CS533 Information Retrieval Systems HW2

Sefa Kılıç

March 23, 2010

1. Consider the following search results for two queries Q1 and Q2 (the documents are ranked in the given order, the relevant documents are shown in bold).

   Q1: **D1**, D2, **D3**, **D4**, D5, **D6**, D7, D8, D9, D10.

   Q2: **D1**, D2, **D3**, D4, D5, **D6**, D7, D8, D9, D10.

   For Q1 and Q2 the total number of relevant documents is, respectively, 4 and 5 (Q2 two of the relevant documents are not retrieved).

   (a) Using TREC interpolation rule, in a table give the precision value for the 11 standard recall levels 0.0, 0.1, 0.2, ... 1.0. Please also draw the corresponding recall-precision graph as shown in the first figure of TREC-6 Appendix A (its link is available on the course web site).

   Please do this for each query separately and obtain one table for both queries using the average of two values at each recall point.

   (b) Find R-Precision (TREC-6 Appendix A for definition) for Query1 and Query2.

   (c) Find MAP for these queries.

Solution:

(a) For Q1, total number of relevant documents is 4. Actual recall& precision table is as follows:

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	Yes	No	Yes	No	No	No	No
Precision	1/1	1/2	2/3	3/4	3/5	4/6	4/7	4/8	4/9	4/10
Recall	1/4	1/4	2/4	3/4	3/4	4/4	4/4	4/4	4/4	4/4

Interpolated recall&precision table is as follows:

Precision	1	1	1	3/4	3/4	3/4	3/4	3/4	4/6	4/6	4/6
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

For Q2, total number of relevant documents is 5 and actual recall&precision table is as follows:

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Yes	No	Yes	No	No	Yes	No	No	No	No
Precision	1/1	1/2	2/3	2/4	2/5	3/6	3/7	3/8	3/9	3/10
Recall	1/5	1/5	2/5	2/5	2/5	3/5	3/5	3/5	3/5	3/5

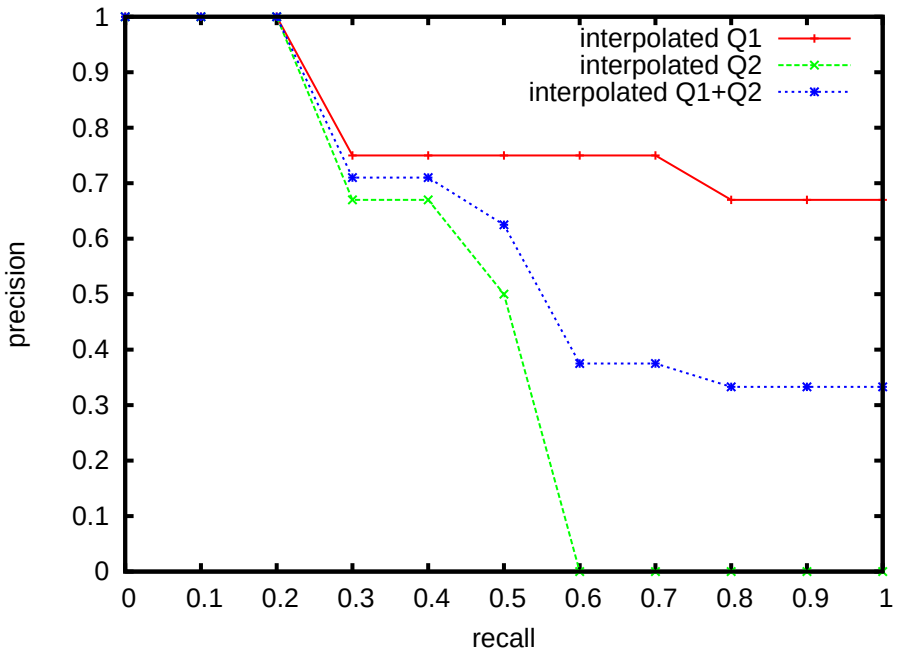
Interpolated recall&precision table is as follows:

Precision	1	1	1	2/3	2/3	3/6	0	0	0	0	0
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

For Q1+Q2, average interpolated recall&precision table is as follows:

Precision	1	1	1	17/24	17/24	5/8	3/8	3/8	1/3	1/3	1/3
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Interpolated recall-precision curves are given below.



(b)
- For Q1, 3 relevant documents are retrieved in the top 4 documents, so R-Precision = 3/4.
- For Q2, 2 relevant documents are retrieved in the top 5 documents, so R-Precision = 2/5.

