Focused (Web) Crawling

Michael Mlivoncic
mlivoncic@inf.ethz.ch

Überblick

- Traditionelles Crawling...
- Focused Crawling, insbesondere
  - Klassifizierung
  - Linkauswertung
  - Schematischer Aufbau
  - Zeit für Diskussion
  - hoffentlich noch Zeit für Mensa :-)

Generelle Funktionsweise

... Web-Crawler / Robots - Spider

- Strategie: Standard-Graphen-Algorithmen zum Erkunden, wie BFS (oder DFS)
- Ausgehend von einem oder mehreren Start-URLs ("seed set")
- Extraktion von Links, Verwaltung der Links, systemat. Abfragen & Verarbeiten der Seiten...

Daten...

- Über 1 Mrd. WWW-Seiten, 6 Mio. WebServer, (Stand Jan 2000, [NEC00])
- täglich vermutlich mehr als 1 Mio. neue Seiten...
- Heute übliche Crawler können "Abdeckung" immer weniger garantieren.
- "freshness" (mehrere Monate/Crawl...)
- selbst "inkrementelle" Crawler...
- "one-size-fits-all" Philosophie

"coverage"...

- Surfer haben verschiedene Ansprüche (allgemeinen od. speziellen Informationsbedarf)
- Für allgemeine Infos genügen die "Portale"
  - Links zu verschiedenen Info-Angeboten aus vielen alltäglichen, "breiten" Bereichen
  - Rapider Wachstum des WWW daher eher unwesentlich
- "Coverage" für spezielle, in die Tiefe gehende Themen wichtig:
  - Relevante Seiten erhalten Wichtigkeit erst, wenn sie **gut** erreichbar sind ("link count"!)

Die Alternative...

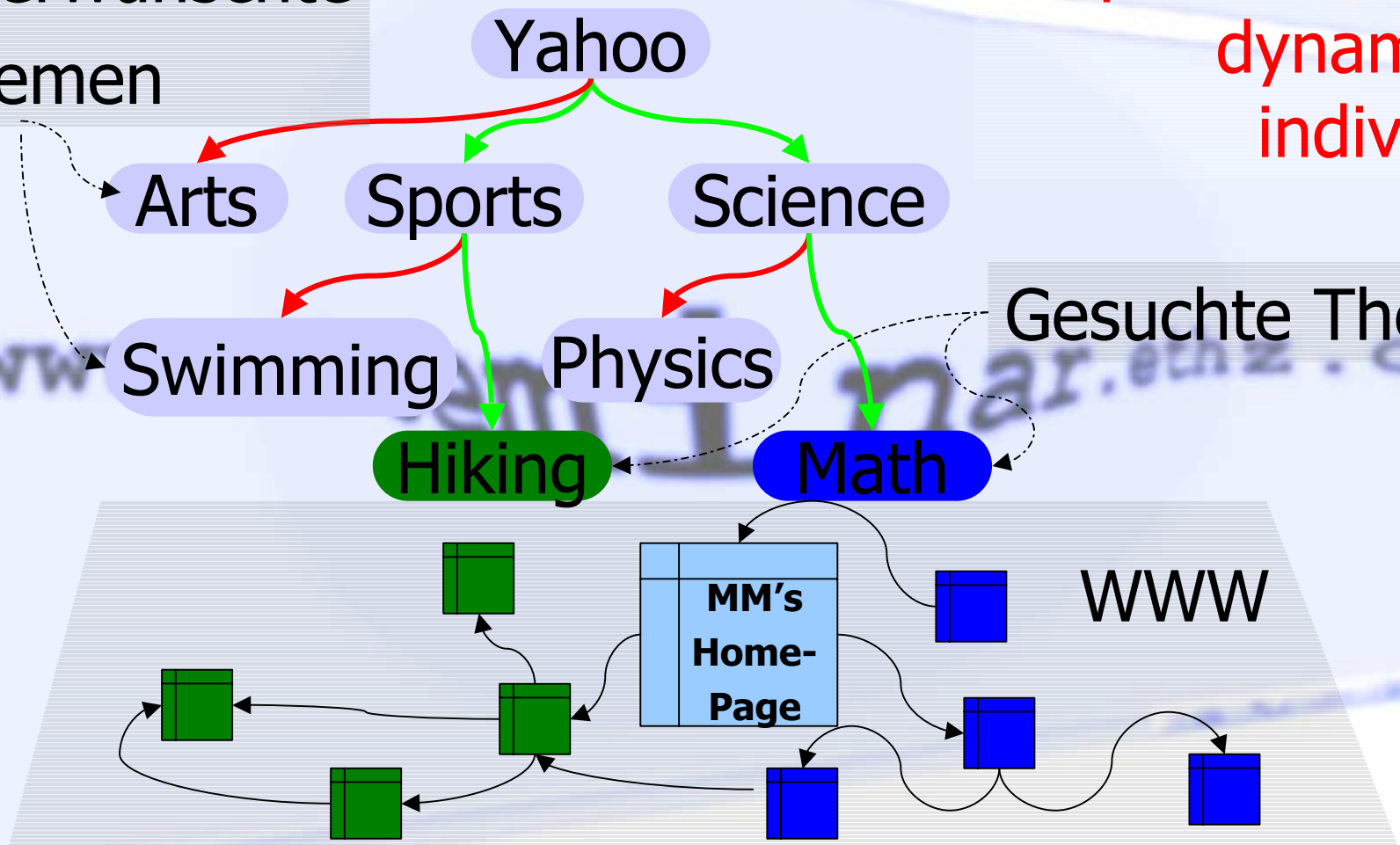
- spezialisierte Systeme oft nützlicher
- Suchmaschine speziell auf Interessengruppe zugeschnitten.
- "Resource discovery" speziell für vorgegebene Interessengebiete.
- Dabei Klassifizierung und auch Filterung nach Relevanz bereits beim Crawlen.

