

**Computer Engineering Department
Bilkent University**

CS533: Information Retrieval Systems

Assignment No. 1 (for Assignment No. 2 please see page 3.)

February 17, 2012

Due date: February 29, 2012; Wednesday, by class time (hardcopy is required)

Notes: Handwritten answers are not acceptable.

1. Consider the following search results for two queries Q1 and Q2 (the documents are ranked in the given order, the relevant documents are shown in bold).

Q1: **D1**, **D2**, D3, **D4**, D5, **D6**, D7, D8, D9, D10.

Q2: **D1**, D2, **D3**, **D4**, D5, D6, D7, D8, **D9**, and D10.

For Q1 and Q2 the total number of relevant documents are, respectively, 6 and 4 (Q1 two of the relevant documents are not retrieved).

- a. Using the TREC interpolation rule, in a table give the precision value for the 11 standard recall levels 0.0, 0.1, 0.2, ... 1.0. Please also draw the corresponding recall-precision graph as shown in the first figure of TREC-6 Appendix A (its link is available on the course web site).

Hint. "Interpolated" means that, for example, precision at recall 0.10 (i.e., after 10% of rel docs for a query have been retrieved) is taken to be MAXIMUM of precision at all recall points ≥ 0.10 . Values are averaged over all queries (for each of the 11 recall levels). These values are used for Recall-Precision graphs. (This paragraph is taken from: http://ir.iit.edu/~dagr/cs529/files/project_files/trec_eval_desc.htm.)

Please do this for each query separately and obtain one table for both queries using the average of two values at each recall point.

- b. Find R-Precision (TREC-6 Appendix A for definition) for Query1 and Query2.
- c. Find MAP for these queries.
2. Consider document-based partitioning and term-based partitioning approaches as define in the Zobel-Moffat paper "Inverted files for text search engines" (*ACM Computing Surveys*, 2006). Please also consider the following document by term binary D matrix for m= 6 documents (rows), n= 6 terms (columns). Describe a two ways of indexing across a cluster of machines.

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Under which condition you would prefer one approach over the other one? Please briefly explain.

3. In this part again consider the Zobel-Moffat paper.

- a. Understand the skipping concept as applied to the inverted index construction.

Assume that we have the following posting list for term a: $\langle 1, 5 \rangle \langle 3, 2 \rangle \langle 9, 2 \rangle \langle 10, 4 \rangle \langle 12, 4 \rangle \langle 17, 4 \rangle \langle 18, 3 \rangle, \langle 22, 2 \rangle \langle 24, 4 \rangle \langle 33, 4 \rangle \langle 38, 5 \rangle \langle 43, 5 \rangle \langle 55, 3 \rangle \langle 64, 2 \rangle \langle 68, 4 \rangle \langle 72, 5 \rangle \langle 75, 1 \rangle \langle 88, 2 \rangle \dots$. The posting list indicates that term-a appears in d1 five times and in d3 twice, etc.

Assume that we have the following posting list for term-b: $\langle 12, 3 \rangle \langle 45, 2 \rangle \langle 66, 1 \rangle$.

Consider the following conjunctive Boolean query: term-a **and** term-b. If no skipping is used how many comparisons do you have to find the intersection of these two lists?

Introduce a skip structure, draw the corresponding figure then give the number of comparisons involved to process the same query.

- b. Give a posting list of term-a (above it is given in standard sorted by document number order) in the following forms: 1), a) ordered by $f_{d,t}$, b) ordered by frequency information in prefix form.

Which method would you prefer and why?

4. What are the components of an information retrieval test collection? Explain the pooling approach?

Is pooling a reliable approach for the construction of test collections? Find the paper by Zobel regarding this issue, it may help.

What is bpref? Is it anyway related to the pooling concept?

5. Please read "Natural language and information layer" lecture paper of Karen Sparck Jones from ACM SIGIR 2007 Conf. proceedings. Sparck Jones says that "My talk has three parts: on the first phase of natural language processing research and its lessons; on subsequent developments up to the present and their lessons; and on where we are now and what I think are the wider implications for the future."

