3.3.5 Datenbanken

3.3.5.1 Personelle Zusammensetzung

Univ.-Professor	Prof. Dr. Erhard Rahm
wiss. Mitarbeiter	Dr. Dieter Sosna
wiss. Mitarbeiter	Timo Böhme
wiss. Mitarbeiter	Robert Müller
wiss. Mitarbeiter (DFG)	Holger Märtens
wiss. Mitarbeiter (DFG)	Ulrike Greiner
wiss. Mitarbeiter (Industrie)	Hong Hai Do
wiss. Mitarbeiter (Industrie)	Thomas Stöhr
Stipendiat (Graduiertenkolleg)	Sergej Melnik
Stipendiatin (Graduiertenkolleg)	Natalya Sklyar (bis Oktober 2001)
Programmierer	Stefan Jusek
Sekretärin	Andrea Hesse

3.3.5.2 Highlights

Im Berichtsjahr sind folgende Ereignisse besonders hervorzuheben:

- Der an der Abteilung entwickelte weltweit erste Benchmark zur XML-Datenverwaltung, XMach-1, wurde im März im Rahmen der BTW-Tagung zusammen mit ersten Messergebnissen vorgestellt.
- Im Rahmen der BTW-Tagung in Oldenburg wurde beschlossen, die 10. BTW-Tagung 2003 nach Leipzig zu vergeben. Die BTW wurde 1985 gegründet.und ist die führende deutschsprachige Konferenz zu Datenbanken und ihren Anwendungen. Nähere Informationen siehe *http://www.btw2003.de*
- Im Rahmen der GI-Jahrestagung in Wien wurde der von Prof. Rahm initiierte neue Arbeitskreis "Web und Datenbanken" offiziell gegründet. Am Jahresende waren bereits über 100 Arbeitskreis-Mitglieder registriert. Nähere Informationen siehe *http://dbs.uni-leipzig.de/webdb*
- Das seit 1997 laufende DFG-Projekt "Datenallokation und dynamische Lastbalancierung in Parallelen Datenbanksystemen" wurde im Oktober 2001 sehr erfolgreich abgeschlossen. Das im Rahmen des Projekts entwickelte Werkzeug *Warlock* zur automatischen Bestimmung einer Datenallokation für parallele Data Warehouses wurde im September auf der 27. Int. VLDB-Tagung in Rom vorgestellt.
- Das Oberseminar der Abteilung wurde im Juni 2001 erstmals und mit großem Erfolg außerhalb von Leipzig durchgeführt, und zwar in der Universitätsaußenstelle Zingst (Ostsee).

- Im Herbst wurde eine mehrtägige F&E-Werkstatt zum Thema "Kundenmonitoring auf der Basis von Web Mining *und* Data Warehousing" für Versicherungsunternehmen mit der Leipziger Gesellschaft für versicherungswissenschaftliche Forschung erfolgreich durchgeführt.
- Die Arbeiten zum Schema Matching kommen gut voran und konnten prominent publiziert werden (VLDB-Journal, VLDB01-Tagung, Data Engineering 2002).
- Die Arbeiten zur Bioinformatik wurden begonnen. Sie sind im Rahmen des nach dem gewonnenen DFG-Wettbewerb im Aufbau befindlichen Interdisziplinären Zentrum für Bioinformatik (IZBI) integriert.

3.3.5.3 Projekte

XML-Unterstützung in Datenbanksystemen (Böhme, Rahm)

Die zunehmende Bedeutung von XML als Datenaustauschformat wird durch eine Reihe von auf XML aufbauenden Informations- und Dienstprotokollen, wie z.B. SOAP oder UDDI, unterstrichen. Weiterhin wird XML in zunehmenden Maße als natives Datenformat eingesetzt und stellt in Form von XHTML den Nachfolger von HTML. Das Ergebnis dieser Entwicklung ist ein sprunghafter Anstieg von XML-formatierten Daten. Eine effiziente Verwaltung dieser Daten mit der Möglichkeit der strukturellen und semantischen Informationsrecherche auf großen Kollektionen bedingt den Einsatz von XML-fähigen Datenbanken. Mit seinem dokumenten-orientierten Ursprung stellt XML besondere Anforderungen an die Datenbanksysteme, da die XML-Daten typischerweise nur semistrukturiert vorliegen, z.B. liegen Schemainformationen nur unvollständig oder nur innerhalb der Daten selbst vor und eine strikte Typvergabe an Elemente oder Attribute ist mit XML selbst nicht möglich.

Aus diesem Grund werden unterschiedliche Architekturen für die Verwaltung von XML-Daten propagiert. Diese lassen sich in zwei große Klassen einteilen: native XML-Datenbanksysteme und relationale bzw. objekt-relationale Datenbanksysteme, die mit einer Erweiterung zum Abspeichern und Manipulieren der XML-Daten versehen sind. Native XML-Datenbanksysteme erlauben den Zugriff auf die Daten über für XML entworfene Schnittstellen, wie z.B. XPath oder DOM und sollten eine weitgehende Unterstützung der spezifischen XML-Anforderungen aufweisen. Demgegenüber bieten relationale Datenbanken mit XML-Unterstützung den Zugriff über SQL-Erweiterungen und sind für die Verwaltung großer strukturierter Datenkollektionen prädestiniert.

