

Vorlesung Netzbasierte Informationssysteme

Information Discovery: Web Mining

Prof. Dr. Adrian Paschke

Arbeitsgruppe Corporate Semantic Web (AG-CSW)
Institut für Informatik, Freie Universität Berlin
paschke@inf.fu-berlin.de
<http://www.inf.fu-berlin.de/groups/ag-csw/>



- Web Mining
 - Web Structure Mining
 - Web Usage Mining
 - Web Content Mining



- Riesige Datenmenge
- Komplexität der Webseiten ist größer als alle traditionellen Textdokumentsammlungen
- Hoch dynamisch
- Große Vielfalt and Nutzern/Autoren
- Nur ein kleiner Teil der Informationen ist wirklich nützlich

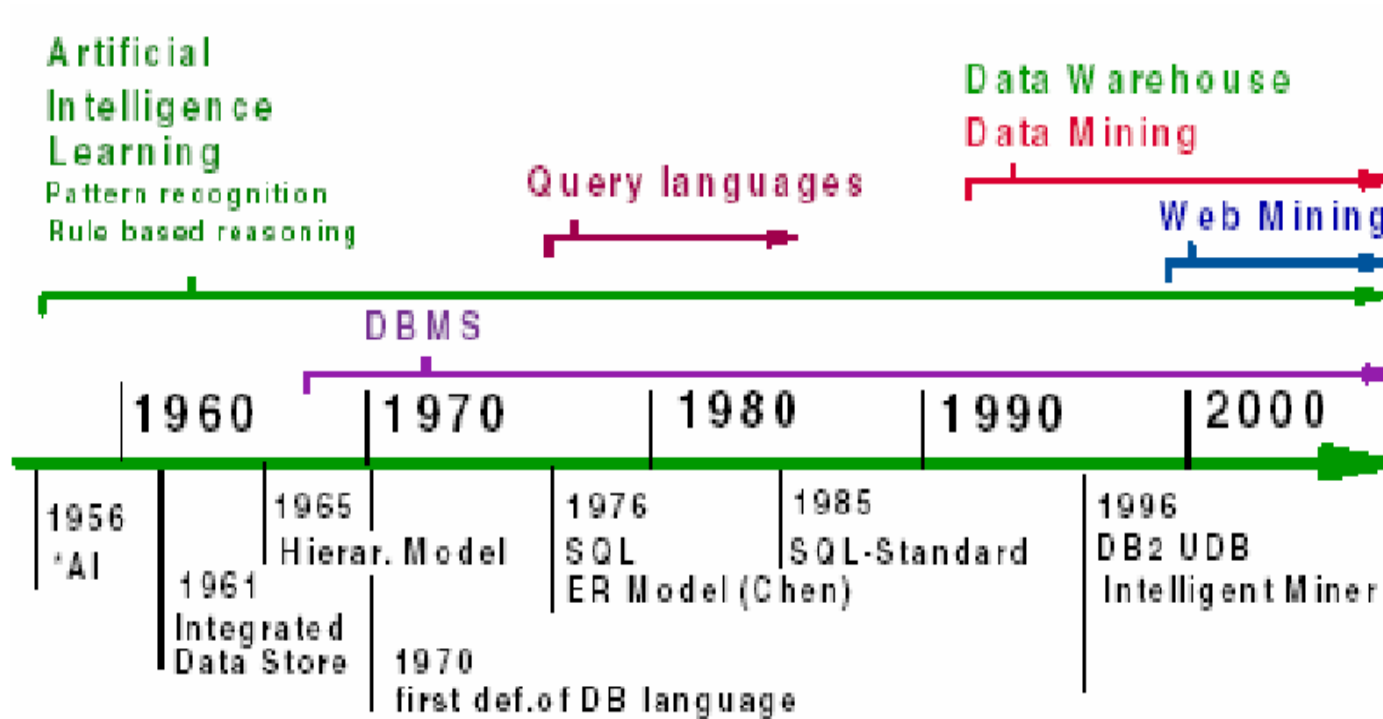


Figure 3-1 A historical view of data mining

- Web Content Mining
 - Entdeckung (discovery) der nützlichen Informationen von Webinhalten, inklusive text, image, audio, video, etc.
 - Finden von Webressourcen
 - Dokumentkategorisierung und Clustering
 - Information Extraktion von Webseiten
- Web Usage Mining
 - Fokussiert die Analyse von Logs wie Search Logs, User Activity Logs
 - Finden von interessanten Patterns der Webnutzung
- Web Structure Mining
 - Studiert das Model, welches der Linkstruktur des Web unterliegt;
 - Normalerweise auf Basis der In- und Out-Link Informationen einer Webseite

HITS

Nach: Jon M. Kleinberg: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5): 604-632 (1999)

<http://www.cs.cornell.edu/home/kleinber/auth.pdf>

- HITS - Hyperlink-Induced Topic Search [Kleinberg, 1998]
- Netzstruktur in einem Hypertext ist selber Information über den Inhalt
- Idee:
Algorithmen entwickeln, um mit Informationen über Graphstruktur besser relevante Informationen zu finden
- Identifiziere die Webseiten mit den meisten In- und Out-Links in einer Menge von Webseiten der selben Domäne.
 - Authorities: Seiten mit den meisten In-Links von Hubs
 - Hubs: Seiten mit den meisten Out-links zu Authorities
- Benutzt z.B. in der "Clever search engine" [Chakrabarti et al, 1999].
- Problem: rekursive Berechnung; berechnungsintensive

Arten von Anfragen

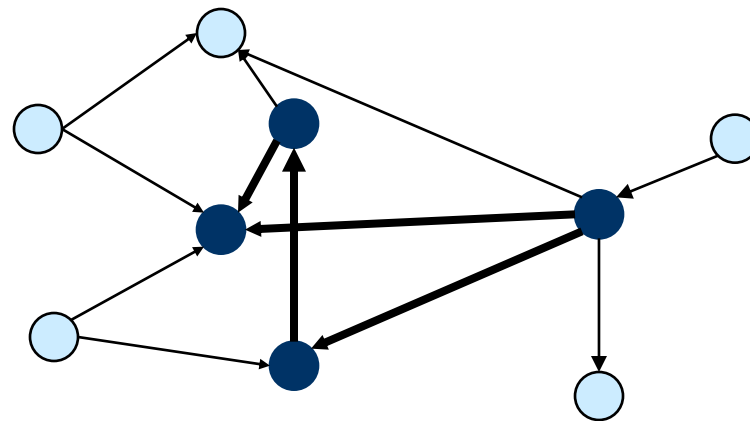
- Spezifische Anfragen:
"Does Netscape support the JDK 1.1 codesigning API?"
Wenig spezifische Antwortseiten, schwer zu finden
- Breite/Vage Anfragen:
"Find information about the Java programming language."
Viele Antwortseiten, wenige relevant
- Ähnlichkeitsanfragen: "Find pages 'similar' to java.sun.com."

Autoritative Quellen

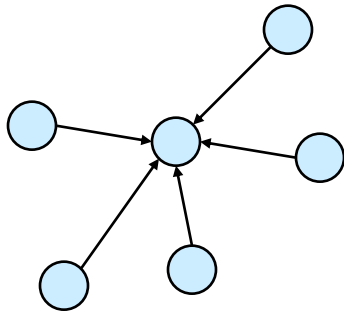
- Autoritative Antworten als Relevanzfilter
- Wie findet man autoritativ Seiten?
 - www.harvard.edu für "Harvard"?
 - Suche nach "Search engine" für Google?
 - Suche nach "automobile manufacturers" für Honda?
- Fachliche Autorität ist nicht ausschließlich endogene Eigenschaft sondern in großem Maß exogen
- Schreiben Hyperlinks dem Ziel eine fachliche Autorität zu?

- Nicht alle Links verfolgen gleiche Absicht
 - Hinweis auf wichtige Seiten
 - Navigationsunterstützung
 - Anzeigen
- Popularität vs. Autorität
 - Viele Seiten die einen Begriff enthalten und oft verlinkt werden sind nicht autoritativ
 - Seiten die allgemeinen Inhalt haben und oft verlinkt werden sind nicht für alle Themen autoritativ (eg. www.yahoo.com)

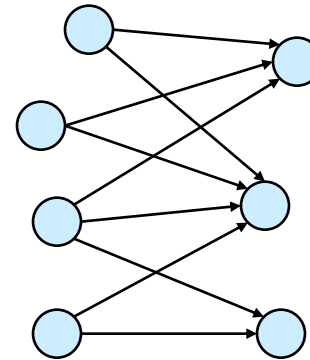
- Web ist $G=(V,E)$
 $(p,q) \in E$ ist Link von p nach q
- *out-degree* einer Seite: Anzahl abgehender Links
- *in-degree* einer Seite: Anzahl eingehender Links
- Mit $W \subseteq V$ meint $G[W]$ einen Subgraphen bei dem Knoten die Seiten aus W sind und Kanten alle Kanten zwischen Seiten aus W



- Autoritative Quellen in G_σ sollten sich dadurch auszeichnen, dass die Mengen der Seiten, die auf sie zeigen überlappen
- *Hubs* verweisen auf mehrere Autoritäten



grosser in-degree



Hubs

Autoritäten

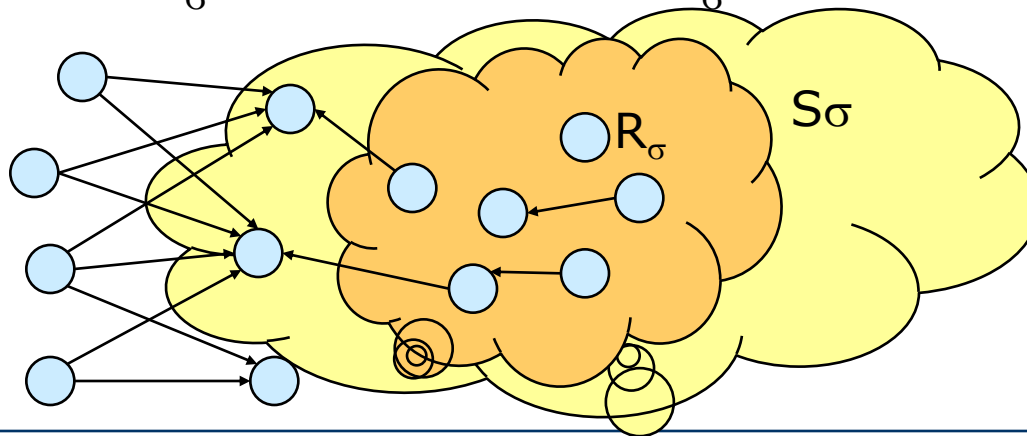
- Guter Hub zeigt auf gute Autoritäten
- Gute Autorität wird von vielen guten Hubs referenziert
- PageRank ermittelt nur populäre Autoritäten

