A PATTERN RECOGNITION DOCUMENT CLASSIFICATION TECHNIQUE (UNE TECHNIQUE DE CLASSEMENT DE DOCUMENTS)

T.Z. Gottlieb and P.I.P. Boulton
Department of Electrical Engineering
Univ. of Toronto, Toronto, Ont. Canada

ABSTRACT

Some of the results obtained in the field of automatic document classification using pattern recognition techniques are presented. The classification problem is shown to be equivalent to the concept of unsupervised learning in pattern recognition theory. Criterion functions are introduced which formally define the notion of cluster quality, and an iterative procedure is presented which allows clusters to be locally optimized. Experimental results show that this approach leads to classes which are at least as good if not better for retrieval than previous techniques. (Quelque resultats obtenus dans le domaine de classification automatique de documents sont presentes).

INTRODUCTION

This paper presents some of the results obtained at the University of Toronto in the field of automatic document classification. Research into classification of such machine readable content vectors for retreival has been restricted mainly to Cornell University (Salton 1971, 1972 and Kerchmer 1971), the University of Maryland (Auguston et al, 1971 and Minker et al 1973) and the Cambridge Language Research Unit (Spark Jones et al, 1968, 1970). The general approach is to place these document vectors into a keyword feature space and use the relative vector positions to measure document similarity. Graph theoretic and other techniques have been used with some success for finding document classes. The pattern recognition approach discussed in the following sections is felt to be an improvement on these techniques.

THE PATTERN RECOGNITION VIEWPOINT

"Classification" in the pattern recognition sense is the assignment of a physical object or event to one of several prespecified categories. Normally, the item cannot be classified directly and so is passed through a feature extractor which examines the item and measures certain properties which are used later to distinguish between classes. The features are passed to a classifier whose output is a decision as to the correct class of the item. The picture of this process is typical of pattern recognition applications and is shown in figure 1.

DOCUMENT CLASSIFICATION

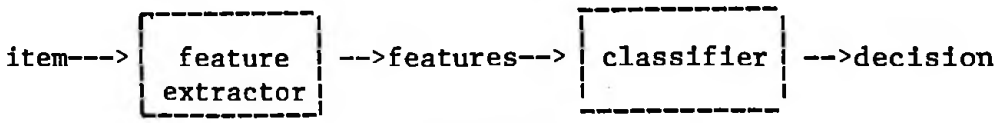

Figure 1 - Pattern Recognition Process

Thus, the main objective of pattern recognition theory is to produce a decision as to the class membership of an item.

The classification of documents for retrieval is a somewhat different concept. Here, the main object is to organize a collection of documents so that those items within a class possess some (hopefully strong) semantic similarity while those in different classes do not. This is achieved through the process shown in figure 2. The documents are analysed by a thesaurus or dictionary in order to reduce the amount of information into keyword vector form. These vectors are in turn analysed by the classifier which uses the information in order to estimate a partition of the vector space in which the documents are represented such that the above conditions of semantic relationships are satisfied.

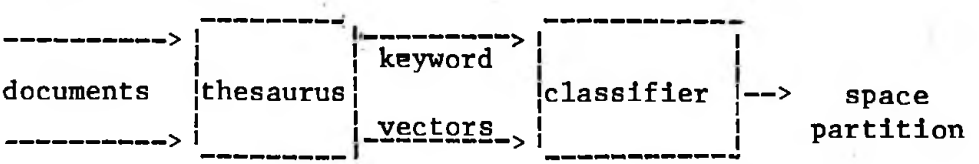

Figure 2 - Document classification process

The process of producing keyword vectors is simply a feature extraction in the pattern recognition sense. The process of classification, on the other hand, has somewhat different connotations in the two fields. In one case it is a clustering process while in the other it is a decision making process.

The basic reason for the difference noted above is that an assumption has been made in the pattern recognition case which does not generally hold true in cluster analysis. The assumption is simply that the type and number of classes is known (i.e. there exists a recognizable pattern). Furthermore, the underlying goal is to determine the single class to which an item belongs. This is specifically not true for document classifiers where an item may logically fit quite well in several classes. Also, the document classification process is only the beginning of retrieval which is to follow. In pattern recognition, the problem is considered terminated once a decision has been made as to the class of an item. This is the underlying reason why pattern recognition classifiers are judged on their ability to minimize the probability of error, i.e. putting an item into the wrong class, whereas document classifiers are judged by the a posteriori retrieval effectiveness that resulted from the classification. The two fields however, are so obviously related that much insight into the information retrieval area may be gained by studying pattern recognition.