Um eine Abschätzung der Leistungsfähigkeit der verschiedenen Architekturen zur Verwaltung von XML Daten zu ermöglichen, wurde in diesem Projekt ein skalierbarer Multi-User Benchmark entwickelt. Dieser, als XMach-1 (XML Data Management Benchmark) bezeichnete Benchmark, modelliert eine XML-Dokumentenkollektion, die unterschiedliche Datentypen enthält, auf welche über eine Web-Schnittstelle zugegriffen wird. Im Benchmark werden die Datenbankstruktur, die Generierung der Daten, sowie Anfrage- und Updateoperationen spezifiziert. Die Durchsatzleistung wird in Xqps (XML queries per second) unter Einhaltung bestimmter Antwortzeitrestriktionen gemessen. Der

Benchmark betont den semistrukturierten Charakter von XML und untersucht schwerpunktmäßig die Eignung der Systeme bzgl. dieser Eigenschaft.

Nach Abschluss der Spezifikation erfolgte die Implementierung des Benchmarks. Duch eine strikte Modularisierung und Aufteilung in einen generischen und einen systemabhängigen Teil, kann der Benchmark mit vergleichsweise geringem Aufwand für bestimmte Datenbanksysteme angepaßt werden, wobei die Vergleichbarkeit der Daten gesichert wird. In Kooperation mit den Herstellern wurde der Benchmark für eine Anzahl von nativen Datenbanksystemen adaptiert und diese Systeme mit dem Benchmark evaluiert. Eine Anpassung an relationale Datenbanksysteme ist in Arbeit.

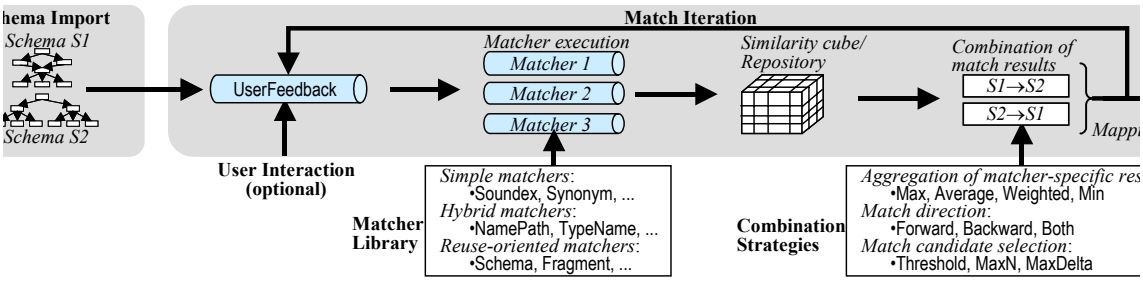
Schema-Matching und Model Management (Do, Melnik, Rahm)

Bei der Kopplung von Datenbanken und Web-Anwendungen spielt die Schema-Integration eine herausragende Rolle. Model-Management-Systeme helfen dabei, Datenbank-Schemas, Kataloge, Mediator-Definitionen usw. zu verwalten und zu manipulieren. Solche Schemas und Definitionen werden als eingenständige Objekte (genannt Modelle) in einer Datenbank abgelegt. Um den Programmieraufwand beim Zugriff auf solche Objekte zu reduzieren, wird eine Reihe von Operationen höheren Abstraktiongrades zur Verfügung gestellt. Der sog. Match-Operator ist dabei einer der mächtigsten Operatoren. Dieser erlaubt es, ähnliche Strukturelemente in Modellen aufzuspüren. Die Match-Ergebnisse können dann unter anderem zur Schema-Integration verwendet werden.

Bisherige Vorschläge zur Realisierung von Schema Matching wurden umfassend analysiert und im Rahmen einer Lösungstaxonomie klassifiziert. Die Ansätze verwenden neben Angaben auf Schemaebene (Namen, Datentypen, Integritätsbedingungen etc.) zum Teil auch Informationen auf Instanzenebene. Zum strukturbasierten Matching wurde ein generischer iterativer Algorithmus namens *Similarity Flooding* entwickelt. Der Flooding-Algorithmus akzeptiert zwei Graphen als Eingabe und liefert eine gewichtete Abbildung zwischen einzelnen Knoten dieser Graphen zurück. Bei der Suche nach strukturähnlichen Elementen wird die Intuition ausgenutzt, daß zwei Knoten ähnlich sind, wenn ihre Umgebungsknoten Ähnlichkeit aufweisen. Der Ablauf des Algorithmus entspricht einer Eigenvektor-Berechung durch ein Fixpunkt-Verfahren. Die Leistungsfähigkeit des entwickelten Ansatzes wurde in einer Benutzer-Studie anhand einer Reihe von Schema-Matching-Aufgaben bestätigt. Ferner wurde begonnen, die Einbindung eines deratigen Match-Operators in einem umfassenderen Model Management-Prototyp zu untersuchen.

