Kapitel 4: Linkanalyse für Autoritäts-Ranking

Zusätzlich zu ihrer inhaltlichen Relevanz können Dokumente auch nach ihrer Autorität, also nach Umfang, Klarheit und Signifikanz der in einem Dokument enthaltenen Information, bewertet werden, und entsprechende Autoritätsmaße können in das Ranking der Resultatsdokumente einer Query einfließen. Durch Kombination von Relevanz- und Autoritätsbewertung ist es möglich, die Präzision von Suchresultaten deutlich zu steigern, indem unter allen Treffern einer Anfrage diejenigen Dokumente mit hoher Autorität möglichst weit vorne in der Rangliste platziert werden.

Eine naheliegende Idee, Autorität von Web-Dokumenten zu quantifizieren, besteht darin, die eingehenden Hyperlinks einer Web-Seite als Zitate durch andere Web-Nutzer zu interpretieren. Einfach nur die Anzahl eingehender Links als Autoritätsmaß zu verwenden, greift jedoch zu kurz, da Links je nach Ursprung unterschiedliches Gewicht haben könnten und ein brauchbarer Ansatz resistent sein sollte gegen Link-Manipulationen, Zitier-Cliquen, u.ä.

4.1 Grundlagen aus der Stochastik

Ein **stochastischer Prozeß** ist eine Familie von Zufallsvariablen $\{X(t) \mid t \in T\}$.
T heißt Parameterraum, und der Definitionsbereich M der X(t) heißt Zustandsraum. T und M können diskret oder kontinuierlich sein.
Ein stochastischer Prozeß heißt **Markov-Prozeß**, wenn für beliebige $t_1, ..., t_{n+1}$ aus dem Parameterraum und für beliebige $x_1, ..., x_{n+1}$ aus dem Zustandsraum gilt:
Ein Markov-Prozeß mit diskretem Zustandsraum heißt **Markov-Kette**. O.B.d.A. werden die natürlichen Zahlen als Zustandsraum gewählt. Als Notation für Markov-Ketten mit diskretem Parameterraum schreiben wir: X_n statt $X(t_n)$ mit n = 0, 1, 2, ...

Die Markov-Kette Xn mit diskretem Parameterraum heißt
homogen, wenn die Übergangswahrscheinlichkeiten $p_{ij} := P[X_{n+1} = j \mid X_n = i]$ unabhängig von n sind
irreduzibel, wenn jeder Zustand von jedem Zustand mit positiver Wahrscheinlichkeit erreichbar ist:
$\sum_{n=1}^{\infty} P[X_n = j \mid X_0 = i] > 0$
aperiodisch, wenn alle Zustände i die Periode 1 haben, wobei die Periode von i der ggT aller Werte n ist, für die gilt: $P[X_n = i \wedge X_k \neq i \text{ für } k = 1,...,n-1 \mid X_0 = i] > 0$

Die Markov-Kette Xn mit diskretem Parameterraum heißt
positiv rekurrent, wenn für jeden Zustand i die Rückkehrwahrscheinlichkeit gleich 1 ist und die mittlere Rekurrenzzeit endlich: $\sum_{n=1}^{\infty} P[X_n = i \wedge X_k \neq i \text{ for } k = 1,...,n-1 \mid X_0 = i] = 1$ und

$\sum_{n=1}^{\infty} n\, P[X_n = i \wedge X_k \neq i \text{ for } k = 1,...,n-1 \mid X_0 = i] < \infty$. Sie heißt
ergodisch, wenn sie homogen, irreduzibel, aperiodisch und positiv rekurrent ist.

Für die **n-Schritt-Transitionswahrscheinlichkeiten** gilt:
$p_{ij}^{(n)} := P[X_n = j \mid X_0 = i] = \sum_k p_{ik}^{(n-1)} p_{kj}$ mit $p_{ij}^{(1)} := p_{ik}$ bzw. allgemeiner: $p_{ij}^{(n)} = \sum_k p_{ik}^{(n-t)} p_{kj}^{(t)}$ für $1 \leq t \leq n-1$

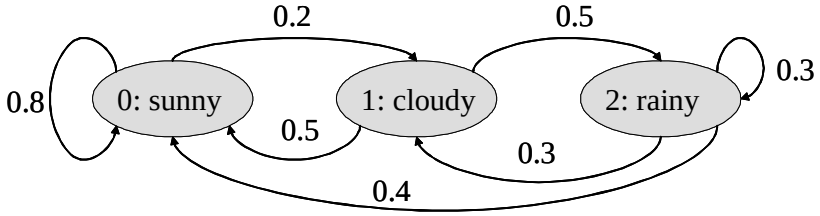
und in Matrixnotation: $P^{(n)} = P^n$.

