

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

- 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 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 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 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

Beispiel: Vergleich von Social Bookmarking Systemen mit Suchmaschinen

Folksonomies

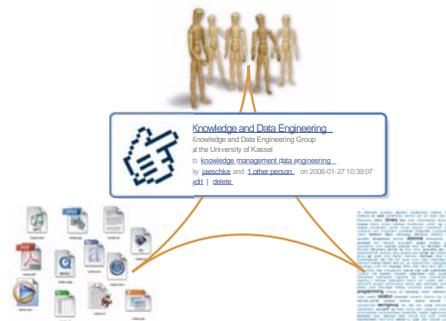


- Allow **users**
- to assign **tags**
- to **resources**
- In order to manage bookmarks, photos, ...
- Lightweight knowledge management
 - less overhead
 - integrated into daily routine

Logsonomies



- Allow **users**
- to query **terms**
- and click the **results**
- Filter information of the web. Simple and fast to use.



Datensatz

Social-Bookmarking-Daten

- del.icio.us crawl
- collection of posts: users, their tags, and URLs
- Zeitstempel zum Erzeugen von Schnappschüssen

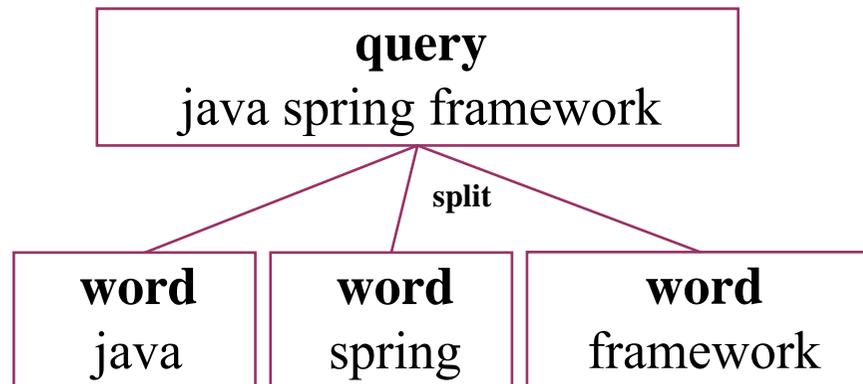
Suchmaschinen-Daten

Datasets	Queries (clickdata)	Rankings
MSN	✓	✓
AOL	✓	✓
Google		✓

Konstruktion des Query/Tag Datensatzes

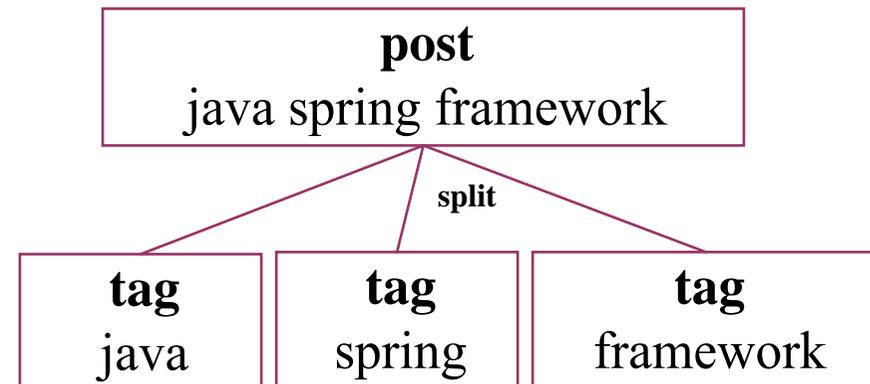
Search engine clickdata

java spring framework



Social bookmarking system

[Springframework.org](#) save this
to java spring framework j2ee programming ... saved by 2854 people



term

- Normalisierung der Terme
- Ergebnisliste hat die Form:
01/05/2006 java 189
01/05/2006 spring 23
02/05/2006 java 222
...

Beziehung zwischen Suchen und Taggen?

