IX. On Some Clustering Techniques for Information Retrieval

J. D. Broffitt, H. L. Morgan, and J. V. Soden

Abstract

Document clustering methods which have been proposed by R. E. Bonner and J. J. Rocchio are compared. Bonner's method is found to give higher precision than Rocchio's method, while the recall for the two methods is about the same. Bonner's method necessitates about twice as many comparisons against a query vector as Rocchio's method; this is to be expected since Rocchio controls the cluster size in order to maximize search efficiency. Manual relevance judgments are used as well as relevance judgments determined by query document cosines. The results are found to be invariant under the two measures.

1. Introduction

The organization of information into homogeneous groups plays a major role in many fields of research. Some areas of application are information retrieval, biological taxonomy, isolation of disease syndromes in medicine, anthropology (categorization of tribes), and business applications such as categorizing TV audiences, sales offices, etc. Indeed, the applications of information organization are numerous, and the particular application being studied dictates the type of classification needed.

Needham[1] has divided classification problems into three types: (1) the assignment of given objects to given classes, (2) the extraction of class characteristics from given classes and their objects, and

(3) the setting up of appropriate classes, clusters, clumps, or groups given a set of objects and some information about them. Even more specifically, this last problem may be viewed as a problem of either structuring the objects into a hierarchy or tree arrangement, or collecting the objects into coherent groups without regard for hierarchical relations among the objects or groups. Both of these problems find their place within the realm of an automatic information retrieval system, where documents are identified by vectors of weighted measures of occurrences of concepts.

Thesaurus construction, for example, while being concerned with the grouping of cooccurring concepts for synonym references, has as a major objective the determination of hierarchies of information among the concepts. On the other hand, document clustering deals with the collection of similar documents into groups based only on similarities among the documents. The application of information organization to document clustering is the concern of this study.

In order effectively to operate an automatic information retrieval system based on vector matching between queries and documents, an efficient document clustering procedure is necessary. The number of documents which must be compared with the query in order reasonably to satisfy the demands of an information request is too large to permit individual comparisons with each document in the collection. Once the documents have been clustered into homogeneous groups, a two-level search procedure greatly reduces the number of comparisons needed to answer a request. This method first compares the query vector with the classification vectors characterizing the groups. The second level compares the query vector with all documents belonging to

the groups with whose classification vector the query vector was found to correlate highly on the first level. While increasing the search efficiency in terms of the number of comparisons made with a query vector, the clustering process inherently decreases the information which categorizes the individual documents.

Thus the problem of document clustering is clear: classifying the documents into homogeneous groups in order to increase search efficiency without seriously sacrificing the ability to retrieve the documents. Various schemes have been suggested to this end. These procedures gather documents into groups based on some type of correlation function which associates similar documents. Since there exists no purely analytical method for comparing the relative efficiency of these various methods, and since this "efficiency" may differ with the use to which the groups are put, it is nccessary to compare the clustering procedures in the context of an automatic information retrieval system, and to evaluate the merits of the clustering procedures by evaluating the overall performance of such a system, with only a change in the clustering method from experiment to experiment. The evaluation of the overall performance of such a system clearly involves the environment in which the system is used, and the users themselves.

Specifically, the present study consists of a comparison of two clustering procedures: a method of R. E. Bonner [2], and the method proposed by J. J. Rocchio [3]. These techniques can reasonably be studied by a comparison of the documents retrieved by an automatic retrieval system using a two level search based on the clusters produced by the methods with those retrieved by a full search of the document collection.

2. Similarity Measures

In the present system, documents and queries are represented as vectors in n-dimensional Euclidean space, where n is the number of allowable concepts or index terms in the system. Documents are retrieved on the basis of their closeness to the query vector, with "closeness" meaning small Euclidean distance between the vectors. Since the document vectors are of varying length, however, perpendicular distance at some fixed distance from the origin may not always be a good measure. Normalization of the document vectors so that their endpoints lie on the unit hypersphere, and use of the arc-length along the hypersphere as a distance measure, removes this problem. The measures used in the present study are functions of this arc length through the cosine of the angle between two document vectors. The measures used are defined in the following ways:

(1) Cosine measure

$$S_{d_1 d_2} = \frac{d_1 \cdot d_2}{|d_1| \cdot |d_2|}$$

(2) Tanimoto's measure [4]

$$S_{d_1 d_2} = \frac{d_1 \cdot d_2}{d_1 \cdot d_1 + d_2 \cdot d_2 - d_1 \cdot d_2}$$

where d_1 and d_2 are document vectors and $S_{d_1 d_2}$ is the similarity of document one with document two.

