UIUCLIS--1999/6+IRG

A Review of Clustering Algorithms as Applied in IR

Qin He

Graduate School of Library and Information Science

University of Illinois at Urbana-Champaign

Table of Content

Abstract – Cluster analysis is an important procedure not only in the social sciences but also in library and information science. For about half of a century, cluster analysis has been studied and employed in many fields. This paper is a review of cluster analysis. The definition of a cluster and the properties of clusters are introduced first. Then choice of variables and similarity measures is discussed. A general review of clustering methods is given in the third part while some related issues are addressed in the fourth part. The conclusion provides a few cautions on and a guide to cluster analysis.

1. Introduction

"Cluster analysis" is the generic name for a wide variety of procedures that can be used to create a classification of objects. For about half of a century, cluster analysis has been developed and used in many fields, from social science to library and information science, etc. This paper will provide a review on cluster analysis and its applications in IR (Information Retrieval).

Before exploring the specific clustering methods, it is necessary to be clear on the definition of a cluster and the properties of clusters.

1.1 The Definition of a Cluster

Several different operational definitions of cluster have been proposed. According to Milligan (1980), one approach has adopted the view that clusters represent mixtures of multivariate normal populations. This approach has been adopted as a rationale and basis for the design of clustering algorithms by some researchers. An unconstrained mixture approach can result in overlapping cluster structures. Another approach has used the concept of an ultrametric space as a basis for the generation of cluster structure. It is noted that most hierarchical clustering routines invoke the ultrametric inequality. Thus, it would be expected that the algorithms should be able to recover structure in this type of space (p. 325).

Everitt (1980) has studied the definitions of a cluster and pointed out the common feature of most proposed definitions is their vague and circular nature, in the sense that

terms such as similarity, distance, alike, etc., are used in the definition, but are themselves undefined. The author describes clusters as continuous regions of a p-dimensional space containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points. Clusters described in this way are sometimes referred to as natural clusters. An advantage of considering clusters in this way is that it does not restrict the shape of clusters as rigidly as other proposed definitions do (p. 60).

1.2 Properties of Clusters

It is clear that clusters have certain properties. In early studies (Cormack, 1971), it is found that clusters should exhibit the properties of external isolation and internal cohesion. External isolation requires that entities in one cluster should be separated from entities in another cluster by fairly empty areas of space. Internal cohesion requires that entities within the same cluster should be similar to each other, at least within the local metric (as cited in Milligan, 1980, p. 326).

Sneath and Sokal (1973) have described a number of properties of a cluster, the most important of which are density, variance, dimension, shape and separation (as cited in Aldenderfer & Blashfield, p. 34).

- *Density* is a property of a cluster that defines it as a relatively thick swarm of data points in a space when compared to other areas of the space that may have comparatively few or no points.
- *Variance* is the degree of dispersion of the points in this space from the center of the cluster. Clusters can be said to be "tight" when all data points are near the centroid, or they may be "loose" when the data points are dispersed from the center.
- *Dimension* is a property closely related to that of variance. If a cluster can be identified, it is then possible to measure its "radius."
- *Shape* is simply the arrangement of points in the space. While the typical conception of the shape of clusters is that they are hyperspheres or ellipsoids, many different kinds of shapes, such as elongated clusters, are possible.
- *Separation* is the degree to which clusters overlap or lie apart in the space.

Taken together, these terms can be used to describe any type of clusters within a space.

1.3 Cluster Analysis

Aldenderfer and Blashfield (1984) define a clustering method as a multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups (p.7). Clustering methods have been recognized throughout this century. In 1963, Sokal and Sneath published the book *Principles of Numerical Taxonomy*. From then on, the literature of cluster analysis has exploded. Aldenderder and Blashfield think there are two reasons for the rapid growth of the literature on cluster analysis: a) the development of high-speed computers, and b) the fundamental importance of classification as a scientific procedure (p. 8).

Aldenderfer and Blashfield's (1984) study shows cluster analysis has been used to achieve four principal goals:

- Development of a typology or classification;
- Investigation of useful conceptual schemes for grouping entities;
- Hypothesis generation through data exploration, and
- Hypothesis testing, or the attempt to determine if types defined through other procedures are in fact present in a data set. (p. 9)

Similarly, Everitt (1980) lists seven possible uses of clustering techniques as follows:

a) Finding a true typology;
b) Model fitting;
c) Prediction based on groups;
d) Hypothesis testing;
e) Data exploration;
f) Hypothesis generating;
g) Data reduction (p.6).

The application of cluster analysis in IR could belong to the last group – data reduction. A IR system usually reduces the information on the whole set of N individuals (documents or terms) to information about k groups (categories) (where hopefully k is

very much smaller than N), so that people can retrieve similar documents or related terms by one query.

2. The Choice of Variables and Similarity Measurements

Choosing the variables and similarity measurements is the first step in a cluster analysis. This is a very important step, since the decision on these issues will affect the final results directly.

2.1 Choice of Variables

In general the raw data used in cluster analysis consist of an (N x P) matrix of measures, X, where X_{ij} is the score on the jth variable (character or attribute) for the ith individual (entity, or object) under study. Everitt (1980) thinks there are several questions that should be asked when we choose the variables (p. 9).

The first question about the variables is whether the correct ones have been chosen in the sense that they are relevant to the type of classification being sought. The second question that might be considered is how many variables are desirable in practical applications of cluster analysis techniques? A further important consideration is whether the data needs to be standardized in some way.

Similarly, before describing popular coefficients used in the calculation of similarity, Aldenderfer and Blashfield (1984) also briefly talk about the choice of variables and about data transformations prior to the calculation of similarity. The choice of variables to be used with cluster analysis is one of the most critical steps in the research process, but, unfortunately, it is one of the least understood as well (p.19). The basic problem is to find that set of variables that best represents the concept of similarity under which the study operates. In most statistical analyses the data are routinely standardized by some appropriate method. If the normality of a variable is in question, a logarithmic or other transformation is often performed.

In the IR field, such as co-word analysis, we also need to decide the variables to describe documents first. They could be keywords, thesaurus terms, subject headings as well as noun phrases from the article. Even if we have decided to use some variables,

such as noun phrases, we still need to decide the features of each variable, such as the length (number of words) of a noun phrase and the type (field) of noun phrases.

