8. Übung zur Vorlesung “Internet-Suchmaschinen” im Sommersemester 2009

Prof. Dr. Gerd Stumme, Wi.-Inf. Beate Krause

15. Juli 2009

1 Bibliometrische Maße

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?

2 HITS Algorithmus

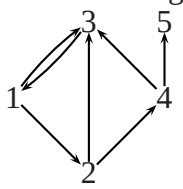
Betrachten Sie die folgenden Webseiten und die Menge der Webseiten, die diese verlinken.
Seite A zeigt auf Seite D und C. Seite B zeigt auf Seite D, C, und E.

1. Zeichnen Sie die zugehörige Adjazenzmatrix.
2. Berechnen Sie die ersten zwei Iterationen des HITS Algorithmus auf dem Teilwebgraphen. Notieren Sie die Autoritäts- und Hub Gewichte für jede Seite jeweils vor und nach der Normalisierung.
3. 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.

3 PageRank Algorithmus

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 Vergleich 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 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?

6 Recommender-Systeme

1. Erklären Sie mit eigenen Worten, was der Pearson-Korrelationskoeffizient aussagt!
2. Betrachten Sie die Bewertungen von Filmen durch die Benutzer Alice (A), Bob (B) und Charlie (C) gemäß der folgenden Tabelle:

Film	Alice	Bob	Charlie
Titanic	7	9	5
High Fidelity	5	7	5
American Beauty	5	7	5
Jarhead	4	6	4
Life of Brian	4	6	4
Three Kings	5		
A Fish Called Wanda			4

   Schätzen Sie – gemäß der vorigen Antwort – ab, wie die Größe der Korrelationskoeffizienten $c_{A,B}, c_{B,C}, c_{A,C}$ relativ zueinander aussehen wird!
3. Berechnen Sie die Korrelationskoeffizienten $c_{A,B}, c_{B,C}, c_{A,C}$!
4. Sagen Sie eine Bewertung der Filme “Three Kings” und “A Fish Called Wanda” durch den Anwender Bob voraus! Ziehen sie dazu jeweils den anderen Anwender heran, der den fraglichen Film bewertet hat. Sie können die Signifikanzgewichtung weglassen.