Forschungsbericht 2010/2011

Abteilung Datenbanken

Universität Leipzig, Institut für Informatik

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

Inhaltsüberblick

1. Personelle Zusammensetzung
2. Highlights
3. WDI-Lab
4. Weitere Projekte
5. Veröffentlichungen / Graduierungsarbeiten
6. Vorträge
7. Mitgliedschaften in Gremien / Redaktionskollegien, Herausgebergremien u. ä.

Abt. Datenbanken im Mai 2011. V.l.n.r. (vorne): Prof. Dr. E. Rahm, Dr. M. Hartung, L. Kolb, A. Chyhir (Studentin), A. Groß, K. Wurdinger, S. Maßmann, H. Köpcke, J. Veshcheva (Studentin); V.l.n.r. (hinten): F. Zhang, C. Wartner, S. Endrullis, E. Peukert, Dr. S. Raunich, D. Aumüller, Dr. A. Thor, S. Kitschke, P. Arnold (Student), T. Gröger (Student).

1. Personelle Zusammensetzung

Univ.-Professor	Prof. Dr. Rahm, Erhard
Wiss. Mitarbeiter (BMBF)	Dr. Algergawy, Alsayed (bis April 2011)
Wiss. Mitarbeiter (BMBF)	Aumüller, David
Wiss. Mitarbeiter (DFG)	Endrullis, Stefan
Wiss. Mitarbeiterin (DFG)	Groß, Anika
Wiss. Mitarbeiter	Dr. Hartung, Michael
Wiss. Mitarbeiter (BMBF)	Heller, Nico
Sekretärin	Hesse, Andrea
Programmierer	Jusek, Stefan
Wiss. Mitarbeiter (IZBI/IMISE)	Dr. Kirsten, Toralf
Wiss. Mitarbeiter (BMBF)	Kitschke, Sven bis Sep. 2011)
Wiss. Mitarbeiter (BMBF)	Kropp, Henning (seit Sep. 2011)
Wiss. Mitarbeiterin (BMBF)	Köpcke, Hanna
Wiss. Mitarbeiterin (BMBF)	König, Kathleen (bis Jan. 2011)
Wiss. Mitarbeiter	Kolb, Lars
Wiss. Mitarbeiterin (BMBF)	Maßmann, Sabine
Doktorand	Peukert, Eric
Wiss. Mitarbeiter (BMBF)	Dr. Raunich, Salvatore (seit April 2010)
Wiss. Mitarbeiterin (BMBF)	Röllig, Carina
Wiss. Mitarbeiter	Dr. Sosna, Dieter (bis Feb. 2010)
Wiss. Mitarbeiter	Dr. Thor, Andreas
Wiss. Mitarbeiter (BMBF)	Wartner, Christian
Wiss. Mitarbeiterin (BMBF)	Wurdinger, Kerstin
Wiss. Mitarbeiter (BMBF)	Zhang, Fan

2. Highlights

Im Berichtszeitraum 2010/2011 sind folgende Ereignisse hervorzuheben:

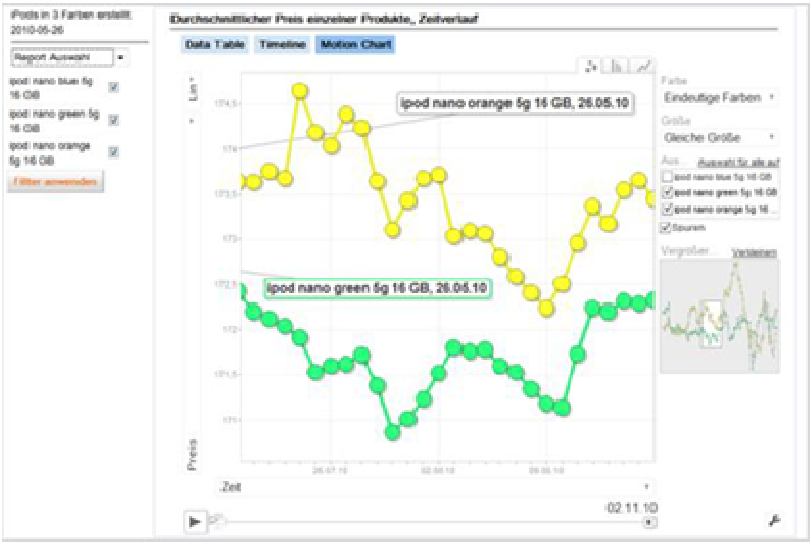
1. Seit Januar 2010 ist das **WDI-Lab** (Web Data Integration Lab) Teil der Abteilung. Es wurde vom BMBF mit ca. 1,4 Millionen Euro gefördert und beinhaltet zehn Stellen für wissenschaftliche Mitarbeiter. In ihm werden vorliegende Forschungsprototypen für den Markteinsatz weiter entwickelt.
2. Ende 2011 wurde mit der Vorbereitung eines WDI-Lab Spinoffs begonnen, der Web Data Solutions GmbH. In ihm werden die im WDI Lab entwickelten Datenintegrationslösungen kommerziell vertreiben.
3. Das Bundesministerium für Wirtschaft und Technologie bewilligte ein Exist-Gründerstipendium an Dr. Nick Golovin zum Thema „Data Virtualizer“ (Mentor: Prof. Rahm). Die Vorbereitungen für eine weitere Ausgründung sind auf einem guten Weg.
4. Auf der VLDB 2011 wurde Prof Rahm mit seinen Koautoren J. Madhavan und Phil Bernstein mit dem renommierten **VLDB Ten Year („Test of Time“) Best Paper Award** ausgezeichnet. Sie gaben eine Keynote zum Thema „Generic Schema Matching – Ten Years Later“.
5. Im Berichtszeitraum wurden zwei Dissertationen verteidigt (Nick Golovin, Michael Hartung)
6. Prof. Rahm veröffentlichte als Mitherausgeber ein neues Buch zu „Schema Matching and Mapping“ im Springer-Verlag
7. Das Oberseminar der Abteilung fand im Juni 2010 sowie im Mai 2011 bereits zum neunten bzw. zehnten Mal an der Uni-Außenstelle in Zingst/Ostsee statt.
8. Dr. Andreas Thor absolvierte zwischen 1/2010 und 04/2011 einen durch die DFG geförderten Forschungsaufenthalt an der University of Maryland, College Park, USA.
9. Es wurde ein EU-Projekt LinkedDesign in Kooperation mit: SAP Research eingeworben.
10. Die Fa. Amazon bewilligte der Abteilung einen Grant zur Nutzung ihrer Cloud-Ressourcen

