Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Suche im Web

Tobias Scheffer

WWW

- 1990 am CERN von Tim Berners Lee zum besseren Zugriff auf Papers entwickelt.
  - HTTP, URLs, HTML, Webserver.
- Verbindet FTP mit der Idee von Hypertext.
- Multilingual (ca. 75% englisch, je 5% japanisch und deutsch).
- Groß, verteilt.
- Volatil: Dokumente erscheinen und verschwinden.
- Unstrukturiert, heterogen.

Entwicklung des WWW

- Hypertext: 1960er Jahre.
- ARPANET / DARPANET / Internet: 1970er.
- FTP: 80er Jahre.
  - Archie: Crawler sammelt Dateinamen von FTP-Servern und ermöglicht Regex-Suche.
  - Gopher.
- Web-Browser: Anfang der 90er von Doktoranden.
  - Mosaic: Marc Andreessen & Eric Bina, UIUC NCSA.
  - Netscape und IE aus Mosaic hervorgegangen.
- Web-Directories: Verzeichnis „Yahoo“, 1994, zwei Stanford-Doktoranden.

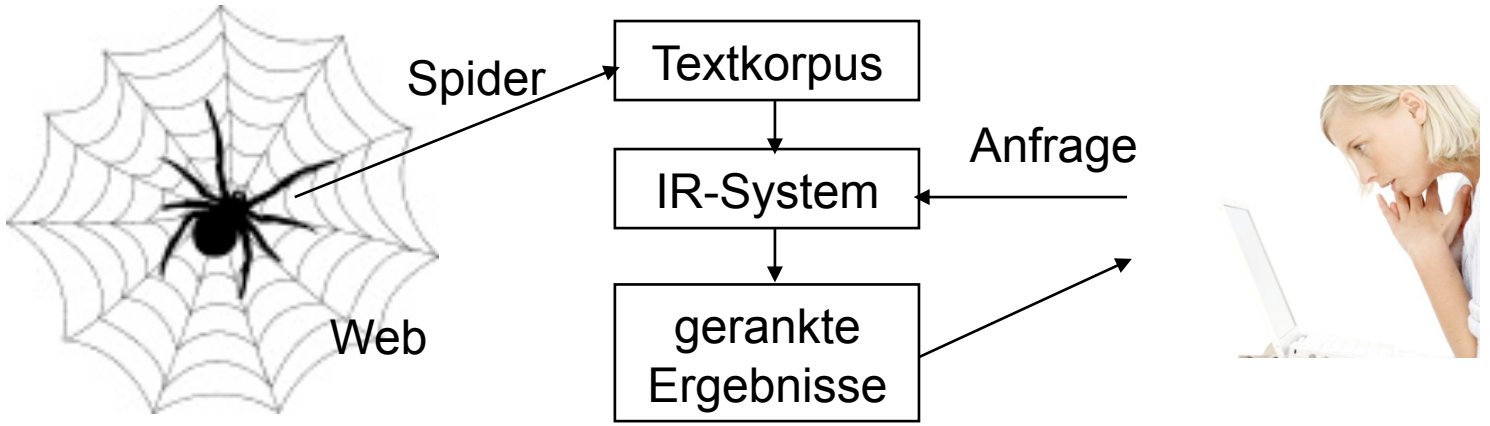
Suche im WWW

- WebCrawler: Programmier-Projekt, Uni Washington, 1994. → Excite, AOL.
- 1994: Doktorand an CMU entwickelt Lycos.
- 1995: DEC entwickelt Altavista.
- 1998: Doktoranden aus Stanford entwickeln Google. Idee: Analyse der Linkstruktur, relevantere Ergebnisse.

Graph-Struktur des WWW

- Große zentrale, stark verknüpfte Komponente.
  - Jede Seite von jeder Seite erreichbar.
- Randgebiete die nur
  - Auf die zentrale Komponente verweisen.
  - Von zentraler Komponente aus verwiesen werden.
  - Isolierte Komponenten.
- Zipfs Gesetz (spezielles „power law“):
  - Wenn n natürlich auftretende Zahlen z_i der Größe nach geordnet werden, dann ist $z_i \sim 1/i$.
  - Beschreibt im Web Länge der Seiten, Anzahl eingehender und ausgehender Kanten, Anzahl Hits etc.