Please summarize/interpret these three parts separately using about five sentences per part. Make sure that you write your own summary/interpretation.

CS533: Information Retrieval Systems

Assignment No. 2

February 27, 2012

Due dates: March 7

5-minute presentation assignment.

Pick a paper from the list given in

Alistair Moffat, Justin Zobel, David Hawking: [Recommended reading for IR research students](#). SIGIR Forum 39(2): 3-14 (2005)

Please read the comments about the papers that appear in the list (see pages below for the citations of the recommended papers).

- 1) Prepare a 5-minute in class presentation using power point. This is a new presentation style used in some CS conferences.
- 2) Make it available on the Web also bring it to the class in a memory stick.
- 3) You must prepare handout (e.g., a poster on A4 paper) and give it to your classmates at the beginning of your presentation.
- 4) Only provide the most essential parts of the paper (you have no other choice).
- 5) Most importantly make us understand the intuition behind it, and its significance.
- 6) There will be two students/presentation, hence there will be $(22/2)=11$ groups (11 x 5 + 10 minutes for transitions= 65 minutes. Please form your groups as soon as practical.
- 7) I will bring a chronometer and stop each presentation when the allocated time expires.
- 8) With your votes we will pick the best presentation. You can only vote for one group. No two groups can present the same paper. (Please do not pick no. 4, 24, 33, 46 since we may cover them in our course in other ways.)
- 9) Please send me an email with the subject line "CS533 5-minute presentation" indicate your group members, and your ranked top-3 preferences. Email me your preferences between February 26 (beginning at 0:00 hours) February 27 (ending at 23:59 hours). I will assign the paper according to FCFS policy.

Future Pointer:

As I have announced we will do something similar for the Hamming book "[The Art of Doing Science and Engineering: Learning to Learn](#)." Please go ahead read some chapters from this book.

The list:

1. S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J. of Molecular Biology*, 215:403--410, 1990.
2. Adam Berger , John Lafferty, Information retrieval as statistical translation, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.222-229, August 15-19, 1999, Berkeley, California, United States [doi>10.1145/312624.312681]
3. Krishna Bharat , Monika R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.104-111, August 24-28, 1998, Melbourne, Australia [doi>10.1145/290941.290972]
4. Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, Proceedings of the seventh international conference on World Wide Web 7, p.107-117, April 1998, Brisbane, Australia
5. Andrei Broder, A taxonomy of web search, *ACM SIGIR Forum*, v.36 n.2, Fall 2002 [doi>10.1145/792550.792552]
6. Chris Buckley , Ellen M. Voorhees, Evaluating evaluation measure stability, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, p.33-40, July 24-28, 2000, Athens, Greece [doi>10.1145/345508.345543]
7. J. Callan. Distributed information retrieval. In W. Bruce Croft, editor, *Advances in Information Retrieval*, chapter 5, pages 127--150. Kluwer Academic Publishers, 2000. URL <http://www-2.ca.cmu.edu/~callan/Papers/ciir00.pa.gz>.
8. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic indexing. *J. of the American Society for Information Science*, 41(6):391--407, 1990.
9. Susan Dumais , Edward Cutrell , JJ Cadiz , Gavin Jancke , Raman Sarin , Daniel C. Robbins, Stuff I've seen: a system for personal information retrieval and re-use, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, July 28-August 01, 2003, Toronto, Canada [doi>10.1145/860435.860451]
10. Abdessamad Echihabi, Daniel Marcu, A noisy-channel approach to question answering, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, p.16-23, July 07-12, 2003, Sapporo, Japan [doi>10.3115/1075096.1075099]
11. D. K. Harman and G. Candela. Retrieving records from a giga-byte of text on a minicomputer using statistical ranking. *J. of the American Society for Information Science*, 41(8):581--589, August 1990.
12. David Hawking, Stephen Robertson, On Collection Size and Retrieval Effectiveness, *Information Retrieval*, v.6 n.1, p.99-105, January 2003 [doi>10.1023/A:1022904715765]
13. M. Hearst. User interfaces and visualization. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 257--323. Addison-Wesley Longman, 1999. URL <http://www.sims.berkeley.edu/~hearst/irbook/chapters/chap10.html>.
14. David G. Hendry, David J. Harper, An informal information-seeking environment, *Journal of the American Society for Information Science*, v.48 n.11, p.1036-1048, Nov. 1997 [doi>10.1002/(SICI)1097-4571(199711)48:11<1036::AID-ASI6>3.3.CO;2-E]
15. William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, Daniel Olson, Do batch and user evaluations give the same results?, Proceedings of the 23rd annual

