

Institut für Informatik

Abteilung Datenbanken

Prof. Dr. Erhard Rahm

<http://dbs.uni-leipzig.de>



Abteilung Datenbanken im Febr. 2003 (von links): Erhard Rahm, Nick Golovin, Dieter Sosna, Thomas Stöhr, Sergey Melnik, Hong Hai Do, Toralf Kirsten, Stefan Jusek, Andreas Thor, Robert Müller, Andrea Hesse, Timo Böhme, Ulrike Greiner

Vorwort

Mit der vorliegenden Broschüre wollen wir einen Einblick in unsere Arbeiten an der Abteilung Datenbanken der Universität Leipzig (Institut für Informatik) geben. Sie finden hier u.a. eine Vorstellung der Mitarbeiterinnen und Mitarbeiter, einen Überblick zu den aktuellen Forschungsprojekten und ausgewählten Publikationen, eine Auflistung des Lehrangebots sowie Kontaktinformationen.

Die Forschungsarbeiten werden derzeit weitgehend über Drittmittel der DFG und Industrie finanziert, u.a. im Rahmen des neuen Bioinformatik-Zentrums der Universität Leipzig sowie im Graduiertenkolleg Wissensrepräsentation. Die Schwerpunkte der Untersuchungen liegen aktuell auf folgenden Gebieten:

- Web und Datenbanken: XML-Datenbanken, Metadaten-Verwaltung, Web Usage Mining, Web-basiertes Lernen
- Datenbankkonzepte in den Lebenswissenschaften
- Workflow-Systeme
- Implementierungskonzepte für Hochleistungs-Datenbanksysteme.

In diesen und verwandten Gebieten sind wir an Kooperationen mit Praxispartnern sehr interessiert, sei es durch studentische Arbeiten (Berufspraktika, Diplomarbeiten), Weiterbildungsveranstaltungen oder gemeinsame Forschungsprojekte. Über eine diesbezügliche Kontaktaufnahme sowie weitere Anregungen würden wir uns sehr freuen.

Ihr



(Prof. Dr.-Ing. Erhard Rahm)



Inhaltsverzeichnis

Vorwort	3
Kontakt	5
Lage	6
Wissenschaftliche Mitarbeiter	8
Mitgliedschaft in Gremien, Redaktionskollegien	12
Forschungsschwerpunkte	13
• Web und Datenbanken	
• Datenbankkonzepte in den Lebenswissenschaften	
• Workflow-Systeme	
• Implementierungskonzepte für Hochleistungs-Datenbanksysteme	
Ausgewählte Bücher und Publikationen	21
Lehrangebot	26



Kontakt

Besucheranschrift / Postanschrift

Institut für Informatik	Institut für Informatik
Universität Leipzig	Universität Leipzig
Hauptgebäude (Raum 3-56)	Postfach 920
Augustusplatz 10-11	04009 Leipzig
04109 Leipzig	

Tel.: 0341/97-32241 / 21
Fax.: 0341/97-32209
E-Mail: rahm@informatik.uni-leipzig.de
Web: <http://dbs.uni-leipzig.de>

Leitung

Prof. Dr.-Ing. Erhard Rahm

Sekretariat

Frau Andrea Hesse

Programmierer

Herr Stefan Jusek

Wissenschaftliche Mitarbeiter

Dipl.-Inform. Timo Böhme	Dr. Robert Müller
Dipl.-Inform. Hong Hai Do	Dr. Dieter Sosna
Dipl.-Inform. Nick Golovin	Dipl.-Inform. Thomas Stöhr
Dipl.-Inform. Ulrike Greiner	Dipl.-Inform. Sergej Melnik
Dipl.-Kfm. (FH) Toralf Kirsten	Dipl.-Inform. Andreas Thor