Suche im Web



- Unterschied zu anderen IR-Systemen:
  - Spider / crawler sammelt Texte für Indexstruktur.
  - Analyse der Linkstruktur liefert Relevanzinformation.
  - Suchmaschinen sind stark verteilte Systeme.

Crawling

- Beginne mit URLs aus zentraler Komponente,
- Folge Links rekursiv und füge neue Seiten in Index.
- Suchstrategien
  - Breadth first, depth first,
  - Hoher PageRank first.
  - Crawlhäufigkeit in Abhängigkeit von Änderungshäufigkeit und PageRank (WallstreetJournal.com vs. HottesSockenShop.de)
  - Crawlhäufigkeit in Abhängigkeit der Zugriffshäufigkeit auf die Seite.
  - Crawling zu den richtigen Zeitpunkten bei zyklischen Updates.

Crawling

- Vor Ablegen testen, ob Seite schon abgelegt ist.
  - Web ist kein Baum sondern Graph.
  - Trie, Hash-Tabelle, schneller Zugriff auf URL.
  - Mehrere (unendlich viele) URLs für dieselben Seiten durch Zyklen in Verzeichnisstrukturen.
  - Seiten mit dynamischem Inhalt.
  - Seiteninhalt von Cookie abhängig.

Crawling

- Gecrawlte Seiten müssen parsiert werden.
  - Finden der Links, die dann wieder rekursiv verfolgt werden.
- Ankertexte werden als Indexterme für die verlinkte Seite abgelegt.
- Beschränkung des crawlbaren Bereichs durch Benutzer möglich.
  - Robots.txt, Robots META Tag

Crawling

- Crawlen muss in parallelen threads erfolgen.
  - Antwortzeit auf jeden HTTP-Request relativ hoch,
  - Ein Crawler-Prozess muss einige hundert Request-Threads verwalten.
- Aber: zuviele Anfragen an denselben Server vermeiden
  - „Denial of Service Attack“.

Benutzerschnittstellen

- HTML-Webinterface
- <form>-Umgebung ermöglicht Übergabe des Suchstrings als Parameter an Suchmaschine.
- Suchmaschine generiert Rückgabeseite,
  - CGI
  - Java Servlet.
- Session Tracking:
  - Speicherung von Präferenzen, Anfrage-Verfeinerung.
  - HTTP ist zustandsloses Protokoll, Tracking nur über Cookies oder dynamische URLs mit Session-ID.
- Meist nur kurze Anfragen (2.35 Wörter).

Benutzerschnittstellen

- Advanced Search:
  - Wird fast nie benutzt.
  - Ähnliche Seiten finden,
  - Links auf Seite finden,
  - Maschinelle Übersetzung, cross-language retrieval.
- Clusterung von Seiten

Retrieval

- Suchterme → Index → Liste von URLs.
- Texte mit mehreren Suchtermen durch Mengenoperationen.
- Aufbau eines Ähnlichkeitsvektors mit verschiedenen Merkmalen

$$\vec{x}_{q,d} = \begin{pmatrix} \varphi_{sim}(q,d) \text{ im Vektorraummodell} \\ \tau_{\#} \text{ gleicher Terme in h1-Tags} \\ \tau_{\#} \\ \tau_{\#} \\ \ldots \\ \tau_{\#} \\ \varphi_{PageRank}(d) \end{pmatrix}$$

- Rückgabe der Seiten $\arg\max_{d}(\vec{w}, \vec{x}_{q,d})$
- Manuelle Konstruktion von w.
- Lernen von w aus Klickdaten.

