# 236621 Search Engine Technology Syllabus

Ronny Lempel, Yahoo! Labs

Winter 2011/12

## 1 Course Objectives

The objectives of the course are to present the theoretical and engineering aspects of search engine technology. The course will cover the data structures and algorithms that are in use in the major components of large scale search engines, as well as characteristics and attributes of the Web corpus and search engine users. Students of the course should exit it with an understanding of both the machinery and ecosystem of search engines.

## 2 Intended Audience and Prerequisites

The course's intended audience includes 4'th year BSc students and graduate students. Its prerequisites are as follows:

- Basic course in Algebra (e.g 104167)
- Basic course in Probability (e.g. 094412)
- Basic course on Data Structures (e.g. 234218)
- Basic course on Algorithms (e.g. 234247)

While the formal prerequisites may be satisfied by 3rd year students, the course requires some mathematical and analytical maturity that 3rd year students typically lack. In particular, basic understanding of stochastic processes (e.g. as taught in 094314 or 044202) is highly recommended.

## 3 Course Structure, Grading and Syllabus

The course consists of 14 weekly lectures (2 hours) and tutorials (1 hours), and is divided into 5 main parts. Grading is based on three regular homework assignments, one of which involves programming, and a final assignment of larger scope that may also involve programming. Some assignments or portions thereof are submitted in pairs, and some are submitted individually.

## 3.1 Introduction to Search Engines and Information Retrieval

This part is based on overview papers [20, 7] and textbooks [62, 12, 53]. A first (dry) homework assignment will be given after this part.

| Week | Lecture | Tutorial |
|---|---|---|
| 1 | Course outline, popular introduction to search engines, technical overview of search engine components | Introduction to Information Retrieval: Boolean model, vector space model, TF/IDF scoring |
| 2 | Probabilistic IR, Neyman-Pearson Lemma | Language models in Information Retrieval |

## 3.2 Inverted Indices

| Week | Lecture | Tutorial |
|---|---|---|
| 3 | Basics: what is an inverted index, how is one constructed efficiently, what operations does it support, what extra payload is usually stored in search engines, the accompanying lexicon | B-Tree lexicon, Min-Heap [30] |
| 4 | Query evaluation schemes: term-at-a-time vs. doc-at-a-time, result heaps, early termination/pruning, WAND [22] | Identification of near-duplicate pages [25] |
| 5 | Index compression [63, 6] and document reordering [61, 65] | The Apache Lucene search library (prerequisite for homework) |
| 6 | Distributed index architectures: global/local schemes [7, 55, 26], combinatorial issues stemming from the distribution of data [49], the Google cluster architecture [14] | |

A second (wet) homework assignment will be given after this part, involving changes to Apache Lucene.

### 3.3 The Web's graph and Link Analysis

| Week | Lecture | Tutorial |
|---|---|---|
| 6 | | Web graph structure: power laws [56], Bow-tie structure [23], self-similarity [32] |
| 7 | Link Analysis basics: Google's PageRank [20], Kleinberg's HITS [42], with some quick overview of Perron-Frobenius theory and ergodicity [35] | Topic-sensitive PageRank [38], SALSA [48] |
| 8 | Stability and similarity of link-based schemes [58, 19, 28, 18, 51], the TKC Effect [48] | Evolutionary models of the Web graph [43, 46] |

### 3.4 Infrastructure Beyond the Index

| Week | Lecture | Tutorial |
|---|---|---|
| 9 | Crawlers - purpose and architecture [40, 47], optimizing crawl order [29, 64, 57], computation of importance metrics during crawl [1] | Bloom Filters [24] |
| 10 | Effective caching and prefetching of query results [54, 50, 49, 11, 9, 33] | |

A third (dry) homework assignment will be given after this part.

### 3.5 Users and Advertising

The computational advertising tutorials and lectures will be based on the pioneering course on this subject taught at Stanford University, `http://www.stanford.edu/class/msande239/`.

| Week | Lecture | Tutorial |
|---|---|---|
| 10 | | Computational advertising: models and definitions. CPM, CPC, CPA; sponsored search (adwords), content match (adsense), display advertising |
| 11 | Computational advertising: auction mechanisms [59, 3, 34, 60] | Computational advertising (TBD) |
| 12 | Mining and tapping implicit user generated content [4, 17] | Query log analysis [21, 39, 8, 41] |
| 13 | Task Completion and Search Assistance - from spell corrections and simple shortcuts to rich media, mashups, query completions and facets [27, 16, 13] | Query suggestions [17, 10, 66, 15] |
| 14 | The Long Tail [5], recommender systems and collaborative filtering [2, 44, 52, 36] | Context-aware search and user modeling [45, 37] |

A fourth homework assignment will be given after this part, also covering the Map-Reduce framework [31].

