A Study on the Clustering Technique Associated with Statistical Term Relatedness in Information Retrieval

by JUN MIN JEONG*

At the present time, the role and importance of information retrieval has greatly increased for two main reasons: the coverage of the searchable collections is now extensive and collection size may exceed several million documents; furthermore, the search results can now be obtained more or less instantaneously using online procedures and computer terminal devices that provide interaction and communication between system and users. The large collection size make it plausible to the users that relevant information will in fact be retrieved as a result of a search operation, and the probability of obtaining the search output without delay creates a substantial user demand for the retrieval services.

INTRODUCTION

It is the objective of all information retrieval systems to optimize the results of the search process by retrieving the most relevant items and holding back the irrelevant ones.

* SCHOOL OF INFORMATION & LIBRARY SCIENCE, CASE WESTERN RESERVE UNIVERSITY

The effort to achieve the above mentioned goal in information retrieval is progressed over years. Examples of early approaches to the improvement of information retrieval system include: classification theory; question analysis; citation analysis; clustering technique (file organization); automatic indexing (term association); etc.

Specially, in the field of file organization, the research developed from the most primitive sequential file organization through inverted files and to the recent work in cluster file organization. The work in this particular area did not focus only on the speed and storage aspects of file organization, but also on bringing as closely as possible the semantically related items and storing them in separate clusters. This concept of clustering is based on the hypothesis that closely associated documents tend to be relevant to the same requests. [Ref. 31 and 32]

On the other hand, the automatic indexing has emphasized on the statistical terms association. Statistical terms association is based on the assumption that the co-occurrence of two terms in a given document may be interpreted as a small piece of probabilistic evidence that the two terms are semantically related. The terms with high co-occurrence with the request terms are then extracted from the collection and used as a feedback in a second expanded search. The main objective the technique as stated by Giuliano [Ref. 11] is to "free the requester from coughing his inquiry in precisely the same terms employed by the indexer."

The purpose of this paper is to propose the utilization of the concepts of clustering and statistical indexing as a unified method for information retrieval optimization.

One of the major drawbacks of traditional information re-

trieval systems is the requirement that the request terms should exactly match the document descriptors before the document can be judged relevant to the request. A document indexed under "Automatic Indexing", for example, will be missed if a searcher requested documents on "Statistical Indexing" even though the two concepts are semantically related. To remedy this situation, a number of statistical techniques have been suggested in the literature of information science. unfortunately, most, if not all, of these techniques were criticized of their inability to enhance the performance of information retrieval systems.

The failure of these techniques is due to the heterogeneousity of the collections used rather than the inefficiency of the association techniques. The use of statistical term association on such collections produces spurious association between semantically and conceptually independent terms. Such association in turn leads to false drops and negatively affect the effectiveness of the system.

STATISTICAL TERM RELATEDNESS

Since the advent of automated information retrieval systems in the late 1950's, the search for information retrieval optimization techniques has occupied the researchers in the fi field. In 1955, Taube has argued that the requester should be provided with all the uniterms that co-occur with the terms of his original request. Taube stated his argument as follows:

In organizing a body of information, we need to be able to recover materials indexed under any heading by searching under that heading, but also to trace all the ideas associated with the heading with which we start through the total system.

[Ref. 29]

Although Taube did not use quantitative measure involving the frequency of co-occurrence, his article might be considered the forerunner of later numerical techniques.

Another important early work in statistical term association techriques came from H. P. Luhn who, in 1958, suggested that the clerical ability of the computer be used to develop statistical frequency counts of text. These counts would then be used to determine "significant terms". [Ref. 18] This suggestion was not followed through until 1960 when Maron and Kuhns published their article on statistical term association as part of a more general methodological attack on the problem of document retrieval. Starting with a statistical matrix of association frequencies, they developed three different measures of closeness of association for index terms. One is the conditional probability that if a term in the original request, Tj, is assigned to a document, then the additional term Tk will be also assigned. The second measure is the inverse conditional probability, that is, the probability that if the additional term Tk is assigned to a document, then the origi nal request term Tj also would be assigned. [Ref. 19]