Für die **Zustandswahrscheinlichkeiten nach n Schritten** gilt:
$\pi_j^{(n)} := P[X_n = j] = \sum_i \pi_i^{(0)} p_{ij}^{(n)}$ mit initialen Zustandswahrscheinlichkeiten $\pi_i^{(0)}$
und in Matrixnotation: $\Pi^{(n)} = \Pi^{(0)} P^{(n)}$ mit einem 1×n-Zeilenvektor Π.
Diese Gleichung nennt man die Chapman-Kolmogorov-Gleichung für Markov-Ketten.

Satz:
Jede homogene, irreduzible, aperiodische Markov-Kette mit endlich vielen Zuständen ist positiv rekurrent und ergodisch.
Für jede ergodische Markov-Kette existieren **stationäre Zustandswahrscheinlichkeiten** $\pi_j := \lim_{n\to\infty} \pi_j^{(n)}$**.** Diese sind unabhängig von $P^{(0)}$ und durch das folgende lineare Gleichungssystem bestimmt: $\pi_j = \sum_i \pi_i p_{ij}$ *für alle* j (Gleichgewichtsgleichungen) und $\sum_j \pi_j = 1$
bzw. in Matrix-Notation: $\Pi = \Pi P$ und $\Pi \vec{1} = 1$ mit einem n×1-Spaltenvektor $\vec{1}$.

Π ist also ein Eigenvektor der Matrix P zum dominanten (d.h. betragsgrößten) Eigenwert 1.
Man kann die Lösung für Π auf zwei Arten berechnen: entweder durch Lösen des Gleichungssystems oder approximativ durch Potenziteration (Power Iteration): $\Pi^{(0)} = (1/n \ldots 1/n)^T$ und $(\Pi^{(i)})^T = (\Pi^{(i-1)})^T P$.

Beispiel:



$\pi0 = 0.8\,\pi0 + 0.5\,\pi1 + 0.4\,\pi2$
$\pi1 = 0.2\,\pi0 + 0.3\,\pi2$
$\pi2 = 0.5\,\pi1 + 0.3\,\pi2$
$\pi0 + \pi1 + \pi2 = 1$

$\Rightarrow \pi0 = 330/474 \approx 0.696$
$\pi1 = 84/474 \approx 0.177$
$\pi2 = 10/79 \approx 0.126$

4.2 Autoritäts-Ranking nach der Methode von Page und Brin (Page-Rank)

Das Web (oder ein Intranet) wird bei dieser Methode, die in der Suchmaschine Google verwendet wird, als gerichteter Graph G = (V,E) gesehen mit Web-Seiten als Knotenmenge V, |V|=n, und Hyperlinks als Kantenmenge E. Für die folgenden Berechnungen wird angenommen, dass der Graph komplett a priori bekannt ist und seine Adjazenzmatrix A, eine n×n-Matrix mit Aij = 1 falls (i,j)∈E, 0 sonst, auf einem Server gespeichert ist. Google baut diesen Graph aufgrund der Ergebnisse seines Crawlers auf; diese sind bei Google mehr als 1 Milliarde Knoten.

Die Kernidee dieser Methode ist, dass die *Autorität (Authority-Score, Authority-Rank)* r(q) einer Web-Seite q proportional zur Summe der Autoritätsbewertungen der Vorgänger von q ist – vorausgesetzt, alle Vorgänger hätten dieselbe Anzahl ausgehender Links. Bei Vorgängern mit vielen ausgehenden Kanten würde man das Autoritätsmaß nur anteilsmäßig auf die Zieldokumente der Kanten umlegen. Diese Überlegung führt auf die folgende erste Arbeitsdefinition:

$$r(q) = k \sum_{(p,q)\in E} r(p) / outdegree(p)$$

mit einer Konstanten k und der Anzahl *outdegree(p)* von p ausgehender Kanten.

Es zeigt sich jedoch, dass diese Definition noch nicht tragfähig ist, da unklar ist, wie man mehreren Zusammenhangskomponenten oder gar isolierten Knoten umgehen sollte. Daher weist die Lösung von Page und Brin jeder Web-Seite unabhängig von ihren Vorgängern eine Mindestautorität ε/n zu und berechnet die (1-ε) gewichtete Restautorität nach dem o.a. Ansatz.

