

Dynamische Informationsfusion

Workshop am Dienstag, 21.9.2004

Veranstaltet von dem GI Arbeitskreis „Web und Datenbanken“

Alfons Kemper (TU München), Erhard Rahm (Universität Leipzig),
Bernhard Seeger (Universität Marburg), Gerhard Weikum (MPI für Informatik, Saarbrücken)

1. Problemstellung

Wir stehen erst am Beginn des „Informationszeitalters“. Alles was wir sehen, lesen, hören, schreiben und messen kann bald im Rechner verfügbar gemacht werden. Über das Internet rücken neben unseren eigenen Daten prinzipiell alle Daten aller Menschen und Organisationen in Zugriffsreichweite. Zusätzlich zu den traditionellen Formen persistent gespeicherter Daten entstehen zunehmend kontinuierliche Ströme von Daten, die durch Sensoren oder „News-Ticker“ gespeist werden. Große Bedeutung haben Peer-to-Peer-Datenverbände erlangt, die durch eine sehr hohe Zahl von typischerweise nur temporär zugänglichen Datenquellen gekennzeichnet sind. Die dynamische, bedarfsgesteuerte Fusion dieser vielfältigen Datenquellen ist für viele Anwendungsbereiche essentiell, dazu zählen insbesondere E-Business, E-Science, Katastrophenmanagement und E-Health.

Verteilte Informationssysteme der aktuellen Generation koppeln Datenquellen mit Anwendungsprogrammen über eine verteilte Softwareinfrastruktur, die als Middleware bezeichnet wird (z.B. J2EE oder Web-Services). Dieser Ansatz erfordert à priori eine Festlegung der gekoppelten Datenbestände sowie vor allem eine intellektuell durch die Anwender selbst zu erarbeitende Integration ihrer Querbeziehungen sowohl auf der Ebene der Datenbeschreibungen (sog. Schemata) als auch der Datenausprägungen. Zu den typischen Vertretern dieser Architektur zählen sog. Data-Warehouses in großen Unternehmen und Web-basierte Wissenschaftsportale wie z.B. SRS für den Zugriff auf spezialisierte Datenbanken mit Genom- und Moleküldaten. Der Schwachpunkt dieses Ansatzes ist der signifikant hohe intellektuelle Aufwand bei der Integration weiterer Datenquellen.

Gegenüber dieser etablierten Architektur ergibt sich akuter Forschungsbedarf aus zwei Gründen: Erstens erweist sich im Zeitalter der Informationsexplosion die statische und nur schwach teilautomatisierte Integrationsmethodik als unzulänglich. Es ergibt sich die Notwendigkeit der dynamischen, hochgradig automatisierten Kopplung vieler Datenquellen. Zweitens werden aktuelle Technologietrends wie Grid-Computing und Web-Services schon bald eine verbesserte Softwareinfrastruktur für den Datenaustausch und die Interoperabilität von Anwendungsdiensten bereitstellen. Diese Mechanismen sind jedoch nur die erste Vorstufe für eine darauf aufzubauende „semantische“ Datenintegration.

Im klassischen Gebiet der Datenintegration bahnt sich daher ein Paradigmenwechsel an, der folgenden Faktoren Rechnung tragen soll:

- *Daten, Ströme, Datenanalysemodelle:* Neben persistent gespeicherten Daten entstehen kontinuierliche Ströme von Daten (z.B. Sensordaten, Webnutzerverhalten), die nur in zeitlichen Ausschnitten verarbeitet werden können. Zusätzlich ergeben sich berechnete Modelle zur Datenanalyse (z.B. Klassifikatoren), die durch Training erworbenes Wissen bereitstellen.
- *Skalierbarkeit:* Über Internet, Grid-Computing und Web-Services befinden sich im Prinzip alle weltweit persistenten Daten, Ströme und Modelle in Zugriffsreichweite. Die effiziente Auswahl und Kopplung der relevantesten Komponenten aus einer großen und schnell wachsenden Gesamtmenge stellt extreme Anforderungen an die Skalierbarkeit.