Bonner uses the coefficient (2) to form his document-document similarity matrices, while Rocchio uses the cosine measure to compare documents.

Rocchio states that since the cosine is used as the matching function to retrieve documents, clustering with it should give better results. To test this, Bonner's method is being tried using both coefficients.

3. Rocchio's Procedure*

The stated objective of Rocchio's procedure is that of jointly maximizing search efficiency and minimizing loss of relevant documents retrieved in the search. It is a heuristically derived algorithm which is meant to be used in conjunction with a two-level search. The input parameters to the algorithm are:

(1) the number of categories desired;

(2) lower and upper bounds on the number of elements to be allowed in a category;

(3) a lower bound on the correlation between a document and a classification vector, below which a document will not be placed in a category.

All documents are first considered unclustered, and pass from this state into one of two other possible states, clustered or loose. The algorithm proceeds as follows. An unclustered document is selected as a possible cluster center. All of the other unclustered and loose documents are correlated with it and the selected document is subjected to a region density test to see if a category should be formed around it. This test specifies that more than N_1 documents should be correlated higher than p_1 with the candidate, and that more than N_2 documents should be correlated higher than p_2 with the candidate. This ensures that documents on the edge of large groups do not become centers of groups. For example, in

* This section is a summary of section 4.5 of reference [3], to which the reader is referred for a more detailed discussion and program flowcharts.

Figure 1 below, document A would pass the test while document B would not (the documents are here represented by their endpoints on the unit hypersphere). If the document passes the

Example of the density test.

Figure 1

region density test, the cutoff on category size is used to find the lowest correlated document that would be included in the group. This figure is used to set up a correlation p_{min}. If a document correlates below p_{min} with a classification vector, it will not be included in the cluster. By using this cutoff, documents that are in an area between two classification vectors are likely to be included only in the cluster to which they are correlated more highly. This means that the boundaries between groups of documents which lie near each other will be sharpened, although some documents may still be included in both groups.

A classification vector is then formed by taking the centroid of all of the document vectors belonging to the cluster at this time. This centroid is then matched against the entire collection, and the cutoff parameters on category size are used to create the cluster. At this point, some documents may be in more than one cluster. Also, some documents which were in a cluster when the centroid was formed may no longer be in the cluster. These documents, as well as those which fail the region density test, are then marked loose, and those in the cluster are marked clustered.

This entire procedure is repeated with all unclustered documents. When this is completed, there is no guarantee that the required minimum number of clusters have been formed. Hence, some documents which failed the region density test in the first pass are chosen as cluster centers. If the number of categories formed in the first pass was too high, the density test may be made stricter and the first pass repeated.

There may still be some relatively isolated documents at the end of the clustering process. These correlate very poorly with any of the classification vectors. In order to properly test the procedures with the limited collections available, these documents are included in the cluster with which they correlate highest.

This procedure has been programmed for the CDC 1604 computer in FORTRAN 63 and CODAP by V. Lesser. The output of this program consists of a deck of punched cards which specify the documents belonging to each category and the classification vector for each category formed. These cards are then used as input to a two-level search procedure program, written by the authors in FORTRAN 63, which compares the queries to the documents in the clusters correlating highest with the query vector.

4. Bonner's Procedure

This algorithm is based upon Clustering Programs I and II and the Cluster Adjustment Program presented by Bonner in reference [2], and has been programmed by the authors for the CDC 1604 computer as FORTRAN 63 subroutines DOCDOC, SIMSIM, CLUSTER, and ADJUSTCL.

