HASSO - PLATTNER - INSTITUTE
for IT Systems Engineering

Datenfusion mit FUSIONPLEX

Seminar “Datenfusion in drei Schritten”
im Wintersemester 2006/2007

Kathleen Haucke

HASSO - PLATTNER - INSTITUTE
for IT Systems Engineering

Inhalt

1 Datenfusion - Inhalt

1.1 die Motivation

- Zusammenführung von Daten aus einer Vielzahl heterogener Datenquellen
- Wichtige Herausforderungen dabei
  - Skalierbarkeit
  - Dynamik
  - Inkonsistenzen
  - Automatisierung
- Drei Generationen von Datenintegrationstools (George Mason University, Fairfax, Virgina, USA)
  - Multiplex
  - Fusionplex
  - Autoplex

1.2 die Grundlagen

Virtuelle Datenbanksysteme

- Ermöglichen transparenten Zugriff auf eine Vielzahl heterogener Daten-quellen für Anwender und Applikationen durch
  - Einsatz eines globalen Integrationsschemas
  - Mapping zwischen diesem globalen Integrationsschema und dem Schema einer eingebundenen Datenquelle

Inkonsistenzen

- Treten auf, wenn sich die Inhalte der Datenquellen überlappen
- Schematische Inkonsistenzen
  - Verschieden strukturierte Tabellen
- Duplikate
  - Mehrfach vorhandene Datensätze
- Datenbezogene Inkonsistenzen
  - Auch als Datenkonflikte bezeichnet
  - Unterteilbar in Widersprüche und Unsicherheiten

2 Multiplex - Inhalt

2.1 die Idee

Fokus

- Schnelle Integration sehr großer, heterogener und dynamischer Quellen

Voraussetzungen

- Annahme über die Konsistenz der Schemata (gilt)
  - Schema einer Datenquelle entspricht einem Abbild der realen Welt
  - Fehler in den Modellen existieren nicht - nur unterschiedliche Schema-Modellierungen
- Annahme über die Konsistenz der Daten (gilt nicht)
  - Daten einer Datenquelle entsprechen den Objekten der realen Welt
  - Fehler in den Daten existieren nicht - nur unterschiedliche Daten-Repräsentationen

2.2 der Lösungsansatz

Unterstützung heterogener Datenquellen in einem dynamischen Umfeld

- Keine Beschränkung auf relationale Datenbanken
- Einbindung jeder Software, die tabellenförmige Antworten erstellt
- Zeitweilige Nicht-Erreichbarkeit von Datenquellen
- Inkonsistenzen zwischen Datenquellen

Annäherung an die optimale Antwort zu einer Anfrage

- Bei fehlenden Daten oder auftretenden Datenkonflikten
- Annäherung 1 „complete answer“: Antwort enthält potentiell „zu viele“ Datensätze - alle DS, für die alle Abfragekriterien gelten sind in der Antwortmenge enthalten
- Annäherung 2 „sound answer“: Antwort enthält potentiell „zu wenig“ DS - mindestens für die DS, die als Antwort ausgegeben werden, gelten alle Abfragekriterien

2.3 die Grenzen

Erkennung von Inkonsistenzen

- Nur auf Datensatzebene möglich
- Zwei DS, die sich nur marginal unterscheiden werden nicht als mögliche Duplikate erkannt

3 Fusionplex - Inhalt

3.1 die Idee

Fokus

- Einbindung wirkungsvoller Mechanismen zur Erkennung und automatischen Lösung von Datenkonflikten auf Attributebene
- Nutzung von Features (Metadaten) zur Bewertung der Datenquellen

Voraussetzungen (analog Multiplex)

- Annahme über die Konsistenz der Schemata (gilt)
- Annahme über die Konsistenz der Daten (gilt nicht)

3.2 die Prinzipien

- Daten sind nicht gleichwertig
  - Unterschiedliche Datenquellen liefern nicht die identische Datenqualität
  - Datenquellen haben jeweils ihre individuellen Vor- und Nachteile
- User hat bei der Lösung von Datenkonflikten ein Stimmrecht
  - Subjektive Bewertung der Wichtigkeit einzelner Features durch den User
  - Möglichkeit die vorhandenen Antworten danach zu bewerten und gegebenenfalls einige Datensätze auszusortieren