Analyseziel

- Anfrage σ
- Ziel: Analyse der Linkstruktur in einem Teil des Web um autoritative Quellen zu finden
- Q_σ (alle Seiten, die σ enthalten)?
 - Zu groß
 - Autoritative Quellen eventuell nicht enthalten
- Gesucht: S_σ so dass:
 - S_σ vergleichsweise klein ist (i)
-> Aufwand begrenzt
 - S_σ viele relevante Seiten enthält (ii)
-> gute Autoritäten auffindbar
 - S_σ die meisten oder viele Autoritäten enthält (iii)

Root Set / Base Set

- Root-set R_σ aus ersten t Antworten einer herkömmlichen Suchmaschine auf Anfrage σ ($t=200$)
 - Erfüllt (i) und (ii): R_σ ist Untermenge von Q_σ
 - Erfüllt nicht (iii), weil Q_σ (iii) schon nicht erfüllt
 - R_σ ist nicht sehr eng verlinkt
- Wenn es aber eine Autorität für die Anfrage gibt, ist es aber wahrscheinlicher, dass auf diese von R_σ aus verwiesen wird
- -> Root-Set R_σ zum Base-Set S_σ erweitern



Errechnung des Base Set

Subgraph($\sigma, \mathcal{E}, t, d$)

σ : a query string, \mathcal{E} : a text-based search engine,
 t, d : natural numbers, Let R_σ denote the top t results of \mathcal{E} on σ .

Set $S_\sigma := R_\sigma$

For each page $p \in R_\sigma$

 Let $\Gamma^+(p)$ denote the set of all pages p points to.

 Let $\Gamma^-(p)$ denote the set of all pages pointing to p .

 Add all pages in $\Gamma^+(p)$ to S_σ .

 If $|\Gamma^-(p)| \leq d$, then

 Add all pages in $\Gamma^-(p)$ to S_σ .

 Else

 Add an arbitrary set of d pages from $\Gamma^-(p)$ to S_σ .

End

Return S_σ

- $G[S_\sigma] = \text{Subgraph}(\sigma, \text{Altavista}, 200, 50)$
 - Erfüllt in der Regel (i), (ii) und (iii)
 - Größe ca. 1000-5000 Seiten
 - Eine Referenz in 200 Seiten "findet" Autorität
- Zwei Arten von Links in $G[S_\sigma]$:
 - Transvers: Zwischen Seiten aus unterschiedlichen IP-Domains
 - Intrinsic: zwischen Seiten aus gleicher Domain
- Heuristik: Intrinsic Links dienen hauptsächlich der Navigation
- Also nur noch G_σ betrachten ($G[S_\sigma]$ ohne intrinsic)

Auffinden von Autoritäten

- Sind Seiten mit höchsten in-degree in G_σ Autoritäten?
- Nein, universell populäre Seiten sind auch in G_σ populär
- Bei Anfrage "Java" auch in G_σ mit hohem in-degree:
 - www.gamelan.com
 - java.sun.com,
 - Reklame für Karibikreisen
 - Home page von Amazon Books

- Für jede Seite p :
 - $x^{<p>}$: Authority weight
 - $y^{<p>}$: Hub weight
 - Normalisiert: $\sum_{p \in S_\sigma} (x^{<p>})^2 = 1$ und $\sum_{p \in S_\sigma} (y^{<p>})^2 = 1$
- Hohes $x^{<p>}$: p ist gute Autorität
- Hohes $y^{<p>}$: p ist guter Hub
- Operatoren \mathcal{P} und \mathcal{O} verrechnen Gewichte:

$$\mathcal{P} : x^{<p>} \leftarrow \sum_{q:(q,p) \in E} y^{<q>}$$

$$\mathcal{O} : y^{<p>} \leftarrow \sum_{q:(q,p) \in E} x^{<q>}$$

Algorithmus: Gegenseitig verstärken bis Fixpunkt erreicht

Iterate(G, k)

G : a collection of n linked pages

k : a natural number

Let z denote the vector $(1, 1, 1, \dots, 1) \in \mathbb{R}^n$.

Set $x_0 := z$.

Set $y_0 := z$.

For $i = 1, 2, \dots, k$

 Apply P operation to (x_{i-1}, y_{i-1}) , obtaining new x-weights x'_i

 Apply O operation to (x_i, y_{i-1}) , obtaining new y-weights y'_i

 Normalize x'_i , obtaining x_i .

 Normalize y'_i , obtaining y_i .

End

Return (x_k, y_k) .

Angewandt zum Filtern

Filter(G, k, c)

G : a collection of n linked pages

k, c : natural numbers

$(x_k, y_k) := \text{Iterate}(G, k)$.

Pages with the c largest coordinates in x_k as authorities.

Pages with the c largest coordinates in y_k as hubs.

- Mit steigendem k konvergieren $\{x_k\}$ und $\{y_k\}$ zu Fixpunkten x^* und y^*
- Konvergenz sehr schnell: Schon $k=20$ reicht, um gefundene beste Autoritäten und Hubs stabil zu halten

- Anfrage "Java"

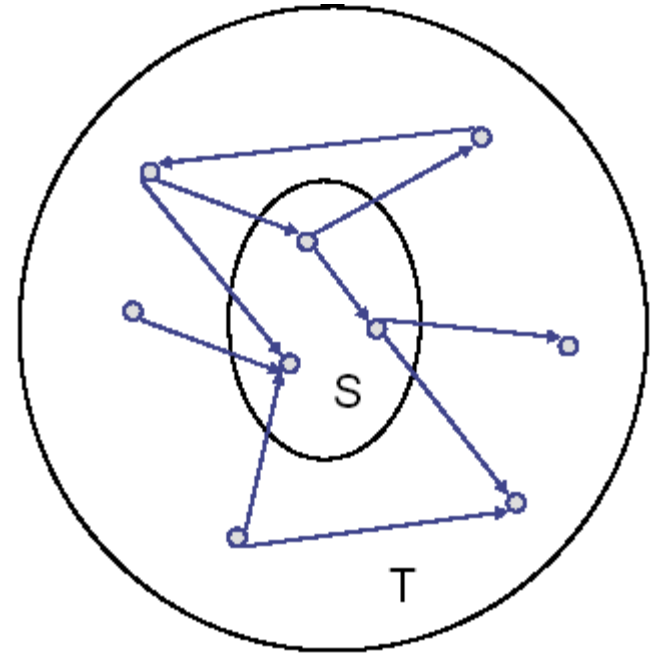
.328	http://www.gamelan.com/	<i>Gamelan</i>
.251	http://java.sun.com/	<i>JavaSoft Home Page</i>
.190	http://www.digitalfocus.com/digitalfocus/faq/howdoi.html	<i>The Java Developer: How Do I . . .</i>
.190	http://lightyear.ncsa.uiuc.edu/~srp/java/javabooks.html	<i>The Java Book Pages</i>
.183	http://sunsite.unc.edu/javafaq/javafaq.html	<i>comp.lang.java FAQ</i>

- Anfrage "Search Engines"

.346	http://www.yahoo.com/	<i>Yahoo!</i>
.291	http://www.excite.com/	<i>Excite</i>
.239	http://www.mckinley.com/	<i>Welcome to Magellan!</i>
.231	http://www.lycos.com/	<i>Lycos Home Page</i>
.231	http://www.altavista.digital.com/	<i>AltaVista: Main Page</i>

Zusammenfassung HITS

- HITS stellt ein Startset S von Seiten zusammen, die mit dem Schlüsselwort übereinstimmen.
- Dann wird S zu einem größeren Basisset T expandiert, indem jede Seite hinzugefügt wird, die auf eine Seite in S zeigt oder von einer Seite in S verlinkt wird.
- Assoziiere mit jeder Seite p ein Hub-Gewicht $h(p)$ und ein Authority-Gewicht $a(p)$, alle initialisiert mit 1.
- Iterative aktualisiere die h 's und a 's
 - $a(p) := \sum_{q \rightarrow p} h(q)$
 - $h(p) := \sum_{p \rightarrow q} a(q)$
- Jede Iteration ersetzt $a(p)$ durch die Summe der $h()$'s von Seiten welche auf p zeigen und $h(p)$ durch die Summe der $a()$'s die von p verlinkt sind.
- Die Iterationen konvergieren zu einem stabilen Set an Authority- und Hub-Gewichten.