Subroutine DOCDOC accepts as input a binary document-term matrix and calculates a document-document similarity matrix S, using either of the

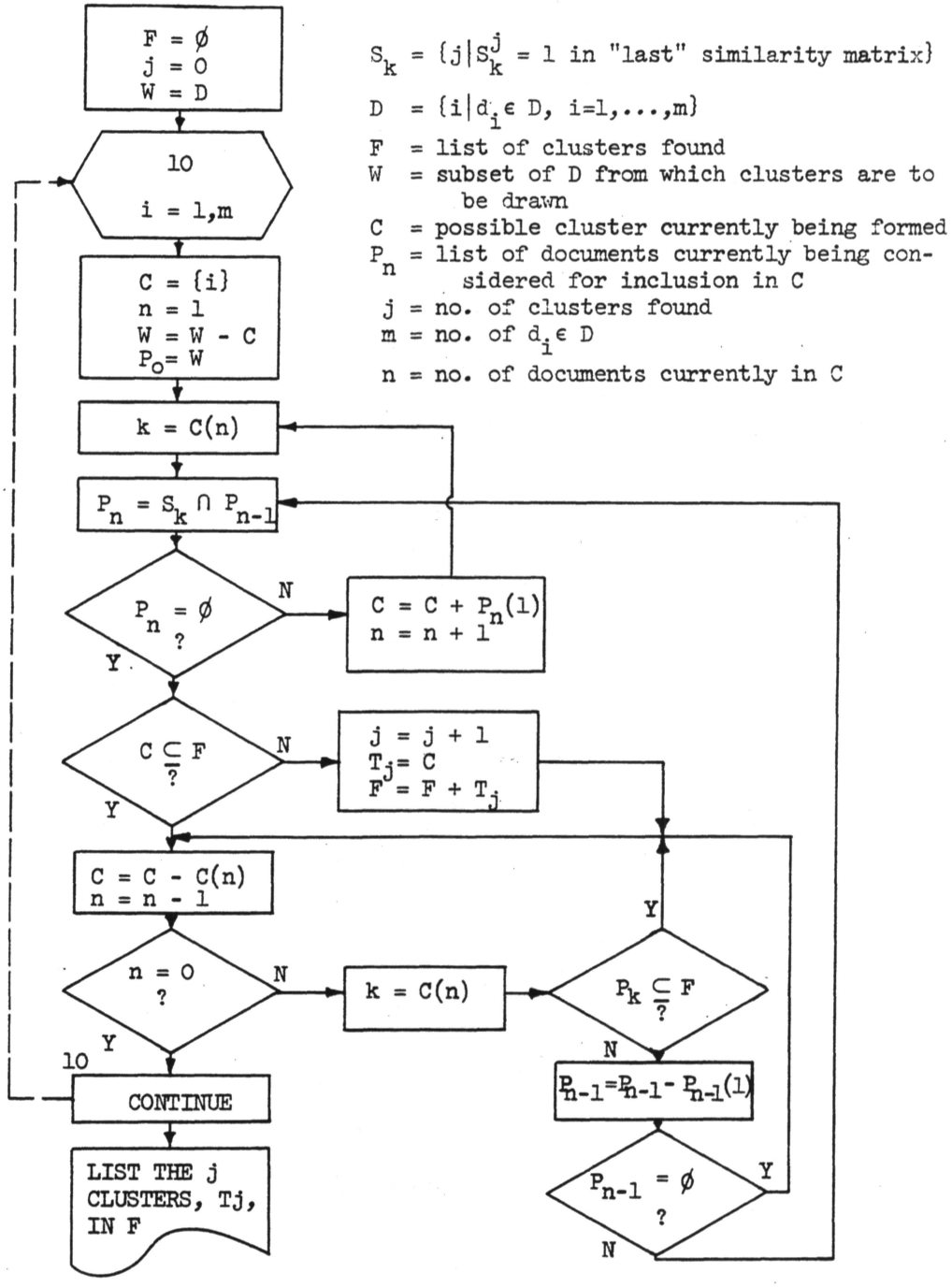
similarity measures discussed in section 2. S is then transformed into a binary matrix whose elements are zero (0) if the calculated similarity coefficient is below an empirically determined threshold, or one (1) otherwise.

