

8. Besondere Datentypen und Anwendungen

Inhalt dieses Kapitels

8.1 Temporal Data Mining

Problemstellung, Sequential Patterns, Modifikation des Apriori-Algorithmus

8.2 Spatial Data Mining

Aufgaben und Probleme, typische Methoden, räumliche Charakterisierung und Trenderkennung

8.3 Text- und Web-Mining

Aufgaben und Probleme, Clustering von Web/Text-Dokumenten, Suchmaschine mit Berücksichtigung der Linkstruktur

8.1 Temporal Data Mining

Problemstellung

- zwei Arten von Methoden

Analyse zeitlicher Zusammenhänge *innerhalb* einzelner Abläufe

Analyse zeitlicher Zusammenhänge *zwischen* verschiedenen Abläufen

- Besonderheit des Temporal Data Mining

komplexe zeitliche Relationen zwischen Zeitpunkten und Zeitintervallen:
„während“, „überschneidend“, „direkt aufeinanderfolgend“ . . .



neue Typen interessanter Regeln

zusätzliche Komplexität der Algorithmen

8.1 Temporal Data Mining

Problemstellung

- Analyse von zeitbezogenen Daten

- Anwendungen

Finanzen: Aktienkurse, Inflationsraten, . . .

Medizin: Blutdruck, . . .

Meteorologie: Niederschläge, Temperaturen, . . .

- ausgezeichnetes Attribut:

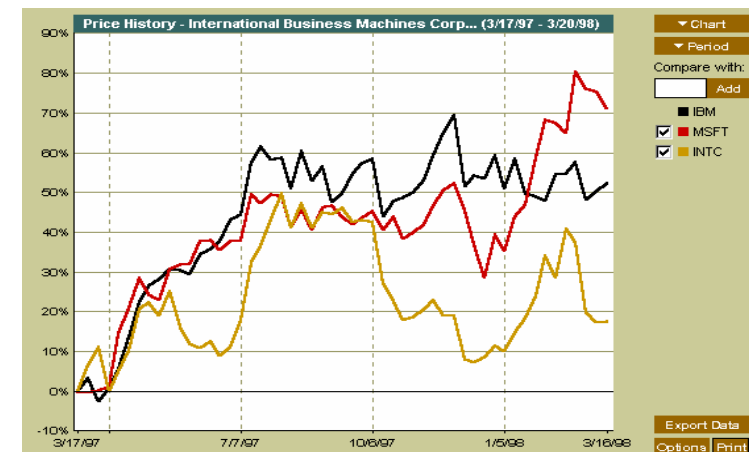
Punkte oder Abschnitte in einem zeitlichen Bezugssystem



impliziert zeitliche Ordnung der Datensätze

8.1 Zeitreihen -Analyse

Beispiel



8.1 Zeitreihen-Analyse

Komponenten von Zeitreihen [Fahrmeier et al.1999]

Trendkomponente

langfristige systematische Veränderung

Konjunkturkomponente

Verlauf von Konjunkturzyklen

Saisonalkomponente

jahreszeitlich bedingte Schwankungen

Restkomponente

Irreguläre Veränderungen, zufällig, relativ gering

8.1 Sequential Patterns

Idee

- nicht einzelne Transaktionen, sondern Mengen von zusammengehörigen und zeitlich geordneten Sequenzen von Transaktionen
- häufige Sequenz:
viele Kunden, die zu einem Zeitpunkt Produkte *A*, *B*, *C* eingekauft haben, haben zu einem späteren Zeitpunkt auch die Produkte *D*, *E* und *F* gekauft
„5% aller Kunden haben zuerst das Buch *Solaris*, danach das Buch *Transfer* und dann *Der Futurologische Kongreß* gekauft.“
- Anwendung
Kunde hat schon *Solaris* gekauft, bestellt jetzt *Transfer*:
➡ empfehle *Der Futurologische Kongreß*

8.1 Zeitreihen-Analyse

Methoden [Fahrmeier et al.1999]

Globale Regression

- Auswahl eines Funktionstyps
- Schätzung der unbekannt Parameter mit Hilfe der Methode der kleinsten Fehlerquadrate

🔍 globaler Trend häufig zu grob

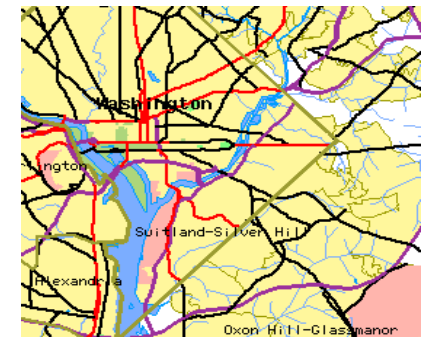
Lokale Methoden

- gleitender Durchschnitt (Moving Window)
Glättung
- lokale Regression
Regressionsfunktion für Umgebung des jeweiligen Punkts

