Information Retrieval

HOMEWORK #3

Information Retrieval

HOMEWORK #3

Section A (Solutions are due to Caner Mercan.)
)

1. First, we need to construct the similarity matrix in order to obtain the corresponding single-link clustering structure. For the similarity, we use overlap coefficient which is,

$$\text{Overlap coefficient}(\text{X}, \text{Y}) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

The similarity matrix obtained by the overlap coefficient is as follows;

$$S = \begin{bmatrix} 1 & 0.5 & 0 & 0.5 & 0.5 \\ - & 1 & 1 & 0.75 & 0.5 \\ - & - & 1 & 1 & 0 \\ - & - & - & 1 & 1 \\ - & - & - & - & 1 \end{bmatrix}$$

Let's have the documents sort pairwise by their similarity values;

Document Pairs	Similarity Values
D_2 D_3	1
D_3 D_4	1
D_4 D_5	1
D_2 D_4	0.75
D_1 D_2	0.5
D_1 D_4	0.5
D_1 D_5	0.5
D_2 D_5	0.5
D_1 D_3	0
D_3 D_5	0

The corresponding dendrogram is as follows;

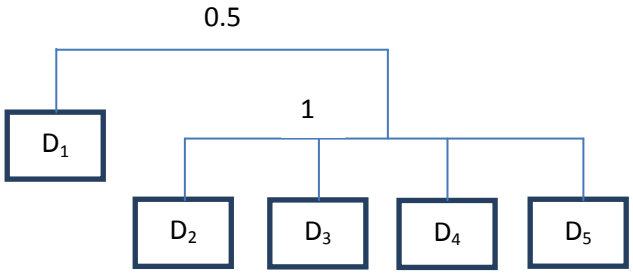

Figure 1 Single-link Clustering Structure

2. In the first question, I have obtained similarity matrix by using the overlap coefficient. Thus, I will focus on the Complete Link structure with the similarity matrix I obtained in the first question. The corresponding dendrogram is shown in Figure 2.

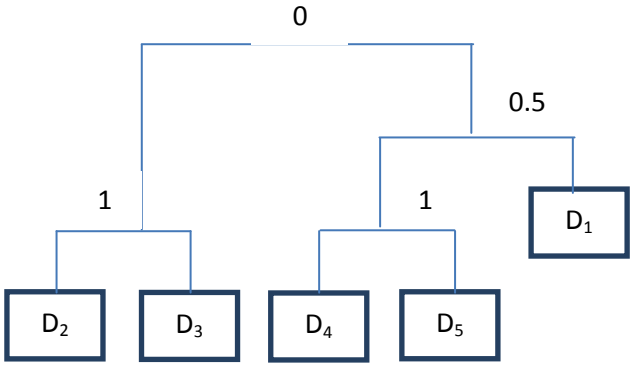

Figure 2 Complete-link Clustering Structure

In a threshold value between 0 and 0.5, we can obtain two different clusters. However, there is also another thing to consider which is the value linking the D_2 D_3 and D_4 D_5 D_1 clusters. It is zero in Complete Link Clustering scheme. Thus, in this scheme, we initially obtain two different clusters with a linkage value being pretty low.

3. In its nature, the complete link clustering structure is order dependent. Since when two sub-clusters have the same similarity value, changing the order of choosing them also alters the whole structure. In our D matrix, we have such situation, even only in the beginning, if we swap the positions of the S_{23} and S_{34}, we change the structure altogether. In order to overcome order dependency, we can prevent any sub clusters with the same document to have same similarity value. This way, the order dependency can be overcome. Since the original similarity matrix is order dependent, I generate another D matrix which is order independent when applied complete link clustering structure (since single link clustering structure is order independent, it does not need any additional constraints). The generated similarity matrix based on the original one is shown below:

$$S' = \begin{bmatrix} 1 & 0.5 & 0 & 0.5 & 0.5 \\ - & 1 & 1 & 0.75 & 0.5 \\ - & - & 1 & 0.9 & 0 \\ - & - & - & 1 & 1 \\ - & - & - & - & 1 \end{bmatrix}$$

The alteration of S_{34} from 1 to 0.9 allowed this similarity matrix to be order independent. In whatever order, when the similarity of two docs are the same, does not change the complete-link clustering structure now.

