CS533 - Assignment 1

Hamzeh Ahangari Email: hamzeh@bilkent.edu.tr

Q1

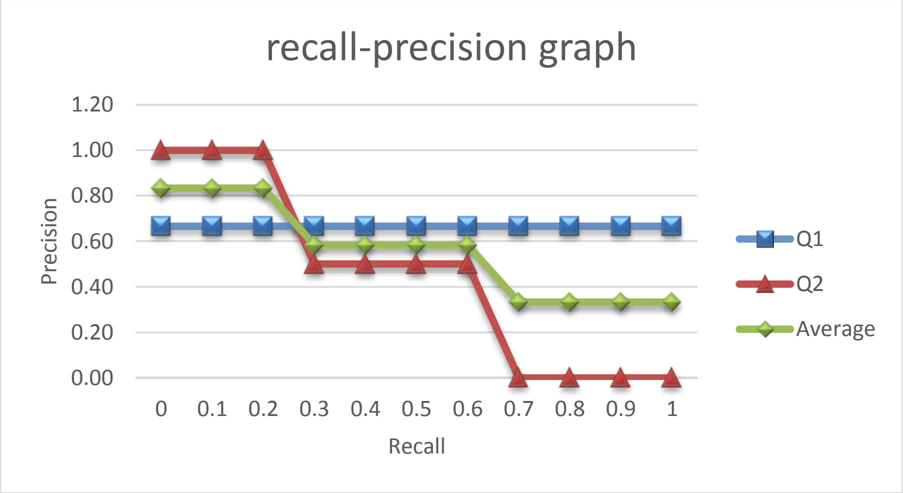
First considering the given queries' results, I made these tables for Q1 and Q2:

Q1 search results (ranked)	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Relevant	N	Y	N	Y	Y	Y	N	N	N	N
Precision	0/1	1/2	1/3	2/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	0/4	1/4	1/4	2/4	3/4	4/4	4/4	4/4	4/4	4/4

Q2 search results (ranked)	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Relevant	Y	N	N	Y	N	Y	N	N	N	N
Precision	1/1	1/2	1/3	2/4	2/5	3/6	3/7	3/8	3/9	3/10
Recall	1/5	1/5	1/5	2/5	2/5	3/5	3/5	3/5	3/5	3/5

A) From above tables and by using TREC "interpolation" rule, interpolated table is calculated and plotted as below:

Recall	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Q1 Prec.	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
Q2 Prec.	1.00	1.00	1.00	0.50	0.50	0.50	0.50	0.00	0.00	0.00	0.00
Avg Prec.	0.83	0.83	0.83	0.58	0.58	0.58	0.58	0.33	0.33	0.33	0.33



b) According to definition, R-Precision is the precision after retrieving R documents, where R is the number of relevant documents for the topic.

- For Q1, R = 4. In first 4 documents, 2 are relevant. Then the R-precision of Q1 is: $\frac{2}{4} = 0.5$
- For Q2, R = 5. In first 5 documents, 2 are relevant. Then the R-precision of Q2 is: $\frac{2}{5} = 0.4$
- Average R-precision = : $\frac{0.5+0.4}{2} = 0.45$

c) According to definition MAP is the arithmetic mean of average precision values for individual information needs. $MPA = \frac{\sum precision\ after\ each\ relevant\ document\ retrieved}{total\ number\ of\ relevant\ documents}$

- For Q1, MAP = $\frac{\frac{1}{2}+\frac{2}{4}+\frac{3}{5}+\frac{4}{6}}{4} = \frac{17}{30} = 0.566$
- For Q2, MAP = $\frac{\frac{1}{1}+\frac{2}{4}+\frac{3}{6}}{5} = \frac{2}{5} = 0.40$
- Average of MAP = $\frac{0.566+0.40}{2} = 0.483$

Q2

- Method 1 (Brute Force) : we go over all Sij by two nested for loop and for each of i-j pair we perform similarity calculation of doc_i & doc_j:

```
for i = 1 to m-1
    for j = i + 1 to m
        find Sij
    end
end
```