Ferner wurde ein Prototyp namens COMA (composite match system) entwickelt, mit dem unterschiedliche Match-Algorithmen in flexibler Weise kombiniert werden können. COMA repräsentiert ein generisches Match-System und unterstützt verschiedene Schematypen, wie z.B. XML und relationale Schemas. Seine Kernkomponenten sind einerseits eine Bibliothek von Match-Algorithmen, darunter auch neuartigen Algorithmen zur Wiederverwendung von existierenden Match-Ergebnissen, und andererseits eine Bi-

bliothek von Kombinationsmöglichkeiten (vgl. Abbildung). COMA ermöglicht die Konfiguration einer zugeschnittenen Match-Strategie durch die Auswahl geeigneter Match-Algorithmen und einer Kombinationsstrategie. Neben dem automatischen Modus unterstützt COMA eine interaktive und iterative Gestaltung des Match-Prozesses unter Nutzerkontrolle. Die Voraussetzung für die Flexibilität von COMA ist die Nutzung eines DBMS-basierten Repository, welches die Schemas, die Zwischenergebnisse einzelner Matchers sowie Endergebnisse durchgeführter Match-Operation verwaltet und für Wiederverwendung bereithält. Die Effektivität einzelner Matchers und Kombinationsmöglichkeiten wurde in einer umfangreichen Evaluierung analysiert und gegenübergestellt.

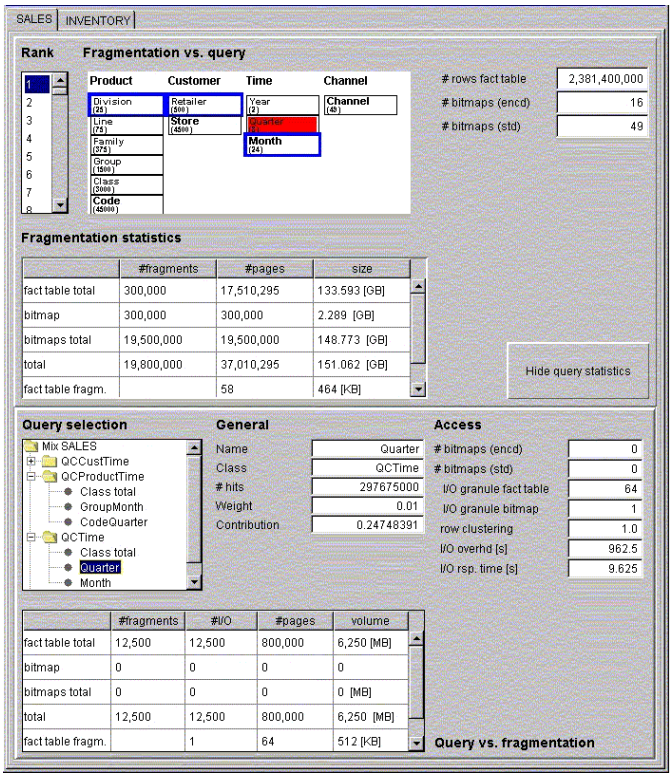


Parallele Datenbanksysteme (Märtens, Rahm, Stöhr)

Unsere von der DFG geförderten Forschungsarbeiten umfassen die Entwicklung und Bewertung von Verfahren zur dynamischer Lastbalancierung und der flexiblen Datenallokation in Parallelen DBS. Wir konzentrieren uns dabei auf sog. Shared-Disk-Architekturen, welche durch die gemeinsame Plattenanbindung ein hohes Potential für diese leistungskritischen Aufgaben aufweisen. Besondere Aufmerksamkeit widmen wir dabei der effektiven Parallelverarbeitung komplexer Anfragen im Mehrbenutzerfall und der dynamischen Behandlung sogenannter Skew-Effekte, welche durch unregelmäßige Werte- und Datenverteilungen zu abweichenden Bearbeitungszeiten einzelner Teilanfragen führen. Ebenso stehen Aspekte der Automatisierung bzw. des "Self-Tuning" von Datenbanksystemen im Vordergrund. Die Leistungsbewertung aller entwickelten Lösungen erfolgt analytisch oder mit praktischen Experimenten im Rahmen eigens entwickelter, umfassender und flexibel parametrisierbarer Simulationsmodelle. Einige Verfahren wurden prototypisch implementiert.

Im Berichtsjahr wurde die begonnene Studie zur Behandlung von Star-Schemas in relationalen Data-Warehouse-Umgebungen fortgesetzt. Ziel sind optimierte Lastbalancierungs- und Datenallokationsstrategien für die in einem solchen Umfeld üblichen hochkomplexen mehrdimensionalen Anfragen (Star-Joins) auf sehr großen Datenmengen.

Bisher gab es keinerlei Werkzeugunterstützung für die komplexe Aufgabe der Bestimmung einer passenden Datenallokation. Auf Basis der im Vorjahr entwickelten mehrdimensionalen, hierarchischen Fragmentierungsmethode für Faktentabellen und Zugriffsstrukturen (MDHF) wurde zunächst ein analytisches Kostenmodell entwickelt, welches die Performance-kritischen Größen E/A-Aufwand und -Antwortzeit eines gewichteten Star-Join-Mix unter MDHF abschätzt. Die integrierte Minimierung beider Größen diente als Basis eines GUI-basierten Prototypen, *Warlock* (s. Abbildung), welcher Fragmentierungskandidaten vorschlägt, das resultierende Anfrageverhalten paralleler Star-Joins analysiert sowie eine Datenalllokation ableitet und visualisiert. Zusätzlich können unterschiedliche Konfigurationen verglichen und optimiert werden. *Warlock* wird über wenige DB- und Hardware-Größen sowie graphische Schema- und Query-Editoren einfach parametrisiert. Eine erste Version des Werkzeugs wurde bei der 27. Int. VLDB-Tagung in Rom im September 2001 präsentiert.



