2004

Comparison of different distance measures on hierarchical document clustering in 2-pass retrieval

Azam Jalali
University of Tehran

Farhad Oroumchian
University of Wollongong in Dubai, farhado@uow.edu.au

Mahmoud Reza Hejazi
Iran Telecommunications Research Center

Comparison of Different Distance Measures on Hierarchical Document Clustering in 2-Pass Retrieval

Azam Jalali
Department of Computer and Electrical Engineering,
Faculty of Engineering,
University of Tehran

Info. Society group
Iran Telecom. Research Center
Jalali@itrc.ac.ir

Farhad Oroumchian

IT Department,
University of Wollongong in Dubai

FarhadOroumchian@uowdubai.ac.ae

Mahmoud Reza Hejazi

Info. Society group
Iran Telecom. Research Center

M_hejazi@itrc.ac.ir

Abstract: - Hierarchic document clustering has been applied to search results (query-specific clustering) on the grounds of its potential improved effectiveness compared both to that of static clustering and of conventional inverted file search (IFS).
In this paper we review and compare the effects of seven different measures of similarity among documents in hierarchic query specific clustering. We have conducted a number of experiments using OHSUMED document collection. The Experiments seems to indicate that the choice of similarity measure effects positively or negatively the quality of clustering.

Key-Words: - cluster-based search; Hierarchical clustering; Distance measures

1 Introduction

The cluster hypothesis states that relevant documents tend to be more similar to each other than to non-relevant documents, and therefore tend to appear in the same clusters [7] Document clustering has been extensively investigated as a methodology for improving document search and retrieval.

In most of the previous research the strategy was to build a static clustering of the entire collection and then match the query to the cluster centeroids [1].On the other hand the behavior and effectiveness of clustering methods when applied to the search results of an IR system (i.e. dynamic, or *query-specific* clustering) have not been extensively investigated[2] .Two broad types of query-specific clustering that have been mainly used in IR: are cluster-based search and cluster based browsing

In cluster-based search, a single cluster is retrieved in response to a query. The documents within the retrieved cluster are not ranked in relation to the query but rather the whole cluster is retrieved as an entity. Cluster representation refers to the formation of cluster representatives, or centroids that attempt to summarize the contents of a cluster for the purpose of retrieving the cluster. Incoming queries are matched against representatives, and the cluster whose representatives are most similar to the query, is retrieved [2] . Three different types of cluster-based searches have been studied in IR: Top-down search, Bottom-up search and optimal cluster search

Cluster-based browsing paradigm clusters documents into topically coherent groups, and presents descriptive textual summaries to the user. The summaries consist of topical terms that characterize each cluster generally, and a number of typical titles that sample the contents of the cluster. Informed by the summaries, the user may select clusters, forming a sub-collection, for iterative examination. The clustering and re-clustering is done on the fly, so that different topics are seen depending on the sub-collection clustered [4] .

In order to cluster documents one must first establish a pairwise measure of document similarity. Numerous document similarity measures have been proposed, all of which treat each document as a set of words, often with frequency information, and measure the degree of word overlap between documents. This paper makes a case for the use of hierarchic query-specific clustering in IR using seven different distance between documents.

We compare effectiveness of these distances on hierarchic query specific clustering.

The rest of the paper is structured as follows: Section 2 details the document clustering algorithms; Section 3 provides the details of the experiment. Section 4 discusses the experimental results. At the end, section 5 outlines the conclusions and the future direction of the work.

2 The Document Clustering Algorithms

Two different types of document clusters can be constructed. One is a flat partition of the documents into a collection of subsets. The other is a hierarchical cluster, which can be defined recursively as either an individual document or a partition of the corpus into sets, each of which is then hierarchically clustered. A hierarchical clustering defines a tree, called a dendrogram, on the documents.

The actual effectiveness of hierarchic clustering can be gauged by cluster based searches that retrieve the cluster that best matches the query [2] Four hierarchical methods were employed in the experiments: single link, complete link, group average, and centeroid method, which differ in how the distance between sub nodes is defined in terms of their members.