Ranking

- Ende der 90er wurde alle Suchmaschinen in Web-Portale mit breitem Informations- und Unterhaltungsangebot umgebaut.
- Gegenannahme von Google: nur gute Suchfunktion ist wichtig, der Rest interessiert niemanden.
- Idee von Kleinberg und Page & Brin: Verweisstruktur zeigt, welche Webseiten irgend jemanden interessieren.
- Suchergebnisse werden nach PageRank sortiert.

PageRank

- Keine Unterscheidung in Hubs und Authorities.
- Ranking der Seiten nur nach Authority.
- PageRank ist umso höher, je mehr andere Seiten auf die Seiten verweisen, und je höher deren PageRank ist.
- Wird auf das ganze Web angewandt, nicht nur auf Ergebnismenge einer Suchmaschine.

Authority-Ranking: PageRank

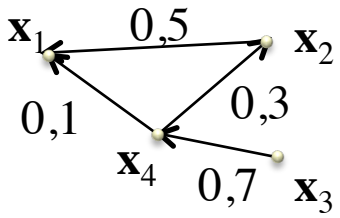
- Gegeben: Trainingsdaten $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..\}^n$ mit gegebenen lokalen *Authority Scores* (Kompetenz-Bewertungen):
  - Authority Score $A_{ij} = a(\mathbf{x}_i, \mathbf{x}_j)$ gibt an wie „kompetent“ $\mathbf{x}_j$ aus Sicht von $\mathbf{x}_i$ ist.
  - Link oder kein Link, Link-Position.

Authority Matrix

- Gesucht: Modell $f : A \in R^{n \times n} \mapsto s \in R^n$ welches für Instanzen $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..\}^n$ globale Authority Scores (Ranking) s_i liefert.
  - Annahme: Je kompetenter $\mathbf{x}_i$ und je höher der Authority Score A_{ij}, desto kompetenter $\mathbf{x}_j$.

Authority-Ranking: PageRank

- Modellierung der (nicht-symmetrischen) Kompetenz-Bewertungen als ungerichteter Graph:
  - Instanzen sind Knoten, Authority Scores sind Kanten-Gewichte
    $\Rightarrow$ Authority Matrix = Adjazenzmatrix $\mathbf{A}$.



$$\mathbf{A} = \begin{pmatrix} 0 & 0,5 & 0 & 0 \\ 0,5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0,3 \\ 0 & 1 & 0,7 & 0 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{pmatrix} \qquad d_{ii} = \sum_{j=1}^{n}$$

  - Intuition: (normierter) Graph beschreibt mit welcher Wahrscheinlichkeit Knoten $\mathbf{x}_i$ Knoten $\mathbf{x}_j$ als „Experten“ nennen würde.
  - Beispiel: $\mathbf{x}_4$ hält $\mathbf{x}_2$ für 3-mal so kompetent wie $\mathbf{x}_1$, $\mathbf{x}_3$ ist aus seiner Sicht kein Experte.