S Matrix	d1	d2	d3	d4	d5	d6
d1	1.0	S12	S13	S14	S15	S16
d2	-	1.0	S23	S24	S25	S26
d3	-	-	1.0	S34	S35	S36
d4	-	-	-	1.0	S45	S46
d5	-	-	-	-	1.0	S56
d6	-	-	-	-	-	1.0

In this method the number of document-document similarity pairs are: m*(m-1)/2 = 15

- Method 3 (most efficient algorithm) : First we should build inverted index file lists :
  - t1 -> d1, d3
  - t2 -> d4
  - t3 -> d1, d2, d3
  - t4 -> d2, d4
  - t5 -> d2, d5, d6
  - t6 -> d5, d6

Now we start the algorithm:

- d1 : constructed **mail-box** is

X	1	2	0	0	0

S12, S13 should be calculated, number of similarity calculation = 2

- d2 : constructed **mail-box** is

X	X	1	1	1	1

S23, S24, S25, S26 should be calculated, number of similarity calculation = 4

- d3 : constructed **mail-box** is

X	X	X	0	0	0

Number of similarity calculation = 0

- d4: constructed **mail-box** is

X	X	X	X	0	0

Number of similarity calculation = 0

- d5 : constructed **mail-box** is

X	X	X	X	X	2

S56 should be calculated, number of similarity calculation = 1

- d6 : constructed **mail-box** is

X	X	X	X	X	X

Number of similarity calculation = 0

Total number of similarity calculation = 2 + 4 + 0 + 0 + 1 + 0 = 7

Q3

After calculating similarities values, we have this D matrix:

S Orig	d1	d2	d3	d4	d5	d6
d1	1.0	2/5	4/4	0	0	0
d2	-	1.0	2/5	2/5	2/5	2/5
d3	-	-	1.0	0	0	0
d4	-	-	-	1.0	0	0
d5	-	-	-	-	1.0	4/4
d6	-	-	-	-	-	1.0

Then after that we sort similarities in descending order:

d1-d3	d5-d6	d1-d2	d2-d3	d2-d4	d2-d5	d2-d6	Others
1	1	0.4	0.4	0.4	0.4	0.4	0

- Dendrogram (cluster tree) structure corresponding to the single-link :

d1-d3 = 1	Clusters combined : 1-3,2,4,5,6
d5-d6 = 1	Clusters combined : 1-3,2,4,5-6
d1-d2 = 0.4	Clusters combined : 1-2-3, 4, 5-6 Because d1-d2 is biggest similarity between 1-3 & 2
d2-d3 = 0.4	Already Combined
d2-d4 = 0.4	Clusters combined : 1-2-3-4, ,5-6 Because d2-d4 is biggest similarity between 1-2-3 & 4
d2-d5 = 0.4	Clusters combined : 1-2-3-4-5-6 Because d2-d5 is biggest similarity between 1-2-3-4 & 5-6
All combined, Finish	

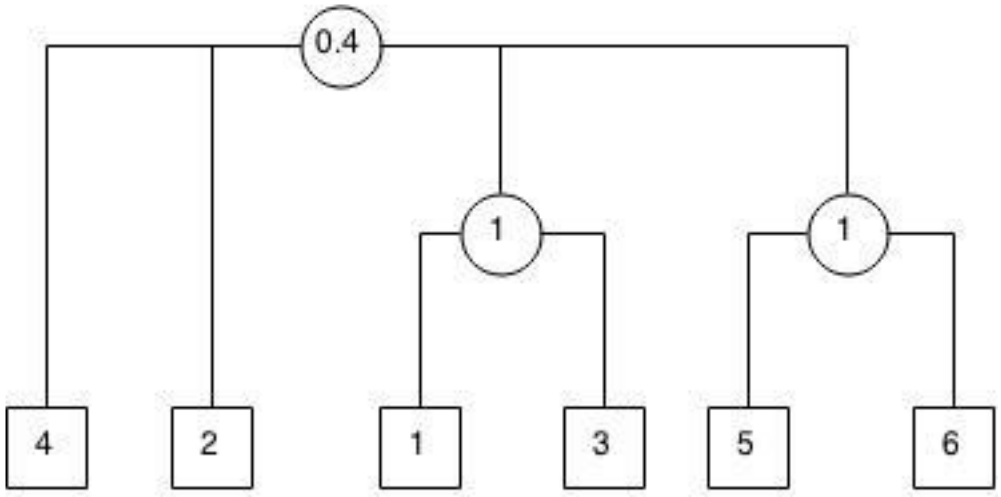


Figure 1. Single-link

Similarity matrix obtained from above clustering tree is:

S Single	d1	d2	d3	d4	d5	d6
d1	1.0	0.4	1	0.4	0.4	0.4
d2	-	1.0	0.4	0.4	0.4	0.4
d3	-	-	1.0	0.4	0.4	0.4
d4	-	-	-	1.0	0.4	0.4
d5	-	-	-	-	1.0	1
d6	-	-	-	-	-	1.0

- $r\ (S\ orig, S\ single) = \frac{cov(S\ orig,S\ single)}{\sqrt{Var(S\ orig).Var(S\ single)}} = \frac{0.1048}{\sqrt{0.1759*0.0743}} = 0.9167$
- I used MATLAB for calculating covariance values.

- Dendrogram (cluster tree) structure corresponding to the complete-link :

d1-d3 = 1	Clusters combined : 1-3,2,4,5,6
d5-d6 = 1	Clusters combined : 1-3,2,4,5-6
d1-d2 = 0.4	Cluster combined : 1-2-3, 4, 5-6 Because d1-d2 is smallest similarity between 1-3 & 2
d2-d3 = 0.4	Already Combined.
d2-d4 = 0.4	No action :Because d2-d4 is not smallest similarity between 1-2-3 & 4
d2-d5 = 0.4	No action :Because d2-d5 is not smallest similarity between 1-2-3 & 5-6
d2-d6 = 0.4	No action : Because d2-d6 is not smallest similarity between 1-2-3 & 5-6
d4-d1 = 0	Clusters combined : 1-2-3-4, 5-6 Because d4-d1 is smallest similarity between 1-2-3 & 4
d5-d1 = 0	Clusters combined : 1-2-3-4-5-6 Because d5-d1 is smallest similarity between 1-2-3-4 & 5-6
All combined, Finish	

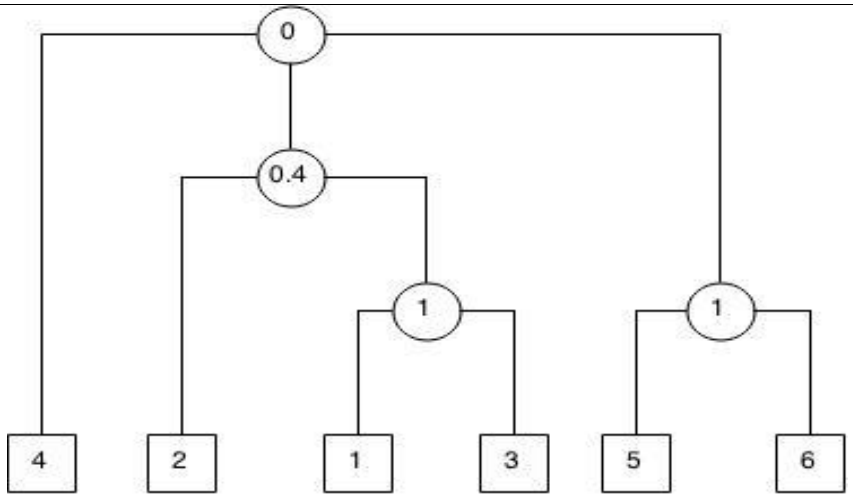


Figure 2. Complete-link

Similarity matrix obtained from above clustering tree is:

S Comp	d1	d2	d3	d4	d5	d6
d1	1.0	0.4	1	0	0	0
d2	-	1.0	0.4	0	0	0
d3	-	-	1.0	0	0	0
d4	-	-	-	1.0	0	0
d5	-	-	-	-	1.0	1
d6	-	-	-	-	-	1.0

- $r\ (S\ orig, S\ Comp) = \frac{cov(S\ orig,S\ single)}{\sqrt{Var(S\ orig).Var(S\ single)}} = \frac{0.1751}{\sqrt{0.1972*0.1759}} = 0.9402$
- It can be seen that Com-link algorithm lasts longer than Simple-link.
- I used MATLAB for calculating covariance values. This number means that S_complete is more correlated to S_Original that S_single. This is because in single-link algorithm we usually loose precision because of enlarged similarity values. While in complete-link usually similarities are more similar to original similarities.

