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?

   Lösungsvorschlag:

   Ko-Zitation und Kopplung entsprechen einander, wenn man für die Semantik der Kanten von “A zitiert B” auf “A wird zitiert von B” übergeht.

2. Auf welches der beiden Maße haben die Autoren der jeweiligen Schriften unmittelbaren Einfluß, auf welches nicht?

   Lösungsvorschlag:

   Die Kopplung können Autoren direkt beeinflussen, da sie selber festlegen, welche Arbeiten sie zitieren, und so auch, mit welchen Arbeiten sie eine große Kopplung haben. Die Ko-Zitation wird erst nach der Veröffentlichung und durch andere Autoren entschieden, so dass der Autor einer Arbeit hierauf keinen (unmittelbaren) Einfluss hat.

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?

   Lösungsvorschlag:

   Wie oben erläutert, ist es leicht, eine hohe Kopplung mit einer beliebigen anderen Schrift zu erreichen. Dagegen wäre eine hohe Ko-Zitation mit einem besonders hochwertigen Artikel ein Indiz dafür, dass die eigene Arbeit einen hohen Stellenwert besitzt.

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.

   Lösungsvorschlag:

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

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.

   Lösungsvorschlag:

   Iteration 1:

   $A = [0.0, 0.0, 2.0, 2.0, 1.0]$

   $H = [4.0, 5.0, 0.0, 0.0, 0.0]$

   Norm $A = [0.0, 0.0, 0.6666666666666666, 0.6666666666666666, 0.3333333333333333]$

   Norm $H = [0.6246950475544243, 0.7808688094430304, 0.0, 0.0, 0.0]$

   Iteration 2:

   $A = [0.0, 0.0, 1.4055638569974547, 1.4055638569974547, 0.7808688094430304]$

   $H = [2.8111277139949093, 3.5919965234379396, 0.0, 0.0, 0.0]$

   Norm $A = [0.0, 0.0, 0.6581451817144176, 0.6581451817144176, 0.36563621206356534]$
   Norm $H = [0.6163082616581106, 0.7875050010075858, 0.0, 0.0, 0.0]$

   Iteration 3:

   $A = [0.0, 0.0, 1.4038132626656963, 1.4038132626656963, 0.7875050010075858]$

   $H = [2.8076265253313926, 3.5951315263389785, 0.0, 0.0, 0.0]$

   Norm $A = [0.0, 0.0, 0.6572842735788008, 0.6572842735788008, 0.36872044615396143]$

Norm $H = [0.615498370759936, 0.7881381576804058, 0.0, 0.0, 0.0]$

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.

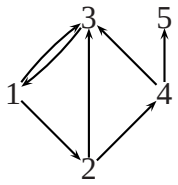
   Lösungsvorschlag:

   Da in jeder zweiten Iteration das Hub-Gewicht, das ein Hub abgibt, wieder an Hubs zurückgegeben wird, wird das Gewicht eines bestimmten Hubs bevorzugt bei jenen Hubs landen, mit denen er eine große Kopplung aufweist, d.h. mit denen er sich eine große Menge von verbundenen Authorities teilt. Die Kopplung ist also ein Indiz dafür, wo (indirekt) das meiste Gewicht zwischen Hubs fließt.

   Symmetrisch ist also die Ko-Zitation ein Maß dafür, wo zwischen Authorities das meiste Gewicht übertragen wird.

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?



   Lösungsvorschlag:

   Das ganze Gewicht wird bei Knoten 5 versickern, da dieser eine PageRank-Senke ("rank sink") ist und kein Gewicht wieder herausgibt.

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

   Lösungsvorschlag:

   Um möglichst viel Gewicht auf eigene Seiten zu ernten, betreiben manche "Optimierer" Linkfarmen, die nur dem Zweck dienen, durch starke Verlinkung untereinander ohne Links nach außen PageRank auf sich zu ziehen.

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).

   Lösungsvorschlag:

   $R_1 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})^T$
   $R_2 = (\frac{1}{4}, \frac{1}{8}, \frac{1}{2}, \frac{1}{8})^T$
   $R_3 = (\frac{1}{2}, \frac{1}{8}, \frac{5}{16}, \frac{1}{16})^T$
   $R_4 = (\frac{5}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{16})^T$
   $R_5 = (\frac{3}{8}, \frac{5}{32}, \frac{11}{32}, \frac{1}{8})^T$
   ...
   Der gesuchte Eigenvektor ist $\frac{1}{11}(4, 2, 4, 1)^T$.

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

   Lösungsvorschlag:

   Es würde sich z. B. anbieten, nach der Konvergenz des PageRank-Algorithmus' für die Rank Sinks noch eine Iteration zu machen. In diesem Fall würde also der Knoten 5 die Hälfte des Gewichts von Knoten 4 erhalten.

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!