# References


[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proc. 12th International WWW Conference (WWW2003), Budapest, Hungary*, pages 280–290, 2003.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[3] G. Aggarwal, S. M. Muthukrishnan, D. Pal, and M. Pal. General auction mechanisms for search advertising. In *Proc. 18th International World Wide Web Conference (WWW'2009)*, pages 241–250, April 2009.

[4] S. Amer-Yahia, M. D. Choudhury, M. Feldman, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *Proc. 21st ACM Conference on Hypertext and Hypermedia (Hypertext'2010)*, pages 35–44, June 2010.

[5] C. Anderson. *The Long Tail - Why the Future of Business is Selling Less of More.* Hyperion Books, New York NY, 2006.

[6] V. Anh and A. Moffat. Index compression using fixed binary codewords. In *Proc. of the 15th Int. Australasian Database Conference*, pages 61–67, 2004.

[7] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, 2001.

[8] R. Baeza-Yates. Graphs from search engine queries. In *Proc. 33rd conference on Current Trends in Theory and Practice of Computer Science*, SOFSEM'07, pages 1–8, 2007.

[9] R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. The Impact of Caching on Search Engines. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2007. ACM Press.

[10] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Improving search engines by query clustering. *J. Am. Soc. Inf. Sci. Technol.*, 58:1793–1804, October 2007.

[11] R. Baeza-Yates, F. Junqueira, V. Plachouras, and H. F. Witschel. Admission Policies for Caches of Search Engine Results. In *SPIRE*, 2007.

[12] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison Wesley, 1999.

[13] Z. Bar-Yossef and M. Gurevich. Mining search engine query logs via suggestion sampling. In *Proc. 34th International Conference on Very Large Data Bases (VLDB 2008)*, pages 54–65, August 2008.

[14] L. A. Barroso, J. Dean, and U. Hölzle. Web search for a planet: The google cluster architecture. *IEEE Micro*, 23(2):22–28, April 2003.

[15] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proc. 6th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD'00, pages 407–416, 2000.

[16] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev. Beyond basic faceted search. In *Proc. 1st ACM Conference on Web Search and Data Mining (WSDM'2008)*, pages 33–43, February 2008.

[17] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proc. 17th ACM Conference on Information and Knowledge Management (CIKM'2008)*, pages 609–618, October 2008.

[18] A. Borodin and H. C. Lee. When the hyperlinked environment is perturbed. In *Ninth International Computing and Combinatorics Conference (COCOON)*, Big Sky, Montana, USA, 2003.

[19] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *Proc. 10th International World Wide Web Conference*, pages 415–429, May 2001.

[20] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th International WWW Conference*, pages 107–117, 1998.

[21] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[22] A. Broder, D. Carmel, M. Herscovichi, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Twelfth International Conference on Information and Knowledge Management (CIKM 2003), New Orleans, LA, USA*, pages 426–434, November 2003.

[23] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proc. 9th International WWW Conference*, pages 309–320, 2000.

[24] A. Broder and M. Mitzenmacher. Network applications of bloom filters: A survey. *Internet Mathematics*, 1(4):485–509, 2004.

[25] A. Z. Broder, S. C. Glassman, and M. S. Manasse. Syntactic clustering of the web. In *Proc. 6th International WWW Conference*, 1997.

[26] B. Cahoon, K. S. McKinley, and Z. Lu. Evaluating the performance of distributed architectures for information retrieval using a variety of workloads. *ACM Transactions on Information Systems*, 18(1):1–43, 2000.

[27] D. Chakrabarti, R. Kumar, and K. Punera. Quicklink selection for navigational query results. In *Proc. 18th International World Wide Web Conference (WWW'2009)*, pages 391–400, April 2009.

[28] S. Chien, C. Dwork, R. Kumar, D. Simon, and D. Sivakumar. Link evolution: Analysis and algorithms. In *Workshop on Algorithms and Models for the Web Graph (WAW)*, Vancouver, Canada, 2002.

[29] J. Cho, H. García-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172, 1998.

[30] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Second Edition.* The MIT Press, 2001.

[31] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[32] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, August 2002.

