

Web-Informationssysteme, WS 2009/10

## Seminar-Übung 3: PageRank, HITS, Tabellarische Daten, Inverted Files

*Besprechung am Mi 11.11.2009*

— SÜ-3.1 —

### Was Ihr wissen müßt über “PageRank & HITS”

Wiederholung

#### PAGERANK

- Was ist und wie wird die **Google-Matrix** (oder PageRank-Übergangsmatrix) berechnet?
- Zeigen Sie das es für diese Matrix einen eindeutigen Eigenwert  $> 0$  gibt mit maximalem Betrag unter allen Eigenwerten der Matrix.
- Skizzieren Sie wie der zugehörige Eigenvektor iterativ berechnet werden kann!
- Die Google-Matrix wird aus der *transponierten* Adjazenzmatrix gebildet. Zeigen Sie wie man den obigen Eigenvektor iterativ berechnen kann ohne diese Transponierung.

#### HITS

- Was ist das wesentliche Unterscheidungsmerkmal des HITS- im Vergleich zum PageRank-Algorithmus.
- Wie wird der Anfangsgraph im HITS Algorithmus berechnet?
- Skizzieren Sie den iterativen HITS Algorithmus. Erklären Sie dabei die beiden Update-Operationen.

— SÜ-3.2 —

## Extraktion von Web-Tabellen

Vertiefung

Gatterbauer, W., Bohunsky, P., Herzog, M., Krüpl, B., and Pollak, B. 2007. *Towards domain-independent information extraction from web tables*. In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM, New York, NY, 71-80.

<http://www2007.org/papers/paper790.pdf> oder <http://doi.acm.org/10.1145/1242572.1242583> oder Anlagen Gatterbauer2007WebTables.pdf.

*“Traditionally, information extraction from web tables has focused on small, more or less homogeneous corpora, often based on assumptions about the use of <table> tags. A multitude of different HTML implementations of web tables make these approaches difficult to scale. In this paper, we approach the problem of domain-independent information extraction from web tables by shifting our attention from the tree-based representation of web pages to a variation of the two-dimensional visual box model used by web browsers to display the information on the screen. The thereby obtained topological and style information allows us to fill the gap created by missing domain-specific knowledge about content and table templates. We believe that, in a future step, this approach can become the basis for a new way of large-scale knowledge acquisition from the current ‘Visual Web.’”*

### FRAGENKATALOG

1. Was versteht man unter “web data extraction” (oder “Web information extraction”)?
2. Warum spielen gerade Tabellen eine wichtige Rolle bei der *web data extraction*?
3. Was unterscheidet den im Paper vorgestellten Ansatz von typischen Ansätzen auf dem Gebiet? Warum wählen die Autoren diesen Ansatz?
4. Fallen Ihnen Gegenargumente gegen den von den Autoren vorgeschlagenen Ansatz ein?
5. Beschreiben Sie grob, die Schritte des Extraktions-Algorithmus! Kommentieren Sie die Annahmen, die diesem Algorithmus zugrunde liegen.
6. Wie wird das Ergebnis der Tabellenextraktion in diesem Ansatz dargestellt?

— SÜ-3.3 —

## Inverted Files im Information Retrieval

Vertiefung, Wiederholung

Zobel, J. and Moffat, A. 2006. *Inverted files for text search engines*. ACM Comput. Surv. 38, 2 (Jul. 2006), 6. Für die Beantwortung der Fragen ist **nur Abschnitt 1–3 notwendig** (Seiten 1–12).

<http://doi.acm.org/10.1145/1132956.1132959> oder <http://www.cs.mu.oz.au/~jz/fulltext/compsurv06.pdf> oder Anlagen Zobel2006InvertedFiles.pdf.

*“The technology underlying text search engines has advanced dramatically in the past decade. The development of a family of new index representations has led to a wide range of innovations in index storage, index construction, and query evaluation. While some of these developments have been consolidated in textbooks, many specific techniques are not widely known or the textbook descriptions are out of date. In this tutorial, we introduce the key techniques in the area, describing both a core implementation and how the core can be enhanced through a range of extensions. We conclude with a comprehensive bibliography of text indexing literature.”*

### FRAGENKATALOG

1. *Once more with feeling*: Erläutern Sie den Unterschied zwischen *matching* (und damit Antworten) in relationalen Datenbanken und in Information Retrieval Systemen.
2. Erklären Sie welche Rolle Ähnlichkeitsmaße in einem IR-System spielen.
3. Wie muss bei großen Dokumentsammlungen das Ähnlichkeitsmaß aussehen, um zu vermeiden, dass alle Dokumente betrachtet werden müssen.
4. Welche Rolle spielt eine Dokument-Index in einem IR-System?
5. Erklären Sie das Prinzip eines *inverted files* (invertierten Index)! Warum heißt dieser Index *invertiert*?
6. Skizzieren Sie wie mit Hilfe von *inverted files* eine Anfrage aus mehreren Termen bearbeitet wird (z.B. “dark keep night”).