   Lösungsvorschlag:

   Google Die Resultate werden per herkömmlicher Suche bestimmt und danach mit PageRank geordnet.

   HITS Es wird per Schlüsselwortsuche eine eingeschränkte Menge von Seiten bestimmt, auf denen dann graphbasiert gerankt wird.

   Personalisierter PageRank Der gesamte Dokumentenbestand wird nach Benutzervorlieben gerankt.

2. Welche der drei Verfahren sind für eine praktisch verwendbare Suchmaschine auf dem gesamten Web nutzbar, welche nicht? Warum?

   Lösungsvorschlag:

   Google Offenbar für Websuche nutzbar ;-), da nur *ein* PageRank offline berechnet werden muß, der für jede Anfrage verwendet werden kann.

   HITS Wird nicht praktisch benutzt, prinzipiell aber nutzbar, da nur kleine Dokumentenmenge in die Berechnung eingeht; diese Berechnung könnte online erfolgen.

   Personalisierter PageRank Da die PageRank-Berechnung viel länger dauert, als ein Benutzer bereit ist zu warten, und sie zudem online durchgeführt werden muss, ist diese Variante (derzeit) nicht für die Websuche einsetzbar.

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.

   Lösungsvorschlag:

   Man muß möglichst viele Links von möglichst hoch bewerteten Seiten auf die eigene Seite ziehen. Im Extremfall führt das zu sogenannten Link-Farmen, die nur dazu dienen, die eigenen Seiten höher zu bewerten.

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

   Lösungsvorschlag:

   Man kann z. B. ein Gegenmaß erfinden (den BadRank), der beschreibt, inwiefern eine Seite in “schlechte Nachbarschaften” verweist. Auch dieser kann in einer Fixpunktiteration zu einem Maß auf alle Seiten fortgesetzt 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?

   Lösungsvorschlag:

- Selbstzitierung (bei einfachen Maßen)
- Selbstplagiate (mehrfache Veröffentlichung von Versionen der gleichen Arbeit)
- Zitierkartelle (mehrere Autoren zitieren sich gegenseitig und erhöhen so gegenseitig ihren Stellenwert)

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

   Lösungsvorschlag:

   - Im Web kostet das Erstellen von Links oder das Aufsetzen von Webseite fast nichts, auch in großen Stückzahlen. Dagegen ist das publizieren einer gedruckten Schrift mit einem gewissen Aufwand verbunden.
   - Viele Webseiten, wie z. B. Blogs oder Social-Bookmarking-Sites, bieten Benutzern die Möglichkeit, Links zu setzen. Dadurch vereinfacht sich das Setzen von Links selbst unter fremdem Namen.
   - Bei den meisten Webangeboten gibt es keine redaktionelle Kontrolle oder unabhängige Begutachtung, im Gegensatz zu Druckerzeugnissen. Dadurch können Manipulationen besonders leicht umgesetzt werden.

6 Recommender-Systeme

1. Erklären Sie mit eigenen Worten, was der Pearson-Korrelationskoeffizient aussagt!

   Lösungsvorschlag:

   Der Pearson-Korrelationskoeffizient zeigt an, wie sehr zwei Bewertungen in den Abweichungen nach oben und unten vom Mittelwert in den jeweiligen Positionen übereinstimmen.

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!

Lösungsvorschlag:

Da Alice und Bob in den Abweichungen nach oben und unten übereinstimmen, sind sie maximal korreliert, also $c_{A,B} = 1$. Weniger stark wird die Korrelation von Alice und Charlie sein, da sie sich bei "Titanic" uneinig sind, obwohl die absoluten Werte nahe beieinander liegen. Die Koeffizienten $c_{A,C}$ und $c_{B,C}$ werden gleich sein, da die Bewertungen von A und B nur verschoben sind und somit die gleichen Abweichungen vom jeweiligen Mittelwert aufweisen.

3. Berechnen Sie die Korrelationskoeffizienten $c_{A,B}, c_{B,C}, c_{A,C}$!

   Lösungsvorschlag:

   Durch Einsetzen in die Formel ergeben sich $c_{A,B} = 1, c_{B,C} = 0.75, c_{A,C} = 0.75$.

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.

   Lösungsvorschlag:

   Three Kings: Alice hat diesen Film mit 5 bewertet. Da wir nur einen weiteren Anwender betrachten, kürzt sich das $w_{Bob,\cdot}$ heraus und man hat

   $p_{Bob,Three\ Kings} = \bar{r}_{Bob} + (r_{Alice,\ Three\ Kings} - \bar{r}_{Alice}) = 7 + (5 - 5) = 7$

   A Fish Called Wanda:

   $p_{Bob,\ Wanda} = \bar{r}_{Bob} + (r_{Charlie,\ Wanda} - \bar{r}_{Charlie}) = 7 + (4 - 4.6) = 6.4$