DOCUMENT CLASSIFICATION

Unsupervised Learning

Pattern recognition assumes that every item belongs exclusively to one defined class. Unfortunately, document classification problems are not so simple. The conditions for class membership are not specifically known and documents may fit into more than one class. Pattern recognition tackles this problem by analysing a training sample and then using their features to build a recognizer assuming the sample to be representative of the classes.

When the training samples used to design the classifier are not labelled to show their category memberships, the learning process is said to be unsupervised. The classifier which results may be thought of as an algorithm for partitioning (clustering) the sample space into distinct classes. This process is pictured in figure 3.

Figure 3 - Building a Classifier

The similarity between figures 2 and 3 should be obvious. The classifiers in both diagrams now represent the same concept. The problem of document classification has been reduced to a classical pattern recognition problem of unsupervised learning. Compared to other more traditional problems, this problem is the most difficult. Let us see then what can be done when all one has is a collection of documents with no previous knowledge of its semantic structure. We first require answers to two important questions.

1) How do we measure the similarity between two documents in a cluster?

2) How do we evaluate a partitioning of documents into clusters?

Similarity Measures. A similarity function $s(x,y)$ may be used to compare two items x and y. It must be a symmetric function whose value is large when x and y are similar. The cosine of the angle between x and y is a commonly used measure. Other measures might include the fraction of shared features or the ratio of shared features to the number of features found in x and y together.

Performance Evaluation. Unfortunately, researchers in information retrieval have largely avoided the problem of evaluating the partitions of their document spaces and have concentrated instead on the problem of measuring only the retrieval characteristics of a partition. Thus if one partition produced better retrieval results than another, it was thought to be better without regard for any natural structure that might exist in the document collection. It may be that one classifier was

better able to find this structure whereas another tried to impose an unnatural structure on the data. It is our belief that more meaningful comparisons may be made given some measure of the goodness of the document space partitions being considered. Criterion functions provide us with such a measure.

CRITERION FUNCTIONS FOR CLUSTERING

Criterion functions can be used to represent document clustering as a well defined mathematical problem. Suppose we have a set D of n documents d1,...,dn which we wish to classify into exactly c clusters D1,...,Dc. The basic qualification for membership is that documents in the same cluster are more similar than documents in different clusters according to some well defined similarity function. A criterion function states this qualification mathematically and the problem is reduced to finding the partition which extremizes this function. Thus we can at least define what is meant by optimal cluster quality with respect to a particular criterion function J and a similarity function s.

Total Squared Distance Criteria. Given a cluster c of n documents and a dissimilarity function, it is easy to compute the inter-document distance matrix Dc.

$$Dc = \begin{vmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & & \vdots \\ \vdots & & & \vdots \\ d_{n1} & \cdots & & 0 \end{vmatrix}$$

Dc is symmetric with zeros on the diagnonal and so may be stored in upper triangular form. A simple criterion function therefore is the sum of the squares of the n(n-1)/2 inter-document distances for all clusters.

$$JT = \sum_{k=1}^{c} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d^2_{ij} = \sum_{k=1}^{c} Jk$$

The disadvantage of the total squared distance criterion function is that it tends to favor clusters of roughly the same size. Large clusters are discouraged even though they may be natural for certain collections. This is because the number of terms contributing to the sum is proportional to the square of the number of documents in the cluster. This advantage prompted the design of the following criterion function which is somewhat more sensitive to cluster size.

Weighted Average Squared Distance Criteria. The weighted average

criterion function is very similar to the total criterion JT defined previously but takes into account the size of clusters. If n_i is the number of documents in cluster i and J_i is sum of the $m_i(m_i-1)/2$ inter-document distances for cluster i then the weighted average c criterion function Jw may be defined by:

$$Jw = \sum_{i=1}^{c} \frac{2n_iJ_i}{n_i(n_i-1)} = \sum_{i=1}^{c} \frac{2J_i}{n_i-1}$$

Other formal pattern recognition criterion functions may be found in Duda and Hart (1973).

