SİNEM SAV
21102143
CS533-Assignment5

Bilkent University CS533: Information Retrieval Systems

Assignment – 5

Q1) For the given D matrix, I calculated TDV value for each term by using the formula $n_c = \frac{m*n}{t}$ where m is the number of documents, n is the number of terms and t is the number of non-zero elements in D and formula,

$$TDV_l = n_c - n_{cl}$$

where n_c is the number of clusters using all terms and n_{cl} is the number of clusters without corresponding term. Then,

$$n_c = \frac{(4*4)}{8} = 2$$

For term1 -> $D = \begin{bmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 1\,1\,0 \\ 0\,0\,1 \end{bmatrix}$ (without term1) -> $n_{c1} = \frac{3*4}{5} = 2.4$

TDV_1= 2 − 2.4 = -0.4

For term2 -> $D = \begin{bmatrix} 1\,0\,1 \\ 1\,0\,0 \\ 1\,1\,0 \\ 0\,0\,1 \end{bmatrix}$ (without term2) -> $n_{c2} = \frac{3*4}{6} = 2$

TDV_2= 2 − 2= 0

For term3 -> $D = \begin{bmatrix} 1\,0\,1 \\ 1\,1\,0 \\ 1\,1\,0 \\ 0\,0\,1 \end{bmatrix}$ (without term1) -> $n_{c3} = \frac{3*4}{7} = 1.71$

TDV_3= 2 − 1.71 = 0.29

For term4 -> $D = \begin{bmatrix} 1\,0\,0 \\ 1\,1\,0 \\ 1\,1\,1 \\ 0\,0\,0 \end{bmatrix}$ (without term4) -> $n_{c4} = \frac{3*4}{6} = 2$

TDV_4= 2 − 2 = 0

Q2) Salton and his co-workers state that to improve recall value, we need to use terms that are least frequent in the document frequency spectrum and to improve precision we need to move to the left (see following figure directly taken from [1]) and to optimize them, we need to use best discriminants with medium frequency level where TDV>0. I agree with this approach since using generating such spectrum will force us to find the best discriminants, (by grouping the least frequent terms etc.) and improve the precision and recall value with this indexing scheme.

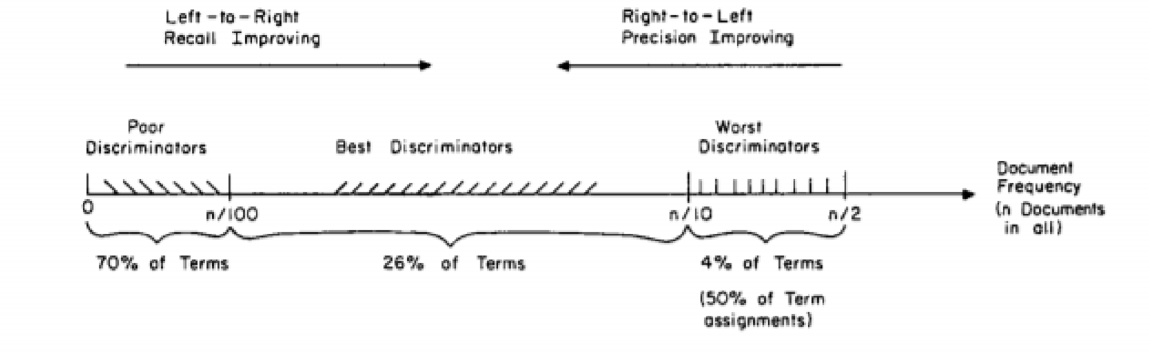


Figure1 – Taken from [1] to show frequency spectrum

Q3) a) For k = 3, we have the following partitioning structure:

000	001	010	011	100	101	110	111
S4	S3	S6		S1, S8	S5, S9	S2, S7	

b) The partitions are considered for false-drop resolution if Q_{sk}& P_{sk} = Q_{sk}

Thus, when we look at first 3 bits of each query and "and" them with partitions:

For Q1 -> activated partitions = (110, 111) = 2

For Q2 -> activated partitions = (011, 111) = 2

For Q3 -> activated partitions = (110, 111) = 2

In sequential environment (turnaround times, assuming each arrived at time 0):

Q1 -> 2 time units

Q2 -> 4 time units

Q3 -> 6 time units

Total = 6 time units

In parallel environment (turnaround times, assuming each arrived at time 0):

PE = Processing Element

SİNEM SAV
21102143
CS533-Assignment5
Table shows the parallel processing sequences:

PE1 (000)				
PE2 (001)				
PE3 (010)				
PE4 (011)	Q2			
PE5 (100)				
PE6 (101)				
PE7 (110)	Q1	Q3		
PE8 (111)	Q1	Q2	Q3	

As each of them are in the sequence of last processing element;