[33] T. Fagni, R. Perego, F. Silvestri, and S. Orlando. Boosting the Performance of Web Search Engines: Caching and Prefetching Query Results by Exploiting Historical Usage Data. *ACM Trans. Inf. Syst.*, 24(1):51–78, 2006.

[34] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *Am. Math Monthly*, 69(1):9–15, 1962.

[35] R. G. Gallager. *Discrete Stochastic Processes*. Kluwer Academic Publishers, 1996.

[36] N. Golbandi, Y. Koren, and R. Lempel. Adaptive bootstrapping of recommender systems using decision trees. In *Proc. 4th ACM Conference on Web Search and Data Mining (WSDM'2011)*, pages 595–604, February 2011.

[37] A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *Proc. 3rd ACM international conference on Web search and data mining*, WSDM'10, pages 221–230, 2010.

[38] T. H. Haveliwala. Topic-sensitive pagerank. In *Proc. 11th International WWW Conference (WWW2002)*, 2002.

[39] D. He, A. Göker, and D. J. Harper. Combining evidence for automatic web session identification. *Inf. Process. Manage.*, 38:727–742, September 2002.

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

[41] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. 17th ACM conference on Information and knowledge management*, CIKM'08, pages 699–708, 2008.

[42] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:5:604–632, 1999.

[43] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The web as a graph: Measurements, models and methods. In *Proc. of the Fifth International Computing and Combinatorics Conference*, pages 1–17, 1999.

[44] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[45] R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar. Searching with context. In *Proc. 15th international conference on World Wide Web*, WWW'06, pages 477–486, 2006.

[46] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. S. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41nd Annual Symposium on Foundations of Computer Science (FOCS 2000), Redondo Beach, California*, pages 57–65, 2000.

[47] H.-T. Lee, D. Leonard, X. Wang, and D. Loguinov. Irlbot: Scaling to 6 billion pages and beyond. In *Proc. 17th International World Wide Web Conference (WWW'2008)*, pages 427–436, April 2008.

[48] R. Lempel and S. Moran. SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2):131–160, April 2001.

[49] R. Lempel and S. Moran. Optimizing result prefetching in web search engines with segmented indices. In *Proc. 28th International Conference on Very Large Data Bases, Hong Kong, China*, pages 370–381, 2002.

[50] R. Lempel and S. Moran. Predictive caching and prefetching of query results in search engines. In *Proc. 12th World Wide Web Conference (WWW2003), Budapest, Hungary*, pages 19–27, May 2003.

[51] R. Lempel and S. Moran. Rank-stability and rank-similarity of link-based web ranking algorithms in authority-connected graphs. Technical Report 2, 2005.

[52] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[53] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[54] E. P. Markatos. On caching search engine query results. In *Proceedings of the 5th International Web Caching and Content Delivery Workshop*, May 2000.

[55] S. Melnik, S. Raghavan, B. Yang, and H. Garcia-Molina. Building a distributed full-text index for the web. In *Proc. 10th International WWW Conference*, 2001.

[56] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Invited Talk in the 39th Annual Allerton Conference on Communication, Control and Computing*, October 2001.

[57] M. Najork and J. L. Wiener. Breast-first search crawling yields high-quality pages. In *Proc. 10th International World Wide Web Conference (WWW'2001)*, pages 114–118, May 2001.

[58] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Stable algorithms for link analysis. In *Proc. 24'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 258–266, 2001.

[59] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[60] L. S. Shapley and M. Shubik. The assignment game 1: The core. *International Journal of Game Theory*, 1(1):111–130, 1971.

[61] F. Silvestri. Sorting out the document identifier assignment problem. In *Proc. of 29th European Conference on Information Retrieval (ECIR'07)*, pages 101–112, Rome, Italy, Apr. 2-5 2007.

[62] C. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[63] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, second edition, 1999.

[64] J. L. Wolf, M. S. Squillante, P. S. Yu, J. Sethuraman, and L. Ozsen. Optimal crawling strategies for web search engines. In *Proc. 11th International World Wide Web Conference (WWW2002)*, pages 136–147, 2002.

[65] H. Yan, S. Ding, and T. Suel. Inverted index compression and query processing with optimized document ordering. In *Proc. 18th International World Wide Web Conference (WWW'09)*, Madrid, Spain, Apr. 20-24 2009.

[66] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proc. 15th international conference on World Wide Web*, WWW'06, pages 1039–1040, 2006.