"Focused Crawling"

Thematische Klassifizierung

Unerwünschte Themen

"Topic taxonomy": dynamisch, individuell



Vorteile...

...gegenüber Positiv/Negativ-Beispielen:

- Negative Kategorien besser beschreibbar...
- Wiederverwendbarkeit eines (allgemeineren) Classifiers
- Entdeckung verwandter Themengebiete

Bedingungen...

- Bzgl. eines Themas "*c*" kann man jeder Seite eine Relevanz zuordnen: $R_{\{c\}}(d)$
- Beispiele für erwünschte und unerwünschte Seiten vorgegeben
- Crawler beginnt mit einem Startset D_0
- Seitenmenge D so erweitern, dass $(1/|D|)$ $\Sigma_{d \in D} R_{\{c\}}(d)$ maximiert wird
- nicht manipulierbar da globale Linkstruktur und Kategorien vorgegeben...

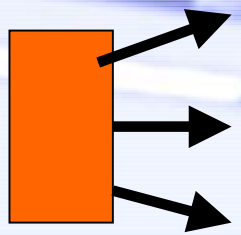
Crawling-Priorität...

- Regel 1: Eine relevante Seite verweist mit hoher Wahrscheinlichkeit auf andere relevante Seiten
  - Gilt nur für kleine Distanzen...
  - Man muss die Relevanz ständig updaten...
- Regel 2 ("co-citation"): Unbesuchte Verweise auf einer Seite mit sonst guten Verweisen sind vielversprechend
  - Wahrscheinlichkeiten... "Bayesian classifier"

"Hubs" und "Authorities"

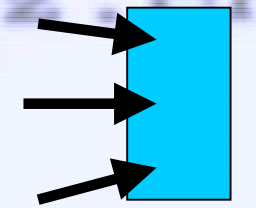
Hubs: gute Quelle für Links

- u.A. handerlesene Link-Sammlungen
- z.B. "super-hubs" wie Yahoo, "Übersichtsartikel"