Analyse einer Fragmentierung mit *Warlock*

Für die auf MDHF aufbauende Anfrageverarbeitung wurden zudem Lastbalancierungsstrategien entworfen, welche sowohl die Prozessor- als auch die Plattenauslastung berücksichtigen. In ersten Ergebnissen einer aktuell laufenden Studie zeigen sich die neuen Verfahren gegenüber herkömmlichen, auf jeweils eine Ressource fokussierten Methoden deutlich überlegen.

In einer neuen, umfassenden Klassifikation wurden die verschiedenen Arten von Skew-Effekten in parallelen DBS den zu ihrer Behandlung geeigneten Lastbalancierungsansätzen gegenübergestellt. Hierbei ergaben sich weitere Belege für die Notwendigkeit hochdynamischer Verarbeitungsmethoden, wie sie in der vergangenen Jahren am Lehrstuhl entwickelt worden sind. Für die immer weiter verbreiteten objektrelationalen DBS wurde ein Leistungsvergleich von relationalen und objektorientierten Methoden der parallelen Join-Verarbeitung vorgelegt. Zusätzlich wurde eine Allokationsstrategie speziell für Klassenhierarchien und mengenwertige Attribute präsentiert, die Join-Operationen bestmöglich unterstützen soll.

Web Usage Mining (Rahm, Stöhr)

Der Analyse des Zugriffsverhaltens auf Web-Seiten kommt eine zunehmende Bedeutung zu, insbesondere um Informationen zu Interessen der Nutzer zu gewinnen. Diese Informationen können genutzt werden, etwaige Performance-Probleme zu beheben, inhaltliche Verbesserungen von Web-Auftritten vorzunehmen und generell die Nutzer besser zu bedienen. Im E-Business-Umfeld kommen einer verstärkten Kundenbindung und der Gewinnung neuer Kunden eine besondere Bedeutung zu. Unser Ansatz für ein umfassendes Web Usage Mining geht von der Nutzung eines Data Warehouse aus, in dem die in Web-Logs und sonstigen Informationsquellen verfügbaren Daten entlang verschiedener Dimensionen strukturiert verwaltet werden. Diese Vorgehensweise ist skalierbar hinsichtlich umfangreicher Web-Sites und sehr großer Zugriffsraten und ermöglicht flexible Aggregierungen und Auswertungen, insbesondere verschiedene Data-Mining-Verfahren. Im Unternehmensumfeld läßt sich ferner eine Kopplung mit Geschäftsdaten erreichen, z.B. zum Customer Relationship Management.

Eine erste Prototyp-Realisierung eines derartigen Web-Usage-Data-Warehouse erfolgte für die Web-Server des Instituts für Informatik unter Nutzung von kommerziell verfügbaren Tools, u.a. durch über Web-Browser nutzbare OLAP-Werkzeuge. Zusätzlich wurde eine GUI-gestützte Anwendung zur Auswertung typischer Web-Usage-Mining-Fragestellungen (Nutzernavigation, Top-Seiten, -Besucher, häufige Fehler etc.) integriert.

Der Lehrstuhl Datenbanken führte unter Mitwirkung des Lehrstuhls E-Business der Handelshochschule Leipzig und unter Organisation der Leipziger Gesellschaft für versicherungswissenschaftliche Forschung mbH eine F+E-Werkstatt zum hochaktuellen Thema "Kundenmonitoring auf der Basis von Web Mining und Data Warehousing" durch. Im Rahmen dieses Projekts wurde mit Partnern aus Versicherungsunternehmen ein Konzept zur Umsetzung einer Data-Warehouse-basierten Web-Zugriffsanalyse entwickelt. Daneben wurde der umfangreiche und unübersichtliche Werkzeugmarkt analysiert und die Funktionalität einiger State-of-the-art-Werkzeuge praktisch evaluiert.

Im Berichtszeitraum wurden außerdem erste Untersuchungen zu adaptivem Workflow-Management für kooperierende Workflows durchgeführt. Dazu wurde eine Klassifikation nach Kooperationsform (synchron und asynchron), Unterstützung von Kooperationsbedingungen und Möglichkeiten der lokalen und workflow-übergreifenden Ausnahmebehandlung erstellt, in die dann verschiedene Forschungsansätze eingeordnet wurden, um einen Überblick über den derzeitigen Stand der Forschung zu erhalten. Weitere Arbeiten sind u.a. zum Einsatz von Web-Services bei der Workflow-Kooperation geplant.

Adaptive Workflow-Systeme (Greiner, Müller, Rahm)

Eine wesentliche Limitation derzeitiger Workflow-Systeme besteht in ihrer unzureichenden Flexibilität, auf unerwartete Ereignisse dynamisch 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. Der zugrundeliegende Ansatz verwendet Regelwissen und temporale Heuristiken, 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.