Wahl der Terme?

hippo

africa

nature

springs



Web

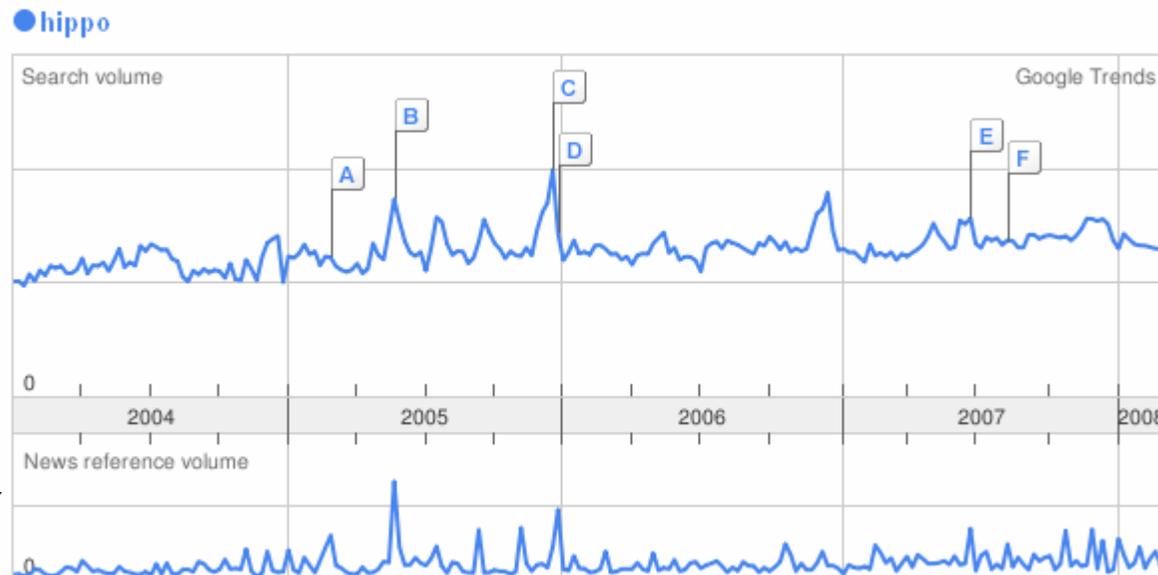
[Nature Conservancy Magazine: Spring 2008 - Flora & Fauna](#) - [[Diese Seite übersetzen](#)]

Nature Conservancy Magazine: **Spring 2008 Nature Conservancy Magazine** ... Once common in the Nile region and West **Africa**, **hippos** are now abundant only in ...
www.nature.org/magazine/spring2008/features/ - 18k - [Im Cache](#) - [Ähnliche Seiten](#)

[Mzima Springs Hippo on Flickr - Photo Sharing!](#) - [[Diese Seite übersetzen](#)]

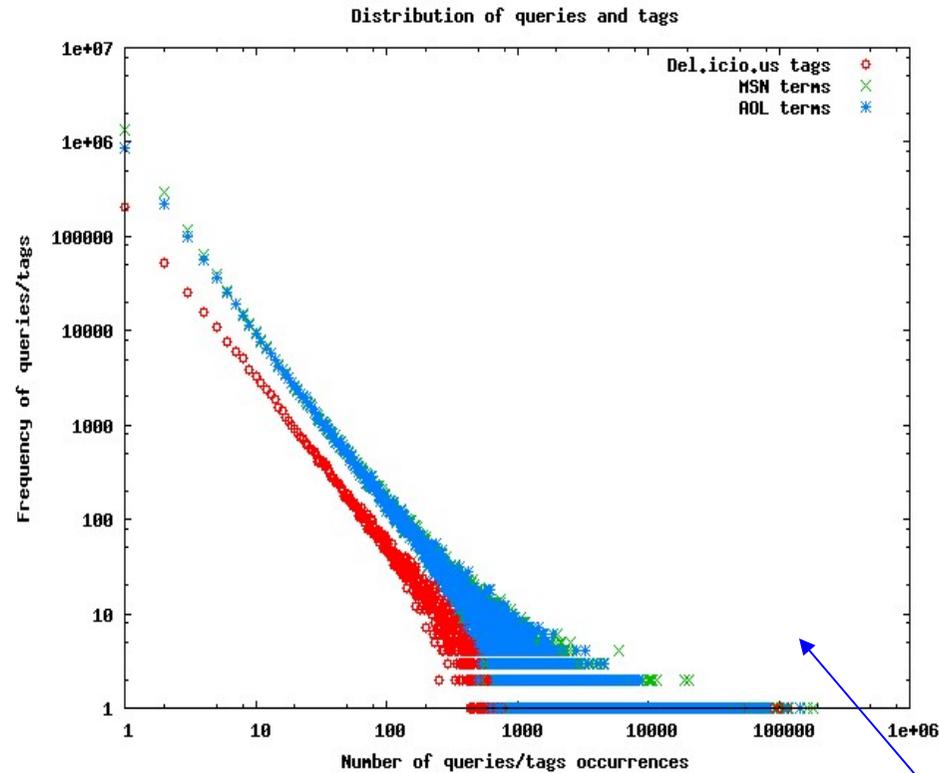
Mzima **Springs Hippo**. To take full advantage of Flickr, you should use a JavaScript-enabled photos. Part of: **Animals**, **Africa** and **Nature** & Wildlife ...
www.flickr.com/photos/54228211@N00/317720996/ - 188k - [Im Cache](#) - [Ähnliche Seiten](#)

