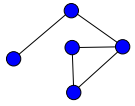
Gliederung

Peer-to-peer Systeme und Datenbanken(SS07)

- Kapitel 1: Einführung
- Kapitel 2: Beispiele
- Kapitel 3: Routing
- Kapitel 4: Schemabasierte p2p-Netzwerke
- Kapitel 5: Integrationsprobleme
  Teil 5-1: Einführung, Gleichheit
  Teil 5-2: Ähnlichkeit - 1
  Teil 5-3: Ähnlichkeit - 2
  Teil 5-4: Mappingbasierte Datenintegration
- Kapitel 6: Anonymität, Authentifikation
- Kapitel 7: Reputation

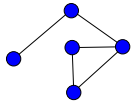
Version vom 4. Juni 2007



Kapitel 5

Integrationsprobleme

- Problembeschreibung
- Vergleiche auf Daten
- Ähnlichkeit auf Daten
  - Ähnlichkeit wegen formaler Merkmale
  - inhaltsbasierte Ähnlichkeit
- Mappingbasierte Datenintegration



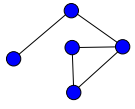
Allgemeine Probleme

vergleichbar mit föderierten DBVS:
Entstehung unabhängig voneinander - Heterogenität:

- Semantische H.: Synonyme - Homonyme in Daten , in Metadaten; Sprachl. H., het. Ontologien
- Strukturelle H.: formal: Datentypen, Zeichensätze
  inhaltlich: het. Ontologien (Struktur d. Graphen)
- Het. Inhalte
- Het. Datenqualität.

Folgt:

- Klass. DB-Prinzipien (z.B. Objekt-Identifikation mit OID oder PS) müssen versagen,
- Manuelle Ansätze aus föd. DB versagen wegen Umfang und Zeitrestriktion.



Matching-Problem

Bisherige P2P-Ansätze und Datenbankansätze weitgehend unterschiedlich.
Anforderungen für P2P-DB:

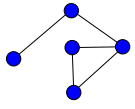
- P2P: Kein globales Wissen, Knotenautonomie, fluktuierende Teilnehmer und Datenbestände.
- DBS: Verarbeitung strukturierter Daten, Metainformationen verfügbar als wesentliche Voraussetzung

Metainformationen:

- syntaktische M.: Struktur der BD, Datentypen, Namen, Fremdschlüsselbeziehungen
- semantische M.: Bedeutung

Matching-Problem: Gewinnung semantischer Informationen aus syntaktischen oder aus Inhalten.

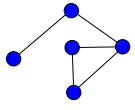
Speziell in P2P-Systemen: schnelle Algorithmen, Wiederverwendung fraglich.



Kapitel 5

Integrationsprobleme

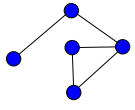
- Problembeschreibung
- Vergleiche auf Daten
- Ähnlichkeit auf Daten
  - Ähnlichkeit wegen formaler Merkmale
  - inhaltsbasierte Ähnlichkeit
- Mappingbasierte Datenintegration



Gleichheit

- Math. Hintergrund: Gleichheitsbegriff -
  reflexiv, symmetrisch, transitiv (rst)
- Informatik:
  zweistufiger Modellierungsprozess: Welt $\Leftrightarrow$ Modell $\Leftrightarrow$ RiD
  Realisierung aller Operationen auf RiD, danach Interpretation im Modell notwendig.

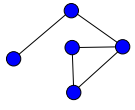
...
ganze Zahlen, Fließkommazahlen, Zeichenketten, logische Größen



Gleichheit Integergrößen

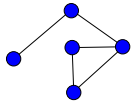
- Abbildung Modell $\Leftrightarrow$ RiD ist eineindeutig,
  aus Gleichheit in RiD kann Gleichheit im Modell (math. Zahl)
  erschlossen werden.
- Probleme: unterschiedliche RID mit unterschiedlichem Wertevorrat,
  d.h. Vergleich kann nicht auf Bit-Niveau geführt werden, Datentypen
  beachten.
- aus Daten und Metainformationen (Wertebereiche, ...) kann mit
  Restrisiko auf Datentyp geschlossen werden.

analog: logische Größen.



Gleichheit Fließkommazahlen

- IEEE 754: single precission (32), double (64), double extendet (80) - round to next, Unendlich, NaN.
- Jede Maschinenzahl repräsentiert ein Intervall; Rechnen: $\rightarrow$ Fehlerfortpflanzung.
  Intervallarithmetik rechnet mit Intervallen $\rightarrow$ Qualität des Ergebnisses sichtbar.
- Vergleiche, bei denen die Exaktheit der F. unterstellt wird bzw. nur der Wert verglichen wird, meist i.O. (wegen der eindeutigen Rundungsregel)
  Vergleiche, in denen gerechnet wird oder mit berechneten Größen gearbeitet wird, erfordern erweiterten Gleichheitsbegriff ($\rightarrow$ Ähnlichkeit).

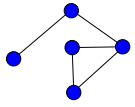


Gleichheit - Zeichenketten

- Vergleiche beruhen (meist) auf bin. Codierung des Textes im gewählten Zeichensatz, ASCII kleinster gemeinsamer Inhalt in vielen europ. Sprachen (trad. Codierung). Problem: nationale Sonderzeichen unterschiedl. codiert.
- Spracherkennung (charakt. Häufigkeiten von Zeichen, Bi- und Trigrammen.)
  $\longrightarrow$ Codierungserkennung (Sonderzeichen der Sprache in vermuteter Codierung darstellen)
  $\longrightarrow$ Überführung in gemeinsame Codierung - UNICODE

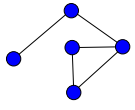
Sprachliche Besonderheiten - Umlaute, Betonung, Trema, ..., Ligaturen, ...

- Umlaute: Codierungsproblem, Sortierproblem (s.u.)
- Betonung: im UNICODE-Zeichensatz Unterscheidung: GR $iota$: ι, ί, ϊ, im Telefonbuch: Gleichbehandlung.
- Trema: auch im DE relevant - ($Haiti$ - $H\ddot{a}iti$, Asteroid, ...)
- Ligaturen (in Dt. ß $\leftarrow$ s+z), Hindi



Vergleiche - Sortierung

- Für Zahlen Vergleichsoperation math. definiert.
  Keine Probleme: Integerformate, unberechnete Fließkommaformate,
  Probleme möglich:
  Vergleich berechneter Fließkommawerte - Qualität der Daten
  (Vorverarbeitung) muss bekannt sein.
- Sortierung - setzt Ordnungsrelation voraus
  Keine Probleme Integerzahlen,
  Fließkommazahlen: i.A. in Ordnung ,
  Unendlich, NaN in vielen Sprachen nicht implementiert



Zeichenketten - lexikographische Sortierung

- Zeichenketten - lexikographischer Vergleich
  Algorithmus: *selbst nachtragen*
- Unterschiedliche Einsortierung der Umlaute:
  - ignorieren: ä wie a (DE: Lexika) - DIN 5007-1
  - in DE: ä wie ae, DIN 5007-2: *Kuciak - Kudies - Küchler* (Telefonbuch)
  - in OE: ä nach az (österr. Telefonbuch)
- Beispiel

DIN 5007-1 Lexika	DIN 5007-2 Telefonb.	Österr.Sort
Göbel	Göbel	Goethe
Goethe	Goethe	Goldmann
Goldmann	Götz	Göbel
Götz	Goldmann	Götz

Quelle: *Wikipedia.Stichwort: Alphabetische Sortierung*.