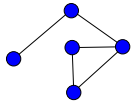
Gliederung

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

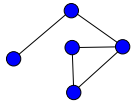
Version vom 22. Juni 2007



Kapitel 5

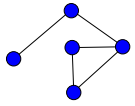
Integrationsprobleme

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



Inhaltsbasierte Ähnlichkeit

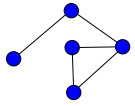
- Idee: Zwei (Objekt-) Klassen sind gleich g.d.h. sich die Instanzen der einen Klasse eineindeutig auf die Instanzen der anderen Klasse abbilden lassen.
- Theoretisch abzählbar viele Instanzen - praktisch nur endl. viele, d.h. keine absolute Sicherheit.
- Probleme (bei Automatisierung):
  Massendatenverarbeitung
  Fehlende Metainformationen (z.B. über Struktur, Semantik)
  Beispiel: $Literatur(\{Autor\}, Titel, Verlag)$
  $Literatur(Verlag, \{\{Autor\}, Titel\})$
  $Literatur(Autor, \{Titel, Verlag\})$
  Mehrere Varianten der interen Struktur von Autor und Verlag.
  Beispiel: $Autor$(<Text>)
  $Autor(Name$ (m. Zusätzen), $\{Vornamen\}$
  $Autor(Name, Zusatz, \{(Vorname|Initiale)\})$
  Zusätze: von, de, ...; aber franz.: $DeFries$.



Einfache Objekte

Definition: Komplexes Objekt g.d.w. in Konstruktionsvorschrift der Klasse wird eine der folgenden Aggregationen *List*, *Array*, *Set*, *Bag*, *Tupel* benutzt.

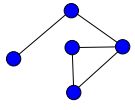
- Ähnlichkeit komplexer Objekte $\Rightarrow$ Schemaintegration.
  Zunächst Ähnlichkeit einfacher (nicht komplexer) Objekte.
- Grundannahme:
  Zu einer Klasse gibt es eine Menge von Instanzen.
  Zwei Möglichkeiten:
  Aus der Menge der Instanz einer Klasse werden Parameter berechnet und mit theoretischen Werten verglichen.
  Aus zwei Mengen von Instanzen werden Parameter berechnet und miteinander verglichen.
- Grundproblem: Der aktuelle ZUstand einer Instanzmenge ist i.a. nur ein möglicher Zustand. Deshalb haben alle Aussagen, die nicht auf allen möglichen Zuständen basieren, das nicht zu vernachlässigende Risiko, falsch zu sein.
  Reduzierung: verschiedene Überprüfungen.



Statistische Tests - Beispiele

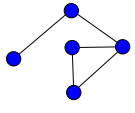
- Parametertests
- Parameterfreie oder nichtparametrische Tests: χ^2-Test:
  Test, ob Stichprobe einer (zuvor angenommenen) W-Verteilung F folgt.
  Kolmogorow-Smirnow-Test

  Test auf Übereinstimmung zweier Wahrscheinlichkeitsverteilungen oder
  Test, ob Stichprobe einer (zuvor angenommenen) W-Verteilung folgt.
  nichtparametrischer Test, sehr stabil, für stetige, diskrete, rangskalierte Merkmale
  sehr flexibel nutzbar $\Rightarrow$ evt. nicht sehr scharf.



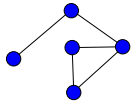
Regressionsanalyse

- Zwischen Attributen einer Instanz bestehe funktionaler Zusammenhang. Für die Daten jeder Menge von Instanzen werden Parameter bestimmt. Stimmen die Parameter für verschiedene Datenmengen überein, ist das ein Hinweis auf gleiche bzw. gleichartige Objekte.
- Aus endlich vielen Werten kann die Funktion nicht bestimmt werden.
  $\Rightarrow$ bei ungünstiger Datenlage falsch positive Ergebnisse.