Übergabe des VLDB Ten Year Best Paper Awards auf der VLDB 2011 (links), Segeltour im Rahmen des Zingst2010-Workshops (rechts)

3. WDI-Lab

E. Rahm, A. Algergawy, D. Aumüller, N. Heller, S. Kitschke, H. Kropp, H. Köpcke, K. König, S. Maßmann, S. Raunich, C. Röllig, C. Wartner, K. Wurdinger, F. Zhang

Im BMBF-geförderten WDI-Lab wurde in einem Team von zehn wissenschaftlichen Mitarbeitern intensiv an der Entwicklung von einsatzreifen Verfahren und Werkzeugen zur semantischen Integration von Daten aus dem Web und aus Unternehmen gearbeitet. Diese semantische Integration der Daten ist aufgrund der Heterogenität und Volumina der Daten nur schwer zu lösen, jedoch für zahlreiche Anwendungsfälle von großer Bedeutung. Ein Schwerpunkt der Arbeiten war die Realisierung von Werkzeugen zur workflowbasierten Datenintegration, welche eine schnelle und kostengünstige Realisierung spezifischer Integrationsanwendungen ermöglicht. Zur Definition ganzer Datenintegrationsworkflows wurde das **Ombat-System** geschaffen, welches über die Ombat Matching Language (OML) die flexible Erstellung individueller Workflows erlaubt. Als Anwendungen wurde u.a. die Monitoring-Lösung **PROOF** zur Überwachung von Produkt- und Flugpreisen realisiert. Zudem wurde ein Fahrradangebote unterschiedlicher Händler und Anbieter in ein gemeinsames Portal integriert und regelmäßig aktualisiert.



Screenshot von PROOF (Product Offer Fusor) zum Preis-Monitoring zweier MP3-Player

Die semantische Integration ist auf zwei Ebenen zu leisten: zum einen auf Ebene von Metadaten, zum anderen auf Ebene der Dateninhalte. Erstere erfordert einen Abgleich von Metadaten, um inhaltliche Entsprechungen zu finden. Die Bestimmung solcher Korrespondenzen wird als Schema- bzw. Ontologie-Matching bezeichnet und wird mithilfe des COMA Systems umgesetzt, das im letzten Jahr hinsichtich Leistung und Funktionalität verbessert wurde (neue Version **COMA 3.0**). Eine in Coma neu entwickelte Funktionalität ist die Möglichkeit, Ontologien zu mischen bzw. zu integrieren (ATOM-Ansatz, s.u.).

Zusätzlich ist die Integration auf Ebene der Dateninhalte notwendig, um äquivalente Objekte zu identifizieren (**Objekt-Matching**). In diesem Bereich wurden flexible Bibliotheken entwickelt, die alle erforderlichen Schritte des Match-Prozesses, u.a. Dateninput und -output, Vorverarbeitung, Blocking-Strategien, Matchalgorithmen sowie Lernverfahren, unterstützen und in Ombat-Workflows verwendbar sind. Insbesondere stehen verschiedene maschinelle Lernverfahren und Verfahren zur automatisierten Auswahl von Trainingsdaten zur Verfügung, um Matcher im Rahmen eines Match-Workflows möglichst automatisiert auszuwählen, zu kombinieren und zu konfigurieren. Zusätzlich wurden komparative Evaluierungen hinsichtlich Effektivität und Effizienz verschiedener existierender Objekt-Matching-Verfahren durchgeführt. Ein weiterer Schwerpunkt lag auf der Behandlung von Produktdaten, deren Matching wegen ihrer Heterogenität sowie oft schlechten Datenqualität eine besondere Herausforderung darstellen. Es wurden daher entsprechende Methoden zur Datenbereinigung entwickelt, u.a. zur Vereinheitlichung von Herstellernamen sowie zur Extraktion herstellerspezifischer Produktcodes. Weiterhin wurde ein Ansatz zur semi-automatischen Kategorisierung von Produkten entwickelt und prototypisch umgesetzt (**Online Product Manager**). Die Ergebnisse werden von Nutzern bewertet und ggf. korrigiert. Dieses Feedback wird einem Lernverfahren zur Verbesserung weiterer Kategorisierungen zur Verfügung gestellt.

Die Ergebnisse des WDI-Labs wurden in mehreren Messebesuchen (u.a. Cebit 2011), Tagungen und zwei selbst organisierten Workshops (Okt. 2010, Dez. 2011) der Öffentlichkeit präsentiert.