ITERATIVE OPTIMIZATION

An optimal solution has been defined as one which extremizes a criterion function. In theory, one can always find the optimal solution by exhaustive enumeration of all possible partitions. However, this approach would require consideration of more than 10^{67} partitionings in order to find 5 clusters in 100 items. Clearly, in most cases, such an approach is not feasible. Fortunately other techniques exist to help find optimal solutions. Suppose we have an unlabelled set of documents which we wish to organize into c clusters according to a particular criterion function. If J_i is the criterion value for cluster i then the criterion value for the entire organization can be represented by

$$J = \sum_{i=1}^{c} J_i$$

The basic approach is to start off with some initial approximation to the c clusters and then transfer documents from one cluster to another if the net effect of the transfer will be to improve the value of J. There are no guarantees using this method that the optimal solution will be reached. Instead, the final partition is said to be locally optimal and different initial partitions usually lead to different solutions. Nevertheless, the process may be repeated from various initial configurations in order to test the sensitivity of the collection to the criterion function and to the starting point. In classifications for retrieval, there is no equivalent concept which might determine the best possible result.

Suppose a document x' currently in cluster i is moved tentatively to cluster j. Then the criterion value of cluster j increases by $(J_j'-J_j)$ and cluster i decreases by (J_i-J_i'). The transfer of x' may be considered beneficial if

$$(J_i-J_i') - (J_j'-J_j) > 0$$

Clearly a good strategy would be to find that cluster j for which the above differences is maximum. This naturally suggests the following clustering procedure:

Procedure: Iterative Optimization.

P1: Select an initial partition of n samples into c clusters and compute J1,...,Jc.
P2: Select the next candidate sample x' in cluster i.
P3: For j = 1,2,...,c compute: $p(j) = (J_j - J_i') - (J_j' - J_j)$
P4: Transfer x' to cluster k if $p(k) \geq p(j)$ for all j.
P5: Update J_j and J_i.
P6: If there have been no transfers in n attempts, stop; otherwise go to P2.

An iteration of this procedure is defined to be one attempt at document transfer. A cycle is completed after n transfer attempts.

EXPERIMENTAL RESULTS

This section presents the results obtained from application of the criterion function classifiers on two document and query collections kindly provided by the SMART system at Cornell University. The collections are in the form of thesaurus vectors consisting of concept-weight pairs. The first is based on the full text of 82 short papers in the field of documentation originally provided by the American Documentation Institute (ADI). The second is based on the abstracts of 200 papers in aerodynamics from Cranfield (CRAN). For further characteristics of these collections, see Gottlieb (1974).

For a number of reasons, Rocchio's clustering method was chosen for comparison against the criterion function classifiers to be tested. Rocchio's method is relatively easy to implement and has been shown to produce performance results representative of results currently available in the field with other techniques (Minker et al, 1973 and Salton, 1971). Rocchio's algorithm was therefore applied to both the ADI and the CRAN document collections and produced 11 and 12 clusters respectively. For each of these organizations a number of random organizations of the same number of clusters were constructed. These random organizations were iteratively optimized with the criterion function classifiers. In addition the organizations produced from Rocchio's method were subjected to the iterative optimization procedure so that the criterion values both before and after optimization could be compared with those obtained from the random initial organizations.

In figures 4 and 5 the reduction in criterion value for several initial clusters is compared graphically with the Rocchio clusters. The random starting all show rapid decreases in criterion value for the first few cycles of the procedure. In terms of total squared distance, the ADI clusters were the poorest in comparison to the random clusters which in most cases converged more quickly and to lower minimum criterion values.

DOCUMENT CLASSIFICATION

The best Rocchio clusters were for the CRAN collection in a weighted average sense. Here the random configurations converged more slowly and did not result in lower criterion values. In summary, the graphs show that Rocchio's method tends to form clusters which are reasonably good in the total squared distance sense but much better in the weighted average measure.

In order to determine the relative effectiveness of the optimized clusters, searches were conducted and compared with results obtained from the unoptimized Rocchio clusters. The search strategy used was to examine the closest 1, 2, or 3 clusters and return the 20 documents closest to the query. These search results are summarized in table 1. The measure used to evaluate retrieval performance is the integrated recall precision measure (IRP) defined by Minker et al (1973), broken down by the number of clusters searched (search depth). The random initial organizations are listed in order of increasing criterion values. The IRP values in the table should be considered with the full search value in mind. If all documents are examined and the top 20 documents retrieved, then the IRP is 0.447 for CRAN and 0.409 for the ADI collection. In some cases the IRP value obtained was greater than the full search value.

