Forschungsbericht 2012/2013

Abteilung Datenbanken

Universität Leipzig, Institut für Informatik

Web: http://dbs.uni-leipzig.de, http://wdilab.uni-leipzig.de

Inhaltsüberblick

Abt. Datenbanken im Juli 2013. *Obere Reihe v.l.n.r:* Marcel Jacob, -/-, Markus Nentwig, André Petermann, Benedict Preßler (Student). *Untere Reihe v.l.n.r:* Ziad Sehili, Prof. Dr. Erhard Rahm, Dan Häberlein (Student), Martin Junghanns (Student), Patrick Arnold, Sergej Sintschilin (Student), Li-Hui Lee, Toni Fröschke (Student), Anika Groß, Peggy Lucke (Studentin), Christian Wartner.

Personelle Zusammensetzung

Univ.-Professor	Prof. Dr. Rahm, Erhard
Wiss. Mitarbeiter	Arnold, Patrick
Wiss. Mitarbeiter (DFG)	Endrullis, Stefan (bis Juni 2012)
Wiss. Mitarbeiterin (DFG)	Groß, Anika
Wiss. Mitarbeiter	Dr. Hartung, Michael (bis Sep. 2013)
Sekretärin	Hesse, Andrea
Programmierer	Jusek, Stefan
Doktorandin	Köpcke, Hanna
Wiss. Mitarbeiter	Kolb, Lars
Wiss. Mitarbeiter (DFG)	Nentwig, Markus (ab Nov. 2012)
Doktorand	Petermann, André (ab Jan. 2013)
Doktorand	Peukert, Eric
Wiss. Mitarbeiter	Sehili, Ziad (ab Okt. 2013)
Wiss. Mitarbeiter	Dr. Thor, Andreas
Wiss. Mitarbeiter (EU)	Wartner, Christian

1. Highlights

Im Berichtszeitraum 2012/2013 sind folgende Ereignisse hervorzuheben:

1. **Forschungstransfer in hochqualifizierte Arbeitsplätze:** Anfang 2012 wurden die beiden Spinoffs *Webdata Solutions GmbH* und *Data VirtualityGmbH* gegründet. Beide Firmen nutzen Technologien, die in langjähriger Forschung an der Abteilung Datenbanken entstanden und für den Markteinsatz weiterentwickelt wurden. Die *Webdata Solutions GmbH* wurde von den ehemaligen Mitarbeiterinnen der Abteilung Datenbanken und des WDI Labs (BMBF Projekt 2010-2011) Hanna Köpcke, Carina Röllig und Sabine Maßmann gegründet. Die Firma *Data VirtualityGmbH* entstand aus dem Exist-Gründerstipendium von Dr. Nick Golovin zum Thema „Data Virtualizer“ (Mentor: Prof. Rahm).

2. Auf der ICDE 2013 wurden Prof. Dr. Erhard Rahm und seiner früherer Doktorand Sergey Melnik (Google) zusammen mit Prof. Dr. Garcia-Molina (Stanford University) mit dem renommierten **10+ Year Most Influential Paper Award** für ihre ICDE 2002 Publikation “Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching” ausgezeichnet. Nach dem „Ten-Year Best Paper Award“ von der VLDB 2011 ist dies bereits die zweite herausragende, internationale Auszeichnung für Forschungsergebnisse der Abteilung im Gebiet Datenbanken bzw. Datenintegration.

3. Auf der ESWC 2013 wurde die Publikation "When to Reach for the Cloud: Using Parallel Hardware for Link Discovery" mit dem **Best Paper Award** ausgezeichnet. Die Publikation im neuen Gebiet „Big Data Integration“ entstand in Kooperation mit Kollegen der Leipziger Semantic-Web-Gruppe AKSW.

4. Im Februar 2012 wurde das WDI-Lab für eine der besten Gründerideen im Rahmen des Leipziger Ideenwettbewerbs für Existenzgründer (LIFE) ausgezeichnet. Die WDI-Lab-Einreichung zum Thema "Automatisierte Aufdeckung von Produktpiraterie“ wurde aus über 40 Einreichungen ausgewählt und mit dem dritten Preis prämiert.

5. 2012 startete das EU-Projekt LinkedDesign in Kooperation mit SAP Research.

6. Die DFG bewilligte 2012 ein Forschungsprojekt zur lernbasierten Link Discovery, das in Kooperation mit der AKSW-Gruppe (Dr. Auer, Dr. Ngonga) durchgeführt wird.

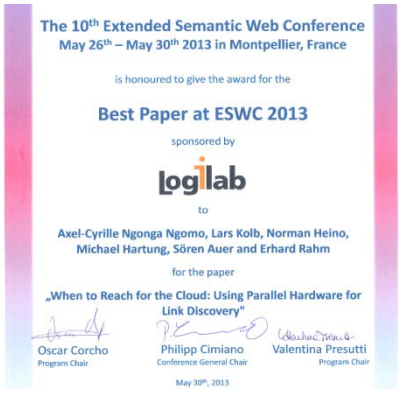
7. Im April 2012 wurde das Teilprojekt „Portal-basiertes Management von Ontologien und ihrer Evolution“ im Rahmen des Verbundprojekts eScience - Forschungsnetzwerk Sachsen (Cluster eSystems) bewilligt. Gefördert wird das Projekt durch den Europäischen Sozialfonds (ESF).

8. Im Sep. 2012 organisierte Prof. Rahm den gut besuchten *Data Integration Day* im Rahmen der SABRE-Konferenz an der Univ. Leipzig

9. Markus Nentwig wurde im Okt. 2012 auf der intern. Konferenz der IBM-Anwendervereinigung GSE (Guide-Share Europe) mit dem **GSE Academic Award for Excellence** für seine Untersuchungen im Rahmen der Masterarbeit ausgezeichnet. Die Masterarbeit bei Prof. Rahm erfolgte in enger Kooperation mit dem IBM-Entwicklungslabor in Böblingen.