Q4

a) We have $D = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$.

$$\begin{matrix} \alpha_1 = 1/2 \\ \alpha_2 = 1/3 \\ \alpha_3 = 1/2 \\ \alpha_4 = 1/2 \\ \alpha_5 = 1/2 \\ \alpha_6 = 1/2 \end{matrix} \quad S = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad \begin{matrix} \beta_1 = 1/2 \\ \beta_2 = 1/1 \\ \beta_3 = 1/3 \\ \beta_4 = 1/2 \\ \beta_5 = 1/3 \\ \beta_6 = 1/2 \end{matrix} \quad S' = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{2} \end{bmatrix}$$

d5	$\frac{1}{2} * t5$	$\frac{1}{3}$ *d2
		$\frac{1}{3}$ *d5
		$\frac{1}{3}$ *d6
	$\frac{1}{2} * t6$	$\frac{1}{2}$ *d5
		$\frac{1}{2}$ *d6

$$C_{52} = \alpha_5 * \left(\sum_{k=1}^{6} d5k * \beta_k * d2k\right) = \frac{1}{2} * \left(1 * \frac{1}{3} * 1\right) = \frac{1}{6} = 0.1667$$

b)

$$C = \begin{bmatrix} \frac{5}{12} & \frac{1}{6} & \frac{5}{12} & 0 & 0 & 0 \\ \frac{1}{9} & \frac{7}{18} & \frac{1}{9} & \frac{1}{6} & \frac{1}{9} & \frac{1}{9} \\ \frac{5}{12} & \frac{1}{6} & \frac{5}{12} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & \frac{5}{12} & \frac{5}{12} \\ 0 & \frac{1}{6} & 0 & 0 & \frac{5}{12} & \frac{5}{12} \end{bmatrix}$$

c) Number of clusters Nc is sum of main diagonal of matrix C, then :

$$n_c = \frac{5}{12} + \frac{7}{18} + \frac{5}{12} + \frac{3}{4} + \frac{5}{12} + \frac{5}{12} = 2.805 \ \approx 3$$

d) We should calculate seed powers :

$$p_1 = \left(1-\frac{5}{12}\right)*\frac{5}{12}*2 = 0.4861 \qquad p_2 = \left(1-\frac{7}{18}\right)*\frac{7}{18}*3 = 0.7130$$
$$p_3 = \left(1-\frac{5}{12}\right)*\frac{5}{12}*2 = 0.4861 \qquad p_4 = \left(1-\frac{3}{4}\right)*\frac{3}{4}*2 = 0.3750$$
$$p_5 = \left(1-\frac{5}{12}\right)*\frac{5}{12}*2 = 0.4861 \qquad p_6 = \left(1-\frac{5}{12}\right)*\frac{5}{12}*2 = 0.4861$$

Seed 1 : d2

Seed 2&3 :

We should select among p1, p3, p5, p6. Since d1=d3 and d5=d6, then we choose:

Seed 2 : d1

Seed 3 : d5

e)

t1 -> d1,d3 t2 -> d4 t3 -> d1,d2,d3

t4 -> d2,d4 t5 -> d2,d5,d6 t6 -> d5,d6

For example for calculation C56 we have:

Initial: C56 = 0;

For t5 : C56 = C56 + (1/2)* (1/3) = 1/6

For t6 : C56 = C56 + (1/2)* (1/2) = 1/6 + ¼ = 5/12

f) For clustering the none-seed documents :

C31 > C32 > c35 => d3 = > d1

C42 > c41 > c45 => d4 = > d2

C65 > c61 > c62 => d6 = > d5

Finally we have these clusters: d1&d3, d2&d4, d5&d6

Q5

For this, we should show that the sum of main diagonal in both matrix C & C' are same. For n_c we have:

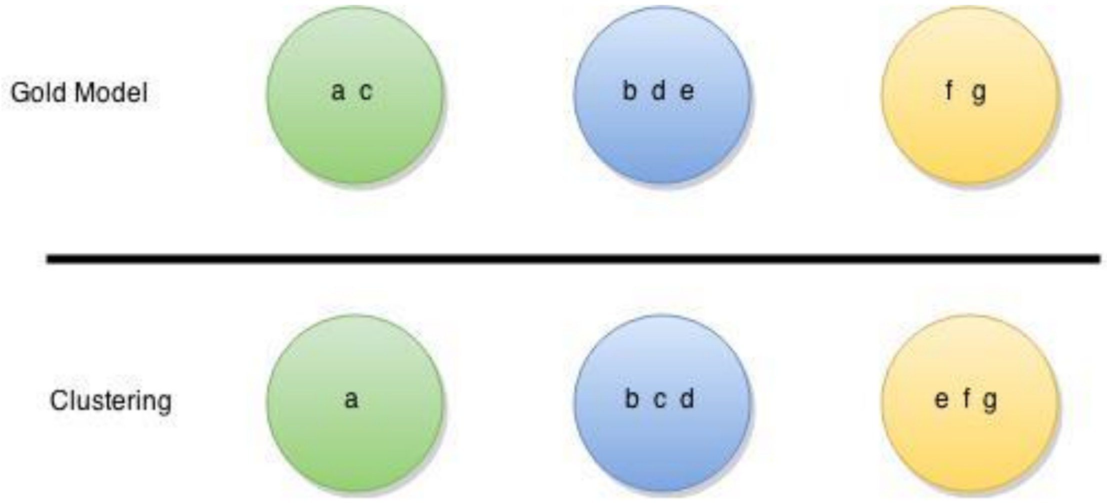
$$n_c = \sum_{i=1}^{m} c_{ii} = \sum_{i=1}^{m} \alpha_i \sum_{k=1}^{n} \beta_k d_{ik} d_{ik} = \sum_{i=1}^{m}\sum_{k=1}^{n} \alpha_i \beta_k d_{ik}$$

Now by swapping the summations we have:

$$n_c = \sum_{k=1}^{n}\sum_{i=1}^{m} \alpha_i \beta_k d_{ik} = \sum_{k=1}^{n} \beta_k \sum_{i=1}^{m} \alpha_i d_{ik} = \sum_{k=1}^{n} \beta_k \sum_{i=1}^{m} \alpha_i d_{ik} d_{ik} = \sum_{k=1}^{n} c'_{ii} = n'_c$$

Q6

I take the first one our Gold Model and second one our imaginary obtained clustering



Total number of pairs = $\binom{7}{2} = 21$

ab	ac	ad	ae	af	ag	bc	bd	be	bf	bg
TN	FN	TN	TN	TN	TN	FP	TP	FN	TN	TN
cd	ce	cf	cg	de	df	dg	ef	eg	fg	
FP	TN	TN	TN	FN	TN	TN	FP	FN	TP	

$$RI = \frac{TN + TP}{ALL} = \frac{12 + 2}{21} = 0.66$$

Q7

a) It is assumed that (mentioned in part b) inverted index list are sorted by document number. Without skipping we should do these algorithm :

```
a_ptr= start; b_ptr = start;
while (not end of b)
    while ( a_ptr_doc_num < b_ptr_doc_num )
        a_ptr = next
    end
    // proper place is found
    b_ptr = next
end
```

For given list of term-a and term-b we need these comparisons **(totally 15):**
b_ptr = <10,2> : is comparted to a_ptr= <1, 2> <3, 1> <8, 2> <10, 3> <12, 4> **(5 comparisons)**
b_ptr = <56,1> : is comparted to a_ptr= <12, 4> <17, 4> <17, 4>, <22, 3> <24, 2> <33, 4> <38, 5> <43, 5> <55, 3><64, 2> **(10 comparisons)**

If we choose the chunk size = 5 : then we have this structure :
Chunk 1 : <1, 2> <3, 1> <8, 2> <10, 3> <12, 4>
Chunk 2 : <17, 4> <17, 4>, <22, 3> <24, 2> <33, 4>
Chunk 3 : <38, 5> <43, 5> <55, 3><64, 2> <68, 4>
Chunk 4 : <72, 5> <75,5> <88, 2>