WDI-Lab während auf der CEBIT2011 in Hannover (links: Messestand, rechts: Vortrag von H. Köpcke)

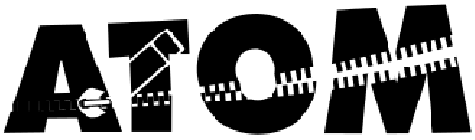
Web Data Integration Workshop 2010 (links), wöchentliches Gruppen-Meeting im WDI-lab (rechts)

4. Weitere Projekte

Ontologie- und Schema-Merging

E. Rahm, S. Raunich

Die Verbreitung von Ontologien und Taxonomien in vielen Domänen führt zu einem wachsenden Bedürfnis, diese zu integrieren. Das Ziel der Ontologie-Integration ist das Zusammenfügen von mindestens zwei Ontologien (mergen genannt), um eine einheitliche Ansicht auf diese zu bieten, wobei alle Informationen der Ausgangsontologien erhalten bleiben sollen. Es wurde ein neuartiger Taxonomie-Merging-Algorithmus entwickelt, der als Eingabe zwei Taxonomien und ein Äquivalenz-Mapping zwischen diesen verarbeitet und automatisch eine integrierte Taxonomie erzeugt. Der Ansatz wird durch die Zieltaxonomie bestimmt (target-driven), das heißt, dass die Ausgangstaxonomie in diese eingefügt wird und dabei die Struktur der Zielontologie so weit wie möglich bewahrt wird. Es wurde darüber hinaus untersucht, inwieweit man den Algorithmus um zusätzliche Informationen, wie z.B. Beziehungen zwischen Ausgangs- und Zielkonzepten, erweitern kann, um so das Endergebnis semantisch zu verbessern. Das gesamte Verfahren wurde im System ATOM (Automatic Target-driven Ontology Merging) prototypisch implementiert und anhand von Szenarien verschiedener Art und Größe evaluiert. Dabei konnten selbst für große Ontologien aus den Lebenswissenschaften effizient integrierte Ontologien berechnet werden. Der Merging-Algorithmus wurde in die neue Version von COMA integriert, so dass erzielte Matchergebnisse direkt zum Mischen von Ontologien verwendet werden können.



Dynamische Fusion verteilter Webdaten

E. Rahm, S. Endrullis, H. Köpcke, 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. Wichtige Teilprobleme dabei sind die Bestimmung von Anfragestrategien bei der Nutzung von Webdatenquellen (z.B. Suchmaschinen, Datenbanken) sowie der semantisch korrekte Abgleich zwischen heterogenen Objektbeschreibungen (Objekt-Matching). Im Rahmen dieses DFG-Projektes wird ein flexibles, service-orientiertes Framework zur Unterstützung der dynamischen Fusion verteilter Webdaten entwickelt und evaluiert, dass sowohl Query-Dienste für den Zugriff auf Datenquellen als auch Objekt-Matching-Dienste für den Abgleich von Objektbeschreibungen bereitstellt. Zusätzlich sollen Self-Tuning-Verfahren zur möglichst automatisierten Auswahl und Konfiguration von Query- und Match-Diensten untersucht werden. Innerhalb des Berichtszeitraums wurde hierzu ein Framework zur Definition und Evaluation von Anfragestrategien für (Entitäts-)Suchmaschinen entwickelt, deren Ziel es ist, die Suche nach einer Menge von Entitäten effektiver und effizienter zu gestalten, d.h. möglichst viele relevante Suchergebnisse in möglichst kurzer Zeit zu erhalten. Zur Verbesserung des Suchergebnisses kann eine Anfragestrategie verschiedene Verfahren zur Query-Generierung (sogenannte Query-Generatoren) nutzen, die sich abhängig von der Eingabemenge effizienter / effektiver oder weniger effizient / effektiv erweisen. Aus diesem Grund wurden verschiedenen Anfrage-Strategien untersucht, darunter auch adaptive Verfahren, die die Auswahl der Query-Generatoren dynamisch unter Berücksichtigung bestimmter Eigenschaften der Eingabemenge und einer automatischen Bewertung der Query-Generatoren vornehmen. Die Praxistauglichkeit der Anfrage-Strategien konnte exemplarisch in den Bereichen der Produktsuche und der Suche nach wissenschaftlichen Publikationen nachgewiesen werden. Im Folgenden sollen die hier entwickelten Anfrage-Strategien nun in ein Mashup-Framework integriert werden.

Cloud-basiertes Objekt-Matching

E. Rahm, L. Kolb

Cloud Computing bezeichnet u.a. die Bereitstellung von IT-Infrastruktur durch externe Dienstleister. In den letzten Jahren hat dieses Konzept die IT-Welt massiv verändert. Durch die Möglichkeit, bei Bedarf in sehr kurzer eine große Menge von Rechenleistung, Speicherplatz und Bandbreite ohne Vorabinvestitionen nutzen zu können, entsteht die Illusion unendlicher on-demand verfügbarer Ressourcen. Gleichzeitig steigt in Unternehmen die Menge der zu verwaltenden und zu analysierenden Daten stetig an. Die Verwaltung solcher Datenmengen bedingt eine verteilte Speicherung und Auswertung in Clustern mit Tausenden von einzelnen Knoten. Als