Clusteranalyse

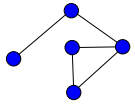
- Datenmenge wird Clusteranalyse unterworfen. Für jedes Cluster werden charakteristische Werte (z.B. Clusterzentren) bestimmt. Zwei Datenmengen sind ähnlich, wenn sie die gleichen (ähnliche) Cluster bilden, d.h. wenn sie ähnliche charakteristische Werte - z.B. die Vektoren der Clusterzentren- der haben.
- weitere Abstandsmaße: s.u.



Abstandsmaße für Mengen (in metr. Räumen)

Gegeben zwei Mengen $\mathcal{A} = \{a\}, \mathcal{B} = \{b\}$.
Abstand der Mengen $d_M(\mathcal{A}, \mathcal{B}) =$

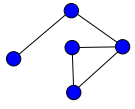
- $\min_{a \in \mathcal{A}, b \in \mathcal{B}}(a, b)$ — Minimaler Abstand zweier Elemente.
- $\max_{a \in \mathcal{A}, b \in \mathcal{B}}(a, b)$ — Maximaler Abstand zweier Elemente.
- $\frac{\sum_{a \in \mathcal{A}, b \in \mathcal{B}} d(a,b)}{\text{card}(\mathcal{A})\text{card}(\mathcal{B})}$ — Durchschnittlicher Abst. aller Elementpaare aus...
- $\frac{\sum_{a,b \in \mathcal{C}, \mathcal{C} = \mathcal{A} \cap \mathcal{B}} d(a,b)}{\text{card}(\mathcal{C})}$ — Durchschn. Abst. aller Paare aus Vereinigungsm.
- $d(\bar{a}, \bar{b})$ — Abst. der Mittelwerte d. Cluster (Centroid-Abst.)
- $\frac{d(\bar{a},\bar{b})}{1/\text{card}(\mathcal{A}) + 1/\text{card}(\mathcal{B})}$ — Zunahme der Varianz beim Vereinigen von $\mathcal{A}$ und $\mathcal{B}$ (Ward'sche Methode).



Clusterverfahren

Gruppierung von Objekten
Abbruchkriterien: z.B. Zahl der Cluster, Mindestabst. der Cluster, ...

- anhäufend:
  1. Anfangs jedes Ojekt eineigenes Cluster.
  2. Schrittweise Zusammenfassung ähnlicher Objekte bzw. Cluster zu einem neuen Cluster.
  3. Abbruchkriterum erfüllt $\rightarrow$ fertig, sonst weiter bei (2).
- teilend:
  1. Anfangs alle Ojekte in einem Cluster.
  2. Teilung der Cluster /eines Cl., so dass Abstand der Teile möglichst groß.
  3. Abbruchkriterum erfüllt $\rightarrow$ fertig, sonst weiter bei (2).
- Wahl des Abbruchkriteriums und des Schrittes (2)$\rightarrow$ verschiedene Verfahren.

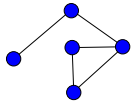


Clusterverfahren (Auswahl)

Partitionierend:

- k-means-Algorithmus (theoret. Schwächen, billig und gut)
- EM-Algorithmus
- Spektral Clusterung (Bildverarbeitung, WEB-Suche)
- Parallele Mehrfachclusterung
- ...

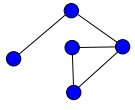
Graphentheoret. Methoden: ... Fuzzy-Clusterung



Clusterverfahren (k-mean)

k Zahl der Cluster - vorgegeben.

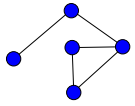
1. (Initialisierung) Auswahl von k initialen Clusterzentren
2. Jedes Objekt wird dem ihm nächsten Zentrum zugeordnet. Neuberechnung der Clusterzentren.
3. Ist jetzt ein Objekt falsch eingeordnet $\rightarrow$ (2).