As predicted by the discussions centered around figures 4 and 5, the retrieval results obtained from the optimized clusters do not show dramatic increases in IRP over Rocchio's clusters. Nevertheless, optimization of Rocchio's clusters produced an increase in IRP in 7 out of 12 cases and only an insignificant decrease in 2 of the remaining 5 cases. In the 60 searches performed with optimized clusters, 36 cases produced IRP values higher than the corresponding unoptimized Rocchio cluster searches. A further 9 cases produced values within 2% of the unoptimized clusters. In 4 cases, the optimized clusters produced better than full search results.

Unfortunately, table 1 does not bring out any direct observable relationship between criterion function value and IRP. Although efforts to minimize the criterion value have resulted (on the whole) in relatively good IRP measures, the recall/precision parameters (upon which IRP is based) are too sensitive to reveal this desirable relationship.

CONCLUSIONS

The formal criterion function classifiers designed specifically for document classification were shown in table 1 to be at least as effective if not better for retrieval than existing clustering techniques. This demonstrates that formal pattern recognition theory can successfully be applied to document retrieval and opens the way for research into other mathematically defined classifiers. Furthermore, the criterion function approach provides a method for measuring the quality of clusters produced by an classifier. This allows direct comparisons to be made between different techniques so that researchers may being to discover why certain classifiers work better on some collections than others.

DOCUMENT CLASSIFICATION

Similarly this quality measure can be used to compare different content analysis techniques so that techniques which produce good quality classes are developed. Only when some of these areas are investigated, will automatic classification begin to see significant use in production document retrieval systems.

REFERENCES

AUGUSTON, J.G., and MINKER, J., 1971, An Analysis of some Graph Theoretical Cluster Techniques. Journal of the ACM, 17(14).

DUDA, R.O., and HART, P.E., 1973, Pattern Classification and Scene Analysis. New York, John Wiley and Sons.

GOTTLIEB, T.Z., 1974, A Pattern Recognition Approach to Automatic Classification and Retrieval of Documents. M.A.Sc thesis, Department of Electrical Engineering, University of Toronto, 124p.

KERCHNER, M.D., 1971, Dynamic Document Processing in Clustered Collections. ISR Report # 19, Cornell University, Ithaca, N.Y.

MINKER, J., PELTOLA, E., and WILSON, G.A., 1973, Document Retrieval Experiments Using Cluster Analysis. Journal of the ASIS, 24(4): 246-257.

SALTON, G., 1971, The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall Inc. Englewood Cliffs, New Jersey.

_____. 1972, Dynamic Document Processing. Communications of the ACM, 15(7): 658-668.

SPARK JONES, K., and JACKSON, D.M., 1970, The Use of Automatically Obtained Keyword Classifications for Information Retrieval. Information Storage and Retrieval, 5(4): 175-201.

_____. and NEEDAM, R.M., 1968, Automatic Term Classification for Retrieval. Information Storage and Retrieval, 4(2): 91-100.

ADI COLLECTION - 11 CLUSTERS

INITIALIZATION & OPTIMIZATION CODE	CRITERION VALUE	TOTAL CENTROID SQUARED DIST	IRP by SEARCH DEPTH 1	2	3
1	185.293	32.005	0.329	0.345	0.364
2	149.837	36.634	0.292	0.371	0.374
3	145.692	37.141	0.345	0.395	0.396
3	147.012	37.635	0.364	0.406	0.395
3	147.575	37.139	0.248	0.342	0.381
3	151.596	36.827	0.357	0.385	0.411
4	52.054	32.005	0.329	0.345	0.364
5	49.627	35.873	0.265	0.324	0.376
6	49.316	39.151	0.323	0.342	0.370
6	51.514	34.631	0.371	0.372	0.392
6	52.829	38.597	0.283	0.309	0.334
6	53.008	32.559	0.316	0.319	0.365