- *Dynamik*: In neuartigen Anwendungen aus den Bereichen E-Business, E-Science, Katastrophenmanagement und E-Health entstehen Kooperationsstrukturen in hochgradig dynamischer Weise. Dies erfordert eine Ad-hoc-Integration der relevanten Daten, Ströme und Modelle.
- *Datenqualität*: Daten haben unterschiedliche Aktualität, Dimensionen, Präzisions- und Vollständigkeitsgrade und erfordern zur korrekten Interpretation oft umfangreiche Zusatzangaben. Methoden für den sauberen Umgang mit unterschiedlichen Datenqualitäten sind daher extrem wichtig.

Die bedarfsgetriebene, skalierbare Kopplung und Integration von Datenbanken, Datenströmen und Datenanalysemodellen bezeichnen wir als *dynamische Informationsfusion*. Im Gegensatz zu früher verfolgten „schemaorientierten“ Integrationsansätzen, die sich lediglich auf die Datenbeschreibungen der Quellen (z.B. das Vorhandensein eines Attributs „Kunde“ in einer Quelle und „Firmenkunden“ oder „Firma“ in einer anderen) abgestützt haben, soll ein methodischer Schwerpunkt auf der „inhaltsorientierten“ Fusion liegen, bei der z.B. durch Data-Mining-Techniken statistisch signifikante Vorkommen gleicher Attributwerte und damit vergleichbare Attribute bestimmt werden können (z.B. im einfachsten Fall häufige Vorkommen von „SAP AG“ sowohl im Attribut „Kunde“ als auch im Attribut „Firma“, ggf. mit Varianten von Attributwerten und Abhängigkeiten verschiedener Werte). Einen zweiten methodischen Schwerpunkt bilden neue Formen der Metadatenverwaltung mit deskriptiven Operatoren zur Analyse und Manipulation unterschiedlichster Metadatenmodelle. Eine Schlüsselrolle kommt dabei der Nutzung ontologischer Metadaten zu (mit deren Hilfe man z.B. ermitteln könnte, dass eine AG eine Firma ist).

2. Anwendungen

E-Science: Unter E-Science versteht man die Internet-basierte dynamische Kooperation verschiedener Forschungsgruppen durch die effektive Nutzung bisher erzielter Ergebnisse (z.B. experimenteller Daten in der Bioinformatik). Einer Gruppe wird dadurch ermöglicht, die für sie relevanten Kooperationen zu finden und deren Daten zu nutzen. Diese Daten sollen künftig als Datenstrom direkt aus einer Simulation oder Messung abgegriffen werden können, wobei die Gewährleistung qualitativ hochwertiger Daten von entscheidender Bedeutung ist. Über die eigentlichen Daten hinaus sind die daraus abgeleiteten Hypothesen und Modelle von großem Interesse. Insgesamt lässt sich durch E-Science die Forschungsproduktivität erheblich verbessern, wobei die Forschungsqualität jederzeit verifizierbar ist.

Katastrophenmanagement: Ein Bereich, in dem dynamische Datenfusion einen besonders kritischen Faktor darstellt, ist das Informationsmanagement bei Katastrophen wie z.B. Fluten, Erdbeben oder Terroranschlägen. Bei einer „Jahrhundertflut“ wie im Sommer 2002 müssen innerhalb kürzester Zeit aktuelle Daten von verschiedensten Messstationen und Einsatzplanungssystemen verglichen und abgestimmt werden. Selbst wenn ein Teil dieser Daten bereits prophylaktisch integriert wurde, erfordert ein grenzübergreifendes Katastrophenmanagement ein hohes Maß an dynamischer Ad-hoc-Fusion. In einer Katastrophensituation kann jede Stunde, die man bei der Fusion und Analyse von Daten verliert, etliche Menschenleben kosten.