Q1 -> 1 time units

Q2 -> 2 time units

Q3 -> 3 time units

Total = 3 time units

Speed-up ratio: 6/3 = 2

Q4) a) block size = 3, Lf = 2/3 (after reaching Lf=2/3 we can add 3*2/3 = 2 more and then update the file, I assume we start with h = 1:

bv=0, h=1

0	S2	S3	
1	S1	S4	

Lf = 4 / 2*3 = 4/6 = 2/3 -> insert 2 more and then update;

0	S2	S3	S5
1	S1	S4	

→

S6		

UPDATE (**bv=1 , h=1** after the update):

00	S3	S5	
1	S1	S4	
10	S2	S6	

Lf = 6 / 3*3 = 6/9 = 2/3 -> insert 2 more and then update;

SİNEM SAV
21102143
CS533-Assignment5

00	S3	S5	
1	S1	S4	S7
10	S2	S6	S8

UPDATE (**bv=0 , h=2** after update):

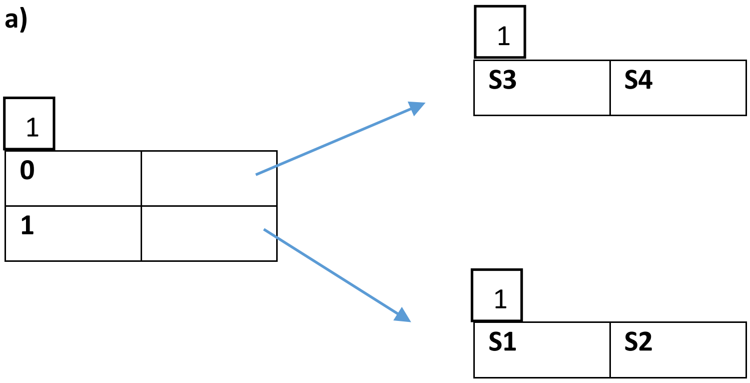
00	S3	S5	
01	S1	S7	
10	S2	S6	S8
11	S4	S9	

Lf = 8 / 4*3 = 8/12 = 2/3 -> insert 2 more (thus, S9 is inserted safely)

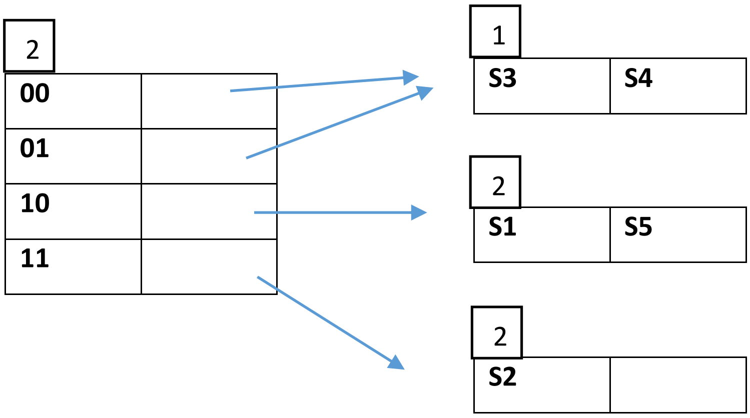
b) Q: 1101 0001 as the last two bits of the query is 01 it will consider the pages **01 and 11** (as it gives $Q_{sk} \& \ P_{sk} = Q_{sk}$)

Q5)

a)

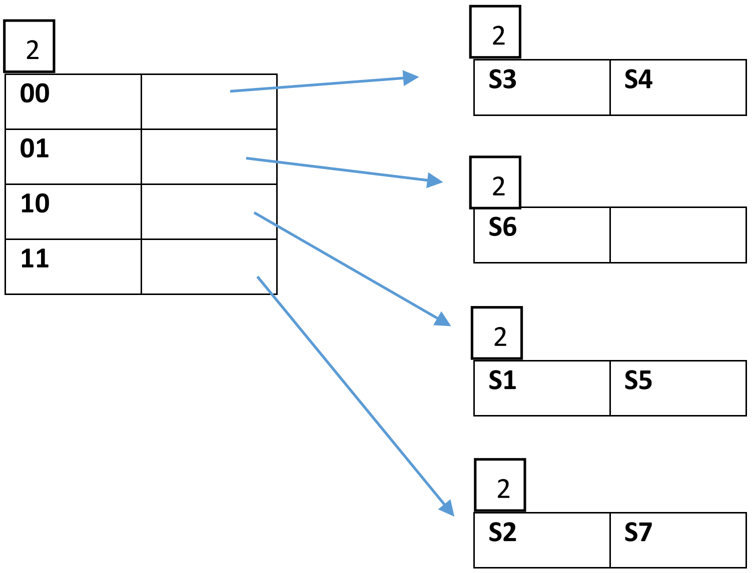


We will insert S5, but no room, thus extend the table:

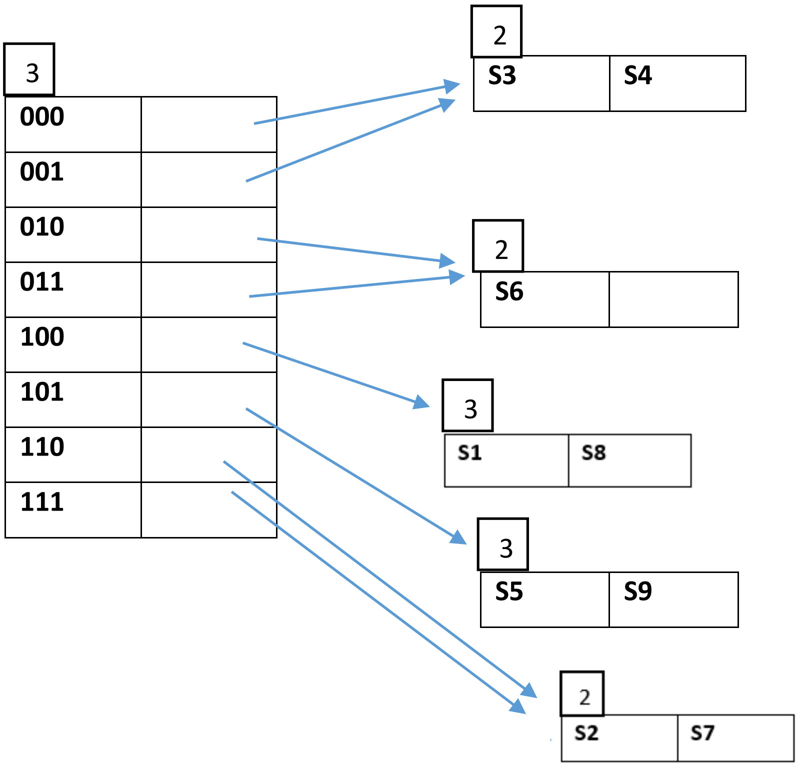


We will insert S6, but no room:

SİNEM SAV
21102143
CS533-Assignment5



We will insert S5, but no room, thus extend the table:

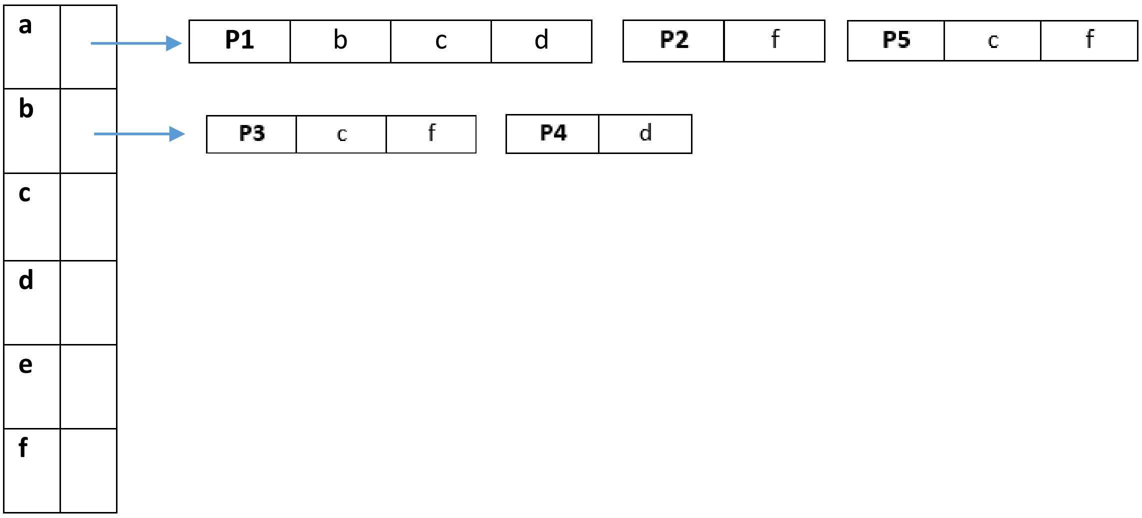


b) For the query Q: 1001 1001 we need to access pages 100,101,110,111 as they give Q_{sk}& P_{sk} = Q_{sk} but 110 and 111 points to the same page already, thus 3 pages accessed.

SİNEM SAV
21102143
CS533-Assignment5

Q6)

a)



b) The intuition behind ranked key method is that the least frequent terms in user profiles are likely to appear most frequently in the documents and with this approach, we ensure that the more frequent words have few number of user profiles associated with them. Therefore, we need to examine fewer number of profiles for an incoming document.

c) Yes, I agree with the statement because as time passes, new documents will be coming. Therefore, one user profile satisfying the incoming document will be matched but with the latest documents, profile terms may also be updated so that new terms that are related to the old ones and the incoming documents (link between them) will be added which will result in higher user satisfaction. In this manner, user will be informed with the new coming documents and with an update.