliche Ansätze zur Visualisierung von Graphen und Analyseergebnissen unterstützt.

2 Community-Beiträge

In der Rubrik „Datenbankgruppen vorgestellt“ gibt der Beitrag *Die Arbeitsgruppe Datenbanken und Informationssysteme an der TU Dortmund* von Jens Teubner einen Überblick über die Aktivitäten dieser Arbeitsgruppe in Forschung und Lehre. Bei ihrer Forschungsarbeit ist das überspannende Thema seit einigen Jahren *Datenbanktechnologie für moderne Systemarchitekturen*. In der Lehre bietet sie neben den DBIS-Vorlesungen im Masterstudium regelmäßig *Projektgruppen* an, in denen 8–12 Studierende über zwei Semester gemeinsam an einem realitätsnahen Software- oder Hardwareprojekt arbeiten. Weiterhin ist die Arbeitsgruppe auch bei der Weiterbildung von Fachkräften aus der Praxis aktiv; so wirkt sie u. a. an einem Zertifikatsstudium *Data Science and Big Data* mit, das vom Dortmunder Zentrum für HochschulBildung (zhb) organisiert wird.

In der Rubrik „Kurz erklärt“ gibt Klaus Meyer-Wegener (Universität Erlangen-Nürnberg) einen kurzen Einblick in ein interessantes Thema, das heute schnell an praktischer Bedeutung gewinnen und damit auch für die DBIS-Gemeinde zunehmend wichtiger werden könnte. Im Beitrag *Wie funktioniert die Blockchain?* beschreibt er, wie auf Basis eines verteilten Datenhaltungssystems oder einer verteilten Datenbank die Blockchain-Technologie entwickelt wurde, um für dezentrale und heterogene Anwendungen mit anonymen Teilnehmern in einem Rechnernetz „Vertrauen zu schaffen“. Er skizziert dabei auch eine Reihe von Problemen, die für den weltweiten Einsatz dieser Technologie noch zu lösen sind.

Die Rubrik „Dissertationen“ enthält in diesem Heft 9 Kurzfassungen von Dissertationen aus der deutschsprachigen DBIS-Community.

Schließlich berichtet die Rubrik „Community“ unter *News* über weitere aktuelle Informationen, welche die DBIS-Gemeinde betreffen.

3 Künftige Schwerpunktthemen

3.1 Data and Repeatability

What is common practice in most natural sciences has only recently entered the database field: Ensuring repeatability (or reproducibility) of experiments, in order to validate scientific results and enable experimental comparisons of methods. In our field, the ability to reproduce and repeat experiments has two main ingredients: First, the software or sufficient method description. Second, the data to run the experiments on. This special issue on data and repeatability places its focus on this second, arguably more challenging part.

Providing data for experimentation must overcome many obstacles. For instance, the data must be non-private and non-proprietary; for many types of experiments, data must be properly labeled or accompanied by a gold-standard; in many cases, data is “massaged” before entering experiments; special properties of the data, such as distributions or size, must be known or even adaptable. Sometimes data is varied to fit specific needs of an experiment, i.e., through upsampling and augmentation. In addition, “input data” for experiments may refer to many different things: from raw data to cleaned data, from sets of non-integrated CSV-files to a fully integrated relational database, etc.

We are calling for non-typical database contributions that report on

- Experiences in handling data for scientific and industrial purposes
- Experiences in handling data in data science/ML/AI workflows
- Efforts to create, evaluate or use data for benchmarking
- Data preparation/cleaning, and data quality war stories
- Data life cycle management
- Long-term data preservation and curation
- Data hubs and repositories
- Description of datasets of general interest and open data
- Data and the law – legally managing data
- Possible impacts on our publishing culture

Submissions can range from single pages, for instance to introduce a dataset, to full-fledged scientific contributions, for instance an experimental analysis of data cleaning methods or war stories.

Expected size of the paper: 8–10 pages, double-column (cf. the author guidelines at www.springer.com/13222). Contributions either in German or in English are welcome.

Deadline for submissions: Feb. 1st, 2019
Issue delivery: DASP-2-2019 (July 2019)

Guest editors:
Jens Dittrich, Universität des Saarlandes
jens.dittrich@cs.uni-saarland.de
Felix Naumann, Hasso Plattner Institut, Universität Potsdam
Felix.Naumann@hpi.de
Norbert Ritter, Universität Hamburg
ritter@informatik.uni-hamburg.de

3.2 Best Workshop Papers of BTW 2019

This special issue of the "Datenbank-Spektrum" is dedicated to the Best Papers of the Workshops running at the BTW 2019 at the University of Rostock. 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, 2019

Deadline for submissions: June 1st, 2019
Issue delivery: DASP-3-2019 (November 2019)

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

3.3 Trends in Information Retrieval Evaluation

Evaluation is a central aspect in the research and development of information retrieval systems. In academia, the quantitative evaluation of such systems is mostly known under the term Cranfield paradigm. This research method has been established for more than 25 years in international evaluation campaigns such as the Text Retrieval Conference (TREC) or the Conference and Labs of the Evaluation Forum (CLEF). Meanwhile industrial research has taken a completely different approach. Many companies are able to access a large number of users and their interactions, which can be recorded and evaluated. These infrastructures allow alternative evaluations like large-scale A/B experiments or other online methods. In the last years, different approaches to go beyond TREC-style evaluations emerged to close the gap and to bring together academic and industrial evaluation.

We are calling for articles that report on novel evaluation efforts, like:

- Living labs
- Evaluation as a service
- Large-scale A/B tests
- Interactive retrieval evaluation
- Session-based evaluation
- User-centered evaluation
- Counterfactual evaluation
- Novel evaluations in application domains such as cultural heritage, digital libraries, social media, expert search, health information, etc.
- Other evaluations that go beyond TREC

Expected size of the paper: 8–10 pages, double-column (cf. the author guidelines at www.springer.com/13222). Contributions either in German or in English are welcome.

Deadline for submissions: Oct. 1st, 2019
Issue delivery: DASP-1-2020 (March 2020)

Guest editors:
Philipp Schaer, Technische Hochschule Köln
philipp.schaer@th-koeln.de
Klaus Berberich, Hochschule für Technik und Wirtschaft des Saarlandes
klaus.berberich@htwsaar.de