PG520 - Webpageranking

Marius Kubatz

12. Oktober 2007

Webpageranking - Quellen

- The PageRank citation ranking: Bringing order to the Web; Page, Brin etal. Technical report, 1998.
- A Unified Probabilistic Framework for Web Page Scoring Systems;
  Diligenti, Gori etal. In IEEE Transactions on Knowledge and Data Engineering, 1, 2004. pages 4-16.

1. Einführung und Motivation
   - Der Webgraph - die endlichen Weiten
   - Anatomie einer Suchmaschine
   - Webpageranking
2. Algorithmen und Modelle
   - Hypertext Induced Topic Search (HITS)
   - Webpageranking mit PageRank
   - Probabilistische Interpretation

Der Webgraph - Definition

Das Web (WWW, Netz und Internet) ist ein mittels HTTP Protokoll realisiertes, virtuelles Netz aus HTML-Dokumenten, die durch Hyperlinks miteinander verbunden sind.

Webgraph - Es sei $G = (V, E)$ ein gerichteter Graph mit:

- V der Knotenmenge statischer Webseiten
- E der Kantenmenge aus Hyperlinks

Eingehende Kanten B bezeichnen wir als Backlinks oder "inedges", ausgehende Kanten F als Forward Links oder "outedges".

Der Webgraph - Struktur

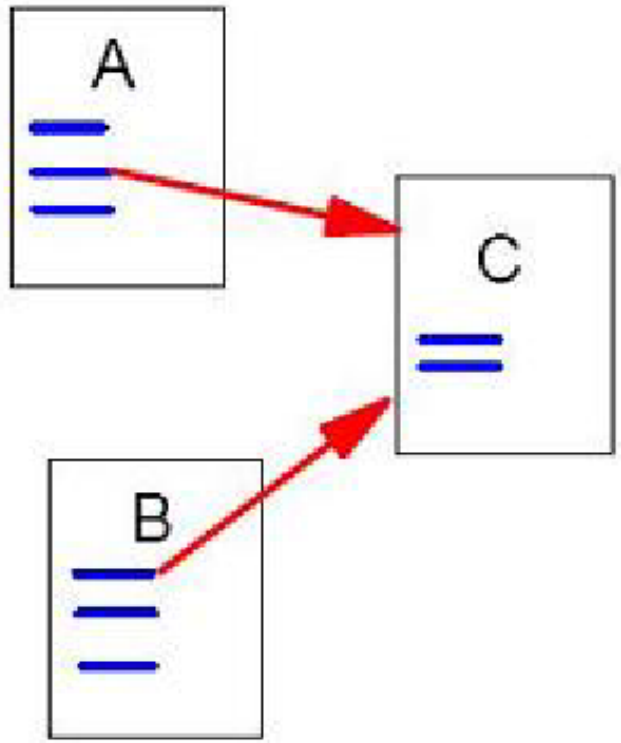


Abbildung: A und B sind Backlinks - C erreichbar durch Forward Links

Der Webgraph - In Zahlen

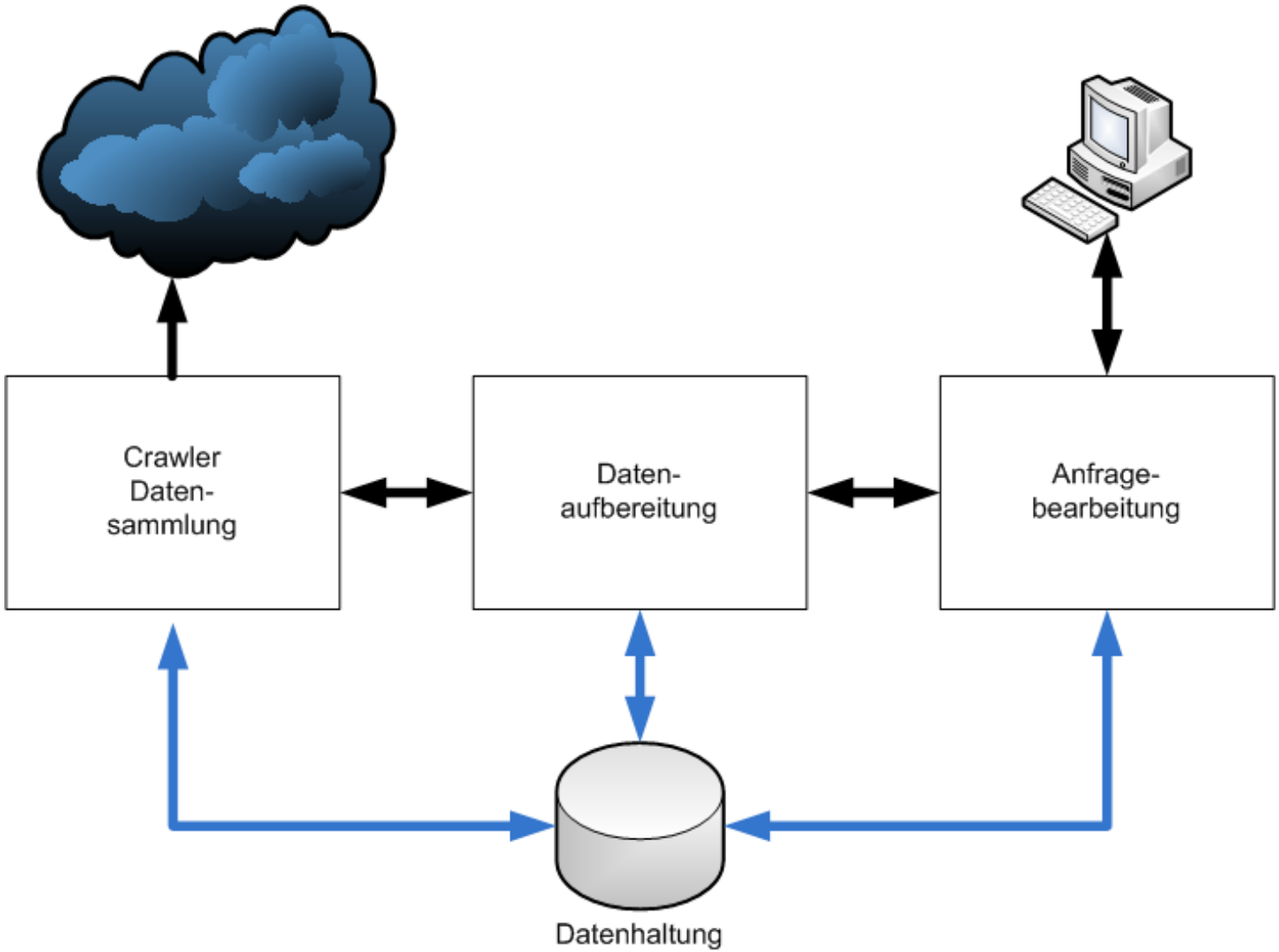
- Laut news.netcraft.com : 135 Mio. Webserver,
- mit monatlichen Zuwachs von 7.2 Mio.
- über 433 Mio. in DNS eingetragene Hosts (www.isc.org).
- Die geschätzte Anzahl indizierbarer Webseiten: 10-30 Mrd.