Studentische Hilfskräfte, Diplomanden

siehe <http://dbs.uni-leipzig.de>

Stand: Februar 2003



Lage

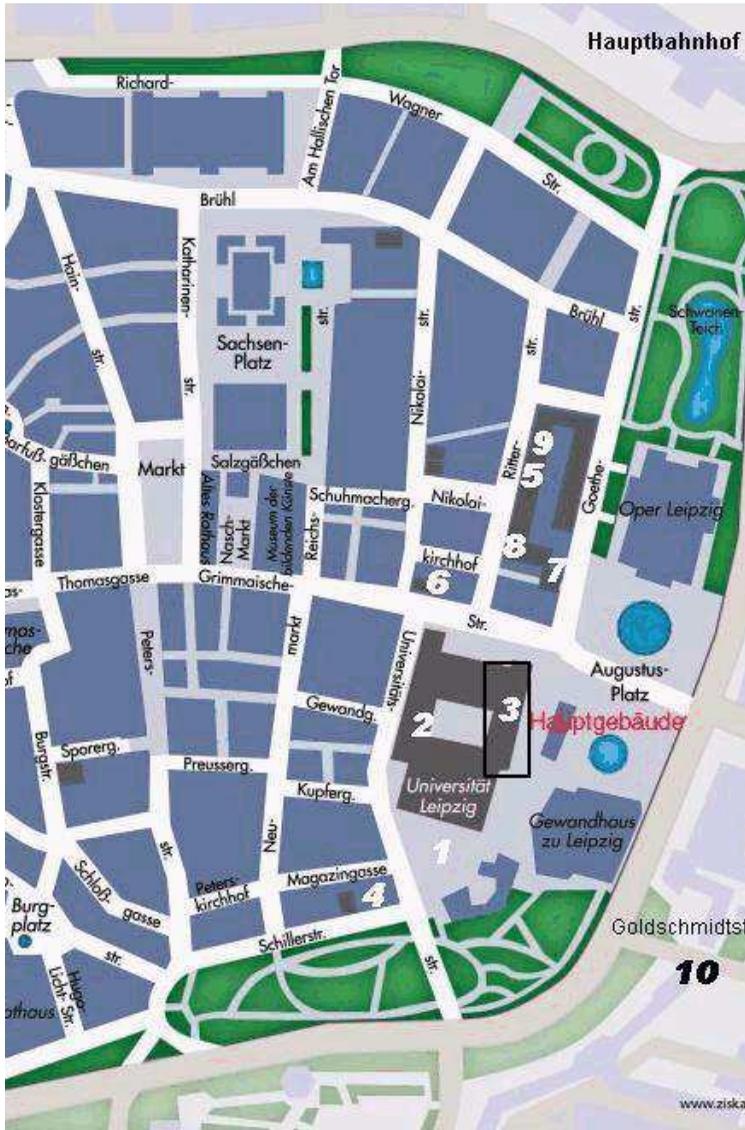
Wir sind in der 3. Etage des Hauptgebäudes der Universität am Augustusplatz untergebracht, im Zentrum von Leipzig in geringer Entfernung zum Hauptbahnhof (s. Lageplan S. 7).



Anreise

- *mit der Bahn:* Wenn Sie auf dem Leipziger Hauptbahnhof angekommen sind, verlassen Sie diesen durch die Osthalle, überqueren den Vorplatz mit den Haltestellen der LVB und erreichen uns zu Fuß durch die Goethestraße in etwa 5 Minuten. Sie können auch in eine Straßenbahn der Linien 4, 7, 15 oder 12 einsteigen und eine Haltestelle in Richtung Augustusplatz fahren. Die Bahnen halten vor dem Hauptgebäude der Universität.
- *mit dem Flugzeug:* Vom Flughafen Schkeuditz (Halle/Leipzig) gelangen sie mit dem Flughafen-Express der Deutschen Bahn direkt zum Hauptbahnhof.
- *mit dem Auto:* Von den entsprechenden Autobahnabfahrten bzw. Landstraßen folgen Sie der Beschilderung "Zentrum" zum Innenstadt-Ring. Am Ring folgen Sie der Beschilderung "Gewandhaus/Oper" bzw. "Tiefgarage Augustusplatz" von Norden und Westen kommend am Hauptbahnhof vorbei, von Süden kommend am Neuen Rathaus vorbei weiter zum Augustusplatz. Vor dem Hauptgebäude der Universität befindet sich die Einfahrt zur Tiefgarage mit zahlreichen Parkplätzen.





Wissenschaftliche Mitarbeiter

Prof. Dr. Erhard Rahm

Geb. 1959 in der Pfalz, verheiratet, drei Kinder. 1979-1984 Studium der Informatik an der Universität Kaiserslautern. Bis zur Promotion 1988 wissenschaftlicher Mitarbeiter am Lehrstuhl von Prof. Härder, Kaiserslautern. 1988/89 ein Jahr Gastwissenschaftler am IBM T.J. Watson Research Center, Hawthorne, NY, im Informatik-Bereich "Large Systems". US-Patent für ein neuartiges Caching-Verfahren zur Nutzung von Erweiterungsspeichern zwischen Haupt- und Plattenspeichern. Bis 1994 wissenschaftlicher Assistent an der Universität Kaiserslautern; Leitung eines DFG-Projekts zu Hochleistungs-Transaktionssystemen; 1993 Habilitation. Seit April 1994 Lehrstuhlinhaber für "Datenbanken" am Institut für Informatik der Universität Leipzig. 1996-1999 geschäftsführender Direktor des Instituts und Prodekan der Fakultät für Mathematik und Informatik. Initiator und erster Sprecher (1997-1999) des Leipziger Informatik-Verbundes (LIV). Sprecher des 2001 gegründeten GI-Arbeitskreises „Web und Datenbanken“. Autor von mehreren Büchern sowie zahlreichen Publikationen; Mitwirkung in den Programmkomitees der international führenden Datenbankkonferenzen. Wissenschaftliche Arbeitsgebiete: Datenbankkonzepte in den Lebenswissenschaften, Metadaten-Verwaltung, Data Warehousing und Datenintegration, XML-Datenbanken, Workflow Management sowie Performance-Bewertung von Computersystemen.