10. Das Ontologie-Matching-System GOMMA erreichte beim OAEI (*Ontology Alignment Evaluation Initiative*) Wettbewerb 2012 die besten Ergebnisse in den Bereichen *Anatomy* und *Library*.
11. Dr. Andreas Thor übernahm von April 2012 bis März 2013 die Vertretungsprofessur am Lehrstuhl für Informatik mit dem Schwerpunkt Informationsmanagement an der Universität Passau. Im Frühjahr 2013 wurde er als Professor an die Hochschule für Telekommunikation Leipzig (HfTL) berufen.
12. Von Mai 2012 bis August 2013 war Frau Li-Hui Lee von der National Yan-Ming-Univ. in Taipeh, Taiwan, als Gastforscherin am Lehrstuhl. Die damit begonnene Kooperation untersucht Datenintegrationslösungen für klinische Laborsysteme, u.a. unter Nutzung von Ontologien wie LOINC.
13. Im Berichtszeitraum verteidigte Eric Peukert erfolgreich seine Dissertation. Außerdem konnte Anika Groß ihre Dissertation einreichen.
14. Das Oberseminar der Abteilung fand im April 2012 sowie im Juli 2013 bereits zum elften bzw. zwölften Mal an der Uni-Außenstelle in Zingst/Ostsee statt.
15. Im September 2012 erfolgte der Umzug der Abteilung Datenbanken aus dem Interimsgebäude des Instituts für Informatik in das Paulinum am Augustusplatz.

Links: Auszeichnung des WDI-Lab-Teams mit dem LIFE-Innovationspreis
Rechts: Übergabe des GSE Academic Award for Excellence an Markus Nentwig



Links: ICDE 10+ Year Most Influential Paper Award
Rechts: Best Paper Award ESWC 2013

Mitarbeiter der Abteilung Datenbanken beim Commerzbank Firmenlauf 2012 am Sportforum Leipzig

Projekte

Semantische Integration von Webdaten

E. Rahm, C. Wartner, P. Arnold

Das im Jahr 2010 gegründete WDI-Lab beschäftigte sich weiter mit der Entwicklung von Werkzeugen und Verfahren zur semantischen Integration von Daten aus dem Web und aus Unternehmen. Ein besonderer >Schwerpunkt lag dabei in der Entwicklung und Optimierung lernbasierter Match-Verfahren für Produktangebote in Online-Shops, um diese z.B. in Vergleichsportalen zugänglich zu machen. Die Arbeitsergebnisse ermöglichten 2012 die erfolgreiche Ausgründung eines neuen IT-Unternehmens, der Webdata Solutions GmbH.

Auch nach der Ausgründung wurden im WDI-Lab der Universität Leipzig verschiedene Projekte im Bereich der (Web-)Datenintegration durchgeführt. Beispielsweise stand die semi-automatische Aufdeckung von Plagiaten und Imitaten in Online-Shops und Auktionsplattformen im Fokus. Ein erster Ansatz zur Lösung dieses Problems wurde 2012 beim Leipziger Ideenwettbewerb für Existenzgründer (LIFE) für seine Innovation und wirtschaftliche Relevanz mit einem Preis ausgezeichnet. Der Lösungsansatz nutzt im WDI-Lab entwickelte Datenintegrationsansätze, um Produktdaten in Online-Shops zu extrahieren und zu integrieren und mit Hilfe von Clustering-Verfahren in vertrauenswürdige und nicht vertrauenswürdige Angebote zu unterteilen.

LinkedDesign - Datenintegration und Datenmanagement in Design, Engineering und Manufacturing

E. Rahm, C. Wartner, P. Arnold

In Zusammenarbeit mit dem Institut für Angewandte Informatik e.V. (InfAI) beteiligt sich die Abteilung seit Januar 2012 an dem EU-Projekt LinkedDesign, das zusammen mit 13 weiteren europäischen Partnern aus Wirtschaft und Wissenschaft durchgeführt wird. In dem auf 42 Monate angelegten Projekt steht Datenintegration und Management von Produkt-Lebenszyklus-Daten im Mittelpunkt. Ziel ist die Schaffung einer Plattform (LEAP - Linked Engineering and Manufacturing Plattform), die eine holistische Sicht auf alle Unternehmensdaten, Personen und Prozesse ermöglichen soll, die im Lebenszyklus eines Produktes relevant sind.

Die Abteilung für Datenbanken beschäftigt sich vor allem mit Methoden und Werkzeugen zur Unterstützung von Schema- und Objekt-Matching Aufgaben. Weitere Schwerpunkte waren die Erforschung von Methoden zur Erzeugung semantischer Graphen, deren Knoten Dokumente, Projektdaten, Personen und andere Objekte aus Design und Fertigungsprozessen sind. Das Ziel ist die Unterstützung eines neuen Frameworks, das Nutzer bei der Navigation und Suche unterstützt und weiterhin die Analyse von Graphen nach Mustern erlaubt, aus denen tiefere Einsichten und neues Wissen abgeleitet werden können.