In pair wise single-linkage clustering, the distance between two nodes is defined as the shortest distance among the pair wise distances between the members of the two nodes.

- In pair wise maximum-linkage clustering, alternatively known as pair wise complete linkage clustering, the distance between two nodes is defined as the longest distance among the pair wise distances between the members of the two nodes.
- In pairwise average-linkage clustering, the distance between two nodes is defined as the average over all pairwise distances between the elements of the two nodes.
- In pairwise centroid-linkage clustering, the distance between two nodes is defined as the distance between their centroids. The centroids are calculated by taking the mean over all the elements in a cluster. As the distance from each newly formed node to existing nodes and items need to be calculated at each step, the computing time of pairwise centroid-linkage clustering may be significantly longer than for the other hierarchical clustering methods.

The main reason behind the choice of these four methods is the fact that they have been extensively used and examined in the context of IR [10] . In order to cluster gene expression data into groups with similar genes or micro arrays, we should first define what exactly we mean by similar. We used C clustering library [11] , seven distance functions are available to measure similarity, or conversely, distance:

1. ***'c' Pearson correlation***

The Pearson correlation coefficient is defined as

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

In which $\bar{x}$ and $\bar{y}$ are the sample means respectively, and σ_x, σ_y are the sample standard deviation of x and y. The Pearson distance is then defined as $d_P = 1 - r$.

2. ***'a' Absolute value of Pearson correlation***

The distance is defined as usual as $d_a = 1 - |r|$. Where, r is the Pearson correlation coefficient.

3. ***'u' Uncentered Pearson correlation*** (equivalent to the cosine of the angle between two data vectors.)

The uncentered correlation is defined as

$$r_u = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\sigma_x^{(0)}}\right)\left(\frac{y_i}{\sigma_y^{(0)}}\right)$$

where

$$\sigma_x^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

$$\sigma_y^{(0)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}$$

This is the same expression as for the regular Pearson correlation coefficient, except that the samples means $\bar{x}$ and $\bar{y}$ are set equal to zero. The distance corresponding to the uncentered correlation coefficient is defined as $d_u = 1 - r_u$. Where r_u is the uncentered correlation.

4. ***'x' Absolute uncentered Pearson correlation*** (equivalent to the cosine of the smallest angle between two data vectors)

The distance measure is defined using the absolute value of the uncentered correlation, $d_u = 1 - |r_u|$; where r_u is the uncentered correlation coefficient.

5. ***'s' Spearman's rank correlation***

The Spearman rank correlation is an example of a non-parametric similarity measure. To calculate the

Spearman rank correlation, each data value is replaced by their rank if the data in each vector is ordered by their value. Then the Pearson correlation between the two rank vectors instead of the data vectors is calculated.
Spearman rank distance measure corresponding to the Spearman rank correlation as
$d_s = 1 - |r_s|$;
where r_s is the Spearman rank correlation.

6. ***'e' Euclidean distance***

The Euclidean distance is the only true metric among the distance functions that are available in the C clustering library, being the only distance function satisfying the triangle inequality. The Euclidean distance is defined as.

$$d = \sum_{i=1}^{n} (x_i - y_i)^2$$

7. ***'h' Harmonically summed Euclidian distance***

The harmonically summed Euclidean distance is a variation of the Euclidean distance, where the terms for the different dimensions are summed inversely (similar to the harmonic mean):

$$d = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{x_i - y_i} \right)^2 \right]^{-1}$$

The characters in front of the distance measures are used in figures.

3 Experimental details

The main objective of these experiments was to compare and to examine effectiveness of different distance between document vectors in clustering.

The OHSUMED test collection [9] was selected as test bed. Four hierarchic agglomerative methods were implemented as the benchmark. Five different values for top n ranked documents also were experimented with in clustering.

