Noname manuscript No.
(will be inserted by the editor)

Der Lehrstuhl Datenbanken an der Universität Leipzig

Erhard Rahm

Received: date / Accepted: date

Zusammenfassung Der Lehrstuhl Datenbanken an der Universität Leipzig befasst sich schwerpunktmäßig mit automatisierten Verfahren zur Integration und Analyse großer Mengen heterogener Daten, v.a. aus dem Web. Im Zusammenhang mit "Big Data" werden unterschiedlichste Hochleistungsstrategien verfolgt, u.a. Skew-resistente Lastbalancierungsmethoden für MapReduce sowie die Nutzung moderner Grafikprozessoren (GPUs). Zum Matching von Modellen (Schemas, Ontologien) und von Instanzdaten wurden leistungsfähige Verfahren und mehrere Prototypen entwickelt. Untersucht werden ferner Methoden zur Evolution von Ontologien und Mappings, um die Auswirkungen von Ontologieänderungen zu minimieren. Der Bericht gibt nach einer Einleitung zur Entwicklung des Lehrstuhls einen Überblick zu den aktuellen Forschungsthemen. Angaben zum Lehrprofil runden die Darstellung ab.

Schlüsselwörter Big Data · Datenintegration · Ontologien · Leipzig

1 Entwicklung

Der Lehrstuhl Datenbanken gehört zum Institut für Informatik an der Fakultät für Mathematik und Informatik der Universität Leipzig und wird seit April 1994 von Prof. Dr. Erhard Rahm geleitet. Erhard Rahm promovierte (1988) und habilitiere (1993) an der Universität Kaiserslautern bei Prof. Dr. Theo Härder zu Themen von Parallelen Datenbanksystemen bzw. Hochleistungs-Transaktionssystemen. Als Post-Doc war er ein Jahr (1988/89) am IBM Research Center in Hawthorne, NY, in der Abteilung "Large Systems" tätig.

Erhard Rahm
Institut für Informatik, Universität Leipzig
Augustusplatz 10, 04109 Leipzig
E-Mail: rahm@informatik.uni-leipzig.de
Website: `http://dbs.uni-leipzig.de`

Nach der Berufung an die Universität Leipzig standen zunächst der Aufbau der eigenen Arbeitsgruppe und als Institutsleiter (1996-99) auch der des erst 1993 gegründeten Instituts für Informatik im Vordergrund. Die Forschungsthemen diversifizierten sich zunehmend und beinhalteten neben parallelen Datenbanksystemen auch Fragestellungen von digitalen Bibliotheken und der Metadatenverwaltung für Data Warehouses.

Einen signifikanten Einfluss auf die weiteren Forschungsschwerpunkte des Lehrstuhls hatte ein Forschungssemester im Jahr 2000 bei Microsoft Research in Redmond, WA. Im Mittelpunkt der Arbeiten stand dort die Ausgestaltung der von Dr. Phil Bernstein formulierten Vision des "Model Management" zur generischen Verarbeitung von Metadaten-Modellen (z.B. DB-Schemas) und Mappings mit mächtigen Operatoren wie Match, Merge und Compose. Zur Realisierung des Matchings wurde 2001 ein Survey veröffentlicht [12], der bis heute gemäß Google Scholar rund 3000-mal zitiert wurde.

Daraufhin wurden in Redmond und Leipzig mit *Cupid*, *Coma* und *SimilarityFlooding* unter Mitwirkung von Erhard Rahm erste Prototypen bzw. neue Algorithmen zum generischen Schema Matching entwickelt, die ebenfalls große Aufmerksamkeit und Bedeutung erlangten. Zwei der entsprechenden Papiere wurden mit einem "Test of Time" Award ausgezeichnet. Für die Cupid-Publikation wurde 2011 der *VLDB Ten-Year Best Paper Award* verliehen (zusammen mit J. Madhavan und P. Bernstein), für das Similarity Flooding-Papier 2013 der *ICDE Influential Paper Award* (zusammen mit S. Melnik und H. Garcia-Molina).