2.2 Similarity Measurements

The first stage in many cluster analyses is to convert the raw data matrix, X, into a matrix of inter-individual similarity (dissimilarity, or distance) measures. There are at least three concepts of similarity and distance, which need to be considered – between entities, between an entity and a group of entities, and between two groups of entities (Everitt, 1980, p. 12). Aldenderfer and Blashfield (1984) use "similarity coefficient" to describe any type of similarity measure and divide it into four groups (p. 17). Different similarity coefficients may have widely different values for the same set of data.

1) *Correlation Coefficients*

   The most popular correlation coefficient is the product moment correlation coefficient suggested by Karl Pearson, which is defined as:

   $$\mathbf{R_{jk}} = S\ \mathbf{(X_{ij} - Xb_j)\ (X_{ik} - Xb_k)\ /\ sqr\ [}S\ \mathbf{(X_{ij} - Xb_j)^2\ (X_{ik} - Xb_k)^2]}$$

   where X_{ij} is the value of variable i for case j, Xb_j is the mean of all values of the variables for case j.

   The correlation coefficient is frequently described as a shape measurement, in that it is insensitive to differences in magnitude of the variables used to compute the coefficient. One of the major drawbacks of the use of the correlation coefficient as a similarity measure is its sensitivity to shape at the expense of the magnitude of differences between the variables. (Aldenderfer & Blashfield, 1984, p.23)

2) *Distance measures*

   Distance measures have enjoyed widespread popularity because intuitively they appear to be dissimilarity measures. Distance measures normally have no upper bounds and are scale-dependent. Euclidean distance is generally used and it is calculated as:

   $$\mathbf{D_{ij} = sqr\ [\ (X_{ik} - X_{jk})^2\ ]}$$

   However, it is not necessarily the most suitable metric in all situations. When Euclidean distance is used on raw data, it may be very unsatisfactory since it is badly affected by changing the scale of a variable. It will not even preserve distance

rankings. Only the Euclidean distance calculated from the standardized variables will preserve relative distance. (Everitt, 1980, p. 17-18)

3) *Association coefficients*

There are three measures that have been used extensively and deserve special consideration, which are the simple matching coefficient, Jaccard's coefficient, and Gower's coefficient (Aldenderfer & Blashfield, 1984, p. 29). All of them are used to establish similarity between objects described by binary variables. Assume we have a 2 x 2 association table as follows.

	1	0
1	a	b
0	c	d

a) The simple matching coefficient is defined as **S = (a+d) / (a+b+c+d)**;

b) Jaccard's coefficient is defined as **S = a / (a + b + c);**

c) Gower's coefficient is defined as $\mathbf{S_{ij}} = \sum \mathbf{S_{ijk}} / \sum \mathbf{W_{ijk}}$, where k = 1 to p and W_{ijk} is a weighting variable valued at 1 if a comparison of variables is considered valid and 0 if it is not.

Gower's similarity coefficient can be used with data containing a mixture of binary, qualitative and quantitative variables. This coefficient has a number of appealing features beyond its ability to accommodate mixed data types. These include its metric qualities and its flexibility, in that, the method can be easily modified to include negative matches in the estimation of similarity by simply modifying the binary weighting system (Aldenderfer & Blashfield, 1984, p. 32).

4) *Probabilistic similarity coefficients.*

Coefficients of this type are quite different from those described above in that the similarity between two cases is not actually calculated. Instead, this type of measure works directly upon the raw data. An important point about probabilistic measures is that they can be used only with binary data. No workable schemes for using this type of measure with quantitative and qualitative variables have been developed yet (Aldenderfer & Blashfield, 1984, p. 33).

A feature of the similarity matrices is the number of "ties", where several cases have the same similarity value. It is worth noting that some clustering methods perform

poorly when many ties are present in the similarity matrix (Aldenderfer & Blashfield, 1984, p. 31). There are four standard criteria can be used to judge whether a similarity measure is a true metric, which include Symmetry; Triangle inequality; Distinguishability of nonidenticals, and Indistinguishability of identicals (Aldenderfer & Blashfield, 1984, p. 18). In addition, allied to the measurement of similarity is the question of weighting of variables (Everitt, 1980).

2.3 The Similarity Measurements in IR

In IR, clustering algorithms are used for two types of studies: clustering documents based on the terms that they have in common and clustering terms based on the documents in which they co-occur. The co-occurrence frequency is always used as the inter-term similarity in term clustering. However, the similarity measurements in document clustering are more complicated.

There are two kinds of document clustering. One of them is citation clustering, which involves measuring the similarity between a pair of documents by the citations which they share in common. The other is to cluster documents based on the index terms that describing the content of the documents. The later one is discussed more frequently using a vector model, in which each document is identified by one or more index terms T_i and represented by a n-dimensional vector $D_i = (d_{i1}, d_{i2}, \ldots, d_{in})$, where d_{ij} is the weight of the *j*th term in document *i* and *n* is the total different index terms in the collection (Salton, Wong, & Yang, 1975). In this model, there are three main classes of similarity coefficients have been used to determine the similarity degree between two documents: distance coefficients, association coefficients, probabilistic coefficients (Willett, 1988). Distance coefficients, such as the Euclidean distance, is not widely used for document clustering due to its limitation that it can lead to two documents being regarded as highly similar to each other, despite the fact that they share no terms at all in common. Conversely, association coefficients have been used more widely. Three commonly used normalized association coefficients are the Dice coefficient, Jaccard coefficient and Cosine coefficient (Rasmussen, 1992). Probabilistic coefficients have been used in El-Hamdouchi's work, in which the main criterion for the formation of a cluster is that the

documents in it have a maximal probability of being jointly corelevant to a query (as cited in Willett, 1988).

Willett (1983) has compared four similarity coefficients (the vector or inner product, the Tanimoto coefficient, the cosine coefficient, and the overlap coefficient) and five weighting functions with three document test collections, using single linkage clustering methods. The results show the cosine and Tanimoto coefficients generally give rather better levels of retrieval effectiveness than the vector product and overlap coefficients. The results also suggest that although inverse frequency weighing may be useful in certain cases, it does not consistently lead to significant increases in performance over the use of unweighted terms. In addition, it is found that it is important to use a measure that is normalized by the length of the document vectors.

