

**Hauptseminar Inside Google - Algorithmen für Suchmaschinen**

**Ausarbeitung von Igor Neyman**

**SALSA: The Stochastic Approach for Link-Structure Analysis**

13. November 2006

Betreuer: Stefan Schwoon

Wintersemester 06/07, Universität Stuttgart

## Inhaltverzeichnis

<b>1. Einführung und Problemstellung .....</b>	<b>3</b>
<b>2. Kleinbergs „Mutual Reinforcement“-Ansatz .....</b>	<b>3</b>
2.1 HITS.....	4
<b>3. PageRank.....</b>	<b>6</b>
<b>4. SALSA.....</b>	<b>7</b>
4.1 Meta-Algorithmus.....	7
4.2 Formelle Definition von SALSA.....	8
4.3 SALSA und der Ein- und Ausgangsgrad von Seiten.....	9
4.4 Der Tightly-Knit-Community-Effekt (TKC- Effekt) .....	11
4.5 Vor- und Nachteile von SALSA.....	13
<b>5. Schlussfolgerung .....</b>	<b>13</b>
<b>6. Literaturverzeichnis .....</b>	<b>13</b>

# 1. Einführung und Problemstellung

Heutzutage besteht das Web aus vielen unsortierten Seiten, die auf irgendeine Weise mittels einer Suchmaschine durchsucht werden. Dabei kann man zwischen „*narrow-topic*“- und „*broad-topic*“-Suchanfragen unterscheiden. Unter „*narrow-topic*“ versteht man ein Thema, für das wenig Stoff vorhanden ist. Das können zum Beispiel einige Zeilen aus irgendeinem Lied sein.

Unter dem zweitgenannten Begriff versteht man eine Suche zu einem oft vorkommenden Thema. Es kann geschehen, dass man hunderttausende Seiten als Ergebnis bekommt. Für den Benutzer wäre es unpraktisch (unmöglich) alles durchzustöbern, um ein paar passende Seiten zu finden.

Die klassischen Probleme für eine Suchmaschine sind:

- „Synonymie“ (für eine Suchanfrage „Wagen“ müssen auch Seiten mit dem Begriff „Auto“ ausgegeben werden),
- „Polysemie“ (wenn die Suchanfrage „Jordan“ lautet, muss man die Seiten ausgeben, die den Fluss „Jordan“ oder „Michael Jordan“ enthalten),
- sowie „Authorship styles“ (für manche Begriffe existieren viele unterschiedliche Schreibweisen, z.B. in Deutsch-Deutschland (Januar) und Deutsch-Österreich (Jänner))[1].

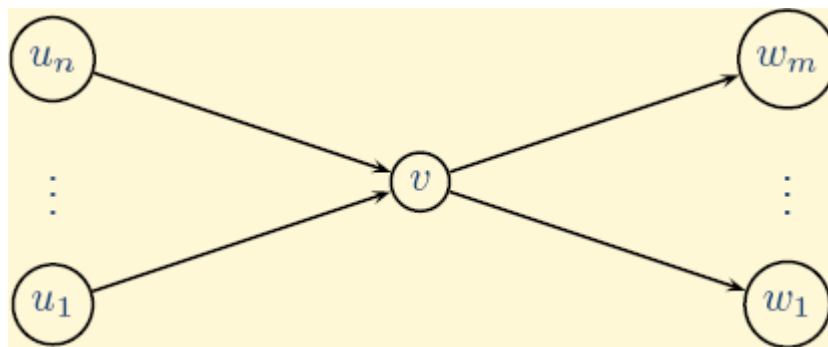
Die Aufgabe von den Suchmaschinen besteht darin, die Ergebnisse möglichst so zu sortieren, dass auf den ersten paar Seiten die „besten“ Links stehen. Außerdem soll die Wartezeit sehr kurz sein.

Im Folgenden werden zwei Verfahren kurz präsentiert, auf denen SALSA basiert.

Genaue Information für diese Verfahren können Sie in entsprechenden Seminararbeiten nachlesen. Im 4. Kapitel steht die formelle Definition von SALSA.

## 2. Kleinbergs „Mutual Reinforcement“-Ansatz

Um die obengenannten Aufgaben zu lösen, entwickelte Jon Kleinberg ein Suchverfahren „*Mutual Reinforcement Approach*“, in dem er die Webseiten in zwei Typen („Authority“ und „Hub“) unterteilte. Auf diese Weise konnte er die Seiten bewerten. Dabei sind „Authorities“ solche Seiten, die Informationen enthalten, und „Hubs“ diejenigen, die auf diese Seiten verweisen. Dabei kann jede Seite sowohl eine „Authority“ als auch ein „Hub“ sein.