A classic paper on statistical terms association was written by Stiles in 1961. In that paper, Stiles introduced the idea of an all computer document retrieval system based on statistical term association. Stiles claimed that his system "can find documents related to a request even though they may not be indexed by the exact terms of the request ...". He later added that "our handicap has been we have had to select the precise terms that were originally used to index documents on various aspects of a particular subject and yet we must group for just the right set of terms". Stiles used a simi-

larity measure, called the association factor, to compute the degree of association between the original request terms and the index terms that frequently co-occur with them. The words with high degree of association with the request terms 'called first generation terms by Stiles' can then be used to expand the original request in a second broad search. The first generation terms can be used in turn to produce second generation terms, and so on. Stiles Claimed that this 'second generation terms' procedure was capable of catching synonymous terms, near synonymous, generics, specifics and other semantically related terms. [Ref. 27] Following the publication of Stiles' paper, a number of other experiments based on statistical term association were carried out. [Refs. 7,9,13 and 17] With the expectation of the work done by Curtice at LeHigh University in 1965, the results were rather desappointing. In his work, Curtice investigated two techniques of statistical association. The first, which is based on Doyle's work, produced an association map that has visually shown the degree of association between index terms, the other technique was based on the association factor introduced by Stiles. In his findings, Curtice stated that "both the association maps and request profiles as methods for representing associations among the terms in the vocabulary could be very useful." [Ref. 9]

In a doctoral dissertation carried out at Case Western Reserve University in 1966, Jean Tague [Ref. 28] studied the effectiveness of statistical term association as a technique for request expansion. She used three association coefficients in her comparative study and tested their efficiencies as methods of file searching. In the study, Tague distinguished between two kinds of term association, 'continuous association' in which the terms bore a real-world relationship

to each other, and 'synonymous associations', in which the terms bore a semantic relationship to each other. Tague summarized her findings in the following points:

1) There is no significant difference in effectiveness among the methods of search strategy formulation.
2) The original request terms had a significantly higher specificity than the other methods.
3) The original request terms had a significantly lower sensitivity than the other methods.

Tague concluded that most of the term associations produced in her study were due to chance than to any other factors. However she did not rule out the possibility that her negative results might be due to the small size of the sample collection used in the study.

In a similar study done at Harvard University in 1969, Lesk [Ref. 17] investigated the effectiveness of association techniques. He used the Smart Retrieval System Programs developed by Salton to carry out his study. The experimental results of Lesk's study indicated that:

1) In small collections, associations were not useful for determining work meanings or relationships, since the majority of the associated pairs depended on purely local meanings of the words and did not reflect their general meaning in technical text.
2) Associative retrieval that a properly constructed thesaurus offered better performance than statistical association methods.

Lesk concluded that a properly constructed thesaurus offered better performance than statistical association methods.

Harry M. Hersh and his colleagues at the Air Force Systems Command also experimented with associative retrieval sys-

tems. As most of their colleagues, their study revealed no significant differences between the statistical techniques and the other traditional approaches to information retrieval systems. They summarized their findings in the following, [Ref. 13]

The statistical aspects of the system were shown to be unnecessary in order to obtain a reasonable level of retrieval performance. That is, the statistical measures used in this study did not appear to add significantly to the retrieval operation.

They suggested that more conceptual and emperical work in the area of linguistic relationships would be necessary in order to develop more intelligent information retrieval systems.

It is obvious from the above discussion that most of the experiments conducted in this area produced negative results regarding the effectiveness of statistical term associations. One of the major problems that contributed to this low performance was the problem of false or spurious associations. In almost all of the experiments studied, the statistical techniques succeeded in producing or extracting the terms that frequently co-occurred with the request terms, but a high percentage of these terms bore no semantical or 'real world' relationships with the request terms.

CLUSTERING TECHNIQUE

Last statement of the above section will bring us to the area of clustering and its application in the field of information retrieval. Clustering techniques attempt to group points in a multidimensional space in such a way that all

points in a single group have a natural relation to one another and points out of same group art somehow different.

