Automatic Structuring and Retrieval of Large Text Files

Gerard Salton*
J. Allan
C. Buckley

TR 92-1286
June 1992

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*Department of Computer Science, Cornell University, Ithaca, NY 14853-7501. This study was supported in part by the National Science Foundation under grant IRI 89-15847.

Automatic Structuring and Retrieval of Large Text Files

Gerard Salton[*], J. Allan, and C. Buckley

Abstract

In many operational environments, large text files must be processed covering a wide variety of different topic areas. Aids must then be provided to the user that permit collection browsing and make it possible to locate particular items on demand. The conventional text analysis methods based on preconstructed knowledge-bases and other vocabulary-control tools are difficult to apply when the subject coverage is unrestricted.

An alternative approach, applicable to text collections in any subject area, is introduced which uses the document collections themselves as a basis for the text analysis, together with sophisticated text matching operations carried out at several levels of detail. Methods are described for relating semantically similar pieces of text, and for using the resulting hypertext structures for collection browsing and information retrieval.

[*]Department of Computer Science, Cornell University, Ithaca, NY 14853-7501. This study was supported in part by the National Science Foundation under grant IRI 89-15847.

1 Introduction

In many practical situations it is necessary to deal with large heterogeneous collections of text. This is the case notably for newspaper files, message collections, dictionaries and encyclopedia, textbook materials, and generally in many library environments. In such situations the subject matter varies widely, often covering large, open-ended slices of knowledge. Normally, selective access is needed to particular items on demand, and file access may be considerably simplified when browsing capabilities are made available that allow a flexible traversal of the text structure.

The following main technical problems must then be considered: how to analyze the text materials and generate representations of text content usable for search and retrieval; how to structure the files in such a way that related texts, or text excerpts, are properly identified; and how to design search strategies capable of finding relevant materials in response to a variety of user needs.

The conventional wisdom is that the keyword-type systems, where the information items are represented by sets of manually or automatically chosen index terms, have run their course: most keywords are believed to be ambiguous and often poorly related to actual text content; furthermore, document content is in any case poorly represented by small collections of individual terms. Accordingly, sophisticated so-called conceptual, text representations are often proposed, including the use of thesauruses (synonym dictionaries) tailored to particular subject areas, and of preconstructed knowledge representations that classify the main entities of interest in a given subject area, and specify the relationships that may hold between the entities in particular areas of application. The contention is that a detailed, deep semantic analysis of available texts will make it possible to relate and link similar texts to each other, and detect relevant texts in answer to available user queries.[1-3]

Unfortunately many conceptual and practical problems arise when deep language analysis systems are considered in unrestricted environments with arbitrary subject matter. Viable methods still do not exist for building large thesauruses automatically, or even manually, and the use of large knowledge bases is hampered by the fact that it is unclear what knowledge is actually needed for particular applications, how to represent the needed knowledge, how to isolate individual pieces of knowledge from an apparently unlimited context, and how effectively to match the content of any existing knowledge base with the available text collections.

In consequence, an unfortunate situation exists in the text processing area. Certain objective text characterizations and text relationships are manageable by fully automatic methods – for example, the recognition of the usual hierarchical text structure, including the subdivision of many texts into sections, paragraphs, sentences, phrases, and so on; and the automatic linking of text excerpts to any footnotes, annotations, and bibliographic references that may be provided. On the other hand, effective, fully automatic methods are hard to come by when the text operations specifically relate to text content. Even human subject experts have difficulty operating successfully in open-ended text environments.[4,5]

Fortunately, an avenue exists for dealing with large classes of heterogeneous text which goes much beyond the standard automatic keyword analysis, but also avoids the use of vast knowledge-representation tools and the assistance of human expertise in the basic text processing tasks. That method takes advantage of the fact that direct access is now available to very large text databases that constitute vast stores of knowledge. A corpus-based text handling approach may then be considered that uses the available text collections themselves to obtain the information needed for text analysis and text characterization.