Anatomie einer Suchmaschine - Google

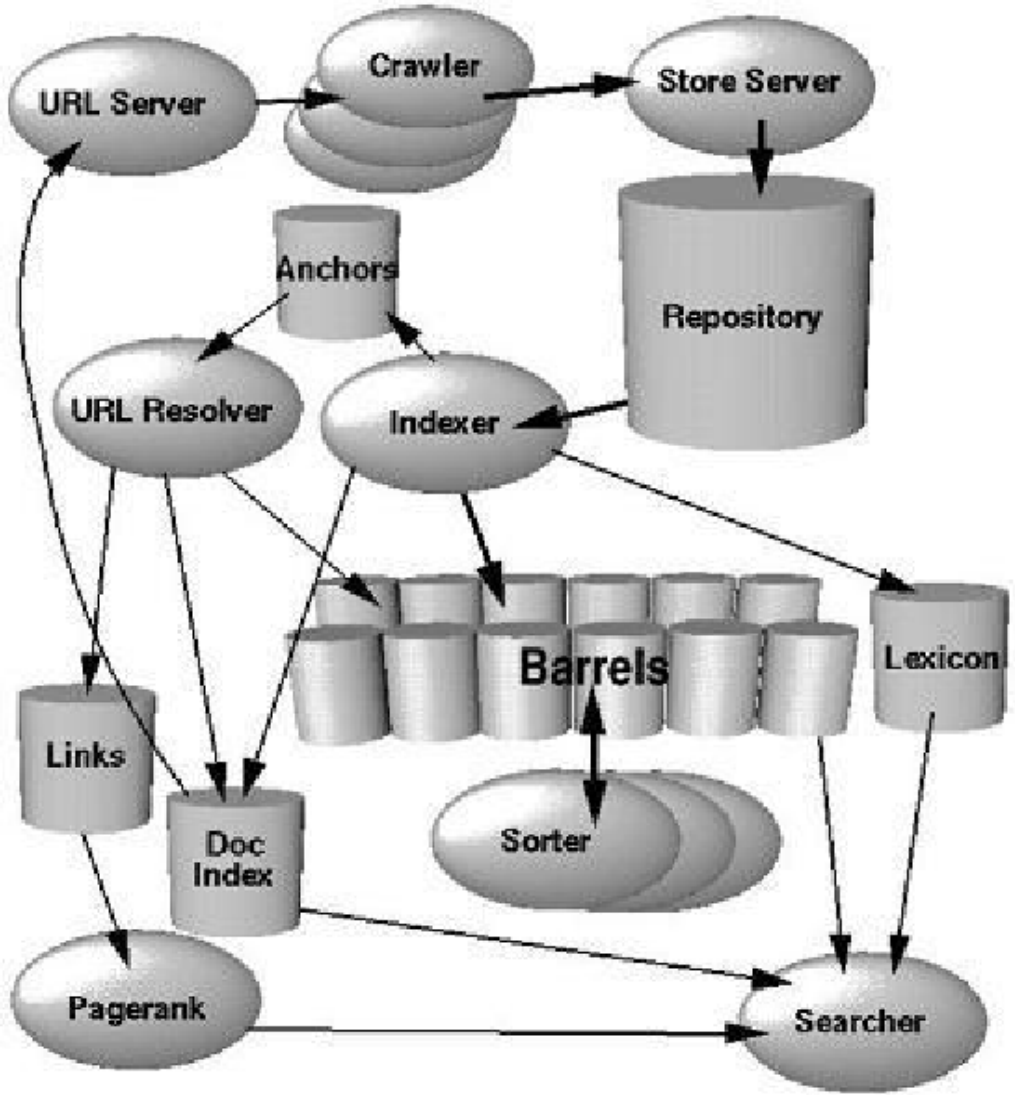
Die größte, kostenlose Suchmaschine der Welt, gegründet 1998 von Larry Page und Sergey Brin. Basiert auf zwei wichtigen Publikationen:

- Anatomy of a Large-Scale Hypertextual Web Search Engine.
- The PageRank Citation Ranking: Bringing Order to the Web.

Anatomie einer Suchmaschine - Simplified



Anatomie einer Suchmaschine - Architektur



Was ist Webpageranking?

Webpageranking - Definition

Pageranking ist das Ergebnis der Sortierung von Webseiten mit Relevanz als Sortierkriterium. Der Internetnutzer erlebt Webpageranking als eine sortierte Linkliste in der Anwort einer Suchmaschine.