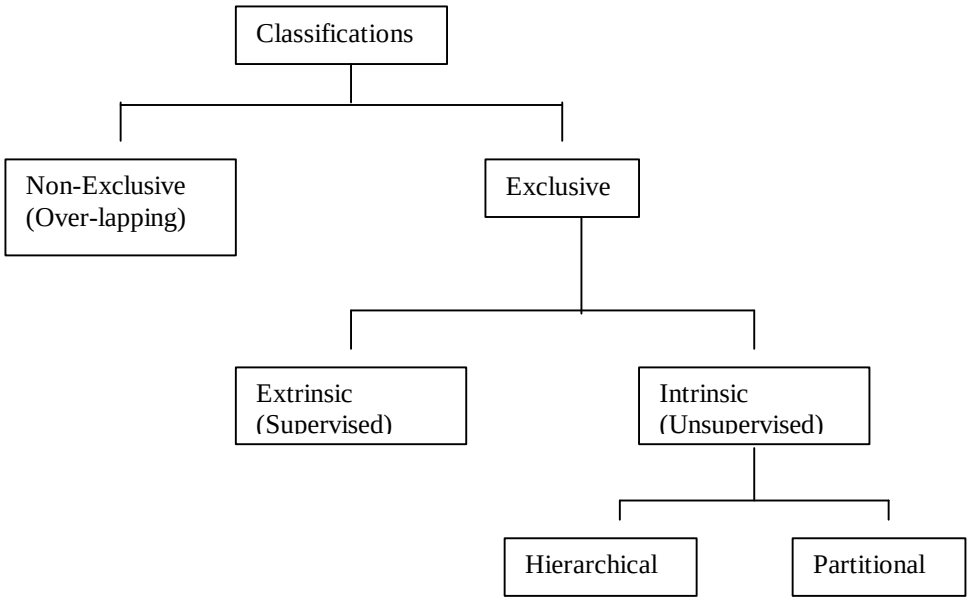
Willett's review (1988) shows that the similarity coefficient may affect the clustering that is obtained, but there seems to be little consensus as to which types of coefficient are most generally applicable.

3. A General Review of Clustering Methods

3.1 Clustering Methods

According to Aldenderfer and Blashfield (1984, p.35), seven major families of clustering methods have been developed, which include: 1) hierarchical agglomerative; 2) hierarchical divisive; 3) iterative partitioning; 4) density analytic; 5) factor analytic; 6) clumping, and 7) graph theoretic.

Jain & Dubes (1988) have shown a tree of classification problems suggested by Lance and Williams in the following figure (p. 56).



However, Everitt (1980) thinks Cormack's paper (1971) remains outstanding and his suggested classification of cluster analysis techniques includes roughly five types: 1) hierarchical techniques; 2) optimization techniques; 3) density or mode-seeking techniques; 4) clumping techniques; 5) others (p, 23). The following section will review the clustering methods by combining Cormack's and Aldenderfer and Blashfield's classification.

3.1.1 Hierarchical Methods

Essentially, hierarchical techniques may be divided into agglomerative methods, which proceed by a series of successive fusions of the N entities into groups, and divisive methods, which partition the set of N entities successively into finer partitions.

1) Agglomerative Methods

There are many different agglomerative methods. At any particular stage, all the methods fuse individuals or groups of individuals which are closest (or most similar). Differences between these methods lie in the different ways of defining distance (or similarity) between an individual and a group containing several individuals, or between groups of individuals (Everitt, 1980, p. 25).

- The nearest neighbor or single link method.

  This method can be used both with similarity measures and with distance measures. Groups are fused according to the distance between their nearest members; the groups with the smallest distance being fused.
- The furthest neighbor or complete linkage method.

  This method is exactly the opposite of the single linkage method, in that distance between groups is now defined as the distance between their most remote pair of individuals.
- Centroid cluster analysis.

  The distance between groups is defined as the distance between the group centroids. A disadvantage of this method is that, if the sizes of the two groups to be fused are very different, the centroid of the new groups will be very close to that of the larger group and may remain within that group. The characteristic properties of the smaller group are then virtually lost.
- Median cluster analysis.

  Assuming that the groups to be fused are of equal size, the apparent position of the new group will then always be between the two groups to be fused. If the centroids of the groups to be fused are i and j, then the distance of the centroid of a third group k from the group formed by fusion of i and j lies along the median of the triangle defined by i, j and k.
- Group average method

  This method defines distance between groups as the average of the distances between all pairs of individuals in the two groups.
- Ward's method

  Ward proposes that the loss of information which results from the grouping of individuals into clusters can be measured by the total sum of squared deviations of every point from the mean of the cluster to which it belongs. At each step in the analysis, union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in the error sum of squares are combined.
- Lance & Williams flexible method

The distance measures between groups used by many cluster analysis methods satisfy a recurrence formula for the distance between a group k and a group (ij) formed by the fusion of group i and group j. This formula is given by Lance and Williams and is defined as:

D k(ij) = a**iDki +** a**j Dkj +** b**Dij +** g**| Dki – Dkj |,**

where Dij is the distance between groups i and j and α, β and γ are parameters whose values are different for different methods above. The recurrence formula makes the computer implementation of agglomerative methods relatively easy.

2) Divisive Methods

Two families of divisive techniques are discussed by Everitt (1980): *monothetic*, which are based on the possession or otherwise of a single specified attribute, and *polythetic*, which are methods based on the values taken by all the attributes.

The most feasible of the *polythetic* divisive techniques is that described by Mac-Naughton-Smithe et al. (as cited in Everitt, 1980, p.35). In this instance a splinter group is accumulated by sequential addition of the entity whose total dissimilarity with the remainder less its total dissimilarity with the splinter group is a maximum. This method has the advantage that the computation required is considerably less than for 'an all possible sub-divisions method'.

Monothetic techniques are usually used in cases where we have binary data. A division of the set of data is then initially into those individuals who possess, and who lack, some one specified attribute. Two *monothetic* methods are discussed in more detail by Everitt (1980) as follows.

- Association Analysis

  This method creates division of a cluster into two sub-sets in terms of the presence and absence of one of the binary characters, say T. It has several variants which differ in the division criterion adopted. When N individuals have been scored on m attributes, and a, b, c and d are the usual cell counts in the fourfold table for attribute j and k, the *m x m* matrix of chi-square coefficients can be obtained, where,

  $C^2_{jk} = (ad - bc)^2 N / (a+b)(a+c)(b+d)(c+d)$

The most common division criterion adopted in association analysis is to divide the data on that attribute k, which makes $\{\sum X^2_{jk}\}$ a maximum. Each division node can be located at a precise hierarchical level on a division tree according to the max $\{\sum X^2_{jk}\}$ value.

- The Automatic Interaction Detector Method (A.I.D.)