Authority-Ranking: PageRank

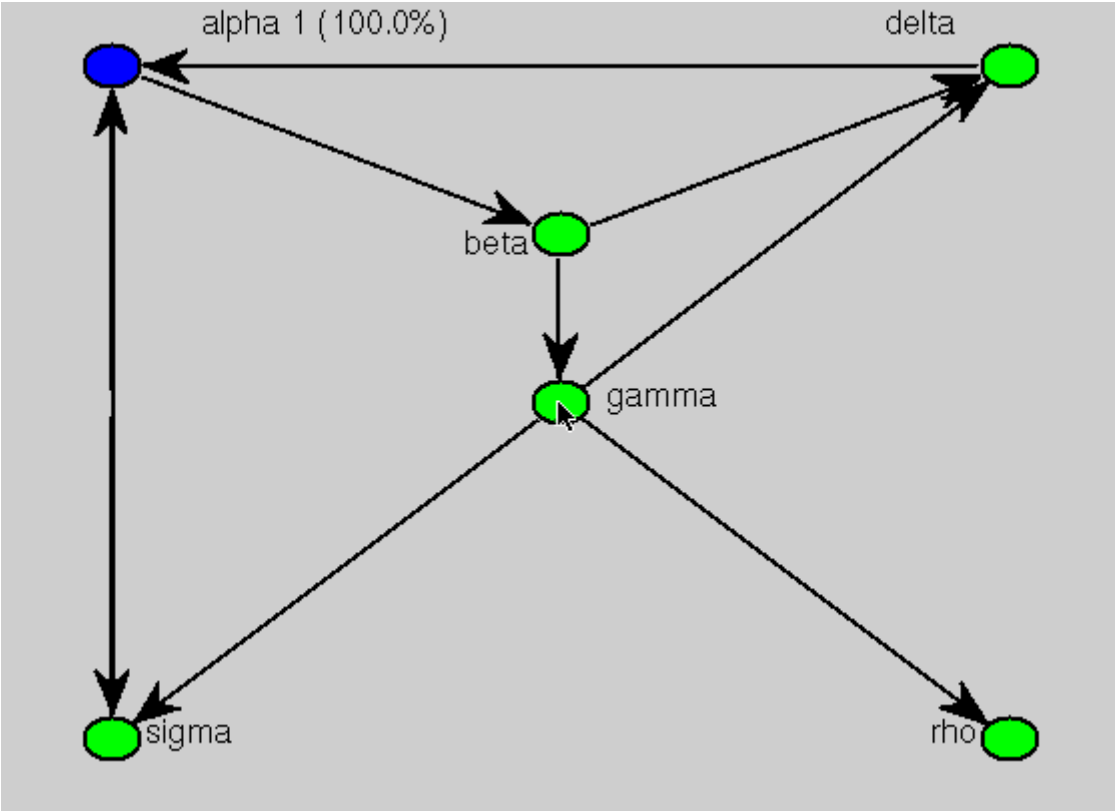
- Random Surfer:
  - Beginnend bei einem beliebigen Knoten folgt der Surfer mit Wahrscheinlichkeit 1-ε einem zufälligen Link.
  - Mit Wahrscheinlichkeit ε startet er neu an Zufallsknoten.
  - Wahrscheinlichkeit für Aufenthalt an einem Knoten = globaler Authority Score.
- Hubs & Authorities (HITS):
  - Jeder Knoten besitzt…
    - *Hub Score:* Wie gut sind seine Verweise auf „kompetente“ Knoten
      = ausgehende Kanten.
    - *Authority Score:* Wie kompetent ist der Knoten, d.h. wie viele kompetente Knoten verweisen auf ihn = eingehende Kanten.

Authority-Ranking: PageRank

- PageRank: Random Surfer Modell zum Ranking von Webseiten.
  - (Urspünglicher) Ranking-Algorithmus von Google.
  - Abhängig von Query werden relevante Webseiten gefunden und nach ihrem globalen Authority Score sortiert.
- Annahmen:
  - Link auf Webseite $\mathbf{x}_i$ verweist auf „kompetente“ Webseite $\mathbf{x}_j$, d.h. Adjazenzmatrix des Webgraphen = Authority Matrix mit $A_{ij} = 1$.
  - Mit Wahrscheinlichkeit $1 - \varepsilon$ folgt der Nutzer (Random Surfer) einem Link auf der Webseite.
  - Mit Wahrscheinlichkeit ε wechselt er auf eine zufällige Webseite.

Authority-Ranking: PageRank

- Beispiel:



Authority-Ranking: PageRank

- Gegeben: Wahrscheinlichkeit dafür, dass der Nutzer von Webseite $\mathbf{x}_i$ zu Webseite $\mathbf{x}_j$ wechselt ist

$$P_{ij} = P(\mathbf{x}_j \mid \mathbf{x}_i) \qquad \frac{A_{ij}}{\sum_{k=1}^{n} A_{ik}} \qquad \frac{1}{n}$$

Transitionswahrscheinlichkeit

und somit $P_{ij} = A_{ij}$ mit $U_{ij} = \frac{1}{n}$.