Due to the fact that cluster analysis is used in almost all fields of endeavor, the literature of this field is both voluminous and diverse. The terminology, however, differs from one field of knowledge to another. Anderberg successfully summarized this in his work on cluster analysis; "numerical taxonomy is frequent among biologists, botanists, and ecologists, while some social scientists may prefer 'typology'. In the fields of pattern recognition, cybernetics, and electrical engineering, the terms 'learning without teacher' and 'unsupervised learning' usually pertain to cluster analysis. In information retrieval and linguistics, 'clumping' refers to a particular kind of clustering. In geography and regional sciences, one may find the term 'regionalization'. Graph theorists and circuit designers prefer 'partition' as a term describing a collection of clusters". [Ref. 1]

In information retrieval the term clustering refers to the process of classifying a collection of documents into groups of related items. A measure of similarity is often to estimate the relationship between pairs of items that constitute the collection. A clustering algorithm is then used to enter into a common class, all items that are found to be sufficiently similar. As Van Rijsbergen stated, the underlying logic behind using clustering in information retrieval is the hypothesis that 'closely associated documents tend to be relevant to the same requests'.

Although most of the work done on clustering is still experimental, the results obtained are very promising. Salton summarizes the advantages and disadvantages of clustering as follows: [Ref. 22]

Advantages

Fact searches are possible, since only a few classes must be handled for each search.

Directed searches are possible by narrowing or broadening the search as required by the user.

Disadvantages

The clustering process is expensive when carried out for large files.

Storage overhead is added because the centroid file must be stored and maintained.

It is clear from the table that most of the disadvantages associated with clustering related to the cost of storing and maintaining the clustered files and their directories. However, due to the continuing decrease in the cost of computer storage, it is widely accepted that the future is for cluster-based information retrieval systems.

The clustering methods used today in information retrieval can be classified into two main categories, the hierarchic and non-hierarchic clustering methods. Most of the work on hierarchic clustering was done by Van Rijsbergen, jardine and Sibson. [Refs. 14 and 15]

Van Rijsbergen explains the whole process as 'a retrieval from a document collection which has been arranged into a hierarchic system of clusters, such that highly associated documents are nested within sets of less highly associated documents. After constructing the hierarchic system, the retrieval process proceeds by first matching clusters at the top of the hierarchy and chooses the best matching cluster. It then matches the request with the cluster immediately included in that cluster and so on, working down the hierarchy until

the optimal match is achieved. [Ref. 30] The documents in the cluster on which the optimal match is achieved are then retrieved.

The non-hierarchic method on the contrary, is a single-level clustering technique. Most of the work in this area was done by Salton and his co-workers in the Smart System project at Harvard and Cornell. In such a system, the retrieval process starts by clustering the collection into a number of single-level clusters. Then for each cluster, a cluster profile is constructed using the index terms that highly characterize the whole cluster. The search request is then matched against each cluster profile to determine the best matching cluster, and finally a serial search is done to determine the relevant set. The output of such a system is usually a ranking of the relevant documents so retrieved. [Refs. 23 and 25]

A number of other clustering methods which can be classified under the single-level clustering methods emerged in the early 1960's. The lack of mathematical models for classification of documents during that period has forced some of the researchers in information retrieval to resort to other fields of knowledge looking for suitable models. Baker, for example, experimented with 'latent class analysis' because of 'the similarity of using keywords in documents to classify the documents and that of classifying human subjects according to their responses to questionnaires'. Latent class analysis was originally developed by Lazarsfeld during World War II to provide a means for categorizing soldiers according to their attitude towards selected topics. Baker claimed that latent class analysis is superior to the other clustering techniques because it assigned documents to their classes on a weighted basis. In other words, instead of

stating that either a document belonged to a given category or not, latent class analysis stated that a document could belong to a category to a degree. [Refs. 2 and 3]

In an article written in 1962, Harold Borko introduced factor analysis as a document clustering technique. This technique has been used successfully by psychologists in their efforts to study the underlying variables of intelligence, personality, creativity, ability, ... etc. After conducting several experiments, Borko proved that factor analysis is a reliable and valied as the rest of the clustering method. [Refs. 4, 5 and 6]

The basic idea of his paper was derived from Maron's automatic indexing [Ref. 19]. Borko hypothesized that if it were possible to derive a set of categories which would provide a best possible description of a domain of documents the task of classification would be simple, and in fact could be done automatically. To test this hypothesis, he was

1) to construct an empirically based mathematically derived classification system,
2) to devise a set of procedure by which documents could be automatically classified into these categories, and
3) to determine the accuracy of the classification by comparison to a criteria.

Underlying the application of factor analysis to this study are the assumptions that documents can be classified on the basis of the terms they contain and that documents containing similar sets of terms belong to the same category.