3.3 die Bewertungskriterien - einige Beispiele für Features

recentness (t)

- Aktualität einer Datenquelle

cost (c)

- Zugriffskosten (z.B. Zugriffszeit) für eine Datenquelle

priority (p)

- Priorität einer Datenquelle

accuracy (s)

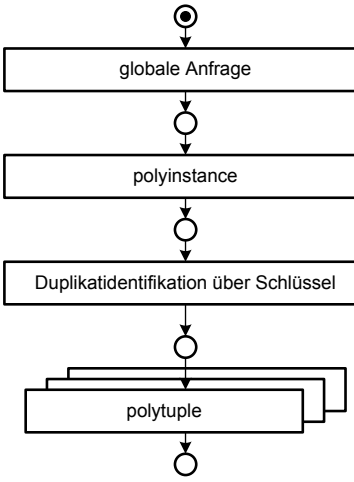
- Genauigkeit einer Datenquelle

availability (v)

- Wahrscheinlichkeit, dass auf eine Datenquelle zugegriffen werden kann

je Anfrage: Definition, welches Gewicht (w_i) die Features haben ⇨ utility

3.4 das Verfahren - mehrstufiger Prozess



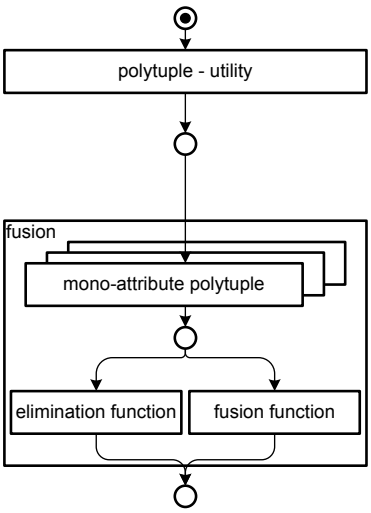
(1) Globale Anfrage ⇨ Antwortfragmente je integrierter Datenquelle

(2) Vereinigung aller Antwortfragmente zu einer polyinstance ⇨ Notwendigkeit diese polyinstance in eine konsistente Antwort zu überführen

(3) Identifikation derjenigen Tupel in der polyinstance, die ein und dasselbe Objekt der realen Welt repräsentieren (Annahme hier: Identifikation über gleiche Schlüsselattribute möglich)

(4) Zusammenfassen dieser identifizierten Tupel zu polytuples (visualisierbar in Tabellenform - Handout enthält Beispiel [1])

3.4 das Verfahren - Fortsetzung



(5) „Komprimierung“ der polytuples

- Utility - Berechnung für die einzelnen Tupel eines polytuple (Attributwerte ungleich *null*)

$u(x)= w_1*t(x)+ w_2*c(x)+ w_3*p(x)+ w_4*s(x)+ w_5*v(x)$

- Aussortieren und Ordnen

(6) Attributweise Bereinigung der verbleibenden Datenkonflikte innerhalb des polytuple ⇨ Fusion

- Elimination functions (feature- oder inhalt-basiert)
  ⇨ min, max, top_five_percent
- Fusion functions (immer inhalt-basiert)
  ⇨ avg, any, avg_without_extreme_values

3.4 das Verfahren - Fortsetzung

- Attributweise Aufsplittung der polytuples in mono-attribute polytuples (visualisierbar in Tabellenform - Handout enthält Beispiel [2])
- Anwendung der globalen Lösungsstrategie auf die mono-attribute polytuples
  - Festgelegte Sequenz aus ein oder mehreren elimination functions und einer fusion function

(7) Zusammenführung der fusionierten Attributwerte zu Tupeln und Berechnung der utility des fusionierten Ergebnisses (näheres dazu im paper)

Prototypische Realisierung mittels Erweiterung der SQL- Syntax

- Durch Festlegung der Feature-Gewichte durch den Anwender wird nicht immer die beste Fusion erreicht
- Möglichkeit die Effektivität der Fusion mittels Standardverfahren zu optimieren ⇨ Berechnung der best fusion