(c) Average precision for Q1: $\frac{1+2/3+3/4+4/6}{4} = 0.77$
Average precision for Q2: $\frac{1+2/3+3/6}{5} = 0.43$
MAP $= (0.77 + 0.43)/2 = 0.6$.

2. Consider the following document by term binary D matrix for $m = 6$ documents (rows), $n = 6$ terms (columns).

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

Consider the problem of constructing a document by document similarity, S matrix. How many similarity coefficients will be calculated using the following methods? For each case

explain your answer briefly, give exact numbers for each document and explain how you come up with those numbers.

(a) Straightforward approach (using document vectors) -the 1st method discussed in the class-.

(b) Using term inverted indexes.

Solution:

(a) Since $S_{ij} = S_{ji}$ and $S_{ii} = 1$, we only need to calculate following matrix elements:

$$\begin{bmatrix} 1.0 & S_{1,2} & S_{1,3} & S_{1,4} & S_{1,5} & S_{1,6} \\ \times & 1.0 & S_{2,3} & S_{2,4} & S_{2,5} & S_{2,6} \\ \times & \times & 1.0 & S_{3,4} & S_{3,5} & S_{3,6} \\ \times & \times & \times & 1.0 & S_{4,5} & S_{4,6} \\ \times & \times & \times & \times & 1.0 & S_{5,6} \\ \times & \times & \times & \times & \times & 1.0 \end{bmatrix}$$

In straightforward approach, $n(n-1)/2 = 15$ similarity calculations are required.

(b)

t1 →d1, d3
t2 →d2, d4
t3 →d1, d3
t4 →d2, d4
t5 →d2, d5, d6
t6 →d5, d6

Consider d1: d1 contains t1, t3
- t1 →d1, d3 ($S_{1,3}$)
- t3 →d1, d3 (-)

Consider d2: d2 contains t2, t4, t5
- t2 →d2, d4 ($S_{2,4}$)
- t4 →d2, d4 (-)
- t5 →d2, d5, d6 ($S_{2,5}, S_{2,6}$)

Consider d3: d3 contains t1, t3
- t1 →d1, d3 (-)
- t3 →d1, d3 (-)

Consider d4: d4 contains t2, t4
- t2 →d2, d4 (-)
- t4 →d2, d4 (-)

Consider d5: d5 contains t5, t6
- t5 →d2, d5, d6 ($S_{5,6}$)
- t6 →d5, d6 (-)

Consider d6: d6 contains t5, t6
- t5 →d2, d5, d6 (-)
- t6 →d5, d6 (-)

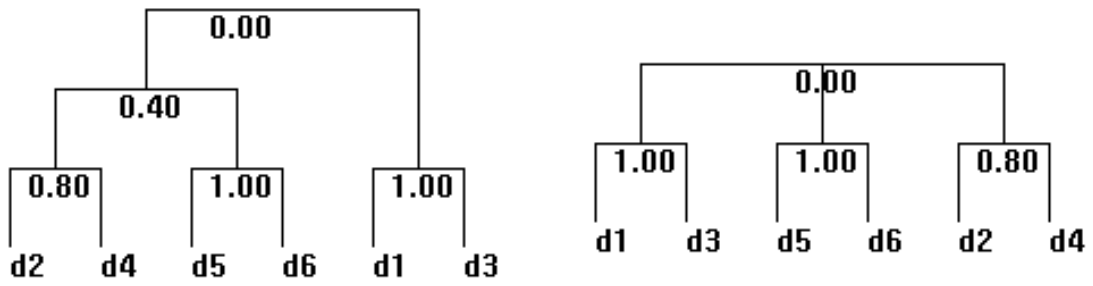
Total similarity calculations: 5.

3. Obtain the similarity matrix S for the above D matrix (you don't need to show your intermediate steps). Use the Dice similarity coefficient. Use the S matrix to construct dendrogram (cluster tree) structure corresponding to the single-link and complete-link clustering methodologies.

Solution: D matrix and corresponding S matrix are as follows:

$$D = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \qquad S = \begin{bmatrix} 1.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ & 1.0 & 0.0 & 0.8 & 0.4 & 0.4 \\ & & 1.0 & 0.0 & 0.0 & 0.0 \\ & & & 1.0 & 0.0 & 0.0 \\ & & & & 1.0 & 1.0 \\ & & & & & 1.0 \end{bmatrix}$$

Dendrograms corresponding to single link and complete link clustering algorithms are given below. Left one is dendrogram for single link and the right one is the dendrogram for complete link clustering algorithm.