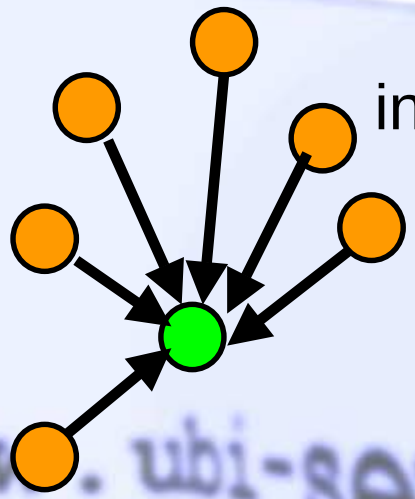
Authorities: gute inhaltl. Quelle

- vergleichbar mit oft zitierten Papers

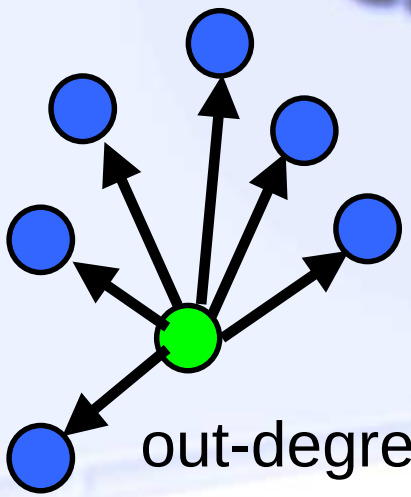


Hubs verbinden nicht nur Informationen zu verwandten Themen.

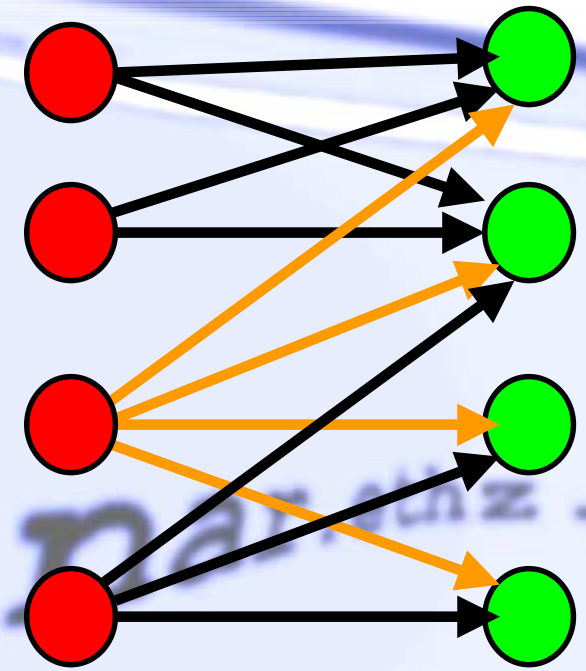
Hubs & Authorities ... (II)



in-degree



out-degree



Hubs

Authorities

..."community"

Regel 1: Relevanz abschätzen

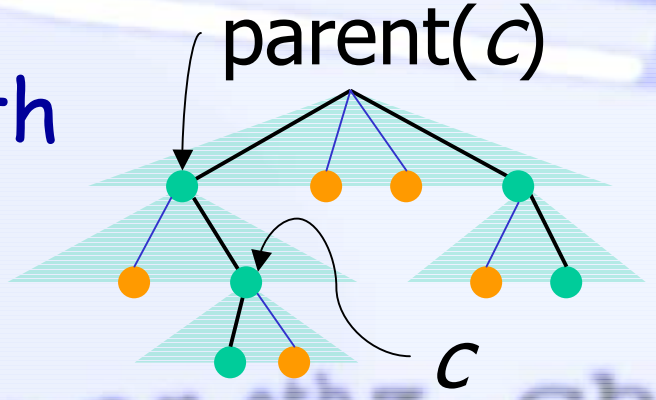
For each document "d":