3.1 Document Collections and Initial Retrieval

For the experiments, the OHSUMED medical abstracts collection was selected as test collection. It is a clinically oriented MEDLINE subset, consisting of 348,566 references covering all references from 270 medical journals over a five-year period (1987-1991). In this study we used a subset of the OHSUMED[9] We selected all documents in 1987, which includes 54,710 documents and 63 queries.

The SMART document retrieval system [12] was used in order to perform the initial retrieval with atc.atc weighting scheme. The atc measure normalizes the weights for document length, giving all documents an equal chance for retrieval.

Formula 1 shows atc weighting scheme of SMART retrieval system. To assign an indexing weight w_{ij} that reflects the importance of each single-term Tj in a document D, different factors should be considered, as follows:

- within-document term frequency tf_{ij}, which represents the first letter of the SMART label.
- collection-wide term frequency df_j , which represents the second letter of the SMART label.

In Formula 1, $idf_j = \log \frac{N}{F_j}$;

where, N represents the number of documents and F_j represents the document frequency of term T_j.

- normalization scheme, which represents the third letter of the SMART label.

$$w_{ij} = \frac{idf_j \times (0.5 + \frac{tf_{ij}}{2 \times \max\_ tf_i})}{\sqrt{\sum_{k=1}^{n} [idf_k \times (0.5 + \frac{tf_{ik}}{2 \times \max\_ tf_i})]^2}} \quad (1)$$

The default SMART stoplist and stemming were also used in indexing all the collections and queries. After the initial retrieval, the top-n ranked documents were used to create the collections that were clustered. Five different values of n were tested: n=100, n=200, n=300, n=400, n=500.

3.2 Optimal cluster evaluation

In these experiments, the E effectiveness function that was proposed by Jardine and Van Rijsbergen is used as optimally criterion [1] [15] [15] [10] . The formula for the measure is given by: $1 - \{(\beta^2 + 1)PR / (\beta^2 P + R)\}$ where P and R correspond to the standard definitions for precision and recall (over the set of documents of a specific cluster), and β is a parameter that reflects the relative importance attached to precision and recall. Three values of this parameter are usually used: 1, 0.5 and 2, the first value attributing equal importance to precision and recall, the second deeming precision twice as important as recall, and the third treating recall twice as important as precision. The E effectiveness measure and these three values of the

parameter β are used in the experiments reported in this paper.

The optimal cluster for any given query is the cluster that yields the least E value for that query. Jardine and Van Rijsbergen named this measure MK1. It is used to measure optimal cluster effectiveness in our experiments.

The main advantage of optimal measures eliminates any bias that may be introduced from external sources. External sources include the choice of a particular cluster-based search strategy that matches queries to clusters, and the ability of a user during a browsing session to choose the cluster, which is most relevant to his/her information need [2].

We have excluded all the queries that had fewer than 10 relevant documents in their top 100 ranked documents. Fourteen queries were left after the above process. For each of the remaining queries we would run the query in SMART system and cluster the set with five different values for top-ranked documents (100, 200, 300, 400, 500).

4 Experimental Results

In this section we analyze the experimental results and discuss their implications.

4.1 Optimal Cluster Evaluation Results

Optimal cluster evaluation results on the MK1 measure for SMART using seven different distance document-clustering methods are reported in the figure 1 through 12. The figures present a comparison of seven clustering methods (Hierarchical with different similarity) based on structures bag of words for 5 different number of documents (from 100 to 500) and three different β values (0.5, 1, 2). These figures compare MK1 values.

The performance of the other 6 measures was similar and it seems that all of them could be used for hierarchical clustering in the second stage of retrieval.

The results that presented in these figures may suggest that there is no significant degradation of effectiveness for decreasing of n with $\beta = 1, \beta = 2$. The effectiveness of applying different values of n for $\beta = 0.5$ (figures 1, 4, 7, 10) appears to increase as n increases, MK1 increases too.

Our results over all experimental conditions indicated that the group average method was the most effective in terms of optimal cluster evaluation. Maximum link and Centroid methods were close to each other with often-negligible differences in effectiveness, while single link displayed the poorest effectiveness of the four.