  This is a multivariate technique for determining those variables, and the categories within them, which combine in defining groups, which are maximally different with respect to some dependent variable. The method proceeds by dividing the sample through a series of binary splits into mutually exclusive monothetic classes. At each binary split the method seeks optimal reduction in the unexplained sum of squares of the dependent variable. The splitting point is that which maximizes the between group sum of squares, B.S.S., which is defined as

  $$\mathbf{B.S.S_{ik} = (N_1 * Y_1^2 + N_2 * Y_2^2) - N_{12} * Y^2_{12},}$$ where

  $N_{12} = N_1 + N_2$ size of parent group; N_1 is size of first sub-group; N_2 is size of second sub-group; Y_1 is mean of criterion in first sub-group; Y_2 is mean of criterion in second sub-group; Y_{12} is mean of the criterion in the parent group being split; k is any predictor variable.

3.1.2 *Optimization techniques (Iterative partitioning methods)*

Optimization techniques differ from the hierarchical techniques in that they admit relocation of the entities, thus allowing the possibility that a poor initial partition might be corrected at a later stage. They also differ from those in that their solutions do not necessarily portray hierarchical relationship among the entities. Most of the methods assume that the number of groups has been decided *a priori* by the investigator, although some do allow the number to be changed during the course of the analysis. Optimization techniques are so called because they seek a partition of the data, which optimizes a predefined numerical measure, high values of which are indicative of a desirable clustering solution. Difference between methods arises with respect to 1) the methods by which they obtain an initial partition of the data and 2) the clustering criterion which they seek to optimize (Everitt, 1980).

1) Techniques used for initiating clusters

Most techniques begin by finding k points in the p-dimensional space, which act as initial estimates of the cluster center. Entities are allocated to the cluster to whose center they are nearest, and the estimate of the center may be updated after the addition of each entity to the cluster. Various procedures have been suggested for choosing the initial points, such as the first k points in the sample, the k points mutually furthest apart and so on.

2) Clustering criteria

Once an initial classification has been found, a search is made for entities whose reallocation to some other group will cause an improvement in the particular clustering criterion. Various clustering criteria have been proposed for this purpose.

- Minimization of Trace (W);
- Minimization of the Determinant of W
- Maximization of Trace (BW-1)

These three criteria are derived from the well-known matrix identity T = W + B, where T is the total scatter or dispersion matrix, W is the matrix of 'within'-groups dispersion – that is $W = \Sigma Wi$ where Wi is the dispersion matrix for group i – and B is the 'between'-groups dispersion matrix. In addition, there are another two criteria proposed as follows.

- Average entity stability.

  This measure is based on defining the 'attraction' of an entity to a group as the average similarity between the entity and the members of the group. If the entity is attracted to an 'outside' group more than to the group it is in, then it is said to be 'unstable'. The measure of entity stability for object i in group G is defined by the following two terms.

  M = attraction of i to other entities in the group G;

  Mb = maximum attrraction of i to any group except G, and

  Entity stability $Oi = S^{*}M - (1-S^{*}) Mb$, for i = 1, …, N, where S^{*} is the attraction of any individual to an empty set.

- Information measure.

This is a function of the data, measurement accuracies and parameters of certain distributions, which the method assumes for the variables. The latter can be multistate or continuous, or a combination of both. Multinomial distributions are assumed for the multistate data, and normal distributions for the continuous data.

Optimization techniques are called iterative partitioning methods by Aldenderfer and Blashfield. Unlike hierarchical agglomerative methods, iterative partitioning methods have not been extensively used or examined (Aldenderfer & Blashfield, 1984). K-means is one iterative method that works directly upon the raw data. It therefore offers the opportunity of handling distinctly larger data sets than hierarchical methods. Moreover, iterative methods make more than one pass through the data and can compensate for a poor initial partition of the data, thereby avoiding one of the major drawbacks of hierarchical agglomerative methods (p. 46).

Most of the heuristic, computational and statistical properties of iterative partitioning methods can be summarized by reference to three major factors: (1) choice of initial partition; (2) type of pass; and (3) statistical criterion (p. 47-48).

3.1.3 Density Search Techniques

If entities are depicted as points in a metric space, a natural concept of clustering suggests that there should be parts of the space in which the points are very dense, separated by parts of low density. This could form the basis for the definition of a natural cluster (Everitt, 1980, p. 46). There are many methods of cluster analysis, which use this approach of seeking regions of high density or modes in the data.

- The Taxmap method of Carmichael and Sneath.

  It attempts to imitate the procedure used by the human observer for detecting clusters in two or three dimensions, that is to compare relative distance between points, and to search for continuous relatively densely populated regions of the space surrounded by continuous relatively empty regions. Clusters are formed initially in a way similar to the single linkage method, but criteria are adopted for judging when additions to clusters should be stopped. One such criterion is to terminate addition if the prospective point is ‘much’ further away than was the last point admitted.

- Gitman and Levine's methods for detecting unimodal fuzzy sets.

  This method also starts in a similar way to the single linkage clustering technique, but considers points for allocation to clusters in a particular order.

- Cartet count method.

  Essentially this method consists of partitioning the multidimensional space and counting the number of points in each cartet (or hyper cube). Fixing a 'significantly high density' count, relative to the average total density, enables clusters to be found (Everitt, 1980, p. 49).

- Mode analysis.

  This is a derivative of single linkage clustering which searches for natural sub-groupings of the data by estimating disjoint density surfaces in the sample distribution. One important difficulty with mode analysis is its failure to identify both large and small clusters simultaneously.

- Method of mixtures

  Since the members of a class differ from one another, it is reasonable to assume the existence of a probability distribution of characteristics for the members. The combined population of several classes has a probability distribution, which is a mixture of distributions. Mixture distributions arise in a wide variety of practical situations ranging from distributions of wind velocities to distributions of physical dimensions of various mass-produced items.

3.1.4 Clumping techniques

In some cases, classification must permit an overlap between the classes if it is to be of any value, because words tend to have several meanings, and if they are being classified by their meanings they may belong in several places. In general, classification techniques which allow overlapping clusters are known as clumping techniques, a terminology which seems to have been first introduced by Sparck Jones and Needham and fellow workers at the Cambridge Language Research Unit.