4. Consider the above D matrix. Cluster the documents using the cover coefficient-based clustering methodology (C^3M). Please

   (a) Show the double-stage probability experiment tree for the second document, and show the calculation of c_{24} of the corresponding C matrix.

   (b) Obtain the C matrix (you do not need to show the intermediate steps).

   (c) Find the number of clusters implied by the C matrix -explain how-.

   (d) Find the cluster seeds.

   (e) Obtain the IISD (inverted index for seed documents)

   (f) Obtain the clusters and explain how.

Solution:

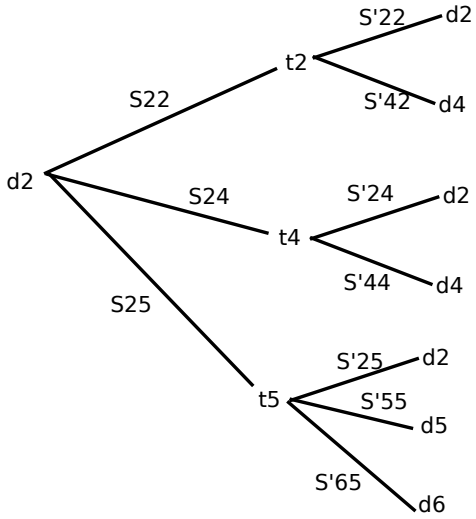
(a)

$$C_{24} = \sum_{k=1}^{6} S_{2k} \times S'_{4k}$$

where S_{ik} is probability of selecting t_k from d_i ($S_{ik} = d_{ik} \times [\sum_{h=1}^{n} d_{ih}]^{-1}$), and S'_{jk} is the probability of selecting d_j from t_k ($S'_{jk} = d_{jk} \times [\sum_{h=1}^{m} d_{hk}]^{-1}$).

$$C_{24} = 0 \times 0 + 1/3 \times 1/2 + 0 \times 0 + 1/3 \times 1/2 + 1/3 \times 0 = 0.33$$

Double stage probability experiment tree for the second document is given below.



(b)

$$C = \begin{bmatrix} 0.50 & 0.00 & 0.50 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.44 & 0.00 & 0.33 & 0.11 & 0.11 \\ 0.50 & 0.00 & 0.50 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.50 & 0.00 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.17 & 0.00 & 0.00 & 0.42 & 0.42 \\ 0.00 & 0.17 & 0.00 & 0.00 & 0.42 & 0.42 \end{bmatrix}$$

(c) $n_c = \sum_i C_{ii} = 0.50 + 0.44 + 0.50 + 0.50 + 0.42 + 0.42 = 2.78 \approxeq 3$

(d)

$$\begin{aligned} P_1 &= 0.50 \times 0.50 \times 2 = 0.5 \\ P_2 &= 0.44 \times (1 - 0.44) \times 3 = 0.74 \\ P_3 &= 0.50 \times 0.50 \times 2 = 0.5 \\ P_4 &= 0.50 \times 0.50 \times 2 = 0.5 \\ P_5 &= 0.42 \times (1 - 0.42) \times 2 = 0.49 \\ P_6 &= 0.42 \times (1 - 0.42) \times 2 = 0.49 \end{aligned}$$

Select D_1, D_2 and D_4 as cluster seeds.

(e) $T1 \rightarrow < D1, 1 >$

$T2 \rightarrow < D2, 1 >, < D4, 1 >$

$T3 \rightarrow < D1, 1 >$

$T4 \rightarrow < D2, 1 >, < D4, 1 >$

$T5 \rightarrow < D2, 1 >$

$T6 \rightarrow <>$

(f) $C_{31} > C_{32} = C_{34}$, D3 goes to cluster with D1.

$C_{52} > C_{51} = C_{54}$, D5 goes to cluster with D2.

$C_{62} > C_{61} = C_{64}$, D6 goes to cluster with D2.

So, final clusters are {D1, D3}, {D2, D5, D6} and {D4}.

5. Consider the incremental version of C^3M: C^2ICM, Cover Coefficient-based Incremental Clustering Methodology, described in Can F, Incremental clustering for dynamic information processing, (ACM TOIS, 1993).

   (a) Briefly explain the algorithm (one paragraph).

   (b) In the paper there is the concept of clustering similarity, explain its purpose within the context of C^2ICM.

