6. Übung zur Vorlesung "Internet-Suchmaschinen" im Wintersemester 2007/2008

Dr. Andreas Hotho, Prof. Dr. Gerd Stumme, MSc. Wi-Inf. Beate Krause

24. Januar 2008

1 Bibliometrische Maße und Link-Analyse

1. Inwiefern sind Ko-Zitation und Kopplung symmetrische Phänomene?
2. Auf welches der beiden Maße haben die Autoren der jeweiligen Schriften unmittelbaren Einfluß, auf welches nicht?
3. Sie schreiben einen wissenschaftlichen Artikel A. Ein Nobelpreisträger hat einen preisgekrönten Artikel B geschrieben. Was wäre Ihnen lieber: eine hohe Ko-Zitation von A und B, oder eine hohe Kopplung von A und B? Warum?
4. Was hat HITS mit Ko-Zitation und Kopplung zu tun? Können Sie den Fortschritt einer HITS-Berechnung mit diesen beiden Maßen beschreiben? Was genau bedeuten Ko-Zitation und Kopplung für den Fluß des Gewichtes im Graphen?

   Tipp: Stellen Sie sich die HITS-Iterationen so aufgeteilt vor, daß jeweils in den ungeraden Schritten das Authority-Gewicht von Hubs zu Authorities, in den geraden Schritten das Hub-Gewicht von den Authorities zu den Hubs fließt.
5. Gegeben sei die folgende Adjazenzmatrix:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Bestimmen Sie $\vec{a}$ und $\vec{h}$ für die ersten drei Iterationen. Sie brauchen dafür nicht zu normalisieren.

2 Spam in der Bibliometrie und bei der Link-Analyse

1. Beschreiben Sie eine einfache Technik, wie bei Link-Analyse-Verfahren wie PageRank eine Seite im Web ihren Rang erhöhen kann.

2. Können Sie sich eine Gegenmaßnahme vorstellen? Wie könnte diese z. B. in PageRank umgesetzt werden?

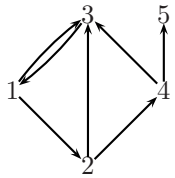
3. Ähnliche Maße wie der Einflußfaktor für Zeitschriften sind auch für die Bewertung der wissenschaftlichen Leistung von Einzelpersonen denkbar (Wie oft wird Autor X zitiert, usw.).

   Welche Tricks könnte es geben, um den eigene Bedeutung in solchen bibliographischen Einflußmaßen künstlich zu erhöhen? Wie kann diesen begegnet werden?

4. Warum sind solche Manipulationen im Web einfacher und effektiver umzusetzen als in der Bibliometrie?

3 Link-Analyse

1. Betrachten Sie den folgenden Web-Graphen. Können Sie vorhersagen, wie der Pagerank der einzelnen Seiten aussehen wird, wenn ohne Gewichtsquelle E gerechnet wird? Welcher Knoten ist der "Schuldige" für dieses Ergebnis? Warum?



2. Wie wird dieses Problem zur Manipulation von Suchergebnissen eingesetzt?

3. Entfernen Sie den verdächtigen Knoten aus dem Graphen und berechnen Sie die ersten 5 Iterationen von PageRank ohne Gewichtsquelle. Das Anfangsgewicht sei bei allen Knoten gleich. Schätzen Sie das Endergebnis ab (Tip: es läßt sich gut in Elfteln ausdrücken).

4. Wenn man den eben entfernten Knoten dennoch gewichten wollte – welches Gewicht würden Sie ihm geben?

4 Link-Analyse

HITS, Google und der personalisierte PageRank beschreiben drei Möglichkeiten, inhaltsbasierte Suchverfahren und Link-Analyse zu verknüpfen.

1. Beschreiben Sie kurz die Art der Verknüpfung und grenzen Sie die drei Varianten voneinander ab!
2. Welche der drei Verfahren sind für eine praktisch verwendbare Suchmaschine auf dem gesamten Web nutzbar, welche nicht? Warum?

5 Inhaltsbasiertes und kollaboratives Filtern

1. Nennen Sie Merkmale (je ≥ 3), die die Gegenstände beim inhaltsbasierten Filtern in den folgenden Anwendungsdomänen beschreiben könnten:
   a) Reisebuchungen
   b) Bücher
   c) Musikdownloads
   d) Filme
   e) Gute Vorsätze
2. Wie könnte eine Kombination aus kollaborativem und inhaltsbasiertem Filtern aussehen?

6 Praxisaufgabe (Abgabe: 6.2.08)

Sie haben bereits einen invertierten Index auf Dokumenten sowie einen Spider implementiert.
Bauen Sie diese beiden Komponenten zu einer Suchmaschine zusammen! Dazu soll auf der Basis von Apache Tomcat mit Hilfe von Java-Servlets oder Java Server Pages eine Web-Schnittstelle entwickelt werden, die folgende Funktionalität anbietet:

- Anfrageformular, in das der Benutzer zu suchende Terme eingeben kann
- Anzeige passender Dokumente in TF-IDF-Ranking
- Eingabe von zu crawlenden Basis-URLs und einer Schranke (Anzahl Seiten, Crawltiefe o.ä.). Diese Seiten sollen aus dem Web geholt und zum Index hinzugefügt werden.

Tomcat gibt es hier: `http://tomcat.apache.org/`