Relevanz ist die geschätzte relative Wichtigkeit einer Webseite. Wir suchen einen möglichst objektiven und maschinell berechenbaren Messwert, der menschliches Interesse und Aufmerksamkeit spiegelt.

Webpageranking - klassisch

Die ersten inhaltsbasierten Suchmaschinen berechneten aus Worthäufigkeiten und Meta-Daten ein Ranking. (Alta Vista, 1995).

Intuitiv: wir zählen, vergleichen, bilden einen Wert.

- Bei 30 Mrd. Webseiten werden es viele Vergleiche sein...
- Kann man Meta-Tags vertrauen?
- Worthäufigkeiten oder bestimmte Meta-Tags sagen nichts über Relevanz aus.

Webpageranking - the second Gen

Idee: Erzeuge ein globales Ranking von Webseiten durch Analyse der Verbindungsstruktur des Netzes.

Wir betrachten nur die Topologie des Webgraphen und ignorieren die Inhalte und den semantischen Kontext einer Webseite. Die Verweise auf andere Webseiten sind unsere einzigen Informationsträger.

Webpageranking - der Clou

Annahme: Wichtige Seiten werden besonders oft von anderen Seiten verlinkt.

Wir hoffen ein vom Wahrheitsgehalt oder Qualität der Informationen, unabhängiges Kriterium gefunden zu haben.
Es zeigt die Bereitschaft eines Webautors eine andere Quelle in den eigenen Inhalten zu zitieren

- daher der Name: Citation Ranking.

Algorithmen und Modelle

Hypertext Induced Topic Search (HITS)

Autor: Jon Kleinberg (1999)

Story: Ein Citation Ranking Algorithmus der zu Forschungszwecken genutzt wird und soll angeblich bei Ask.com laufen.

Ziel: Findet Autoritäten, also die wichtigen Informationsträger im Web.

Dieser Algorithmus fokussiert auf Webgemeinden in denen Autoritäten (authorities) also wichtige Webseiten durch Verteiler (hubs) miteinader verlinkt werden.

HITS - Preprocesing

Bestimme die Knotenmenge des zu betrachtenden Teilgraphen.

- Bestimme S_q als die Menge der besten t Suchergebnisse einer Inhaltsbasierten Suchmaschine zur Query q,
- Die Autoritäten und Verteiler sollen auf der zunächst leeren Menge $T_q = \emptyset$ berechnet werden.
- Für jeden Knoten $u \in S_q$ füge alle seine Nachfolger und eine durch d beschränkte zufällige Anzahl seiner Vorgänger in die Menge T_q

HITS - Berechnung

Berechne Autoritäten und Verteiler auf T_q

- Bestimme Gewicht a_u von u als Autorität, als Summe aller Verteiler die auf u verweisen
  $a_u = \sum_{(v,u) \in E} h_v$
- Bestimme Gewicht h_u von u als Verteiler, als Summe aller Autoritäten auf die u verweist
  $h_u = \sum_{(u,v) \in E} a_v$

HITS - Summary

- Der Algorithmus wird erst bei der Anfrage ausgeführt.
- Berechnet zwei Werte pro Webseite.
- Betrachtet nur Teilgraphen. Tendiert dazu sogenannte "Tightly Knit Communities" zu bevorzugen und degradiert umliegende Webseiten.
- Abhängig von den Inhalten einer Webseite: läuft auf Ergebnissen einer Suchanfrage.
- Gut geeignet zum entdecken und zerpflücken von Webcommunities.

PageRank

Autoren: Page, Brin, Motwani, Winograd (1998)

Story: Ein Citation Ranking Algorithmus, entstanden an der Stanford University und wichtiger Bestandteil der Google Suchmaschine.

Ziel: Finde das Maß für absolute Wichtigkeit aller Webseiten, unabhängig von der Anfrage.