- international ACM SIGIR conference on Research and development in information retrieval, p.17-24, July 24-28, 2000, Athens, Greece [doi>10.1145/345508.345539]
16. W. R. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, C. Moesback, and D. Kraemer. Factors associated with success for searching MED-LINE and applying evidence to answer clinical questions. *J. of the American Medical Informatics Association*, 9(3):283--293, May/June 2002. URL <http://madir.ohsu.edu/~hersh/jamia-02-irfactors.pdf>.
 17. J. R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. E. Stickel, and M. Tyson. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*, pages 383--406. MIT Press, 1996. URL <http://citeseer.nj.nec.com/hobbs96fastus.html>.
 18. P. Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *J. of Documentation*, 52(1):3--50, 1996.
 19. Jon M. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM)*, v.46 n.5, p.604-632, Sept. 1999 [doi>10.1145/324133.324140]
 20. Victor Lavrenko, Martin Choquette, W. Bruce Croft, Cross-lingual relevance models, *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, August 11-15, 2002, Tampere, Finland [doi>10.1145/564376.564408]
 21. Victor Lavrenko, W. Bruce Croft, Relevance based language models, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, p.120-127, September 2001, New Orleans, Louisiana, United States [doi>10.1145/383952.383972]
 22. V. Lavrenko, W. B. Croft. Relevance models in information retrieval. In W. Bruce Croft and John Lafferty, editors, *Language Modelling for Information Retrieval*, pages 11--56. Kluwer Academic Publishers, 2003.
 23. David D. Lewis, Karen Spärck Jones, Natural language processing for information retrieval, *Communications of the ACM*, v.39 n.1, p.92-101, Jan. 1996 [doi>10.1145/234173.234210]
 24. Alistair Moffat, Justin Zobel, Self-indexing inverted files for fast text retrieval, *ACM Transactions on Information Systems (TOIS)*, v.14 n.4, p.349-379, Oct. 1996 [doi>10.1145/237496.237497]
 25. Douglas W. Oard, Bonnie J. Dorr, A survey of multilingual text retrieval, University of Maryland at College Park, College Park, MD, 1996
 26. Douglas W. Oard, Julio Gonzalo, Mark Sanderson, Fernando López-Ostenero, Jianqiang Wang, Interactive Cross-Language Document Selection, *Information Retrieval*, v.7 n.1-2, p.205-228, January-April 2004 [doi>10.1023/B:INRT.0000009446.22036.e3]
 27. Michael Persin, Justin Zobel, Ron Sacks-Davis, Filtered document retrieval with frequency-sorted indexes, *Journal of the American Society for Information Science*, v.47 n.10, p.749-764, Oct. 1996 [doi>10.1002/(SICI)1097-4571(199610)47:10<749::AID-ASIS>3.3.CO;2-U]
 28. Jay M. Ponte, W. Bruce Croft, A language modeling approach to information retrieval, *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p.275-281, August 24-28, 1998, Melbourne, Australia [doi>10.1145/290941.291008]
 29. Stephen Robertson, Hugo Zaragoza, Michael Taylor, Simple BM25 extension to multiple weighted fields, *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, November 08-13, 2004, Washington, D.C., USA [doi>10.1145/1031171.1031181]