Van Rijsbergen [17] , and also Sneath and Sokat [14] , emphasized that the various association and distance measures are monotone with respect to each other. Consequently, a clustering method that depends only on the rank ordering of the resemblance values would give similar results for all such measures. In contrast to this claim, we perceived that the results of hierarchical document clustering using seven distance functions were different. Harmonic method produced results worse than other methods.

The results suggest that, with the exception of the smallest value of n for the hierarchical methods, there is no significant increase in effectiveness for larger numbers of top-ranked documents. Consequently, if one were to choose a unique value for n, one would also have to consider practical issues. It may be advantageous from an efficiency point of view to cluster the top-200 or top-300 documents returned from a search rather than, for example, the top-500 documents. Moreover, it can be argued that if the resulting cluster structure was to be presented to a user in an interactive task environment, then a reduced document space may be advantageous (e.g. allowing the user to easily and quickly find a few relevant documents which could start a relevance feedback iteration or satisfy the user's information need).

Table 1: Mean Relevant documents per query for different best-ranked documents

Top -n	n Relevant documents per q
100	13. 58
200	16.35
300	16.5
400	16.64
500	16.64

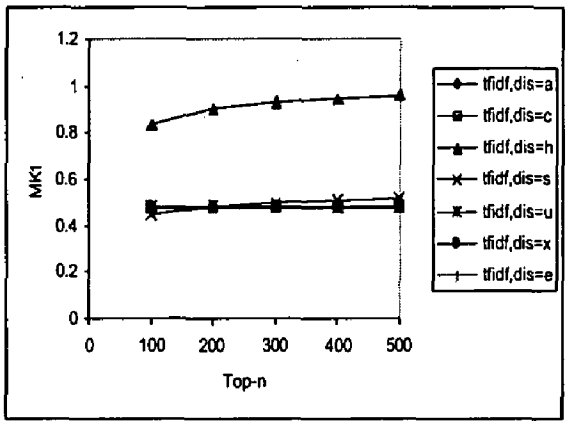

Fig. 1: Mk1 for **average-linkage clustering** with $\beta = 0.5$