PageRank - Definition

PageRank Sei u eine Webseite und

- F_u die Menge aller Forward Links von u,
- B_u die Menge aller Backlinks von u,
- $N_u = |F_u|$ die Anzahl aller Nachfolger von u
- d soll Damping Faktor zur Normierung des Ranks sein.

R ist eine Rankingfunktion von u und ihrer Vorgängerin v, der Form:

$$R(u) = d \sum_{v \in B_u} \frac{R(v)}{N_v} + dE(u)$$

PageRank - Vereinfacht

Betrachtet die Backlinks einer Webseite, allerdings werden nicht alle eingehenden Links gleich bewertet! Der Rank einer Seite ist die Rank-Summe ihrer Backlinks.

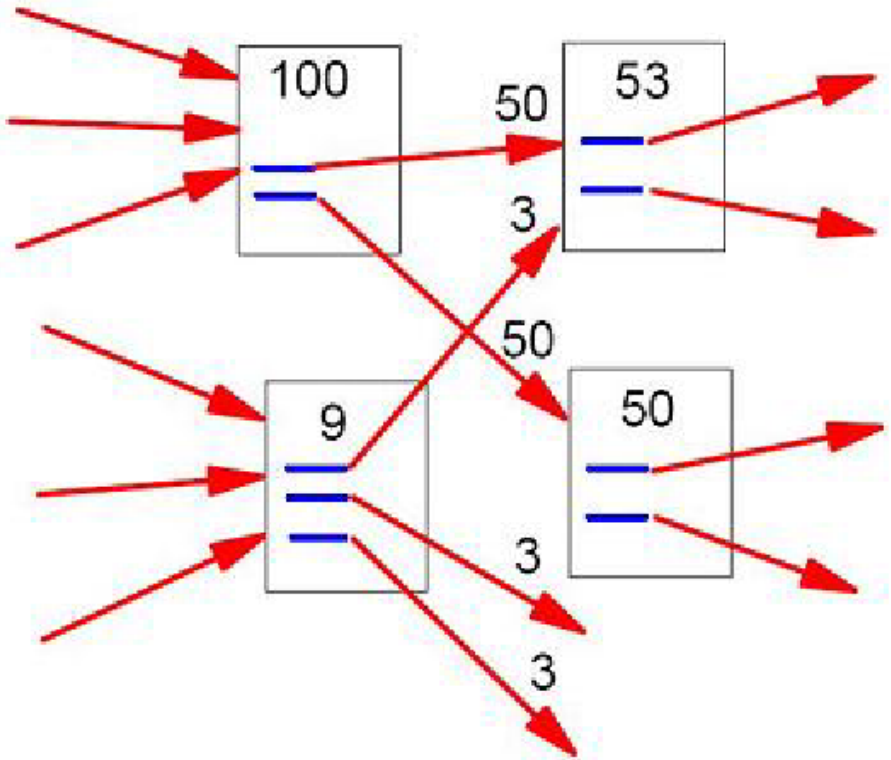


Abbildung: Eine Iteration der vereinfachten Rankingfunktion

PageRank - Was wäre wenn ...

Beobachtung: Ohne den Damping Faktor und der Initialbelegung würde der Rank bei jeder Iteration zu den Senken fliessen. Also zu Seiten

- die auf keine anderen Seiten verlinken (Dangling Links)
- die auf dem Webgraph Kreise bilden (Rank Sinks)

