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	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precisio	0.6	0.6	0.6	0.6	0.5	0.0	0.0	0.0	0.0	0.0	0.0

2. (8 pts.) Indexing questions.

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

A ***stop list*** contains high frequency words that should be removed before indexing.

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.

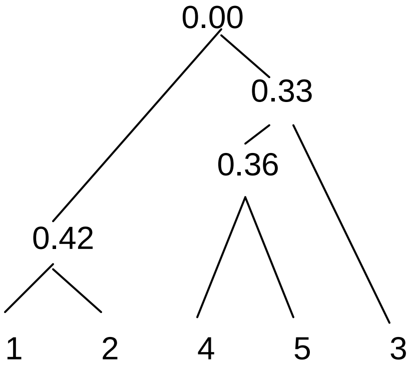
Stop word list terms are frequent, so their TDVs<0 according to similarity, calculation method.

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

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

a. Obtain the corresponding complete-link dendrogram.

S[12]:0.42
S[45]:0.36
S[34]:0.33
S[45]:0.36
S[34]:0.33
S[25]:0.20
S[14]:0.10
S[15]:0.10
S[24]:0.07



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

$$S' = \begin{array}{cccccc} d1 & d2 & d3 & d4 & d5 & \\ 1.00 & 0.42 & 0.00 & 0.00 & 0.00 & d1 \\ X & 1.00 & 0.00 & 0.00 & 0.00 & d2 \\ X & X & 1.00 & 0.33 & 0.33 & d3 \\ X & X & X & 1.00 & 0.36 & d4 \\ X & X & X & X & 1.00 & d5 \end{array}$$

4. Consider a D matrix with 10 documents, d , (1<=I<=10), and assume that document d , contains I number of terms, so d contains 1 term, d contains 2 terms etc.. In the corresponding cover coefficient matrix C, the value of c = 0.25. What is the value of c52? Show the steps of your calculations.

Answer:
$C_{25} =;_2 \Sigma d_{2k}\ \beta_k\ d_{5k} = 0.20$
$C_{52} =;_5 \Sigma d_{5k}\ \beta_k\ d_{2k}$
$;_2 = ½ = 0.5$
$;_5 = 1/5 = 0.2$

So, we can have the value of $\Sigma d_{2k}\ \beta_k\ d_{5k}$
$\Sigma d_{2k}\ \beta_k\ d_{5k} = C_{25}/;_2 = 0.2/0.5 = 0.4$
$C_{52} =;_5 \Sigma d_{5k}\ \beta_k\ d_{2k} = 0.2 * 0.4 = 0.08$

5. 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.**
No.
The complete-link algorithm works as follows: the similarity between the most dissimilary member is taken as the cluster similarity. So before forming complete-link cluster, we have to consider the all the similarity between members in one cluster. In complete-link clusters will not be very scattered.

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

Yes. Because according to the algorithm of cover coefficient, the members of clusters are similar to the seed, it means that the non-seed members of the same clusters may have little similarity. So using C^3M, the members of the clusters may be scattered.

6. 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 developend 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).

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

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)

Answer:
K=1, m = (1-(nd-1+1)/(n-1+1)) = m(1-d) = m(1-(1-1/m)) = m*1/m = 1
So, Yao's formula implies the same result.

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

$$\begin{array}{cc} & \begin{array}{cccccc} t1 & t2 & t3 & t4 & t5 & t6 \end{array} \\ D= & \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

Consider the problem of construction the term by term 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.
Answer: 1+2+3+4=10.
The S matrix has 5*5 entries. We only have to calculate upper half of the matrix.

b. Using the knowledge of document distributions in the terms. Explain your answer briefly.
Answer:
D1-> <t1,1>,<t2,1>,<t5,1>
D2-> <t1,1>,<t2,1>,<t4,1>,<t5,1>
D3-> <t6,1>
D4-> <t2,1>,<t3,1>,<t6,1>

D5-> <t3,1>,<t4,1>,<t6,1>

Consider t1->d1,d2-> (t1,t2,t5) $\bigcup$ (t1,t2,t4,t5) = t1,t2,t4,t5
So we need to calculate S12,S14,S15.

Consider t2->d1,d2,d4->(t1,t2,t5) $\bigcup$ (t1,t2,t4,t5) $\bigcup$ (t2,t3,t6) = t1,t2,t3,t4,t5,t6.
So we need to calculate S23.S24,S25,S26.

Consider t3->d4,d5->(t2,t3,t6) $\bigcup$ (t3,t4,t6) = t2,t3,t4,t6
So we need to calculate S34,S36

Consider t4->d2,d5->(t1,t2,,t4,t5) $\bigcup$ (t3,t4,t6) = t1,t2,t3,t4,t5,t6
So we need to calculate S45,S46.

Consider t5->d1, d2->(t1,t2,t5) $\bigcup$ (t1,t2,t4,t5) = t1,t2,t4,t5
So we don’t have to calculate anything.
We don’t have to consider t6 either.

So we need to calculate S12,S14,S15, S23,S24,S25,S26, S34,S36, S45,S46.
Totally we have to calculate 11 similarity coefficients.

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

$$\begin{array}{cc} & \begin{array}{cccccc} t1 & t2 & t3 & t4 & t5 & t6 \end{array} \\ D= & \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

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.

Answer: Nc = m*n/t = 5*6/15 = 2.

b. What is the TDV of t6 according to the cover coefficient concept? Again answer the question without obtaining the C matrix, i.e., state approximate value. Use part a of this question as a hint.

Answer: remove the column t6.
Nc6 = m*n/t = 5*5/12 = 2.08
TDV = Nc – Nc6 = 2 – 2.08 =-0.08

c. According to your calculation is t6 a good discriminator or bad discriminator? Explain why.

Answer: T6 is a bad discriminator, because its TDV value is negative.

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?

Answer: Space= 4500 * 8 = 36000 bytes
Disk= 36000 / 1000 = 36

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

Answer: To enhance the precision we can use Special words.
For example: Information…retrieval →Information retrieval
To enhance the recall we can use stem.
For example computer,computing→compute

c. State two advantages of stemming.

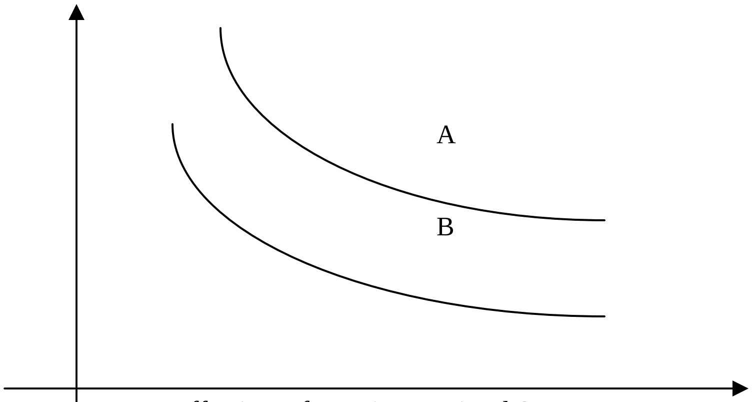
Answer: Store efficient.
Improve recall.

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

Answer:	C	1	O	
	CO	1	M	
	COM	1	P	
	COMP		1	U
	COMPU1		T	
	COMPUT		3	A,E,I

The string “comput” have 3 successor varieties, they are A,E and I.

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



Answer: A represents a more effective Information Retrieval System.
It should not use “efficient” but “effective”.