Dr. Robert Müller

Diplom-Mathematiker (Universität Mainz 1994). 1994 - 1996 wissenschaftlicher Mitarbeiter am Institut für Medizinische Biometrie, Epidemiologie und Informatik der Universitätsklinik Mainz (Tätigkeitsschwerpunkte: Medizinische Informationssysteme, Wissensbasierte Systeme in der Onkologie). Seit 1996 wissenschaftlicher Mitarbeiter am Lehrstuhl "Datenbanken". (Tätigkeitsschwerpunkte: Workflow-Management, E-Services, Bioinformatik). 2002 Promotion zum Doktor-Ingenieur (Titel der Dissertation: *Event-Oriented Dynamic Adaptation of Workflows: Model, Architecture and Implementation*)

**Dr. Dieter Sosna**

Diplom-Mathematiker, 1975 Promotion auf dem Gebiet der nichtlinearen Funktionalanalysis (Univ. Leipzig), 1986 - 1994 Computergraphik, seit 1994 Abteilung Datenbanken. Forschungsschwerpunkte: Elektronische Bibliotheken, Lernen mit digitalen Medien, Geoinformationssysteme, Datenschutz und Datensicherheit, Bedienungstheorie.

**Timo Böhme**

Diplom-Informatiker, Univ. Leipzig 1999. Seit Dezember 1999 an der Abteilung Datenbanken. Forschungsschwerpunkte: Verwaltung von XML-Daten, XML-Datenbanksysteme.



Hong-Hai Do

Magister Artium (Germanistik), 1996, Diplom (Informatik), Univ. Leipzig 1999. 1999-2001 wissenschaftlicher Mitarbeiter am Institut für Informatik. Seit 2001 wissenschaftlicher Mitarbeiter am Interdisziplinären Zentrum für Bioinformatik. Forschungsschwerpunkt: Daten- und Metadatenintegration in der Bioinformatik.

**Nick Golovin**

Magister Diplom (Informatik), Dnepropetrovsk 2001. Seit Februar 2003 an der Abteilung Datenbanken innerhalb des Graduiertenkollegs Wissensrepräsentation.

**Ulrike Greiner**

Diplom-Informatikerin, Univ. Leipzig 2000. Seit November 2000 an der Abteilung Datenbanken. Forschungsschwerpunkte: Adaptives Workflow-Management, kooperative Workflows, E-Services.

**Toralf Kirsten**

Diplom-Kaufmann (FH), HTWK Leipzig 1999. Seit 2002 am Interdisziplinären Zentrum für Bioinformatik. Forschungsschwerpunkte: Einsatz von Datenbanken in der Bioinformatik, Datenbankintegration von Analysemethoden und -verfahren, Algorithmen zur Optimierung von Microarrays und zur Bestimmung des Genexpressionsgrades.



Sergey Melnik

Diplom-Informatiker, Univ. Leipzig 1999. Gastforscher in der Stanford Database Group: 1999-2002. Forschungsschwerpunkte: Metadaten-Verwaltung, Datenbank-Optimierung, Information Retrieval, Semantic Web.



Thomas Stöhr

Diplom-Informatiker, Universität Kaiserslautern 1993. 1993-1994 wiss. Mitarbeiter an der Universität Kaiserslautern. Seit 1994 wiss. Mitarbeiter und Doktorand am Lehrstuhl Datenbanken. Z.Z. Promotion auf dem Gebiet Datenallokation in Parallelen Datenbanksystemen. Forschungsschwerpunkte: Parallele Datenbanken, Data Warehousing, Web Usage Mining.



Andreas Thor

Diplom-Informatiker, Univ. Leipzig 2002. Seit Januar 2003 an der Abteilung Datenbanken innerhalb des Graduiertenkollegs Wissensrepräsentation. Forschungsschwerpunkte: Web Mining, Personalisierung von Webseiten.



Mitgliedschaft in Gremien, Redaktionskollegien

Prof. Dr. Erhard Rahm

- Leitungsgremium Fachbereich "Datenbanken und Informationssysteme" der Gesellschaft für Informatik
- Sprecher des Arbeitskreises "Web und Datenbanken" der Gesellschaft für Informatik (<http://dbs.uni-leipzig.de/webdb/>)
- Vorstandsmitglied Interdisziplinäres Zentrum für Bioinformatik (IZBI) der Universität Leipzig
- Advisory Board Europar Conference
- Mitherausgeber der Zeitschrift "Datenbank-Spektrum"
- Mitherausgeber der Zeitschrift "Parallel Computing"
- Tagungsleiter BTW2003 (www.btw2003.de)
- Programmkomitee zahlreicher Konferenzen, u.a. VLDB 2003, NOD 2003, BTW 2003, SIGMOD 2002, Web Databases 2002, NOD 2002, Web-Datenbanken 2001, Parallel Computing 2001, BTW 2001, Data Engineering 2000)

Timo Böhme

- Programmkomitee Workshop on the Efficiency and Effectiveness of XML Tools, and Techniques (EEXTT2002)

Sergey Melnik

- Eingeladener Experte in RDF Core Working Group am World Wide Web Consortium (W3C)
- Programmkomitee: Semantic Web Workshop 2002, WWW 2002 Posters



Forschungsschwerpunkte

Web und Datenbanken

XML-Datenverwaltung

Timo Böhme, Erhard Rahm

Die massive Verbreitung von XML verlangt zunehmend eine effiziente Verwaltung großer Mengen von XML-Daten. Hierzu wurden einerseits native XML-Datenbanken entwickelt, andererseits mit XML-Erweiterungen versehene relationale bzw. objekt-relationale Datenbanksysteme.