Specifically, the available texts can be processed automatically and the text structure can be examined in detail, including the words and expressions included in the texts, the contexts in which the text elements are used, the decomposition of the text into sections, paragraphs and sentences, the structure of the various text components, and so on. By judiciously combining information obtained from the available texts, reasonable text characterizations may then be generated. Such a corpus-based text analysis approach is reminiscent of the so-called memory-based reasoning systems, where problems are solved by comparing problem statements with available databases, finding the most similar situation described by the stored data, and using the corresponding information to handle the new problem.[6]

In this study, viable methods are proposed for handling large text databases in arbitrary subject areas. Procedures are described for analyzing text content, structuring the text by linking text excerpts covering related subject matter, and retrieving text items in response to available user queries. For experimental purposes, the 25,000 articles included in the 29 volumes of the Funk and Wagnalls encyclopedia (65 Mb of text) are used as a database.[7]

2 Standard Vector Processing Model

In conventional information retrieval environments, keywords (terms) are manually or automatically assigned to the information items, and queries are formulated by using terms interconnected by Boolean operators. The documents retrieved in response to a query such as "A *and* B" will contain both term A and term B, whereas "A *or* B" locates items containing either one or both of the terms. Although widely used, the Boolean retrieval model is not ideally suited to the information retrieval task. Most obviously, ordinary users find it hard to generate useful Boolean queries that will retrieve just the right type and amount of information. In addition, the retrieved items are presented to the users in a random order that does not correspond to any presumed order of relevance or usefulness. This makes it difficult for the user to construct improved query formulations because the most useful information may not be presented to the user at all. Term weights reflecting term importance should be used in retrieval, but such weights are awkward to incorporate into Boolean systems in a consistent way. Finally, the operations of Boolean logic are unusually inflexible in information retrieval environments: for example, in response to *and*-queries such as "A *and* B *and* ... Z", a document containing all but one of the query terms is treated just as badly as a document containing no query term at all, because neither item would ever be brought to the users' attention.[8,9]

The vector-processing model represents an alternative possibility for handling information retrieval operations. In that case, both the stored documents as well as the search requests are represented by sets of terms (term vectors) without Boolean operators. Different vectors can be compared with each other, and a vector similarity coefficient is obtainable reflecting similarity in the term assignment for different vectors. In the vector processing model of retrieval, parallel operations can be used for collection structuring (by comparing pairs of document vectors with each other and identifying documents pairs found to be sufficiently similar), and for information retrieval (by comparing query vectors with the vectors representing the stored items and retrieving items found to be similar to the queries). When similarity measurements are performed between a query vector and the stored document vectors, the output can be ranked in decreasing order of the computed query similarity. This makes it possible to retrieve the most important items (those most similar to the user queries) first. Furthermore, term weights are easily accommodated because vectors of weighted terms are manipulated as easily as binary term vectors where weights are restricted to 1 for assigned terms and 0 for missing terms.[10,11]

The two most crucial tasks in a vector processing environment are the indexing operation used to construct the term vectors for the documents and texts under consideration, and the vector similarity computations used to identify groups of similar texts and select items for retrieval in response to available information queries.

In the Smart system environment [11], the following standard automatic indexing procedure is used for the term vector construction starting with the text of the individual documents:[12,13]

a) The text words are recognized, and certain common function words (such as "and", "or", "but", etc.) are eliminated by consulting a short list of "stop words".

b) The remaining words are reduced to word stem form by suffix removal and/or truncation. This reduces words such as "analysis", "analyzer", "analyzing", etc., to a common form such as "analy".

c) Optionally, term co-occurrence criteria are used to construct term phrases for sets of words that tend to co-occur frequently in the texts under consideration.

d) Term weights are assigned to the remaining terms (word stems and/or phrase stems). In particular, a term weight w_{ik} is assigned to each term T_k occurring in document (or text) D_i, and the text is represented by a term vector of the form $D_i = (w_{i1}, w_{i2}, ..., w_{it})$, where t terms in all are assumed to be available in the system. A zero weight is used for terms absent from a document, and positive weights characterize the terms actually assigned in a given case.