For each node *c* on a path to a chosen node in topological order:

Find Pr[*c*|*d*, parent(*c*)]

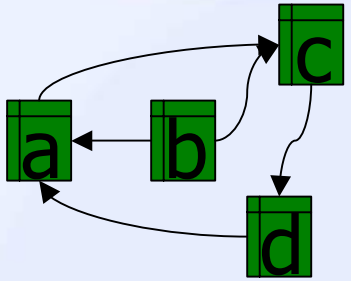
Pr[c|d]=Pr[c|d,parent(c)]Pr[d|parent(c)]

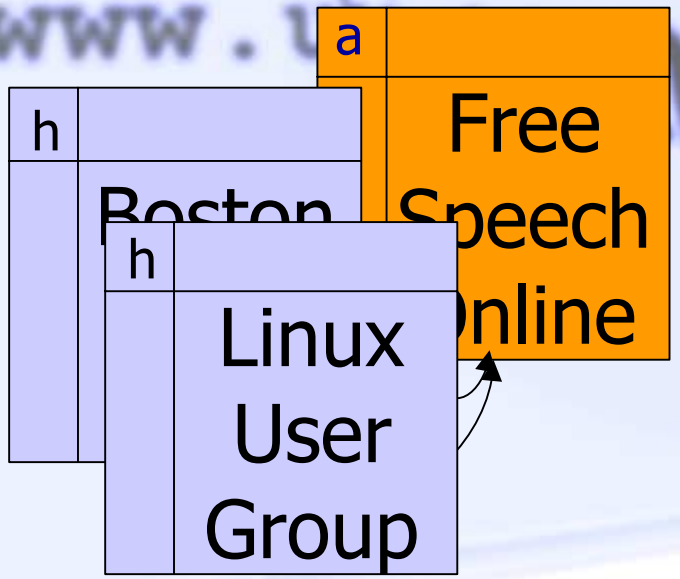
Find $R_{\{c\}}(d) = \sum_{\{c\}} Pr[c|d]$ for chosen $\{c\}$



Regel 2: Links bewerten

E	a	b	c	d
a			1	
b	1		1	
c				1
d	1			





HITS (Kleinberg 1997):