Forschungsschwerpunkt wurde begonnen, die Nutzung des MapReduce-Konzepts zur automatischen Parallelisierung großer Objekt-Matching- Aufgaben zu untersuchen. Es wurde begonnen, ein MapReduce-basiertes Framework zur verteilten Berechnung von Match-Workflows zu implementieren. Dazu wurden zunächst typische Blocking-Methoden wie z.B. das Sorted Neighborhood-Verfahren in MapReduce umgesetzt sowie verschiedene Methoden zum Vergleich einzelner Objekte integriert. Darauf aufbauend, wurden verschiedene Probleme des MapReduce-basierten Object Matchings, wie Lastbalancierung bei Datenungleichverteilung und Speicherbehandlung untersucht. Die entwickelten Lösungsansätze sind, unabhängig vom Object Matching, für jede Art paarweiser Ähnlichkeitsberechnung anwendbar. Im Folgenden wurden Möglichkeiten zum parallelen Lernen eines Machine-Learning-Modelles sowie dessen kostenintensive Anwendung auf das kartesische Produkt zweier Eingabedatenquellen betrachtet. Dazu wurden Methoden zur effizienten MapReduce-basierten Auswertung des kartesischen Produktes entwickelt. Aktueller Forschungsgegenstand ist die Unterstützung von Redundanz-basierten Blocking und die Untersuchung von Strategien zum Vermeiden redundanter Objekt-Vergleiche. Darüber hinaus wird zurzeit ein Tool entwickelt, welches die komfortable Definition von Matching Workflows, die Submission in MapReduce Cluster sowie die Auswertung der Ergebnisse, ermöglicht.

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 und Publikationsorgan. Einfache Statistiken möglichen u.a. die Erstellung von Rankings der meistzitierten Arbeiten pro Autor oder Publikationsorgan sowie aggregierte Kennzahlen (z.B. h-Index) zur vergleichenden Analyse. Die Betrachtung der zitierenden Arbeiten lässt weitere Rückschlüsse auf den Einfluss einer wissenschaftlichen Publikation zu. So lassen sich z.B. Publikationen dahingehend vergleichen, ob sie von selbst vielzitierten Arbeiten zitiert wurden. Hierzu wurden entsprechende Webanwendungen entwickelt. Neben der Bewertung einzelner Publikationen und Autoren kommt auch der Aggregation der Daten auf Ebene der Forschungsinstitutionen bzw. Autor-Affiliations eine wachsende Bedeutung zu, u.a. für Rankings von Forschungsinstitutionen. Komplexere Analysen setzen Zitierungszahlen mit Ergebnissen eines Peer-Review-Verfahrens in Verbindung, um die Wirksamkeit des Begutachtungsprozesses bei Zeitschriften und Konferenzen zu evaluieren. Die automatische Ermittlung von Zitierungszahlen sowie die Erfassung der Affiliations aus Volltextdokumenten erfordert eine Integration der Daten aus unterschiedlich verfügbaren Quellen. Für zu analysierende Publikationen müssen zunächst entsprechende Suchanfragen gestellt. Anschließend erfolgt ein aufwändiges Matching, um gleiche Publikationseinträge in unterschiedlichen Datenquellen identifizieren und einheitliche Affiliation-Schreibweisen generieren zu können. Ein in diesem Kontext verwandter Service versetzt den Nutzer in die Lage, PDF-Volltexte mit bibliographischen Metadaten aus Webdatenquellen anzureichern. Als Ergebnis liegen u.a. eine Affiliation-Referenzdatenbank und Hilfsmittel zur inkrementellen Erweiterung vor, um so beliebige bibliometrische Analysen zu ermöglichen.

Schema- und Ontologie-Matching

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

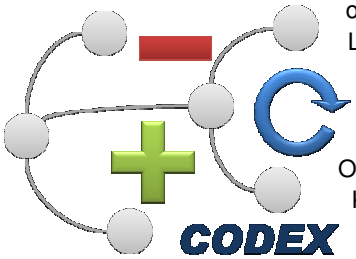
Im Bereich des Schema Matchings wurde untersucht, wie die Auswahl geeigneter Matching-Algorithmen, deren Kombination und Konfiguration vereinfacht werden kann. Ein neuartiges Modellierungswerkzeug zur Erstellung und Parametrisierung sogenannter Matching-Prozesse (AMC) wurde dabei entwickelt. Das Werkzeug vereinfacht die Anpassung von Matching-Prozessen an spezifische Domänen und wendet neuartige Visualisierungstechniken zur Darstellung von Matching-Prozessen an. Prozesse lassen sich schrittweise parametrisieren, und Zwischenergebnisse können analysiert werden. Existierende Matching-Algorithmen können im Werkzeug integriert, und in einer gemeinsamen Umgebung evaluiert werden. Darüber hinaus wurden existierende Matching-Systeme und deren Matching-Prozesse analysiert. Diese nutzen eine Anzahl wiederkehrender Prozessmuster die häufig kombiniert werden um robuste Matching Systeme zu entwickeln. Dennoch bleibt die Erstellung und das Parametrisieren von Matching Prozessen eine manuelle Aufgabe. Wir haben daher ein selbstkonfigurierendes Matching-System entwickelt, dass sich automatisch an ein gegebenes Mapping-Problem anpassen kann. Unser Ansatz analysiert Eingabe- Schemas und Zwischenergebnisse bereits ausgeführter Matcher und führt danach sog. Matching Regeln aus. 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.