30. S. E. Robertson and K. Sparck Jones. Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, Cambridge Computer Laboratory, May 1997. URL <http://www.cl.cam.ac.uk/TachReports/UCAM-CL-TR-356.pdf>.
31. S. E. Robertson, C. J. van Rijsbergen, M. F. Porter, Probabilistic models of indexing and searching, Proceedings of the 3rd annual ACM conference on Research and development in information retrieval, p.35-56, June 23-27, 1980, Cambridge, England
32. S. E. Robertson, S. Walker, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In Proc. TREC-3, November 1994. URL <http://trec.nist.gov/pubs/trec3/papers/city.ps.gz>. NIST Special Publication 500-225.
33. Gerard Salton, Christopher Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management: an International Journal, v.24 n.5, p.513-523, 1988 [doi>10.1016/0306-4573(88)90021-0]
34. Linda Schamber, Michael Eisenberg, Michael S. Nilan, A re-examination of relevance: toward a dynamic, situational definition, Information Processing and Management: an International Journal, v.26 n.6, p.755-776, 1990 [doi>10.1016/0306-4573(90)90050-C]
35. Amit Singhal, Chris Buckley, Mandar Mitra, Pivoted document length normalization, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.21-29, August 18-22, 1996, Zurich, Switzerland [doi>10.1145/243199.243206]
36. Amit Singhal , Fernando Pereira, Document expansion for speech retrieval, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.34-41, August 15-19, 1999, Berkeley, California, United States [doi>10.1145/312624.312645]
37. Arnold W. M. Smeulders, Marcel Worring, Simone Santini , Amarnath Gupta , Ramesh Jain, Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence, v.22 n.12, p.1349-1380, December 2000 [doi>10.1109/34.895972]
38. K. Sparck Jones, S. Walker, S. E. Robertson, A probabilistic model of information retrieval: development and comparative experiments, Information Processing and Management: an International Journal, v.36 n.6, p.779-808, Nov.06.2000 [doi>10.1016/S0306-4573(00)00015-7]
39. Anastasios Tombros, Mark Sanderson, Advantages of query biased summaries in information retrieval, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.2-10, August 24-28, 1998, Melbourne, Australia [doi>10.1145/290941.290947]
40. C. J. van Rijsbergen, Towards an information logic, Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval, p.77-86, June 25-28, 1989, Cambridge, Massachusetts, United States [doi>10.1145/75334.75344]
41. Various Authors. Collected papers about TREC-2. Information Processing and Management, 31(3):269--453, May 1995. URL <http://www.sciencedirect.com/science/journal/03064573>.
42. Ellen M. Voorhees, Variations in relevance judgments and the measurement of retrieval effectiveness, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.315-323, August 24-28, 1998, Melbourne, Australia [doi>10.1145/290941.291017]
43. Ellen M. Voorhees, Variations in relevance judgments and the measurement of retrieval effectiveness, Information Processing and Management: an International Journal, v.36 n.5, p.697-716, Sept. 2000 [doi>10.1016/S0306-4573(00)00010-8]

44. Oren Zamir, Oren Etzioni, Grouper: a dynamic clustering interface to Web search results, Proceeding of the eighth international conference on World Wide Web, p.1361-1374, May 1999, Toronto, Canada
45. Chengxiang Zhai, John Lafferty, A study of smoothing methods for language models applied to Ad Hoc information retrieval, Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.334-342, September 2001, New Orleans, Louisiana, United States [doi>10.1145/383952.384019]
46. Justin Zobel, How reliable are the results of large-scale information retrieval experiments?, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.307-314, August 24-28, 1998, Melbourne, Australia [doi>10.1145/290941.291014]
47. Justin Zobel, Alistair Moffat, Exploring the similarity space, ACM SIGIR Forum, v.32 n.1, p.18-34, Spring 1998 [doi>10.1145/281250.281256]