Aus einer riesigen Knotenmenge (Web) wird für eine Suchanfrage  $t$  eine „kleine“ Knotenmenge herausgezogen. Der Knoten  $v$  enthält benötigte Information. Die Knoten  $u_1, \dots, u_n$  gehören zur Hub-Menge.

$$a(v) = \sum_{i=1}^n h(u_i)$$

$$h(v) = \sum_{i=1}^m a(w_i)$$

- Knoten (Webseite) –  $v$
- Authority -  $a(v)$
- Hub -  $h(v)$

Wie man leicht aus den beiden Formeln erkennt, zeigt ein „guter Hub“ (höhere Bewertungszahl) auf viele „Authorities“, und dementsprechend wird auf eine „gute Authority“ von vielen „Hubs“ verwiesen. Die Bewertungen hängen voneinander stark ab.

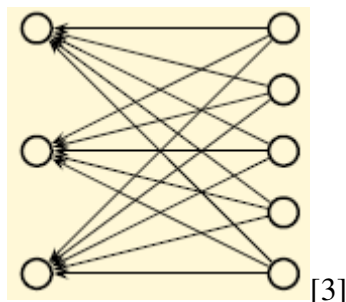
## 2.1 HITS

Der HITS-Algorithmus (Hypertext Induced Topic Search) ist die Implementierung des „Mutual Reinforcement Approach“.

### Der Tightly-Knit-Community-Effekt

HITS bevorzugt *dicht vernetzte* Gruppen von Seiten.

Lempel und Moran (2001): HITS hat eine Tendenz dazu, kleine vollständige bipartite Graphen hoch zu bewerten.



Experimente bestätigen diese Tendenz für Gruppen von Seiten im Web, die nicht bipartit sind.

### Konsequenz: Spamming

Die Rangordnung kann durch eingehende und ausgehende Links leicht verändert werden.

### Topic Drift

Es kann passieren, dass zu einer Suchanfrage wird die Seite in die Seitensammlung aufgenommen, die in einem anderen Gebiet eine hohe Rangordnung besitzt. So kann die Seite die Authority- und Hub-Gewichte beeinflussen, was dazu führen kann, dass man in ein fremdes Gebiet abdriftet.

### Beispiele:

- „Java“ driftet zu EarthWeb ab (Lempel & Moran, 2000) [Tabelle 1](#).
- „movies“ driftet zu go.msn.com ab (L&M) [Tabelle 2&3](#).
- „genetics“ bevorzugte Genetic Algorithms (L&M) [Tabelle 4](#).

Die Bedeutung von Spalten in den Tabellen:

- (1) Die URL
- (2) Der Titel der URL
- (3) Der Wert der Koordinate von der URL im (normierten) Haupteigenvektor der Authority-Matrix

Anfrage zu einem bestimmten Begriff „Java“

URL	Title	Weight
<a href="http://www.jars.com/">http://www.jars.com/</a>	EarthWeb's JARS.COM Java Review Service	0.3341
<a href="http://www.gamelan.com/">http://www.gamelan.com/</a>	Gamelan - The Official Java Directory	0.3036
<a href="http://www.javascripts.com/">http://www.javascripts.com/</a>	Javascripts.com - Welcome	0.2553
<a href="http://www.datamation.com/">http://www.datamation.com/</a>	EarthWeb's Datamation.com	0.2514
<a href="http://www.roadcoders.com/">http://www.roadcoders.com/</a>	Handheld Software Development@RoadCoders	0.2508
<a href="http://www.earthweb.com/">http://www.earthweb.com/</a>	EarthWeb	0.2494
<a href="http://www.earthwebdirect.com/">http://www.earthwebdirect.com/</a>	Welcome to Earthweb Direct	0.2475
<a href="http://www.itknowledge.com/">http://www.itknowledge.com/</a>	ITKnowledge	0.2469
<a href="http://www.intranetjournal.com/">http://www.intranetjournal.com/</a>	intranetjournal.com	0.2452
<a href="http://www.javagoodies.com/">http://www.javagoodies.com/</a>	Java Goodies JavaScript Repository	0.2388

[Tabelle 1](#). Authorities für Suchanfrage „Java“ [2]

- Die ersten 10 Authorities sind Teile von EARTHWEB Inc. Network ([Tabelle 1](#)).
- Sie sind untereinander verbunden (interconnected), aber weil sie unterschiedliche Domainnamen haben, wurden sie nicht ausgefiltert.
- Einige der Seiten sind für die Abfrage hoch relevant, aber erscheinen hauptsächlich in dem Hauptfeld nur wegen ihrer EarthWeb-Verbindung.

