Max-Planck-Institut Saarbrücken
AG5: Datenbanken und Informationssysteme
Dozenten: Prof. Dr.-Ing. G. Weikum, Dr. Ralf Schenkel
Betreuer: Dipl.-Inform. Sergej Sizov, Dipl.-Inform. Christian Zimmer

Informationssysteme (SS 05)
Übungsblatt 3
Part II

Beispiellösungen

Zur Vorwärmung könnten (optional) folgende einfache Fragen besprochen werden:

- effiziente Implementierung der Datenstrukturen für Linkanalyse (dünne Matrizen etc.)
- was kann rauskommen wenn HITS auf dem Page-Rank Graphen berechnet wird
- welche Probleme können auftreten wenn Page-Rank auf dem HITS Graphen berechnet wird

Aufgabe 9: Sonderfälle bei Page-Rank und HITS

a) Welchen Einfluss haben Knoten mit Eingangsgrad 0 beim Page-Rank- und beim HITS-Verfahren? Welchen Einfluss haben Knoten mit Ausgangsgrad 0? Ergeben sich daraus ggf. Optimierungen für die Berechnung der Autoritäts-Scores?
b) Lassen sich das Page-Rank-Verfahren und das HITS-Verfahren auch auf ungerichtete Graphen anwenden? Macht das Sinn?

a)
Im einem Page-Rank Modell ohne Random-Jump würden solche Knoten offensichtlich AuthScore=0 haben, da sie nicht erreichbar sind. Im Modell mit Random-Jumps haben alle solche Knoten den „gleichen“ AuthScore, der kleiner ist als AuthScore jedes Knoten mit Indegree >0.
Der Beweis kann, zum einen, durch formale Betrachtung der Transitionsmatrix durchgeführt werden Es gibt einen solchen Beweis im Paper von Ding (Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, Horst Simon: Page-Rank, HITS, and a Unified Framework for Link Analysis SIAM International Conference on Data Mining, 2003. http://www.nersc.gov/research/SCG/cding/papers_ps/sigpage6b.ps)

Einfacher ist es allerdings, die Struktur der jeweiligen Markov-Kette zu betrachten. Die Wahrscheinlichkeit, in einem Knoten k zum Zeitpunkt t den Random-Surfer zu finden, ist gleichzeitig die Summe der Wahrscheinlichkeiten, dass er zum Zeitpunkt t-1 in einem seiner Vorgänger war. multipliziert mit entsprechenden Übergangswahrscheinlichkeiten. Daraus ergibt sich direkt die o.a. Aussage.
Für Page-Rank sind Knoten mit Indegree=0 also insignifikant, was zur Beschleunigung der Berechnung genutzt werden kann. Die Motivation einer Optimierung durch Rausschmeißen der Indegree=0 (also das es viele gibt und die Verbesserung signifikant sein könnte) gibt u.a. ein Paper von Broder (Graph Structure of the Web). Er hat gezeigt, dass die Wahrscheinlichkeiten

- indegree: durch n Seiten referenziert zu werden ~ n^(-2.1) ist und
- outdegree: n links selber zu enthalten ~ n^(-2.72) ist.

Aus diesen Power Laws kann man auch sehen, dass 2.1 < 2.72 ist. Das ist übrigens der versteckte Grund, warum HITS in seiner Ursprungsform alle ausgehenden Links, aber nicht alle eingehenden Links bei der Modellierung berücksichtigt.
HITS hat prinzipiell keine Random-Jumps, und daher
- Authscore mit Indegree=0 ist immer 0
- Hubscore mit Outdegree=0 ist immer 0

An dieser Stelle kann auch diskutiert werden, ob es sinnvoll wäre, bestimmte Random-Jumps für das HITS Modell einzuführen.

b)
Der prinzipielle Unterschied besteht darin, dass im neuen Modell alle Verbindungen symmetrisch sind. Die Matrix der Übergänge ist dementsprechend auch symmetrisch.
Meine Vermutung ist, dass in diesem Fall eine 'Verwässerung' der Hubs/Authorities auftreten wird, d.h. es wird 'Blöcke' (z.B. Domänen bzw. stark vernetzte Communities) mit gleichen Scores innerhalb eines solchen Blocks geben.

Aufgabe 10: Themenspezifisches HITS-Verfahren

Entwerfen Sie - analog zum themenspezifischen Page-Rank-Verfahren - eine Methode, die themenspezifische Authorities und Hubs nach einem geeignet erweiterten HITS-Verfahren ermittelt.

Es gibt ein gutes Paper, dass dieses Thema behandelt:

Taher Haveliwala:
Topic-Sensitive Page-Rank: A Context-Sensitive Ranking Algorithm for Web Search
IEEE Transactions on Knowledge and Data Engineering, to appear in 2003.
(http://www.stanford.edu/~taherh/papers/topic-sensitive-pagerank-tkde.pdf)

Prinzipiell basieren alle Modifikationen des HITS auf der Anpassung der Weise, wie Auth-Score und Hubscore iterativ berechnet werden. Man arbeitet typischerweise mit zusätzlichen Koeffizienten, die abhängig von z.B Klassifikationsgüte des Dokuments innerhalb de Klasse gesetzt werden.