Subroutine SIMSIM then calculates the similarity matrix of the document-document matrix in a manner completely analagous to DOCDOC. This is obviously equivalent to calculating the similarity matrix of another similarity matrix and is repeated as many times as necessary to better define the clusters which result.

Next, subroutine CLUSTER uses the algorithm as outlined in Figure 2 to define a "cluster" as that set of documents of maximum size where each document is similar (in the final document-document matrix of SIMSIM) to all other members of the set. In addition, no cluster can be a subset of another cluster.

After CLUSTER has found the "tight" clusters or, in graph theoretic terms, the maximal complete subloops of the document collection, subroutine ADJUSTCL attempts to redefine the clusters in a manner more conducive to search optimization. This is accomplished by transferring a member of a cluster with membership less than a predetermined constant into the large cluster to which it is most similar. This similarity is the percentage of documents in the large cluster to which the transferring document is linked in the first document-document similarity matrix. Before the transfer is made, however, this percentage is checked against a predetermined value (SIMTHRES), below which the transfer will not be made.

After all shifting is completed, the classification vector is calculated. This is merely the centroid vector of all of the document vectors



Bonner's Cluster Building Algorithm II

Figure 2

in the cluster. This set of classification vectors is used in the two-level search procedure, as described at the end of section 3.

5. The Experiment

The experimental environment consisted of 82 documents and 20 queries from an American Documentation Institute (ADI) collection. These documents were automatically indexed by the SMART automatic document retrieval system.[5] This system provided 82 document vectors and 20 query vectors in 601-dimensional Euclidean space. These vectors are then normalized to length one and used as input to both Bonner's and Rocchio's clustering procedures.

Since each of these procedures depends on several parameters, many runs were planned with each method in order empirically to determine desirable values for these parameters. The results presented below are based mainly on six computer runs, four using Bonner's clustering method and two using Rocchio's clustering method.

Each run consists of a clustering process for the 82 documents, resulting in the generation of the classification vectors, followed by a two-level search for each of the 20 queries. In addition, a run was made to match each query with the entire document collection using the cosine matching function, so as to obtain an ordering of the documents by this correlation for each of the 20 queries.

At the end of the first level of the search, only the two clusters which correlate highest with a query are retained. This is done because of the small size of the collection. If more than two clusters are retained, the total number of comparisons with a given query vector becomes

too large to obtain any savings over a search of the entire collection. In larger collections, one would probably wish to look not only at a certain number of clusters, but to take into account their correlations with a query as well, when determining how many clusters to search.

6. Evaluation

The results of the experiment are evaluated by using the recall and precision measures defined in reference [5]*, and the mean number of comparisons per query. In calculating the recall and precision measures, the question of which documents are relevant to a given query arises. For the 82 document ADI collection, judgments have previously been made by searching the entire collection manually to identify all documents relevant to a given query. It is not clear that recall and precision calculated by using these manual relevance judgments provide a good measuring device for comparing the clustering procedures. Hence, an additional set of relevance judgments has been made based upon the results of the search of the entire collection described in the previous section. For each query, all documents correlating above .30 with a query (.30 was chosen to give the same average number of relevant documents per query as the manual judgments gave) are considered relevant. In the results presented below, the recall and precision figures for both the manual and the "automatic" relevance judgments are given. The authors feel that both sets of measures should be considered in evaluating the performance of a clustering procedure as part of an overall automatic information retrieval system.

---

* Recall = $\frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents in the collection}}$

Precision = $\frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}}$

7. Results and Conclusions

The results of the study are summarized in Table 1. One striking characteristic of Bonner's method is the large number of clusters it produces. This is to be expected since Bonner refuses to associate dissimilar documents, whereas Rocchio allows dissimilar documents to be associated in order to build clusters of a size conducive to search efficiency.