In 1984, the kernel technique, which combined the strengths of non-hierarchic method (by Salton) and the factor analysis of which structure is hierarchic, is introduced. [Ref. 16]

A directed graph G = (V, A) or simply G is an ordered pair consisting of a nonempty set v = (vi,..., vj) of vertices and a set $A \subseteq V \times V$ of ordered pairs (vi, vj) called edges. The vertex u of an edge (u, v) is called the predecessor of v and v is called the successor of u. For any $x \in v$, $S \subseteq V$, the mapping Γ maps x and S into their adjacent vertex sets $\Gamma(x)$ and $\Gamma(S)$ defined by

$$\Gamma(x) = \{v : (xv) \subset s\}, \quad \Gamma(s) = \bigcup_{k \in s} \Gamma(k)$$

a set $k \subset V$ which satisfies the properties of internal and external stability expressed by 1) and 2) is called a kernel or nucleus of the graph G = (V, A).

$$\Gamma(u) \cap k = \phi, \forall u \in k \qquad 1)$$

$$\Gamma(u) \cap k = \phi, \forall u \in \bar{k} \cap v \qquad 2)$$

The vertex set K which satisfies 1) and 2) is also called stable (independent) and absorbant (domination) respectively. And not every graph has a kernel, and if a graph possesses a kernel, this kernel is not necessarily unique.

The characteristics of a kernel are defined in terms of independent and domination. In a connected graph, the elements in a kernel set should be independent of each other and these elements should dominate the other elements in the graph. That is, any two elements in a kernel may have no relationship directly, but they can be reached each other through the other elements which are outside the kernel set. Hence, the elements in a kernel can represent the total graph.

UNIFIED METHOD IN INFORMATION RETRIEVAL

It is clear from the above discussion that tremendous works have been done on both the areas of clustering and sta-

tistical terms association, however little if no effort at all was made to combine them in a unified information retrieval system. So, it can be hypothesized that the use of statistical terms association within clustered files will significantly enhance the performance of information retrieval systems. By clustering the file and bringing all the semantically related documents together, the problem of spurious association will be greatly reduced. In other words, the clustering technique is capable of creating a semantically appropriate environment for the application of statistical terms association.

In this paper, it is considered how the methodology of unified technique can be generated. First, the random sample as a test data is selected which consists of citations and their abstracts. From each abstracts, the significant words are counted and stored. Before to do this, all terms should be stemmed and normalized with the same root but differing suffix.

After this frequency count and syntactical analysis, a Term/Document Matrix [Fig 1] is constructed.

	T1	T2	T3	T4	T5	Tn
D1	0	1	0	1	1	0
D2	1	0	0	1	1	0
D3	0	0	1	1	0	1
D4	1	1	1	0	0	1
D5	1	0	0	0	1	1
.						
.						
Dn	0	0	0	1	1	1

Figure 1. A Term/Document Matrix

The columns of Figure 1 represent the index terms that best describe the documents shown in the rows of the matrix. The matrix is a binary one since only zeros and ones are used to

represent the presence or absence of a given term in a given document.

A similarity coefficient for each pair of documents using the Term/Document Matrix as an input is computed. The similarity measure is based on the following correlation coefficient,

$$Rij = \frac{N(wj.wj)}{N(wi) + N(wj) - N(wi.wj)}$$

where in a binary matrix, the numerator represents the frequency of co-occurrence of matching non-zero properties and the denominator represents the number of destinct non-zero properties in each document vector. The function was originally suggested by Doyle [Ref. 10] and is used extensively in information retrieval experiments.

After computing the similarity coefficient for each pair of documents, a Document/Document Matrix is built [Fig 2].

	D1	D2	D3	D4	Dn
D1		.59	.70	.31	.69
D2	.40		.60	.67	.83
D3	.39	.19		.29	.59
D4	.51	.41	.23		.36
.					
.					
.					
Dn	.73	.41	.36	.17	

Figure 2. A Document/Document Matrix

The Cells of the matrix represent the degree of similarity between each pair of documents rather than a binary state of zeros and ones.