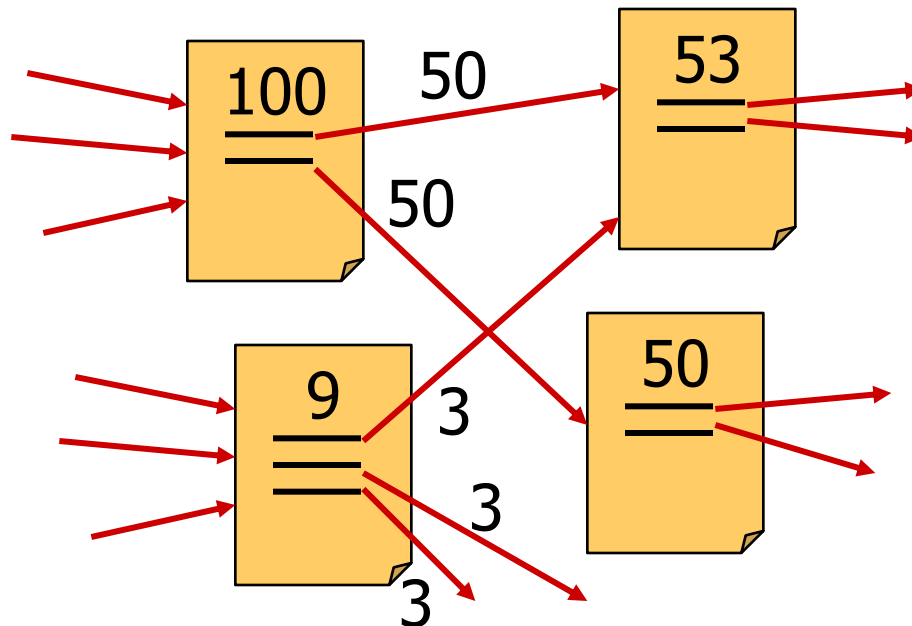


Pagerank

Page, L., Brin, S., Motwani, R., Winograd, T.L.:
The PageRank Citation Ranking: Bringing Order to the Web
<http://newdbpubs.stanford.edu:8090/pub/1999-66>

- PageRank Verfahren: Bewertung aller Web-Seiten nach ihrer relativen Popularität [Brin and Page '98]
- Kerntechnologie von Google
- Viele Verweise auf eine Seite legen nahe, dass die Seite wichtig ist
- Setzen eines Links ist Einschätzung der Wichtigkeit der referenzierten Seite (ähnlich einem Zitat)
 - Sie ist aber eigentlich nur populär!
- Links von wichtigen Seiten erhöhen Wichtigkeit
 - Eine Seite hat hohen PageRank, wenn die PageRanks der Seiten, die auf sie verweisen, hoch sind
- Seiten mit hohem Pagerank werden in Ergebnismengen zuerst gelistet

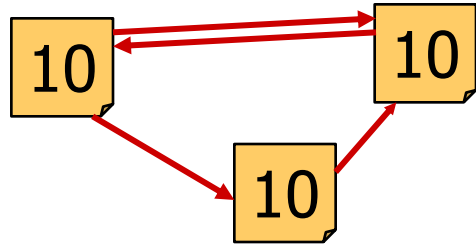
- Jede Seite "verteilt" ihren Pagerank auf die von ihre referenzierten Seiten



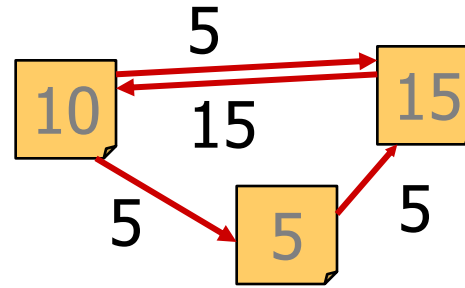
- Wird iterativ errechnet
- Notwendig: Komplette Linkstruktur des Web

Beispieliterationen

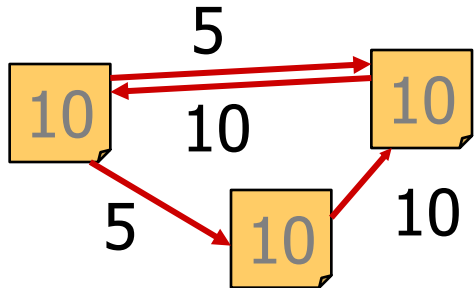
a) Start



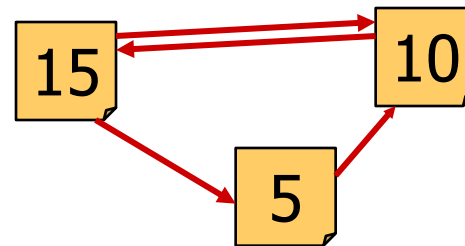
d) Ranks verteilen



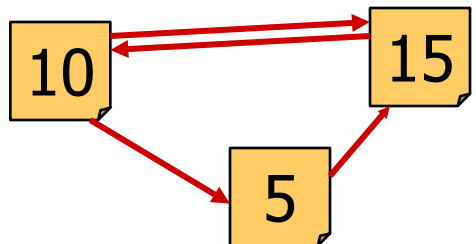
b) Ranks verteilen



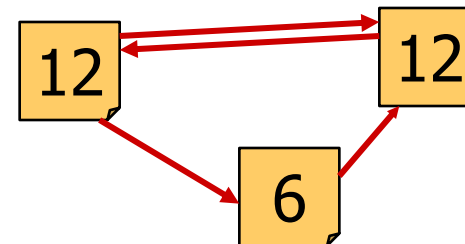
e) Zwischenstand



c) Zwischenstand

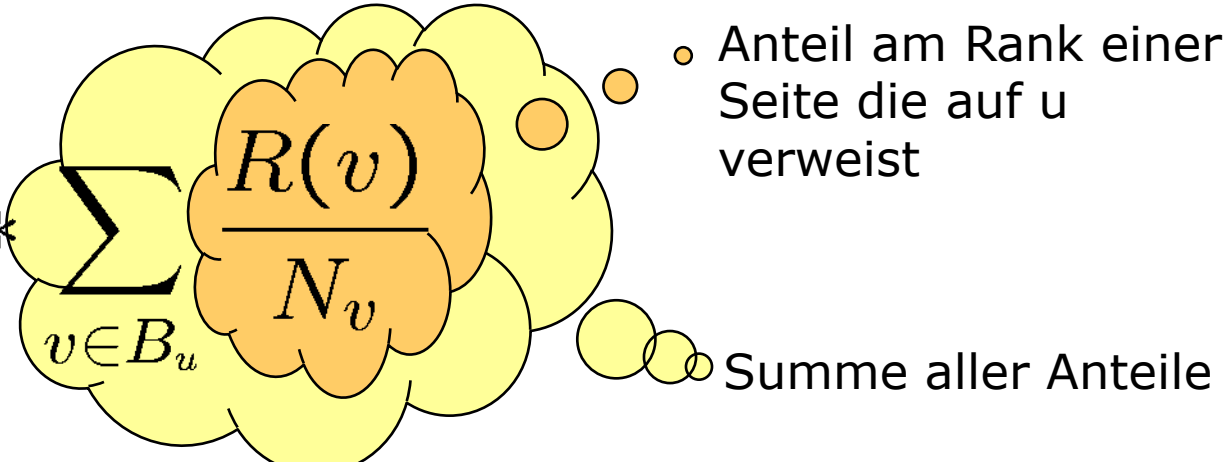


...x) Konvergierter Stand



PageRank

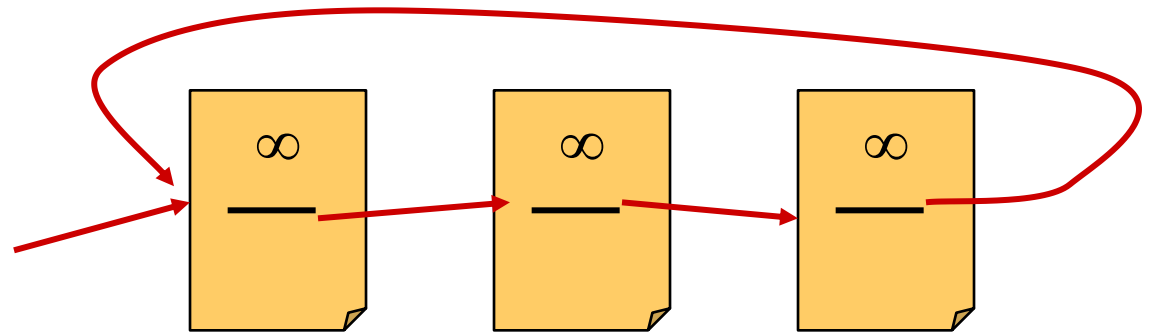
- Wir suchen den Pagerank für eine Web-Seite u
- F_u : Menge der Seiten, auf die u verweist
- B_u : Menge der Seiten, die auf u verweisen
- $N_u = |F_u|$: Anzahl der Links von u (out-degree)
- c : Faktor zur Normalisierung
- Vereinfachter PageRank $R(u)$ der Seiten u :

$$R(u) = c * \sum_{v \in B_u} \frac{R(v)}{N_v}$$


- Anteil am Rank einer Seite die auf u verweist
- Summe aller Anteile

- c normalisiert die Summe aller PageRanks zu 1

- Problem Schleifen:



- Ist ein „Ranksink“, weil aus der „Schleife“ heraus keine Ranks abgegeben werden
- Lösung: Ausgleichen durch „Ranksource“, die ein Vektor $E(u)$ aus Webseiten mit Rank ist
- Dieser Vektor wird zu den PageRanks hinzugerechnet

$$R'(u) = c * \sum_{v \in B_u} \frac{R'(v)}{N_v} + c * E(u)$$

so dass c maximiert ist und $\|R'\|_1 = 1$

Vollständiger PageRank

- Intuition: Einfache Version modelliert Zufalls-Surfer und die Wahrscheinlichkeit, dass Seiten von ihm durch zufälliges Verfolgen von Links besucht werden
- In einer Schleife springt der Surfer zu einer beliebigen andern Seite. E modelliert die Verteilung dieser zufälligen Auswahl
- PageRank ist die Verteilung der Wahrscheinlichkeit eine bestimmte Seite durch zufällige Navigation zu erreichen
- Bestimmung aus globaler Sicht