Zur Beschleunigung der Implementierung wurde im Berichtszeitraum das Workflow-System ADEPT der Universität Ulm hinsichtlich eines Einsatzes im Projekt evaluiert. ADEPT bietet eine umfangreiche Java-API, über die sowohl eine temporale Abschätzung von Workflows als auch ein automatischer Umbau von Workflows möglich ist. Es wurden verschiedene Testimplementierungen erstellt, nach deren Auswertung dann entschieden wurde, ADEPT als Grundlage für die Realisierung eines Prototypen zur ereignisgesteuerten, automatischen Workflow-Adaptation zu verwenden.

Für den Prototypen, der in Zusammenarbeit mit dem Institut für Medizinische Informatik, Statistik und Epidemiologie der Universitätskliniken Leipzig (Prof. M. Löffler, Dr. B. Heller, Prof. A. Winter, J. Ramsch) entwickelt wird, wurde ein Architekturansatz erarbeitet Im Mittelpunkt steht ein Programmodul zur dynamischen Workflow-Adaptation, das über die zu verwendende Adaptationsstrategie entscheidet, temporale Abschätzungen durchführt und Workflow-Instanzen automatisch umbaut. Diese Aktivitäten werden über ein Monitoring-Modul auf der Patienten- und Labor-Datenbank ausgelöst. Außerdem sollen eine Wissensbasis mit den Regeln und ein Monitoring-Modul zur Überwachung von Workflow-Instanzen und Erkennung von Ausnahmen implementiert werden. Die Kommunikation zwischen diesen Teilen erfolgt über eine genau definierte Schnittstelle auf der Basis von XML-Nachrichten.

Die Adaptations-Ansätze sind vor allem durch medizinische Anwendungen aus dem Bereich der Krebsbehandlung motiviert, lassen sich aber auch auf andere Anwendungsfelder übertragen. Für die Zukunft ist daher vor allem geplant, die Verwertbarkeit der Ansätze für den E-Commerce-Bereich zu untersuchen.

Im Berichtszeitraum wurden außerdem erste Untersuchungen zu adaptivem Workflow-Management für kooperierende Workflows durchgeführt. Dazu wurde eine Klassifikation nach Kooperationsform (synchron und asynchron), Unterstützung von Kooperationsbedingungen und Möglichkeiten der lokalen und workflow-übergreifenden Ausnahmebehandlung erstellt, in die dann verschiedene Forschungsansätze eingeordnet wurden, um einen Überblick über den derzeitigen Stand der Forschung zu erhalten. Weitere Arbeiten sind u.a. zum Einsatz von Web-Services bei der Workflow-Kooperation geplant.

Web-basiertes Lernen (Rahm, Sosna)

Die Untersuchungen zu elektronischen Biblotheken bzw. web-basiertem Lernen sowie die Nutzung erfolgter Implementierungen wurden fortgesetzt. Beim Betrieb des Dokumentenservers (URL: http://dol.uni-leipzig.de), der es ermöglicht, unterschiedlichste an der Universität erstellte Dokumente im Volltext zu verwalten und flexibel zugänglich zu machen, wurden bisherige Erkenntnisse bei Administration und Nutzung zusammengefaßt und publiziert. Die Untersuchungen zur Verwaltung von Annotationen wurden fortgeführt, insbesondere im Hinblick auf quantitative Anforderungen einer geeigneten Server-Architektur (technischer Bericht).

Ein neues Projekt, "*Internetgestützter SQL-Trainer*", wurde im Rahmen des Verbundprojekts Bildungsportal Sachsen gestartet (gefördert vom SMWK). Es soll zur Unterstützung in der Datenbank-Lehre eine web-basierte Einführung in die standardisierte Datenbankabfragesprache SQL realisieren. 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.

Verteilte Volltext-Indexierung für das WWW (Melnik)

Neben dem effektiven Einsatz von iterativen Ranking-Algorithmen spielt Volltext-indexierung nach wie vor eine signifikante Rolle bei der Web-Suche. Die im WWW vorhandene Sammlung von Volltext-Dokumenten erreichte einen gewaltigen Umfang. Dabei ist ein Großteil der globalen Dokumentenkollektion durch extreme Schnellebigkeit charakterisiert. Diese Faktoren machen verteilte und parallele Indexierungsansätze erforderlich. Im Rahmen dieses Projektes wurden drei Aspekte der verteilten Indexerstellung untersucht:

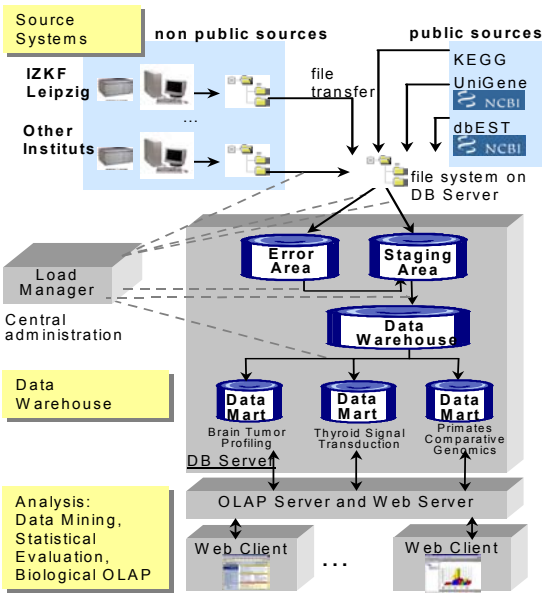
- Wir entwickelten ein Pipelining-Verfahren, welches die Indexierungszeiten erheblich verringert. Die Netzwerk-, CPU-, und Festplattenzugriffe werden parallelisiert, so daß eine höhere Effizienz der Ressourcennutzung möglich wird. Es wurde gezeigt, wie eine optimale Pipeline zu konstruieren ist, die den theoretisch bestmöglichen Durchsatz liefert. Die Ergebnisse wurden experimentell bestätigt.
- Als zweiter Schwerpunkt wurde der Einsatz eingebetteter Datenbanksysteme für die Speicherung und effiziente Suche in invertierten Index-Dateien untersucht. Wir entwickelten und verglichen mehrere Speicherungsverfahren von Indexen in eingebetteten Datenbanken. Die Nutzung solcher Datenbanksysteme reduziert die Komplexität der Indexierungs-Software ermöglicht, die vorhandene Datenbank-Funktionalität wiederzuverwenden.
- Viele Ranking-Methoden, die zur Bestimmung der Reihenfolge von Anfrageergebnissen dienen, setzen detaillierte statistische Informationen über globale und lokale Termhäufigkeiten voraus. Ein dritter Aspekt, der im Rahmen des Projekts bearbeitet wurde, widmete sich der effizienten Sammlung solcher Informationen während der Volltexterstellung. Mehrere Verfahren wurden vorgestellt und evaluiert.

Der entwickelte Prototyp eines verteilten Volltextindexierungssystem, an dem wir die obigen Fragestellungen erforscht haben, wird als Bestandteil des WebBase-Systems an der Stanford University weitergenutzt.

Gene Expression Warehousing (Kirsten, Do, Rahm)

Die Genexpressionsanalyse ist von grundlegender Bedeutung für zahlreiche molekularbiologische Fragestellungen, insbesondere zur Funktionsbestimmung von Genen sowie für evolutionäre Vergleichsstudien. Für praktische Untersuchungen steht den Kooperationspartnern ein Affymetrix-Microarray System zur hochparallelen Analyse von Genexpressionsdaten am Interdisziplinären Zentrum für klinische Forschung (IZKF) der Universität Leipzig zu Verfügung. Die Microarray-Technologie ermöglicht hier eine neue Qualität der Untersuchungen, da mit ihr simultan die Expression tausender Gene bzw. Sequenzen gemessen werden kann. Die damit generierten Massen an Daten sind mit herkömmlichen Methoden nicht sinnvoll auszuwerten und erfordern eine leistungsfähige Datenbank-Lösung, die neben den von Microarrays generierten Daten vielfältige weitere Daten - insbesondere aus im Web zugänglichen Datenquellen - integriert und mit einem Spektrum leistungsfähiger Data Mining-Verfahren analysiert.

Wir verfolgen hierzu einen innovativen Data Warehouse-Ansatz, der eine semantische Integration der heterogenen Daten ermöglicht und eine mehrdimensionale Verwaltung der Daten zur Unterstützung unterschiedlicher Vergleichsstudien verfolgt. Neben der gezielten Abfrage nach der Expression einzelner Gene sind besonders Fragestellungen des globalen Clustering der Expression von Proteinfamilien aufgrund struktureller oder funktioneller Parameter möglich. Dadurch ergeben sich Einblicke, die bei herkömmlichen Einzelanalysen in der grossen Datenmenge nicht erkennbar sind. Der Aufbau eines solchen Data Warehouse ist sehr aufwendig, ermöglicht jedoch eine hohe Leistungsfähigkeit und Flexibilität hinsichtlich der Unterstützung unterschiedlichster Studien.



In der ersten Phase des Projektes wurde in Zusammenarbeit mit einer Anwendergruppe mit einer umfangreichen Anforderungsanalyse begonnen. Es wurde eine Systemarchitektur entwickelt, die auf dem Data Warehouse Ansatz aufbaut und die lokalen Gegebenheiten im IZKF berücksichtigt. Darüber hinaus entstanden ein Entwurf für ein konzeptionelles Datenmodell für Genexpressionsdaten zur Implementierung in der Warehouse-Datenbank und ein Load-Managment-Konzept zur Population und periodischen Aktualisierung der Datenbank (s. Abbildung).

DBMS-basierte Verwaltung und Analyse von EST-Daten für unterschiedliche Gewebe (Do, Rahm)

Expressed Sequence Tags (ESTs) sind kurze Sequenzen von typischerweise 300-500 Basen, die durch Ansequenzieren von cDNA-Klonen generiert werden und in öffentlichen Sequenzdatenbanken zur Verfügung stehen. Die Häufigkeit des Vorhandenseins von ESTs in cDNA-Banken bestimmter Gewebe bzw. Organe kann Information über Spezifität der in den entsprechenden Geweben bzw. Organen exprimierten Gene geben. Die in den öffentlich zugänglichen Datenbanken verfügbaren grossen Mengen an EST-Daten bieten daher eine geeignete Basis für die Identifikation und vergleichende und funktionale Analyse von spezifisch exprimierten Genen und erlauben die Erstellung qualitativer und quantitativer Aussagen zur Genexpression in verschiedenen Geweben.