8.2 Spatial Data Mining

Problemstellung

- Analyse von raumbezogenen Daten
- ausgezeichnetes Attribut:
Lage und Ausdehnung in einem 2- oder 3-dimensionalen Raum
➡ Punkte, Linien, Polygone, Polyeder
- Anwendungen
Geographie: Topologische Karten, Thematische Karten, . . .
Biologie: Proteine, . . .



8.2 Spatial Data Mining

Problemstellung

- Aufgaben
 - Analyse von *einzelnen* räumlichen Verteilungen bestimmter Attribute
 - Analyse von Abhängigkeiten *zwischen* räumlichen Verteilungen von Attributen
- Anwendungen
 - Geo-Marketing
 - Verkehrssteuerung
 - Umweltschutz . . .
- Besonderheit des Spatial Data Mining
 - Attribute von Nachbarn beeinflussen ein gegebenes Objekt
 - Einfluß hängt ab von räumlichen Nachbarschaftsbeziehungen

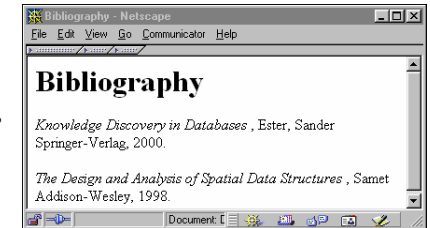
8.3 Text- und Web-Mining

Problemstellung

- Text
 - Transformation eines Dokuments D in Vektor $r(D) = (h_1, \dots, h_d)$
 - $h_i \geq 0$: die Häufigkeit des Terms t_i in D
 - Reduktion der Anzahl der Terme
 - Stop-Listen, Stemming, Entfernen besonders häufiger bzw. seltener Terme

- Präsentation (HTML)

```
<h1> Bibliography </h1>
<p> <i>Knowledge Discovery in Databases</i>,
Ester, Sander <br>
Springer-Verlag, 2000. </p>
. . .
```



8.3 Text- und Web-Mining

Problemstellung

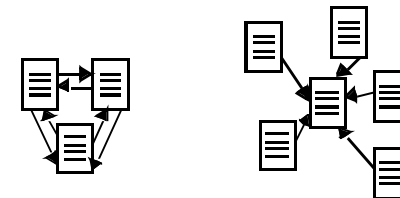
- Analyse von Text- und Hypertext-Daten sowie ihrer Benutzung
- Anwendungen
 - elektronische Mails einer Firma
 - Newsgroup-Artikel
 - Webseiten aus dem Internet oder dem Intranet einer Firma
- Text- und Hypertext-Daten
 - Text
 - Präsentation
 - Inhalt
 - Hyper-Links

8.3 Text- und Web-Mining

Problemstellung

- Inhalt (XML)

```
<bibliography> <book> <title> Knowledge Discovery in Databases </title>
<author> Ester </author> <author> Sander </author>
<publisher> Springer-Verlag </publisher>
<year> 2000 </year>
</book>
. . .
</bibliography>
```
- Hyper-Links




8.3 Text- und Web-Mining

Problemstellung

- Aufgaben
 - Analyse von *Inhalt* und *Struktur* von Hypertext-Dokumenten
 - Analyse der *Link-Struktur* einer Menge von Hypertext-Dokumenten
 - Analyse der *Benutzung* einer Menge von Hypertext-Dokumenten
- Besonderheit des Text- und Web-Mining
 - ➔ Diversität des Vokabulars, z.B. verschiedene Sprachen
 - Vagheit der Texte
 - Unterschiedliche Qualität der Texte
 - ➔ Link-Struktur

8.3 Clustering der Antwortmengen von Suchmaschinen

Motivation

- Ergebnisse von Web-Suchmaschinen
 - im allgemeinen in Form einer Liste
- Probleme
 - Antwortlisten typischerweise sehr lang
 - viele Terme treten in ganz verschiedenen Kontexten auf
 - sehr unübersichtliche Darstellung
-  z.B. „Cluster“: Datenanalyse, Rechnernetze, Astronomie, . . .
- Ziel
 - Clustering der Antwortmengen nach Kontexten
 - Browsen der Clustering statt der Antwortliste

Siehe Vorlesung Internet-Suchmaschinen im Winter