Semantische Erweiterung von Mappings und Ontologie-Merging

E. Rahm, P. Arnold, S. Raunich

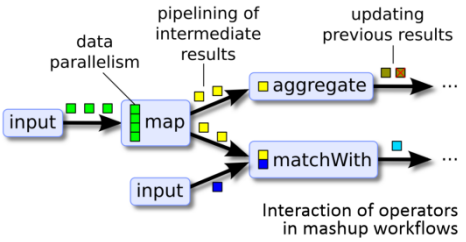
Die semantische Erweiterung vorhandener Schema- und Ontologie-Mappings spielt eine wesentliche Rolle bei der Unterstützung von typischen Datenintegrationsaufgaben wie z.B. der Integration mehrerer Ontologien oder der Adaptierung von Mappings infolge von Ontologieänderungen. Ausgehend von einem einfachen Äquivalenz-Mapping, werden konkrete semantische Beziehungen (equal, is-a, part-of, related) zwischen Konzeptpaaren bestimmt. Im Gegensatz zu bisherigen Arbeiten verwendet unser Ansatz linguistisch sauber definierte Beziehungstypen (Synonyme, Hyperonyme/Hyponyme, Meronyme/Holonyme, Kohyponyme). Es wurde eine neuartige 2-Schritt-Architektur (Matching und Nachbearbeitung) entwickelt, welche in verschiedenen Evaluationen deutlich besser als verwandte Systeme abschneiden konnte. Die Bestimmung des Beziehungstyps geschieht mittels linguistischer und struktureller Verfahren sowie unter Verwendung von Hintergrundwissen. Bspw. konnte die Verwendung von WordNet, UMLS und OpenThesaurus, die Qualität der angereicherten Mappings verbessern. Im Rahmen aktueller und zukünftiger Forschung wird das automatische Extrahieren linguistischer Beziehungen aus Wikipedia und weiteren Quellen wie Wiktionary sowie deren Integration in ein individuelles Repository verfolgt. Das so gewonnene Hintergrundwissen soll zur semantischen Erweiterung von Schema- und Ontologie-Mappings eingesetzt werden.

Derartige Mappings werden benötigt, um zwei oder mehrere Ontologien zu einer einheitlichen Ontologie zusammenzufügen (mergen genannt). Dabei sollen alle Informationen der Ausgangsontologien erhalten bleiben. Das Ontology Merging Tool ATOM (Asymmetric Target-driven Ontology Merging) kann semantische Beziehungen (z.B. *equal*, *is-a*) zwischen zwei unterschiedlichen Ontologien nutzen, um eine integrierte Ontologie zu erstellen. Anstelle eines üblichen symmetrischen Ansatzes, verwendet ATOM eine der beiden Ontologien als fixes Zielsystem (target-driven). Die Bewertung der Qualität verschiedener Ontology-Merging-Verfahren gestaltet sich schwierig. Daher wurden zunächst Maße zur Evaluierung der Qualität der erzeugten Ontologie wie z.B. die relative Abdeckung der Eingabe-Ontologien vorgestellt. Im Berichtszeitraum wurde ATOM weiterentwickelt und evaluiert.

Dynamische Fusion verteilter Webdaten

E. Rahm, S. Endrullis, A. Thor

Interaktive Webanwendungen, z.B. sogenannte Mashups, erfordern oft eine schnelle Fusion heterogener Daten zur Laufzeit. Eine solche dynamische Datenfusion ist jedoch angesichts der Heterogenität von Datenquellen sowie Qualitätsproblemen von Webdaten eine große Herausforderung.



Ein wesentlicher Untersuchungsschwerpunkt war die Realisierung und Evaluierung neuer adaptiver Suchstrategien, um anwendungsrelevante Objekte aus Datenquellen wie Entity-Suchmaschinen und Web-Datenbanken effektiv und effizient (d.h. mit minimalem Kommunikationsaufwand) abzurufen. Die Suchverfahren nutzen pro Datenquelle mehrere Query-Generatoren zur automatisierten Erzeugung unterschiedlicher Anfragen. Die Anfragetechnik wurde in ein neu entwickeltes Mashup-Framework namens WETSUIT (Web EnTity Search and fUsIon Tool) integriert und somit zur Erstellung von Datenintegrationsanwendungen verfügbar gemacht. WETSUIT stellt neben der Anfragegenerierung weitere mächtige Operatoren bereit, u.a. zum Objekt-Matching. WETSUIT unterstützt ferner die parallele, überlappende Ausführung mehrerer Operatoren, verbunden mit einem Streaming von Datenobjekten zwischen den Operatoren. Damit konnten mehrere anspruchsvolle und hoch interaktive Datenintegrationsanwendungen realisiert werden.

Link Discovery

E. Rahm, M. Nentwig, M. Hartung, Lars Kolb

Linked (Open) Data (LOD) verwendet offene Web-Standards wie RDF und HTTP zur Speicherung, Publikation und Verknüpfung von heterogenen Daten. Ein wichtiges Ziel ist die Identifikation von semantischen Zusammenhängen (Links) zwischen verschiedenen LOD-Datenquellen, da zwar sehr viele Daten veröffentlicht werden, diese jedoch nicht ausreichend miteinander verknüpft sind. Da manuelle Techniken aufgrund der Größe der Quellen meist zu aufwändig sind, werden (semi-)automatische Verfahren zum Abgleich der Quellen (Link Discovery) eingesetzt. Bisherige Verfahren erzeugen jedoch meist unvollständige oder ungenaue Mappings. Um eine Übersicht zu erhalten, wurden zunächst bestehende Ansätze zur (semi-) automatischen Erstellung von Links bzgl. ihrer

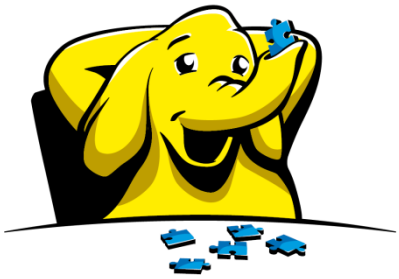
Stärken und Schwächen untersucht. Insbesondere wurden der allgemeine Ablauf des jeweiligen Ansatzes sowie die verwendeten Matching-Strategien und Suchraumoptimierungen verglichen. Weiterhin wurden Methoden vorgestellt, die anhand der Komposition existierender Links eine Bestimmung neuer Links zwischen LOD Datenquellen ermöglichen. Eine Evaluierung anhand realer Datensätze zeigte den hohen Wert der Wiederverwendung und Komposition sowie die hohe Effektivität der Verfahren.

Zudem wurde mit der Implementierung eines frei zugänglichen Portals zur zentralen Verwaltung neu berechneter Links aus verschiedenen Domänen begonnen. Ziel dieses Portals ist unter anderem eine Anreicherung von Mappings mit Metadaten wie dem Ersteller, Datum und verwendeten Algorithmus, um die Herkunft der Daten sowie deren Versionierung besser nachvollziehen zu können. Zukünftig soll das Portal u.a. als Datenquelle für den Vergleich von Benchmark-Ergebnissen verwendet werden.