Im Bereich der Lebenswissenschaften hat das Matching großer Ontologien stark an Bedeutung gewonnen. Häufig existieren Ontologien mit zumindest teilweise überlappenden Informationen. 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 zunächst das parallele Matching von Ontologien unter Nutzung von Multicore-Prozessoren untersucht. Dazu wurde eine verteilte Infrastruktur implementiert, welche paralleles Matching ermöglicht und eine große Auswahl verschiedener Match-Techniken zur Verfügung stellt: GOMMA (Generic Ontology Matching and Mapping Management). Des Weiteren wurden verschiedene Strategien zum parallelen Matching von Ontologien untersucht. Diese erlauben eine parallele Ausführung ganzer Matcher sowie die interne Parallelisierung der Matcher unter Verwendung einer Partitionierung der Eingabedaten. Dieser Ansatz ermöglichte eine signifikante Reduzierung der Ausführungszeit sowie eine gute Lastbalancierung, Skalierbarkeit und eingeschränkte Speicheranforderungen. Außerdem wurde die Wiederverwendung bereits existierender Mappings zur effizienten Berechnung neuer Mappings untersucht. Ein neu entwickelter auf Mapping-Komposition basierender Ansatz gestattet es qualitativ hochwertige Mappings zu erzeugen. Der Ansatz wurde erfolgreich für das Anatomie-Matchproblem der Ontology Alignment Evaluation Initiative (OAEI) evaluiert. Dabei wurde das Mapping zwischen den Ontologien Adult Mouse Anatomy (MA) und NCI Thesaurus (NCIT) indirekt durch Komposition von Mappings zu vier weiteren großen Anatomie-Ontologien bestimmt.

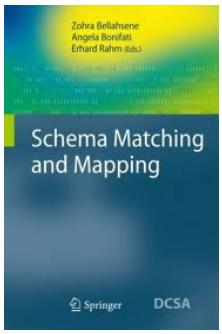
Evolution von Ontologien und Mappings

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

Ontologien werden insbesondere in den Lebenswissenschaften zur eindeutigen semantischen Beschreibung (Annotation) von Objekten wie z.B. Proteinen oder Genen eingesetzt. Bedingt durch neue wissenschaftliche Erkenntnisse oder aufgrund neuer Anforderungen unterliegen die Ontologien ständigen Änderungen, welche sich auf abhängige Datenquellen, Mappings und Anwendungen auswirken. Um mit einer derartigen Evolution umgehen zu können, besteht ein wichtiger Schritt in der Bestimmung der Differenz (des DIFF) zwischen zwei Versionen einer Ontologie. Gefundene Änderungen können dann zur Anpassung abhängiger Daten usw. verwendet werden. Im Projekt wurde ein regelbasierter Diff-Ansatz COntoDiff (Complex Ontology Diff) zur Bestimmung eines vollständigen und ausdrucksstarken Diff Evolution-Mapping zwischen zwei Ontologieversionen entwickelt. Insbesondere erlaubt die webbasierte Applikation CODEX (Complex Ontology Diff Explorer) eine online Berechnung und Analyse von DIFFs für zahlreiche Ontologien in den Lebenswissenschaften. Aufgrund der enormen Größe der Ontologien wurde darüber hinaus ein neuartiger Ansatz zur Bestimmung änderungsintensiver bzw. stabiler Ontologieregionen entwickelt. Nutzer können sich u.a. einen kompakten Überblick über die Evolution in verschiedenen Teilen einer Ontologie verschaffen und Trends in der Evolution aufspüren. In einer Kooperation mit dem Max-Planck-Institut für evolutionäre Anthropologie werden zudem die Auswirkungen von Ontologie- und Annotationsevolution auf die Ergebnisse abhängiger Experimente und Analysen erforscht.



5. **Veröffentlichungen / Graduierungsarbeiten**



Buch

- Bellahsene, Z.; Bonifati, A.; Rahm, E. (Eds.): *Schema Matching and Mapping*. Springer-Verlag, Data-Centric Systems and Applications, 2011

Zeitschriften

- Algergawy, A; Nayak, R; Saake, G.: *Element similarity measures in XML schema matching*. Information Sciences, Volume 180, Issue 24, 15, Pages 4975-4998, December 2010
- Aumüller, D.; Rahm, E.: *Affiliation analysis of database publications*. ACM SIGMOD Record, Volume 40, Issue 1, Pages 26-31, 2011
- Bornmann, L.; Marx, W.; Schier, H.; Thor, A.; Daniel, H.-D.: *From black box to white box at open access journals: Predictive validity of manuscript reviewing and editorial decisions at Atmospheric Chemistry and Physics*. Research Evaluation, Volume 19, Issue 2, Pages 105-118, June 2010
- Hartung, M; Loebe, F; Herre, H.; Rahm, E.: *Management of Evolving Semantic Grid Metadata Within a Collaborative Platform*. Information Sciences Volume 180, Issue 10, Pages 1837-1849, May 2010
- Kirsten, T.; Groß, A.; Hartung, M.; Rahm, E.: *GOMMA: A Component-based Infrastructure for managing and analyzing Life Science Ontologies and their Evolution*. Journal of Biomedical Semantics, Volume 2, paper 6, 2011
- Köpcke, H.; Rahm, E.: *Frameworks for entity matching: A comparison*. Data & Knowledge Engineering, Volume 69, Issue 2, Pages 197-210, February 2010
- Köpcke, H.; Thor, A.; Rahm, E.: *Evaluation of learning-based approaches for matching web data entities*. IEEE Internet Computing, Volume 14, Issue 4, Pages 30-38, July 2010
- Rahm, E.: Semantische integration von Webdaten. GFFT-Jahresbericht 2009/10. März 2010
- Rahm, E.: Integration von Webdaten, Transferbrief leipzig, 2011
- Thor, A.; Bornmann, L.: *The calculation of the single publication h index and related performances measures: A Web application based on Google Scholar*. Online Information Review, Volume 35, Issue 2, pages 291-300, 2011