$h(u) \leftarrow 1$ for all u

For each v: $a(v) = \sum_{u \to v} h(u)$

Normalize $\sum_v a(v)$ to 1

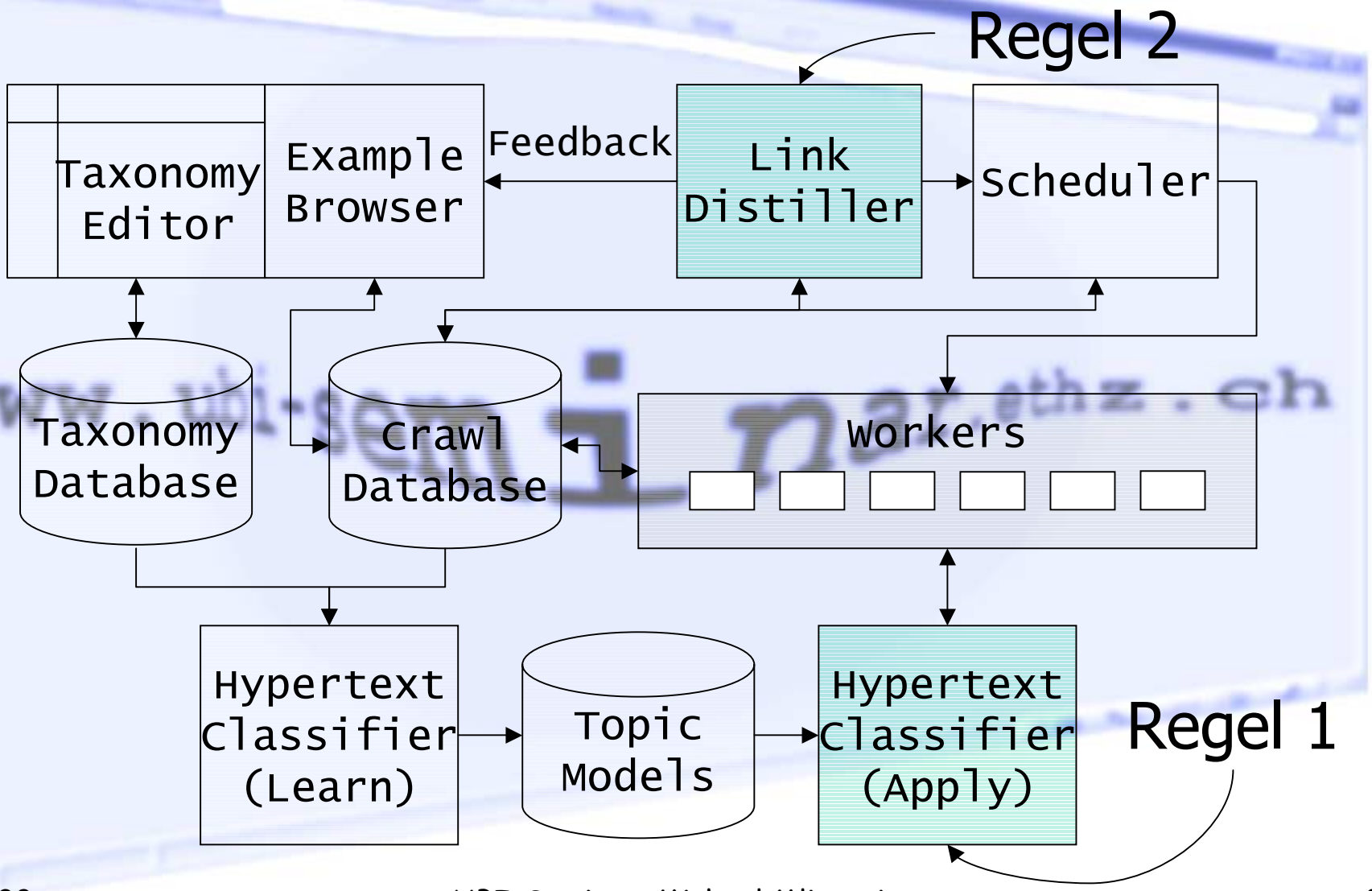
For each u: $h(u) = \sum_{u \to v} a(v)$