Benutzung über die Zeit



Vorlesung Know

Einfache Statistische Analyse



Del.icio.us

MSN

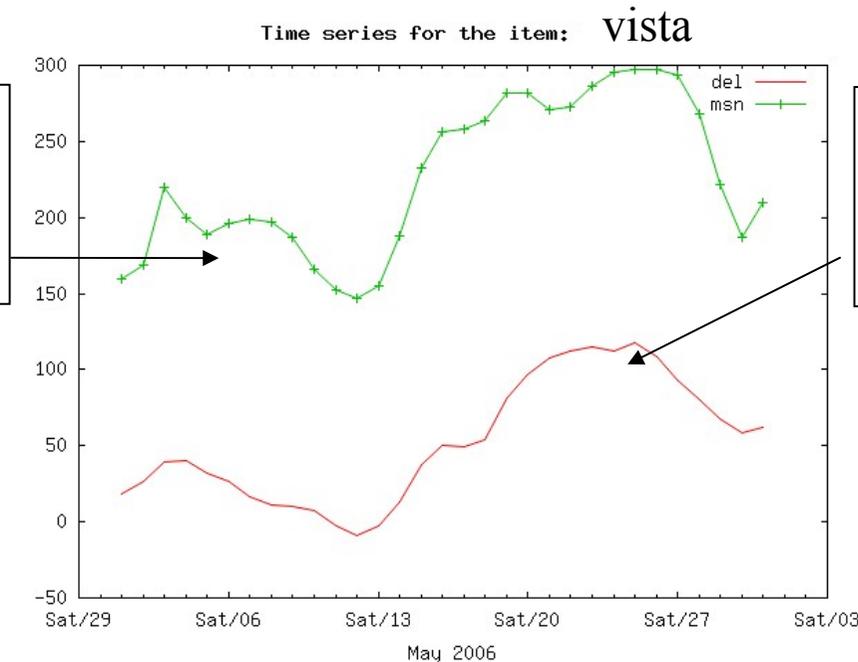
AOL

Top tags	Frequency	Top Terms	Frequency	Top Terms	Frequency
design	119,580	yahoo	181,137	free	145,585
blog	102,728	google	166,110	google	116,537
software	100,873	free	118,628	http	84,376
web	97,495	county	118,002	county	77,798
reference	92,078	myspace	107,316	pictures	75,977

Beispiel Zeitreihe

- Vergleich der MSN Worte mit den del.icio.us tags
- Normalisierung, Pearson Korrelationskoeffizient, T-Test
- 1003 Terme: für 307 von 1003 Termen wurde eine signifikante Korrelation (5 % level) beobachtet