Proceedings

- Algergawy, A.; Massmann, S.; Rahm, E.: *A Clustering-based Approach For Large-scale Ontology Matching*. Proc. 15th Intl. Conference on Advances in Databases and Information Systems (ADBIS), LNCS (Lecture Notes in Computer Science) 6909, pages 415-428, 2011
- Algergawy, A; Nayak, R; Siegmund,N; Koppen, V; Saake,G.: *Combining Schema and Level-Based Matching for Web Service Discovery*. 10th International Conference of Web Engineering (ICWE), 2010
- Aumüller, D.; Rahm, E.: *PDFMeat: Managing Publications on the Semantic Desktop*. Proc. of 20th ACM Conference on Information and Knowledge Management (CIKM), ACM, pages 2565-2568, 2011
- Bernstein, P.A.; Madhavan, J.; Rahm, E.: *Generic Schema Matching, Ten Years Later*. Proc. of the VLDB Endowment, Volume 4, Issue 11, pages 695-701, 2011
- Groß, A.; Hartung, M.; Kirsten, T.; Rahm, E.: *On Matching Large Life Science Ontologies in Parallel*. Proc. 7th Int. Conference on Data Integration in the Life Sciences (DILS), Springer LNCS (Lecture Notes in Computer Science) 6254, 2010
- Groß, A.; Hartung, M.; Kirsten, T.; Rahm, E.: *Mapping Composition for Matching Large Life Science Ontologies*. Proc. 2nd International Conference on Biomedical Ontology (ICBO), pages 109-116, 2011
- Hartung, M.; Groß, A.; Kirsten, T.; Rahm, E.: *Discovering Evolving Regions in Life Science Ontologies*. Proc. 7th Int. Conference on Data Integration in the Life Sciences (DILS), Springer LNCS (Lecture Notes in Computer Science) 6254, 2010
- Hartung, M.; Terwilliger, J.; Rahm, E.: *Recent advances in schema and ontology evolution*. Schema Matching and Mapping, Springer Data-Centric Systems and Applications, pages 149-190, 2011
- Kirsten, T.; Kiel, A.: *Ontology-based Registration of Entities for Data Integration in large biomedical Research Projects*. Proc. GI-Workshop - Informationsintegration in Service-Architekturen, 2010

- Kirsten, T.; Kolb, L.; Hartung, M.; Groß, A.; Köpcke, H.; Rahm, E.: *Data Partitioning for Parallel Entity Matching*. Proc. 8th International Workshop on Quality in Databases in conjunction with VLDB, 2010
- Köpcke, H.; Thor, A.; Rahm, E.: *Evaluation of entity resolution approaches on real-world match problems*. Proc. 36th Intl. Conference on Very Large Databases (VLDB), 2010
- Kolb, L.; Köpcke, H.; Thor, A.; Rahm, E.: *Learning-based Entity Resolution with MapReduce*. Proc. of 3rd International CIKM Workshop on Cloud Data Management (CloudDB), ACM, pages 1-6, 2011
- Kolb, L.; Thor, A.; Rahm, E.: *Parallel Sorted Neighborhood Blocking with MapReduce*. Proc. 14. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), LNI (Lecture Notes in Informatics) 180, pages 45-64, 2011
- Kolb, L.; Thor, A.; Rahm, E.: *Block-based Load Balancing for Entity Resolution with MapReduce*. Proc. of 20th ACM Conference on Information and Knowledge Management (CIKM), ACM, pages 2397-2400, 2011
- Massmann, S.; Raunich, S.; Aumueller, D.; Arnold, P.; Rahm, E.: *Evolution of the COMA Match System*. Proc. of 6th International Workshop on Ontology Matching (OM), CEUR Proceedings Vol. 814, pages 49-60, 2011
- Peukert, E.; Berthold, H.; Rahm, E.: *Rewrite Techniques for Performance Optimization of Schema Matching Processes*. Proc. of 13th Intl. Conference on Extending Database Technology (EDBT), 2010
- Peukert, E.; Eberius, J.; Rahm, E.: *AMC - A framework for modelling and comparing matching systems as matching processes*. Proc. of 27th IEEE International Conference on Data Engineering (ICDE), IEEE Computer Society, pages 1304-1307, 2011
- Peukert, E.; Eberius, J.; Rahm, E.: *Rule-based construction of matching processes*. Proc. of 20th ACM Conference on Information and Knowledge Management (CIKM), ACM, pages 2421-2424, 2011
- Peukert, E; Massmann, S.; König, K.: *Comparing Similarity Combination Methods for Schema Matching*. GI-Workshop - Informationsintegration in Service-Architekturen, 2010
- Peukert, E.; Rahm, E.: *Restricting the Overlap of Top-N Sets in Schema Matching*. Proc. EDBT Workshop on New Trends in Similarity Search (NTSS), ACM, pages 20-25, 2011
- Rahm, E.: *Evolution and Merging of Real-Life Ontologies*. Proc. of Italian Database Conference (SEBD), 2011
- Rahm, E.: *Towards large-scale schema and ontology matching*. Schema Matching and Mapping, Springer Data-Centric Systems and Applications, pages 3-27, 2011
- Raunich, S.; Rahm, E.: *Automatic Target-driven Ontology Merging*. Proc. of 27th IEEE International Conference on Data Engineering (ICDE), IEEE Computer Society, pages 1276-1279, 2011
- Thor, A., Anderson, P.; Raschid, L.; Navlakha, S.; Saha, B.; Khuller, S.; Zhang, X.-N.: *Link Prediction for Annotation Graphs using Graph Summarization*. Proc. International Semantic Web Conference (ISWC), LNCS (Lecture Notes in Computer Science) 7031, pages 714-729, 2011
- Thor, A., Rahm, E.: *CloudFuice: A flexible Cloud-based Data Integration System*. Proc. of10th International Conference on Web Engineering (ICWE), LNCS (Lecture Notes in Computer Science) 6757, pages 304-318, 2011
- Wartner, C., Kitschke, S.: *PROOF: Produktmonitoring im Web*. Proc. 14. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW), LNI (Lecture Notes in Informatics) 180, pages 722-725, 2011