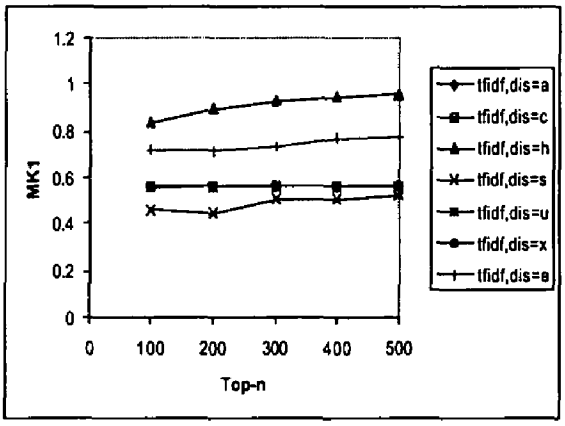

Fig. 4: Mk1 for **centroid-linkage clustering** with $\beta = 0.5$

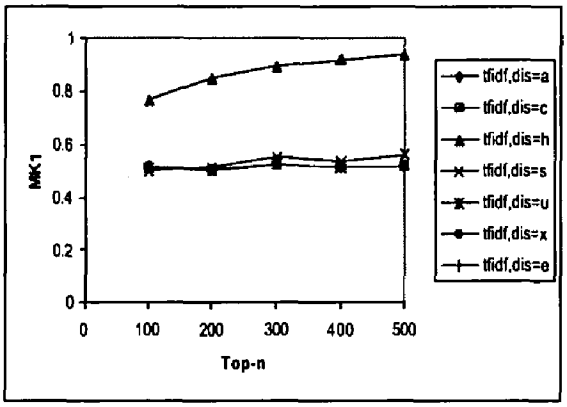

Fig. 2: Mk1 for **average-linkage clustering** with $\beta = 1$

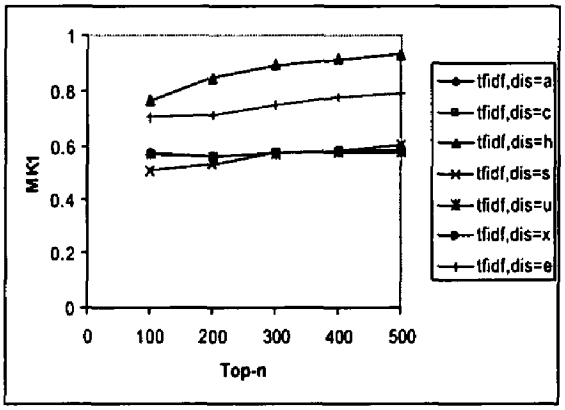

Fig. 5: Mk1 for centroid-linkage clustering with $\beta = 1$

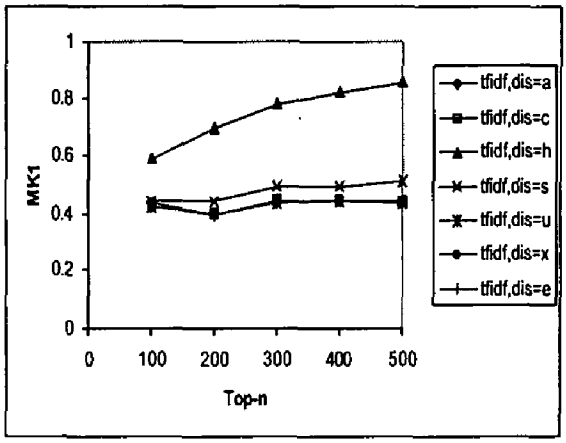

Fig. 3: Mk1 for **average-linkage clustering** with $\beta = 2$

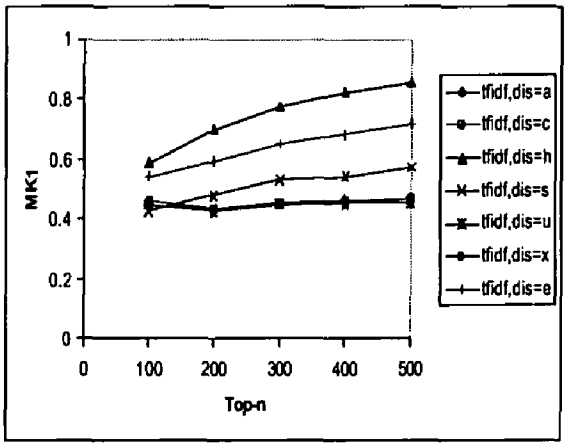

Fig. 6: Mk1 for **centroid-linkage clustering** with $\beta = 2$

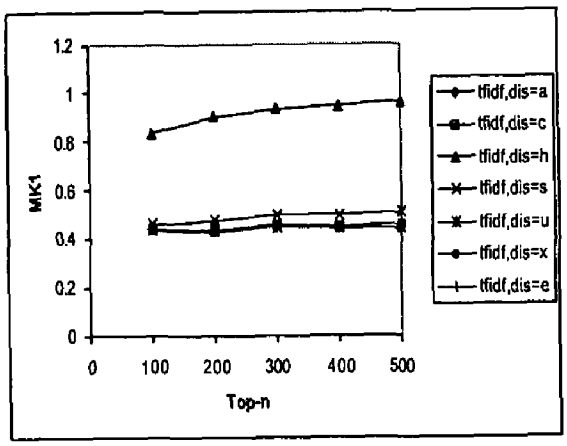

Fig. 7:Mk1for maximum-linkage clustering with $\beta = 0.5$

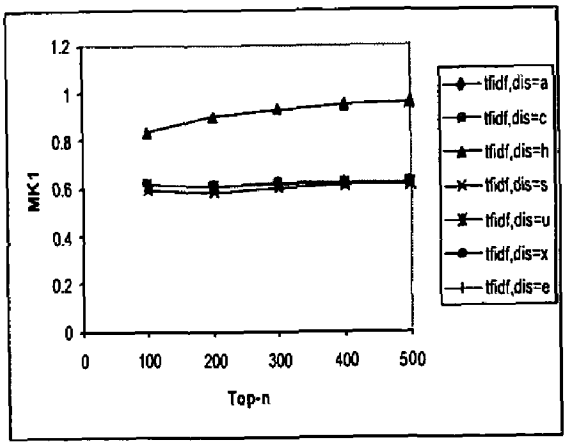

Fig.10: Mk1for **single-linkage clustering** with $\beta = 0.5$

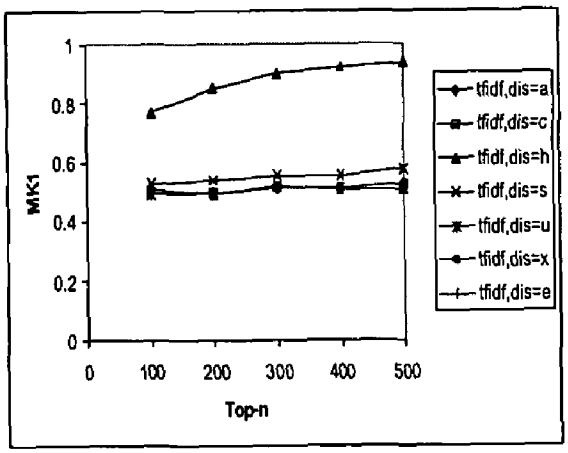

Fig. 8: Mk1 for **maximum-linkage clustering** with $\beta = 1$

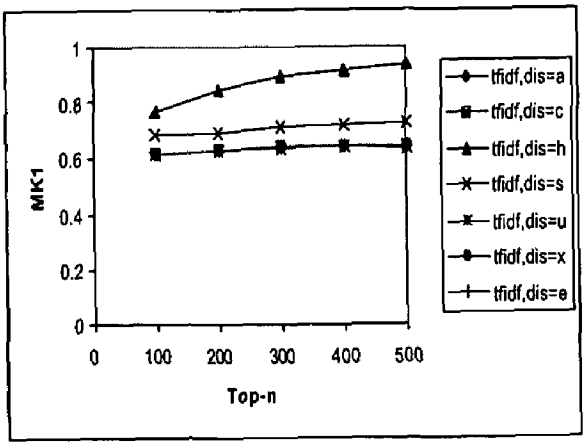

Fig.11: Mk1 for **single-linkage clustering** with $\beta = 1$

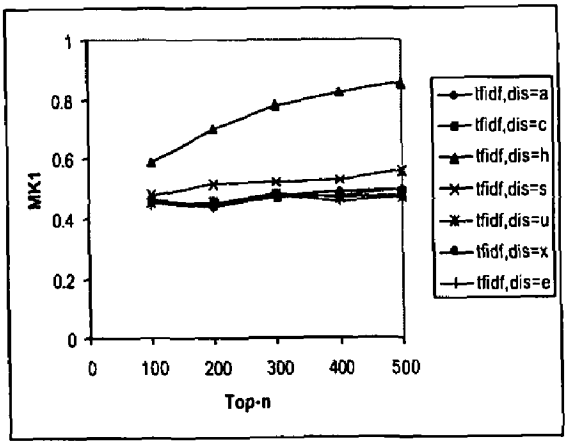

Fig.9: Mk1for **maximum-linkage clustering** with $\beta = 2$

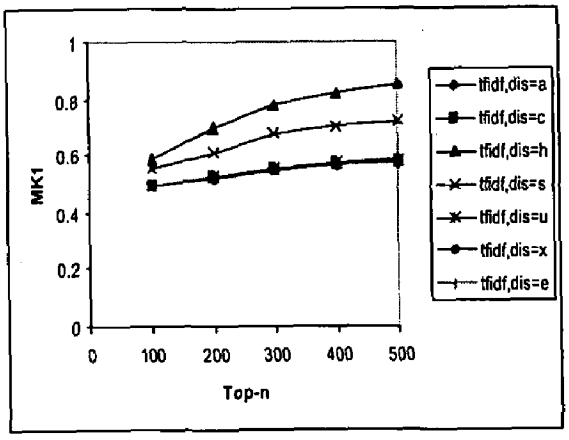

Fig.12: Mk1for **single-linkage clustering** with $\beta = 2$

5 Future Work and Conclusions

We have perceived that the results of hierarchical document clustering depend on features values and measure similarity between documents. The results produced by Harmonic Distance were worse than the other measures.