In der ersten Projektphase ging es um eine Abschätzung der Leistungsfähigkeit der verschiedenen Architekturen zur Verwaltung von XML Daten. Dafür wurde ein skalierbarer **Multi-User Benchmark, XMach-1** (XML Data Management Benchmark), entwickelt. Er basiert auf einer Web-Anwendung und spezifiziert Datenbankstruktur, die Generierung der Daten, sowie Anfrage- und Update-Operationen. Der Benchmark wurde für verschiedene Architekturen implementiert. Dabei zeigten die Evaluierungsergebnisse teilweise erhebliche Leistungsprobleme derzeitiger Systeme, insbesondere bezüglich der Skalierbarkeit sowie bei der Verwaltung dokumentorientierter XML-Daten.

Auf der Grundlage dieser Ergebnisse erfolgt in der zweiten Projektphase die Entwicklung optimierter Strukturen für die effiziente XML-Datenverwaltung mit Schwerpunkt auf objektrelationalen Datenbanksystemen. Es werden speziell die Potentiale dieser Datenbanksysteme für eine performante, skalierbare und generische Verwaltung dokumentorientierter XML-Daten untersucht.



Metadaten-Verwaltung und Schema-Matching

Hai Hong Do, Sergey Melnik, Erhard Rahm

Bei der Kopplung von Datenbanken und Web-Anwendungen spielt die Metadaten-Verwaltung eine herausragende Rolle. Model-Management-Systeme helfen dabei, Datenbank-Schemas, Kataloge, Mediator-Definitionen usw. sowie Abbildungen (Mappings) zwischen ihnen in einheitlicher Weise zu verwalten und zu manipulieren. Um den Programmieraufwand beim Zugriff auf solche Objekte zu reduzieren, wird eine Reihe von Operationen höheren Abstraktionsgrades zur Verfügung gestellt.

Der sog. Match-Operator ist dabei einer der bedeutendsten Operatoren. Bisherige Vorschläge zur Realisierung von Schema Matching wurden umfassend analysiert und im Rahmen einer Lösungstaxonomie klassifiziert. Zum strukturbasierten Matching wurde ein generischer iterativer Algorithmus namens Similarity Flooding entwickelt. Ferner wurde ein Prototyp namens **COMA** (composite match system) entwickelt, mit dem unterschiedliche Match-Algorithmen in flexibler Weise kombiniert werden sowie eine Wiederverwendung (Reuse) früherer Match-Ergebnisse ermöglicht wird. Die Effektivität einzelner Algorithmen und Kombinationsmöglichkeiten wurde in einer umfangreichen Evaluierung analysiert.

Für ein umfassenderes Model-Management wurde ein detailliertes Verarbeitungskonzept mit generischen Operationen (Match, Merge, Compose, etc.) auf Modellen und Mappings spezifiziert. Ferner wurde ein umfassender Prototyp namens **RONDO** entwickelt, der eine Implementierung der Operatoren enthält und ihre Verwendung innerhalb von Skript-Programmen ermöglicht, mit denen komplexe Metadaten-basierte Anwendungen mit geringem Aufwand realisiert werden können.

Die Arbeiten erfolgen zum Teil in Kooperation mit Microsoft Research (Phil Bernstein). Zum Schema Matching wurde ein Projekt mit SAP begonnen.



Web Usage Mining

Erhard Rahm, Thomas Stöhr, Nick Golovin, Andreas Thor

Insbesondere im E-Business-Umfeld dient die Analyse des Zugriffsverhaltens auf Web-Seiten einer verstärkten Kundenbindung bzw. der Gewinnung von Neukunden. Es können inhaltliche Verbesserungen des Web-Auftritts erzielt, Performance-Probleme behoben und generell eine bessere Bedienung der Nutzer erreicht werden.

Unser Web Usage Mining-Ansatz nutzt ein Data Warehouse, in dem die Daten aus Web-Logs und sonstigen Informationsquellen entlang verschiedener Dimensionen strukturiert verwaltet werden. Diese Vorgehensweise ist skalierbar hinsichtlich umfangreicher Web-Sites und sehr großer Zugriffsraten und ermöglicht flexible Aggregationen und Auswertungen, insbesondere Data-Mining-Verfahren. Im Unternehmens-Umfeld lässt sich ferner eine Kopplung mit Geschäftsdaten erreichen (CRM). Eine Prototyp-Realisierung erfolgte für die Web-Server des Instituts für Informatik unter Nutzung von kommerziell verfügbaren Tools, wie z.B. Web-fähiger OLAP-Tools, und einer eigenentwickelten, GUI-gestützten Anwendung zur Auswertung von Navigationspfaden, Top-Seiten etc.