Anfrage zu einem bestimmten Begriff „movies“

URL	Title	Weight
<a href="http://go.msn.com/npl/msnt.asp">http://go.msn.com/npl/msnt.asp</a>	MSN.COM	0.1673
<a href="http://go.msn.com/bql/whitepages.asp">http://go.msn.com/bql/whitepages.asp</a>	White Pages - msn.com	0.1672
<a href="http://go.msn.com/bsl/webevents.asp">http://go.msn.com/bsl/webevents.asp</a>	Web Events	0.1672
<a href="http://go.msn.com/bql/maps.asp">http://go.msn.com/bql/maps.asp</a>	Microsoft Expedia Maps-Home	0.1672

[Tabelle 2](#). Mutual Reinforcement Authorities für Suchanfrage „movies“ [2]

Authorities, die man nach der Anwendung der Mutual Reinforcement Methode bekommt, sind von Position 1 bis 30 alle „go.msn.com“- Seiten. Alle außer der ersten Authority haben die gleiche Bewertung. Um besser zu verstehen, wie es dazu kam, dass die Seiten so gut bewertet wurden, schauen wir die [Tabelle 3](#) mit Hubs an. Auf den ersten Blick sind sie ganz verschieden, aber sie gehören alle zum Microsoft-Network (msn.com). Da alle Hubs (ohne Ausnahme) auf alle Authorities zeigen, kommt es zur gleichen Bewertung der Authorities (siehe [Tabelle 2](#)). Weil andere Seiten nicht so gut miteinander verlinkt sind, können sie nicht mit den Microsoft-Seiten konkurrieren, obwohl sie vielleicht besser zur Suchanfrage passen. An dem Beispiel sieht man, dass die msn-Seiten überbewertet wurden.

URL	Title	Weight
<a href="http://denver.sidewalk.com/movies">http://denver.sidewalk.com/movies</a>	movies: denver.sidewalk	0.1692
<a href="http://boston.sidewalk.com/movies">http://boston.sidewalk.com/movies</a>	movies: boston.sidewalk	0.1691
<a href="http://twincities.sidewalk.com/movies">http://twincities.sidewalk.com/movies</a>	movies: twincities.sidewalk	0.1688
<a href="http://newyork.sidewalk.com/movies">http://newyork.sidewalk.com/movies</a>	movies: newyork.sidewalk	0.1686

Tabelle 3. Mutual Reinforcement Hubs für Suchanfrage „movies“ [2]

Jetzt schauen wir ein Beispiel mit einer Suchanfrage zu einem mehrdeutigen Begriff „genetic“  
Tabelle 4. (Zusammenhang mit Gentechnologie, genetischer Algorithmus, menschliches Erbgut, ...)

URL	Title	Weight
<a href="http://www.aic.nrl.navy.mil/galist/">http://www.aic.nrl.navy.mil/galist/</a>	The Genetic Algorithms Archive	0.2785
<a href="http://alife.santafe.edu/">http://alife.santafe.edu/</a>	Artificial Life Online	0.2762
<a href="http://www.yahoo.com/">http://www.yahoo.com/</a>	Yahoo!	0.2736
<a href="http://www.geneticprogramming.com">http://www.geneticprogramming.com</a>	The Genetic Programming Notebook	0.2559
<a href="http://gal4.ge.uiuc.edu/illigal.home.html">http://gal4.ge.uiuc.edu/illigal.home.html</a>	illiGAL Home Page	0.2357
<a href="http://www.cs.gmu.edu/research/gag/">http://www.cs.gmu.edu/research/gag/</a>	The Genetic Algorithms Group...	0.2012
<a href="http://www.scs.carleton.ca/~csgs/resources/gaal.html">http://www.scs.carleton.ca/~csgs/resources/gaal.html</a>	Genetic Algorithms and Artificial Life Resources	0.1813
<a href="http://lancet.mit.edu/ga/">http://lancet.mit.edu/ga/</a>	GALib: Matthew's Genetic Algorithms Library	0.1812

Tabelle 4. Authorities für Suchanfrage “genetic” [2]

Wie Sie sehen, haben fast alle Seiten irgendwie mit genetischen Algorithmen zu tun. In dem Kapitel 4.4 können Sie die Ergebnisse sehen nach der Anwendung des SALSA-Algorithmus.

### Schwächen von Hits (Zusammenfassung)

- Ranking muss bei jeder Anfrage neu berechnet werden.
- *Tightly-Knit-Community-Effekt*.
- Konsequenz: Einfaches Spamming.
- Hub-Ranking kann durch ausgehende Links leicht beeinflusst werden.
- Dadurch wird auch der Autoritätswert der verlinkten Seiten gesteigert.
- Konsequenz: *Topic Drift*. [3]

## 3. PageRank

Entwicklung von S. Brin und L. Page im Jahre 1998

- Rangordnung (*Ranking*) von Webseiten durch „random walk“ (stochastischer Ansatz). Beim „random walk“ wird ein Link zufällig aus den vorhandenen ausgehenden Links mit gleicher Wahrscheinlichkeit ausgewählt.
- Die Rangordnung von Webseiten hängt nicht von der Suchanfrage ab (globales Ranking).