Full search integrated recall/precision 0.409

CRAN COLLECTION - 12 CLUSTERS

INITIALIZATION & OPTIMIZATION CODE	CRITERION VALUE	TOTAL CENTROID SQUARED DIST	IRP by SEARCH DEPTH 1	2	3
1	865.230	27.919	0.328	0.412	0.446
2	740.466	30.814	0.362	0.448	0.442
3	739.455	30.726	0.401	0.461	0.450
3	743.116	30.718	0.333	0.409	0.435
3	744.762	30.355	0.356	0.434	0.430
3	751.915	30.492	0.357	0.437	0.444
4	102.403	27.919	0.328	0.412	0.446
5	98.537	30.447	0.338	0.424	0.440
6	101.315	32.382	0.386	0.401	0.415
6	103.303	28.820	0.329	0.412	0.419
6	104.773	27.365	0.367	0.450	0.443
6	104.778	28.301	0.325	0.415	0.428

Full search integrated recall/precision 0.447

CODES	INITIAL CLUSTERS	OPTIMIZATION PERFORMED
1	Rocchio	none
2	Rocchio	total
3	random	total
4	Rocchio	none
5	Rocchio	average
6	random	average

Table 1 - Search Results

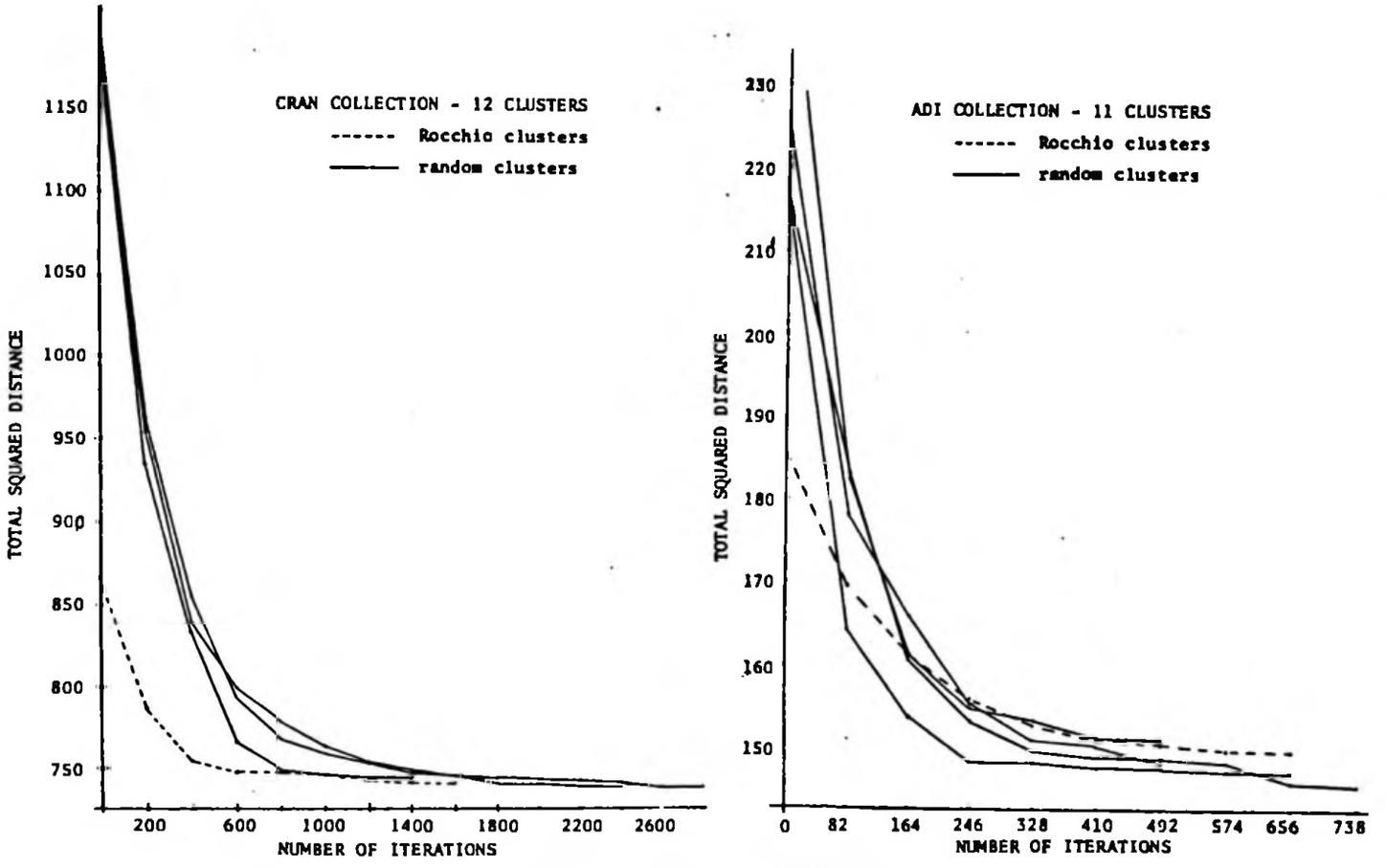

Figure 4 - Reduction of total squared distance.