   (c) The paper mentions a measure called Rand coefficient (and cites the classic book of Jain & Dubes: Algorithms for Clustering Data). Obtain the (regular) Rand similarity of the clustering structures

   CS1 = {{a, b, c}, {d, e}, {f, g}}
   CS2 = {{a}, {b, c, d}, {e, f, g}}

   -where the last cluster of CS2 contains the members e, f and g-. Optional: you may also obtain the corrected Rand coefficient using these two clustering structure. Show the contingency table that needs to be corrected for the Rand coefficients.

   (d) Explain the difference between Rand and corrected Rand coefficients.

Solution:

(a) In the original version, C^3M, cover coefficient matrix, C, is introduced. Individual entries of C, c_{ij} indicates the probability of selecting any term of d_i from d_j. In C^3M, for clustering this C matrix is computed, number of clusters (n_c) is calculated, some documents are selected as cluster seeds using C matrix. Other documents are assigned to any cluster according to similarities to seed documents. In the incremental version, C^2IM, at each update on document set, there is no need to run the entire C^3M algorithm. Instead, only diagonal elements are recalculated. Seed documents are reselected. New documents and documents in falsified old clusters are reassigned. Falsified cluster is cluster where its seeds are not seed any more or cluster containing documents that are not seeds initially, but assigned as seeds in the last seed selection phase. Algorithm does not need to calculate entire C matrix at each step.

(b) The clusterings generated incrementally by using C^2IM and reusing C^3M at each step are compared. The similarity between two methods is using to check how well the C^2IM generates clustering as C^3M while performing with lower cost.

(c) CS1 = {{a, b, c}, {d, e}, {f, g}}

CS2 = {{a}, {b, c, d}, {e, f, g}}

- a: number of pairs of objects that are in the same group in both partitions. $a = 2$ {(b,c), (f,g)}
- b: number of pairs of objects that are in the same group in CS1 but in different groups in CS2. $b = 3$ {(a,b), (a,c), (d,e)}
- c: number of pairs of objects that are in different groups in CS1 but in the same group in CS2. $c = 4$ {(b,d), (c,d), (e,f), (e,g)}
- d: number of pairs of objects that are in different groups in both partitions. $d = \binom{n}{2} - a - b - c = 12$

Contingency table, rand similarity and corrected rand similarity are given below, In contingency table, X1, X2, X3 are clusters of CS1 and Y1, Y2, Y3 are clusters of CS2.

	D1	D2	D3	
C1	1	2	0	3
C2	0	1	1	2
C3	0	0	2	2
	1	3	3	7

$$\text{Rand similarity} = \frac{a+d}{\binom{n}{2}} = \frac{14}{21} = 0.67$$

$$\begin{aligned}\text{Corrected Rand sim.} &= \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[1/\binom{n}{2}\right]\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}}{(1/2)\left[\sum_j \binom{n_{.j}}{2} + \sum_i \binom{n_{i.}}{2}\right] - \left[1/\binom{n}{2}\right]\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}} \\ &= 0.14\end{aligned}$$

(d) In the corrected rand index, regular rand index is normalized so that its value is 0 if partitions are selected by chance and 1 when a perfect match is achieved.

6. In this part consider the paper J. Zobel, A. Moffat, "Inverted files for text search engines.", *ACM Computing Surveys*, Vol. 38, No. 2, 2006.

(a) Understand the skipping concept as applied to the inverted index construction.

Assume that we have the following posting list for term-a: $< 1, 2 >, < 3, 1 >, < 9, 2 >, < 10, 3 >, < 12, 4 >, < 17, 4 >, < 18, 3 >, < 22, 2 >, < 24, 2 >, < 33, 4 >, < 38, 5 >, < 43, 5 >, < 55, 3 >, < 64, 2 >, < 68, 4 >, < 72, 5 >, < 75, 5 >, < 88, 2 >$. The posting list indicates that term-a appears in d1 twice and in d3 once, etc.

Assume that we have the following posting list for term-b: $< 12, 2 >, < 66, 1 >$.

Consider the following conjunctive Boolean query: term-a **and** term-b. If no skipping is used how many comparisons do you have to find the intersection of these two lists?

Introduce a skip structure, draw the corresponding figure then give the number of comparisons involved to process the same query.

State the advantages and disadvantages of large and small skips in the posting lists. Note that in the paper it is assumed that compression will be used. The skip idea is applicable in an uncompressed environment too.