The assignment of useful term weights capable of distinguishing the important terms from the less important ones is crucial to the success of the automatic indexing process. A high performance term weighting system assigns large term weights to terms that occur frequently in particular documents, but rarely on the outside, because such terms are able to distinguish the items in which they occur from the remainder of the collection. A typical term weight of this type, known as a $tf \times idf$ weight (term frequency times inverse document frequency) may be defined as

$$w_{ik} = \frac{tf_{ik} \cdot log(N/n_k)}{\sqrt{\sum_{j=1}^{t}(tf_{ij})^2 \cdot (log(N/n_j))^2}} \tag{1}$$

where tf_{ik} is the frequency of occurrence of term T_k in D_i, $tf_{ik} = 0$ for terms not assigned to D_i, N is the size of the document collection, and n_k represents the number of documents in the collection with term T_k. The summation in the denominator, taken over all terms in a particular vector, is used for length normalization purposes to insure that all documents have equal chance of being retrieved. (Without length normalization, the longer documents with more assigned terms and higher term frequencies would generate higher document similarities, and exhibit higher retrieval potential than the shorter items.)[14]

After term vectors are available for all documents and information requests, global vector comparisons can be used to obtain similarity measurements, $sim(D_i, Q_j)$, for each document D_i and each query Q_j. Assuming that D_i is represented by vector $(w_{i1}, w_{i2}, ..., w_{it})$, and Q_j by vector $(w_{j1}, w_{j2}, ..., w_{jt})$, the inner product function $sim(D_i, Q_j) = \sum_{k=1}^{t} w_{ik} \cdot w_{jk}$ computes a similarity measure between 0 and 1 when the normalized term weights of expression (1) are used in the term vectors. The size of the vector similarity will then depend on the proportion and the weight of matching terms in the vectors.

For retrieval or text linking purposes, a threshold value must be chosen in the vector similarity to distinguish the potentially interesting items from those that can be rejected. Items whose vector similarity falls below the stated threshold are then considered to be insufficiently related to warrant further attention. On the other hand, when the global vector similarity between two documents, or between a query and a stored document, is sufficiently large, a presumption of item similarity exists. It is possible in these circumstances that the corresponding texts are indeed semantically related. Alternatively, the matching vocabulary may be highly ambiguous, and no definite conclusions may be warranted concerning any semantic text relationship. For example, ambiguous terms such as "base", "reach", "hit", and "strike" might occur in documents dealing with baseball, as well as in texts about military maneuvers. For this reason, global text comparisons are best used as initial filters leading to the rejection of items that are clearly dissimilar. When the global similarity between items is sufficiently large, additional processing steps may be needed to insure that the common vocabulary is actually indicative of semantic relatedness. A local context check is proposed for this purpose.

3 Local Context Processing

A) "Use Theory" of Meaning

It has been claimed that two different aspects of text meaning are important in considering the content of text expressions and utterances. There is first the purely semantic aspect that deals with an intrinsic meaning of words and expressions taken out of context. The linguistic environment in which the words are used is disregarded in considering the intrinsic semantics. In addition, the background information will affect text meaning including the manner and place in which expressions are used, and the purpose served by using certain words in particular environments. This latter aspect of text meaning is known as the pragmatic aspect of meaning.[15-17]

Questions have been raised about the practical importance of an intrinsic notion of word environments. For example, Searle has noted that there is no common notion of "cutting" in expressions such as "he cut the grass", "he cut the meat", "he cut a class", "cut it out", etc.[17] Accordingly, the pragmatic aspect of language understanding and the context and purpose of linguistic statements become all important. A number of influential philosophers of language have enunciated what might be called a "use theory" of meaning which states that text meaning is effectively determined by text use.[18-20] The following quote from Wittgenstein is indicative in this respect:[19]

> "For a large class of cases – though not for all – in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language."

A good case can be made for claiming that in information retrieval the major concern should be with word use rather than with intrinsic word meaning. In fact, in retrieval it may not matter whether the word "base", for example, refers to a lamp base or an army base in a particular instance. What does matter is to insure that the meaning of words such as "base" is the same (whatever that meaning might be) in the queries and in the corresponding retrieved documents. If the meaning of different occurrences of words and expressions is reasonably similar, then it is legitimate to determine text similarity by using vector comparison method of the type previously described. On the other hand, if the meaning of the common vocabulary for different texts is distinct, then a text comparison will provide an erroneous indication of text similarity.