- Gesucht: Wahrscheinlichkeit dafür, dass man auf Webseite $\mathbf{x}_i$ ist, d.h. $sp_i = P(\mathbf{x}_i)$.

Aufenthaltswahrscheinlichkeit

Authority-Ranking: PageRank

- Algorithmus: Beginnend mit initialem Ranking Scores $\mathbf{s}$ iterativ neue Scores bestimmen mit

$$\mathbf{s} \leftarrow \frac{1}{c} \mathbf{P}^T \mathbf{s} \quad \text{wobei} \quad c = \left\| \mathbf{P}^T \mathbf{s} \right\| \quad \left\| \cdot \right\| = 1$$

- Konvergenz von PageRank bei $\mathbf{s} \leftarrow \mathbf{s}$, sodass gilt

$$\mathbf{s} = \frac{1}{\lambda} \mathbf{P}^T \mathbf{s}$$

  d.h. $\mathbf{s}$ ist ein Eigenvektor von $\mathbf{P}^T$ mit Eigenwert λ.
- Man kann zeigen, dass $\mathbf{s}$ der Eigenvektor mit größtem Eigenwert λ ist.

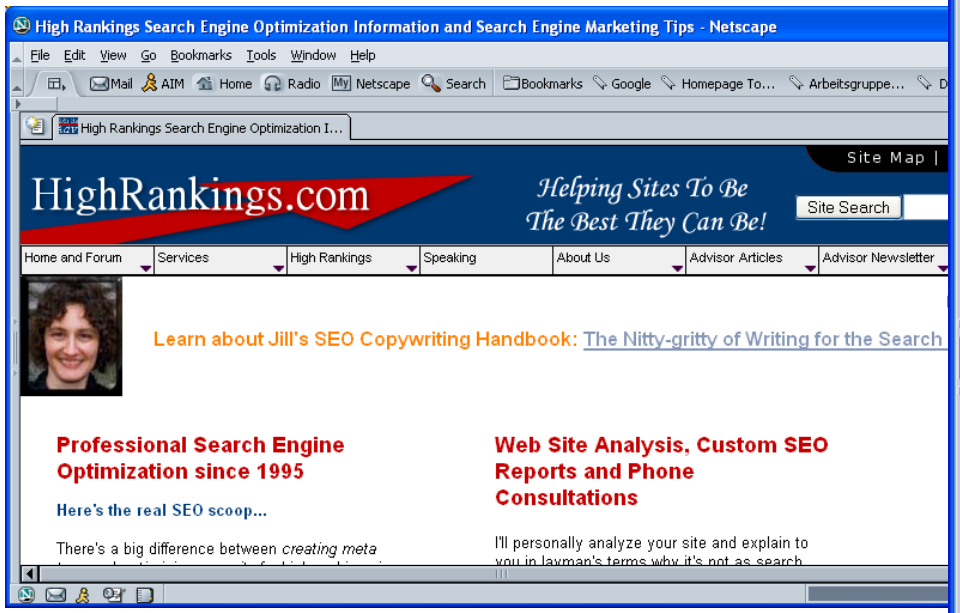
Authority-Ranking: PageRank

- Vorteile:
  - Leicht und effizient berechenbar.
  - Existenz & Eindeutigkeit der Lösung sowie Konvergenz des Algorithmus ist garantiert für $0 \leq k < 1$.
- Nachteile:
  - Links können schlechte Indikatoren für Kompetenz-Bewertung sein:
    - ★ Kompetenz zweier Instanzen kann sich gegenseitig verstärken.
    - ★ Automatisch generierte Links haben kaum Aussagekraft.
    - ★ „Künstliche“ Links (z.B. Link-Spam) verändern Ranking.
  - Eigenschaften der Instanzen fließen nicht ins Ranking ein.