These methods seek a partition of the entities into two groups, the smaller of which is generally considered to be the class sought. Partitions are found by minimizing

a cohesion function between the two groups. Needham considered a symmetric cohesion function, $G_1(A)$ given by:

$$\mathbf{G_1(A) = S_{AB} / (S_{AA}S_{BB})}$$

And Parker-Rhodes and Jarckson suggested a modification $G_2(A)$ given by

$$\mathbf{G_2(A) = [S_{AB} / S_{AA}] [na(na-1)/ S_{AA} - S_{AA} /p\ na(na-1)]}$$

Where A and B refer to the two groups, A being the putative clump. Sxy is the sum of the similarities between members of x and y. na is the number of entities in A, and p is an arbitrary parameter, which allows the user some control over the size of clumps and the amount of overlap. The cohesion function $G_1(A)$ is designed to find a good partition of the set of entities, while $G_2(A)$ allows the internal similarities of A and the separation of A from B to be adjusted relative to each other by the parameter p.

3.1.5. Factor analysis variants

Most of the methods in this group are known as factor analysis variants, inverse factor analysis, or Q-type factoring. These methods start by forming a correlation matrix of similarities among cases. Conventionally, factor analysis is performed on a P x P correlation matrix, but when used to define clusters, it is performed on the N x N correlation matrix. Factors are extracted from the correlation matrix, and the cases are assigned to clusters based on their factor loadings (Aldenderfer & Blashfield, 1984, p. 49).

Criticisms of factor analytic methods of clustering include the implausible use of a linear model across cases, the problem of multiple factor loadings (what is to be done with a case that has high loadings on more than one factor), and the double centering of the data (Everitt, 1980; Aldenderfer & Blashfield, 1984).

'Inverse' or Q factor analysis: in this method individuals and variables are interchanged with respect to a normal factor analysis, so that correlations become correlations between individuals rather than between variables (Everitt, 1980).

In addition, there are some other clustering methods, such as Ling's generalization of single and complete link clustering to find what are termed (k, r) clusters (Ling, 1972).

It is worth noting that the results obtained from different methods on the same data can be very different. Certain families of methods have been found to be particularly useful in specific sciences. What is important to remember, when faced with the difficult choice of which clustering method to use, is that the method must be compatible with the desired nature of the classification, the variables to be used, and the similarity measure used to estimate the resemblance between cases if one is required (Aldenderfer & Blashfield, 1984, p. 35).

3.2 Problems of Some Clustering Techniques

There are problems existing in some of the clustering methods. Everitt (1980) has studied this issue and found the following results:

3.2.1 Problems of hierarchical clustering techniques

There are several hierarchical clustering techniques that give rise to a property called chaining, which refers to the tendency of the method to cluster together at a relatively low level objects linked by chains of intermediates. This is often viewed as a defect in the methods. Hierarchical clustering techniques have a general disadvantage since they contain no provision for reallocation of entities which may have been poorly classified at an early stage in the analysis (Everitt, 1980, p.68). Because of the arbitrariness involved in scaling and combining different variables, there is rarely any justification for using the particular values rather than values obtained from some monotonic transformation.

3.2.2 Problems of optimization techniques

The most commonly used numerical criteria for these methods is minimization of trace (W) which nevertheless suffers from a number of serious problems. Firstly, the method is transformation dependent. A further problem is that its use implies that we are interested in finding spherical homogeneous clusters, and clusters of this shape will be found even if the natural clusters in the data are of other shapes (Everitt, 1980, p. 69). A further practical problem of optimization type clustering techniques is that they consume

large amounts of computer time, which makes them unsuitable for use with very large data sets.

3.2.3 Problem of density-seeking methods

Methods such as the fitting of mixtures of multivariate normal distributions also suffer from the problem of sub-optimal solutions, since there may be more than one solution to the maximum likelihood equations. A further difficulty with this method lies in the assumption of multivariate normal distributions. Another problem with this approach arises when the assumption that the component normal distributions have the same variance-covariance matrix is abandoned. Other density-seeking techniques suffer from the problem of containing various parameters which control the techniques, and which have to be arbitrarily chosen by the investigator.

3.2.4 Problem with other methods

In Q factor analysis, a problem arises when a person has sizable factor loadings on more than one factor, making classification difficult. Ordination methods such as principal components and multidimensional scaling also have their problems. Multidimensional scaling techniques also suffer from the problems of sub-optimal solutions corresponding to those found with the optimization clustering techniques.

3.3 The Applications of Clustering Algorithms in IR

According to Rasmussen (1992), cluster analysis can be performed on documents in several ways:

- Document clustering. Documents are clustered on the basis of the terms that they contain. The aim of this approach has usually been to provide more efficient or more effective retrieval, though it has also been used after retrieval to provide structure to large sets of retrieved documents. Many experiments with clustering algorithms fall in this category and Croft's test (1977) is an early example.
- Co-citation analysis. In this case, documents are clustered based on co-occurring citations in order to provide insights into the nature of the literature of a field. An example is Braam, Moed, and Raan's (1991) work, which uses a special clustering

routine to aggregate clusters of cited documents by sequentially linking together all selected pairs of cited documents that have at least one cited document in common.

- Term clustering. Terms may be clustered on the basis of the documents in which they co-occur, in order to aid in the construction of a thesaurus or in the enhancement of queries. An example is Crouch's paper (1988), which describes an approach to the automatic generation of global thesauri, based on the discrimination value model of Salton, Yang , and Yu and on a complete link clustering algorithm.

As it is shown above, there are many clustering methods available. The choice of the method will determine the outcome, and the choice of algorithms will determine the efficiency with which it is achieved. It is worthy to find out which methods can be and have been used in IR.

3.3.1 The Application of Nonhierarchical Methods

The nonhierarchical methods were used for most of the early work in document clustering, such as the clustering experiments carried out in the SMART project (Salton, 1971). With limited computational resources at that time, the emphasis of these experiments is on the efficiency of best matching search in IR systems due to the time and storage requirements of these methods are much lower than those of the hierarchical methods and much larger data sets could be processed.

Single pass is one of the nonhierarchical methods with the advantage of simplicity (Rasmussen, 1992). However, it is often criticized for its tendency to produce large clusters early in the clustering pass, and because the clusters formed are not independent of the order in which the data set is processed. The cover coefficient algorithm (Can & Ozkarahan, 1984) is an example of a single pass algorithm developed for document clustering. In this algorithm, a set of documents is selected as cluster seeds, and then each document is assigned to the cluster seed that maximally covers it. For a document, the cover coefficient is a measure that incorporates the extent to which it is covered by this document and the uniqueness of this document, that is, the extent to which it is covered by itself. A multi-pass algorithm using Dice's coefficient for document

assignment to the cluster seeds is also studies in this work to show the validity of the new cover coefficient algorithm.