Number of times the split word was submitted to a search engine on one day

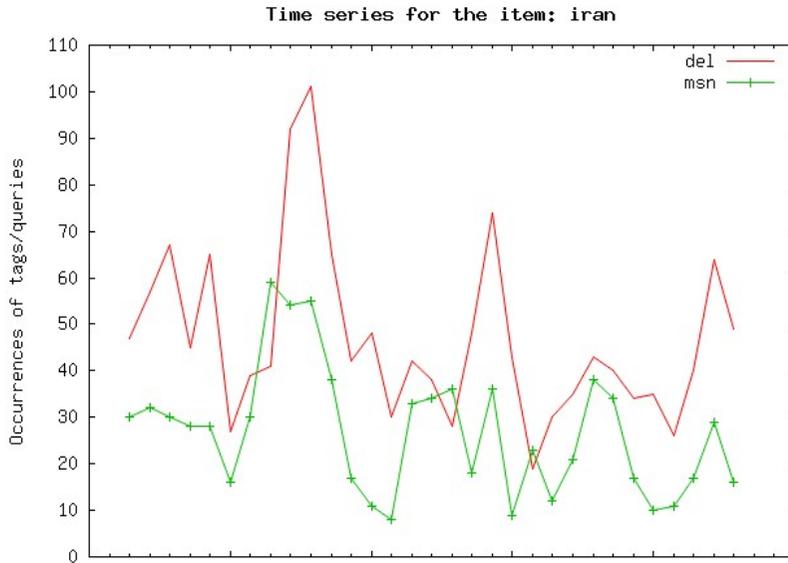


Number of times the tag was added to a resource on one day

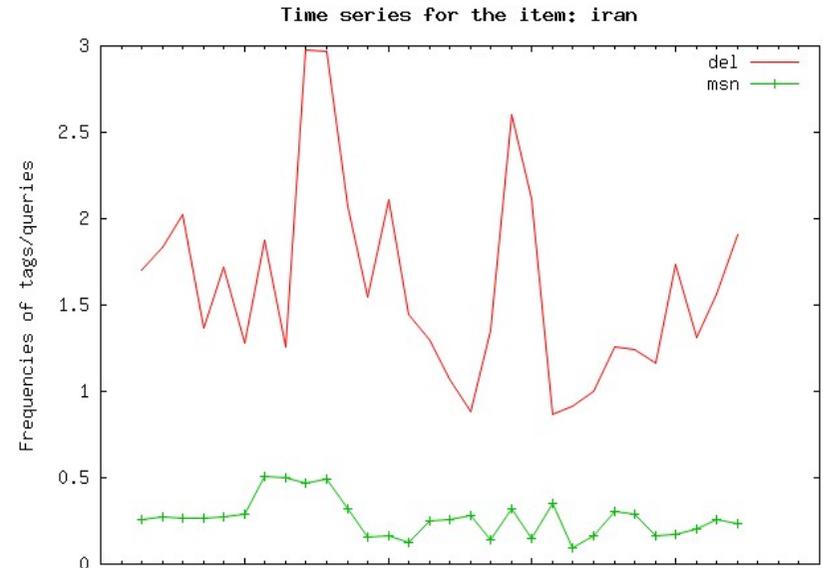
- Nutzer Suchen und Taggen das Gleiche zur gleichen Zeit
- Taggen und Suchen wird durch ähnliche Motivationen angetrieben

Vergleich von Zeitreihenmethoden (Term: iran)