In future, more experiments will be conducted with other collections and different methods. Another direction would be to use this online clustering method and use it as a method of presentation in an information retrieval system. In such a scenario, after clustering the best representatives of the clusters could be presented to the user. In this case user can use these cluster representatives for choosing the browsing direction.

References:

[1] Anastasios Tombros, "The effectiveness of query-based Hierarchic clustering of documents for information retrieval", Ph.D. Thesis, Department of Computing Science Faculty of Computing Science, Mathematics and Statistics, University of Glasgov, 2002.

[2] Anastasios Tombros, Robert Villa, C.J. Van Rijsbergen, "The effectiveness of query-specific hierarchic clustering in information retrieval", Information Processing and Management 38 ,pp. 559–582,2002.

[3] Croft, W. B."A model of cluster searching based on classification", Information Systems, 5, 189–195, 1980.

[4] Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. "Scatter/Gather: A cluster-based approach to browsing large document collections", In Proceedings of the 15th annual ACM SIGIR conference, Copenhagen, Denmark pp. 126–135, 1992.

[5] Ellis, D., Furner-Hines, J., & Willett, P. "Measuring the degree of similarity between objects in text retrieval systems", Perspectives in Information Management, 3(2), 128–149, 1993.

[6] El-Hamdouchi, A., & Willett, P. "Techniques for the measurement of clustering tendency in document retrieval systems", Journal of Information Science, 13, 361–365, 1987.

[7] Gordon, A.D. "A review of hierarchical classification", Journal of the Royal Statistical Society, Series A, 150(2):119-137, 1987.