The reallocation method (Rasmussen, 1992) is an improvement of single pass and similar to the K-means method (which will be reviewed later in detail). However, the results are still arbitrary in operation since the final clusters may depend on the order in which the document file is processed, the random selection of documents as initial cluster centers, or the exact parameter values that are used. So, the nonhierarchical methods are not used as frequently as the hierarchical methods any more in late IR systems.

3.3.2 The Application of Hierarchical Methods

With improvements in computational resources, the easy available of software packages for cluster analysis, and improved algorithms, hierarchical agglomerative clustering methods have been used more frequently in late IR systems. Willett (1988) provided a very comprehensive review on this.

Since the commonly used hierarchical methods, such as single linkage, complete linkage, group average linkage, and Ward's method, have high space and time requirements, some efficient algorithms been implemented to optimized each method and make them more suitable for document clustering. Based on Willett's (1988) and Rasmussen's (1992) article, a brief review of the optimized algorithms for each method is as followings.

- Single linkage. The SLINK algorithm of Sibson (1973) is a single linkage method that operates by progressively updating a hierarchy. It has a low time and storage requirement and been shown to be optimally efficient. Minimal spanning tree (MST) is another algorithm closely related to the single link method. A spanning tree is a set of N-1 similarities that links all of the N objects in the data set together into a connected graph without any circuits. MSTs have been used for the clustering of index terms in studies of query expansion (van Rijsbergen, 1977; van Rijsbergen, Harper, & Porter, 1981). Voorhees (1986) has also devised an algorithm to provide an efficient way of generating single linkage classifications. This algorithm "operates by successively identifying that document not currently

linked into the hierarchy which has the greatest similarity to a document that is so linked (as cited in Willett, 1988, p. 583)."

- Complete linkage. CLINK is a modification of Sibson's SLINK with the complexities for the complete linkage method (Defays, 1977). However, this algorithm has been shown to give very poor levels of retrieval effectiveness. The reason is that the CLINK algorithm does not seem to generate an exact complete linkage hierarchy. Voorhees (1986) has also implemented a complete linkage algorithm, which has been tested with large collections and has yielded much more satisfactory levels of retrieval effectiveness. However, this algorithm is very demanding of storage and involves extended execution times when used with large document collections (as cited in Willett, 1988).
- Group average linkage. Voorhees (1986) has noted that the sum of the inter-document similarities is equal to the inter-centroid similarity if a suitable weighting of the centriod elements is used. She has made use of this fact in her group average algorithm and improved the performance of this method (as cited in Willett, 1988).
- Ward's method. The mathematical properties of Ward's method make it a suitable candidate for a *reciprocal nearest neighbor* algorithm. El-Hamdouchi and Willett (1986) have used this algorithm to generate Ward classifications of several document test collections.

It is noted that all the algorithms above require a very large amount of memory. SLINK is the one with least overheads and quite simple to implement (Willett, 1988).

Over all, with limited computational resources, clustering analysis was initially used in IR systems for efficiency, that is, document collections were partitioned, using nonhierarchical methods, and queries were matched against cluster centroids, which reduced searching time. However, studies of retrieval from partitioned document collections showed that there was a decrease in retrieval effectiveness, even though retrieval efficiency was achieved. Thus, with the improvements in computation resources, subsequent IR studies have concentrated on the effectiveness of retrieval from hierarchically clustered collections. Many hierarchical clustering algorithms have been

devised and tested. However, there is still no algorithm found to be able to perform well in all cases.

4. Related Issues on Cluster Analysis

4.1 Determining the Number of Clusters

A problem common to all clustering techniques is the difficulty of deciding the number of clusters present in the data.

Aldenderfer and Blashfield (1984, p. 53) claim that the most important reasons that little progress has been made toward the solution of the problem are the lack of a suitable null hypothesis and the complex nature of multivariate sampling distributions. Much of the difficulty in creating a workable null hypothesis has been the lack of a consistent and comprehensive definition of the structure and content of a cluster.

Everitt (1980) states the main difficulties with deriving formal significance tests in this area appear to be the specification of a suitable null hypothesis, the determination of the sampling distribution of the distance or similarity measure used, and the development of a flexible test procedure.

In social science, two basic approaches to determining the number of clusters present have evolved: heuristic procedures and formal tests (Aldenderfer & Blashfield, 1984, p. 54). Heuristic procedures are by far the most commonly used methods. However, the procedures are hardly satisfactory because they are generally biased by the needs and opinions of the researcher as to the "correct" structure of the data. The development of formal statistical tests has not been appreciably slowed by the formidable problems of complex multivariate sampling distributions, but few tests have been widely accepted.

4.2 Comparing Clustering Methods

Since different clustering methods can produce different results when applied to the same data, it is important to explore some of the reasons this occurs.

According to Aldenderfer and Blashfield (1980, p. 59), most comparisons of clustering methods have been based upon the evaluation of how well different clustering

methods recover the structure of data sets with known structure. The results of all these studies are difficult to summarize, because each of them emphasized different combinations of data structure and methods tested. However, four factors appear to influence greatly the performance of clustering methods:

- Elements of cluster structure: cluster shape, cluster size and the number of cluster
- The presence of outliers and the degree of coverage required: if complete coverage of a classification is required.
- The degree of cluster overlap, and
- Choice of similarity measure.

4.3 Validation Techniques

Validation of the results from clustering methods is one of the most important issues in cluster analysis. In Aldenderfer and Blashfield's book (1984, p. 62-74), the authors have discussed five techniques for validating a cluster analysis solution.

1) Cophenetic correlation

The cophenetic correlation was first proposed by Sokal and Robill and is the major validation measure advocated by the numerical taxonomists. This measure is appropriate only when a hierarchical agglomerative method of clustering is used. The cophenetic correlation is used to determine how well the tree or dendrogram resulting from a hierarchical method actually represents the pattern of similarities or dissimilarities among the entities.

The basic idea is to create an implied similarity matrix. Each value in this matrix represents the similarity value at which the respective pair of entities was merged into a common cluster. The cophenetic correlation is the correlation between the values in the original similarity matrix and the values in the implied similarity matrix.

The method has some problems. First, the use of the product-moment correlation assumes normal distributions of the values in the two matrices being correlated. This assumption is generally violated for the values in the implied similarity matrix. Second, since the number of unique values in the implied similarity matrix (N-1) is much smaller than the number of unique values in the original similarity matrix

(N*(N-1)/2), the amount of information contained in the two matrices is quite different.