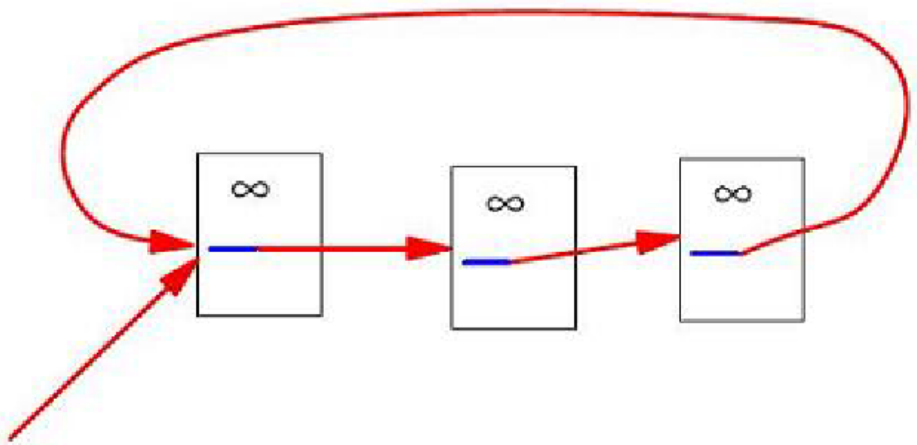


Abbildung: Ein Kreis im Webgraph: Rank Sink

PageRank - Der Algorithmus

Betrachte die Adjazenzmatrix A des Webgraphen mit:

- $a_{v,u} = \frac{1}{N_v}$ falls eine Kante von v nach u existiert.
- $a_{v,u} = 0$ sonst.

Algorithmus Starte in $R_0 = E$ der Initialbelegung und laufe durch die folgende Schleife bis R_j konvergiert.

- $R_{i+1} \leftarrow AR_i$
- $d \leftarrow ||R_i||_1 - ||R_{i+1}||_1$
- $R_{i+1} \leftarrow R_{i+1} + dE$

PageRank - Summary

- Unabhängig von den Inhalten einer Webseite.
- Algorithmus konvergiert schnell und liefert selbst nach wenigen Iterationen gute Ergebnisse.
- Gute Laufzeit 45 Iterationen bei 161 Mio. Links und nur 52,5 Iterationen bei 322 Mio. Links.
- Wird bei der Datenaufbereitung ausgeführt.
- Manipulationssicher??

Probabilistische Interpretation

Random Surfer und Web Page Scoring Systems

Dilligenti, Gori etal. (2004) modellieren mit Hilfe des Random Surfer Modells die beiden klassischen Algorithmen, sowie deren themen-orientierte Varianten in einem Framework aus Scoring Systemen. Sie unterscheiden zwischen:

- Horizontal Scoring mit Algorithmen die sich nur der Topologie des Webgraphen bedienen.
- Vertical Scoring für fokussierte Algorithmen, die nur Seiten betrachten deren Thema der Anfrage entspricht.

Zusätzlich definieren sie zwei Modelle des Random Surfers:

- den Single Surfer Walk (PageRank),
- den Multisurfer Walk, mehrerer kooperierender Surfer (HITS).

Random Surfer Modell - Aktionen

Webpageranking-Algorithmen lassen sich auch als Verfahren zur Abbildung von Benutzer-Verhalten interpretieren.

Wir modellieren die atomaren Aktionen eines generischen Websurfers als $O = (j, l, b, s)$ mit

- j - (jump) als Sprung zu einer beliebigen Webseite,
- l - (link) als einen Schritt entlang der Forward Links,
- b - (back) als einen Rückschritt zu einer Vorgänger Seite mit einem Back Link
- s - (stay) zum Verweilen auf der aktuellen Seite

Random Surfer Modell - Wahrscheinlichkeiten

Das Verhalten des Surfers hängt von der Webseite ab, auf der sich der Surfer gerade befinden, seine Aktionen werden als bedingte Wahrscheinlichkeiten für jede Aktion $o \in O$ modelliert:

- $x(j \mid q)$ - Sprung aus der Seite q,
- $x(l \mid q)$ - ein Schritt entlang der Forward Links von q,
- $x(b \mid q)$ - Rückschritt zu einer Vorgänger Seite von q
- $x(s \mid q)$ - Verweilen auf der aktuellen Seite q

$$\sum_{o \in O} = 1$$

Random Surfer Modell

Random Walk ist ein durch Iteration eines Algorithmus beschriebener Prozess, also eine Reihe von Aktionen eines Random Surfers der an einer beliebigen Seite q startet und zufällig gleichverteilt eine Aktion wählt die ihn zu einer Seite p führt.