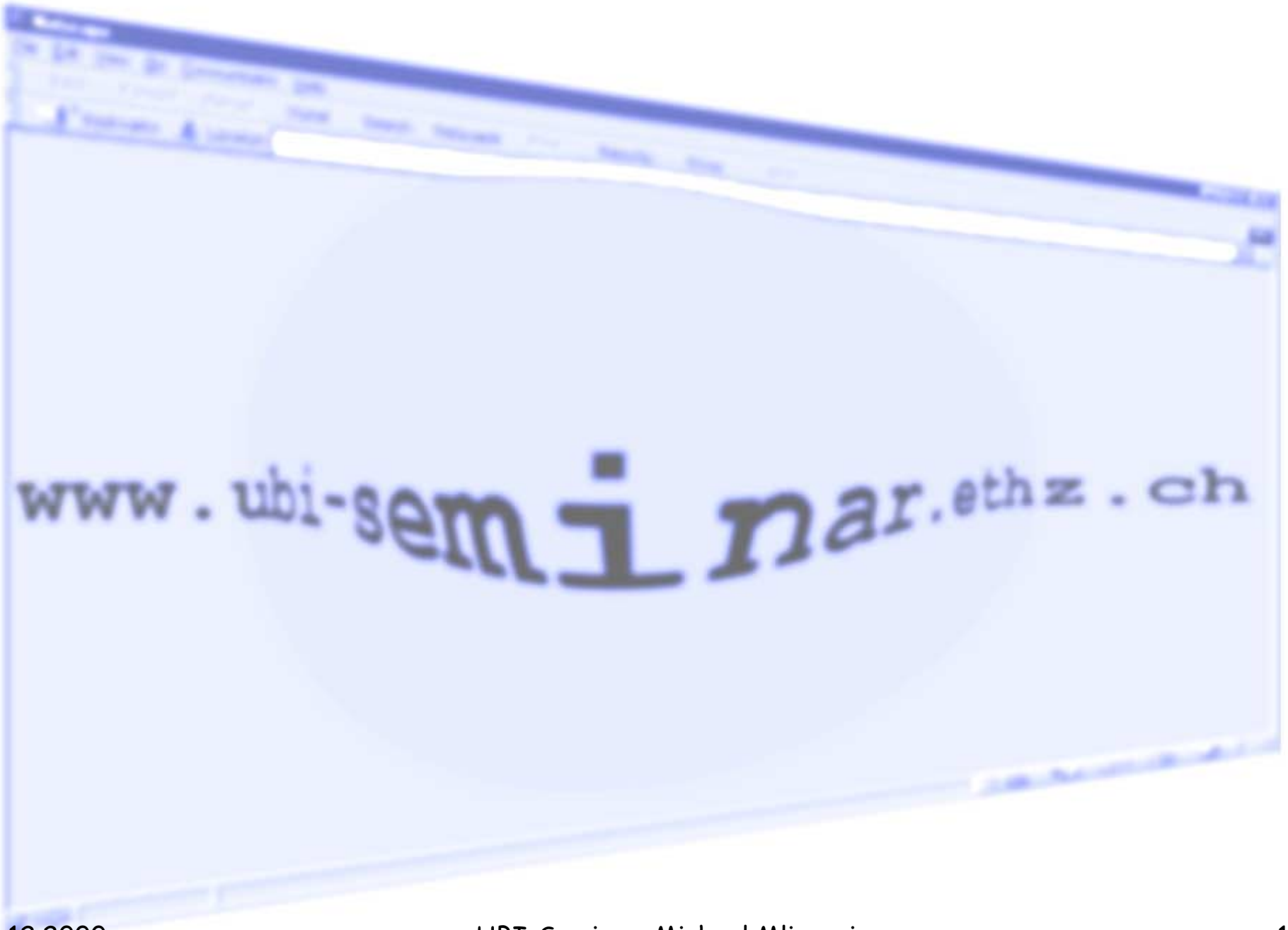
Normalize $\sum_u h(u)$ to 1

$$E_{\{c\}}[u,v] = R_{\{c\}}(v)$$

$$E^{\top}_{\{c\}}[u,v] = R_{\{c\}}(u)$$

Focused Crawler-Aufbau
(aus [Chak99])




www.ubi-seminar.ethz.ch

Referenzen

- [Chak99] Soumen Chakrabarti, Martin van den Berg, Byron Dom: Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. WWW8/ Computer Networks 31(11-16): 1623-1640 (1999) [http://www8.org/w8-papers/5a-search-query/crawling/]
- [NEC00] "Web surpasses one billion documents: Inktomi/NEC press release.", available at http://www.inktomi.com, Jan 18 2000.
- [Law00] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, Marco Gori: Focused Crawling Using Context Graphs. VLDB 2000
- [Cho00] Junghoo Cho and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. Submitted to VLDB 2000, Experience/Application track, 2000.

Referenzen (Hubs,HITS, etc.)

- [Klein97] J. Kleinberg. "*Authoritative sources in a hyperlinked environment*", Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998 (also IBM Research Report RJ 10076, May 1997).
- [Chak98] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan, *"Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text"*, Proceedings of the 7th World-Wide Web conference, 1998.
- [Chak99] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, *"Hypersearching the web"*, Scientific American, June, 1999.