

Reading List (605.744 Information Retrieval)

Each student is expected to present one paper to the class. I am especially interested in seeing papers marked by “[*]” presented, but you may suggest other papers, even ones not in this list. Your presentation should be brief (i.e., not over 15 minutes), should probably use prepared slides, and must occur between 3/5/07 and 4/23/07. These paper reviews are separate from the oral presentations during the last two weeks of class that are part of the research projects. (I’ll hand out information about projects next week.)

During the next week send me an email telling me which paper you want to present and your top two dates for giving a presentation. You are allowed to present a paper that isn't on this list, subject to my approval. I suggest you pick something that sounds interesting to you, or that might be related to a topic you’d like to investigate for your research project. Many of these papers are available on the web. If you have trouble locating any online, I should be able to obtain a copy for you. A few of the papers on this list are intended to be required readings and those are indicated with a check mark; since the whole class will read these, they aren’t good choices for a presentation.

Introductory

- ✓ M Lesk, *The Seven Ages of Information Retrieval*, 1995.
- M. Lesk, *How Much Information is there in the World?*, 1997 (<http://lesk.com/mlesk/ksg97/ksg.html>)

Models of Retrieval

- [*] D Hiemstra and A de Vries, *Relating the new language models of information retrieval to the traditional retrieval models*, *CTIT Technical Report TR-CTIT-00-09*, 2000, (<http://wwwhome.cs.utwente.nl/~hiemstra/papers/>)
- W. Cooper, A. Chen, F. Gey, *Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression*, TREC-2, 1993, <http://trec.nist.gov/pubs/trec2/papers/txt/05.txt>
- S Deerwester, S Dumais, G Furnas, T Landauer, R Harshman, *Indexing by Latent Semantic Analysis*, JASIS, 1990.
- [*] Clarke, Cormack, and Tudhope, *Relevance ranking for one to three term queries*, IPM 36:291-311, 2000.
- [*] Amati and van Rijsbergen, *Probabilistic models of information retrieval based on measuring the divergence from randomness*, ACM TOIS 20(4):357-389, 2002.

Text processing

- W Frakes, *Stemming Algorithms* (Chapter 8 in *Information Retrieval: Data Structures and Algorithms by Frames and Baeza-Yates*), 1992
- M. Porter, *An Algorithm for Suffix Stripping* (In *Readings in IR*)
- M. Damashek, *Gauging Similarity with ngrams: Language-Independent Categorization of Text*. *Science*, Vol. 267, 10 February, 843 – 848, 1995
- G. Grefenstette and P. Tapanainen, *‘What is a word, What is a sentence? Problems of Tokenization’*, Rank Xerox Research Center Tech Report (MLTT-004), 1994
- [*] K. Church, *One Term or Two?*, SIGIR-95 (on stemming and term weighting)

Indexing / Efficiency

- ✓ G. Salton and C. Buckley, *Term Weighting Approaches in Automatic Text Retrieval*.
- J. Zobel and A. Moffat, *Adding Compression to a Full-Text Retrieval System*, *Software Practice and Experience* 25(8), 1995
- [*] David Carmel et al., *Static Index Pruning for Information Retrieval Systems*, SIGIR 2001, pp. 43-50.
- [*] Shieh et al., *‘Inverted file compression through document identifier reassignment’*, *Information Processing and Management*, 39(1), 117-131, 2003.
- [*] D. Bahle., H. E. Williams, and J. Zobel, *‘Efficient Phrase Querying with an Auxiliary Index.’* SIGIR-02, pp. 215-221, 2002.

Evaluation

- E. Voorhees, D. Harman, 'Overview of the Eighth Text REtrieval Conference (TREC-8)', TREC-8 (available at: http://trec.nist.gov/pubs/trec8/t8_proceedings.html)
- J. Zobel, How Reliable Are the Results of Large-Scale Information Retrieval Experiments?, SIGIR-98
- [*] Sanderson and Zobel, Information retrieval system evaluation: effort, sensitivity, and reliability, ACM SIGIR 2005, pp. 162-169.
- [*] C. Buckley and E. Voorhees, 'Retrieval Evaluation with Incomplete Information', SIGIR-2004
- [*] Yilmaz and Aslam, Estimating Average Precision with Incomplete and Imperfect Judgments, CIKM-06, pp. 102-111.

Query Processing

- [*] D. Harman, 'Relevance Feedback Revisited.' SIGIR-92
- J. Xu and W. B. Croft, Query Expansion Using Local and Global Document Analysis, SIGIR-96
- V. Lavrenko and B. Croft, 'Relevance-Based Language Models', SIGIR-2001.

Web

- ✓ S Lawrence and L Giles, 'Searching the World Wide Web, Science v280 pp 98-100, 1998 (available at: <http://citeseer.nj.nec.com/lawrence98searching.html>)
- Jon Kleinberg. Authoritative sources in a hyperlinked environment. Technical Report RJ 10076, IBM, May 1997. <http://citeseer.nj.nec.com/kleinberg99authoritative.html>
- S. Lawrence and L. Giles, "Accessibility of Information on the Web", Nature, Vol. 400, pp. 107-109, 1999.
- Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine, available at <http://citeseer.nj.nec.com/brin98anatomy.html>
- J. Cho, H. Garcia-Molina, L. Page, "Efficient Crawling Through URL Ordering," Proceedings of the 7th World Wide Web conference, April 1998.
- [*] B.E. Brewington, G. Cybenko, "How Dynamic Is the Web?" 9th World Wide Wide Conference, May, 2000.
- [*] A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani, Adwords and generalized online matching, JACM 54(5), 2007.