Unter Mitwirkung des Lehrstuhl E-Business der HHL Leipzig, der Leipziger Gesellschaft für versicherungswissenschaftliche Forschung mbH und Partnern aus Versicherungsunternehmen wurde eine F+E-Werkstatt zum Thema "Kundenmonitoring auf der Basis von Web Mining und Data Warehousing" durchgeführt. Dabei wurden ein Konzept zur Web-Zugriffsanalyse für Versicherungsunternehmen entwickelt sowie der umfangreiche Werkzeugmarkt analysiert.

Aktuelle Untersuchungen befassen sich mit der Nutzung der Analyseergebnisse zur Personalisierung von Websites.



Web-basiertes Lernen

Erhard Rahm, Dieter Sosna

Für den von uns entwickelten Dokumenten-Server der Universität (<http://dol.uni-leipzig.de>) wurden bisherige Erkenntnisse bei Administration und Nutzung zusammengefasst und publiziert.

Im Projekt "Internetgestützter SQL-Trainer" im Rahmen des Verbundprojekts Bildungsportal Sachsen entstand ein web-basierter Prototyp zum Erlernen und Üben der standardisierten Datenbankabfragesprache SQL. Dabei sind mehrere Arbeitsmodi für die Nutzer (Studenten, etc.) vorgesehen, die sowohl ein freies Üben mit einer SQL-Datenbank als auch eine Bearbeitung von vorgefertigten Anfragen mit und ohne Hilfe bei Fehlern gestatten. Über die Leistungen eines Nutzers soll Protokoll geführt werden können. Damit wird auch eine Auswertung für die Vergabe von Übungsscheinen usw. ermöglicht. Die erreichten Ergebnisse wurden publiziert.

Zusätzlich wurden erste Untersuchungen zur Realisierung eines Online-Wörterbuchs Neugriechisch - Deutsch in Zusammenarbeit mit Prof. Henrich (Institut für Klassische Philologie) durchgeführt.



Datenbankkonzepte in den Lebenswissenschaften

Hong-Hai Do, Toralf Kirsten, Erhard Rahm

Die Univ. Leipzig konnte als einzige ostdeutsche Universität den DFG-Wettbewerb zur Einrichtung neuer Bioinformatik-Zentren gewinnen. In dem 2001 gegründeten Leipziger Interdisziplinären Zentrum für Bioinformatik (IZBI, www.izbi.de) beschäftigt sich eine Arbeitsgruppe unter Leitung von E. Rahm mit Datenbanken und Datenintegration in den Lebenswissenschaften.

Ein Schwerpunkt der bisherigen Arbeiten war die Realisierung eines eigenen Data Warehouse-Systems zur Genexpressionsanalyse. Hintergrund ist die Bereitstellung einer flexiblen Datenbanklösung für die Verwaltung und Analyse von an der Univ. Leipzig auf Basis der Microarray-Technologie (hier hauptsächlich Hard- und Software der Fa. Affymetrix) zahlreich erzeugten Genexpressionsdaten.

Zu Beginn wurden eine umfangreiche Anforderungsanalyse durchgeführt sowie bereits implementierte Genexpressionsdatenbanken evaluiert. Aufbauend auf den Ergebnissen wurden eine Systemarchitektur und ein konzeptionelles Datenmodell zur Speicherung von Genexpressionsdaten in einem Data Warehouse entworfen. Durch das multidimensionale Datenmodell kann die Datenbank sowohl die Rohdaten als auch Daten verschiedenster Normalisierungsmethoden flexibel speichern und auswerten. Die Plattform wird derzeit in die Produktivphase überführt. Weitere Untersuchungen befassen sich mit der Integration unterschiedlichster Annotationen und Analyse-Werkzeuge.

Die Arbeiten erfolgen in enger Kooperation mit biologischen Anwendern, insbes. Prof. Pääbo (MPI für Evolutionäre Anthropologie), Prof. Horn und Dr. Krohn (Medizinische Fakultät).



Adaptive Workflow-Systeme

Ulrike Greiner, Robert Müller, Erhard Rahm

Eine wesentliche Limitation derzeitiger Workflow-Systeme besteht in ihrer unzureichenden Flexibilität, auf unerwartete Ereignisse zu reagieren. Im Rahmen eines DFG-geförderten Forschungsprojekts wird daher das Workflow-System **AGENTWORK** entwickelt, welches die ereignisorientierte und automatische Laufzeit-Adaptation von Workflows unterstützt. Primäres Anwendungsgebiet hierbei ist die computergestützte Krebsbehandlung. Der zugrundeliegende Ansatz verwendet Regelwissen auf der Basis einer temporalen Logik, um Ereignisse daraufhin zu bewerten, ob sie die dynamische Adaptation eines Workflows - wie z.B. das Hinzufügen oder Entfernen von Aktivitäten - erforderlich machen. Insbesondere wird zwischen einer reaktiven und einer prädiktiven Adaptationsstrategie unterschieden. Während die reaktive Strategie Workflow-Adaptationen nur bei akutem Bedarf durchführt, wird bei der prädiktiven Strategie ein Workflow temporal abgeschätzt und vorausschauend umgebaut, damit sich die an der Workflow-Ausführung beteiligten Benutzer frühzeitig auf die veränderte Situation einstellen können. Weiterhin werden Ansätze zur Modellierung von Workflow-Kooperationen und zur Behandlung von Inter-Workflow-Abhängigkeiten bei Adaptationen erarbeitet.