In Kooperation mit der AKSW-Gruppe (Institut für Informatik, Universität Leipzig) wurde das Link Discovery Verfahren HR³ auf verschiedenen verteilten Rechnersystemen unter Ausnutzung von Graphical Processing Units (GPUs) und MapReduce-Plattformen implementiert. Eine umfangreiche Performance-Evaluierung bietet Richtlinien, um zu entscheiden, welche Hardware zur Realisierung von Link Discovery Aufgaben geeignet ist. Die Arbeit wurde mit dem *Best Paper Award* der 10. Extended Semantic Web Conference (ESWC 2013) ausgezeichnet.

Big Data Integration

E. Rahm, L. Kolb, Z. Sehili



Die Verwaltung der weltweit steigenden Datenmengen bedingt eine verteilte Speicherung und Auswertung in Clustern mit Tausenden von einzelnen Knoten. Als Forschungsschwerpunkt wurde in den vergangenen beiden Jahren die Nutzung des MapReduce-Konzepts zur automatischen Parallelisierung daten- und rechenintensiver „Big Data“-Anwendungen untersucht.

Im Zuge dieser Forschungsarbeiten entstand das Hadoop-basierte Datenintegrations-Framework Dedoop. Es erlaubt die High-Level Spezifikation komplexer Object Matching-Workflows, die anschließend automatisch in eine Menge von MapReduce Jobs überführt und auf verschiedenen Hadoop-Clustern zur Ausführung gebracht werden können. Dazu können verschiedene Blocking-Methoden zur Reduzierung der Kandidatenpaare selektiert werden, für die entsprechende MapReduce-Algorithmen entwickelt wurden. Wesentlicher Forschungsschwerpunkt war die Konzeption und Umsetzung von Strategien zur Lastbalancierung und zur Vermeidung redundanter Datensatzvergleiche um die zu Verfügung stehenden Ressourcen eines Rechnerclusters optimal ausnutzen zu können. Dedoop unterstützt die Verwendung von Cloud-Diensten wie Amazon S3 und Amazon EC2 um bei Bedarf eine Menge virtueller Maschinen in entfernten Rechenzentren starten und zur Berechnung nutzen zu können. Zuletzt wurde die Möglichkeit zur Ausführung iterativer Graphalgorithmen für große Datenmengen in Dedoop integriert. Aktueller Forschungsgegenstand ist Kombination der MapReduce-basierte Parallelisierung von Object Matching-Workflows mit der Beschleunigung von Ähnlichkeitsberechnungen durch massiv-parallele Graphikprozessoren (GPUs).

Graphbasierte Geschäftsdatenanalyse

E. Rahm, A. Petermann

In den vergangenen Jahren sind leistungsfähige Graphdatenbanken entstanden. Sie ermöglichen eine flexible Integration von heterogenen Geschäftsdaten und, darauf basierend, neuartige Verfahren zur Datenanalyse. Zwar existieren bereits Forschungsarbeiten über graphbasierte Lösungen für spezifische analytische Probleme, jedoch kein genereller Ansatz. Unser Ziel ist es, beginnend von der Integration von Quelldaten bis hin zur analytischen Benutzeroberfläche eine graphbasierte Ergänzung oder Alternative zu etablierten Data Warehouse-basierten Systemen zu untersuchen. Dafür wurde mit der Entwicklung des Frameworks BIIIG (Business Intelligence with Integrated Instance Graphs) begonnen. Die Datenobjekte einzelner Quellsysteme werden dabei weitgehend automatisiert in einem übergreifenden Instanzgraphen integriert und durch ein vereinheitlichtes Metadaten-Modell beschrieben. In dem integrierten Instanzgraphen können ferner Teilgraphen, die bestimmte Geschäftsvorgänge betreffen, automatisiert extrahiert und für fokussierte Analysen zugänglich gemacht werden. Damit lassen sich neben einfachen Aggregationen auch kausale Beziehungen er-

mitteln, z.B. welche Mitarbeiter besonders häufig und wie an erfolgreichen Projektakquisitionen beteiligt waren. Es existiert bereits ein erster Prototyp, basierend auf einer produktiven Graphdatenbank.

Schema- und Ontologie-Matching

E. Rahm, E. Peukert, A. Groß, M. Hartung, T. Kirsten

Das teil-automatisierte Matching von Schemata und Ontologien ist meist sehr aufwändig und erfordert die Kombination mehrerer Techniken zur Berechnung der syntaktischen, semantischen oder strukturellen Ähnlichkeit von Elementen. Wichtige Teilprobleme sind dabei die Auswahl geeigneter Matching-Algorithmen, deren Kombination und Konfiguration. Wir haben im Berichtszeitraum weiter an einem selbstkonfigurierenden Matching-System (AMC) gearbeitet, dass sich automatisch an ein gegebenes Mapping-Problem anpassen kann. Unser Ansatz basiert dabei auf der Analyse der Eingabe- Schemas und von Zwischenergebnissen bereits ausgeführter Matcher. Verschiedene Matching- Regeln nutzen die Analyseergebnisse, um automatisch einen Matching-Prozess zu konstruieren und anzupassen. Die Evaluation zeigt, dass unser System in der Lage ist, Mapping-Probleme aus verschiedenen Domänen in guter Qualität zu lösen. Der entwickelte Ansatz konnte auf der ICDE 2012-Konferenz veröffentlicht werden. Darüber hinaus wurden die in der Arbeit gewonnen Erkenntnisse im letzten Jahr als Dissertation eingereicht und im November 2013 erfolgreich verteidigt.