Thus, one would expect that the mean number of matches made with a query vector using Rocchio's method would be less than the mean number of matches made using Bonner's method. The results support this hypothesis since approximately twice as many matches are required using Bonner's method as when using Rocchio's method.

Next, restricting our attention to the recall and precision results obtained using the manual relevance judgments, it is apparent that Bonner's method exhibits higher precision than, and nearly equivalent recall to Rocchio's method. One would expect the higher precision since there are far fewer members in each cluster, and hence, when a cluster is retrieved, it is more likely to contain a high percentage of relevant documents. Also, there is a higher similarity between members of the same cluster with Bonner's method than with Rocchio's method. Hence, if the cluster is similar, more of the documents in it are likely to be relevant since they are all very similar. The nearly equivalent recall between the two methods is somewhat surprising, as one would expect the large number of documents retrieved by using Rocchio's method to include more of the relevant ones. This is the case if the two highest ranking clusters are used, but is not true if only the highest ranking cluster is used, although using only the

No.	Procedure	Cutoff	Simthres	No. of Clusters	Mean No.of Matches	Mean No. of Documents Retrieved	Mean Recall[1]	Mean Precision[1]	Mean Recall[2]	Mean Precision[2]
1	Bonner (2) Cosine	.40	.60	45	52.6	7.6	.43	.26	.70	.70
2	Bonner (2) Tanimoto	.25	.60	51	59.2	8.2	.36	.22	.63	.70
3	Bonner (2) Tanimoto	.25	.40	49	57.4	8.4	.34	.20	.63	.63
4	Bonner (2) Tanimoto	.22	.60	40	52.0	11.4	.36	.19	.73	.56
5	Rocchio(1)			8	18.4	10.4	.28	.10	.49	.36
6	Rocchio(2)			8	29.7	20.8	.41	.06	.76	.30
7	Rocchio(1)			10	23.5	13.5	.43	.11	.59	.37
8	Rocchio(2)			10	35.8	25.0	.57	.08	.88	.27

[1]Measure based on manual relevance judgments.

[2]Measure based on "automatic" relevance judgments.

(1): Documents in highest correlated cluster only are retrieved.

(2): Documents in two highest correlating clusters are retrieved.

Summary of Results Using 82 Document ADI Collection with 20 Queries

Table 1

highest ranking clusters makes the number of documents retrieved closer to the number retrieved by Bonner's method.

The superiority of the cosine similarity measure is evidenced by entry number one in Table 1. In this run, the cosine coefficient was used to obtain the similarity matrices for Bonner's procedure. This entry clearly dominates the next three entries for any reasonable measure of "goodness". Further research using Bonner's method should therefore focus attention upon this similarity measure.

If we compare the results for all of the entries using the manual relevance judgments with those using the "automatic" relevance judgments, it is clear that the relative ranking of the entries is not changed. These results are, for all practical purposes, invariant under either set of relevance judgments.

The above results should be viewed in their proper light. The document collection is very small and the number of computer runs obtained so far too few to make any strong statements. It is clear that a more thorough analysis of the effects of the parameters of these clustering procedures should be made before any solid conclusions are drawn regarding the relative efficiency of the two methods tested.

References

[1] R. M. Needham, Applications of the Theory of Clumps, Mechanical Translation, Vol. 8, Nos. 3 and 4, June-October 1965.

[2] R. E. Bonner, On Some Clustering Techniques, IBM Journal of Research and Development, Vol. 8, No. 1, January 1964.

[3] J. J. Rocchio, Harvard University Doctoral Thesis, Report No. ISR-10 to the National Science Foundation, April 1966.

[4] D. Rogers and T. Tanimoto, A Computer Program for Classifying Plants, Science, Vol. 132, pp. 1115-1118, October 1960.

[5] G. Salton, et. al. Information Storage and Retrieval, Report No. ISR-9 to National Science Foundation, Harvard Computation Laboratory, August 1965.

[6] J. Rubin, Optimal Classification into Groups: An Approach for Solving The Taxonomy Problem, (unpublished IBM report), March 1966.

[7] W. D. Fisher, Two-Way Cluster Analysis, (preliminary report) Kansas State University, 1966.