2) Significance tests on variables used to create clusters

A multivariate analysis of variance (MANOVA) of the variables used to generate the solution is performed in order to test for the significance of the clusters. In contrast to the use of the cophenetic correlation that attempts to analyze the accuracy of the hierarchical tree, the performance of standard significance tests is concerned with the quality of the cluster solution as a partition of the data set. Thus, the MANOVA technique can be used on solutions from any clustering technique that creates partitions.

"However, the use of discriminant analysis (or MANOVA or multiple ANOVA) in this fashion is inappropriate statistically (Aldenderfer & Blashfield, 1984, p. 64)."

3) Replication

This technique involves the estimation of the degree of replicability of a cluster solution across a series of data sets. It is a check for the internal consistency of a solution. To show that the same clusters appear across different subsets when the same clustering method is used is not strong evidence for the validity of a solution. In other words, the failure of a cluster solution to replicate is reason for rejecting the solution, but a successful replication does not guarantee the validity of the solution.

4) Significance tests on external variables

The procedures included in this category are probably among the better ways to validate a clustering solution, but, unfortunately, the approach has been little used despite its potential importance. Basically, the procedure is to perform significance tests that compare the cluster on variables not used to generate the cluster solution. The power of external validation is that it directly tests the generality of a cluster solution against relevant criteria. One reason this approach to validation is not used frequently in cluster analysis research is that the methodological design necessary to collect relevant criterion data is usually expensive. Another likely reason is that in many cases it is difficult to define a set of relevant external criteria because the necessary theory surrounding the classification process has not yet been refined sufficiently to determine what is truly relevant to the intended classification.

5) Monte Carlo Procedures

Basically, this approach is to use Monte Carlo procedure, using random number generators, to create a data set with general characteristics matching the overall characteristic of the original data but containing no clusters. The same clustering methods are used on both real data and the artificial data, and the resulting solutions are compared by appropriate methods.

Dubes and Jain (1979) provided a semi-tutorial review of the state-of-the-art in cluster validation, or the verification of results from clustering algorithms. The authors divided the validation into several steps.

- Data are first checked for clustering tendency. Four approaches are introduced to determine the randomness of the data, which include Fillenbaum and Repoport's test; Ling's test; Ling and Killough's (1976) index, and Strauss's method. Only if the data tend to be non-random is clustering attempted. The validation process judges the success of an algorithm in imposing a structure as well as the suitability of the structure for the data.
- The global fit of hierarchy is one test aspect in order to test whether a hierarchical clustering method provides a valid conceptualization of the data. Two indices for this purpose are reviewed which include Sneath and Sokal's cophenetic correlation coefficient (CPCC) and Hubert's Goodman-Kruskal γ-statistic.
- The global fit for partitions is another test aspect. Five specific studies of cluster validity for partitions are summarized, which include:
  1. Choosing the proper value of cluster number K;
  2. Hartigan's measure of splitting the data;
  3. Sneath's test for distinctiveness of clusters in terms of an overlap index and a disjunction index;
  4. A theoretical model based on minimizing a squared error criterion, and
  5. Mountfor's mathematical model.
- Validity of individual clusters is another step. This is done based on the isolation criterion in a hierarchy and the compactness criterion in a partition.

The authors conclude that "a user of clustering algorithms interested in cluster validity would be well advised at present to apply several clustering approaches and check for common clusters instead of searching for a technical measure of validity for an individual clustering (Dubes & Jain, p. 252)."

4.4 Errors in Clustering

In (Milligan, 1980), an evaluation of several clustering methods was conducted in order to detect errors in clustering. A classic error model has been used to perturb the interpoint distances with error in some manner. However, it seems desirable to consider other types of errors as well. For example, it seems likely that outliers would be present in empirical data sets, which do possess distinct clustering. Thus, outliers or intermediates between clusters could be considered a different form of error perturbation. Another form of error would occur when a researcher inadvertently includes in a cluster analysis a variable, which is irrelevant to the true cluster structure.

In this study, the author generates some ideal data first. Then eleven agglomerative hierarchical algorithms and four nonhierarchical centroid sorting procedures are tested with the data for six types of errors as follows.

1. Error-free parent data sets. It is found all methods produced mean recovery values about .90 except for MacQueen's algorithm.
2. Data sets with outliers. Some methods (e.g. complete link and Ward's) exhibited fairly noticeable decrements in cluster recovery. Some exhibited only slight decreases. The four nonhierarchical algorithms were virtually unaffected by the addition of outliers.
3. Error perturbation of the distance. Except single linkage, which was strongly affected, all the other methods were only slightly affected by the distance perturbation conditions.
4. Addition of random noise dimensions. The addition of one or two random noise dimensions produced fairly strong decrements in cluster recovery for all fifteen methods.
5. Computation of the distance with a noneuclidean index. This tended to produce only slight decrements.

6. Standardization of the variables. This tended to produce only slight decrements.

The results from step 5 and 6 show all of the fifteen algorithms appear to be robust with respect to the errors involving noneuclidean indexes and variable standardization. The overall result also shows the rank order performance of the methods is not consistent across the error conditions. This leads to the conclusion that a significant method by condition interaction would be obtained in an analysis of variance of the data. Further, the K-means algorithms performed rather poorly in all error conditions.

It is also interesting to note that the hierarchical methods which have been more commonly used in psychology, namely the single link, the complete link, and Ward's method, did not even place among the top four procedures. Further, three of the nonhierarchical procedures never fell into the superior group in any error condition. Thus, the K-means algorithms do not even seem to be very desirable if random starting seeds must be used (Milligan, 1980, p. 334).

4.5 The Generation of Test Clusters

In order to validate the results of a clustering algorithm, many researchers tend to approach the problem from a Monte Carlo framework. The advantage of this approach is that the researchers are able to use artificial data with known structure. The validation study is then reduced to the problem of determining how well the various clustering techniques or statistics detected the true (known) structure in the data (Milligan, 1985, p. 123).

Milligan (1985) has proposed an algorithm for generating artificial test clusters. There are 9 steps in this algorithm with a three-way factorial design. The three factors are the number of clusters, the number of dimensions, and the density level. The clusters, which are generated, are well-separated, (mildly) truncated multivariate normal mixtures. As such, the resulting structure could be considered to consist of "natural" clusters, which exhibit the property of external isolation and internal cohesion suggested by Cormack (1971). Three different verification activities have been conducted on the output of the algorithm. In addition, the advantage of this algorithm is the inclusion of numerous error perturbation routines.

Dubin (1996) has implemented and used Milligan's algorithm to generate test data in his dissertation. At the same time, the author has found "Milligan's algorithm cannot be predicted on the basis of independence, on the basis of their projections on the coordinate axes, nor on the basis of simple diagnostic measures like mean and standard deviation. The cluster structure in such spaces depends on a more subtle dependency between the variables (p. 58)."