Das Matching großer Ontologien hat im Bereich der Lebenswissenschaften in den letzten Jahren stark an Bedeutung gewonnen. Häufig existieren mehrere Ontologien in der gleichen Sub-Domäne, welche zumindest teilweise überlappende Informationen enthalten. Ziel ist es, die Beziehungen zwischen den verschiedenen Ontologien zu bestimmen, um somit u.a. die Integration heterogener Quellen oder das Merging von Ontologien zu ermöglichen. Die Erzeugung qualitativ hochwertiger Matchergebnisse erfordert eine kombinierte Ausführung verschiedener Matcher, was jedoch eine sehr zeit- und speicher-intensive Berechnung darstellt. Innerhalb des Berichtszeitraumes wurde das Matching-System GOMMA zweimal zur Ontology Alignment Evaluation Initiative (OAEI 2011.5 und 2012) eingereicht. GOMMA (Generic Ontology Matching and Mapping Management) ist eine Infrastruktur zum Management und zur Evolutionsanalyse großer Ontologien und Mappings im Bereich der Lebenswissenschaften. Insbesondere wird das Matching sehr großer Ontologien durch skalierbare Matchtechniken ermöglicht (paralleles Ontologiematching, indirekte Berechnung von Ontologiemappings durch Wiederverwendung und Komposition existierender Mappings, Blocking zur Reduktion des Suchraums). In der OAEI 2012-Evaluierung erreichte GOMMA die besten Ergebnisse im Anatomy- und Library-Track. Insgesamt erzeugte GOMMA Mappings von hoher Qualität innerhalb guter Laufzeiten. Detaillierte Ergebnisse werden auf der OAEI-Ergebnisseite zur Verfügung gestellt:
http://oaei.ontologymatching.org/2012/results/index.html

GOMMA wurde weiterhin um eine neuartige „Multi-part“-Match-Strategie erweitert und für das Matching der klinischen Ontologie LOINC mit anderen in Krankenhäusern eingesetzten Labor-Terminologien evaluiert. Das Verfahren bezieht Kombinationshäufigkeiten einzelner LOINC-Konzeptteile ein und erzeugte Mappings von sehr hoher Qualität. Darüber hinaus wurde die Verwendung von GPUs (Graphical Processing Units) für eine massiv parallele Ausführung von String-Matching-Algorithmen realisiert. Eine Evaluierung anhand sehr großer biomedizinischer Ontologien zeigte signifikante Verbesserungen bzgl. der Ausführungszeiten aufgrund der Verwendung von GPUs.

Evolution von Ontologien und Mappings

E. Rahm, A. Groß, M. Hartung, T. Kirsten

Ontologien werden u.a. in den Lebenswissenschaften zur eindeutigen semantischen Beschreibung (Annotation) von Objekten wie z.B. Proteinen oder Genen eingesetzt. Aufgrund neuer wissenschaftlicher Erkenntnisse und veränderter Anforderungen werden die Ontologien ständig modifiziert. Derartige Änderungen haben wiederrum Auswirkungen auf abhängige Datenquellen, Mappings und Anwendungen, so dass diese aktualisiert werden müssen. Um mit der Evolution der Ontologien umgehen zu können, besteht ein wichtiger Schritt in der Bestimmung der Differenz (des DIFF) zwischen zwei Versionen einer Ontologie. Im Berichtszeitraum wurde der COnto-Diff (Complex Ontology Diff) Ansatz zur Bestimmung eines vollständigen und ausdrucksstarken Diff-Evolution-Mappings zwischen Ontologieversionen weiterentwickelt. Insbesondere erlaubt die webbasierte Applikation CODEX (Complex Ontology Diff Explorer) eine interaktive Berechnung und

Analyse von DIFFs für zahlreiche Ontologien in den Lebenswissenschaften. Zudem wurden mit der Entwicklung der Webapplikation REx beonnen, welche eine Untersuchung stark veränderlicher und stabiler Ontologieregionen erlauben soll.

Unter Verwendung von COnto-Diff wurde der Einfluss von Ontologieänderungen auf Ontologiemappings untersucht. Insbesondere konnten die Änderungsoperationen des Diffs zur Adaptierung veralteter Ontologiemappings genutzt werden. Es wurden zwei generische Algorithmen zur (semi-) automatischen Adaptierung von Ontologiemappings entwickelt. Ein Ansatz basiert auf der Komposition von Ontologiemappings, wohingegen der andere Ansatz Diff-Evolution-Mappings zur individuellen Adaptierung der Mappings nutzt. Beide Verfahren ermöglichen die Wiederverwendung unbeeinflusster, bereits bestätigter Mappingteile und vermeiden eine vollständige Neubestimmung bestehender Mappings. Eine Evaluierung für sehr große, biomedizinische Ontologien und Mappings zeigte, dass beide Verfahren qualitativ hochwertige Ergebnisse produzieren. Unter Verwendung der an der Abteilung entwickelten Algorithmen und Prototypen wurden in einer Kooperation mit dem Max-Planck-Institut für evolutionäre Anthropologie die Auswirkungen der Ontologie- und Annotationsevolution auf die Ergebnisse sogenannter funktionaler Analysen großer biologischer Datensätze erforscht. Es konnte gezeigt werden, dass sich Analyseergebnisse durchaus ändern, jedoch sind die untersuchten statistischen Verfahren insgesamt relativ robust gegenüber der Ontologie- und Annotationsevolution.

Bibliometrische Analysen

E. Rahm, D. Aumüller, A. Thor

Bibliometrische Analysen untersuchen wissenschaftliche Publikationen hinsichtlich ihrer Zitierungshäufigkeiten sowie den üblichen bibliografischen Angaben wie z.B. Autoren, Forschungsinstitut und Publikationsorgan. Evaluierungen ermöglichen u.a. die Erstellung von Rankings der meistzitierten Arbeiten pro Autor oder Publikationsorgan sowie aggregierte Kennzahlen (z.B. h-Index) zur vergleichenden Analyse. Weiterhin lassen sich Auswertungen für Forschergruppen oder Institute durchführen. Hierzu wurden entsprechende Anwendungen entwickelt, die ausgehend von einer Publikationsdatenbank mit bibliografischen Metadaten (u.a. Titel, Autoren, Publikationsorgan) sowie auswertungsbezogenen Eingabedaten die bibliometrischen Daten aus unterschiedlichen Quellen automatisiert ermitteln und konsolidieren.