Assuming last element is indicator of each chunk, we need these comparisons **(totally 12):**
b_ptr = <10,2> : is comparted to a_ptr= <12, 4>_<1, 2> <3, 1> <8, 2> <10, 3> **(5 comparisons)**
b_ptr = <56,1> : is comparted to (continued from last a_ptr) a_ptr= <12, 4> <33, 4> <68, 4><38, 5> <43, 5> <55, 3><64, 2> **(7 comparisons)**

If the size of chunks are big, then skipping is done very fast and effectively, however search inside the chunks is time consuming. On the other hand, if the size of chunks are small (and number them be large) search inside chunks is fast, but skipping is not done very effective. Consequently there is an **optimum point** for size of chunks which should be found for each system.

b)

- Posting list ordered by f(d,t) in descending order: <75,5> <72, 5> <43, 5> <38, 5><68, 4> <33, 4> <17, 4> <17, 4> <12, 4> <55, 3> <22, 3> <10, 3> <88, 2> <64, 2> <24, 2> <8, 2> <1, 2> <3, 1>
- Posting list frequency information in prefix form : <5:4, 75,72,43,38> <4:5, 68,33,17,17,12> <3:3, 55,22,10> <2:5, 88,64,24,8,1> <3:1, 1>

The advantage of approach b over approach a is : approach b saves more space by not storing frequency numbers multiple times.

Q8

a) Pattern recognition like image or voice processing, is an important area in which clustering is in use. Data with similar patterns, like similar type of eyes or hair, are clustered together to train a system. Later fresh test data are given to system and by using previously recognized and clustered data, patterns in these new test data are detected.

b) Clustering tendency is the intrinsic inclination of a data set to be clustered into some groups. It means that some groups of data have a discriminative characteristic that differentiate them from others. By this characteristic, those groups of data tend to be grouped together inside one cluster. Using a gauge of “clustering tendency” during clustering makes sense. When there is nothing that differentiate data from each other, it is logical to avoid cluster them anymore. I guess factors like covariance which shows the dependency between data can be used as a gauge for this purpose somehow.

Q9

Our approaches to incorporating a new data point into a cluster hierarchy incrementally can be divided into two stages. During the first stage, the algorithm locates a node in the hierarchy that can host the new data point. The second stage performs hierarchy restructuring.

1. Locating proper place :
   We start from the leaf with the highest similarity with the new document. Starting from the parent of the closest leaf node, perform upward search to locate a cluster (or create a new cluster hierarchy) that can host the new point with minimal density changes and minimal disruption of the hierarchy monotonicity.
2. inserting new document to proper place :
   The insertion algorithm is called reconstructing operation. Five different situations (split,merge, demote,…) are possible for structure update, which are explained in source paper.

Source:
Widyantoro, Dwi H., Thomas R. Ioerger, and John Yen. "An incremental approach to building a cluster hierarchy." Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002.

Q10

For a specific IR system, test collection is composed of a set of queries and a set of documents with known relevance. This collection is used by a test system to investigate the ability, performance or quality of an IR system to detect relevant objects.

In pooling approach top k documented returned from IR systems are assumed to be relevant, collected inside a so called pool, and considered for evaluation.

Mentioned article has proposed some criticisms about pooling method. First the data collected inside pool are expected to be unbiased, however in practice they are not as expected. Another important factor which is discussed is pooling depth. Writer discussed what will happen if pooling depth is not suitable. Some experimental test are provided. Then dynamically varying pool size, for a new query is assessed.

Q11

Cluster Hypothesis definition as we had in the class: documents similar to each other would be relevant to the same query and would appear in the same cluster.

Real life examples which follows this hypothesis:

- When we check a booking site for hotel reservation, the booking site usually shows us similar and relevant options.
- When we decided to buy a particular laptop, online or from the shop, we usually see similar (same class) laptops around the one we want to buy.
- In nature, dependent on the climate and geographical position, known species of animals live around each other.

Real life examples which don't follows this hypothesis:

- When we go to a cell phone shop, and seeking for a cheap one, we usually see phones with all kind of prices near each other.
- Sometimes we expect products (like car), with nearly same quality should has almost same prices, but it is not always true. Perhaps because of a special discount which we are not aware of, something similar is very cheaper. But we miss to buy it, since we think it has lower quality.

Finish