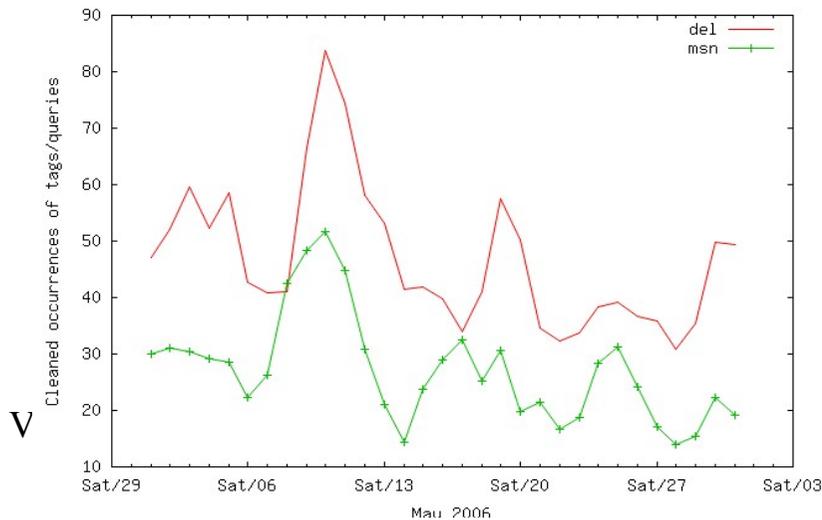
Original dataset



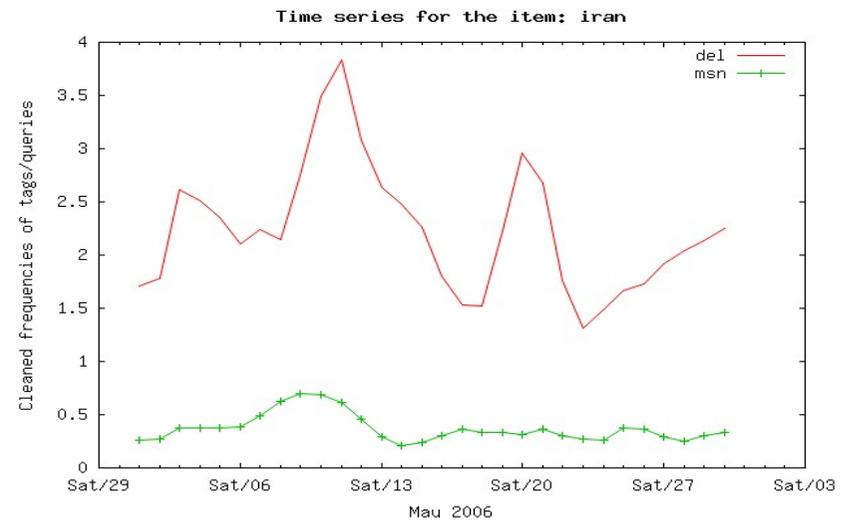
Frequency normalized



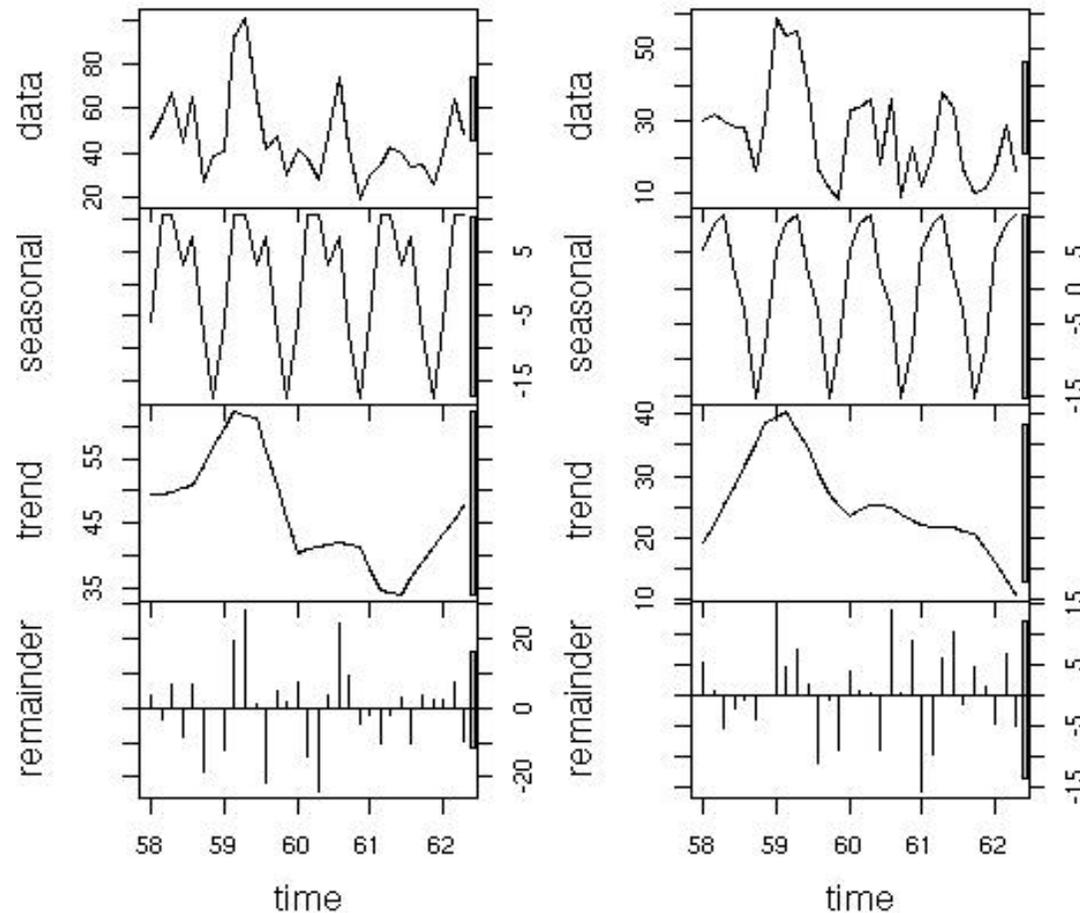
Exp. Smoothing



Frequency normalized + Sine Smoothing



Saisonale Zerlegung des Terms “iran” mit LOESS



LOESS ist eine Zerlegungsmethode in R basierend auf lokaler Regression.