Zur Anwendbarkeit der Ansätze im Bereich E-Business wurden erste Untersuchungen auf Basis von *Web-Services* durchgeführt. Im Mittelpunkt steht hierbei eine flexible Ausnahmebehandlung im Rahmen kooperativer Workflows.

Die medizinorientierten Arbeiten erfolgen zusammen mit dem Institut für Medizinische Informatik, Statistik und Epidemiologie der Universitätskliniken Leipzig (Prof. M. Löffler, Prof. A. Winter, Dr. B. Heller).



Implementierungskonzepte für Hochleistungs-Datenbanksysteme

Parallele Datenbanken

Erhard Rahm, Thomas Stöhr

Unsere Forschungsarbeiten umfassen die Entwicklung und Bewertung von Verfahren zur dynamischen Lastbalancierung und der flexiblen Datenallokation in Parallelen DBS. Wir konzentrieren uns auf Shared-Disk-Architekturen und deren hohes Potential, das durch die gemeinsame Plattenanbindung gegeben ist. Im Vordergrund stehen eine effektive Parallelverarbeitung im Mehrbenutzerfall, die dynamische Behandlung von Skew-Effekten sowie Aspekte des „Self-Tuning“. Eine Leistungsbewertung unserer Verfahren erfolgt analytisch oder mit eigenentwickelten, komplexen Simulationssystemen. Einige Verfahren wurden prototypisch realisiert.

Ein Schwerpunkt unserer Arbeit sind Verfahren für hochkomplexe, mehrdimensionale Anfragen (Star-Joins) auf sehr großen Datenmengen im Data Warehouse-Umfeld. Auf Basis einer von uns entwickelten mehrdimensionalen, hierarchischen Fragmentierungsmethode für Faktentabellen und Zugriffsstrukturen (MDHF) und eines analytisches Kostenmodells wurde der GUI-basierte Prototyp **Warlock** entwickelt. Dieses Werkzeug schlägt Fragmentierungskandidaten vor, analysiert das resultierende Anfrageverhalten paralleler Star-Joins, leitet eine Datenallokation ab und visualisiert diese. Zusätzlich können unterschiedliche Konfigurationen verglichen und optimiert werden.

Wir haben zudem Lastbalancierungsstrategien entwickelt, welche die Prozessor- und Plattenauslastung sowie Skew-Effekte integriert berücksichtigen. Für die parallele Join-Verarbeitung in objektrationalen DBS wurde ein Leistungsvergleich von relationalen und objektorientier-



ten Verarbeitungsmethoden durchgeführt sowie eine angepasste Allokationsstrategie entwickelt.

Optimierung von Set-Containment-Joins

Sergey Melnik, Hector Garcia-Molina (Stanford University)

Ein Set-Containment-Join ist ein Verbund zwischen mengenwertigen Attributen zweier Relationen, dessen Verbund-Bedingung mittels des Teilmengenoperators spezifiziert wird. Wir haben einen neuartigen Partitionierungsalgorithmus, genannt Divide-and-Conquer Join (DCJ), entwickelt, welcher die effiziente Ausführung von Set-Containment-Joins unterstützt und die bekannten Algorithmen in vielen Fällen übertrifft. Wir haben eine detaillierte Analyse von DCJ und der bereits existierenden Algorithmen vorgestellt und ihr komparatives Verhalten mit Hilfe einer prototypischen Implementierung studiert.



Ausgewählte Bücher und Publikationen

Bücher

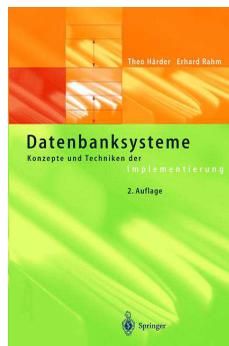
Rahm, E., Vossen, G. (Hrsg.): **Web & Datenbanken** - Konzepte, Architekturen, Anwendungen. dpunkt-Verlag, 2003. Gebunden. 504 Seiten

Härder, T., Rahm, E.: **Datenbanksysteme - Konzepte und Techniken der Implementierung**. Springer-Verlag, März 1999, 2. Auflage 2001. ISBN 3-540-42133-5, Gebunden. 600 Seiten.

Rahm, E.: **Mehrrechner-Datenbanksysteme - Grundlagen der verteilten und parallelen Datenbankverarbeitung**. Addison-Wesley-Verlag / Oldenbourg, ISBN 3-486-24363-2, Okt. 1994, Gebunden. 443 Seiten.

Weikum, G., Schöning, H., Rahm, E. (Hrsg.): **Datenbanksysteme für Business, Technologie und Web**. Proc. der 10. BTW-Tagung, Leipzig, Feb. 2003

Chaudhri, A., Jeckle, M., Rahm, E., Unland, R. (eds.), **Web, Web-Services, and Database Systems**. Springer-Verlag, Lecture Notes in Computer Science (LNCS) 2593, 2003



Ausgewählte Publikationen (seit 2000)

Die meisten der im folgenden genannten Publikationen sind im Volltext auf den Web-Seiten unserer Abteilung (<http://dbs.uni-leipzig.de>) bzw. auf dem Dokumenten-Server der Universität (<http://dol.uni-leipzig.de>) verfügbar.