Die Arbeiten zur Metadatenverwaltung wurden am Lehrstuhl zunehmend ausgedehnt auf das Matching sowie Merging von Ontologien sowie Ansätze zur Diff-Berechnung nach Ontologieänderungen. Einen neuen Schwerpunkt bilden zudem Ansätze zum Data Cleaning und Objekt-Matching (Entity Resolution), um zu ganzheitlichen Datenintegrations-

lösungen zu kommen. Hierzu werden auch zunehmend lernbasierte Methoden eingesetzt, u.a. um den manuellen Konfigurationsaufwand zu minimieren. Zur dynamischen Integration von Webdaten wurde bereits 2005 ein erstes Mashup-ähnliches Framework (*iFuice*) realisiert.

Weitere Forschungsarbeiten befassten sich mit XML-Datenbanken sowie mit der Realisierung von adaptiven Workflow-Änderungen zur flexiblen Ausnahmebehandlung während der Ausführung von Workflows (System *AgentWork*). Zur Generierung von *Recommendations* auf Webseiten, z.B. in E-Commerce-Shops, wurde ein adaptiver Ansatz namens *Awesome* entwickelt (VLDB 2004), der den Erfolg vergangener Recommendations berücksichtigt.

Seit 2002 erfolgen viele Untersuchungen bezüglich der Anwendungsdomäne der Lebenswissenschaften in enger Kooperation mit dem Leipziger Bioinformatikzentrum IZBI (Interdisziplinäres Zentrum für Bioinformatik). So wurden mehrere Plattformen mit physischer, virtueller und hybrider Datenintegration implementiert und evaluiert. Zudem werden viele Ontologiethemen für biomedizinische Ontologien bearbeitet. 2004 wurde die bis heute bestehende DILS-Konferenzserie (Data Integration in the Life Sciences) in Leipzig ins Leben gerufen.

Die umfangreichen Arbeiten und Protoytyp-Realisierungen zur Datenintegration ermöglichten die Einwerbung einer großen BMBF-Förderung zur Einrichtung des sogenannten *Web Data Integration (WDI)* Labs. In dem seit 2010 bestehenden Innovationslabor wurden zunächst die vorliegenden Prototypen zurn Schema- und Objekt-Matching sowie zur Mashup-artigen Integration von Webdaten für den Markteinsatz weiter entwickelt. Die Ergebnisse wurden an ein Anfang 2012 gegründetes, in Leipzig ansässiges Spinoff transferiert, mit dem eng kooperiert wird. Im Jahr 2012 wurde ein zweites Startup von einem ehemaligen Doktoranden des Lehrstuhls (Dr. Golovin) in Leipzig gegründet, das zunächst durch ein Exist-Stipendium vorbereitet wurde. In diesem Unternehmen geht es um die optimierte Nutzung von Data Warehouse-Technologien.

Nach dem Neubau des Innenstadt-Campus ist der Lehrstuhl wie das Institut für Informatik seit 2012 wieder am angestammten Platz am Augustusplatz angesiedelt, jetzt in dem sogenannten Paulinum am Ort der zu DDR-Zeiten gesprengten Paulinerkirche. Die Arbeitsgruppe umfasst neben einer Sekretärin und einem technischen Mitarbeiter derzeit elf wissenschaftliche Mitarbeiter bzw. Doktoranden sowie mehrere wissenschaftliche und studentische Hilfskräfte. Im Folgenden werden zunächst die aktuellen Forschungsthemen dargestellt. Abschließend wird das Lehrkonzept skizziert. Detail-Informationen sowie die Publikationen finden sich auf der Lehrstuhl-Website http://dbs.uni-leipzig.de.

2 Forschungsthemen

2.1 Integration und Analyse von Webdaten

Das im Jahr 2010 als Lehrstuhlerweiterung gegründete WDI-Lab beschäftigt sich mit der Entwicklung von Werkzeugen und Verfahren zur semantischen Integration von Daten aus dem Web und aus lokalen (Unternehmens-) Quellen. Nach Auslaufen der BMBF-Förderung werden in dem Labor neue Drittmittelprojekte durchgeführt, derzeit ein EU-Projekt (*LinkedDesign*) und ein DFG-Projekt zur Link Discovery. Wir besprechen beispielhaft aktuelle Arbeiten zur dynamischen Datenfusion, des Matchings von Produktangeboten und der Erkennung von Produktplagiaten sowie zur Link Discovery.