- Basierend auf Vektorensicht:

$$R_0 = S$$

loop:

$$R_{i+1} = AR_i$$

$$d = ||R_i||_1 - ||R_{i+1}||_1$$

$$R_{i+1} = R_{i+1} + dE$$

$$\delta = ||R_{i+1} - R_i||_1$$

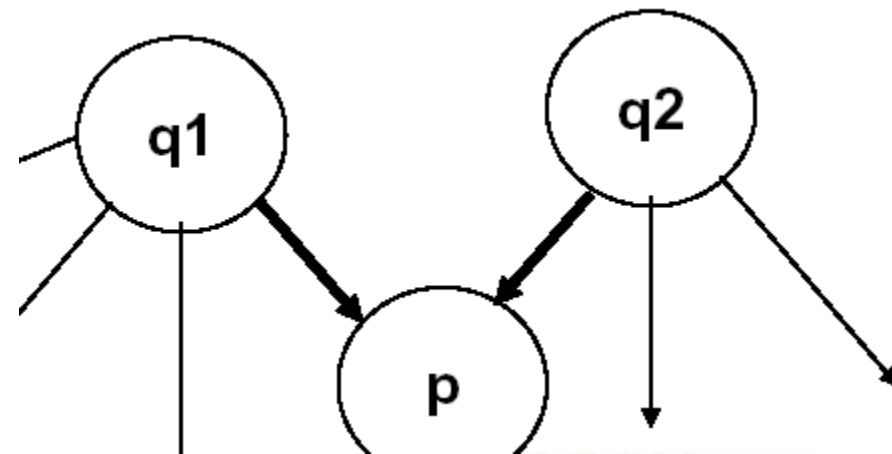
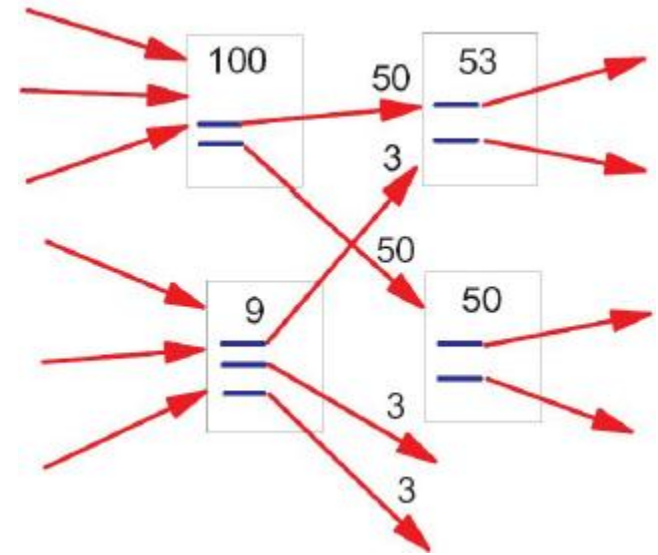
while $\delta > \varepsilon$

- Die Vektoren sind aber *sehr* groß
(=Anzahl der besuchten Seiten des Web)
- Verschiedene Verfahren zur effizienten Errechnung
vorhanden

Zusammenfassung: Page Rank

- Annahme: Ein Link von Seite q1 zu Seite p ist eine Empfehlung von Seite p durch den Author von q1 (p ist ein *Nachfolger* von q1)
- Qualität der Seite bezieht sich auf den In-degree der Seite
- Qualität einer Seite bestimmt durch
 - ihren in-degree, und durch
 - die *Qualität* von Seiten die auf sie zeigen
- Ergebnisse in "Eigenvector centrality"

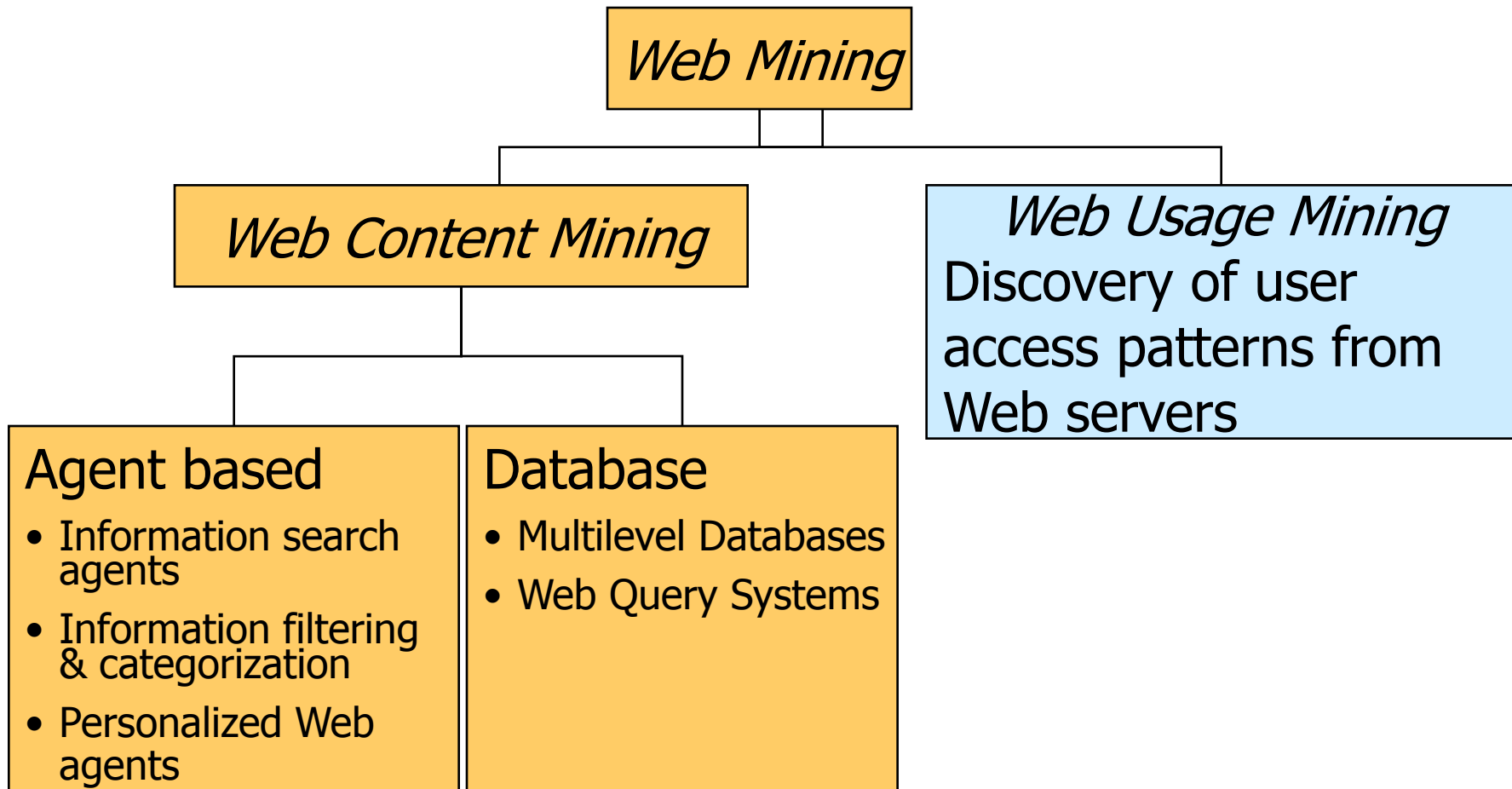
$$PR(p) = \left(\frac{PR(q1)}{4} + \frac{PR(q2)}{3} \right) d + 1 - d$$



- Web Content Mining
 - Entdeckung (discovery) der nützlichen Informationen von Webinhalten, inklusive text, image, audio, video, etc.
 - Finden von Webressourcen
 - Dokumentkategorisierung und Clustering
 - Information Extraktion von Webseiten
- Web Usage Mining
 - Fokussiert die Analyse von Logs wie Search Logs, User Activity Logs
 - Finden von interessanten Patterns der Webnutzung
- Web Structure Mining
 - Studiert das Model, welches der Linkstruktur des Web unterliegt;
 - Normalerweise auf Basis der In- und Out-Link Informationen einer Webseite

- Nutzer von Web-Sites sind für den Server anonym
 - Keine Identifikation des tatsächlichen Rechners:
Proxies, Caches, private Netze, dynamische IP-Nummern
 - Keine Identifikation des Nutzerprozesses:
Mehrbenutzerrechner, Proxies, Caches
 - Keine Identifikation des Nutzers:
Account-Informationen lokal
- Informationen über Nutzer sind aber nützlich
 - Personalisierung
 - Optimierung des Angebots
 - Grundlage des Geschäftsmodells

- *Web Mining*: The discovery and analysis of useful information from the Web



Logfiles auf Web-Servern

- Logfiles werden zeilenweise geschrieben
- Mögliches Format: Common Logfile Format
 - remotehost: IP-Nummer oder Name des Client-Rechners
 - rfc1413: Nutzer-ID auf Quellrechner (ident Dienst)
 - authuser: Nutzer-ID für Web-Session
 - [date]: Datum des Eintrags
 - "request": HTTP-Request Zeile
 - status: HTTP Antwortcode
 - bytes: Größe der Antwort

- - - - [19/Dec/2002:10:07:30 +0100] ↵
 "GET /~paschke/coo12.gif HTTP/1.1" 200 4942
- - - - [19/Dec/2002:10:08:06 +0100] ↵
 "GET /~paschke/%22http://www.dcs.ed.ac.uk/home/cdw/
 phdproject/SECD/Applet/lispkit.html%22 HTTP/1.1" 404 -

- Extended Common Logfile Format
 - CLF Felder
 - "referer": Seite von der Link verfolgt wurde
 - "user agent": Client-Software
- - - - [19/Dec/2002:10:07:30 +0100] ↵
"GET /~paschke/coo12.gif HTTP/1.1" 200 4942 ↵
"http://grunge.cs.tu-berlin.de/~paschke/vmlanguages.html" ↵
"Mozilla/4.0%20(compatible;%20MSIE%206.0;%20windows%20NT%205.1)"

- Probleme:
 - remotehost:
Nummer des Rechners, der einen Socket zum Server aufbaut ist noch nicht Rechner an dem der Nutzer ist
 - rfc1413:
Läuft ident-Dienst überhaupt? Was soll man mit Ergebnis anfangen?
 - [date]:
Nicht eindeutig bei vielen Zugriffen in kurzen Abständen
 - "request":
GET mit IfModifiedSince-Header, Caches
 - "referer":
Nicht bei Direkteingabe, Bookmarks
 - "user_agent":
Keine zuverlässige Angabe, was ist mit Crawlern?