Spektral Clusterung - Skizze

Literatur:Tutorial given at ICML 2004: Spectral Clustering.
URL: http://crd.lbl.gov/ cding/Spectral/

- Initialzustand: Alle Objekte in einem Cluster.
  Ähnlichkeit der Objekte i, j: $\{w_{i,j}\}$ Adjazenzmatrix der Ä.
  Schnitt S teilt Objekte in zwei Cluster A, B.
  Schnittgewicht G_S = gewichtete Kantensumme d. durchtrennten Kanten.
- Gesucht: Schnitt mit minimalem Gewicht: führt auf Eigenwertproblem für pos. semidef. Operator. Zweitkleinster Eigenwert ist die Lösung unserer Aufgabe, dazu Eigenvektor q_2.
  Mengentrennung: $A = \{i : q_2(i) < 0\}, B = \{i : q_2(i) > 0\}$,
- Da Lösung unabhängig von add. Konstante im Gewicht: Sortiere Objekte nach $q_2(i)$ und trenne in der Mitte.
- Wiederhole mit A bzw. B, wenn Abbruchkriterium nicht erfüllt oder teile weiter mit höheren EW.



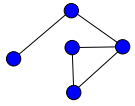
Ähnlichkeit von Mengen

Gegeben zwei Mengen $\mathcal{A} = \{a_i\}, \mathcal{B} = \{b_j\}$.
Ähnlichkeitsmaße:

- Base: $s_{Base}(\mathcal{A}, \mathcal{B}) = \begin{cases} 1 & : \quad \mathcal{A} \cap \mathcal{B} \neq \emptyset \\ 0 & : \quad sonst \end{cases}$
- Dice: $s_{Dice}(\mathcal{A}, \mathcal{B}) = \frac{2 \times \text{card}(\mathcal{A} \cap \mathcal{B})}{\text{card}(\mathcal{A}) + \text{card}(\mathcal{B})}$.
- Min : $s_{Min}(\mathcal{A}, \mathcal{B}) = \frac{\text{card}(\mathcal{A} \cap \mathcal{B})}{\min(\text{card}(\mathcal{A}) + \text{card}(\mathcal{B}))}$. Entsprechend: $s_{Max}(\mathcal{A}, \mathcal{B})$.

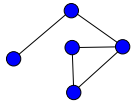
Es gilt: $s_{Max}(\mathcal{A}, \mathcal{B}) \leq s_{Dice}(\mathcal{A}, \mathcal{B}) \leq s_{Min}(\mathcal{A}, \mathcal{B}) \leq s_{Base}(\mathcal{A}, \mathcal{B})$
Definition: Zwei Mengen heißen ähnlich nach dem Maß μ

mit dem Schwellwert s_0 g.d.w $s_\mu > s_0$.



Kommentare

- *Dice* vernachlässigt gemeinsames Nichtenthaltensein $\rightarrow$ Faktor 2.
  Auch andere Ähnlichkeitsmaße für Bitvektoren übertragbar.
- In ähnlicher Weise auch Teilmengenbeziehungen über den Inhalt definierbar:
  $s(\mathcal{A} \subseteq \mathcal{B}) = \frac{\text{card}(\mathcal{A} \cap \mathcal{B}) - \text{card}(\mathcal{A} \setminus \mathcal{B})}{\text{card}(\mathcal{A})}$.



Ähnlichkeit von Bildern

Charles E. Jacobs: Fast Multiresolution Image Querying. Proc. SIGGRAPH 1995.

- aus Inhalt charakteristische Daten errechnet:
  Farbmodell YIQ, Wavelettransformation (Haar Wavelets)
- Idee d.Metrik: gewichtete L_1-Norm von bearbeiteten WL-Koeffizienten der Bilder Q, T für jeden Kanal des Farbmodells:
  $\|Q,T\| = w_{0,0}|Q(0,0) - T(0,0)| + \sum_{i,j} w_{i,j}|\tilde{Q}(i,j) - \tilde{T}(i,j)|$
- Praktische Metrik noch vereinfacht (Symmetrieverlust)- ist dann im math Sinn keine Metrik.
- u.a. Vergleiche zwischen Kinderzeichnungen und Photographien möglich.