Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 2 (for Assignment No. 3 please see page 3.)
March 4, 2010
Due date: March 22, 2010; Monday, by noon time (12:00 o'clock) (hardcopy is required)

Notes: Handwritten answers are not acceptable. The number of questions in this version is more than the earlier version that I gave in the classroom, please use this version. Printing the first four pages for assignment no. 2 & 3 would be enough. The next assignment may overlap with this one.

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 is, respectively, 4 and 5 (Q2 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).

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 the following document by term binary D matrix for m= 6 documents (rows), n= 6 terms (columns).

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

Consider the problem of constructing a document by document similarity, S, matrix. How many similarity coefficients will be calculated using the following methods? For each case explain your answer briefly: give exact numbers for each document and explain how you came up with those numbers.

a. Straightforward approach (using document vectors) -the 1st method discussed in the class-.

b. Using term inverted indexes.

3. Obtain the similarity matrix S for the above D matrix (you don't need to show your intermediate steps). Use the Dice similarity coefficient. Use the S matrix to construct the dendrogram (cluster tree) structure corresponding to the single-link and complete link clustering methodologies.
Note: For verifying your answer you may experiment with the toy clustering algorithms program which is available on our course web site: cluster.exe. Some hints about its use is as follows –please

experiment with it-. Use **Options**: for setting the dimensions of D, **Matrices**: for entering your D matrix (after entering a value hit "Enter" on keyboard or use the arrow keys), c) **Hiearchical/Single-Link**: for obtaining the single-link, d) **Exit**: for going back to the main menu. (New Windows operating systems and this program may not be compatible with each other.)

4. Consider the above D matrix. Cluster the documents using the cover coefficient-based clustering methodology (C^3M). Please a) Show the double-stage probability experiment tree for the second document, and show the calculation of c_{24} of the corresponding C matrix, b) obtain the C matrix (you do not need to show the intermediate steps), c) find the number of clusters implied by the C matrix –explain how-, d) find the cluster seeds, e) obtain the IISD (inverted index for seed documents), f) obtain the clusters and explain how you them.

5. Consider the incremental version of C^3M: C^2ICM, Cover Coefficient-based Incremental Clustering Methodology, described in Can F, Incremental clustering for dynamic information processing, ACM TOIS, 1993).

a. Briefly explain the algorithm (one paragraph).

b. In the paper there is the concept of clustering similarity, explain its purpose within the context of C^2ICM.

c. The paper mentions a measure called Rand coefficient (and cites the classic book of Jain & Dubes: Algorithms For Clustering Data, http://www.cse.msu.edu/~jain/Clustering_Jain_Dubes.pdf, pp. 172-177). Obtain the (regular) Rand similarity of the clustering structures CS1= { {a, b, c}, {d, e}, {f, g}} and another clustering structure CS2= {{a}, {b, c, d}, {e, f, g}} -where the last cluster of CS2 contains the members e, f, and g-. Optional: you may also obtain the corrected Rand coefficient using these two clustering structure. Show the contingency table that needs to be corrected for the Rand coefficients.

d. Explain the difference between Rand and corrected Rand coefficients.

You may want to consider the presentation by Uluçınar, Özcan and Canım for some examples etc. http://www.cs.bilkent.edu.tr/~canf/CS533/CS533Spr06stuPresent/DMOZ Clustering Sunum.ppt

6. In this part consider the paper J. Zobel, A. Moffat, "Inverted files for text search engines." *ACM Computing Surveys*, Vol. 38, No. 2, 2006.

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

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

Assume that we have the following posting list for term-b: <12, 2> <66, 1>.

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.

State the advantages and disadvantages of large and small skips in the posting lists. Note that in the paper it is assumed that compression will be used. The skip idea is applicable in an uncompressed environment too.

b. Give a posting list of 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. What are the advantages of the approaches a and b? Do they have any practical value?

7. In this part consider the paper A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review" *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.

a. Please explain the stages of clustering as defined in this paper.

b. Consider fuzzy clustering and introduce and idea that we can use fuzzy clustering approach in connection with C^3M.

c. In connection with simulated annealing the authors mention "tabu search." What does it mean? Explain its use within the context of simulated annealing-based clustering.

d. What are the components of a typical clustering task? Explain each step within the framework of an information retrieval environment.

e. In connection with the above question (section d) please also explain what is meant by clustering tendency? Does it make sense to use clustering tendency in some stage(s) of clustering? What would you propose to use for identifying clustering tendency? Please try to be creative. For this purpose you may do a literature search and borrow some ideas and use them after some modification.

8. Is the complete-link clustering method order-independent? Explain/prove your claim. (You may see a related formal proof for the single-link method on our course web site).

9. What are the components of an information retrieval test collection? Explain the pooling approach? Please read the paper by Zobel (How Reliable Are the Results of Large-Scale Information Retrieval Experiments?) and give some reflections of his criticism of this approach.

10. Please reexamine the 11 Watt/Google Query Legend. Is it real or not? Please write your findings based on research – please specify your resources-. Assuming that the claim is true please calculate how much KW you spend in a typical year and also calculate its TL equivalent. (You may also calculate the same cost for a person who lives in New York, NY or any other famous foreign city for comparison.) Explain your reasoning.

If you provide a nicely written answer that can be longer than a paragraph I plan to publish it on our course web site as your answer to this claim.

CS533: Information Retrieval Systems
Assignment No. 3
March 4, 2010
Due date: April 14, 2010; Wednesday

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).

Here is the paper (please open it and read the comments), see pages 5-9 below for the citations of the recommended papers.
http://delivery.acm.org/10.1145/1120000/1113344/p3-moffat.pdf?key1=1113344&key2=7974077621&coll=ACM&dl=ACM&CFID=78637105&CFTOKEN=20859108

1) Prepare a 5-minute in class presentation using power point.

2) Make it available on the Web also bring it to the class in a memory stick.

3) You may also give a handout (e.g., a poster on A4 paper) to your classmates.

4) Only provide the most essential part (according to you or some other people) of the paper.

5) Most importantly make us feel the intuition behind it, and its significance.

6) There will be two students/presentation, hence there will be Floor(25/2)= 12 groups, 12 x 5 + 10 minutes for transitions= 70 minutes.

7) I will bring a chronometer and stop each presentation at the end of 5th minute.

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. I will collect your top three preferences in ranked order next week on March 10. (Please do not pick no. 4, 24, 33, 46 since we cover them in our course in other ways.)

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 and 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-ASI3>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]