The modified kernel technique is used for clustering. Modified kernel is rather non-hierarchic structure than hierarchic. Each cluster can be generated by isolating the elements in the kernel of total documents set. After that, cluster profiles which represent the property of each cluster are identified by the general kernel technique, for the purpose of matching that cluster against the incoming requests. The cluster profile can be a set of document or can be a set of terms in documents of kernel set. In this study, the later case is chosen. Request is matched against the cluster profile that best describes the members of the cluster. The cluster that best matches the request vector can be the best candidate for searching.

For the purpose of comparing the effectiveness of statistical terms association as a method of search optimization, a search on both clustered and sequential files can be done. First, the original request terms are used in a serial and clustered searches and secondly, the recombined request terms which were expanded with the terms that frequently co-occur with them in the file. The expanded search strategies can be used in a clustered and a serial search referred to as the association factor, will be measured using the following formula suggested by Salton [Ref. 26].

$$Vij = \frac{N(wi/wj)}{N(wi)\ N(wj)}$$

where, N (wi.wj) is the number of documents assigned both terms wi and wj

N(wi) is the number of documents assigned term wi

N(wj) is the number of documents assigned term wj

After computing the association factor for each original index term, a threshold value is used to determine the significant level of association.

The data generated from the search procedures described above is analyzed using a method developed at the Western reserve University Center for Documentation and Communication Research. [Ref. 12] In that method, the performance of a retrieval system is defined as a function of the time and efficiency, where efficiency is some cost function of time and effectiveness, a function of two variables, sensitivity and specificity. Sensitivity is the ability of the system to provide the user with the relevant documents he desires, and specificity is the ability not to provide him with the non-relevant documents. Using Tague's notation [Ref. 28], sensitivity (Se), specificity(Sp) and effectiveness(U) are defined by the following formulas,

$$Sp = \frac{N(A' \cap I')}{N(I')}$$

$$Se = \frac{N(A \cap I)}{N(I)}$$

$$U = (Se + Sp) - 1$$

DISCUSSION

In this paper, the literature of statistical terms association and clustering technique is briefly reviewed. From the previous works, it is revealed that the statistical terms association is not sufficient for searching and classification due to the heterogeneousity of the collections rather than the

inefficiency of the association techniques. To support this problem, the clustering technique is introduced and it is proposed that the use of statistical terms association within clustered files will significantly enhance the performance of information retrieval systems.

The unified technique is theoretically introduced and how performed it will be is described. Stiles' association factor is adapted for the automatic terms association and the kernel technique is supported for its cluster environment. But this approach is theoretical only not involves experiment and evaluation. For further work, it is suggested to test this proposed method and evaluate it with its performace.

REFERENCES

[1] Anderberg, M. R. Cluster Analysis for Applications (New York: Academic, 1973).

[2] Baker, F. B. "Information Retrieval Based Upon Latent Class Analysis." J of the ACM 9(1962): 512-521.

[3] Baker, F. B. "Latent Class Analysis as an Association Model for Information Retrieval." J of the ACM 11(1963): 816-824.

[4] Borko, H. "The Construction of an Empirically Based Mathematically Derived Classification System." Proc. Spring Joint Computer Conference 21(1962): 279-289.

[5] Borko. H. "Studies on the Reliability and Validity of Factor - Analytically Derived Classification Categories." Statistical Association Methods for Mechanized Documentation -- Symposium Proceedings, 1965.

[6] Borko, H. and M. D. Bernic. "Automatic Document Classification." J or the ACM 10(1963): 151-162.

[7] Borko, H., Donald A. Blankenship and Robert C. Burket. "On-line Information Retrieval Using Associative Indexing." Technical Report #RADC-TR-68-100, (New York: Air Force Systems Command, 1968).

[8] Can, F. and E. A. Ozkarahan. "Similarity and Stability Analysis of the Two Partitioning Type Clustering Algorithms." JASIS 36(1985): 3-14.

[9] Curtice, R. M. and Victor Rosenberg. "Optimizing Retrieval Results with Man-machine Interaction." Technical Report (Center for the Information Science, LeHigh University, 1965).

[10] Doyle, L. B. "Indexing and Abstracting in Association." American Documentation 13(1962): 378-390.

[11] Giuliano, V. E. and P. E. Jones. Linear Associative Information Retrieval, Vistas in Information Handling. (Washington, D.C.: Spartan Books, 1963).