4.6 Spatial vs. Tree Representation of Proximity Data

According to Pruzansky, Tversky and Carroll (1982), most representations of proximity data belong to one of the two families of models:

- Continuous spatial models. Objects are embedded in some coordinate space so that the metric distances between points represent the observed proximities between the respective objects. The most widely used spatial representation is the two-dimensional Euclidean space, or plane.
- Discrete network models. Each object is represented as a node in some graph so that the relations among the nodes in the graph reflect the proximities among objects. Three is the most common network model.

Pruzansky and her colleagues have studied these two kinds of models to develop guidelines for comparing these representations, and to discover properties that could help diagnose which representation is more appropriate for a given set of data. In their study, artificial data generated either by a plane or a tree are scaled using procedures for fitting either a plane or a tree. The results show that the appropriate model fits the data better than the inappropriate model for all noise levels and the two models are roughly comparable. In addition, two properties of the data proved to be useful in distinguishing between the models: the skewness of the distribution of distances and the proportion of elongated triangles, which measures departures from the ultrametric inequality.

In his dissertation, Dubin (1996) has studied the skewness and elongation of data in detail and employed these two properties to study the clustering tendency of data.

6. Conclusions

It is worth to noting that cluster analysis itself is exploratory in nature and a tool for discovery rather than an end in itself. Validation of the results is a necessary step. A few cautions about cluster analysis have been given by Aldenderfer and Blashfield (1984, p. 14-16), which include:

- Most cluster analysis methods are relatively simple procedures that, in most cases, are not supported by an extensive body of statistical reasoning;
- Cluster analysis methods have evolved from many disciplines and are inbred with the biases of these disciplines;
- Different clustering methods can and do generate different solutions to the same data set;
- The strategy of cluster analysis is structure-seeking although its operation is structure-imposing.

In addition, Aldenderfer and Blashfield (1984, p. 80) have also provided a guide to reporting cluster analysis studies, which includes:

- An unambiguous description of the clustering method should be provided;
- The choice of similarity measure (or statistical criterion if an iterative method is used) should be clearly stated;
- The computer program used should be stated;
- The procedures used to determine the number of cluster should be explained;
- Adequate evidence of the validity of the cluster analysis solution should be presented.

From nonhierarchical methods to a variety of hierarchical methods, cluster analysis has been employed to create groups of documents and terms in many IR studies. In addition to the implementation of all kinds of efficient algorithms, studies have also been carried out to:

- test the hypothesis, as in Jardine and Van Rijsbergen's study (1971);
- validate the results, as in Shaw's study (1986), and;
- update the cluster structure, as in Can and Ozkarhan's study (1989).

During the process of this literature review, the author has also tested some hierarchical clustering methods and the k-means method on big document collections. It is found most of the methods either assign the documents into a huge cluster or randomly assign the documents into the clusters pre-specified. One of the important reasons for this is because the input data matrix is very sparse. A conclusion of these experiments is that we need to combine the clustering methods among themselves or with some neural network algorithms to cluster big document collections.

Acknowledgements:

This report is finished under the instruction of my advisor Linda C. Smith. The student would like to thank her for her kindly help in the whole process. The student would also like to thank the other members in her advisory committee, Professor Bruce R. Schatz, P. Bryan Heidorn, David S. Dubin for providing the readings and useful suggestions.

Reference:

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster Analysis. Sage Publications, Inc.

Braam, R. R., Moed, H. F., & van Rann, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I. structure aspects, JASIS, 42(4): 233-251.

Can, F. & Ozkarahan, E. A. (1984). Two partitioning type clustering algorithms. JASIS. 35(5): 268-276.

Can, F. & Ozkarahan, E. A. (1989). Dynamic cluster maintenance. Information Processing & Management, 25(3): 275-291.

Cormack, R. M. (1971). A review of classification. Journal of the Royal Statistical Society (Series, A), 134, 321-367.

Croft, W. B. (1977). Clustering large files of documents using the single-link method. JASIS, 11, 341-344.

Crouch, C. J. (1988). A cluser-based approach to thesaurus construction. 11th International Conference on Research and Development in Information Retrieval. New York: ACM, 309-320.

Defays, D. (1977). An efficient algorithms for a complete link method. Computer Journal, 20:93-95.

Dubes, R., & Jain, A. K. (1979). Validity studies in clustering methodologies. Pattern Recognition, 11, 235-254.

Dubin, D. (1996). Structures in Document Browsing Spaces. Ph.D. dissertation. University of Pittsburgh.

El-Hamdouchi, A. & Willett, P. (1986). Hierarchic document clustering using Ward's method. Proceedings of the Ninth International Conference on Research and Development in Information Retrieval, 149-156.

Everitt, B. (1980). Cluster Analysis. 2nd ed. New York: Halsted Press.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice-Hall, Inc.

Jardine, N., & Van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. Information Storage and Retrieval, 7, 217-240.

Ling, R. F. (1972). On the theory and construction of k-clusters, Computer Journal, 15, 326-332.

Ling, R. F., & Killough, G. G. (1976). Probability tables for cluster analysis based on a theory of random graphs. Journal of the American Statistical Association, 71(345), 293-300.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45(3), 325-342.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. Psychometrika, 50(1), 123-127.

Pruzansky, S.; Tversky, A.; & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. Psychometrika, 47(1), 3-24.

Rasmussen, E. (1992). Clustering Algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, N. J.: Prentice Hall.

Salton, G. (ed.) (1971). The SMART Retrieval System. Englewood Cliffs, N.J.: Prentice Hall.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing, Communications of the ACM, 18(11), 613-620.

Shaw, W. M. (1986). An investigation of document partitions. Information Processing & Management, 22: 19-28.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single link cluster method. Computer Journal, 16: 30-34.

Sneath, P., & Sokal, R. (1973). Numerical Taxonomy: the Principles and Practice of Numerical Classification. San Francisco: W. F. Freeman.

Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation, 33: 106-119.

Van Rijsbergen, C. J.; Harper, D. J.; & Porter, M. F. (1981). The selection of good search terms. Information Processing & Management, 17: 77-91.

Voorhees, E. M. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. Information Processing & Management, 22:465-476.

Willett, P. (1983). Similarity coefficients and weighting functions for automatic document classification: an empirical comparison, International Classification, 10, 138-42.

Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. Information Processing & Management, 24(5), 577-597.