[8] Hearst, M. A., & Pedersen, J. O. "Re-examining the Cluster Hypothesis: Scatter/Gather on retrieval results", In Proceedings of the 19th Annual ACM SIGIR conference, Zurich, Switzerland pp. 76–84, 1996.

[9] Hersh, W. R., Buckley, C., Leone, T. J. and Hickam, D. H. "OHSUMED: An interactive retrieval evaluation and new large test collection for research", In Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94), 192-201, 1994.

[10] Jardine, N., & Van Rijsbergen, C. J. "The use of hierarchical clustering in information retrieval", Information Storage and Retrieval, 7, 217–240, 1971.

[11] Michiel de Hoon, Seiya Imoto, Satoru Miyano, " The C Clustering Library" , The university of Tokyo, Institute of Medical Science, Human Genome Center, 2003.

[12] Salton, G. "The SMART retrieval system – experiments in automatic document retrieval", Englewood Cliffs, NJ: Prentice-Hall, 1971.

[13] Sneath, P.H.A. and Sokal, R.R.. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: W.H. Freeman, 1973.

[14] Stefan Ruger, Susan Gauch, "Feature Reduction for Document Clustering and Classification", Technical Report DTR 2000/8; Department of Computing, Imperial College; London, England, 2000.

[15] Van Rijsbergen, C. J. "Further experiments with hierarchic clustering in document retrieval", Information Storage and Retrieval, 10, 1–14, 1974.

[16] Van Rijsbergen, C. J., & Croft, W. B. "Document clustering: An evaluation of some experiments with the Cranfield 1400 Collection", Information Processing & Management, 11, 171–182, 1975.

[17] Van Rijsbergen, C.J. Information Retrieval. London: Butterworths, 2nd Edition, 1979.