4. In order to obtain the document by document similarity matrix with the term inverted indexes, first we need to create the posting list. The posting list of the terms for the D matrix given in the first problem is as follows:

$t_1 \rightarrow$ <2,1>,<4,1>
$t_2 \rightarrow$ <2,1>
$t_3 \rightarrow$ <1,1>,<2,1>,<4,1>,<5,1>
$t_4 \rightarrow$ <1,1>
$t_5 \rightarrow$ <4,1>,<5,1>
$t_6 \rightarrow$ <2,1>,<3,1>,<4,1>

And the doc length info is as follows; $D_1 = 2$, $D_2 = 4$, $D_3 = 1$, $D_4 = 4$, $D_5 = 2$.

For D_1;

It contains t_3 and t_4, thus we should proceed with these;

	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
Process t_3	X	1	0	1	1
Process t_4	X	1	0	1	1

By looking at the table, we can see that the only similarity values we should compute are **S_{12}** and **S_{14}**. With the overlap coefficient;

S_{12} = ½ = 0.5, S_{14} = ½ =0.5 and S_{15} = ½ =0.5

For D_2;

It contains t_1, t_2, t_3 and t_6, thus we should proceed with these;

	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}
Process t_1	X	X		1	
Process t_2	X	X			
Process t_3	X	X		2	1
Process t_6	X	X	1	3	

By looking at the table, we can see that the similarity values we should compute are **S_{23}, S_{24}** and **S_{25}**. With the overlap coefficient;

S_{23} = 1/1 = 1, S_{24} = ¾ =0.75 and S_{25} = ½ =0.5

For D_3;

It contains t_6, thus we should proceed with these;

	S_{31}	S_{32}	S_{33}	S_{34}	S_{35}
Process t_6	X	X	X	1	

By looking at the table, we can see that the similarity values we should compute are **S_{34}.** With the overlap coefficient;

S_{34} = 1/1 = 1.

For D_4;

It contains t_1, t_3, t_5 and t_6 thus we should proceed with these;

	S_{41}	S_{42}	S_{43}	S_{44}	S_{45}
Process t_1	X	X	X	X	
Process t_3	X	X	X	X	1
Process t_5	X	X	X	X	2
Process t_6	X	X	X	X	

By looking at the table, we can see that the similarity values we should compute are **S_{45}.** With the overlap coefficient;

S_{45} = 2/2 = 1.

In the light of the similarity values, the newly constructed Similarity Matrix by using the term inverted indexes is as follows;

$$S'' = \begin{bmatrix} 1 & 0.5 & 0 & 0.5 & 0.5 \\ - & 1 & 1 & 0.75 & 0.5 \\ - & - & 1 & 1 & 0 \\ - & - & - & 1 & 1 \\ - & - & - & - & 1 \end{bmatrix}$$

And, this is obviously the same with the similarity matrix we obtained in the first question. But, this time, it's been in a more efficient manner.

5. a. Here, the C matrix is obtained by the multiplication of inverse of row sums and the transpose of inverse of column sums.

The inverse of row sum matrix is as follows;

$$IRS = \begin{bmatrix} 0 & 0 & 0.50 & 0.50 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0 & 0 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0.25 & 0 & 0.25 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0.50 & 0 & 0.50 & 0 \end{bmatrix}$$

And the inverse of column sum matrix;

$$ICS = \begin{bmatrix} 0 & 0 & 0.25 & 1 & 0 & 0 \\ 0.50 & 1 & 0.25 & 0 & 0 & 0.33 \\ 0 & 0 & 0 & 0 & 0 & 0.33 \\ 0.50 & 0 & 0.25 & 0 & 0.50 & 0.33 \\ 0 & 0 & 0.25 & 0 & 0.50 & 0 \end{bmatrix}$$

After constructing these matrices, we can find the C matrix by using inner product on IRS and the transpose of ICS;

$$C = IRS\ x\ ICS^{T}$$

$$C = \begin{bmatrix} 0.6250 & 0.1250 & 0 & 0.1250 & 0.1250 \\ 0.0625 & 0.5200 & 0.0825 & 0.2700 & 0.0625 \\ 0 & 0.3300 & 0.3300 & 0.3300 & 0 \\ 0.0625 & 0.2700 & 0.0825 & 0.3950 & 0.1875 \\ 0.1250 & 0.1250 & 0 & 0.3750 & 0.3750 \end{bmatrix}$$

b. The number of clusters is calculated with the following formula;

$$n_c = \sum_{i=1}^{n} c_{ii}$$

$$n_c = 2.2450 \cong 2$$

The number of clusters is most likely either 2 or 3.