- Auf Basis von Logfiles lassen sich verschiedene Aussagen über die Nutzung einer Site treffen
- Insbesondere sind diese Aussagen Basis für die Preisfindung der Werbewirtschaft
- Diese Aussagen sind von unterschiedlicher Güte

- Hits
 - Anzahl der Abrufe von Informationen
 - Summe der Anzahl der Requests mit 200 und 304 Antwort
 - Nicht sehr aussagekräftig, weil nicht jede Datei eigenständige Informationseinheit
- Pageviews/Page Impressions
 - Anzahl der abgerufenen HTML-Seiten
 - Anzahl der Hits mit HTML Dateien als Antwort
 - Beschränkt auf einen Medientyp

- 4 Hits, 1 Pageview:
 - - - [19/Dec/2002:12:05:51 +0100]
"GET /~paschke/vmlanguages.html HTTP/1.1" 200 81671
"http://search.msn.com/results.asp?FORM=SCPN&RS=CHECKED&un=doc
&v=1&q=java%20window%20commands"
"Mozilla/4.0%20(compatible;%20MSIE%206.0;%20windows%20NT%205.1)"
 - - - [19/Dec/2002:12:05:51 +0100]
"GET /~paschke/unclear.gif HTTP/1.1" 200 988
"http://flp.cs.tu-berlin.de/~paschke/vmlanguages.html"
"Mozilla/4.0%20(compatible;%20MSIE%206.0;%20windows%20NT%205.1)"
 - - - [19/Dec/2002:12:05:51 +0100]
"GET /~paschke/new.gif HTTP/1.1" 200 907
"http://flp.cs.tu-berlin.de/~paschke/vmlanguages.html"
"Mozilla/4.0%20(compatible;%20MSIE%206.0;%20windows%20NT%205.1)"
 - - - [19/Dec/2002:12:05:51 +0100]
"GET /~paschke/cool2.gif HTTP/1.1" 200 4942
"http://flp.cs.tu-berlin.de/~paschke/vmlanguages.html"
"Mozilla/4.0%20(compatible;%20MSIE%206.0;%20windows%20NT%205.1)"
- Pageviews und Framesets
 - Erster Abruf des Framesets ist 1 Pageview
 - Jedes Neuladen eines Inhaltsframes ist 1 Pageview
 - Zum Messen immer nur einen Frame neuladen (DMMV)

- Visits / Sessions
 - Zusammenhängende Abrufe in einem Zeitraum
 - Navigationspfade aus Logfile
 - Nicht zuverlässig identifizierbar
 - Problem: Wann ist Visit beendet?
- Heuristiken
 - Zeitorientiert:
 - Gesamtdauer einer Visit ist nach oben begrenzt
 - Verweildauer auf einer Seite ist nach oben begrenzt
 - Navigationsorientiert
 - Topologische Begrenzung: Sitzungsende, wenn Seite nicht von vorherigen Seiten aus erreicht werden konnte
 - Begrenzung durch Referrer: Sitzungsende, wenn Seite nicht durch Navigation von vorheriger Seite erreicht wurde

- Unique Visitors
 - Abrufe von gleicher IP Adressen als 1 Besucher gezählt
 - Objektiv nicht aussagefähig (Proxies, Dynamische IP Adressen, etc.)
- AdImpressions / Clickthroughs
 - Klick auf Werbebanner
 - Messbar beim Werbekunden
 - Quelle durch Referer ermittelbar
 - Abrechnung
 - Preis nach Attraktivität des Werbeträgers: Pageviews und Visits als Maß
 - Preis nach Effizienz des Werbemittels: Clickthroughs als Maß

- Viewtime
 - Dauer des Verweilens auf einem Angebot
 - Kaum aus Logfile messbar
 - Klientenseitige Unterstützung notwendig (z.B. Scripting)
 - Sitzt der Nutzer vor dem Rechner?
- Durch zusätzliche direkte Befragung ermittelbar:
 - Qualified visits: Bestätigte Besuche
 - Regionale Herkunft
 - Alter, Geschlecht etc.
 - Interessen
 - Akzeptanz

Wer misst?

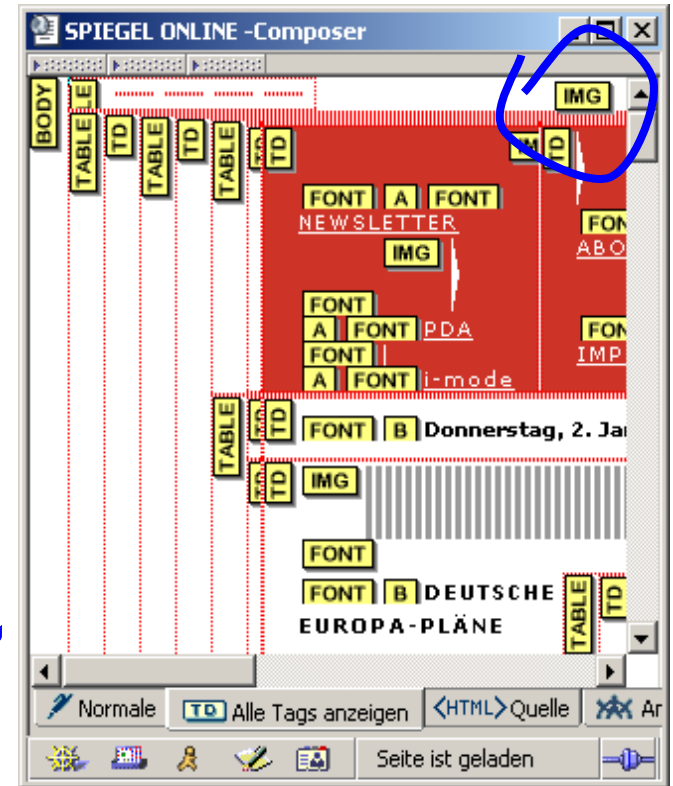
- Serverbetreiber nach eigenem Verfahren und eigener Auswertung
- Serverbetreiber oder Externer nach standardisiertem Verfahren und Auswertung
 - „Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V.“ (IVW) (<http://www.ivwonline.de/>)
 - Messung z.B. durch transparente Grafiken („IVW-Pixel“) auf Seiten
 - ``
 - ``
 - Lösen Messung aus
 - IVW Zahlen sind Grundlage für Preisgestaltung

Aus www.spiegel.de/index.html

```
<body bgcolor="#ffffff" text="#000000"
 link="#b20a15" vlink="#b20a15" alink="#ff0000"
 marginheight="0" marginwidth="4" leftmargin="4"
 topmargin="0" rightmargin="4" bottommargin="0">
<!-- IVW VERSION="1.2" -->
<script language="JavaScript">
<!--
 var IVW="http://spiegel.ivwbox.de/cgi-bin/ivw/CP/
   spiegel;/home/c-18/be-PB64-aG9tZXBhZ2UvY2VudGvy";
 document.write('<IMG SRC="' + IVW + '?r=' +
   escape(document.referrer) + '" WIDTH="1,,
   HEIGHT="1" BORDER="0" ALIGN="RIGHT">');
// -->
</script>

<noscript>
 <IMG SRC="http://spiegel.ivwbox.de/cgi-bin/ivw/CP/spiegel;
   /home/c-18/be-PB64-aG9tZXBhZ2UvY2VudGvy"
   WIDTH="1" HEIGHT="1" BORDER="0" ALIGN="RIGHT">
</noscript>
<!-- /IVW -->
<!-- IVW VERSION="prev" -->

<!-- /IVW -->
```



IVW Messungen

IVW Online Nutzungsdaten 09-2007	Info		Basisdaten		Kategorien					
Gemeldete Angebote: 506	Info	Local - Liste	Visits	PageImpressions	Redaktioneller Content	User generierter ...	E - Commerce	Kommunikation	Suchmaschinen, Verzeichnisse ...	Spiele
Gemeldete PageImpressions: 18.361.197.827										
Angebote										
Netzwerke										
Vermarktungsgemeinschaften										
swinger-tageblatt.de			126.262	1.312.753	1.249.362	7.300	12.147	1.167	40.941	4
Spektrum der Wissenschaft/Wissenschaft online			377.983	1.364.186	1.025.841	—	259.485	15.135	7	—
SPIEGEL ONLINE			68.659.980	405.467.226	395.959.105	5.922.523	2.643.487	73.551	—	—
Spieletipps.de			4.574.826	39.568.845	38.561.141	62.843	—	237.346	—	—
Spin.de			7.731.337	308.934.589	117.483	166.982.170	—	136.671.311	—	4.754.858
Sport Auto			109.071	737.631	715.600	20.671	1.330	—	—	—
Sport1			24.544.218	174.416.190	161.198.918	—	—	1.481.473	—	10.853.759
Hinweis: Aus technischen Gründen konnten für diesen Monat keine vollständigen Nutzungsdaten ermittelt werden										
STADTPLANDIENST			1.307.977	16.759.249	12.309	—	18.801	—	16.727.932	—
STAR FM 87.9 MAXIMUM ROCK!			124.213	448.072	441.577	—	—	—	5.226	—
Stellen Online			57.584	248.899	13.253	—	197.455	—	38.167	—
stellenanzeigen.de			995.518	6.112.430	2.018.916	—	3.917.382	176.048	—	—
stern.de			12.759.503	130.459.351	108.625.499	20.269.217	32.818	1.137.166	190.344	—
studieren.de			126.602	1.933.270	1.933.254	—	—	—	—	—
StudiVZ			111.637.977	3.666.027.724	17.293.349	3.648.217.686	429.792	—	—	—
Stuttgarter Zeitung online & Stuttgarter Nachricht ...			1.139.171	8.337.212	3.882.245	2.425.551	1.842.704	10.523	46.172	40.035
Stylepark			82.520	818.153	29.782	—	—	—	788.370	—