2002/2003

Melnik, S., Rahm, E., Bernstein, P.: *RONDO - A Programming Platform for Generic Model Management*. **Proc. ACM SIGMOD Conf.**, San Diego, 2003

Do, H.H., Rahm, E.: *COMA - A System for Flexible Combination of Schema Matching Approaches*. **Proc. 28th Intl. Conference on Very Large Databases (VLDB)**, Hongkong, Aug. 2002

Melnik, S., Garcia-Molina, H., Rahm, E.: *Similarity Flooding: A Versatile Graph Matching Algorithm*. **Proc. 18th International Conference on Data Engineering (ICDE)**, San Jose, 2002 (Best Student Paper Award)

Melnik, S., Garcia-Molina, H.: *Adaptive Algorithms for Set Containment Joins*. **ACM Trans. on Database Systems**, Vol. 28 (2), June 2003

Melnik, S., Garcia-Molina, H.: *Divide-and-Conquer Algorithm for Computing Set Containment Joins*, Proc. 8th Int. Conf. on Extending Database Technology (EDBT02), Prague, March 2002

Böhme, T., Rahm, E.: *Multi-User Evaluation of XML Data Management Systems with XMach-1*. Proc. VLDB Workshop Efficiency and Effectiveness of XML Tools, and Techniques, Hongkong, Aug. 2002, Lecture Notes in Computer Science (LNCS) 2590, Springer-Verlag, 2003

Böhme, T., Rahm, E.: *Benchmarking von XML-Datenbanksystemen*. In: Web & Datenbanken. Dpunkt-Verlag, 2003



Do, H.H., Melnik, S., Rahm, E.: *Comparison of Schema Matching Evaluations*. Proc. GI-Workshop "Web and Databases", Erfurt, Oct. 2002, In: Lecture Notes in Computer Science (LNCS) 2593, 2003

Do, H.H., Kirsten, T., Rahm, E.: *Comparative Evaluation of Microarray-based Gene Expression Databases*. Proc. 10. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW 2003), Leipzig, Feb. 2003

Greiner, U., Rahm, E.: *WebFlow: Ein System zur flexiblen Ausführung webbasierter, kooperativer Workflows*. Proc. 10. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW 2003), Leipzig, Feb. 2003

Märtens, H., Rahm, E., Stöhr, T.: *Dynamic Query Scheduling in Parallel Data Warehouses*. Proc. EURO-PAR 2002, Springer-Verlag, LNCS 2400, Paderborn, Aug. 2002

Melnik, S.: *Generic Model Management: Experience and Open Questions*. Doctoral Poster, VLDB Conf. 2002, Hongkong, Aug. 2002

Rahm, E.: *Web Usage Mining*. Datenbank-Spektrum, Vol. 2, Heft 2, Feb. 2002

Rahm, E., Stöhr, T.: *Data Warehouse-Einsatz zur Web-Zugriffsanalyse*. In: Web & Datenbanken. Dpunkt-Verlag, 2003

Sosna, D., Rahm, E.: *Web-basiertes SQL-Training im Bildungsportal Sachsen*. Proc. BTW-Workshop "Datenbanken und E-Learning", Leipzig, Feb. 2003

Spruth, W., Rahm, E.: *Sysplex-Cluster-Technologien für Hochleistungs-Datenbanken*. Datenbank-Spektrum, Vol. 2, Heft 3, Mai 2002



2000/2001

Madhavan, J., Bernstein, P.A., Rahm, E.: *Generic Schema Matching with Cupid*. **Proc. 27th Intl. Conference on Very Large Databases (VLDB)**, Rome, Italy, Sep. 2001

Stöhr, T., Rahm, E.: *Warlock: A Data Allocation Tool for Parallel Warehouses*. **Proc. 27th Intl. Conference on Very Large Databases (VLDB)**, Rome, Italy, Sep. 2001 (software demonstration)

Rahm, E., Bernstein, P.A.: *A Survey of Approaches to Automatic Schema Matching*. **VLDB Journal**, Vol. 10, No. 4, Dec. 2001

Stöhr, T.; Märtens, H.; Rahm, E.: *Multi-Dimensional Database Allocation for Parallel Data Warehouses*. **Proc. 26th Intl. Conference on Very Large Databases (VLDB)**, Cairo, Egypt, Sep. 2000

Bernstein, P.A., Rahm, E.: *Data Warehousing Scenarios for Model Management*. Proc. 19th Int. Conf. on Entity-Relationship Modelling (ER'00), LNCS 1920, Springer-Verlag, Salt Lake City, Oct. 2000

Böhme, T., Rahm, E.: *Benchmarking XML Database Systems - First Experiences*. Proc. 9th Int. Workshop High Performance Transaction Systems (HPTS), Sep. 2001

Böhme, T., Rahm, E.: *XMach-1: A Benchmark for XML Data Management*. Proc. 9. Fachtagung Datenbanksysteme für Büro (BTW 2001), Technik und Wissenschaft, Springer-Verlag, Oldenburg, März 2001