If meaning is largely determined by word use, it becomes necessary to determine how words are used in particular instances. In practice, it may not always be possible to ascertain precisely the purposes and linguistic actions that control word use. It is, however, normally possible to study the linguistic contexts in which the words occur in substantial detail, and to argue that when words and expressions appear in local contexts that are substantially similar, the word meanings are also the same. This suggests the following refined method of text comparison:[21-22]

a) A global text similarity is computed first by comparing the respective text vectors as previously explained.

b) Text pairs without sufficient global similarity are discarded and not subjected to further processing.

c) The local text environments are next considered for texts with sufficient global similarity, by comparing individual text sections, paragraphs, and sentences.

d) When text pairs are found that are globally similar and also contain a sufficient number of locally similar substructures, then the conclusion is reached that the texts are indeed related.

An interactive text comparison process which first uses large text segments, such as complete articles, and then considers smaller text units such as sections, paragraphs and sentences should be expected to produce a high degree of retrieval precision in the sense that very few semantically dissimilar texts will pass the multiple vocabulary filter. In particular, when different texts contain similar vocabulary patterns used in similar local contexts, then retrieval of one of the texts appears in order when the other is submitted as a query. Alternatively, the two texts may be treated as related by placing an indicator of semantic relationship (a text link) between them.

The recall question is more difficult to settle. In principle, different texts may exist that cover substantially related subject matter in completely different terms. In that case, a vocabulary comparison method is not adequate for text identification. This leads to the rejection, or nonretrieval, of apparently relevant texts. In practice, such situations may be rare, and the recall performance of the multiple text comparison method should also be relatively high, because ultimately it is near impossible to describe similar topics without using substantially overlapping vocabularies. Such an overlap is then detectable by properly designed text matching methods. A particular known mishap may be discussed by saying "railway" or "railroad" instead of "train", and "crash" or "derailment" or "malfunction" instead of "accident". But other components of the story will be directly comparable when a particular, common incident is covered.[2] Swanson has made extensive studies of document classes covering related subject matter that are formally disconnected, in the sense that no common items are included in the bibliographies attached to the different classes of items. Nevertheless, in all such cases, substantial overlap exists in the vocabulary patterns used in the different classes.[23]

B) Local Context Operations

The methods described earlier for the computation of the global text similarities can in principle be used also to obtain the local similarities in restricted contexts such as paragraphs or sentences. That is, term vectors can be generated for the local text environments, and a vector similarity measure can again be computed between vector pairs representing local text constructs. The term weights used in the local environments will not, however, be identical with those used earlier for the full text comparisons. In particular, the term frequency factors now refer to the local occurrence frequencies of the terms within particular paragraphs or sentences, and the number of text paragraphs, or text sentences must now be used to obtain the document frequency factors rather than the number of full documents.

For short text segments such as sentences and short paragraphs it is also preferable to use unnormalized term weights – that is, the $tf \times idf$ term weight formula of expression (1) is used without the normalization factor in the denominator. When unnormalized term weights are used, the text similarity depends on the number (rather than the proportion) of matching terms. This implies that the similarities between very short sentences which often prove to be meaningless ("consider the following figure", "see the example in figure x", "consider figure x", etc.) will be disregarded. Such fragments would be treated as highly similar in a matching system using normalized term weights.

In the examples used in this study, normalized term weights serve for the comparison of full texts and text sections. The similarity measurements then depend on the proportion of matching terms, and the similarity measurements lie between 0 and 1. Unnormalized weights are used for the local text comparisons between text paragraphs and text sentences. The text similarities then depend on the number and weight of matching terms and the similarity measurements are normally much larger than 1.

