

Seminar „Suchmaschinen“
WS 2007/08

Martin Sauerhoff

Funktionsweise von Suchmaschinen –
hier im Wesentlichen theoretische Aspekte.

- Aufbau von Suchmaschinen (Grundlagen).
- Ranking mittels Linkanalyse (lineare Algebra).
- Analyse des Webs: Änderungshäufigkeit von Seiten, Samplingverfahren (statistische Methoden).
- Random Walks, Hashing (randomisierte Algorithmen).

- Einarbeitung in das Thema:
vorgegebene Literatur + eigene Recherche;
- Aufbereitung des Themas für den Vortrag:
 - Stoffauswahl;
 - evtl. Anpassung der Notation;
 - Aufbereiten/Kürzen von Beweisen;
- Erstellen einer Kurzzusammenfassung;
- Präsentation.

- 90 Minuten Zeit insgesamt, 10 Minuten für Fragen;
Termin: Do., 12:15 – 13:45 Uhr, R. 205, OH 16.
- Ziel: Verständlichkeit für [Seminarteilnehmer](#);
- technisch einwandfreie Folienpräsentation mit Beamer.

- 2 Seiten, erstellt mit \LaTeX ;
- Abgabe bei mir bis spätestens 2 Wochen vor dem eigenen Vortrag, Einarbeitung von Korrekturen;
- wird vor dem Vortrag an die Teilnehmer verteilt.

1. **Klassisches Information-Retrieval:**

R. Baeza-Yates, B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999. Auswahl aus den Kapiteln 2, 3, 7 und 8.

D. Lewandowski. Web Information Retrieval. *Information: Wissenschaft und Praxis*, 56(1):5–12, 2005.

Themen: Mathematische Modelle für Dokumente, Bewertung von Retrieval, Ähnlichkeitsmaße, Indizierung, Textaufbereitung, Vergleich Klassisches IR – Web-IR.

2. **Suchmaschinen-Grundlagen:**

A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, S. Raghavan. Searching the Web. *ACM Trans. on Internet Technology*, 1(1):2-43, 2001.

3. **Crawler-Architektur am Beispiel:**

A. Heydon, M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.

A. Heydon, M. Najork. High-performance web crawling. Technischer Bericht, SRC-RR-173, 2001.

4. **Konvergenz von Ranking-Algorithmen:**

A. Farahat, T. Lofaro, J. C. Miller, G. Rae, L. A. Ward. Authority rankings from HITS, PageRank, and SALSA: Existence, uniqueness, and effect of initialization. *SIAM Journal on Scientific Computing*, 27(4):1181–1201, 2006.

5. **PageRank im Detail:**

A. N. Langville und C. D. Meyer. Deeper inside PageRank. *Internet Mathematics*, 3(1):335–380, 2003.

6. **HITS- und PageRank-Varianten:**

K. Bharat, M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. of 21st SIGIR*, 104–111, 1998.

T. H. Haveliwala. Topic-sensitive PageRank. In *Proc. of 11th WWW Conference*, 517–526, 2002.

7. **Automatische Spamererkennung 1:**

Z. Gyöngyi, H. Garcia-Molina, J. Pedersen. Combating Web spam with TrustRank. In *Proc. of 30th VLDB*, 576–587, 2004.

8. **Automatische Spamererkennung 2:**

Z. Gyöngyi, H. Garcia-Molina, P. Berkhin, J. Pedersen. Link spam detection based on mass estimation. In *Proc. of 32nd VLDB*, 439–450, 2006.

9. **Schätzung der Änderungshäufigkeit von Webseiten:**

J. Cho, H. Garcia-Molina. Estimating frequency of change. *ACM Transactions on Internet Technology (TOIT)*, 3(3):256–290, 2003.

10. **Rabin-Fingerprinting und Min-Hashing:**

A. Z. Broder. Some applications of Rabin's fingerprinting method. In *Sequences II: Methods in Communications, Security, and Computer Science*, R. Capocelli, A. De Santis and U. Vaccaro (Hrsg.), 143–152. Springer-Verlag, 1993.

A. Broder. On the resemblance and containment of documents. In *Proc. of Compression and Complexity of Sequences*, 21–29, 1997.

11. **Seitenvergleich mit randomisierten Projektionen:**
M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. of 34th STOC*, 380–388, 2002.
12. **Sampling von Webseiten mit Random Walks:**
Z. Bar-Yossef, A. Berg, S. Chien, J. Fakcharoenphol, D. Weitz. Approximating aggregate queries about web pages via random walks. In *Proc. of 26th VLDB*, 535–544, 2000.

- 13. Anfragebasiertes Sampling von Webseiten:**
Z. Bar-Yossef und M. Gurevich. Random sampling from a search engine's index. In *Proc. of 15th WWW Conference*, 367–376, 2006.
- 14. Finden von Webgemeinden:**
Y. Dourisboure, F. Geraci, M. Pellegrini. Extraction and classification of dense communities in the Web. In *Proc. of 16th WWW Conference*, 461–470, 2007.

Literatur (ohne Buchkapitel für Thema 1):

<http://ls2-www.cs.uni-dortmund.de/sauerhof/sm0708/lit>

Wichtige Quelle für Hintergrundinfos:

Folien zu „Internet-Algorithmen“, Sommer 2007.

<http://ls2-www.cs.uni-dortmund.de/lehre/sommer2007/ia>

Vortragszuordnung demnächst auf der Seminar-Homepage:

<http://ls2-www.cs.uni-dortmund.de/sauerhof/sm0708/lit>

Erster Vortrag: 8.11.2007, Raum 205, OH 16.