Computer Science and Systems Analysis Department
Miami University

SAN626: Information Retrieval Systems
Midterm
July 18, 2000 **NAME:**

GOOD LUCK!

Notes: **1. There are 100 points and 10 questions on 6 pages..**
2. Please read the questions.
3. For the formulas please refer to the appendix (see the last page).

1. (10 pts.) Consider a query with total of 10 relevant documents in the collection. For this query assume that a user examines ten documents and states that 2nd, 4th, 5th, and 8th documents are relevant. In the following table give the precision values at 11 standard recall values using the TREC evaluation approach.

Recall											
Precision											

2. (8 pts.) Indexing questions.

a. What is meant by stop word list and how do we use it? A short answer is enough.

b. What would be the TDVs of the stop word list terms if they were used for indexing purposes? Explain briefly and consider the space density, similarity, based calculation method.

3. (10 pts.) Consider the following document by document similarity matrix.

$$S = \begin{array}{c} \begin{matrix} d_1 & d_2 & d_3 & d_4 & d_5 \end{matrix} \\ \begin{bmatrix} 1.00 & 0.42 & 0.00 & 0.10 & 0.10 \\ - & 1.00 & 0.00 & 0.07 & 0.20 \\ - & - & 1.00 & 0.33 & 0.33 \\ - & - & - & 1.00 & 0.36 \\ - & - & - & - & 1.00 \end{bmatrix} \end{array} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix}$$

a. Obtain the corresponding complete-link dendrogram.

b. Give the similarity matrix implied by the dendrogram obtained in section a.

4. (8 pts.) Consider a D matrix with 10 documents, d_i, ($1 \leq i \leq 10$), and assume that document di contains i number of terms, so d_1 contains 1 term, d_2 contains 2 terms etc.. In the corresponding cover coefficient matrix C, the value of c_{25}= 0.20. What Is the value of c_{52}? Show the steps of your calculations.

5. (10 pts.) Answer the following questions.

a. Is It true that In "complete-link clusters can be very scattered (e.g., two members of the same cluster may have very little in common)?" Explain your answer.

b. Can the members of the clusters be scattered if the clusters are generated by the cover coefficient based clustering algorithm? Explain your answer.

6. (10 pts.) In a partitioned clustered file environment for a given query if there is only one document to retrieve then we have to access only one cluster. This is obvious. Does the formula developed by Yao imply the same?

According to Yao the total number of clusters (blocks) to be accessed for k documents is given by the following formula. (In this question and in the following formula it is assumed that all clusters have the same size).

$$m \cdot \left[1 - \prod_{i=1}^{k} \frac{nd - i + 1}{n - i + 1} \right]$$

Where
m : No. of clusters
n : No. of documents
d : (1 - 1/m)
k : No. of documents to be retrieved for the query (in this question k= 1).

7. (12 pts.) Consider the following D matrix. In this matrix rows and columns, respectively, indicate document and term vectors.

$$D = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix}$$

Consider the problem of constructing the <u>term by term</u> similarity, S, matrix. (Please read the previous sentence once more.) Notice that this question is the inverted version of the problem that we have studied in the class.

How many similarity coefficients will be calculated by using the following approaches?

a. Straight forward approach (using term vectors). Explain your answer briefly.

b. Using the knowledge of document distributions in the terms. Explain your answer briefly.

8. (12 pts.) Consider the following D matrix. In this matrix rows and columns, respectively, indicate document and term vectors.

$$D = \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix}$$

a. Can you guess the number of clusters implied by D matrix using the concepts of the cover coefficient-based clustering algorithm; however, without obtaining the C matrix? If so give that number.

b. What is the TDV of t_6 according to the cover coefficient concept? Again answer the question without obtaining the C matrix, i.e., state the approximate value. Use part a of this question as a hint.

c. According to your calculation is t_6 a good discriminator or bad discriminator? Explain why.

9. (20 pts.) Short questions.

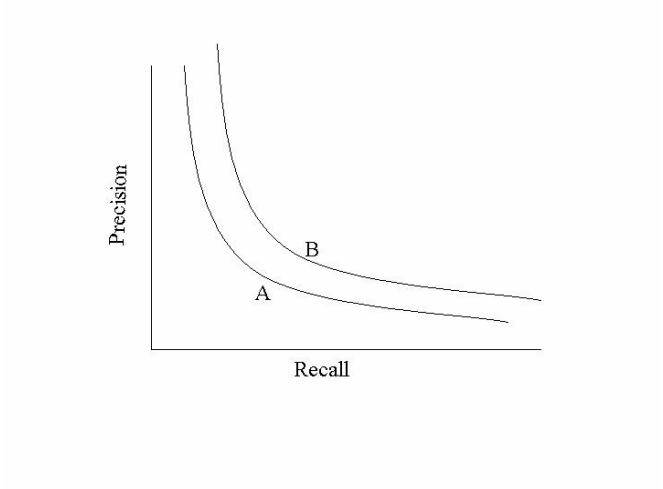
a. If each entry of a posting list is allocated 8 bytes of storage how much space is needed if a posting list contains 4500 entries and how many disk blocks are required to store this posting list if the size of page is 1 KB?

b. According to Salton what can be used to enhance precision and recall? (Give one example for each.)

c. State two advantages of stemming.

d. For the corpus {computability, computation, computed, computers, computing} how many successor varieties does the string "comput" have? What are they?

e. In the figure given below which of the two (A or B) represents a more efficient Information Retrieval System? If there is something wrong in the previous question. What is it?



APPENDIX

The definitions of c_{ij} and c_{ij}'

$$c_{ij} = \alpha_i \times \sum_{k=1}^{n} (d_{ik} \times \beta_k \times d_{jk}) \qquad c'_{ij} = \beta_i \times \sum_{k=1}^{m} (d_{ki} \times \alpha_k \times d_{kj})$$

m: No. of documents
n: No. of terms