- Bisherige Realisierungen computergestützter (in silico) Analyseverfahren von EST-Daten zur Suche nach unbekannten Genen basieren auf einer einfachen, meist dateiorientierten Datenverwaltung. Dies führt zu erheblichen Einschränkungen, u.a. kleine Datenmengen, Beschränkung auf ein bzw. wenige Gewebe bzw. Organ(e), fixierte und unkomfortable Abfragemöglichlichkeiten sowie manuelle Ergebnisbewertung. Das Hauptziel dieses Projekts umfasst den Entwurf und die Implementierung einer neuartigen und umfassenden Datenbanklösung zur Überwindung der Beschränkungen bisheriger Studien. Basierend auf einer zentralen, DBMS-basierten Datenverwaltung sollen mächtige Werkzeuge zur weitgehenden Automatisierung der EST-Clusterung und -Analyse entwickelt werden, mit denen die Suche nach neuen, funktionsspezifischen Genen unterstützt werden soll. Die EST-Datenbank stellt eine sehr nützliche Ergänzung des Genexpressions-Data Warehouse dar.

3.3.5.4 Publikationen

- Härder, T., *Rahm, E.*: Datenbaksysteme: Konzepte und Techniken der Implementierung, 2. Auflage. 600 Seiten, Springer 2001
- *Böhme, T.; Rahm, E.*: XMach-1: A Benchmark for XML Data Management, Proc. of BTW2001, Springer, Berlin 2001.
  http://dol.uni-leipzig.de/pub/2001-1
- *Böhme, T.; Rahm, E.*: Benchmarking XML Database Systems - First Experiences, Position Paper, Ninth International Workshop on High Performance Transaction Systems (HPTS), Pacific Grove, California, 2001. http://dol.uni-leipzig.de/pub/2001-31
- *Greiner, U.*: Adaptives Workflow-Management für kooperierende Workflows. In Tagungsband zum 13. GI-Workshop "Grundlagen von Datenbanken", Juni 2001, Gommern
  http://dol.uni-leipzig.de/pub/2001-26
- *Märtens, H.; Rahm, E.*: On Parallel Join Processing in Object-Relational Database Systems. Proc. 9. Fachtagung Datenbanksysteme für Büro, Technik und Wissenschaft (BTW 2001), Oldenburg, Springer-Verlag, März 2001.
  http://dol.uni-leipzig.de/pub/2001-4

- *Märtens, H.:* A Classification of Skew Effects in Parallel Database Systems. Proc. 7th International Euro-Par Conference (Euro-Par 2001), Manchester, LNCS 2150, Springer-Verlag, August 2001.
  http://dol.uni-leipzig.de/pub/2001-20
- *Melnik, S.,* Garcia-Molina, H.; *Rahm, E.*: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching, Proc. 18th ICDE, San Jose CA, Feb 2002 (Best Student Paper Award)
- *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
- *Melnik, S.*, Raghavan, S., Yang, B.; Garcia-Molina, H.: Building a Distributed Full-Text Index for the Web, 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. GI-Jahrestagung, Wien, Sep. 2001. http://dol.uni-leipzig.de/pub/2001-27
- Madhavan, J., Bernstein, P.A., *Rahm, E.:* Generic Schema Matching with Cupid. Proc. 27th Intl. Conference on Very Large Databases (VLDB), Rom, Sep. 2001, 49-58. http://dol.uni-leipzig.de/pub/2001-28
- *Rahm, E.*, Bernstein, P.A.: A Survey of Approaches to Automatic Schema Matching. VLDB Journal 10 (4): 334-350 (2001)
- *Rahm, E.*: Web Usage Mining, Datenbank-Spektrum, Nr. 2, Feb. 2002
- *Stöhr, T.:* Analytische Bestimmung einer Datenallokation für Parallele Data Warehouses. Proc. 9. Fachtagung Datenbanksysteme für Büro, Technik und Wissenschaft (BTW 2001), Oldenburg, Springer-Verlag, März 2001.
  http://dol.uni-leipzig.de/pub/2001-3
- *Stöhr, T.; Rahm, E.*: *Warlock*: A Data Allocation Tool for Parallel Warehouses. Proc. 27th Intl. Conference on Very Large Databases (VLDB), Rom, Sep. 2001 (Software-Demonstration)
  http://dol.uni-leipzig.de/pub/2001-23
- *Do, H.H., Rahm, E.:* COMA - a flexible system for combining schema match approaches. Techn. Bericht, Dez. 2001
- *Greiner, U.*: Adaptives Workflow-Management für kooperierende Workflows- ein Überblick. Ifi-Report 5-2001, Mai 2001, Universität Leipzig
  http://dol.uni-leipzig.de/pub/2001-18
- *Märtens, H.:* A Classification of Skew Effects in Parallel Database Systems. Ifl-Report 3/2001, Universität Leipzig, Mai 2001.
  http://dol.uni-leipzig.de/pub/2001-21
- *Märtens, H.; Rahm, E.; Stöhr, T.:* Dynamic Query Scheduling in Parallel Data Warehouses. Ifl-Report 6/2001, Universität Leipzig, November 2001.
  http://dol.uni-leipzig.de/pub/2001-38
- *Rahm, E.*; Bernstein, P.A.: On Matching Schemas Automatically. Techn. Report.
  http://dol.uni-leipzig.de/pub/2001-5