(b) Give a posting list of term-a (above it is given in the standard sorted by document number order) in the following forms: 1) ordered by $f_{d,t}$, 2) ordered by frequency information in prefix form. What are the advantages of the approaches 1 and 2? Do they have any practical value?

Solution:

(a) (i) Without skipping: term-a has 18 entries and term-b has 2 entries. $18 \times 2 = 36$ comparisons are needed.

(ii) With skipping: Let the chunk size be at most 5. Sort each chunk in reverse order according to the document IDs. The resulting chunks

term-a:

Chunk1: $< 12, 4 >, < 10, 3 >, < 9, 2 >, < 3, 1 >, < 1, 2 >$

Chunk2: $< 33, 4 >, < 24, 2 >, < 22, 2 >, < 18, 3 >, < 17, 4 >$

Chunk3: $< 68, 4 >, < 64, 2 >, < 55, 3 >, < 43, 5 >, < 38, 5 >$

Chunk4: $< 88, 2 >, < 75, 5 >, < 72, 5 >$

term-b:

Chunk1: $< 66, 1 >, < 12, 2 >$

Compare elements in term-a list and term-b list.

* Retrieve first element of term-b list: $< 66, 1 >$.
  1. Chunk1: $66 > 12$: skip chunk.
  2. Chunk2: $66 > 33$: skip chunk.
  3. Chunk3: $66 < 68$: OK. Search the rest of Chunk3:
     (a) $66 \stackrel{?}{=} 64$
     (b) $66 \stackrel{?}{=} 55$
     (c) $66 \stackrel{?}{=} 43$
     (d) $66 \stackrel{?}{=} 38$
* Retrieve next element of term-b list: $< 12, 2 >$.
  1. Chunk1: $12 = 12$: found.

Total 8 comparisons. Better than without skipping.

With large size chunks, greater number of documents would be skipped with one chunk skip, but more comparisons would occurred in one chunk. With small size chunks, smaller number of documents would be skipped with one chunk skip, but less comparison would occurred in one chunk.

(b) Posting list of term-a

- ordered by $f_{d,t}$: $< 38, 5 >, < 43, 5 >, < 72, 5 >, < 75, 5 >, < 12, 4 >, < 17, 4 >, < 33, 4 >, < 68, 4 >, < 10, 3 >, < 18, 3 >, < 55, 3 >, < 1, 2 >, < 9, 2 >, < 22, 2 >, < 24, 2 >, < 64, 2 >, < 88, 2 >, < 3, 1 >$.
- ordered by frequency information in prefix form: $< 5 : 4 : 38, 43, 72, 75 >, < 4 : 4 : 12, 17, 33, 68 >, < 3 : 3 : 10, 18, 55 >, < 2 : 6 : 1, 9, 22, 24, 64, 88 >, < 1 : 1 : 3 >$.

Documents with higher term-frequency values are considered important. By storing in frequency order, documents with frequencies higher than a threshold can be processed without fetching all list. By fetching posting list block-by-block, performance gain is achieved.

7. In this part consider the paper, A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review" *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.

   (a) Please explain the stages of clustering as defined in this paper.

   (b) Consider fuzzy clustering and introduce and idea that we can use fuzzy clustering approach in connection with C^3M.

   (c) In connection with simulated annealing the authors mention "tabu search". What does it mean? Explain its use within the context of simulated annealing-based clustering.

   (d) What are the components of a typical clustering task? Explain each step within the framework of an information retrieval environment.

   (e) In connection with the above question (section d) please also explain what is meant by clustering tendency? Does it make sense to use clustering tendency in some stage(s) of clustering? What would you propose to use for identifying clustering tendency?

Solution:

(a) The stages of clustering are

- Feature selection/extraction: The most effective subset of original features is identified and features are transformed to new representations.
- Inter-pattern similarity: Some distance function defined on pairs of patterns is used to measure the similarity/dissimilarity between patterns.
- Grouping: Using inter-pattern similarity, similar patterns are assigned to same groups while not similar one are assigned to different groups.

(b) Let M be the number of objects to be clustered and n_c be the number of clusters. Determine n_c using C^3M algorithm (Compute C matrix and $n_c \approxeq \sum_{i=1}^{M} c_{ii}$). Define an $M \times n_c$ membership matrix, U. Entry u_{ij} in U represents the grade of membership of document d_i in cluster c_j. $u_{ij} \in [0, 1]$. $\sum_{j=1}^{n_c} u_{ij} = 1$.