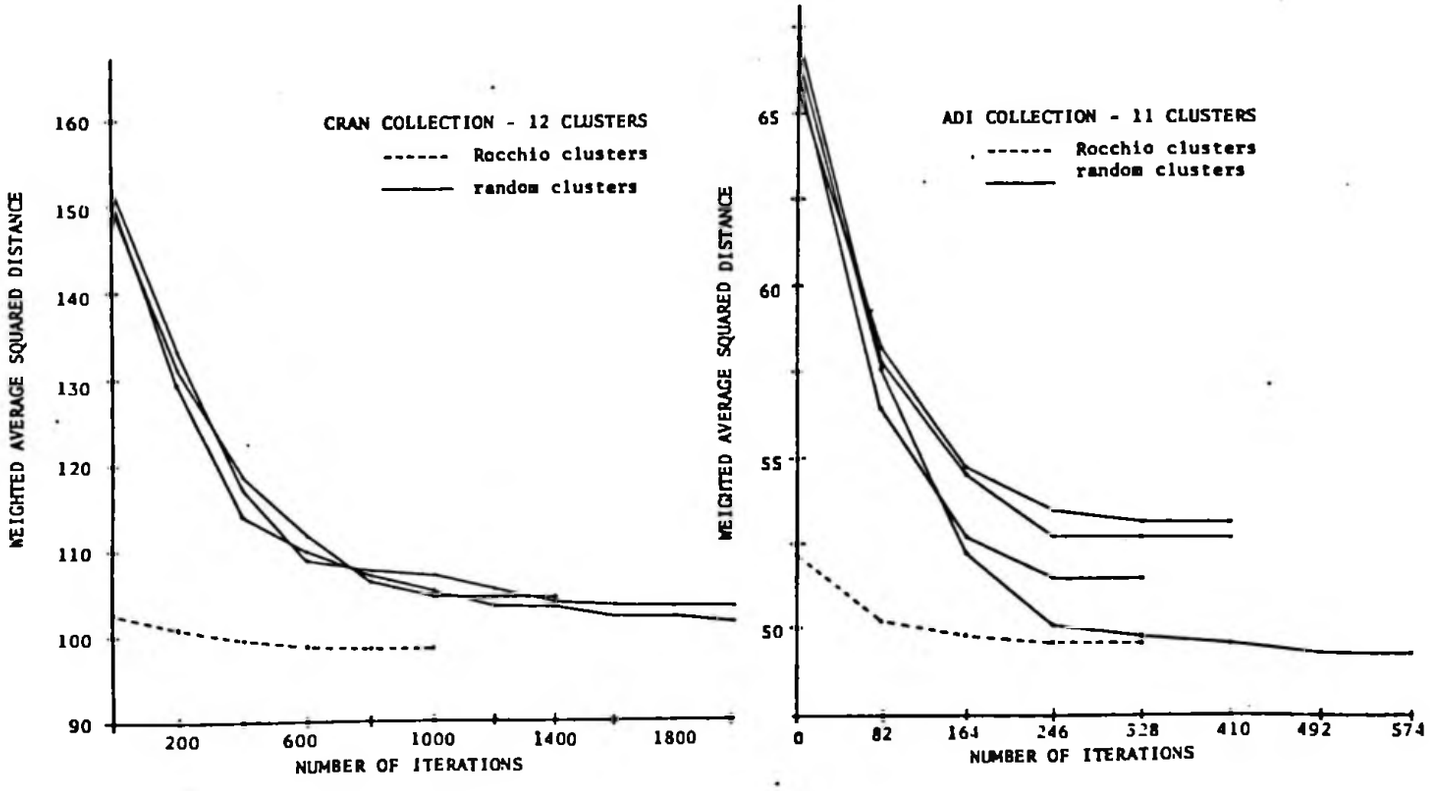

Figure 5 - Reduction of weighted average squared distance.