Do, H. H.; Stöhr, T.; Rahm, E.; Müller, R.; Dern, G.: *Evaluierung von Data Warehouse-Werkzeugen*. Proc. Data Warehousing (DW) 2000, Friedrichshafen, Nov. 2000



Märtens, H., Rahm, E.: *On Parallel Join Processing in Object-Relational Database Systems*. Proc. 9. Fachtagung Datenbanksysteme für Büro (BTW 2001), Technik und Wissenschaft, Springer-Verlag, Oldenburg, März 2001

Märtens, H.: *A Classification of Skew Effects in Parallel Database Systems*. Proc. 7th Intl. Euro-Par Conference (Euro-Par 2001), LNCS 2150, Springer-Verlag, Manchester, August 2001

Melnik, S.: *Declarative Mediation in Distributed Systems*, Proc. 19th Int. Conf. on Entity-Relationship Modelling (ER'00), LNCS 1920, Springer-Verlag, Salt Lake City, Oct. 2000

Melnik, S., Raghavan, S., Yang B., Garcia-Molina, H.: *Building a Distributed Full-Text Index for the Web*, Proc. 10th WWW Conf., Hong Kong, May 2001. Extended version in: ACM Trans. on Information Systems (TOIS), Vol. 19, No. 3, pp. 217-241, July 2001

Melnik, S., Rahm, E., Sosna, D.: *DOL - an Interoperable Document Server*. Proc. 31. GI-Jahrestagung, Wien, Sep. 2001

Müller, R., Rahm, E.: *Dealing with Logical Failures for Collaborating Workflows*. Proc. CoopIS 2000, LNCS 1901, Springer-Verlag, Eilat, Israel, Sep. 2000

Rahm, E., Do, H.H.: *Data Cleaning: Problems and Current Approaches*. IEEE Techn. Bulletin on Data Engineering, Dec. 2000

Stöhr, T.; Rahm, E., Quitzsch, S.: *OLAP-Auswertung von Web-Zugriffen*. Proc. GI-Workshop Internet-Datenbanken, Berlin, Sep. 2000

Stöhr, T.: *Analytische Bestimmung einer Datenallokation für Parallele Data Warehouses*. Proc. 9. Fachtagung Datenbanksysteme für Büro, Technik und Wissenschaft (BTW 2001), Springer-Verlag, Oldenburg, März 2001



Lehrangebot

Vorlesungen

Grundstudium

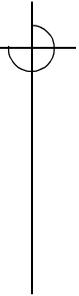
- Datenbanksysteme 1
- Datenbanksysteme 1 für Wirtschaftsinformatiker
- Datenbanksysteme 2 für Wirtschaftsinformatiker
- Programmierung und Programmiersprachen
- Algorithmen und Datenstrukturen 1
- Algorithmen und Datenstrukturen 2

Hauptstudium

- Datenbanksysteme 2
- Data Warehousing und Data Mining
- Mehrrechner-Datenbanksysteme
- Implementierung von Datenbanksystemen 1
- Implementierung von Datenbanksystemen 2
- Datenschutz und Datensicherheit
- Workflow-Management-Systeme
- Geoinformationssysteme 1+2

Zu den Vorlesungen im Grundstudium und den Vorlesungen Datenbanksysteme werden Übungen angeboten. Umfassende Lernmaterialien (Vorlesungsskripte, Übungsaufgaben, Literaturempfehlungen) finden sich unter <http://dbs.uni-leipzig.de>.





Praktika und Seminare

Zusätzlich zu den Vorlesungen wird regelmäßig ein Datenbank-Praktikum für Studenten im Hauptstudium angeboten. Es beinhaltet den Entwurf einer Datenbank sowie die Realisierung von Datenbankanwendungen unter praxisnahen Bedingungen. Die Aufgaben sind mit dem kommerziellen SQL-Datenbanksystem DB2 im Rahmen einer Unix-basierten Client/Server-Umgebung zu bearbeiten.

Desweiteren findet ein Problemseminar zu wechselnden Themen statt. In den letzten Jahren lag der Schwerpunkt in den Bereichen E-Commerce, Datenintegration und Bio-Datenbanken.

Diplomarbeiten

Die Abteilung vergibt laufend attraktive Diplomarbeiten, u.a. zu

- Datenbankeinsatz in der Bioinformatik
- XML-Datenbanken
- Web Usage Mining
- Workflow Management / Web Services
- innovativen web-basierten DB-Anwendungen.

Im Rahmen studentischer Arbeiten sind bereits mehrere komplexe datenbankbasierte Implementierungen entstanden, u.a.

- der Dokumenten-Server der Universität
- die Web-Präsentation des Musikinstrumentenmuseums der Universität
- das E-Procurement-System der Universität zur Online-Materialbestellung.

Das in jeden Semester stattfindende Diplomanden-Seminar findet im Sommer in Verbindung mit dem Mitarbeiter-Oberseminar in Zingst/Ostsee im ehemaligen Observatorium der Universität statt.



<http://dbs.uni-leipzig.de>