

SIGIR 2008 Tutorial : Web Mining for Search: Bibliography

Ricardo Baeza-Yates and Rosie Jones, Yahoo! Research

1 Web search and Web Mining

- characteristics of the Web estimates of size: [17] [9] [11] [31]
- Link analysis [57] [41] [18] [30] [47] [52] [71] [10] [33] [46] [53]
- modeling Web query distribution [64] [65, 37] [58].
- architecture of a Web search engine [15] [6]
- evaluation [38]
data collections available [26] [69] [24]

2 Web Crawling and Indexing

- identify page change frequency [55] [8]
- identify web pages with high page-rank / in-degree [15] [35]
- duplicate detection: shingling [14] [36]
- language identification [25]
- mining for index layout [7]
- spam detection (link and content based) [34] [32] [54] [2]

3 Query Processing and Ranking

- query taxonomies [16] [61] [62]
- mining queries and clicks for caching answers/index [12] [4]
- ranking as a machine learning problem [19] [48]
- mine clicks to infer relevance judgements / preference judgements [39] [23] [22]

4 User Interface

- generating result summary snippets [67] [68] [70] [27]
- generating query recommendations [29] [5] [40]
- spelling collection: constructing language models, mining spell corrections from logs [63] [13] [28]
- mining stemming [60] [59]
- mining query intention [3]

5 Other Retrieval Related Tasks

- mining anchor text [43] [49]
- identify multi-word phrases from query logs (proximity and phrase queries [51])
- identify named-entities [56]
- stream of queries is very large: streaming algorithms for calculating most-frequent, etc. [50] [20]

6 Advertising

- identify related keywords [21]
- page-template identification [42]
- keyword extraction [45] [66]

7 Discussion

- privacy issues in web search data [1] [44]
- future trends and challenges

References

- [1] E. Adar. User 4xxxxx9: Anonymizing query logs. In *Query Log Workshop, WWW2007*, 2007.
- [2] R. A. Baeza-Yates, P. Boldi, and C. Castillo. Generalizing pagerank: damping functions for link-based ranking algorithms. In *SIGIR*, pages 308–315, 2006.
- [3] R. A. Baeza-Yates, L. Calderón-Benavides, and C. N. González-Caro. The intention behind web queries. In *SPIRE*, pages 98–109, 2006.
- [4] R. A. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. The impact of caching on search engines. In *SIGIR*, pages 183–190, 2007.
- [5] R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza. Query clustering for boosting web page ranking. In *AWIC*, pages 164–175, 2004.
- [6] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
- [7] R. A. Baeza-Yates and F. Saint-Jean. A three level search engine index based in query log distribution. In *SPIRE*, pages 56–65, 2003.
- [8] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: towards an understanding of the web’s decay. In *WWW*, pages 328–337, 2004.
- [9] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine’s index. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 367–376, New York, NY, USA, 2006. ACM.
- [10] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

- [11] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. *Comput. Netw. ISDN Syst.*, 30(1-7):379–388, 1998.
- [12] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: evidence and implications. In *INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 126–134, 1999.
- [13] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. In *ACL*, 2000.
- [14] S. Brin, J. Davis, and H. Garcia-Molina. Copy detection mechanisms for digital documents. In *SIGMOD Conference*, pages 398–409, 1995.
- [15] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.
- [16] A. Z. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
- [17] A. Z. Broder, M. Fontoura, V. Josifovski, R. Kumar, R. Motwani, S. U. Nabar, R. Panigrahy, A. Tomkins, and Y. Xu. Estimating corpus size via queries. In *CIKM*, pages 594–603, 2006.
- [18] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.
- [19] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.
- [20] L. S. Buriol, G. Frahling, S. Leonardi, A. Marchetti-Spaccamela, and C. Sohler. Counting triangles in data streams. In *PODS '06: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 253–262, New York, NY, USA, 2006. ACM.
- [21] J. J. Carrasco, D. C. Fain, K. J. Lang, and L. Zhukov. Clustering of bipartite advertiser-keyword graph. In *International Conference on Data Mining*, 2003.
- [22] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there. In *ECIR*, pages 16–27, 2008.
- [23] B. Carterette and R. Jones. Evaluating search engines by modeling the relationship between relevance and clicks. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 217–224. MIT Press, Cambridge, MA, 2008.
- [24] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.
- [25] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994.
- [26] J. Cho, H. Garcia-Molina, T. Haveliwala, W. Lam, A. Paepcke, S. Raghavan, and G. Wesley. Stanford webbase components and applications. *ACM Trans. Interet Technol.*, 6(2):153–186, 2006.
- [27] C. L. A. Clarke, E. Agichtein, S. T. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *SIGIR*, pages 135–142, 2007.
- [28] S. Cucerzan and E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *EMNLP*, 2004.
- [29] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *WWW*, pages 325–332, 2002.