Zur Integration heterogener Webdaten wurde ein neuartiges Mashup-Framework namens WETSUIT (Web EnTity Search and fUsIon Tool) realisiert [3]. Damit können interaktive Webanwendungen entwickelt werden, die Daten unterschiedlicher Quellen zur Laufzeit anfragen, fusionieren und analysieren. Ein wesentliches Merkmal ist dabei die Unterstützung adaptiver Suchstrategien, um anwendungsrelevante Objekte aus Datenquellen wie Entity-Suchmaschinen und Web-Datenbanken effektiv und effizient (d.h. mit minimalem Kommunikationsaufwand) abzurufen [2]. Die Suchverfahren nutzen pro Datenquelle mehrere Query-Generatoren zur automatisierten Erzeugung unterschiedlicher Anfragen. WETSUIT stellt neben der Anfragegenerierung weitere mächtige Operatoren bereit, u.a. zum Objekt-Matching. Es unterstützt ferner die parallele, überlappende Ausführung mehrerer Operatoren, verbunden mit einem Streaming von Datenobjekten zwischen den Operatoren. Damit konnten bereits mehrere anspruchsvolle und hoch interaktive Datenintegrationsanwendungen entwickelt werden.

Objekt-Matching (Deduplizierung, Entity Resolution) ist ein wesentlicher Schritt zur Datenintegration. Es bezeichnet die Aufgabe, Instanzen in verschiedenen Datenquellen zu identifizieren, die dasselbe Objekt der realen Welt beschreiben. Im WDI-Lab werden vor allem Verfahren zum Abgleich von Produktangeboten aus unterschiedlichen Web-Shops entwickelt, was aufgrund der hohen Heterogenität sowie oft schlechten Datenqualität der Angebotsbeschreibungen eine besondere Herausforderung darstellt. Es wurden daher entsprechende Methoden zur Datenbereinigung realisiert, u.a. zur Vereinheitlichung von Herstellernamen sowie zur Extraktion herstellerspezifischer Produktcodes aus dem Titel oder Beschreibungstext eines Produktangebots. Zum Matching werden ferner pro Produktkategorie eigene lernbasierte Match-Strategien verwendet, welche zur verbesserten Vorhersage mehrere Einzelverfahren kombiniert einsetzen [7].

Ein neues, mit dem Matching von Produktangeboten verwandtes Thema ist das semi-automatische *Aufspüren von Produktfälschungen*, die in Online-Shops oder Auktionsplatt-

formen zunehmend angeboten werden. Hierbei geht es neben dem Abgleich unterschiedlicher Angebote darum, verdächtige Abweichungen z.B. in der Beschreibung oder sonstige Hinweise auf Fälschungen zu erkennen und bei einer Plagiatbewertung zu berücksichtigen. Das Thema verspricht angesichts des wachsenden Umfangs der Produktpiraterie im Internethandel eine hohe wirtschaftliche Relevanz.

Im sog. Semantic Web werden immer mehr Datenquellen verschiedener Anwendungsbereiche im RDF-Format veröffentlicht und im Rahmen einer "Linked Open Data" (LOD)-Initiative vernetzt, um eine übergreifende Auswertung und Integration der Daten zu ermöglichen. Eine aktuelle Herausforderung ist "Link Discovery", d.h. die Identifizierung neuer semantischer Links zwischen Datenquellen. Ein erster Ansatz versucht dabei bereits bestehende Links wiederzuverwenden, um effektiv neue Links zu erzeugen. Aufgrund der Größe und Komplexität der LOD Cloud gilt es dabei nur die vielversprechendsten Links auswählen, um eine effiziente Match-Verarbeitung zu ermöglichen. Zu diesem Zweck wurden verschiedenartige Selektions- sowie Graphalgorithmen erforscht. Daneben sollen neue lernbasierte Verfahren Verwendung finden, insbesondere des "aktiven Lernens" um mit möglichst wenig Nutzerfeedback effektive Strategien zur Link Discovery zu bestimmen.

2.2 Objekt-Matching für Big Data

In den vergangenen Jahren haben Cloud-Infrastrukturen und Programmiermodelle wie MapReduce enorm an Bedeutung gewonnen. Sie ermöglichen die effiziente Bearbeitung sehr datenintensiver Probleme ("Big Data") durch Parallelverarbeitung in Cluster-Umgebungen mit bis zu Tausenden von Knoten unter Abstraktion von den Details des verteilten Rechnens sowie der Behandlung von Hardwarefehlern.