E-Health: Im Gesundheitswesen eröffnet eine dynamische Fusion patientenbezogener Daten große Optimierungspotentiale für die Patientenversorgung und Abwicklung von Verwaltungsvorgängen. Über geschützte Web-Plattformen bereitgestellte und integrierte Patienteninformationen wie etwa allgemeine Personenangaben, ärztliche Diagnosen, Röntgenaufnahmen, Laborbefunde oder erfolgte Impfungen ermöglichen autorisierten Nutzern in Kliniken und Arztpraxen einen optimalen Datenaustausch zur Vermeidung von Mehrfachuntersuchungen, Berücksichtigung von Vorerkrankungen, etc. Im Fall der akuten Versorgung von Unfallopfern kommt es darüber hinaus auf die dynamische Datenfusion unter Echtzeitbedingungen an.

E-Business: Technologien für die offene, verteilte Datenintegration haben eine enorme wirtschaftliche Bedeutung. Deutsche Software-Hersteller wie SAP sind weltweit führend in der Verwaltung betriebswirtschaftlicher Daten von Unternehmen und großen Organisationen. In Zukunft ist jedoch eine firmen- und organisationsübergreifende Informationsverarbeitung notwendig, um international wettbewerbsfähig zu bleiben. Eine wesentliche Voraussetzung dafür liegt in der Integration der jeweils relevanten Daten. Diese können sowohl statischer Natur sein (wie z.B. Produktkataloge) als auch dynamisch in der Form von Datenströmen generiert werden wie im Falle kontinuierlich fortgeschriebener Logistikinformationen. Neben der Datenintegration spielt die adaptive Integration von Datenanalyse- und Geschäftsprozessen eine wichtige Rolle.

3. Für den Workshop besonders relevante Themen

Architekturansätze zur dynamischen Informationsfusion: Klassische Ansätze zur Datenintegration berücksichtigen eine feste Zahl passiver Datenquellen. Neue Architekturkonzepte sind erforderlich zur Fusion von Metadaten und Daten *aktiver* Daten-Provider, die kontinuierliche Datenströme wie Sensor- und Monitoring-Daten, News-Feeds, etc. erzeugen. Ebenso soll eine flexible Zahl *volatiler* Daten-Provider, wie sie etwa in Peer-to-Peer-Umgebungen auftritt, effektiv unterstützt werden.

Management von Metadaten: Neue Fusionsansätze erfordern die effektive Nutzung einer Vielzahl von Metadaten wie Datenbankschemata, Datenstrombeschreibungen, Nutzerprofile und Ontologien, welche in unterschiedlichen Sprachen repräsentiert sein können (XML, RDF, OWL, SQL etc.). Diese Metadaten müssen bei der dynamischen Fusion aufeinander abgebildet und transformiert werden. Dazu muss eine möglichst generische Verwaltung von Metadaten und deren Abbildungen entwickelt werden, bei denen mächtige Operatoren, wie z.B. das Herstellen von Korrespondenzen zwischen zwei Datenbankschemata oder Ontologien, weitgehend automatisiert ablaufen.

Verarbeitung von Datenströmen: Wichtige Einsatzfälle für die Fusion von Datenströmen sind Publish-Subscribe-Anwendungen (mit dynamischer Zuordnung von Datensätzen zu Interessensprofilen) und die Verarbeitung von E-Business-XML-Nachrichten mit hohen Skalierbarkeitsanforderungen. Besondere Herausforderungen entstehen bei der Echtzeitverarbeitung volatiler Datenströme, die aufgrund begrenzter Ressourcen innerhalb kürzester Zeit analysiert werden müssen, z.B. zur sofortigen Erkennung von Lastspitzen, Denial-of-Service-Attacken oder Missbrauchsversuchen in Netzwerken.

