Computer Engineering Department
Bilkent University

CS533: **Information Retrieval Systems**
Assignment No. 5 - Progressive
May 20, 2014
Due date: May 20, 2014; Tuesday

1. Consider the following document collection containing four documents (rows) defined by four terms (columns)

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

A query is submitted and its vector is defined as follows:

$$Q = [1 \quad 0 \quad 0 \quad 0]$$

Assume that we want to use the MMR algorithm for selecting the best matching first two documents. After each case what is the cohesiveness (similarity) and diversity among the selected documents and how can we measure it? Does the MMR algorithm provide what it promises. For each case please show your steps explicitly. For similarity calculations use the Dice coefficient.

a. Use $\lambda = 1.00$ and indicate the selected documents.

b. Use $\lambda = 0.00$ and indicate the selected documents. What is the diversity among the selected documents and how can we measure it? Do we have more diversity with respect to the first case above? Please explain.

c. Use $\lambda = 0.50$ and indicate the selected documents. What is the diversity among the selected documents and how can we measure it? Do we have more diversity with respect to the first case above? Please explain.

d. Please repeat the steps a to c this time assume that we want to select three documents.

2. Consider a possible use of the MMR approach in news portals for front-page news selection (actually for any information aggregator). In news portals on the front pages we would like to reflect the news agenda and the news agenda is undefined and changes dynamically. In defining the agenda diversification is important, but is it the only thing?

Suggest a way of using the MMR approach for news portal front page selection.

3. Consider the following search engines A, B, C, and D and ranking provided by them for the documents a, b, c, d, e, and f.

A= {a, b, c, d}
B= {b, a, d, f}
C= {b, c, d, a}
D= {a, c, d, e}

Sort the documents according to the following data fusion methods.

a. Reciprocal rank,
b. Borda count,
c. Condorcet.

4. Study the page rank algorithm. Consider its use for social network construction. Define a page rank-based approach for obtaining social network using a collection of news articles.

How can we measure the effectiveness of your method? Please try to define a measurement method.

5. For the following D matrix calculate the TDV by using the cover coefficient concepts (use the approximate method).

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

6. Salton and his co workers define a way of using TDVs for increasing recall and precision in IR. Define their methods. Do you agree or disagree please explain?

Progressive Part

7. Consider the following signatures.
S1: 1000 1001
S2: 1100 0010
S3: 0011 1100
S4: 0000 1111
S5: 1011 0100
S6: 0100 1010
S7: 1100 0101

a. Use the fixed suffix method to partition the above signatures. Take k (key length) as 2. We didn't study this method in class but it is easy to imagine how it works (it is a version of the fixed prefix method but this time it is based on suffixes).

b. Now consider the following queries.
Q1: 1101 0001
Q2: 0110 0011
Q3: 1100 1100
Use the partitions of section-a to calculate the time needed (turnaround time) to process the queries in sequential and parallel environments. (Use the assumptions that we used in the class room, e.g., the processing of one page signature requires 1 time unit, etc.). What is the speed up ratio for the parallel environment (defined as ratio (parallel processing time for all queries / sequential processing time for all queries)?

8. Partition the signatures of question 7 using the following partitioning methods. (To answer this question consider the paper Lee and Leng 1989 paper in ACM TOIS, "Partitioned signature files: Design issues and performance evaluation," or "Signature files: An integrated access method for formatted and unformatted databases" by Aktug & Can. The second one is available on our course Web site. If a password is needed use myirnotes

a. EPP (take z= 2).

b. FKP (take k= 2).

c. To process the following queries which pages need to be accessed and why?

Q1: 1101 0001
Q2: 0110 0011
Q3: 1100 1100

9. Consider the following information filtering profiles used in a Boolean environment.

P1= a, b, c, d
P2= a, f
P3= b, c, f
P4= b, d
P5= a, c, f

Assume that when the terms are sorted in frequency order according to their number of occurrences in documents term a is the least frequently used term in the documents and is also the most frequently used term in the user profiles. The sorted term list continues as b, c … f.

a. Consider the ranked key method explained in the paper by Yan and Garcia-Molina (Index structures for selective dissemination of information under the Boolean model) and draw the directory and the posting lists for the ranked key method.

b. What is the intuition behind the ranked key method: how does it improve the filtering efficiency?

c. Do you agree with the following statement: As time passes automatic update of user profiles may provide higher user satisfaction. Please explain your answer. If your answer is yes suggest an algorithm for this purpose.