Technische Berichte und Poster

- Hartung, M.; Groß, A.; Rahm, E.: *Rule-based Generation of Diff Evolution Mappings between Ontology Versions*. Technical Report, arXiv:1010.0122, 2010 - http://arxiv.org/abs/1010.0122
- Raunich, S.; Rahm, E.: *Target-driven merging of Taxonomies*. Technical Report, arXiv:1012.4855, 2010 - http://arxiv.org/abs/1012.4855

Dissertationen

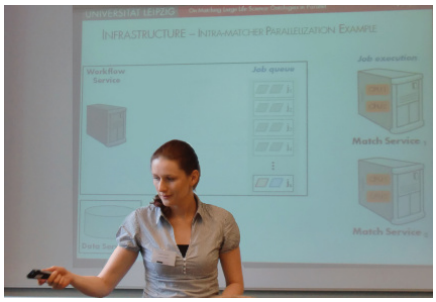
- Golovin, N.: Web Recommendations for E-Commerce Websites, Univ. Leipzig, 2010
- Hartung, M.: Evolution von Ontologien in den Lebenswissenschaften, Univ. Leipzig, 2011

Nach der erfolgreichen Verteidigung: N. Golovin (März 2010) und M. Hartung (April 2011) mit Doktorhut

Bachelor-, Master- und Diplomarbeiten

1. Arnold, P.: The Basics of Complex Correspondences and Functions and their Implementation and Semi-automatic Detection. Masterarbeit, Univ. Leipzig, 2011
2. Bremer, A.: Erstellung eines XQuery-Tutorials für das LOTS. Bachelorarbeit, Univ. Leipzig, 2011
3. Chyhir, A.: Analyse der Evolution von Mappings anhand ausgewählter Matchprobleme in den Lebenswissenschaften. Bachelorarbeit, Univ. Leipzig, 2011
4. Do, V.H.: Automatische Wiederverwendung auf dem Element-Level im Kontext von Schema Matching. Masterarbeit, Univ. Leipzig, 2011
5. Fiedler, M.: Ein hybrider Ansatz zur flexiblen Speicherung von XML-Daten in RDBMS am Beispiel von PubMed. Diplomarbeit, Univ. Leipzig, 2010
6. Gäbel, J.: Konzeption und Realisierung einer Rich-Internet-Applikation als Web-basierte Benutzerschnittstelle für das GeWare-System. Bachelorarbeit, Univ. Leipzig, 2010
7. Gassner, M.: Realisierung eines Extraktionswerkzeugs zur Untersuchung von Trends in Themengebieten der Lebenswissenschaften. Bachelorarbeit, Univ. Leipzig, 2011
8. Gröger, T.: Entwicklung eines Data-Warehouse-Systems zur Analyse von Online-Wohnungsanzeigen. Masterarbeit, Univ. Leipzig, 2011
9. Huang, L.: Performance-Vergleiche bei der Speicherung von XML-Daten in relationalen DBVS. Diplomarbeit, Univ. Leipzig, 2010
10. Jundin, E.: Vergleich von mySQL und CouchDB. Bachelorarbeit, Univ. Leipzig, 2010
11. Kähler, M.: Entwicklung und Implementierung eines Software-Tools zur Visualisierung von Flugplänen. Bachelorarbeit, Univ. Leipzig, 2011
12. Kropp, H.: Berechnung von Diff-Evolution-Mappings zwischen geänderten Produktkatalogen. Diplomarbeit, Univ. Leipzig, 2011
13. Phan, D.: Machine Learning Matching mit COMA++. Diplomarbeit, Univ. Leipzig, 2010
14. Sander, F.: Automatisches Matching von Hoteldatensätzen. Diplomarbeit, Univ. Leipzig, 2011
15. Thomas, S.: Preprocessing für das Matchen von Produktangeboten. Bachelorarbeit, Univ. Leipzig, 2010
16. Veshcheva, Y.: Object Matching für Linked Data. Masterarbeit, Univ. Leipzig, 2011
17. Xia, C.: Web-basierte Methoden zur Untersuchung von Affiliation-Angaben wissenschaftlicher Papiere. Bachelorarbeit, Univ. Leipzig, 2010
18. Xuzhi, Z.: Broken Link-Erkennung in Linked Data-Datenquellen. Bachelorarbeit, Univ. Leipzig, 2010

Vorträge



Vorträge von L. Kolb auf der BTW2011 in Kaiserslautern (links), S. Maßmann auf der ADBIS2011 in Wien (Mitte) sowie A. Groß auf der DILS2010 in Göteborg (rechts)