[12] Goffman, William and Vaun A. Newill. "The Complex Problem of Medical Information Retrieval, 1. A Methodology for a Comparative Systems Laboratory." (Cleveland, OH: WRU Center for Documentation and Communication Research, 1964): 10-15.

[13] Hersh, H. M. "Statistical Methods for Technical Documentation Retrieval." Technical Report # RADC-TR-77-217 (New York: Air Force Systems Command, 1979).

[14] Jardine, N. and R. Sibson. "The Construction of Hierarchic and Non-hierarchic Classifications." The Computer J. 11(1968): 177-184.

[15] Jardine, N. and C. J. Van Rijsbergen. "The Use of Hierarchic Clustering in Information Retrieval." Info. Stor. and Ret. 7(1971): 217-240.

[16] Jeong, Jun Min. The Ecology of the Scientific Literature and Information Retrieval (Ph. D. Dissertation. Case Western Reserve University, 1985).

[17] Lesk, M. E. "Word-word Associations in Document Retrieval Systems." American Documentation 16(1969): 27-37.

[18] Luhn, H. P. "Auto-encoding of Documents for Information Retrieval Systems." Symposium on Documentation (University of Southern California, 1958).

[19] Maron, M. E. and J. L. Kuhns. "On Relevance, Probabilistic Indexing and Information Retrieval." J of the ACM 7(1960): 216-244.

[20] Murray, D. M. Document Retrieval Based on Clustered Files. (Ph. D. Dissertation, Cornell University, 1972).

[21] Rocchio, J. J. Document Retrieval Systems Optimization and Evaluation. (Ph. D. Dissertation, Harvard University, 1966).

[22] Salton, G. Automatic Information Organization and Control. (New York: McGraw-Hill, 1968).

[23] Salton, G. Dynamic Information and Library Processing. (Englewood Cliffs, NJ: Prentice-Hall, 1975).

[24] Salton, G. Introduction to Modern Information Retrieval. (New York: McGraw-Hill, 1983).

[25] Salton, G. The Smart Retrieval System: Experiments in Automatic Document Processing. (Englewood Cliffs, NJ: Prentice-Hall, 1971).

[26] Salton, G. "Some Hierarchical Models for Automatic Document Retrieval." American Documentation 10(1963): 213-222.

[27] Stiles, H. E. "The Association Factor in Information Retrieval." J of the ACM 8(1961): 271-279.

[28] Tague, J. M. Statistical Measures of Term Association in Information Retrieval. (Ph. D. Dissertation, Case Western Reserve University, 1966).

[29] Taube, Mortimer. "Storage and Retrieval of Information by Means of the Association of Ideas." American Documentation 6(1955): 1-18.

[30] Van Rijsbergen, C. J. "Further Experiments with Hierarchic Clustering in Document Retrieval." Info. Stor. and Ret. 8(1973): 319-327.

[31] Van Rijsbergen, C. J. Automatic Information Structuring and Retrieval. (Rh. D. Dissertation), University of Cambridge, 1972).

[32] Van Rijsbergen, C. J. Information Retrieval, 2nd ed. (London: Butterworth, 1979).

〈國 文 要 約〉

情報檢索에 있어서 用語의 統計的 關聯性을 應用한 클러스터링技法

本 論文에서는 統計的 用語組合과 클러스터링技法에 관한 文獻을 간단히 살펴보았다. 先行硏究들로부터 統計的 用語組合은 組合技法의 非效率性때문이 아니라 文獻集團의 異質性때문에 檢索과 分類에 적당치 않다는 事實을 發見할 수 있다. 그 結果로부터 情報檢索의 最適化를 위한 組合技法으로서 클러스터링과 統計的 索引의 槪念을 利用할 수 있다.

本 論文의 假說은 클러스터파일內에서 統計的 用語組合을 使用함으로써 情報檢索시스템의 性能을 상당히 向上시킬 수 있다는 것이다. 달리말해서, 파일들을 모으고 意味的으로 關聯있는 모든 文獻들을 함께 모아줌으로써, 類似組合(spurious association)의 問題를 상당히 解決할 수 있을 것이다.

實際的으로, 本 論文에서는 組合技法의 方法論을 어떻게 生成할 수 있을 것인가를 고려했다. 自動用語 相關性을 위하여 스타일(stiles)의 組合因子를 利用했으며 클러스터링 環境을 위해 커널技法(kernel method)을 使用했다.