Definition:

Die Autorität der Web-Seite q im Web-Graphen G=(V,E) ist gegeben durch

$$r(q) = \varepsilon / n + (1-\varepsilon) \sum_{(p,q)\in E} r(p) / outdegree(p) .$$

Dabei ist ε ein Kalibrierungsparameter, für den lt. Page und Brin $0 < \varepsilon \leq 0.25$ gelten sollte.

Satz:

Mit einer modifizierten Matrix A' mit A'_{ij} = 1/outdegree(j) falls $(j,i)\in E$ und 0 sonst, gilt

$\vec{r} = \vec{\varepsilon} / n + (1-\varepsilon)A'\vec{r}$ und äquivalent dazu $\vec{r} = \left(\frac{\vec{\varepsilon}}{n}\vec{1}^T + (1-\varepsilon)A' \right)\vec{r}$.

Dabei ist $\vec{r}$ ein n×1-Spaltenvektor, und $\vec{\varepsilon}$ und $\vec{1}$ sind n×1-Spaltenvektoren, die komplett mit dem Wert ε bzw. 1 besetzt sind.

Man sieht somit, dass der Spaltenvektor $\vec{r}$ der Autoritätswerte Eigenvektor einer modifizierten Transitionsmatrix ist. Eine approximative Berechnung durch Potenziteration ist:

1) $\vec{r}^{(0)} = \vec{1} / n$
2) Wiederhole bis sich die größten Werte von $\vec{r}$ kaum noch ändern:
   $\vec{r}^{(i+1)} = \vec{\varepsilon} / n + (1-\varepsilon)A'\vec{r}^{(i)}$

Die besten Autoritäten sind dann diejenigen Komponenten von $\vec{r}$ mit den größten Werten.

Diese Methode funktioniert in der Praxis verblüffend gut. Google z.B. rechnet ca. 100 Iterationen und speichert dann die Web-Seiten-spezifischen Autoritätswerte in seinem Index. Bei Anfragen berechnet Google pro Trefferseite einen anfragespezifischen Relevanz-Score und kombiniert diesen in einer gewichteten Summe mit dem vorberechneten Autoritäts-Score. Die Resultatsliste ist der Präfix der nach dieser gewichteten Summe absteigend sortierten Liste. Die relativen Gewichte von Relevanz und Autorität sind per Trial-and-Error anhand typischer Google-Queries ermittelt.

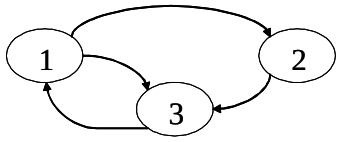
Auch die Methode in der Praxis sehr gut funktioniert, ist vor allem der Ad-hoc-Charakter der oben eingeführten Gewichtungen, insbesondere auch der etwas willkürlichen Einführung von ε, wissenschaftlich unbefriedigend. Eine alternative Herleitung derselben Lösung verwendet einen sog. Random Walk auf dem Web-Graphen und mathematische Resultate über Markov-Ketten.

Autoritätsbewertung mittels Random Walk

Das Verfahren von Page und Brin modelliert einen Random Walk über den Web-Graphen, bei dem man mit Wahrscheinlichkeit (1-ε) von der aktuell besuchten Web-Seite zu einem der Nachfolger wandert und mit Wahrscheinlichkeit ε einen „Random Jump“ zu irgendeiner Web-Seite macht. Im ersten Fall wird die als nächstes besuchte Seite gemäß einer Gleichverteilung unter den Nachfolgern der aktuellen Seite ausgewählt; im zweiten Fall wird eine Seite unter allen Web-Seiten gemäß einer Gleichverteilung ausgewählt.

Für den so definierten Random Walk kann man eine Markov-Kette angeben, die beweisbar ergodisch ist. Der Autoritäts-Score einer Seite nach dem Verfahren von Page und Brin ist die **stationäre Besuchswahrscheinlichkeit der Web-Seite** in dieser Markov-Kette.