Text Classification and Filtering

- [*] David Lewis. Naive bayes at forty: The independence assumption in information retrieval. In Proc. 10th European Conference on Machine Learning ECML-98, pages 4-15, 1998.
- [*] Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Proc. 10th European Conference on Machine Learning ECML-98, pages 137-142, 1998. (this paper and others at: <http://www.cs.helsinki.fi/research/doremi/categorization/bibliography.html>)
- D Hull and S Robertson, The TREC-8 Filtering Track Final Report, TREC-8, (<http://trec.nist.gov/pubs/trec8/papers/filtering.pdf>)
- T. Yan and H. Garcia-Molina. SIFT -- A tool for wide-area information dissemination. In Proc. USENIX Winter 1995 Technical Conference, New Orleans, January 1995. <http://citeseer.nj.nec.com/yan95sift.html>
- Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In Proc. 14th International Conference on Machine Learning, pages 412-420, 1997.
- James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In Proc. ACM SIGIR, pages 37-45, 1998.
- ✓ Goodman, Cormack, and Heckerman, Spam and the ongoing battle for the inbox, CACM 50(2), pp24-33, 2007.

Cross-Language Information Retrieval

- D. Hull and G. Grefenstette, Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval, (Readings in IR)
- A. Pirkola, T. Hedlund, H. Keskusalo, and K. Järvelin, 'Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings.' *Information Retrieval*, 4:209-230, 2001.
- L Ballesteros and W B Croft, Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval, SIGIR-97

- P McNamee and J Mayfield, Comparing cross-language query expansion techniques by degrading translation resources, SIGIR 2002 (available at: <http://apl.jhu.edu/~paulmac/publications.html>)
- [*] Scott McCarley, Should we Translate the Documents or the Queries in Cross-language Information Retrieval?, ACL-99 (available at: <http://acl.ldc.upenn.edu/P/P99/>)
- ✓ K. Kishida, Technical issues of cross-language information retrieval: a review, IPM 41, pp. 433-455, 2005.

Distributed Retrieval

- D. Cutting, D. Karger, J. Pedersen, and J.W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, Proceedings of the 15th Annual International ACM/SIGIR Conference, 1992.
- [*] J. Callan. Distributed information retrieval. In W.B. Croft, editor, Advances in information retrieval, chapter 5, pages 127-150. Kluwer Academic Publishers, 2000. <http://citeseer.nj.nec.com/callan00distributed.html>
- C. M. Bowman, Peter B. Danzig, Darren R. Hardy, Udi Manber and Michael F. Schwartz. Harvest: A Scalable, Customizable Discovery and Access System. Technical Report CU-CS-732-94, Department of Computer Science, University of Colorado, Boulder, July 1994.
- [*] Andoni and Indyk, Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions, CACM 51(1), pp. 117-122, 2008. (or a similar paper)

Information Extraction / Question Answering

- Leek, T. R. 1997. Information extraction using hidden Markov models. Master's thesis, UC San Diego. <http://citeseer.nj.nec.com/leek97information.html>
- E. Voorhees, D. Tice, 'The TREC-8 Question Answering Track Evaluation' (available at: http://trec.nist.gov/pubs/trec8/t8_proceedings.html)
- S Harabagiu, D Moldovan, et al. FALCON: Boosting Knowledge for Answer Engines, TREC-9, 2000, (available at <http://trec.nist.gov/pubs/trec9/papers/smu.pdf>)
- S. Soderland, "Learning to Extract Text-based Information from the World Wide Web," in Proceedings of Third International Conference on Knowledge Discovery and Data Mining (KDD-97).
- [*] A Culotta, A McCallum and J Betz, 'Integrating Probabilistic Extraction Models and Relational Data Mining to Discover Relations and Patterns in Text'. HLT-NAACL, 2006.
- [*] Y. Shinyama and S. Sekine. Preemptive Information Extraction Using Unrestricted Relation Discovery. HLT-2006.
- [*] M. Pasca, B. van Durme, and N. Garera, The Role of Documents vs. Queries in Extracting Class Attributes from Text, CIKM 2007. (available at: <http://www.cs.jhu.edu/~ngarera/publications/attribExtractionCIKM07.pdf>)

Multimedia Retrieval

- Garofolo, J., Auzanne, C. G. P. and Voorhees, E. M., The TREC Spoken Document Retrieval Track: A success story, Proceedings of the Eighth Text Retrieval Conference, Gaithersburg, MD, November, 1999, pp. 107-130. Available at <http://trec.nist.gov/pubs/trec8/papers/trec8-sdr-overview.pdf>
- [*] Ng, K., "Subword-based Approaches for Spoken Document Retrieval," Ph.D. Thesis, MIT, February 2000. (<http://citeseer.ist.psu.edu/article/ng99subwordbased.html>)
- Gudivada, V.N. and Raghavan, V.V., Modeling and Retrieving Images by Content, Information Processing and Management, Vol 33, No 4, pp. 427-452, 1999.
- [*] Murat Saraclar, Richard Sproat. "Lattice-Based Search for Spoken Utterance Retrieval." HLT-NAACL 04, Boston, May 2004.
- Something on Music Retrieval from the ISMIR conferences.

Natural Language Processing and IR

- T. Strzalkowski, Robust Text Processing in Automated Information Retrieval, (Readings in IR)
- R. Mihalcea and V. Nastase, Letter Level Learning for Language Independent Diacritics Restoration, CoNLL-2002
- Manning and Schutze, Foundations of Statistical Natural Language Processing, Chapter 10 (Part of Speech tagging using Hidden Markov Models)
- ✓ M. Sanderson, Retrieving with Good Sense, Information Retrieval 2(1), 2000 (available online at http://dis.shef.ac.uk/mark/cv/publications/papers/my_papers/)
- [*] Ellen M. Voorhees: Using WordNet to Disambiguate Word Senses for Text Retrieval. SIGIR 1993: 171-180