3.5 ein Beispiel

existing data sources

SSN	Name	Age	Salary	timestamp	cost	availability
326435218	Smithson	38	75000	0,8	0,5	1
777535289	Miller	58	85000	0,8	0,5	1
777535289	Millar	58	*85000*	0,8	0,5	1
311116679	Stevens	20	*null*	0,8	0,5	1

SSN	Name	Age	Salary	timestamp	cost	availability
326435218	Smith	35	*null*	0,7	0,2	0,8
777535289	Miller	55	77500	0,7	0,2	0,8

SSN	Name	Age	Salary	timestamp	cost	availability
326435218	Schmidt	35	77000	0,7	0,8	1
777535289	Miller	56	*null*	0,7	0,8	1
311116679	Stevenson	25	*null*	0,7	0,8	1

SSN	Name	Age	Sex	timestamp	cost	availability
326435218	Schmidt	30	male	0,2	0,8	0,2
326435218	Smithson	31	male	0,2	0,8	0,2

SSN	Name	Age	Salary	timestamp	cost
326435218	Smithson	36	75000	0,7	0,8

Stufe (1)

3.5 ein Beispiel - Fortsetzung

resulting polyinstance

SSN	Name	Age	Salary	timestamp	cost	availability
326435218	Smithson	38	75000	0,8	0,5	1
777535289	Miller	58	85000	0,8	0,5	1
777535289	Millar	58	*85000*	0,8	0,5	1
311116679	Stevens	20	*null*	0,8	0,5	1
326435218	Smith	35	*null*	0,7	0,2	0,8
777535289	Miller	55	77500	0,7	0,2	0,8
326435218	Schmidt	35	77000	0,7	0,8	1
777535289	Miller	56	*null*	0,7	0,8	1
311116679	Stevenson	25	*null*	0,7	0,8	1
326435218	Schmidt	30	*null*	0,2	0,8	0,2
326435218	Smithson	31	*null*	0,2	0,8	0,2
326435218	Smithson	36	75000	0,7	0,8	*null*

Stufe (2) bis (5)

3.5 ein Beispiel - Fortsetzung

ordered polyinstance

SSN	Name	Age	Salary	timestamp	cost	availability
326435218	Smithson	38	75000	0,8	0,5	1
326435218	Smith	35	*null*	0,7	0,2	0,8
326435218	Schmidt	35	77000	0,7	0,8	1
326435218	Schmidt	30	*null*	0,2	0,8	0,2
326435218	Smithson	31	*null*	0,2	0,8	0,2
326435218	Smithson	36	75000	0,7	0,8	*null*
777535289	Miller	58	85000	0,8	0,5	1
777535289	Millar	58	*85000*	0,8	0,5	1
777535289	Miller	55	77500	0,7	0,2	0,8
777535289	Miller	56	*null*	0,7	0,8	1
311116679	Stevens	20	*null*	0,8	0,5	1
311116679	Stevenson	25	*null*	0,7	0,8	1

example of a polytuple

SSN	Name	Age	Salary	timestamp (w=2)	cost (w=1)	availability (w=3)	utility
326435218	Smithson	38	75000	0,8	0,5	1	**5,1**
326435218	Smith	35	*null*	0,7	0,2	0,8	**4**
326435218	Schmidt	35	77000	0,7	0,8	1	**5,2**
326435218	Schmidt	30	*null*	0,2	0,8	0,2	**1,8**
326435218	Smithson	31	*null*	0,4	0,7	0,2	**2,1**
326435218	Smithson	36	75000	0,7	0,8	0,4	**3,4**

Stufe (2) bis (5)

3.5 eine Möglichkeit zur Bereinigung von mono-attribute polytuples [3]

SSN	Name	Age	Salary	timestamp	cost	availability
326435218	Smithson	38	75000	0,8	0,5	1
326435218	Smith	35	*null*	0,7	0,2	0,8
326435218	Schmidt	35	77000	0,7	0,8	1

for Name **fuse** any
for Age **fuse** avg
for Salary **keep** min() **fuse** any

SSN	Name	timestamp	cost	availability
326435218	Smithson	0,8	0,5	1
326435218	Smith	0,7	0,2	0,8
326435218	Schmidt	0,7	0,8	1