Beispiel:



$\varepsilon = 0.2$ $\quad P = \begin{pmatrix} 0.0 & 0.5 & 0.5 \\ 0.1 & 0.0 & 0.9 \\ 0.9 & 0.1 & 0.0 \end{pmatrix}$

$$\Pi^{(0)} \approx \begin{pmatrix} 0.333 \\ 0.333 \\ 0.333 \end{pmatrix}^T \Rightarrow \Pi^{(1)} \approx \begin{pmatrix} 0.333 \\ 0.200 \\ 0.466 \end{pmatrix}^T \Rightarrow \Pi^{(2)} \approx \begin{pmatrix} 0.439 \\ 0.212 \\ 0.346 \end{pmatrix}^T \Rightarrow \Pi^{(3)} \approx \begin{pmatrix} 0.332 \\ 0.253 \\ 0.401 \end{pmatrix}^T$$

$$\Rightarrow \Pi^{(4)} \approx \begin{pmatrix} 0.385 \\ 0.176 \\ 0.527 \end{pmatrix}^T \Rightarrow \Pi^{(5)} \approx \begin{pmatrix} 0.491 \\ 0.244 \\ 0.350 \end{pmatrix}^T$$

$\pi1 = 0.1\ \pi2 + 0.9\ \pi3$
$\pi2 = 0.5\ \pi1 + 0.1\ \pi3$
$\pi3 = 0.5\ \pi1 + 0.9\ \pi2$
$\pi1 + \pi2 + \pi3 = 1$ $\quad \Rightarrow \pi1 \approx 0.3776,\ \pi2 \approx 0.2282,\ \pi3 \approx 0.3942$

4.3 Autoritäts-Ranking nach der HITS-Methode von Kleinberg

Bei dem HITS-Verfahren nach Kleinberg (Hyperlink-Induced Topic Search) analysiert man die Linkstruktur eines relativ kleinen Untergraphen, der aufgrund seiner thematischen Relevanz bestimmt wird. Man berechnet zunächst aufgrund einer klassischen Relevanzbewertung eine Menge von Wurzelseiten, z.B. die Top 100 einer Anfrage bei Google oder AltaVista, und fügt zu dieser alle Nachfolger und möglichst viele Vorgänger hinzu sowie alle Kanten zwischen den betrachteten Seiten; der Graph G=(V,E), |V|=n, der so enstandene Menge von – z.B. einigen Tausend – Seiten V bildet die Basis der Linkanalyse. Anders als beim Verfahren nach Page und Brin ist dieser Graph anfragespezifisch.

Die Arbeitshypothese des HITS-Verfahrens ist, dass es in einem solchen Graph außer guten *Autoritäten (Authorities)* auch gute *Referenzen (Hubs)* gibt:Linksammlungen, die Verweise auf die besten Autoritäten eines bestimmten Themas enthalten. Das Verfahren berechnet für jede Seite p im Graphen sowohl ein *Autoritätsgewicht (Authority-Score)* x_p als auch ein *Referenzgewicht (Hub-Score).* y_p. Zwischen Autoritäten und Referenzen gibt es eine wechselseitige Rekursion: eine Autorität ist umso besser, hat also höheres Referenzgewicht, je besser die Autoritäten sind, auf die sie verweist, und eine Autorität ist umso besser, hat also höheres Autoritätsgewicht, je besser die Referenzen sind, die auf sie verweisen. Wenn man postuliert, dass dieser Zusammenhang zwischen Autoritätsgewichten x_p und Referenzgewichten y_p linear ist, kommt man auf folgende Gleichungen:

$$x_q = \sum_{(p,q)\in E} y_p \quad \text{und} \quad y_p = \sum_{(p,q)\in E} x_q$$

bzw. in Matrix-Notation:

$\vec{x} = A^T \vec{y}$ und $\vec{y} = A\vec{x}$, wobei A die Adjazenzmatrix von G ist.

Setzt man die rechte Seite der y-Gleichung in die rechte Seite der x-Gleichung ein (und analog für y), so erhält man

$$\vec{x} := A^T \vec{y} := A^T A \vec{x} \quad \text{und} \quad \vec{y} := A\vec{x} := AA^T \vec{y},$$

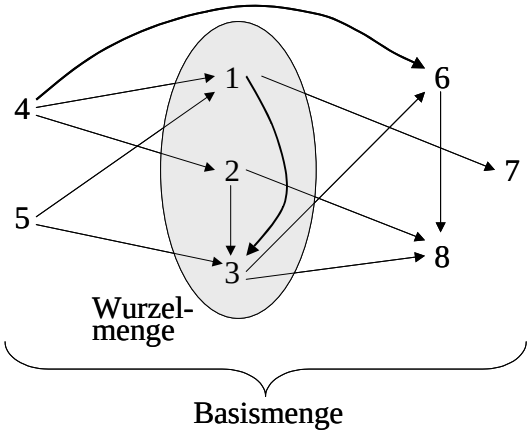
und wir erkennen, dass die Vektoren x und y Eigenvektoren der Matrizen A^TA bzw. AA^T sind.