Dateninhaltsorientierte Fusion: Bisher verfolgte Integrationsansätze basieren auf Datenbankschemata, mit sehr wechselhaftem Erfolg. Künftig sollen dateninhaltsorientierte Fusionsansätze untersucht werden, insbesondere auch für Datenströme und Daten mit fehlenden oder begrenzt aussagekräftigen Metadaten. Data-Mining-Techniken und statistische Lernverfahren könnten dazu beitragen, Zuordnungen von Daten verschiedener Quellen zu finden, z.B. aufgrund typischer Muster in Datensätzen.

Skalierbarkeit und Selbstorganisation: Die potentiell in die Millionen gehende Zahl an Daten-Providern und Nutzern sowie die damit generierten Datenvolumina stellen extreme Anforderungen hinsichtlich der Skalierbarkeit dynamischer Fusionsansätze. Skalierbarkeit lässt sich nur in Verbindung mit der Minimierung manueller Eingriffe zur Systemverwaltung erreichen. Alle beteiligten Systemkomponenten müssen somit selbstüberwachend, selbstoptimierend und selbstadministrierend ausgelegt sein.

Datenqualität und Auswertungsqualität: Die Problematik der Datenqualität ist bereits in derzeitigen Architekturen wie Data-Warehouses unzureichend behandelt, steigt jedoch aufgrund der Vielfalt und Dynamik der Metadaten und Daten überproportional mit der Zahl der beteiligten Partner und Quellen. Einerseits werden geeignete Metriken und Modelle zur Spezifikation und Bewertung von Datenqualität und -konsistenz und andererseits werden Methoden (z.B. verallgemeinerte Formen von „Data-Cleaning“) für den Umgang mit unterschiedlichen Datenqualitäten unter Berücksichtigung der Herkunft abgeleiteter Daten benötigt.

Bewertung von Fusionierungsansätzen: Zur Bewertung unterschiedlicher Ansätze zur dynamischen Datenfusion müssen geeignete Metriken entworfen und untersucht werden, in die wesentliche Kennzahlen der Datenquellen und Metadaten einfließen. Dazu bedarf es einer Benchmark-Umgebung mit geeigneten Anwendungsdaten zur vergleichenden Analyse verschiedener Verfahren.

Datensicherheit und Datenschutz fusionierter Daten: Datenschutzprobleme entstehen durch die umfassende Verknüpfbarkeit der Daten, beispielsweise durch die Protokollierung des Nutzungsverhaltens im Web und in Peer-to-Peer-Netzwerken. Diese Daten könnten zur Erstellung umfassender Profile zusammengeführt und mit personenbezogenen Daten kombiniert werden. Von besonderer Relevanz für den GI-Workshop sind mit der Informationsfusion zusammenhängende Sicherheitsmaßnahmen wie z. B. Entdeckung von Datenmissbrauch, Anonymisierung von Nutzerdaten und -profilen, dynamische Autorisierung, etc.

Langzeitarchivierung und Evolution: Neue Techniken zur langfristigen Archivierung fusionierter Informationen sind notwendig. Es genügt nicht, Daten persistent zu speichern, da sie möglicherweise im Originalformat und -kontext gar nicht mehr interpretierbar sind. Archivierte Daten müssen dauerhaft zugänglich und verarbeitbar bleiben.

Programmkomitee:

Karl	Aberer	EPFL Lausanne
Stefan	Conrad	Univ. Duesseldorf
Klaus	Dittrich	Univ. Zürich
Norbert	Fuhr	Univ. Duisburg
Georg	Gottlob	Technische Universität Wien
Gerti	Kappel	TU Wien
Alfons	Kemper	TU München
Achim	Kraiss	SAP AG
Wolfgang	Lehner	TU Dresden
Wolfgang	Nejdl	Univ. Hannover
Erhard	Rahm	Univ. Leipzig
Kai-Uwe	Sattler	TU Ilmenau
Harald	Schöning	Software AG
Bernhard	Seeger	Univ. Marburg
Rudi	Studer	Univ. Karlsruhe
Gerhard	Weikum	MPI Saarbrücken