Zu jeder atomaren Aktion gibt es eine Wahrscheinlichkeit x für die Bewegung zwischen zwei Webseiten in der Form:

- $x(p \mid q, o)$ - Der Surfer bewegt sich mit der Aktion o von q nach p

Random Walks - Iteration durch die Zeit

Sei $x(t)$ sei die Wahrscheinlichkeitsverteilung für alle Seiten und $x_q(t)$ die Wahrscheinlichkeit
mit der sich der Random Surfer auf Seite q zum Zeitpunkt t befindet.

Die Wahrscheinlichkeit $x_p(t)$ zum Zeitpunkt t wird bei jedem Schritt berechnet. Dies kann man bezüglich der in Form einer rekursiven Gleichung darstellen:

$$x_p(t+1) = \sum_{q \in G} x(p \mid q) * x_q(t)$$

Diese Gleichung kann man auch für jede einzelne Aktion aufsplittern.

Random Walks - Definition der zugehörigen Matrizen

Neben der Adjazenzmatrix W des Webgraphen G definieren wir für jede Aktion eine $(V \times V)$ Matrix:

- Δ - für Bewegungen entlang der Forward Links
- Γ - für die Backlinks
- Σ - für Sprünge

Die jeweiligen Einträge an Position $a_{p,q}$ mit Werten aus $[0 \ldots 1]$ ergeben die Wahrscheinlichkeiten für den jeweiligen Übergang $x(p \mid q, o)$.
Der Eigenvektor dieser Matrizen darf nicht höher sein als 1.

Random Walks - Darstellung als Matrizen

Den Verlauf des Random Walks sammeln wir in diagonalen Matrizen D_o in denen an Stelle $a_{q,q}$ die Wahrscheinlichkeit $x(o \mid q)$ einer Aktion auf der Seite q eingetragen wird.
Diese Matrizen fassen wir zur einer Transitionsmatrix T zusammen:

$$T = (\Sigma * D_j + \Delta * D_l + \Gamma * D_b + D_s)$$

diese kann unter berücksichtigng der Zeit auch $x(t+1) = T + x(t)$ beschrieben werden. (T' ist die zustandsübergangs Matrix der Markov Kette)

PageRank - Single Surfer Walk

PageRank wird als ein einzelner Single Surfer Walk modelliert der:

- keine Backlinks nimmt $x(b \mid p) = 0$
- nie stehen bleibt $x(s \mid p) = 0$
- $x(l \mid p) = d$ folgt einem Forward Link
- $x(j \mid p) = 1 - d$ springt zu einer neuen Seite

PageRank - Single Surfer Walk - Gleichung

Die rekursive Gleichung für den modellierten PR Random Walk ist

$$x_p(t+1) = \frac{(1-d)}{|G|} \sum_{q \in G} x_q(t) + d \sum_{q \in B_p} \frac{1}{N_q} * x_q(t)$$

Mit G als Graph, B_p als Backlinks von p und
N_q als Nachfolger von q

Abschliessend:
Der PageRank Single Surfer Walk muss speziell modifiziert werden um Dangling Links und Rank Sinks zu vermeiden.

Summary Webpageranking

- Webgraph und seine Topologie als Informationsquelle.
- Suchmaschinen die Webpageranking als globales Kriterium bei der Sortierung der Ergebnisse nutzen.
- Den PageRank Algorithmus der die Relevanz berechnet.
- Schließlich: eine alternative, stochastische Interpretation, welche das Nutzerverhalten modelliert. Der Single Surfer Walk simuliert einen menschlichen Surfer der auf wunderbare Weise immer die "relevanten" Webseiten besucht.

Danke

Das wars ...

Danke für Eure Aufmerksamkeit.