- Datenaufbereitung
 - Extraktion relevanter Zugriffe, also z.B. nicht Hits auf eingebettete Daten etc. (jpg, map, robots.txt)
 - Hinzufügen verloren gegangener Zugriffe
 - Zusammenführen mit Cookie-Informationen
 - Zusammenführen mit Registrierungsinformationen
 - Heuristiken zur Cache Nutzung
- Sitzungsermittlung
 - Ziel: Sequenzen von zusammengehörigen (gleicher Nutzer, gleiche Nutzung) Zugriffen als Sitzung (Session, Visit) identifizieren
 - Problem vergleichbar mit dem Problem der Identifizierung eines Nutzers

- Path analysis

- Ermittlung von Pfaden in Graphen (oder Graphen aus Pfaden), die Web-Site repräsentieren
 - Link-Struktur einer Site
 - Ähnlichkeitsstruktur von Seiten einer Site
 - Linkverfolgungsstruktur einer Site
- Weitere Zusammenhänge ermitteln:
 - 70% der Nutzer, die `/inst/ag-nbi/lehre/08/S_SW/` zugegriffen haben kamen über den Pfad `/inst, /inst/ag-nbi` (20% über `/lehre/, ...`)
 - 5% der Nutzer haben ihren Besuch bei `/inst/ag-nbi` begonnen
 - 70% der Nutzer haben ihre Sitzung nach einem Pfad der Länge 5 beendet
- Nutzung dieser Zusammenhänge für die Struktur der Site

- Association rule
 - Ermittlung von Korrelationen zwischen Zugriffen einer Sitzung
 - 30% der Nutzer die `/inst/ag-nbi` besucht haben, haben auch `/inst/ag-tech` besucht
 - 2% der Nutzer von `/inst/ag-nbi/lehre/0809/V_NBI/` haben sich danach in die Mailingliste eingetragen auf http://lists.spline.inf.fu-berlin.de/mailman/listinfo/nbi_v_nbi
- Sequential pattern
 - Ermittlung von Zusammenhängen zwischen Sitzungen
 - 20% der Nutzer, die sich über http://lists.spline.inf.fu-berlin.de/mailman/listinfo/nbi_v_nbi in eintragen, haben das innerhalb von 10 Tagen über http://lists.spline.inf.fu-berlin.de/mailman/listinfo/nbi_s_xml auch für die andere Mailingliste eingetragen

- Classification rules
 - Ermittlung von Profilen von Nutzergruppen
 - 80% derjenigen, die sich unter http://lists.spline.inf.fu-berlin.de/mailman/listinfo/nbi_v_nbi eingetragen haben, studieren Diplom-Informatik
 - Bachelor-Studierende besuchen eher Seiten unter /inst/ag-nbi als unter /inst/ag-bio
- Clustering
 - Gruppierung ähnlicher Nutzer und Daten
 - Interesse an hochpreisiger Consumer-Electronic
 - Nutzung für Marketing und Site-Personalisierung (z.B. amazon)

- Ermittlung von Informationen über Nutzer und Nutzung notwendig
- Logfiles als Datenbasis bei Servern, verschiedene Format
- Verschiedene Messgrößen verbreitet
- Ermittlung teilweise sehr schwer
- Web Usage Mining zur Ermittlung komplexerer Zusammenhänge

- *Common Logfile Format*.
<http://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format>
- Mike StJohns. *Identification Protocol*. Request for Comments 1413. February 1993
<http://www.ietf.org/rfc/rfc1413.txt?number=1413>
- DMMV. *Messgrößen*.
http://www.dmmv.de/de/7_pub/homepagedmmv/themen/emarketing/media/zielemedia.cfm
- Bettina Berendt, Bamshad Mobasher, Myra Spiliopoulou, and Jim Wiltshire. *Measuring the accuracy of sessionizers for web usage analysis*. In Workshop on Web Mining at the First SIAM International Conference on Data Mining, pages 7-14, April 2001.
<http://maya.cs.depaul.edu/~mobasher/papers/wm-siam01.pdf>
- R. Cooley, B. Mobasher, J Srivastava. *Web Mining: Information and Pattern Discovery on the World Wide Web*. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.
<http://maya.cs.depaul.edu/~mobasher/papers/webminer-tai97.ps>

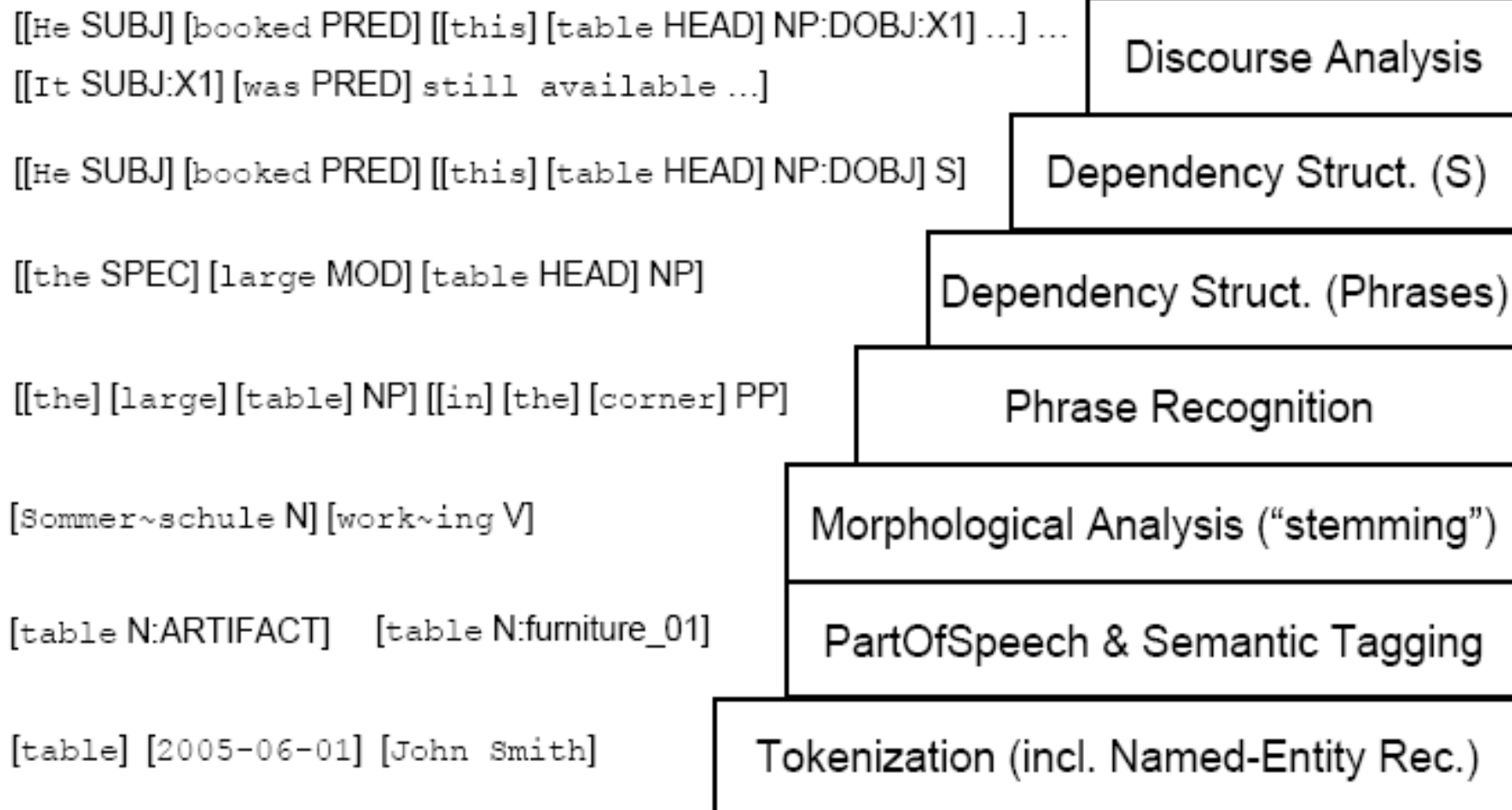
- Web Content Mining
 - Entdeckung (discovery) der nützlichen Informationen von Webinhalten, inklusive text, image, audio, video, etc.
 - Finden von Webressourcen
 - Dokumentkategorisierung und Clustering
 - Information Extraktion von Webseiten
- Web Usage Mining
 - Fokussiert die Analyse von Logs wie Search Logs, User Activity Logs
 - Finden von interessanten Patterns der Webnutzung
- Web Structure Mining
 - Studiert das Model, welches der Linkstruktur des Web unterliegt;
 - Normalerweise auf Basis der In- und Out-Link Informationen einer Webseite

- **Web Content Mining** befasst sich mit der Erkennung von Regularitäten in den Inhalten einer Webressource
- Web Content Mining von multiplen Datentypen wird als „**Multimedia Data Mining**“ bezeichnet
- Web Content Mining ist ein Anwendungsgebiet für das Textmining
 - Text Mining kann als Instanz und übergeordnetes Forschungsgebiet von Web Content Mining verstanden werden
 - Verwendete Methoden sind allgemeine Text/Data-Mining-Methoden, wobei statistische und computerlinguistische Verfahren die Transformation der Texte in eine (für das Text/Data Mining) adäquate Form realisieren