- Algergawy, A.: Combining Schema and Level-Based Matching for Web Service Discovery. ICWE, Wien, 2010
- Groß, A.: On Matching Large Life Science Ontologies in Parallel. 7th Intl. Conference on Data Integration in the Life Sciences, Gothenburg (Sweden), 2010
- Groß, A.: Mapping Composition for Matching Large Life Science Ontologies. International Conference on Biomedical Ontologies (ICBO), Buffalo (USA), 2011
- Hartung, M.: Discovering Evolving Regions in Life Science Ontologies. 7th Intl. Conference on Data Integration in the Life Sciences, Gothenburg (Sweden), 2010
- Hartung, M.: Evolution von Ontologien in den Lebenswissenschaften. Gastvortrag Forschungsseminar WBI - DBIS, Humboldt Universität, Berlin, 2011
- Köpcke, H.: Evaluation of entity resolution approaches on real world match problems. VLDB, Singapur, 2010
- Köpcke, H.: Object Matching für Linked Data. Semantic Web Tag, Leipzig, 2011
- Kolb, L.: Data Partitioning for Parallel Entity Matching. 8th International Workshop on Quality in Databases (QDB), Singapur, September 2010
- Kolb, L.: Parallel Sorted Neighborhood Blocking with MapReduce. 14. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web, Kaiserslautern, März 2011
- Kolb, L.: Parallele Link Discovery in der Cloud. 3. Leipziger Semantic Web Tag (LSWT), Leipzig, Mai 2011
- Kolb, L.: Learning-based Entity Resolution with MapReduce. 3rd International Workshop on Cloud Data Management (CloudDB), Glasgow, Oktober 2011
- Maßmann, S.: A Clustering-based Approach For Large-scale Ontology Matching. Advances in Databases and Information Systems, Wien, 2011
- Maßmann, S.: Evolution of the COMA Match System. Ontology Matching Workshop @ ISWC, Bonn, 2011
- Peukert, E.: Rewrite Techniques for Performance Optimization of Schema Matching Processes. 13th International Conference on Extending Database Technology (EDBT), Lausanne (Switzerland), 2010
- Peukert, E.: Comparing Similarity Combination Methods for Schema Matching. GI-Workshop Informationsintegration in Service-Architekturen, Leipzig, 2010
- Peukert, E.: Restricting the Overlap of Top-N Sets in Schema Matching. Workshop on New Trends in Similarity Search (NTSS),Uppsala (Sweden), 2011
- Rahm, E.: Semantische Integration von Webdaten. Keynote GFFT-Jahrestagung, Frankfurt, März 2010
- Rahm, E.: WDI-Lab – Innovationslabor zur semantischen Integration von Webdaten. Semantic Web Day, Leipzig, Mai 2010
- Rahm, E.: Evolution and merging of real-life ontologies. Keynote Italian Database Conf. SEBD, Maratea, June 2011
- Rahm, E.: Generic Schema Matching – Ten Years later. VLDB 10-Year Best Paper Award Keynote, Seattle, Sep. 2011
- Rahm, E.: Paralleles Object Matching in der Cloud. FGDB-Workshop, Potsdam, Nov. 2011

- Thor, A.: Web Data Integration. University of Maryland, USA, 2010
- Thor, A.: Toward an adaptive String Similarity Measure for Matching Product Offers. GI-Workshop Informationsintegration in Service-Architekturen, Leipzig, 2010
- Thor, A.: CloudFuice: A flexible Cloud-based Data Integration System. ICWE, Zypern, 2011
- Thor, A.: Link Prediction for Annotation Graphs using Graph Summarization and PSL. University of Maryland, USA, 2011
- Thor, A.: Data Integration in the Cloud. University of Waterloo, Canada, 2011
- Thor, A.: Data Integration in the Cloud. University of Toronto, Canada, 2011
- Thor, A.: Link Prediction for Annotation Graphs using Graph Summarization. ISWC, Bonn, 2011

Mitgliedschaften in Gremien/Redaktionskollegien, Herausgebergremien u.ä.

Rahm, E.:

- Stv. Sprecher des Fachbereichs "Datenbanken und Informationssysteme" der Gesellschaft für Informatik
- Advisory Board Europar Conference
- Mitherausgeber der Zeitschrift "Datenbank-Spektrum" (bis 2010)
- Vice PC Chair Int. Conf. On Data Engineering 2012
- Programmkomitee verschiedener Konferenzen (u.a. SIGMOD 2012, EDBT 2011, BTW 2011, ESWC 2011, VLDB 2010, DILS 2010, WebDB 2010, OM2010)
- Gutachter für diverse Zeitschriften und Forschungsgesellschaften
- Vorstandsmitglied IZBI (Interdisz. Zentrum für Bioinformatik, Leipzig)
- Vorsitzender Prüfungsausschuß Informatik

Aumüller, D.:

- Programmkomiteemitglied: ESWC 2010 - Semantic Web in Use Track, SemWiki 2010
- Gutachter für Zeitschrift Journal on Computing and Cultural Heritage

Hartung, M.:

- Programmkomiteemitglied: IDC 2011/2012
- Gutachter für Zeitschrift IEEE Transactions on Knowledge and Data Engineering
- Mitglied im Fakultätsrat

Thor, A.:

- Programmkomiteemitglied: ESWC 2012
- Gutachter für Tschechische Forschungsgemeinschaft (2010)