c. Cluster seed power is calculated with the following formula;

$$p_i = \delta_i . \psi_i . x_{d_i}$$

where

$$\delta_i = c_{ii}$$

$$\psi_i = 1 - c_{ii}$$

$$x_{d_i} = depth\ of\ indexing\ for\ d_i$$

Thus, in the light of the formula, the results are as follows;

$$p_1 = 0.625 \,.\, (1 - 0.625) \,.\, 2 = 0.469$$

$$p_2 = 0.520 \,.\, (1 - 0.520) \,.\, 4 = 0.998$$

$$p_3 = 0.333 \,.\, (1 - 0.333) \,.\, 1 = 0.217$$

$$p_4 = 0.395 \,.\, (1 - 0.395) \,.\, 4 = 0.956$$

$$p_5 = 0.375 \,.\, (1 - 0.375) \,.\, 2 = 0.469$$

d. We concluded that the number of clusters is either 2. I have chosen 2 as the cluster seeds. In the previous part, we have obtained the cluster power seeds and by the light of the results, it is clear that d_2 and d_4 are cluster seeds as p_2 and p_4 returned the highest values in terms of cluster seed power. Thus, cluster seeds are d_2 and d_4 and the non-seeds are d_1, d_3 and d_5.

e. The IISD I have constructed is as follows;

$t_1 \rightarrow$ <2,1>,<4,1>
$t_2 \rightarrow$ <2,1>
$t_3 \rightarrow$ <2,1>,<4,1>
$t_4 \rightarrow$ <>
$t_5 \rightarrow$ <4,1>
$t_6 \rightarrow$ <2,1>,<4,1>

f. We can obtain the values from the c matrix we previously obtained. However, explicit computation is as follows;

$$c_{ij} = \alpha_i * \sum_{k=1}^{6} d_{ik} * \beta_k * d_{jk}$$

What we need to compute are;

$t_1 \rightarrow c_{54}$

$t_3 \rightarrow c_{52}, c_{54}$

$t_5 \rightarrow c_{54}$

$t_6 \rightarrow c_{52}, c_{54}$

In other words, only c_{52},c_{54} need to be calculated;

$$C_{52} = \frac{1}{2} * \left(0 + 0 + \frac{1}{4} + 0 + 0 + 0\right) = 0.125$$
$$C_{54} = \frac{1}{2} * \left(0 + 0 + \frac{1}{4} + \frac{1}{2} + 0 + 0\right) = 0.375$$

Since it has the highest value, the cluster of d_4 is chosen for d_5.

g. Like in the previous example, we need to look at the C matrix to find the clusters. We have found that d_4 and d_5 should be in the same cluster and by looking at the C matrix, we see that for d1 and d3, it does not matter whether we choose C_2 or C_4. We can conclude that, the two clusters with the seeds d_2 and d_4 are as follows;

{d_1, d_2, d_3}, {d_4, d_5}.

6. The problem of cluster maintenance deals with the modification of clustering structures due to the addition of new documents or deletion of old documents (or both). A close examination of clustering algorithms reveals that most of them are unsuitable for cluster maintenance. In the literature there are very few maintenance algorithms. In general, these algorithms are developed for growing databases; most of them, however, can also be used for document deletion. We can use C^3M to autonomously maintain the clustering structure when a new sample arrives or a sample deleted from the cluster space. Since the C^3M allows us to play with the clusterin structure based on the C matrix, it is very easy to add or remove a new instance to the cluster space.

7. a. In the case of documents being randomly distributed among the clusters;

$$P_j = 1 - \prod_{i=1}^{k} \frac{m_j - i + 1}{m - i + 1}$$

And, the expected number of blocks to be assessed;

$$E = \sum_{i=1}^{m} Pi$$

a. Yao's formula for the probability of accessing the cluster C_j is as follows;

$$E = m * \left[1 - \prod_{i=1}^{k} \frac{n - n/m - i + 1}{n - i + 1}\right]$$

where,

- E *is the expected number of blocks to be assessed*
- m *is the number of blocks*
- k *is the number of relevant records*
- n *is the number of records*

With the given values placed into the equation; the result is computed as;

$$E = 10 * \left[1 - \prod_{i=1}^{5} \frac{150 - 15 - i + 1}{150 - i + 1}\right] = 4.1795$$