A typical example of the global/local text matching operations is shown in Tables 1-4. A search can be started by using a natural-language statement – for example, "I am interested in information about President John F. Kennedy". Alternatively, an existing text describing the subject matter of interest can be used as an initial search statement. In the illustration of Table 1, the full text of encyclopedia article 12941 entitled "Kennedy, John Fitzgerald" is used as a search request. The table shows the top 15 retrieved encyclopedia articles in decreasing order of similarity with the query, including article numbers, similarity coefficients with the query, and article titles. It is not surprising that the top retrieved article in response to query article 12941 (John F. Kennedy) is again article 12941 with a perfect query similarity of 1.00. The next few retrieved items (12943 "Robert F. Kennedy", 12942 "Joseph P. Kennedy") also exhibit high similarities above 0.50 with the query.

Only a single global vector similarity computation for pairs of encyclopedia articles was used to obtain the output of Table 1. The ranked output list makes it clear that most of the retrieved items are closely related to the query subject "John F. Kennedy". Substantial success in relating similar subject matter is thus achievable even by unmodified global text matching methods. The same effect was noted also for even simpler vocabulary matching systems that operate on raw texts without any term weighting or document indexing steps.[24-25]

One serious error must be noted in the output of Table 1: the inclusion of document 12939 "Anthony M. Kennedy", relating to the current Supreme Court Justice. In addition, some possibly marginal items, such as 4341 "Cape Canaveral", are also listed. Table 2 contains the common terms (word stems) present in documents 12939 and 12941. Each term is characterized by a uniquely chosen "concept number", as well as the term weight in vector 1 (document 12941) and vector 2 (document 12939), and finally a combined weight computed as the product of the weights in the two individual vectors.

Although the two Kennedys are of course unrelated, the output of Table 2 reveals a substantial congruence in the background of the two persons: they both had ties to Harvard University, they both had associations with U.S. Presidents (JFK was President, and AMK was appointed to the Court by President Reagan), and of course they both carry the name Kennedy. Thus, a global vector match is not capable of distinguishing the two articles.

The situation changes when the context of the common vocabulary is examined. Table 3 shows typical contexts for the term "Kennedy" taken from the text of the two documents. As expected, these contexts have little in common: AMK is described as conservative and associated with President Reagan; JFK obviously would be characterized very differently with associations to Cuba, Berlin, and Khrushchev. A new search for query document 12941 (JFK) now requiring an additional local context similarity (at least one matching pair of sentences between the query and any retrieved document, with a computed vector similarity not smaller than 75.0) will reject document 12939 together with a few other possibly marginal items. The corresponding restricted search results are shown in the output of Table 4.

Concretely, the local context check is carried out by comparing all pairs of sentences from the candidate document pairs, arranging the results in decreasing order of the computed sentence similarity, and rejecting all items whose maximum sentence similarity with the query document falls below the stated threshold. The output of Table 4 demonstrates that clearly relevant items such as Robert F. Kennedy (JFK's brother), or Joseph P. Kennedy (JFK's father), do in fact pass the required context check without difficulty.

4 Effect of Local Context Operations

A comparison of the unrestricted search results based on global term vector comparison only, with the restricted searches utilizing additional local context matches reveals the following important properties: First, the rejection of very short documents that are often nonspecific and usually unwanted. For the encyclopedia environment, this means that cross-reference articles that list a title with an additional reference ("Athenian League", see "Delian League") are rejected because such items do not include matchable sentences or paragraphs. This effect is shown in the output of Table 5 for query article 6944 "Delian League" (the Delian League was a Federation of Greek city-states founded around 500 BC as a safeguard against aggression from the Persians). The restricted search output requiring one matching sentence pair shows that the cross-reference articles 1653 "Athenian League" and 17416 "Pallas Athena" are now rejected.

A further, important property of the local context match is the rejection of obviously nonrelevant materials. This effect, already described in the JFK-AMK example, is illustrated in Table 5(b) by the rejection of articles 13673 "League of Nations", and 23373 "Urban League". The League of Nations, created after World War I, had much in common with the Delian League (an association of states whose aim was to prevent war). However most experts would not consider the association between the two items sufficiently close to warrant retrieval. The Urban League is designed to help inner-city minorities in the United States. Such an item would not be considered relevant in response to a query about the Delian League.