Objekt-Matching ist ein ideales Problem zur Ausführung auf Cloud-Infrastrukturen, da es oft auf sehr großen Datenmengen (z.B. Kundendaten oder Produktangeboten) auszuführen ist und eine inhärent quadratische Komplexität aufweist. Der wünschenswerte Einsatz lernbasierter Verfahren führt zu zusätzlichen Leistungsanforderungen, da diese unoptimiert auf großen Datenmengen oft prohibitiv hohe Laufzeiten verursachen. Eine Herausforderung bei der Cloud-basierten Lösung von Objekt-Matching-Aufgaben ist die effektive Parallelisierung kompletter Match-Workflows für unterschiedlichste Anwendungsfälle und Datenverteilungen. Hierzu wurde ein umfassendes MapReduce-basiertes Framework namens *Dedoop* (Deduplication with Hadoop) realisiert, mit dem zuvor spezifizierte Match-Workflows automatisch auf unterschiedlichen Hadoop-Cluster zur Ausführung kommen [5]. Der Prototyp unterstützt dabei eine Vielzahl verschiedener Blocking-Techniken zur Reduzierung des Suchraums sowie Match-Techniken. Insbesondere werden trainingsbasierte, maschinelle Lernverfahren unterstützt. Match-Workflows sowie Parametrisierungen der Hadoop-Cluster lassen sich bequem über eine Web-GUI einstellen.

Zur MapReduce-Realisierung der Match-Workflows wurden u.a. mehrere Techniken zur Lastbalancierung bzw. Skew-Behandlung entwickelt [6]. Dabei wird ein Analyse-MapReduce-Job zur Erfassung der Datenverteilungen und Partitionsgrößen eingesetzt, um danach in der eigentlichen Match-Verarbeitung angepasste Datenumverteilungen zwischen Map- und Reduce-Tasks für eine gleichmäßige Auslastung der Knoten zu erreichen. Weitere Optimierungen betreffen die Vermeidung redundanter Match-Vergleiche im Fall überlappender Partitionierungen zu vergleichender Objekte.

Als Komplementärstrategie zum Einsatz von Cloud-Infrastrukturen wurde begonnen, die Beschleunigung von Match-Aufgaben durch Portierung auf Grafikprozessoren (GPUs) zu untersuchen, die eine inhärent hohe Parallelisierung zu niedrigen Kosten unterstützen. In einem ersten Schritt wurde hierzu die GPU-Portierung für häufig auszuführende Funktionen der Stringvergleiche (z.B. auf Basis von n-gram-Vergleichen) untersucht. Trotz der Beschränkungen der Grafikprozessoren konnten dabei aufgrund der hohen Parallelisierung signifikante Leistungssteigerungen erzielt werden. Für die Zukunft erscheint sinnvoll für Big Data, die Parallelverarbeitung auf mehreren Stufen zu unterstützen: innerhalb eines Knotens durch Multiprocessing mit mehreren Prozessoren bzw. Cores und durch Grafikprozessoren sowie knotenübergreifend innerhalb von Cloud-Infrastrukturen wie Hadoop-Cluster [9].

2.3 Matching und Merging großer Ontologien

Zum generischen Schema-Matching wurde bereits 2002 mit *Coma* ein leistungsfähiger Prototyp entwickelt, der zur Qualitätsverbesserung mehrere Match-Techniken (Matcher) kombiniert einsetzt. Außerdem wird eine Wiederbenutzung (Reuse) bereits vorliegender Mappings unterstützt, insbesondere durch Mapping-Komposition (ein Mapping zwischen Schema A und C kann z.B. durch Komposition von zwei Mappings A-B und B-C mit einem Schema B erfolgen). Im Nachfolgesystem *Coma* $++$ wurden neben einem GUI weitere Verbesserungen insbesondere zur Unterstützung großer Schemas und Ontologien [11] integriert. Hervorzuheben ist ein Partitionierungsansatz, mit dem zur Suchraumreduzierung und damit zur Beschleunigung nur noch die ähnlichsten Schemafragmente miteinander verglichen werden. Die derzeitige, weiterentwickelte Version *Coma* 3.0 [8] ist in einer Teilversion als Open Source nutzbar.

