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, 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).

   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) Hierarchical/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\(^3\)M). 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\(^3\)M: C\(^2\)ICM, 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\(^2\)ICM.

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.


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 term-a (above it is given in standard sorted by document number order) in the following forms: 1) a) ordered by \( f_{dt} \), b) ordered by frequency information in prefix form. What are the advantages of the approaches a and b? Do they have any practical value?


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'M.

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

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:


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]


25 Douglas W. Oard, Bonnie J. Dorr, A survey of multilingual text retrieval, University of Maryland at College Park, College Park, MD, 1996


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]


37 Arnold W. M. Smeulders, Marcel Woring, Simone Santini, Aarnoud Cruijsen, 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]


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]