SSN	Age	timestamp	cost	availability
326435218	38	0,8	0,5	1
326435218	~~35~~	0,7	0,2	0,8
326435218	35	0,7	0,8	1

SSN	Salary	timestamp	cost	availability
326435218	75000	0,8	0,5	1
326435218	*null*	0,7	0,2	0,8
326435218	77000	0,7	0,8	1

SSN	Name	Age	Salary	timestamp	cost	availability
326435218	Smith	36	75000	0,7	0,2	0,8

Verwendung des Minimums

Stufe (6) und (7)

4 Autoplex - Inhalt

4.1 die Idee

Fokus

- Automatisierung der Integration neuer Datenquellen in ein bestehendes virtuelles Datenbanksystem

Mehrwert

- Zeitersparnis, da die Integration neuer Datenquellen automatisch abläuft
- Für das automatische Mapping notwendigen Informationen über die Bedeutung der Daten werden aus den Inhalten der Datenquellen gewonnen

5 Konkurrenzansatz - Inhalt

5.1 die Idee

Fokus

- Umgang mit Datenkonflikten (speziell mit Unsicherheiten) und nur teilweise bekannten Attributwerten bei der Integration verschiedener Datenquellen in ein virtuelles Datenbanksystem
  - Datenquellen können *null* - Werte enthalten
  - Verschiedene Datenquellen können Attribute enthalten, die sich nicht 1-zu-1 aufeinander abbilden lassen (Beispiel: Staat und Region - Deutschland und Bayern)

Lösungsansatz

- Bewertung der vorhandenen Attributwerte ⇨ Wahrscheinlichkeiten
- Qualifizierung der erzeugten Ergebnisse über diese Wahrscheinlichkeiten

5.2 ein kurzer Einblick

Datenquelle 1

Name	Region	Fachgebiet	Alter	Abschluss
Andy	Deutschland	AI	*null*	MS
Frank	Deutschland	DB	26	PhD
Jesse	Deutschland	SE	30	MS
Mike	Frankreich	DB	32	PhD
John	Schweiz	SE	28	PhD

Datenquelle 2

Name	Stadt	Fachgebiet	Alter	Organisation
Andy	Berlin	CS	25	NTU
Frank	Hamburg	CS	28	NCTU
Annie	Karlsruhe	CS	27	NCKU
Paul	Berlin	EE	30	NTU
Lisa	Hamburg	Math	26	NTHU

Suche nach computer sience (CS) - Spezialisten in Deutschland

Name	Region	Fachgebiet	Alter	Abschluss
Andy	Deutschland	AI	*null*	MS
Frank	Deutschland	DB	26	PhD
Jesse	Deutschland	SE	30	MS

Name	Stadt	Fachgebiet	Alter	Organisation
Andy	Berlin	CS	25	NTU
Frank	Hamburg	CS	28	NCTU
Annie	Karlsruhe	CS	27	NCKU

Integration der spezialisierten Informationen über Region und Fachgebiet

Name	Region	Fachgebiet	Alter	Abschluss
Andy	[Berlin-Hamburg-Karlsruhe]	AI	*null*	MS
Frank	[Berlin-Hamburg-Karlsruhe]	DB	26	PhD
Jesse	[Berlin-Hamburg-Karlsruhe]	SE	30	MS

Name	Stadt	Fachgebiet	Alter	Organisation
Andy	Berlin	[AI-DB-SE]	25	NTU
Frank	Hamburg	[AI-DB-SE]	28	NCTU
Annie	Karlsruhe	[AI-DB-SE]	27	NCKU

Spezialisierung der Suche: Suche nach Datenbank (DB) - Spezialisten in Hamburg, die älter als 27 sind

Name	Region	Fachgebiet	Alter	Abschluss	Organisation	Status
Andy	[Berlin-Hamburg-Karlsruhe]	[AI-DB-SE]	*null*	MS	NTU	möglich
Frank	[Berlin-Hamburg-Karlsruhe]	[AI-DB-SE]	[26,28]	PhD	NCTU	möglich

5.2 ein kurzer Einblick - Fortsetzung

Bewertung der spezialisierten Informationen über Region und Fachgebiet
(inklusive Bewertung der Wahrscheinlichkeit)