8. The similarity matrix implied by the dendrogram (the single-link clustering structure) is as follows;

$$S^d = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ - & 1 & 1 & 1 & 1 \\ - & - & 1 & 1 & 1 \\ - & - & - & 1 & 1 \\ - & - & - & - & 1 \end{bmatrix}$$

And the original similarity matrix;

$$S = \begin{bmatrix} 1 & 0.5 & 0 & 0.5 & 0.5 \\ - & 1 & 1 & 0.75 & 0.5 \\ - & - & 1 & 1 & 0 \\ - & - & - & 1 & 1 \\ - & - & - & - & 1 \end{bmatrix}$$

The product moment correlation between X and Y is defined as follows;

$$r = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}$$

For our problem, when the equation is calculated, the resulting correlation is given as follows;

$$r(X, Y) = 0.5435$$

9. He says that selecting an appropriate cluster method and implementing it are two separate problems and thus, Cluster methods are more profitably discussed at the level of abstraction at which relations are discussed in connection with data bases, that is, in a thoroughly data independent way. This way of looking at clustering file structures makes sense when delving into deeper on the hardware file structure implementations. It is one of the basic things to cluster file structures without actually having the files. The clustering methodology might not be dependent on the data itself. But, on the other hand, it is pretty mainstream to cluster file structures based on the data they cover. All in all, the two approaches have their upsides and downsides.

Section B

1. In the supervised classification scheme, we have the initially labeled training data which have the true labels given the data and another data which is called test data that consist of only data and no other additional label information. In the case of training, the classifier tries to learn a modal based on the training samples and the label information and then, when encounters with new data, come up with a label. However, in the case of the clustering, it is impossible to train a sample to learn a modal since we have no predefined labels for any of the instances. In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data.
2. In the paper, it mentions about five components of a clustering task;
    - Pattern representation: The term and document matrices of the information.
    - Definition of a pattern proximity measure appropriate to the data domain; similarity measures such as dice coefficient.
    - Clustering or grouping; Such as Agglomerative hierarchical clustering scheme
    - Data abstraction; extracting a good representation of the dataset, e.g. excluding the typos from the dataset.
    - Assessment of output; Evaluation of the significance of the output. Paired t-test and many more can be used for this purpose.
3. Cluster tendency analysis is defined as follows; the input data are examined to see if there is any merit to a cluster analysis prior to one being performed.
4. The process of K-Means is as follows; Choose k cluster centers to coincide with k randomly-chosen patterns or k randomly defined points inside the hyper volume containing the pattern set. Assign each pattern to the closest cluster center. Recompute the cluster centers using the current cluster memberships. If a convergence criterion is not met, go to assigning step. Typical convergence criteria are: no (or minimal) reassignment of patterns to new cluster centers, or minimal decrease in squared error.
5. (1) Image segmentation
   (2) Object recognition
   (3) Document retrieval
   (4) Data mining

In image segmentation, the image segmentation is dependent on the similarity and dissimilarity of the consecutive pixel values(given the image is clear) Thus, clustering the image with respect to spatial information, color, gradient etc. is of utmost importance in segmentation of the image. There are also hierarchical clustering algorithms to segment the image by choosing different threshold values or clustering the image in the frequency domain etc.
In object recognition, there are many possible views of a 3D object and one goal of the studies is to avoid matching an unknown input view against each image of each object. A common theme in the object recognition literature is indexing, wherein the unknown view is used to select a subset of views of a subset of the objects in the database for further comparison, and rejects all other views of objects. One of the approaches to indexing employs the notion of view classes; a view class is the set of qualitatively similar views of an object. The view classes are identified by clustering.

Section C

1. The experimental findings reported in the literature are in general favorable to clustering engines, suggesting that they may be more effective than plain search engines both in terms of the amount of relevant information found and the time necessary to find it. However, we must be cautious because these results may be biased by unrealistic usage assumptions or unrepresentative search tasks. To date, the evaluation issue has probably not yet received sufficient attention and there is still a lack of conclusive experimental findings for demonstrating the superior retrieval performance of clustering engines over search engines, at least for some type of queries.
2. I tried most of the cluster based search engines mentioned in the paper but among the ones that have a web page with a search engine, none of them worked properly, I got “no results” or “no document returned” and such. Additionally, I could not find the search engine at all. I believe this is due to the time passed on those studies. Most of them are no longer in working condition, or cannot even accessed on web.