- *Rahm, E., Märtens, H., Stöhr, T.:* Abschlussbericht des DFG-Projektes "Datenallokation und dynamische Lastbalancierung in Parallelen Datenbanksysteme", Nov. 2001
- *Rahm, E.,* Spiliopoulou, M., *Stöhr, T.,* Pohle, C., Winkler, K: Kundenmonitoring auf der Basis von Web Mining und Data Warehousing. Abschlussbericht der gleichnamigen F+E-Werkstatt, Dez. 2001
- *Sklyar, N.*. Survey of existing Bio-Ontologies, Techn. Report 5/2001, Dept. of Comp. Science, Univ. of Leipzig. http://dol.uni-leipzig.de/pub/2001-30
- *Sosna, D.*: Der Annotationsserver als mathematischer Bedienprozeß, Teil 1. (Techn.Bericht). http://dol.uni-leipzig.de/pub/2001-46

3.3.5.5 Vorträge

- *Böhme, T.*: XMach-1: A Benchmark for XML Data Management, BTW 2001, Oldenburg, 8. März 2001.
- *Böhme, T.:* XMach-1 & XML-Forschung. DB-Workshop, Zingst, Juni 2001
- *Böhme, T.*: Benchmarking XML Database Systems - First Experiences, Ninth International Workshop on High Performance Transaction Systems (HPTS), Pacific Grove, California, Okt. 2001
- *Do, H.H.*: Generisches Metadaten-Management - Realisierung des Modellmanagements. DB-Workshop, Zingst, Juni 2001
- *Do, H.-H.*: *DBMS-based EST Clustering and Profiling for Gene Expression Analysis*. 1. Workshop Computational Biology in Saxony: Problem and Perspectives. Dresden, Nov. 2001
- *Greiner, U.*: Adaptives Workflow-Management für kooperierende Workflows. GI-Workshop "Grundlagen von Datenbanken", Juni 2001, Gommern
- *Greiner, U.:* Adaptives Workflow-Management für kooperierende Workflows. DB-Workshop, Zingst, Juni 2001
- *Märtens, H.:* Parallele Join-Verarbeitung in objektrelationalen DBS. BTW 2001, Oldenburg, 8. März 2001.
- *Märtens, H.:* Dynamische Lastbalancierung in parallelen Datenbanksystemen. DB-Workshop, Zingst, Juni 2001
- *Märtens, H.:* A Classification of Skew Effects in Parallel Database Systems. Euro-Par 2001, Manchester, 29. August 2001.
- *Melnik, S.*: Semantic Web: A Database Perspective, IBM Almaden, San Jose, Dec 2001
- *Rahm, E.*: Metadaten-Verwaltung für heterogene Informationssysteme. Eingeladener Vortrag. GI-Workshop "Grundlagen von Datenbanken", Juni 2001, Gommern
- *Rahm, E.*: DOL – An Interoperable Document Server. GI-Jahrestagung, Sep. 2001, Wien
- *Rahm, E.*: Benchmarking von XML-Datenbanksystemen. VERTIS-Tagung, Bamberg, Okt. 2001

- *Rahm, E.*: Einführung in Web Usage Mining. Institut für Versicherungswissenschaften, Leipzig, Sep. 2001
- *Rahm, E.*: Umsetzungskonzept zur Web-Zugriffskontrolle für Versicherungsunternehmen. Institut für Versicherungswissenschaften, Leipzig, Nov. 2001
- *Sklyar, N.*: Ontology-based Integration of Genomic Resources. DB-Workshop, Zingst, Juni 2001
- *Sosna, D.*: Der Annotationsserver als mathematischer Bedienprozeß. DB-Workshop, Zingst, Juni 2001
- *Stöhr, T.*: Analytische Bestimmung einer Datenallokation für Parallele Data Warehouses. 9. Fachtagung Datenbanksysteme für Büro, Technik und Wissenschaft (BTW 2001), Oldenburg, 8. März 2001
- *Stöhr, T.*: Datenallokation in parallelen Datenbanksystemen. DB-Workshop, Zingst, Juni 2001
- *Stöhr, T.*: *Warlock*: A Data Allocation Tool for Parallel Warehouses. 27th Intl. Conference on Very Large Databases (VLDB), Rom, 12./13. Sep. 2001 (Software-Demonstration)
- *Stöhr, T.*: Werkzeuge für Web Usage Mining (Vortrag+Demos). Institut für Versicherungswissenschaften, Leipzig, 2001

3.3.5.6 Implementierungen

- *Müller, R., Greiner, U.*: AGENTWORK: Ein adaptives Workflow-Management-System.
- *Melnik, S., Rahm, E., Sosna, D.*: Dokumenten-Server der Universität , http://dol.uni-leipzig.de
- *Do, H.-H., Rahm, E.*: COMA - ein System zum automatischen Schema-Matching
- *Märtens, H.; Stöhr, T.*: komplexe Simulationssysteme von parallelen DBS
- *Stöhr, T.*: Werkzeug zur Bestimmung einer Datenallokation für parallele Data Warehouses (*Warlock*)
- Prototypische Realisierung eines Web-Usage-Data-Warehouse (Diplom- und Hiwi-Arbeiten)