- $$PageRank(p) = (1 - d) + d \left( \sum_{i \text{ out degree of } q_i}^k \frac{PageRank(q_i)}{\text{out degree of } q_i} \right)$$

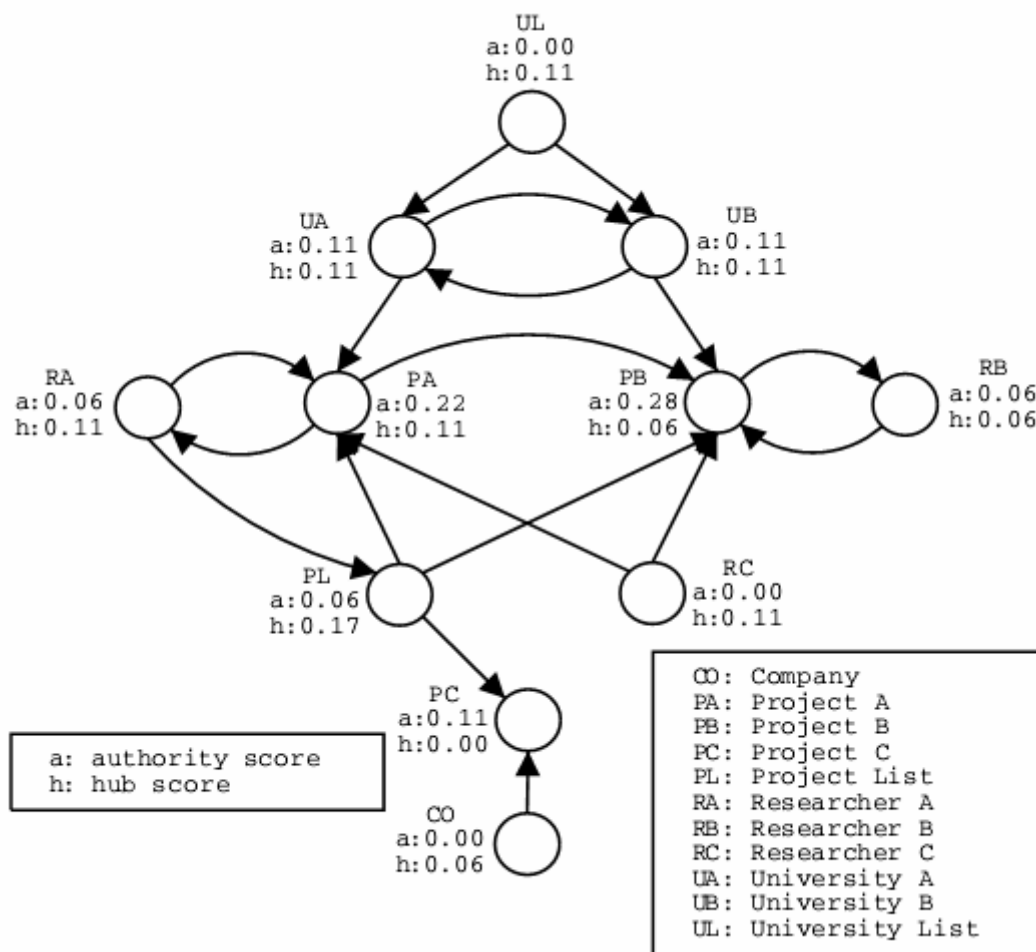
p – Webseite; d – Dämpfungsfaktor ( $0 \leq d \leq 1$ );

### Nachteil

Die Algorithmen können manipuliert werden. Dies geschieht sicher in hohem Maße, da ein erhebliches wirtschaftliches Interesse daran besteht, eigene Seiten bei vielen Internet-Suchen an vorderen Ranking-Positionen zu sehen. [5]

## 4. SALSA

In diesem Kapitel wird eine stochastische Methode zur Analyse der Linkstruktur vorgestellt, die einen „random walk“ auf einer aus einem Graph abgeleiteten Linkstruktur ausführt.



Beispiel für SALSA [4]

### 4.1 Meta-Algorithmus

SALSA benutzt den gleichen Meta-Algorithmus wie auch HITS.