	Web Content Mining	
	IR View	DB View
View of Data	<ul style="list-style-type: none"> - Unstructured - Semi structured 	<ul style="list-style-type: none"> - Semi structured - Web site as DB
Main Data	<ul style="list-style-type: none"> - Text documents - Hypertext documents 	<ul style="list-style-type: none"> - Hypertext documents
Representation	<ul style="list-style-type: none"> - Bag of words, n-grams - Terms, phrases - Concepts or ontology - Relational 	<ul style="list-style-type: none"> - Edge-labeled graph (OEM) - Relational
Method	<ul style="list-style-type: none"> - TFIDF and variants - Machine learning - Statistical (including NLP) 	<ul style="list-style-type: none"> - Proprietary algorithms - ILP - (Modified) association rules
Application Categories	<ul style="list-style-type: none"> - Categorization - Clustering - Finding extraction rules - Finding patterns in text - User modeling 	<ul style="list-style-type: none"> - Finding frequent sub-structures - Web site schema discovery

Kosala, Raymond und Blockeel, Hendrik (2000), Web Mining Research: A Survey, SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, Volume 2, Issue 1, o.O., Seite 1-10

Beispiel: Linguistische Methoden zur Termextraktion



- Nutzung von Bewertungen in der Termextraktion
 - MI (Mutual Information) - Cooccurrence Analysis

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

Informations Entropy – Erwartungswert des Informationsinhalts von X

$$H(X) = E(I(X)) = -\sum_{i=1}^n p(x_i) \log_2 p(x_i)$$

- TFIDF - Termgewichtung

$$tfidf(w) = tf \cdot \log\left(\frac{N}{df(w)}\right)$$

- χ^2 (Chi-square) - Cooccurrence Analysis & Term Weighting

$$\chi^2 = \sum \frac{(obs - exp)^2}{exp}$$

Beispiel: Informations Entropy

Information can be a message which was received and understood

Information can be the knowledge acquired through study or experience or instruction.

Information can be a formal accusation of a crime.

Information can be a collection of facts from which conclusions may be drawn.

Information a numerical measure of the uncertainty of an outcome.

$$H_a(\text{Information}) = -5/53 \log_2 5/53$$

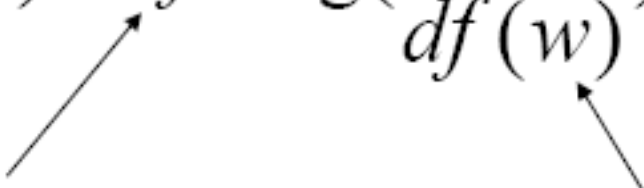
$$= \mathbf{0.32 \text{ bit}}$$

$$H_b(\text{uncertainty}) = -1/53 \log_2 1/53$$

$$= \mathbf{0.10 \text{ bit}}$$

Count(Information)	5
Total Number of words	53

most popular weighting schema
(normalized word frequency)

$$tfidf(w) = tf \cdot \log\left(\frac{N}{df(w)}\right)$$


The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

$tf(w)$ term frequency (number of word occurrences in a document)

$df(w)$ document frequency (number of documents containing the word)

N number of all documents

$tfidf(w)$ relative importance of the word in the document

Example: TFIDF

Information can be a message which was received and understood

Information can be the knowledge acquired through study or experience or instruction.

Information can be a formal accusation of a crime.

Information can be a collection of facts from which conclusions may be drawn.

Information a numerical measure of the uncertainty of an outcome.

$$\text{tfidf}(X) = \text{tf}(X) * \log (N/\text{df}(X))$$

$$\text{tfidf}_{\text{doc1}}(\text{message}) = 1 * \log (5/1) = 0.7$$

$$\text{tfidf}_{\text{doc1}}(\text{be}) = 1 * \log (5/4) = 0.1$$

$$\text{tfidf}_{\text{doc1}}(\text{Information}) = 1 * \log (5/5) = 0, \\ \text{with } \log(0) \text{ defined as } 0$$

$$\text{tfidf}_{\text{doc1}}(\text{message}) = 1 * \log (16\text{M}/16\text{K}) = 3$$

$$\text{tfidf}_{\text{doc1}}(\text{Information}) = 1 * \log (16\text{M}/489\text{K}) = 1.51$$

$$\text{tfidf}_{\text{doc1}}(\text{be}) = 1 * \log (16\text{M}/5,6\text{M}) = 0.54$$

$$\text{tfidf}(\text{Information}) = 5 * \log (16\text{M}/489\text{K}) = 7,55$$

$$\text{tfidf}(\text{message}) = 1 * \log (16\text{M}/16\text{K}) = 3$$

$$\text{tfidf}(\text{be}) = 4 * \log (16\text{M}/5,6\text{M}) = 2,16$$

Count(Information)	5
Total Number of words	53

McCain calls timeframe for Iraq withdrawal 'not too important'

Republican presidential nominee John McCain this morning categorised the time frame for US troops withdrawal from Iraq as "not too important" and suggested he is prepared for a long-term commitment there.

"We are succeeding" in Iraq, he said. "And it's fascinating that Senator Obama doesn't realise that."

The remarks provided an immediate opportunity for presumptive Democratic nominee Barack Obama's campaign to bring the war back to the forefront of the race. He has been a steadfast opponent of the war since 2002.

His campaign responded immediately and forcefully this morning, lining up a team of surrogates who said McCain's statements show he doesn't understand the nature of the conflict there and is out of touch with the desire of the American people to see an end to the deployment.

Phrases gewichtet mit TFIDF

iraq
campaign
war
McCain's statements
Republican presidential nominee John McCain
Senator Obama
US troops withdrawal
immediate opportunity
long-term commitment
presumptive Democratic nominee Barack Obama's campaign
steadfast opponent
american people
forefront
surrogates
time frame
remarks
deployment
touch
desire
conflict
race
team
nature
morning
end

- Entity Recognition (ER) ist die Aufgabe der Erkennung von Ausdrücken in natürlichsprachigen Dokumenten die Entitäten bezeichnen
- ER kann als eine Klassifikationsaufgabe gesehen werden, wobei jedes Wort oder Token im Text einer Klasse zugeordnet wird, z.B.

Erkennung von Personennamen :

John Denver will be playing at the Denver Buffalo Company in Denver

P **P** N N N N N N N N N N

Erkennung von Firmennamen:

John Denver will be playing at the Denver Buffalo Company in Denver

N N N N N N N **C** **C** **C** N N

- z.B. part-of-speech tagging

John Denver will be playing at the Denver Buffalo Company in Denver

NP NP MD VB VBG IN DT NP NP NP IN NP

- Penn Treebank Tags (Word level)

NP proper noun, singular

MD modal verb

VB verb, base form

VBG verb, gerund

IN preposition

DT determiner

Entity Recognition (ER)

"John Denver will be playing at the Denver Buffalo Company in Denver, CO."



"John Denver will be playing at the Denver Buffalo Company in Denver, CO."



"John Denver will be playing at the Denver Buffalo Company in Denver, CO."

is a Person

is a Company

is a Place

Namesvariationen:

- Lexikalisch: tumour, tumor
organisation, organization
lorry, truck
- Orthographisch: a helix, α -helix, alpha helix
Rocky II, Rocky 2
- Strukturell: lung cancer, cancer of the lung
- Morphologisch: kick, kicks, kicked (person, time)
human, humans; mouse, mice (plural, singular)

- Mehrdeutigkeiten:
 - People vs. Companies: Ford, Philip Morris
 - People vs. Location: Paris, Washington, JFK
- Abkürzungen:
 - PCR - Polymerase Chain Reaction
 - Program Clock Reference
 - Patient Care Report
 - Payload Changeout Room
 - Photochemical Reaction
 - ...

- Anaphora:

- (Ein Element welches für seine Referenz von der Referenz auf ein anderes Element abhängt)

"Jack went to Paris last year.

He loves to go there. His brother did it last year too."

- Metonymy:

- (Ein figürlicher Ausdruck der aus der Benutzung des Namens für ein Ding als Bezeichner für etwas anderes steht)

"The orchestra played well. Only the trumpet didn't know his part."

"He payed with plastic."

Vorgehen im ER

- Benötigt oft pre-processing (PDF zu text, HTML zu text, ...)
- Aufteilung des Textes and Satzgrenzen:
- $\text{text} = \langle \text{sentence1}, \text{sentence2}, \dots, \text{sentencen} \rangle$
z.B.: Trenne an Leerzeichen nach Endzeichen (.,?,!) wenn der nächste Buchstabe groß geschrieben und der nächste Buchstabe kein Punkt ist.
- Besser jedoch: Lerne die Grenzen!
- Tokenization:
 $\text{sentence} = \langle \text{token1}, \text{token2}, \dots, \text{tokenm} \rangle$
z.B.: Trenne an Leerzeichen und zwischen Wörtern mit Punctuationen

- Word stemming:

Activities, activity, activate => activ

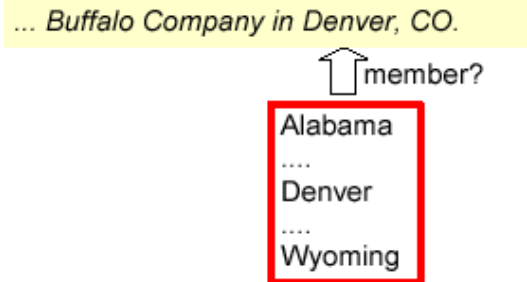
- Porter Stemming Algorithmus :

- Rule based for English:

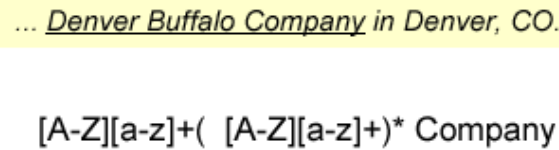
- SSES -> SS (caresses -> caress)
 - ATIONAL -> ATE (relational -> relate)
 - IZATION -> IZE (organization -> organize)
 - IZE -> (organize -> organ)

Organization und organ haben beide die Wurzel organ!