Semi-automatische Match-Tools wie Coma sind zwar sehr mächtig, aber oft nur schwer bezüglich der Auswahl anzuwendender Matcher und Konfigurierung anderer Parameter für neue Match-Aufgaben einzustellen. Wir haben daher an einem selbstkonfigurierenden Match-System gearbeitet, dass sich automatisch an ein gegebenes Match-Problem anpassen kann. Der Ansatz basiert auf der Analyse der Eingabe-

Schemas und auch von Zwischenergebnissen bereits ausgeführter Matcher [10]. Die damit gewonnenen Kenngrößen werden in verschiedenen Regeln genutzt, um eine optimierte Auswahl der Matcher und anderer Einstellungen zu gewinnen bzw. zur Laufzeit anzupassen. Erste Evaluationsergebnisse zeigen, dass damit Match-Probleme aus verschiedenen Domänen in guter Qualität gelöst werden können.

Derzeit untersuchen wir ferner, wie die Semantik der Mappings für Ontologien verbessert werden kann, in dem nicht nur Gleichheits-Korrespondenzen, sondern auch Is-a- und Part-of-Beziehungen zwischen Konzepten zweier Ontologien erkannt werden sollen. Hierzu verwenden wir zunächst einen Enrichment-Ansatz, der Ergebnisse eines herkömmlichen Match-Tools wie Coma nachbearbeitet. Wenn z.B. mehrere Konzepte der ersten Ontologie in einer vermeintlichen Match-Beziehung mit einem Konzept der zweiten Ontologie stehen, liegen tatsächlich oft Part-of- bzw. Is-a-Beziehungen vor. Auch lassen sich durch linguistische Analysen der Konzeptbezeichnungen häufig semantische Beziehungen zwischen Konzepten erkennen. So können Is-a-Beziehungen vorliegen, wenn ein Konzeptname ein Teilbegriff eines anderen Konzeptnamens ist (z.B. *Keyboards* Is-A *KeyboardsAndComputers* oder *ActionGame* Is-A *Game*).

Neben dem Matching ist das Zusammenführen bzw. Merging verschiedener Schemas oder Ontologien eine wichtige, und trotz langjährigen Arbeiten zur Schemaintegration noch unzureichend gelöste Herausforderung bei der Informationsintegration. Zum Merging von Ontologien bzw. Taxonomien wurde ein Ansatz namens ATOM (Automatic Target-driven Ontology Merging) realisiert, der auf einem Match-Mapping zwischen den Ontologien aufbaut und weitgehend automatisch eine integrierte Ontologie erzeugt. Der Ansatz ist asymmetrisch und geht davon aus, dass eine der Eingabeontologien die Zielontologie darstellt, in welche die andere Ontologie eingefügt wird, wobei die Struktur der Zielontologie so weit wie möglich bewahrt wird. Der Ansatz kann zudem semantische Is-a-Korrespondenzen zwischen den Eingabeontologien zur Verbesserung des Merge-Ergebnisses nutzen.

Das Matching von Ontologien hat zuletzt v.a. in den Lebenswissenschaften große Bedeutung erlangt, da dort eine Vielzahl oft großer und überlappender Ontologien Verwendung findet. Für diese Domäne wurde – aufbauend auf den Erfahrungen mit Coma – ein weiteres Match-System im Rahmen unseres Systems zur Ontologie-Evolution GOMMA entwickelt. Es verfügt über einige Besonderheiten, insbesondere die Unterstützung eines parallelen Matchings auf Mehrprozessorknoten sowie eine weitergehende Unterstützung einer Mapping-Komposition zur Nutzung bereits vorliegender Match-Ergebnisse. Im Rahmen der Ontology Alignment Evaluation Initiative (OAEI) 2012 gehörte GOMMA zu den besten Systemen.

2.4 Evolution von Ontologien und Mappings

Ontologien werden in verschiedenen Wissenschaftsdisziplinen sowie in kommerziellen Anwendungen zunehmend zur einheitlichen und semantischen Annotation von Objekten verwendet. Bedingt durch neue wissenschaftliche Erkenntnisse oder aufgrund neuer Anforderungen unterliegen die Ontologien ständigen Änderungen, welche sich auf abhängige Datenquellen, Mappings und Anwendungen auswirken. Im Rahmen eines DFG-Projekts wurde mit GOMMA (Generic Ontology Matching and Mapping Management) eine umfassende Infrastruktur zum Management und zur Evolutionsbehandlung großer Ontologien und Mappings, vorrangig im Bereich der Lebenswissenschaften, realisiert.