*Verlauf des Algorithmus:*

- Zu einem gegebenen Thema (Suchanfrage)  $t$ , wird eine Seitensammlung  $C$  aufgebaut, die viele  $t$ -Hubs und  $t$ -Authorities enthalten soll, aber nicht viele andere Hubs und Authorities für eine andere Suchanfrage  $t'$ . Sei  $n = |C|$ .
- Man leitet von der Seitensammlung  $C$  und ihrer Linkstruktur zwei  $n \times n$ -Matrizen (Hub-Matrix  $H$  und Authority-Matrix  $A$ ) ab. Sei  $M$  eine dieser Matrizen, dann besitzt  $M$  einen eindeutigen reellen positiven Eigenwert  $\mu(M)$ , dabei gilt  $\mu(M) > |\mu'(M)|$ ,  $\mu \neq \mu'$ . Der dazu gehörige Eigenvektor  $v_{\mu(M)}$  (positiver Vektor) wird Haupteigenvektor (*principal eigenvector*) genannt.
- Seiten, die den größten Koordinaten in  $v_{\mu(A)}$  entsprechen, bilden eine Hauptauthority-Gruppe (*principal algebraic community of authorities*), und Seiten, die den größten Koordinaten in  $v_{\mu(H)}$  entsprechen, bilden eine Haupthub-Gruppe (*principal algebraic community of hubs*).

Zwei Freiheitsgrade, die der Meta-Algorithmus erlaubt, sind:

- die Methode, wie man die Seitensammlung  $C$  bekommt
- und wie man die Matrizen definiert.

## 4.2 Formelle Definition von SALSA

Wir konstruieren einen ungerichteten bipartiten Graph  $\tilde{G} = (V_h, V_a, E)$  aus unserer Seitensammlung und ihrer Linksstruktur.

- $V_h = \{s_h \mid s \in C \text{ und } \text{Ausgangsgrad}(s) > 0\}$  (Die Hub-Seite von  $\tilde{G}$ )
- $V_a = \{s_a \mid s \in C \text{ und } \text{Eingangsgrad}(s) > 0\}$  (Die Authority-Seite von  $\tilde{G}$ )
- $E = \{(s_h, r_a) \mid s \rightarrow r \text{ in } C\}$

Jede nicht isolierte Seite  $s \in C$  ist in  $\tilde{G}$  durch einen oder zwei Knoten ( $s_h$  und  $s_a$ ) repräsentiert. In der [Abbildung 1](#) ist die Konstruktion vom bipartiten Graph aus der Seitensammlung dargestellt. (Dort kann man die Hub-Seiten  $\{1, 2, 4, 5\}$  erkennen, Authority analog  $\{2, 3, 4, 6\}$ ).

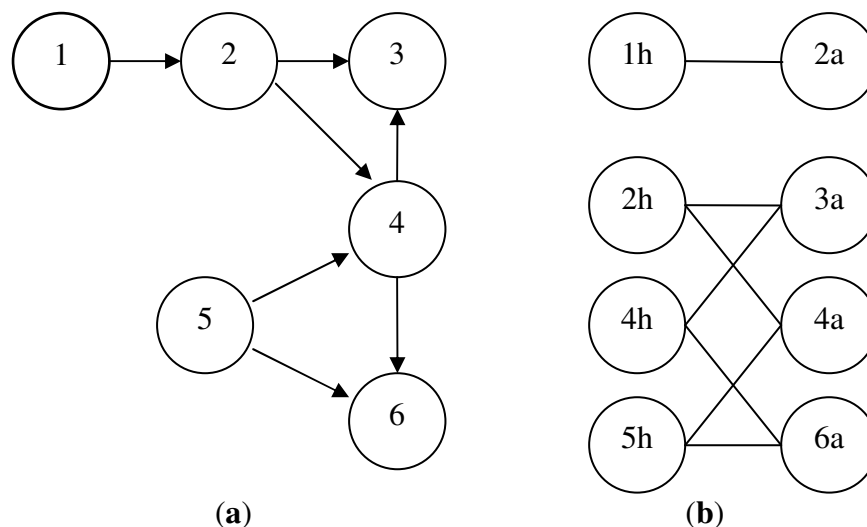


Abbildung 1: Umwandlung der Seitensammlung  $C$  (a) in den bipartiten Graph  $\tilde{G}$  (b)

Im Unterschied zu PageRank führen wir an dieser Stelle zwei „random walks“ (Autoritätenmarsch und Hubmarsch) aus. Durch Überqueren von Pfaden besucht jeder Marsch nur Knoten von einer Seite des Graphen. Dabei werden in einem Schritt zwei Kanten durchlaufen. Mit einer Kante landet man immer auf unterschiedlichen Seiten des Graphen  $\tilde{G}$ , da es keine Kante in der Knotenmenge  $V_h$  bzw.  $V_a$  gibt, die von einem Knoten zu einem anderen in  $V_h$  bzw.  $V_a$  führt. Die Knoten zur Anfrage  $t$  müssen „gut sichtbar sein“ (von vielen Knoten leicht erreichbar sein, direkt oder durch einen kleinen Pfad). Wir können erwarten, dass die  $t$ -Authorities bzw.  $t$ -Hubs zu den Knoten gehören, die am meisten durch den „random walk“ von  $V_h$  bzw.  $V_a$  besucht werden.