- [30] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.
- [31] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW (Special interest tracks and posters)*, pages 902–903, 2005.
- [32] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *AIRWeb*, pages 39–47, 2005.
- [33] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Databases*, pages 576–587. Morgan Kaufmann, 2004.
- [34] Z. Gyöngyi, H. Garcia-Molina, and J. O. Pedersen. Combating web spam with trustrank. In *VLDB*, pages 576–587, 2004.
- [35] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, pages 517–526, 2002.
- [36] N. Heintze. Scalable document fingerprinting. In *1996 USENIX Workshop on Electronic Commerce*, November 1996.
- [37] B. J. Jansen, A. Spink, and J. O. Pedersen. A temporal comparison of altavista web searching. *JASIST*, 56(6):559–570, 2005.
- [38] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.
- [39] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.
- [40] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW*, pages 387–396, 2006.
- [41] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [42] A. Kolcz and W. tau Yih. Site-independent template-block detection. In *PKDD*, pages 152–163, 2007.
- [43] R. Kraft and J. Zien. Mining anchor text for query refinement. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 666–674, New York, NY, USA, 2004. ACM.
- [44] R. Kumar, J. Novak, B. Pang, and A. Tomkins. On anonymizing query logs via token-based hashing. In *WWW*, pages 629–638, 2007.
- [45] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. A. Ribeiro-Neto. Learning to advertise. In *SIGIR*, pages 549–556, 2006.
- [46] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):387–401, 2000.
- [47] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.
- [48] T.-Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *SIGIR 2007 workshop: Learning to Rank for Information Retrieval*, 2007.
- [49] W.-H. Lu, L.-F. Chien, and H.-J. Lee. Anchor text mining for translation of web queries. In *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 401–408, Washington, DC, USA, 2001. IEEE Computer Society.

- [50] G. S. Manku and R. Motwani. Approximate frequency counts over data streams. In *VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases*, pages 346–357. VLDB Endowment, 2002.
- [51] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR*, pages 472–479, 2005.
- [52] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2), 2004.
- [53] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Stable algorithms for link analysis. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 258–266, New York, NY, USA, 2001. ACM.
- [54] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *WWW*, pages 83–92, 2006.
- [55] C. Olston and S. Pandey. Recrawl scheduling based on information longevity. In *WWW*, pages 437–446, 2008.
- [56] M. Paşca. Weakly-supervised discovery of named entities using web search queries. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 683–690, New York, NY, USA, 2007. ACM.
- [57] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
- [58] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *The First International Conference on Scalable Information Systems*, 2006.
- [59] F. Peng, N. Ahmed, X. Li, and Y. Lu. Context sensitive stemming for web search. In *SIGIR*, pages 639–646, 2007.
- [60] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [61] D. Rose and D. Levinson. Understanding user goals in web search. In *WWW 2004*, 2004.
- [62] M. Sahami. Mining the web to determine similarity between words, objects, and communities. In *FLAIRS Conference*, pages 14–19, 2006.
- [63] C. Shannon. Communication in the presence of noise. 1949.
- [64] C. Silverstein, M. R. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.
- [65] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, March 2002.
- [66] W. tau Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *WWW*, pages 213–222, 2006.
- [67] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR*, pages 2–10, 1998.
- [68] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams. Fast generation of result snippets in web search. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134, New York, NY, USA, 2007. ACM.

- [69] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.
- [70] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Learning query-biased web page summarization. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 555–562, New York, NY, USA, 2007. ACM.
- [71] U. G. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character.*, 213:21–87, 1925.