PageRank: Link Spam

- PageRank ist beeinflussbar.
- Link Farmen:
  - Felder künstlich generierter Seiten, deren PageRank E(u) „geerntet“ und zu Zielseite geleitet wird.
- Guestbook Spam:
  - Generierte Einträge in Gästebüchern und Blogs, die Verweis auf Zielseite enthalten.
- Link Exchange Services:
  - Seiten mit Links auf (thematisch nicht verwandte) Seiten, meist gegen Geld.
- Partner-Programme: z.B. Amazon, Ebay.
  - Link auf Produktseite gegen Umsatzbeteiligung.

Link Spam


High Rankings Search Engine Optimization Information and Search Engine Marketing Tips - Netscape
HighRankings.com
Helping Sites To Be The Best They Can Be!
Site Search
Home and Forum
Services
High Rankings
Speaking
About Us
Advisor Articles
Advisor Newsletter
Learn about Jill's SEO Copywriting Handbook: The Nitty-gritty of Writing for the Search
Professional Search Engine Optimization since 1995
Here's the real SEO scoop...
There's a big difference between creating meta
Web Site Analysis, Custom SEO Reports and Phone Consultations
I'll personally analyze your site and explain to


Google Search: link:www.highrankings.com - Netscape
Google
Web Images Groups News Froogle Local more
link:www.highrankings.com
Search
Web Results 1 - 10 of about 780 linking to www.highrankings.com. (0.25 seconds)
Traffick - Internet Search Enlightenment: October 27, 2002
Traffick has covered the world of internet search engines and web portals since 1999 in our inimitable, irascible and thought-provoking way.
www.traffick.com/blog/archive/2002_10_27_archive.asp - 32k - Cached - Similar pages
Web Design: Wisconsin Web Design - Madison - Milwaukee - Wisconsin
Pixel Resource Network, Wisconsin, Web Design, Web Site Development, Pixel Resources, Madison, Milwaukee, Watertown, WI, web developer, web designer, ...
www.pixelresources.net/search.php - 40k - Cached - Similar pages
Online Malls
VirtualBazaar.com: Online shopping bazaar featuring collectibles, crafts, gifts, hobbies, household goods, art; AlienMall.com: An out of this world shopping ...
www.reference.com/Dir/Shopping/Online_Malls/ - 30k - Cached - Similar pages
Resources for search engine optimization and internet marketing by ...
The Bloch School of Business & Public Administration at UMKC has developed a facility that houses 19 programs that serve small businesses and ...
www.bloch.umkc.edu/4747troost/ delivery/egrc/istrategystudio/resources.asp - 43k - Cached - Similar pages
Beginner's Guide to lifetime commissions affiliate programs at ...
What are the most rewarding affiliate programs that pay commissions for life? What are residual commissions? Find out, plus other Frequently Asked Questions ...
www.lifetimecommissions.com/guide.shtml - 21k - Cached - Similar pages
www.fateback.com/news/search_engine/index26.html
Similar pages
traductores ingles-español - guia de traduccion para PYMES y ... - [
Done

Link Spam: BadRank

- Wenn PageRank beeinflussbar wird, dann verliert er seine Korrelation zur Relevanz der Seiten.
- Link Spam sollte bei der Berechnung des Page Ranks nicht so berücksichtigt werden wie „natürliche“ Links.
- Suchmaschinenbetreiber haben „Blacklists“.
  - URLs von Link-Spam-Seiten.
  - Werden manuell erstellt.

BadRank

- Invertierter PageRank-Algorithmus, „bestraft“ Seiten, die auf Link Spam verweisen.
- Initialisierung: B(u) = 1, wenn u auf Blacklist.
- Für alle Seiten:

$$B'(u) = \sum_{v \in B(u)} \frac{B(v)}{N_v}$$

$$B(u) = \frac{B'(u)}{|B'(u)|}$$

- BadRank wirkt wie negativer PageRank.

Browsing

- Web-Verzeichnisse
  - Dmoz.org
  - Yahoo, Google.
- „Taxonomien des menschlichen Wissens“.

Browsing: Webverzeichnisse

- Dmoz.org,
- Yahoo,
- Google.