Im Berichtszeitraum erfolgte u.a. die bibliometrische Impact-Evaluierung eines ausländischen Informatik-Instituts, bezogen auf alle Publikationen der letzten fünf Jahre. Zudem wurde eine vergleichende Evaluierung solcher Quellen (u.a. Google Scholar und Web of Science) begonnen, die wegen ihrer unterschiedlichen Datenakquise (u.a. automatisches Web-Crawling bei Google Scholar vs. Indizierung ausgewählter Publikationsorgane bei Web of Science) unterschiedliche Zitierungshäufigkeiten für dieselbe Publikation ermitteln. Dabei soll ermittelt werden, ob trotz unterschiedlicher Werte zur Zitierungshäufigkeit dennoch vergleichbare Rankings von Publikationslisten (z.B. alle Publikationen eines Autors oder eines Zeitschriftenjahrgangs) resultieren.

Zudem wurden – in Kooperation mit der Hochschule Wallis und der Universität Kopenhagen - Scholar-Zitierungszahlen zur Impact-Analyse der CLEF-Tagungsreihe genutzt.

5. Veröffentlichungen / Graduierungsarbeiten

Zeitschriften

- Groß, A.; Hartung, M.; Prüfer, K.; Kelso, J.; Rahm, E.: *Impact of Ontology Evolution on Functional Analyses*. Bioinformatics 28 (20): 2671-2677, 2012
- Hartung, M.; Groß, A.; Rahm, E.: *CODEX: Exploration of semantic changes between ontology versions*. Bioinformatics 28 (6): 895-896, 2012
- 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
- Kolb, L.; Rahm, E.: *Parallel Entity Resolution with Dedoop*. Datenbank-Spektrum 13 (1), 2013
- Kolb, L.; Thor, A.; Rahm, E.: *Multi-pass Sorted Neighborhood Blocking with MapReduce*. Computer Science - Research and Development 27(1), 2012
- Lee, L. H.; Groß, A.; Hartung, M.; Liou, D. M.; Rahm, E.: *A Multi-Part Matching Strategy for Mapping LOINC with Laboratory Terminologies*. Journal of the American Medical Informatics Association (JAMIA), published online first, 2013
- Ouzzani, M.; Papotti, P.; Rahm, E.: *Introduction to the special issue on data quality*. Information Systems. Vol. 38(6): 885-886, 2013
- Rahm, E.: *Der Lehrstuhl Datenbanken an der Universität Leipzig*. Datenbank-Spektrum 13 (2), 2013
- Raunich, S.; Rahm, E.: *Target-driven Merging of Taxonomies with ATOM*. Information Systems, Vol. 42, 2013

Proceedings

- Anderson, P.; Thor, A.; Benik, J.; Raschid, L.; Vidal, M.-E.*: PAnG - Finding Patterns in Annotation Graphs*. Proc. Intl. Conference on Management of Data (SIGMOD), 2012 (demo paper), 2012
- Arnold, P.: *Semantic Enrichment of Ontology Mappings: Detecting Relation Types and Complex Correspondences*. 25. GI-Workshop Grundlagen von Datenbanken, 2013
- Arnold, P.; Rahm, E.: *Semantic Enrichment of Ontology Mappings: A Linguistic-based Approach*. 17th East-European Conference on Advances in Databases and Information Systems (ADBIS), 2013
- Benik, J.; Chang, C.; Raschid, L.; Vidal, M. E.; Palma, G.; Thor, A.: *Finding Cross Genome Patterns in Annotation Graphs*. Proc. 8th Intl. Conference on Data Integration in the Life Sciences (DILS), 2012
- Benik, J.; Palma, G.; Raschid, L.; Thor, A.; Vidal, M. E.: *Mining Patterns from Clinical Trial Annotated Datasets by Exploiting the NCI Thesaurus*. Proc. 11th Intl. Semantic Web Conference (ISWC), Demo, 2012
- Christen, V.: *REx – eine Webapplikation zur Visualisierung der Evolution von Ontologien in den Lebenswissenschaften*. Studentenkonferenz Informatik Leipzig (SKIL), 2012
- Endrullis, S.; Thor, A.; Rahm, E.: *Entity Search Strategies for Mashup Applications*. Proc. 28th Intl. Conference on Data Engineering (ICDE), 2012
- Endrullis, S.; Thor, A.; Rahm, E.: *WETSUIT: An Efficient Mashup Tool for Searching and Fusing Web Entities*. Proc. 38th Intl. Conference on Very Large Databases (VLDB) / Proc . of the VLDB Endowment 5(12), 2012 (demo), 2012
- Groß, A.; Dos Reis, J.C.; Hartung, M.; Pruski, C.; Rahm, E.: *Semi-Automatic Adaptation of Mappings between Life Science Ontologies*. Proc. 9th Intl. Conference on Data Integration in the Life Sciences (DILS), 2013
- Groß, A.; Hartung, M.; Kirsten, T.; Rahm, E.: *GOMMA Results for OAEI 2012*. Seventh International Workshop on Ontology Matching @ ISWC, 2012
- Groß, A.; Hartung, M.; Thor, A.; Rahm, E.: *How do computed ontology mappings evolve? - A case study for life science ontologies*. Joint Workshop on Knowledge Evolution and Ontology Dynamics @ ISWC, 2012
- Hartung, M.; Groß, A.; Kirsten, T.; Rahm, E.: *Effective Composition of Mappings for Matching Biomedical Ontologies*. ESWC 2012 Workshops, Revised Selected Papers (LNCS), 2012
- Hartung, M.; Groß, A.; Kirsten, T.; Rahm, E.: *Effective Mapping Composition for Biomedical Ontologies*. Semantic Interoperability in Medical Informatics @ ESWC, 2012
- Hartung, M.; Groß, A.; Rahm, E.: *Composition Methods for Link Discovery*. Proc. of 15. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), 2013
- Hartung, M.; Groß, A.; Rahm, E.: *Determining and Analyzing Semantic Ontology Changes with CODEX*. Demo @ Intl. Conference on Data Integration in the Life Sciences (DILS), 2012
- Hartung, M.; Kolb, L.; Groß, A.; Rahm, E.: *Optimizing Similarity Computations for Ontology Matching - Experiences from GOMMA*. Proc. 9th Intl. Conference on Data Integration in the Life Sciences (DILS), 2013
- Kolb, L.; Thor, A.; Rahm, E.: *Dedoop: Efficient Deduplication with Hadoop*. Proc. 38th Intl. Conference on Very Large Databases (VLDB) / Proc. of the VLDB Endowment 5(12), 2012
- Kolb, L.; Thor, A.; Rahm, E.: *Don't Match Twice: Redundancy-free Similarity Computation with MapReduce*. Proc. 2nd Intl. Workshop on Data Analytics in the Cloud (DanaC), 2013
- Kolb, L.; Thor, A.; Rahm, E.: *Load Balancing for MapReduce-based Entity Resolution*. Proc. 28th Intl. Conference on Data Engineering (ICDE), 2012