Die Matrix $M^{(auth)} := A^TA$ kann als Cocitation-Matrix interpretiert werden: $M^{(auth)}_{ij}$ ist die Anzahl der Web-Seiten, die auf i und auf j verweisen. Die Matrix $M^{(hub)} := AA^T$ kann als Bibliographic-Coupling-Matrix interpretiert werden: $M^{(hub)}_{ij}$ ist die Anzahl der Web-Seiten, auf die sowohl i als auch j verweisen

Die wechselseitige Rekursion kann iterativ berechnet werden, indem man sowohl den x- als auch den y-Vektor zunächst mit dem Wert 1/n in allen Komponenten initialisiert und dann abwechselnd x und y neu berechnet, bis sich die größten Komponenten nur noch geringfügig ändern. In jedem Iterationsschritt werden die erhaltenen x- und y-Vektoren auf die Länge 1 normalisiert. Dieses Iterationsverfahren konvergiert (unter bestimmten Bedingungen, die in der Regel gegeben sind) gegen die Eigenvektoren von A^TA und AA^T, die zum betragsgrößten Eigenwert gehören.

Der HITS-Algorithmus sieht insgesamt also folgendermaßen aus:
1) Bestimme hinreichend viele (z.B. 50-200) „Wurzelseiten“
   per Relevanz-Ranking (z.B. mittels tf*idf-Ranking)
2) Füge alle Nachfolger von Wurzelseiten hinzu
3) Füge für jede Wurzelseite max. d Vorgänger hinzu (durch Zufallsauswahl unter allen Vorgängern)
4) Erzeuge den Graph G=(V,E), |V|=n, für die so erhaltene Basismenge
5) Initialisiere alle Komponenten von $x^{(0)}$ mit 1/n und die Komponenten von $y^{(0)}$ mit 1/n
6) Solange noch keine hinreichende Konvergenz erreicht ist
   (oder für eine feste Anzahl von Iterationen) wiederhole
   $x^{(i)} := A^T y^{(i-1)}$ und $y^{(i)} := A x^{(i-1)}$
7) Gib Seiten nach absteigend sortierten Authority-Scores aus
   (z.B. die 10 größten Komponenten von x)

Illustration der Schritte 1 bis 4:



Das HITS-Verfahren hat eine gewisse Anfälligkeit gegenüber Themendrifts: auch wenn die Wurzelmenge nur für das ursprünglich gegebene Thema (die Anfrage) relevant ist, könnte der Algorithmus auf sehr starke Autoritäten und Referenzen zu einem anderen Thema stoßen, so dass das Endresultat durch das andere Thema dominiert sein könnte. Um dies zu verhindern, kann man – in einer erweiteten Variante – in jedem Iterationsschritt thematische Relevanzwerte (z.B. tf*idf-basierte Ähnlichkeiten zur ursprünglichen Anfrage) als multiplikative Gewichte der Vorgänger- bzw. Nachfolger-Gewichte einbauen.

Das HITS-Verfahren kann auch benutzt werden, um zu einer gegebenen Web-Seite ähnliche Web-Seiten zu ermitteln; dabei bezieht sich Ähnlichkeit auf die Linkstruktur, z.B. wären zwei Seiten sehr ähnlich, wenn sie genau dieselben Vorgänger und Nachfolger hätten. Für diese Art der Ähnlichkeitsanalyse bestimmt man zu einer gegebenen Web-Seite zunächst alle Nachfolger, alle oder eine beschränkte Anzahl der Vorgänger sowie (eine beschränkte Zahl von) Vorgänger(n) der Nachfolger und Nachfolger(n) der Vorgänger. Auf dieser Basismenge führt man dann den HITS-Algorithmus aus, und die ermittelten Autoritätsgewichte liefern eine Rangliste ähnlicher Seiten.