Zunächst wurde die Evolution von Ontologien, Annotationen und Ontologie-Mappings für zahlreiche biomedizinische Ontologien quantitativ analysiert. In den Ontologien wurden neben zahlreichen Ergänzungen auch häufiger Löschvorgange und Revisionen festgestellt, welche die Stabilität der Ontologie reduzieren und Folgeänderungen für Mappings und Anwendungen verursachen können.

Um mit den Auswirkungen der Änderungen besser umgehen zu können, besteht ein wichtiger Schritt in der Bestimmung der Differenz (des Diff) zwischen zwei Versionen einer Ontologie. Hierzu wurde ein regelbasierter Diff-Ansatz entwickelt, der auf Basis eines Match-Ergebnisses ein semantisch ausdrucksstarkes Evolutions-Mapping bestimmt [4]. Darin sind neben einfachen Änderungen wie das Hinzufügen oder Löschen von Konzepten auch komplexere Änderungen wie das Splitting oder Mischen von Konzepten erfasst. Über die Web-Anwendung CODEX (Complex Ontology Diff Explorer) kann die Diff-Berechnung für zahlreiche biomedizinische Ontologien für Auswertungen veranlasst werden.

Eine weitere Untersuchung befasst sich mit der semiautomatischen Adaption von Ontologie-Mappings auf Basis eines zuvor berechneten Evolutions-Mappings. Das Ziel ist dabei, vorhandene Mappings zum Großteil weiter zu nutzen und nur die von Ontologieänderungen betroffenen Korrespondenzen anzupassen. Der neue Ansatz nutzt dabei pro Änderungsoperation konfigurierbare Mapping-Anpassungen bzw. verlangt Nutzer-Feedback, wenn mehrere Adaptationsalternativen bestehen (z.B. nach dem Löschen oder Split eines Ontologiekonzepts).

2.5 Bibliometrische Analysen

Als interessante Analyseanwendung auf Basis integrierter Daten verschiedener Quellen haben wir uns mit verschiedenen bibliometrischen Auswertungen von Publikationen, vorrangig im Datenbankbereich, befasst. Hierzu wurden bibliographische Daten aus mehreren Web-Portalen (DBLP, ACM Digital Library, u.a.) und der Entity-Suchmaschine

Google Scholar extrahiert, bereinigt und u.a. mit Verfahren des Objekt-Matchings integriert.

Untersucht wurden u.a. die Zitierungshäufigkeiten wichtiger Datenbank-Zeitschriften und Tagungen über einen Zeitraum von zehn Jahren; für die BTW-Konferenz sogar seit dem Bestehen (1985). Ein Ergebnis dabei war, dass die Publikationen der Top-Tagungen VLDB und SIGMOD im Mittel deutlich häufiger als die der Zeitschriften VLDB Journal oder TODS zitiert werden, wobei generell pro Tagung und Zeitschrift sehr große Unterschiede in der Zitierungshäufigkeit einzelner Publikationen bestehen. Zur schnellen Zitierungsanalyse wurde ein Mashup namens OCS (Online Citation Service) implementiert, mit dem für jede(n) in DBLP gelistete Tagung, Zeitschriftenjahrgang oder Autor die Zitierungszahlen von Google Scholar ermittelt werden können.

Eine weitere Untersuchung befasste sich mit der Analyse der Herkunftseinrichtungen (Affiliations) der in den Top-Tagungen und Top-Zeitschriften erschienenen Publikationen [1]. Hier besteht die Herausforderung, die Affiliation-Angaben zunächst aus den Publikationen bzw. Web-Portalen zu extrahieren, zu bereinigen und zu konsolidieren. Hierzu wurde u.a. eine Referenz-Datenbank mit allen gefundenen Affiliation-Bezeichnungen aufgebaut. Eine Erkenntnis war, dass die Research-Labs von IBM und Microsoft eine dominierende Rolle einnehmen und dass im universitären Bereich die traditionell starken Standorte in den USA (Stanford, Madison, Berkeley) bzgl. der Publikationsanzahl von asiatischen Universitäten in Singapur und Hongkong eingeholt worden sind. Die deutschen Einrichtungen waren vor ca. 10 Jahren noch die Nr. 2 hinter den USA bzgl. der Publikationszahl in den Top-Tagungen und -Zeitschriften. Mittlerweile sind sie hinter China und auch Kanada zurückgefallen.