- Köpcke, H.; Thor, A.; Thomas, S.; Rahm, E.: *Tailoring entity resolution for matching product offers*. Proc. 15th Intl. Conference on Extending Database Technology (EDBT), 2012
- Ngonga Ngomo A.-C.; Kolb, L.; Heino, N.; Hartung, M.; Auer, S.; Rahm, E.: *When to Reach for the Cloud: Using Parallel Hardware for Link Discovery*. Proc. 10th Intl. Extended Semantic Web Conference (ESWC), 2013
- Palma, G.; Vidal, M. E.; Haag, E.; Raschid, L.; Thor, A.: *Measuring Relatedness Between Scientific Entities in Annotation Datasets*. ACM International Conference on Bioinformatics, Computational Biology, and Biomedical Informatics (BCB), 2013
- Palma, G.; Vidal, M.E.; Raschid, L.; Thor, A.: *Exploiting Semantics from Ontologies and Shared Annotations to Find Patterns in Annotated Linked Open Data*. 3rd International Workshop on Linked Science (LISC@ISWC), 2013
- Peukert, E.; Eberius, J.; Rahm, E.: *A Self-Configuring Schema Matching System*. Proc. 28th Intl. Conference on Data Engineering (ICDE), 2012
- Pfeifer, K.; Peukert, E.: *Mapping Text Mining Taxonomies*. Proc. 6th International Conference on Knowledge Discovery and Information Retrieval (KDIR) , 2013
- Raschid, L.; Palma, G.; Vidal, M.E.; Thor, A.: *Exploration Using Signatures in Annotation Graph Datasets*. AAAI 2013 Fall Symposium Series (Discovery Informatics: AI Takes a Science-Centered View on Big Data), 2013
- Raunich, S.; Rahm, E.: *Towards a Benchmark for Ontology Merging*. Proc. 7th OTM Workshop on Enterprise Integration, Interoperability and Networking (EI2N'2012), Springer LNCS, 2012
- Tsikrika, T.; Larsen, B.; Müller, H.; Endrullis, S.; Rahm, E.: *The Scholarly Impact of CLEF (2000-2009)*. Proc. Conference and Labs of the Evaluation Forum (CLEF), LNCS 8138, 2013

Technische Berichte und Poster

- Groß, A.; Hartung, M.; Thor, A.; Rahm, E.: How do Ontology Mappings Change in the Life Sciences? Selected Poster @ Intl. Conference on Data Integration in the Life Sciences (DILS), 2012
- Groß, A.; Hartung, M.; Thor, A.; Rahm, E.: How do Ontology Mappings Change in the Life Sciences? CoRR abs/1204.2731, 2012

Dissertationen

- Peukert, E.: Process-based Schema Matching: From Manual Design to Adaptive Process Construction, Univ. Leipzig, 2012

November 2013 - nach der erfolgreichen Verteidigung von Eric Peukert

Bachelor-, Master- und Diplomarbeiten

1. Chill, S.: Realisierung einer Datenbank zur Erfassung von PA-Fragebögen und Matching zur ICF, Bachelorarbeit, Univ. Leipzig, 2013
2. Christen, V.: Visualisierung der Evolution von Ontologien in den Lebenswissenschaften, Bachelorarbeit, Univ. Leipzig, 2012
3. Do, V.H.: Automatische Wiederverwendung auf dem Element-Level im Kontext von Schema Matching, Masterarbeit, Univ. Leipzig, 2012
4. Fischer, A.: Implementierung eines File Managers für das Hadoop Distributed Filesystem und Realisierung einer MapReduce Workflow Submission-Komponente, Bachelorarbeit, Univ. Leipzig, 2012