Wir definieren zwei stochastische Matrizen (Übergangsmatrizen von zwei Markov-Ketten) wie folgt:

- Die Hubmatrix  $\tilde{H}$ :

$$\tilde{h}_{i,j} = \sum_{\{k|(i_h, k_a), (j_h, k_a) \in \tilde{G}\}} \frac{1}{\deg(i_h)} \cdot \frac{1}{\deg(k_a)}$$

- Die Authoritymatrix  $\tilde{A}$ :

$$\tilde{a}_{i,j} = \sum_{\{k|(k_h, i_a), (k_h, j_a) \in \tilde{G}\}} \frac{1}{\deg(i_a)} \cdot \frac{1}{\deg(k_h)}$$

$\mathbf{a}_{i,j}$  bedeutet dabei die Übergangswahrscheinlichkeit, mit der man von der Seite  $\mathbf{i}$  in zwei Schritten die Seite  $\mathbf{j}$  erreicht. ( $\mathbf{h}_{i,j}$  interpretiert man analog).

### 4.3 SALSA und der Ein- und Ausgangsgrad von Seiten

In diesem Kapitel wird gezeigt, dass zur Berechnung der stationären Verteilung man keine Iterationen (wie beim HITS) benötigt. Die Verteilung der 2 Markov-Ketten wird direkt über den Eingangs- und Ausgangsgrad der Seiten berechnet.

- Es gibt verschiedene Möglichkeiten, die Links zu bewerten. Zum Beispiel:
  - man analysiert den „Ankertext“ um einen Link herum;
  - der Benutzer gibt selbst die „Ankerseiten“ an, die er passend zum einen Begriff findet;
  - man analysiert den Standort des Links.

Sei  $\mathbf{G} = (\mathbf{H}; \mathbf{A}; \mathbf{E})$  ein positiv gewichteter, gerichteter, bipartiter Graph, der keine isolierten Knoten hat, und seien alle Kanten von Seiten in  $\mathbf{H}$  zu Seiten in  $\mathbf{A}$  gerichtet.

- Gewichteter Eingangsgrad der Seite  $\mathbf{i} \in \mathbf{A}$ :

$$d_{\text{in}}(i) = \sum_{\{k \in H | k \rightarrow i\}} w(k \rightarrow i)$$

- Gewichteter Ausgangsgrad der Seite  $k \in H$ :

$$d_{\text{out}}(k) = \sum_{\{i \in A | k \rightarrow i\}} w(k \rightarrow i)$$

- Summe der Kantengewichte

$$W = \sum_{i \in A} d_{\text{in}}(i) = \sum_{k \in H} d_{\text{out}}(k)$$

- Übergangswahrscheinlichkeit von  $i$  nach  $j$ , mit  $i, j \in A$ :

$$P_A(i, j) = \sum_{\{k \in H | k \rightarrow i, k \rightarrow j\}} \frac{w(k \rightarrow i)}{d_{\text{in}}(i)} \cdot \frac{w(k \rightarrow j)}{d_{\text{out}}(k)}$$

- Übergangswahrscheinlichkeit von  $k$  nach  $l$ , mit  $k, l \in H$ :

$$P_H(k, l) = \sum_{\{i \in A | k \rightarrow i, l \rightarrow i\}} \frac{w(k \rightarrow i)}{d_{\text{out}}(k)} \cdot \frac{w(l \rightarrow i)}{d_{\text{in}}(i)}$$

$M_A$  - Markov-Kette

**Behauptung:** Wann auch immer  $M_A$  eine nicht reduzierbare Kette (hat eine nicht „reduzierbare Komponente“\*) ist, hat sie eine eindeutige stationäre Verteilung  $\Pi = (\Pi_1, \dots, \Pi_{|A|})$ , die

$$\Pi_i = \frac{d_{\text{in}}(i)}{W} \quad \forall i \in A$$

erfüllt.

Analog, wann auch immer  $M_H$  eine nicht reduzierbare Kette ist, hat sie eine eindeutige stationäre Verteilung  $\Pi = (\Pi_1, \dots, \Pi_{|H|})$ , die

$$\Pi_k = \frac{d_{\text{out}}(k)}{W} \quad \forall k \in H$$

erfüllt.

**Beweis:** siehe [1]

---

\* Früher wurde angenommen, dass der Graph zusammenhängend ist, um den stationären Verteilungsvektor durch die Summe der Gewichte der Eingangs- und Ausgangskanten zu bestimmen. Da es nicht immer der Fall sein kann, wurde in [1] gezeigt, dass man das „random walk“ auch auf nicht zusammenhängenden Komponenten durchführen kann, um die stationäre Verteilung zu bestimmen.