Select seed documents using the same method in C^3M algorithm. Let D_s be the set of seed documents selected, then membership of d_i to cluster c_j is

$$u_{ij} = \begin{cases} 0 & \text{if } d_j \notin D_s \\ C_{ij} / \sum_{s \in D_s} c_{is} & \text{if } d_j \in D_s \end{cases}$$

(c) Like simulated annealing, tabu search is a mathematical optimization method to find an approximation to the global optimum in a search space. In tabu search, while the method is trying to move the system to a neighbouring state, to lower energy, it keeps a "taboo list" to avoid previously seen states.

(d) Components of a clustering task are given below.

- Pattern representation (optionally including feature extraction and/or selection): Features which are appropriate for the purpose of clustering are selected, represented properly for the clustering algorithm. Some transformations are used to produce new features.

- Definition of a pattern proximity measure appropriate to the data domain: A distance function to measure similarity/dissimilarity between patterns. In information retrieval, document to document similarity is used. Cosine similarity, Dice's coefficient, Jaccard index are some functions used in information retrieval to compare documents, represented using, for instance, tf-idf vectors.
- Clustering or grouping: Based on similarity function selection and decision of one of the several different clustering algorithms, patterns are grouped. At the end of this step, similar patterns belong to the same groups while dissimilar ones are in different groups. In IR, documents are clustered based on their similarity to each other.
- Data abstraction: Used to represent dataset in a simple and compact way. A typical data abstraction in the field of IR is choosing some documents as centroids (cluster prototypes, representative samples).
- Cluster validity: This step is about evaluating the result obtained based on choice of similarity function and clustering algorithm.

(e) Clustering tendency is degree of suitability of data to be clustered (i.e. the data contains natural groupings or not). Clustering tendency analysis is performed not before actual clustering to analyse and decide weather it is useful to cluster data or not.

Sometimes, clustering tendency analysis is useful before actual clustering operations. For instance, all clustering algorithms do not determine the number of clusters internally. For example, C^3M determines the number of clusters as a part of the process, but, in k-means clustering algorithm, k, the number of clusters is given as input to the algorithm. For such a case, performing clustering tendency analysis to identify number of natural clusters would be useful. Some visualization tools can be used to identify number of clusters manually.

8. Is the complete-link clustering method order-independent? Explain/prove your claim.

9. What are the components of an information retrieval test collection? Explain the pooling approach. Please read the paper by Zobel (How Reliable Are the Results of Large-Scale Information Retrieval Experiments?) and give some reflections of his criticism of this approach.

Solution: A test collection is comprised of a set of documents, a set of queries, and relevance information about each document with respect to each query.

The key obstacle of using large sized test collections is topic assessment. Since assessing the whole document set is unpractical, a certain number of documents are selected for judgement. This procedure is called pooling. In his work, Zobel investigates pooling approach using the data generated in the TREC experiments. He shows that many (50% - 70%) relevant documents have not been found in experiments. The reason for low recall value, he points the queries that have large numbers of answers and the strategy of assuming unjudged documents as irrelevant.

10. Please re-examine the 11 Watt/Google Query Legend. Is it real or not? Please write your findings based on research -please specify your resources-. Assuming that the claim is true please calculate how much KW you spend in a typical year and also calculate its TL equivalent. (You may also calculate the same cost for a person who lives in New York, NY or any other famous foreign city for comparison.) Explain your reasoning.

Solution: In Google's official blog[1], it is stated that the amount of energy is 0.0003 kWh and in terms of greenhouse gases, one Google search is equivalent to 0.2 grams of CO2, not 7 grams as claimed.

The author of the article that raised the debate "two Google queries generate about the same amount of CO2 as boiling a kettle for a cup of tea", Alex Wissner-Gross says that he never mentions Google in his study, in a TechNewsWorld article[2]. He states that they found that it takes about 20mg of CO2 per second to visit a web site.

Google says that they are committed to helping to build a clean energy future and minimize their impact on the Earth's climate. On the other hand, Wissner-Gross acknowledges that Google is interested in energy-efficient infrastructure because energy consumption may be a high fraction of their infrastructure costs.

According to U.S. Energy Information Administration[3], average monthly residential electricity consumption in NY is 604 kWh and average monthly bill is $103.25.

(1) http://googleblog.blogspot.com/2009/01/powering-google-search.html

(2) http://www.technewsworld.com/story/Harvard-Prof-Sets-Record-Straight-on-Internet-Carbon-Study-65794.html

(3) http://www.eia.doe.gov/cneaf/electricity/esr/table5.xls