4.4 Themenspezifisches Page-Rank-Verfahren

Alle Verfahren zur Autoritätsanalyse verursachen signifikanten Berechnungsaufwand, so dass sie bei großen Web-Suchmaschine nicht für jede Anfrage neu berechnet werden können. Bei Google etwa werden Page-Rank-Werte nur gelegentlich berechnet und im Index abgespeichert. Autorität ist also in diesem Fall Query- und damit Benutzer-unabhängig. Eine Methode, Autorität in einer für den jeweiligen Benutzer spezifischen Form auszunutzen, ist das themenspezifische Page-Rank-Verfahren. Dabei werden Anfragen durch Klassifikation auf eine vorgegebene Menge von Themen wie z.B. Sport, Politik, Elektronik, Internet, etc. abgebildet. Für jedes Thema gibt es eine intellektuell zusammengestellte Menge von spezifischen Autoritäten, und ein themenspezifischer Page-Rank-Vektor wird dadurch berechnet, dass man die Random Jumps so modifiziert, dass die Autoritäten des jeweiligen Themas mit höherer Wahrscheinlichkeit angesprungen werden.

Sei T_k die Menge der vorgegebenen Autoritäten für das k-te Thema. Die Page-Rank-Gleichung für Thema k lautet dann:

$$\vec{r}_k = \varepsilon \vec{p}_k + (1-\varepsilon) A' \vec{r}_k$$

mit einem Zielvektor $\vec{p}_k$ für Random Jumps, dessen i-te Komponente $1/|T_k|$ ist, wenn i in T_k liegt und 0 sonst (und dieser Ansatz ließe sich noch weiter verallgemeinern, um die relativen Autoritätswerte der Seiten in T_k zu berücksichtigen). A'_{ij} ist 1/outdegree(i), wenn es einen Link von Seite i zur Seite j gibt, und 0 sonst.

Für jedes Thema werden die themenspezifischen Page-Rank-Werte aller Web-Seiten vorausberechnet. Zum Zeitpunkt der Ausführung einer Query q werden folgende Schritte ausgeführt:

1) Berechne Wahrscheinlichkeiten bzw. normalisierte Konfidenzwerte w_k, dass q zur Klasse c_k gehört. (Dies kann man z.B. mit einem Naive-Bayes-Klassifikator tun.)
2) Setze den Autoritätswert einer für q interessanten, potentiellen Trefferseite d auf $\sum_k w_k \, r_k(d)$.

Alternativ könnte man auch Benutzer aufgrund von Profilen oder expliziter Registrierung einem oder mehreren Themen zuordnen und daraus Gewichte w_k ableiten.

Wenn man bei der **Personalisierung** des Suchmaschinenverhaltens noch weitergehen will, könnte man jeden Benutzer selbst einen persönlichen Random-Jump-Vektor $\vec{p}_k$ anlegen lassen. Auf den ersten Blick sieht das nach einem monströsen Verwaltungs-Overhead aus, der für eine Websuchmaschine mit Millionen von Benutzern nicht mehr leistbar ist. Wenn jedoch die Gesamtheit der Random-Jump-Ziele aller Benutzer eine nicht zu große Menge T bildet, kann man jeden der persönlichen Page-Rank-Vektoren als Linearkombination von elementaren Page-Rank-Vektoren berechnen. Dabei ist ein elementarer Page-Rank-Vektor der Vektor der Autoritätswerte aller Webseiten, der sich aus der Gleichung $\vec{r}_i = \varepsilon \vec{e}_i + (1-\varepsilon) A' \vec{r}_i$ ergibt mit dem Basisvektor $\vec{e}_i$, der als i-te Komponente den Wert 1 hat und in allen anderen Komponenten den Wert 0. Bei diesem Random Walk gibt es also nur ein einziges Random-Jump-Ziel, die Seite i. Hat man nun $\vec{r}_i$ für jede Seite i aus T berechnet, kann man für jeden Benutzer mit Random-Jump-Profil $\vec{p} = \sum_{i \in T} \alpha_i \vec{e}_i$ - mit Gewichten $\alpha_i \geq 0$ - den persönlichen Page-Rank-Vektor $\vec{r}$ als konvexe Linearkombination der $\vec{r}_i$ berechnen: $\vec{r} = \sum_{i \in T} \alpha_i \vec{r}_i$.

Ergänzende Literatur

- S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, WWW Conference 1998.
- J.M. Kleinberg: Authoritative Sources in a Hyperlinked Environment, Journal of the ACM Vol.46 No.5, 1999
- C. Ding, X. He, P. Husbands, H. Zha, H. Simon: PageRank, HITS, and a Unified Framework for Link Analysis, SIAM Int. Conf. on Data Mining, 2003.
- A. Borodin, J. S. Rosenthal, G. O. Roberts, P. Tsaparas: Finding Authorities and Hubs From Link Structures on the World Wide Web, WWW Conference 2001
- Taher Haveliwala: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search, IEEE Transactions on Knowledge and Data Engineering Vol. 15 No. 4, 2003

...