Web-scale information retrieval and data mining
List of papers

Charles Elkan
elkan@cs.ucsd.edu

April 7, 2008

This document is an incomplete and tentative list of papers that students may choose to present for CSE 291 in Spring 2008.

Recommendation systems. [DDGR07].

Collaborative filtering. [BK07, BKV07, Gor06].

Search engines. [Hav03, SMBR04, LLC05, SH06, NWM07].

Focused web crawling. [RM99].

Sponsored search. [BMS06, Tuz06, RB07, AS07, EO07, LN07, Rad07].

Search engine optimization. [XL06, Kö6].

Adversarial classification. [DDM+04, DS05].

Online experimentation. [KHS07, LP07].

Topic models for documents. [BNJ03, SG07].

Multilabel classifier learning. [GM05].

Customer value optimization. [SSN06].

Social networking. [LAH07].

References

[AS07] Kamal Ali and Mark Scarr. Robust methodologies for modeling web click distributions. In Williamson et al. [WZPSS07], pages 511–520.

[BK07] Robert Bell and Yehuda Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 43–52, 2007.

[BKV07] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 95–104, 2007.

[BMS06] Kevin Bartz, Vijay Murthi, and Shaji Sebastian. Logistic regression and collaborative filtering for sponsored search term recommendation. In *Proceedings of the Second Workshop on Sponsored Search Auctions*, 2006.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[DDGR07] Abhinandan Das, Mayur Datar, Ashutosh Garg, and ShyamSundar Rajaram. Google news personalization: scalable online collaborative filtering. In Williamson et al. [WZPSS07], pages 271–280.

[DDM+04] Nilesh N. Dalvi, Pedro Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *KDD*, pages 99–108. ACM, 2004.

[DS05] Isabel Drost and Tobias Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo, editors, *ECML*, volume 3720 of *Lecture Notes in Computer Science*, pages 96–107. Springer, 2005.

[EO07] Benjamin Edelman and Michael Ostrovsky. Strategic bidder behavior in sponsored search auctions. *Decision Support Systems*, 43(1):192–198, 2007.

[GM05] Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, pages 195–200, 2005.

[Gor06] Genevieve Gorrell. Generalized Hebbian algorithm for incremental singular value decomposition in natural language processing. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy*, pages 97–104, 2006.

[Hav03] Taher H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, 2003.

[KÖ6] Johan Köhne. Optimizing a large dynamically generated website for search engine crawling and ranking. Master's thesis, Technical University of Delft, Netherlands, 2006.

[KHS07] Ron Kohavi, Randal M. Henne, and Dan Sommerfield. Practical guide to controlled experiments on the web: Listen to your customers not to the HiPPO. In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, *KDD*, pages 959–967. ACM, 2007.

[LAH07] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), 2007.

[LLC05] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In Allan Ellis and Tatsuya Hagino, editors, *WWW*, pages 391–400. ACM, 2005.

[LN07] Yury Lifshits and Dirk Nowotka. Estimation of the click volume by large scale regression analysis. In Volker Diekert, Mikhail V. Volkov, and Andrei Voronkov, editors, *CSR*, volume 4649 of *Lecture Notes in Computer Science*, pages 216–226. Springer, 2007.

[LP07] Diane Lambert and Daryl Pregibon. More bang for their bucks: Assessing new features for online advertisers. In *Proceedings of the AdKDD Workshop*, San Jose, California, August 2007.

[NWM07] Zaiqing Nie, Ji-Rong Wen, and Wei-Ying Ma. Object-level vertical search. In *CIDR*, pages 235–246. www.crdrdb.org, 2007.

[Rad07] Filip Radlinski. Addressing malicious noise in clickthrough data. In *Learning to Rank for Information Retrieval Workshop at SIGIR 2007*, 2007.

[RB07] Oliver J. Rutz and Randolph E. Bucklin. A model of individual keyword performance in paid search advertising. Unpublished paper available at `http://ssrn.com/abstract=1024765`, 2007.

[RM99] Jason Rennie and Andew McCallum. Efficient web spidering with reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 335–343, 1999.

[SG07] Mark Steyvers and Tom Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, 2007.

[SH06] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In Les Carr, David De Roure, Arun Iyengar, Carole A. Goble, and Michael Dahlin, editors, *WWW*, pages 377–386. ACM, 2006.

[SMBR04] Mehran Sahami, Vibhu O. Mittal, Shumeet Baluja, and Henry A. Rowley. The happy searcher: Challenges in web information retrieval. In Chengqi Zhang, Hans W. Guesgen, and Wai-Kiang Yeap, editors, *PRICAI*, volume 3157 of *Lecture Notes in Computer Science*, pages 3–12. Springer, 2004.

[SSN06] Duncan I. Simester, Peng Sun, and John. N.Tsitsiklis. Dynamic catalog mailing policies. *Management Science*, 52(5):683–696, May 2006. Available at `w3.mit.edu/jnt/www/Papers/-J102-06-catalog.pdf`.

[Tuz06] Alexander Tuzhilin. Lane's Gifts v. Google report. Available at `http://googleblog.blogspot.com/pdf/-Tuzhilin_Report.pdf`, July 2006.

[WZPSS07] Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors. *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*. ACM, 2007.

[XL06] Bo Xing and Zhangxi Lin. The impact of search engine optimization on online advertising market. In Mark S. Fox and Bruce Spencer, editors, *ICEC*, volume 156 of *ACM International Conference Proceeding Series*, pages 519–529. ACM, 2006.