Wenn der ungerichtete Graph zusammenhängend ist, dann erhält jede Seite ein Authority-Gewicht, das proportional zur Summe der Gewichte von den eingehenden Kanten ist. Das Hub-Gewicht von jeder Seite ist proportional zur Summe der Gewichte von den ausgehenden Kanten. Wenn die Seitensammlung aus ungewichteten Links besteht, dann sind das Authority- und das Hub-Gewicht einfach proportional zum Eingangs- und Ausgangsgrad der Seite.

Um Webseiten mit SALSA zu ranken, muss für alle Seiten die Summe der Gewichte der eingehenden bzw. ausgehenden Links berechnet werden. [6]. Wie man sieht, braucht man keine aufwändigen Iterationen, um den Haupteigenvektor der Übergangsmatrix der Markov-Ketten zu berechnen.

#### 4.4 Der Tightly-Knit-Community-Effekt (TKC- Effekt)

Im Folgenden stehen die Ergebnisse für dieselben Suchanfragen wie im Kapitel 2.1.1, aber diesmal wurde der SALSA-Algorithmus benutzt. Wie man leicht erkennt, deckt die Seitensammlung ein größeres Quellenfeld ab.

Anfrage zu einem bestimmten Begriff „Java“

URL	Title	Weight
<a href="http://java.sun.com/">http://java.sun.com/</a>	Java(tm) Technology Home Page	0.3653
<a href="http://www.gamelan.com/">http://www.gamelan.com/</a>	Gamelan - The Official Java Directory	0.3637
<a href="http://www.jars.com/">http://www.jars.com/</a>	EarthWeb's JARS.COM Java Review Service	0.3039
<a href="http://www.javaworld.com/">http://www.javaworld.com/</a>	IDG's magazine for the Java community	0.2173
<a href="http://www.yahoo.com/">http://www.yahoo.com/</a>	Yahoo	0.2141
<a href="http://www.javasoft.com/">http://www.javasoft.com/</a>	Java(tm) Technology Home Page	0.2031
<a href="http://www.sun.com/">http://www.sun.com/</a>	Sun Microsystems	0.1874
<a href="http://www.javascripts.com/">http://www.javascripts.com/</a>	Javascripts.com - Welcome	0.1385
<a href="http://www.htmlgoodies.com/">http://www.htmlgoodies.com/</a>	htmlgoodies.com - Home	0.1307
<a href="http://javaboutique.internet.com/">http://javaboutique.internet.com/</a>	The Ultimate Java Applet Resource	0.1181

Tabelle 5. Authorities für Suchanfrage „Java“

Anfrage zu einem bestimmten Begriff „movies“

URL	Title	Weight
<a href="http://us.imdb.com/">http://us.imdb.com/</a>	The Internet Movie Database	0.2533
<a href="http://www.mrshowbiz.com/">http://www.mrshowbiz.com/</a>	Mr Showbiz	0.2233
<a href="http://www.disney.com/">http://www.disney.com/</a>	Disney.com-The Web Site for Families	0.2200
<a href="http://www.hollywood.com/">http://www.hollywood.com/</a>	Hollywood Online:...all about movies	0.2134
<a href="http://www.imdb.com/">http://www.imdb.com/</a>	The Internet Movie Database	0.2000
<a href="http://www.paramount.com/">http://www.paramount.com/</a>	Welcome to Paramount Pictures	0.1967
<a href="http://www.mca.com/">http://www.mca.com/</a>	Universal Studios	0.1800
<a href="http://www.discovery.com/">http://www.discovery.com/</a>	Discovery Online	0.1550
<a href="http://www.film.com">http://www.film.com</a>	Welcome to Film.com	0.1533
<a href="http://www.mgmua.com/">http://www.mgmua.com/</a>	mgm online	0.1300

Tabelle 6. Authorities für Suchanfrage „movies“

Suchanfrage zu einem mehrdeutigen Begriff „genetic“ [Tabelle 7](#). (Zusammenhang mit Gentechnologie, genetischer Algorithmus, menschliches Erbgut, ...)

URL	Title	Weight
<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>	The National Center for Biotechnology Information	0.2500
<a href="http://www.yahoo.com/">http://www.yahoo.com/</a>	Yahoo	0.2278
<a href="http://www.aic.nrl.navy.mil/galist/">http://www.aic.nrl.navy.mil/galist/</a>	The Genetic Algorithms Archive	0.2232
<a href="http://www.nih.gov/">http://www.nih.gov/</a>	National Institute of Health (NIH)	0.1947
<a href="http://gdbwww.gdb.org/">http://gdbwww.gdb.org/</a>	The Genome Database	0.1770
<a href="http://alife.santafe.edu/">http://alife.santafe.edu/</a>	Artificial Life Online	0.1724
<a href="http://www.genengnews.com/">http://www.genengnews.com/</a>	Genetic Engineering News (GEN)	0.1416
<a href="http://gal4.ge.uiuc.edu/illegal.home.html">http://gal4.ge.uiuc.edu/illegal.home.html</a>	illiGAL Home Page	0.1326

Tabelle 7. Authorities für Suchanfrage „genetic“

- Mehrdeutigkeit wurde besser erkannt.