3 Lehre

Der Lehrstuhl deckt die Ausbildung in Datenbank- und Informationsystemen in den Bachelor- und Masterstudiengängen der Informatik und Wirtschaftsinformatik der Universität Leipzig ab. Die Basisangebote DBS1, DBS2 und das Datenbank-Praktikum bilden jeweils ein eigenes Modul im Bachelorstudiengang. Daneben werden unter Mitwirkung promovierter Lehrstuhl-Mitarbeiter z.Zt. acht vertiefende Vorlesungen (u.a. DBS-Implementierung, Cloud Data Management, Mehrrechner-DBS, Data Warehousing, Datenintegration, Bio Data Management), ein Data Warehouse-Praktikum sowie Seminare mit wechselnden Inhalten regelmäßig angeboten, die flexibel Modulen zugeordnet werden können. Damit können sich bereits Bachelor-Studenten im DB-Bereich vertiefen und besser auf eine spezifische Bachelorarbeit vorbereiten. Die Vorlesungen zur DBS-Implementierung und Mehrrechner-DBS basieren natürlich auf den Lehrbüchern des Autors. Zu allen Vorlesungen gibt es Online-Übungen zur automatisierten Überprüfung der Lerninhalte. Im Rahmen des neuen Uni-Campus am Augustusplatz mit Institutsgebäude, Hörsaalgebäude, Mensa, Bibliotheken etc. bieten sich für die Informatik-Studierenden optimale Lernbedingungen und direkter Kontakt mit den Mitarbeitern im Zentrum von Leipzig.

Zur Nachwuchsförderung zeichnet der Lehrstuhl seit 2008 jährlich die drei besten Studierenden aus, die in dem Jahr mindestens zwei DBS-Lehrveranstaltungen mit sehr gutem Ergebnis absolviert haben. In jedem Semester wird ein Oberseminar durchgeführt, in dem die Studenten über ihre am Lehrstuhl laufenden Bachelor- bzw. Masterarbeiten und die Mitarbeiter über ihre Forschungsarbeiten berichten. Im Sommersemester findet dieses Seminar seit 2001 jeweils an der Außenstelle der Uni Leipzig in Zingst, in unmittelbarer Nähe zur Ostsee, statt.

Literatur

1. Aumüller, D., Rahm, E.: Affiliation analysis of database publications. SIGMOD Record **40**(1), 26–31 (2011)
2. Endrullis, S., Thor, A., Rahm, E.: Entity search strategies for mashup applications. In: ICDE (2012)
3. Endrullis, S., Thor, A., Rahm, E.: Wetsuit: An efficient mashup tool for searching and fusing web entities. PVLDB **5**(12), 1970–1973 (2012)
4. Hartung, M., Groß, A., Rahm, E.: Conto-diff: generation of complex evolution mappings for life science ontologies. Journal of Biomedical Informatics **46**(1), 15–32 (2013)
5. Kolb, L., Rahm, E.: Parallel entity resolution with dedoop. Datenbank-Spektrum **13**(1), 23–32 (2013)
6. Kolb, L., Thor, A., Rahm, E.: Load balancing for map/reduce-based entity resolution. In: ICDE, pp. 618–629 (2012)
7. Köpcke, H., Thor, A., Thomas, S., Rahm, E.: Tailoring entity resolution for matching product offers. In: EDBT, pp. 545–550 (2012)
8. Maßmann, S., Raunich, S., Aumüller, D., Arnold, P., Rahm, E.: Evolution of the coma match system. In: OM (2011)
9. Ngomo, A.C.N., Kolb, L., Heino, N., Hartung, M., Auer, S., Rahm, E.: When to reach for the cloud: Using parallel hardware for link discovery. In: ESWC (2013)
10. Peukert, E., Eberius, J., Rahm, E.: A self-configuring schema matching system. In: ICDE, pp. 306–317 (2012)
11. Rahm, E.: Towards large-scale schema and ontology matching. In: Schema Matching and Mapping, pp. 3–27 (2011)
12. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB J. **10**(4), 334–350 (2001)