Protokoll zur Sitzung am 10.7.17 (Kolloquium)

Anastasiia Bespala. WebSuche: PageRank und HITS

Die Vortragende hat über 2 ähnliche Ansätze zur Bestimmung der Relevanz einer Webseite berichtet: PageRank und HITS.

Grundlage moderner Suchalgorithmen bleibt immer noch das Matchen der zu suchenden Begriffe mit einem möglichst großen Satz indizierter Internetseiten (z.B. durch tf.idf). Zufriedenstellend sind diese Treffer aber meistens nicht, hauptsächlich deswegen, weil zu viele Ergebnisse geliefert werden.

Beide Ansätze (PageRank und HITS) können helfen, Infos über die Relevanz der Webseiten zu erhalten: dies wird anhand der Linkstruktur des Internets ermittelt. Es wird davon ausgegangen, dass Betreiber der Webseiten andere Seiten verlinken, die sie selbst für besonders wichtig halten: die Anzahl der eingehenden Links wird also als Maßstab für Relevanz verwendet. Allerdings werden nicht alle Links als gleichwertig betrachtet: eine Seite ist um so wichtiger, je mehr und wichtigere Seite auf sie verlinken (PageRank), bzw. wird nur einen kleinen Teilgraphen des Internets betrachtet und jedem Dokument darin 2 Werte zugeschrieben - einen als *authority* und einen als *hub* (HITS). Ein guter *hub* verlinkt viele wichtige Seiten mit hohem *authority* Wert und eine gute *authority* ist eine Seite, die von guten *hubs* verlinkt wird.

Nachdem man diese Ansätze kurz vorgestellt hatte, wurden die etwas detaillierter besprochen. U. A. wurde mit Hilfe Herrn Schulz die Formel, die das PageRank-System mathematisch beschreibt, erklärt:

$$R(u) := c \sum_{v \in B_u} \frac{R(v)}{N_v} + cE(u)$$

Hier wird das Internet als einen endlichen, gerichteten Graphen dargestellt: $G = (V, E)$ mit $V = v_1, ..., v_n$ und $E \subseteq V \times V$. $(v_i, v_j) \in E$ gdw. Webseite v_i hat einen Link auf Internetseite v_i. Die Menge der Seiten, auf die u einen Link hat ist $F_u = \{w|(u, w) \in E\}$, die Menge der Seiten, die einen Link auf u haben ist $B_u = \{w|(w, u) \in E\}$. Außerdem $N_u = |F_u|$ - die Anzahl der Links von u. c ist ein Normalisierungsfaktor. $E(u)$ ist dabei ein Faktor für alle Webseiten. Dieser Faktor simuliert das Verhalten eines **nicht-***Random Surfers* (Dieser startet bei einer Seite und klickt wahllos auf irgendwelche Links. Ein realer Surfer sucht aber eine andere Webseite auf und hält sich nicht in einer Schleife von wenigen Webseiten).

Es wurde kurz ein Rechenbeispiel zu diesem Modell gezeigt. Danach hat die Vortragende über das HITS-System gesprochen, das sehr ähnlich dem PageRank ist, reduziert allerdings die Betrachtung au einen aussagekräftigen Subgraphen. Um die Idee zu erläutern, hat die Vortragende versucht ein anderes Rechenbeispiel zu erklären.

Am Ende des Vortrages wurden die beiden Ansätze kurz miteinander verglichen:

- Die entsprechenden Matrizen sind nicht gleich groß (PageRank - enorm groß, kann nicht zur Laufzeit berechnet werden; HITS - kleiner Subgraph des Internets, wird zur Laufzeit berechnet).
- Beide Verfahren sind für Manipulationsversuche anfällig (z.B. Bannerwerbung; wirkt sich bei PageRank weniger stark).
- Beide Methoden sind anfällig für Abschweifungen vom eigentlichen Thema, PageRank funktioniert sogar absolut unabhängig von der Suchanfrage.