*An dieser Stelle zeigen wir mathematische Hintergründe des TKC-Effekts:*

Wir bauen eine Seitensammlung  $C_k$  ( $k \geq 3$ ), die zwei Gruppen ( $C_s, C_1$ ) enthält. Gruppe  $C_s$  hat eine kleine Anzahl von Hubs und Authorities, außerdem zeigt jeder Hub auf alle Authorities. In der viel größeren Gruppe  $C_1$  zeigen Hubs nur auf ein Teil der Authorities.

Das Thema, das von  $C_1$  abgedeckt wird, ist wahrscheinlich das meist gesuchte. Da es sehr viele Authorities in  $C_1$  gibt, zeigt kein Hub auf alle diese aufgelisteten Seiten. Im Gegensatz dazu ist  $C_s$  gut verlinkt.

Der TKC-Effekt tritt auf, wenn die Seiten aus  $C_s$  werden höher bewertet werden als die von  $C_1$ .

$C_k$  hat folgende Struktur:

- $C_1$  hat  $n = (k + 1)^2$  Authorities.
- $C_s$  hat  $m = (k + 1)$  Authorities.
- Es gibt  $h_1 = \binom{n}{k}$  Hubs in  $C_1$ . Jeder dieser Hub zeigt auf eine Untermenge von  $C_1$  mit  $k$  Elementen.
- Es gibt  $h_s = \binom{n-1}{k-1}$  Hubs in  $C_s$  und jeder von denen zeigt auf alle  $C_s$ -Authorities.
- Es gibt also  $n \cdot m$  Hubs  $g_{1,1}, \dots, g_{n,m}$ . Jeder von diesen Hubs zeigt auf  $C_1$ -Authority  $i$  und  $C_s$ -Authority  $j$ .

**Behauptung:** SALSA wird die  $C_1$ -Authorities der Seitensammlung  $C_k$  höher ranken als die  $C_s$ -Authorities.

**Beweis:** [6]

## 4.5 Vor- und Nachteile von SALSA

### *Vorteile*

- Keine aufwändigen Iterationen nötig, um stationäre Verteilung zu bestimmen.
- Weniger anfällig für den TKC-Effect, für Spamming (Linksstruktur wird auch mitanalysiert).

### *Nachteile*

- Abhängig von der Suchanfrage (er muss zum Zeitpunkt der Anfrage ausgeführt werden), Zeitverlust.
- SALSA ist nicht so wenig anfällig durch Linkspamming wie PageRank, dadurch kann man auf den Bewertung Einfluss nehmen. [6]

## 5. Schlussfolgerung

In dieser Arbeit wurde gezeigt, wie man durch die Analyse von Eingangs- und Ausgangsgrad von Webseiten, diese sortiert. Außerdem wurde vorgestellt, welche Vor- und Nachteile dieses Verfahren mit sich bringt.

## 6. Literaturverzeichnis

- [1] R. Lempel, S. Moran, SALSA: The Stochastic Approach for Link-Structure Analysis, ACM Transaction on Information Systems, Vol. 19, No. 2, April 2001, Pages 131-160  
<http://www.cs.technion.ac.il/~moran/r/PS/lm-feb01.ps>
- [2] R. Lempel, S. Moran, The stochastic approach for link-structure analysis (SALSA) and the TKC effect, Computer Networks 33 (2000) 387-401  
<http://www9.org/w9cdrom/175/175.html>
- [3] Vorlesung Web-Algorithmen, gehalten von Dr. M. Brinkmeier, SS 2005  
<http://www.tu-ilmeneau.de/fakia/Web-Algorithmen.wa.0.html>
- [4] D. Olmedilla, Finding Hubs for Personalised Web Search Different ranks to different users, Tribunal de Estudios Avanzados (TEA). Universidad Autónoma de Madrid, September 2003.  
[http://www.l3s.de/~olmedilla/pub/2003/2003\\_TEA.pdf](http://www.l3s.de/~olmedilla/pub/2003/2003_TEA.pdf)
- [5] T. Mandl, Evaluierung von Internet-Verzeichnisdiensten mit Methoden des Web-Mining, (ISI 2002), Konstanz: UVK Verlagsgesellschaft mbH, 2002. S. 239 – 257  
<http://www.inf-wiss.uni-konstanz.de/infwiss/download/isi2002/cc-isi2002-art16.pdf>
- [6] C. Beyer, WS 2005/06, HS Inside Google, Uni-Stuttgart.  
<http://www.informatik.uni-stuttgart.de/fmi/szs/teaching/ws0506/google/ausarbeitungen/beyer.pdf>