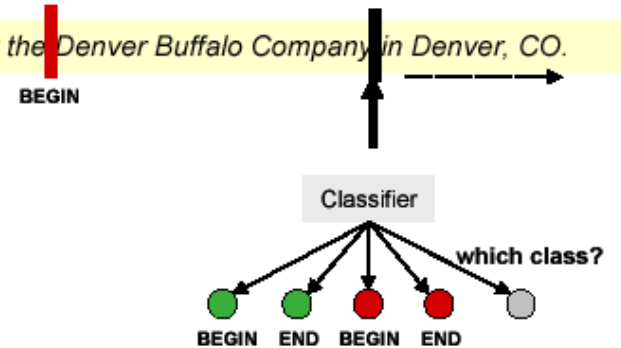
Lexicons



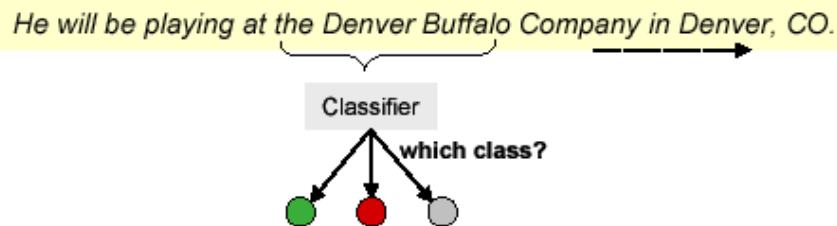
Regular Expressions



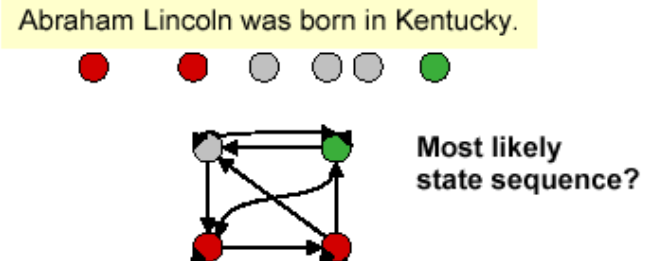
Boundary Classifier



Sliding Window Classifier

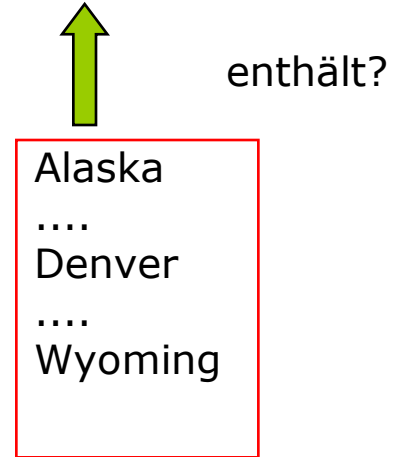


Finite State Machines



John Denver will be playing at the Denver Buffalo Company in Denver, CO.

Einfaches nachsehen, nicht Kontext sensitiv!



- **Woher kommen Lexikons?**
 - z.B.: Telefonbuch mit Personennamen
 - Gelbeseiten mit Firmennamen
 - Atlas mit Städtenamen
- **Sind Lexikons jemans komplett?**
 - - L.A.
 - - Frisco
 - - Big Apple
 - - ...
- **Vorteil: Gibt zusätzliche Auskünfte, z.B. Postcode**
- **Nachteil: Pflege des Lexikons benötigt, kein Bezug auf den Kontext**

Reguläre Ausdrücke

- Personennamen:

`J(\.|ohn|ack|im)\s(Smith|Jones|Taylor|Miller)`

`[A-Z][a-z]+([A-Z][a-z]+)*` (Ford, Philip Morris, John Hancock !!)

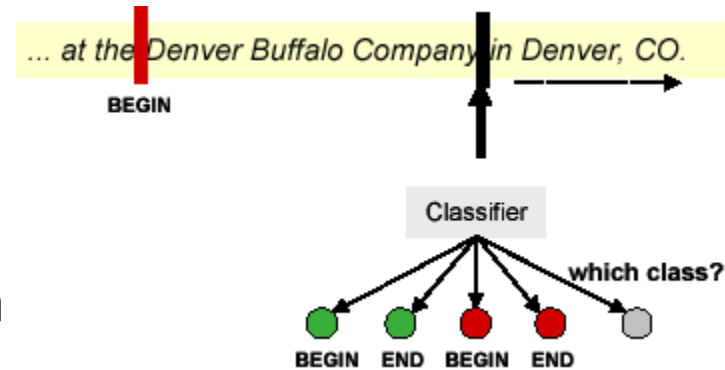
- E-mail Adressen ?

`[A-Za-z0-9](([_\.\\-]?[a-zA-Z0-9]+)*)@([A-Za-z0-9]+)(([\\\.\\-]?[a-zA-Z0-9]+)*)\.[A-Za-z]{2,}`

- Pro: Generalisierung von konkreten Beispielen
- Contra:
 - Kein Kontext-sensitives Matching
 - Erstellung und Debugging ist creation and debugging ist umständlich

Boundary Classifier

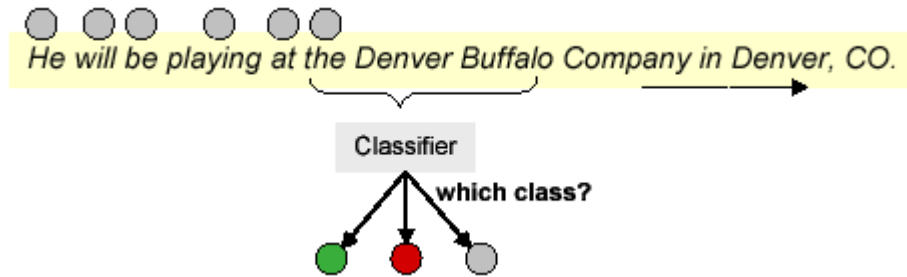
- Klassifiziere Lücken zwischen Tokens:
 - finde Begin und Ende von Entitäten im Text
 - finde Satzgrenzen
- Repräsentiere jede Lücke als Featurevektor, Features vom vorhergehenden und nexten Token (Kontext) z.B:
 - “previous or next starts with upper case” [1|0]
 - “previous is all upper case” [1|0]
 - “next is all digits” [1|0]
 - ...



- Trainingsdaten geben Lückenvektor für jede Klasse
- Gegeben eine neue Tokensequenz => klassifiziere jede Lücke, z.B. mit maximum margin classifier
- Behande Überlappungen (falls vorhanden)
- **Pro: Ausnützung von (lokalem) Kontext**
- **Contra: Benötigt Trainingsbeispiele; was sind die richtigen Features?**

Sliding Window Classifier

- Klassifiziere jedes Token durch Nutzung des lokalen Kontexts (Fenster) und der Historie (vorhergehendes Tag im Fenster)
- Fenstergröße? (z.B. 3, 5, 7)
- Durchlaufe von links nach rechts oder von rechts nach links?
- Features aus den Tokens innerhalb des Fensters
 - words, stems, POS tags
 - n-grams (3-grams for 'Denver' => Den, env, nve, ver)
 - Präfixe, Suffixe
 - Vorher zugeordnete Klassen
- Durchlaufen der Trainingsbeispiele gibt Featurevektor für jede Klasse
 - Binary classification: innerhalb von Namen vs. außerhalb von Namen
 - Multiclass classification: Anfang des Namens vs. innerhalb von Namen vs. außerhalb von Namen
- **Pro: Nutzt lokalen Kontext**
- **Contra: Benötigt Trainingsbeispiele; was sind die richtigen Features?**



- Viele weitere Methoden, z.B.
 - Maximum margin classifier
 - Multiclass classification with binary classifiers
 - Hidden Markov Models (HMMs)

- Intra-Klassen Disambiguierung

John Smith's interest in **law** was developed whilst he was serving in the **British** Royal Artillery. After leaving **military service** he read **law** at Cambridge **University** and became a **barrister** in 1950.

1. John Smith: CEO of General Motors, boardmember of Delta Airlines
2. John Smith: **British soldier**, Sailor, Leader of Virginia Colony
3. John Smith: Wrestler, Oklahoma State **University**
4. John Smith: Professor of **Law** at **University** of Nottingham, Criminal **Law** Revision Committee, **British** Academy

- Foundations of Statistical Natural Language Processing“, Manning & Schütze
- Kosala, Raymond und Blockeel, Hendrik (2000), Web Mining Research: A Survey, SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, Volume 2, Issue 1, o.O., Seite 1-10, Online: <http://www.acm.org/sigs/sigkdd/explorations/issues/2-1-2000-06/kosala.pdf>
- Mehler, Alexander und Wolff, Christian (2005), Einleitung: Perspektiven und Positionen des Text Mining, In: Zeitschrift für Computerlinguistik und Sprachtechnologie, Band 20, Heft 1, Seite 1-18, Regensburg, Deutschland.

- **Web Mining** ist schwierig da häufig keine Struktur der Daten vorhanden
- Daten beinhalten Text Dokumente, Hypertext Dokumente, Linkstrukturen, Logs
- Mining XML Dokumente (Spezielle Strukturen...)

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR View	DB View		
View of Data	- Unstructured - Semi structured	- Semi structured - Web site as DB	- Links structure	- Interactivity
Main Data	- Text documents - Hypertext documents	- Hypertext documents	- Links structure	- Server logs - Browser logs
Representation	- Bag of words, n-grams - Terms, phrases - Concepts or ontology - Relational	- Edge-labeled graph (OEM) - Relational	- Graph	- Relational table - Graph
Method	- TFIDF and variants - Machine learning - Statistical (including NLP)	- Proprietary algorithms - ILP - (Modified) association rules	- Proprietary algorithms	- Machine Learning - Statistical - (Modified) association rules
Application Categories	- Categorization - Clustering - Finding extraction rules - Finding patterns in text - User modeling	- Finding frequent sub-structures - Web site schema discovery	- Categorization - Clustering	- Site construction, adaptation, and management - Marketing - User modeling

Kosala, Raymond und Blockeel, Hendrik (2000), Web Mining Research: A Survey, SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, Volume 2, Issue 1, o.O., Seite 1-10



Fragen?