Name	Region	Fachgebiet	Alter	Abschluss
Andy	[Berlin$^{1/3}$-Hamburg$^{1/3}$-Karlsruhe$^{1/3}$]	AI	[·^{1}]	MS
Frank	[Berlin$^{1/3}$-Hamburg$^{1/3}$-Karlsruhe$^{1/3}$]	DB	26	PhD
Jesse	[Berlin$^{1/3}$-Hamburg$^{1/3}$-Karlsruhe$^{1/3}$]	SE	30	MS

Name	Stadt	Fachgebiet	Alter	Organisation
Andy	Berlin	[A$^{1/3}$-DB$^{1/3}$-SE$^{1/3}$]	25	NTU
Frank	Hamburg	[A$^{1/3}$-DB$^{1/3}$-SE$^{1/3}$]	28	NCTU
Annie	Karlsruhe	[A$^{1/3}$-DB$^{1/3}$-SE$^{1/3}$]	27	NCKU

Kombination der Datenquellen 1 und 2 und ihrer individuellen Wahrscheinlichkeitswerte

Name	Region	Fachgebiet	Alter	Abschluss	Organisation
Andy	[Berlin$^{4/6}$-Hamburg$^{1/6}$-Karlsruhe$^{1/6}$]	[A$^{4/6}$-DB$^{1/6}$-SE$^{1/6}$]	[25$^{1/2}$,·$^{1/2}$]	MS	NTU
Frank	[Berlin$^{1/6}$-Hamburg$^{4/6}$-Karlsruhe$^{1/6}$]	[A$^{1/6}$-DB$^{4/6}$-SE$^{1/6}$]	[26$^{1/2}$,28$^{1/2}$]	PhD	NCTU
Jesse	[Berlin$^{1/3}$-Hamburg$^{1/3}$-Karlsruhe$^{1/3}$]	SE	30	MS	[*1]
Annie	Karlsruhe	[A$^{1/3}$-DB$^{1/3}$-SE$^{1/3}$]	27	[·^{1}]	NCKU

Spezialisierung der Suche: Suche nach Datenbank (DB) - Spezialisten in Hamburg, die älter als 27 sind
(inklusive Bewertung der Wahrscheinlichkeit)

Name	Region	Fachgebiet	Alter	Abschluss	Organisation	Wahrscheinlichkeit
Andy	[Berlin$^{4/6}$-Hamburg$^{1/6}$-Karlsruhe$^{1/6}$]	[A$^{4/6}$-DB$^{1/6}$-SE$^{1/6}$]	[25$^{1/2}$,·$^{1/2}$]	MS	NTU	0 ≤ p ≤ 1/6 × 1/6 × 1/2
Frank	[Berlin$^{1/6}$-Hamburg$^{4/6}$-Karlsruhe$^{1/6}$]	[A$^{1/6}$-DB$^{4/6}$-SE$^{1/6}$]	[26$^{1/2}$,28$^{1/2}$]	PhD	NCTU	4/6 × 4/6 × 1/2 = 2/9

Quellen

- **Motro, Berlin und Anokhin** (SIGMOD Record, Vol. 33, No. 4)
  Multiplex, Fusionplex and Autoplex - Three Generations of Information Integration
- **Motro und Anokhin** (Information Fusion, Vol. 7 - 2006)
  Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources
- **Motro, Anokhin und Acar** (IQIS 2004)
  Utility-based Resolution of Data Inconsistencies
- **Tseng, Chen und Yang** (Distributed and Parallel Databases, Vol. 1 - 1993)
  Answering Heterogeneous Database Queries with Degrees of Uncertainty
- **Bleiholder und Naumann** (http://www2.informatik.hu-berlin.de/mac/publications/HUB-IB-197.pdf)
  Conflict Handling Strategies in an Integrated Information System

HASSO - PLATTNER - INSTITUTE
for IT Systems Engineering

Vielen Dank für die Aufmerksamkeit!

„When you use information from one source, it‘s plagiarism;
when you use information from many sources, it‘s information fusion.“
(Belur V. Dasarathy - seventh international conference on information fusion)