8.1 Sequential Patterns

Idee

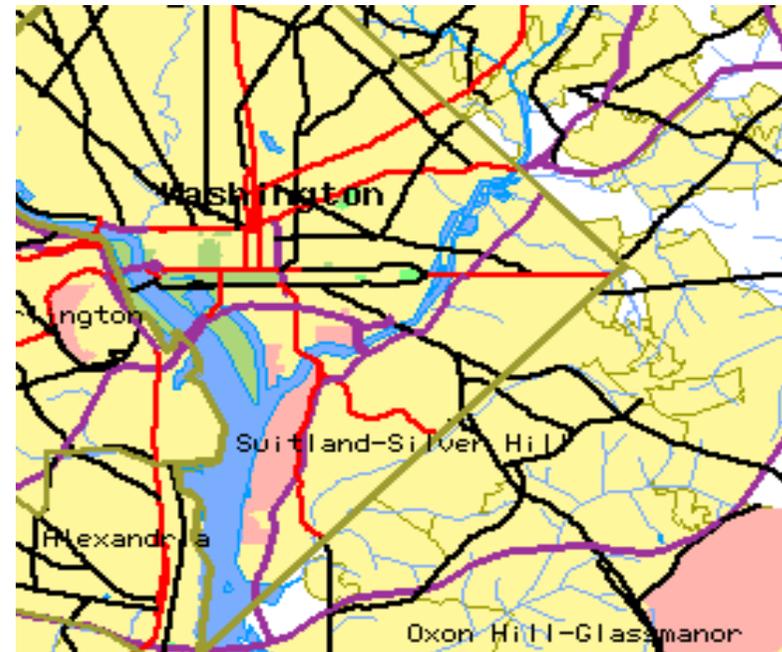
- nicht einzelne Transaktionen, sondern Mengen von zusammengehörigen und zeitlich geordneten Sequenzen von Transaktionen
- häufige Sequenz:
viele Kunden, die zu einem Zeitpunkt Produkte *A*, *B*, *C* eingekauft haben, haben zu einem späteren Zeitpunkt auch die Produkte *D*, *E* und *F* gekauft

„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*

8.2 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

- 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

- 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)

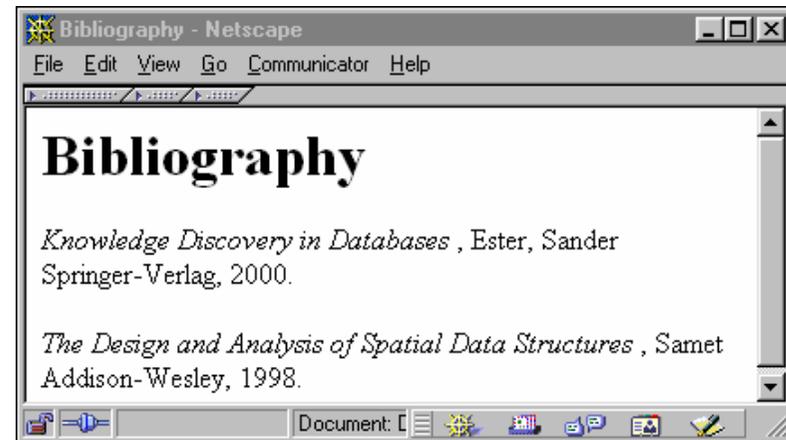
`<h1> Bibliography </h1>`

`<p> <i>Knowledge Discovery in Databases</i>,</p>`

`Ester, Sander
`

`Springer-Verlag, 2000. </p>`

`...`



8.3 Text- und Web-Mining

Problemstellung

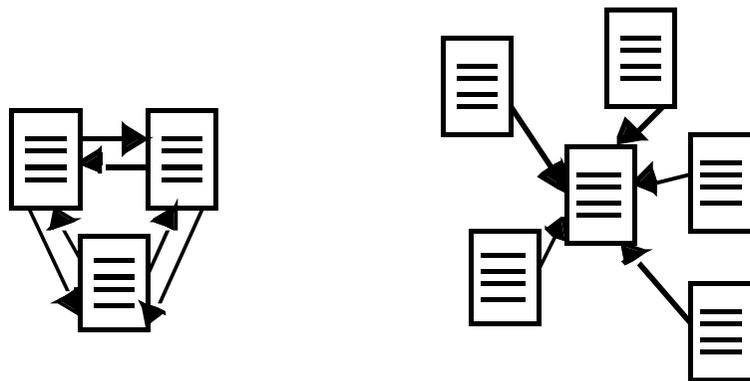
- Inhalt (XML)

```
<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

Siehe Vorlesung Internet-
Suchmaschinen im Sommer



z.B. „Cluster“: Datenanalyse, Rechnernetze, Astronomie, ...

- Ziel
Clustering der Antwortmengen nach Kontexten
Browsen der Clusterung statt der Antwortliste