5. Häberlein, D.: Migration und Extraktion von Datensätzen mittels spaltenorientierter Datenbanken am Beispiel von Apache HBase, Bachelorarbeit, Univ. Leipzig, 2013
6. Jacob, M.: Untersuchung von Hintergrundwissen zur Verbesserung von semantischen Mappings, Bachelorarbeit, Univ. Leipzig, 2013
7. Kubitzky, S.: Übersicht über Crowdsourcing-Ansätze und Plattformen zur Beurteilung von Matchergebnissen, Univ. Leipzig, 2013
8. Lucke, P.: Effektive Link Discovery für veränderliche LOD Quellen, Bachelorarbeit, Univ. Leipzig, 2013
9. Nentwig, M.: SPSS Modeler Integration mit IBM DB2 Analytics Accelerator, Masterarbeit, Univ. Leipzig, 2012
10. Sehili, Z.: Evaluierung und Erweiterung von MapReduce-Algorithmen zur Berechnung der transitiven Hülle ungerichteter Graphen für Entity Resolution Workflows, Masterarbeit, Univ. Leipzig, 2013
11. Sintschilin, S.: Wiederverwendung berechneter Matchergebnisse für MapReduce-basiertes Object Matching, Bachelorarbeit, Univ. Leipzig, 2013
12. Stehmann, F.: Automatische Erkennung von Plagiaten im Internethandel, Bachelorarbeit, Univ. Leipzig, 2013
13. Tran, N.H.: Erstellung eines Benchmarks zu augewaehlten Match-Problemen im Linked Open Data, Bachelorarbeit, Univ. Leipzig, 2013
14. Uhlich, R.: Erstellung von praktischen Lehrbeispielen im Rahmen eines Datenbanken-Seminars für Nicht-Informatiker, Bachelorarbeit, Univ. Leipzig, 2012

Vorträge

- Arnold, P.: Semantic Enrichment of Ontology Mappings, A Linguistic-based Approach. 17th East-European Conference on Advances in Databases and Information Systems (ADBIS), Genoa (Italy), 2013
- Arnold, P.: Semantic Enrichment of Ontology Mappings: Detecting Relation Types and Complex Correspondences. 25. GI-Workshop Grundlagen von Datenbanken, Elgersburg (Germany), 2013
- Endrullis, S.: Entity Search Strategies for Mashup Applications. Intl. Conference on Data Engineering (ICDE), Washington (USA), 2012
- Endrullis, S.: WETSUIT - An Efficient Mashup Tool for Searching and Fusing Web Entities. Data Integration Day @ SABRE, Leipzig, 2012
- Groß, A.: On^{2}Vers - Online Ontology Versioning. 2nd eScience Network Conference, Dresden, 2013
- Groß, A.: GOMMA Results for OAEI 2012. Seventh International Workshop on Ontology Matching (OM) at ISWC, Boston (USA), 2012
- Groß, A.: How do computed ontology mappings evolve? - A case study for life science ontologies. Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn) at ISWC, Boston (USA), 2012
- Groß, A.: How do Ontology Mappings Change in the Life Sciences? Eighth International Conference on Data Integration in the Life Sciences (DILS), College Park, Maryland (USA), 2012
- Hartung, M.: Composition Methods for Link Discovery. 15. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), Magdeburg, 2013
- Hartung, M.: Effective Mapping Composition for Biomedical Ontologies. Semantic Interoperability in Medical Informatics @ ESWC, Heraklion (Greece), 2012
- Hartung, M.: Effective Mapping Composition for Link Discovery. Data Integration Day @ SABRE, Leipzig, 2012
  Hartung, M.: Large-scale ontology matching on massively parallel hardware (Keynote). 1. Workshop “Data Management in the Cloud” @ BTW, Magdeburg 2013
- Hartung, M.: Optimizing Similarity Computations for Ontology Matching - Experiences from GOMMA. 9th International Conference on Data Integration in the Life Sciences (DILS), Montreal (Canada), 2013
- Kolb, L.: Dedoop: Efficient Deduplication with Hadoop. Data Integration Day @ SABRE, Leipzig, 2012
- Kolb, L.: Don't Match Twice: Redundancy-free Similarity Computation with MapReduce. Workshop on Data Analytics in the Cloud (DanaC), New York (USA), 2013
- Peukert, E.: A Self-Configuring Schema Matching System, 28th Intl. Conference on Data Engineering (ICDE), Washington D.C. (USA), 2012.
- Rahm, E.: Scalable Matching of Real-World Data. Keynote VLDB Workshop on Quality in databases (QDB), Istanbul (Turkey), Aug. 2012
- Rahm, E.: Big Data und Datenintegration. Transfer-Meeting der Univ. Leipzig, Juli 2013
- Rahm, E.: Big Data Analytics: Herausforderungen und Systemansätze. Symposium für neue IT, Leipzig, Sep. 2013
- Rahm, E.: Datenintegration für Big Data. Datenbank-Stammtisch, HTW Dresden, Okt. 2013

- Wartner, C.: Erkennung von Produktplagiaten in Online-Shops. Data Integration Day @ SABRE, Leipzig, 2012

Mitgliedschaften in Gremien/Redaktionskollegien, Herausgebergremien u.ä.

Rahm, E.:

- Stv. Sprecher des Fachbereichs "Datenbanken und Informationssysteme" der Gesellschaft für Informatik
- Steering Committee DILS conference series (Data Integration in the Life Sciences)
- Advisory Board Europar conference series
- Organisator des ersten Data Integration Day, Workshop im Rahmen der SABRE conference, Leipzig, Sep. 2012
- Gasteditor Information Systems (Special issue on Data Quality, 2013)
- PC-Chair DILS 2014
- Programmkomitee verschiedener Konferenzen (u.a. SIGMOD 2012, ICDE 2012, EDBT 2013, BTW 2013, DILS 2012, WebDB 2013)
- Gutachter für diverse Zeitschriften und Forschungsgesellschaften
- Vorstandsmitglied IZBI (Interdisziplinären Zentrum für Bioinformatik, Leipzig)
- Vorsitzender Prüfungsausschuß Informatik

Hartung, M.:

- Programmkomiteemitglied: IDC 2012/2013, SKIL 2012, Bio-Ontologies 2013
- Gutachter für Zeitschriften: u.a. BMC Bioinformatics, PLOS ONE, Journal of Biomedical Semantics
- Mitglied im Fakultätsrat

Thor, A.:

- Programmkomitee verschiedener Konferenzen (SCDM 2013, ESWC 2012/2013, DOA-Trusted Cloud 2013, AICCSA 2013, ReD 2012/2013)
- Workshop Chair Data Management in the Cloud (DMC), 2013
- Gast-Editor Datenbankspektrum 2/2014 (Datenmanagement in der Cloud)