The items rejected by the restricted search of Table 5 are characterized in Table 5(c), and the newly retrieved items replacing them are shown in Table 5(d). The tables show that the rejected items are replaced by documents closely related to Greece, and especially to Athens, the leading member of the Delian League.

Another noticeable effect of the restricted search strategy is the replacement of texts with a somewhat general, or global query relationship by texts with a narrower more specific relation with the query. This effect is illustrated by the search output of Table 6. In this case, the search covers text sections rather than full documents (a section is defined as a text segment appearing between two adjacent subheadings together with the respective section titles). The search query of Table 6 is section 76585, corresponding to document 10300.13 (section 13 of document 10300 "Greece"). The section titles are Greece/Population/Political Divisions ("Greece", main section "Population", subsection "Political Divisions"). The unrestricted search output is shown in Table 6(a) and the items rejected in a restricted search using local sentence comparisons are included in Table 6(b).

Table 6(b) shows that once again a number of cross-references are eliminated, as are some extraneous items (section 119642 "Municipal Government / Weak-Mayor-Council Plan"). In addition, section 91581 "Italy/Population/Political Divisions" retrieved in rank 3 of the original search is also eliminated by the restricted search strategy. This section might be related to "Greece/Population/Political Divisions" because both Italy and Greece had divided and unstable governments during long stretches of the 20th century. However, the association may be considered somewhat remote and unspecific. Table 6(c) shows that the rejected items are replaced by new texts more closely related to the Greek situation.

Overall, the local search strategy proves beneficial because large numbers of nonrelevant or otherwise questionable items are removable from the output produced by the unrestricted text comparisons.

Some instances do exist, however, where the local context check will not eliminate unwanted items. The first problem is due to occurrences of standard vocabulary terms that are often closely related to text content but may also carry a number of different meanings. Terms such as "river", "valley", "mountain", "wheel", etc., may be quite subject-specific and hence highly weighted in the documents in which they occur. Matches of such terms sometimes produce sufficiently high sentence similarities to lead to the retrieval of the corresponding documents. When the interpretation of such terms differs in the two matching texts, retrieval errors may occur. The sample sentences at the top of Table 7 are taken from documents about Gemstones and Airplanes, respectively. The term "wheel" is prominent in both sentences, but the meanings differ, referring to grinding wheels and airplane wheels, respectively. The sentence similarity will exceed the standard threshold of 75.0 in that case, showing that a document about airplanes might not be rejectable in response to a query about gemstones, and vice-versa.

The second example of Table 7 illustrates the same problem for ambiguous proper nouns. Most proper nouns occur rarely in text collections, but when they occur they are often closely related to text content. Because they are concentrated in a few documents only, they tend to receive high weights. When such terms carry multiple meanings, problems may occur. In the example of Table 7, a document about Georgia matches another about the French Bourbon dynasty because a class of democratic politicians active in Georgia in the late 19th century were known as the Bourbons. Retrieval mistakes of that type normally cause few problems because the illegitimate output is easily identified.

The third example of Table 7 covers a difficulty that is system-specific and occurs occasionally when truncated terms are used in the document vectors instead of full word forms. Term truncation, or suffix removal, is attractive in retrieval because root forms and truncated terms represent broader concepts than full word forms, and search recall tends to increase. When suffixes are removed from a term like "genetics", "genet" is obtained, which happens to be a small animal similar to a weasel. Once again, an item erroneously retrieved because of such a difficulty is normally easily identified.

The last example of Table 7 illustrates a more subtle problem. Here, substantially similar events are covered, and the computed local text similarities are perfectly justifiable. However, a close examination of the texts reveals that the congruent events take place in different time frames. The appropriateness of retrieval may then be questioned. The example of Table 7 relates to hostilities involving a number of Greek cities, including Athens, Thebes, and Sparta, which took place in the first and the fifth centuries BC, respectively. Whether retrieval is appropriate in such cases depends on the context, and on the interests of the user.

Overall, problem situations such as those mentioned in Table 7 remain rare, and they affect the performance of the local text matching strategy only in minor ways.

5 Evaluation of Local Context Processing

The evaluation of information retrieval, or text linking, operations is a major unsolved problem. Normally, the effectiveness of text processing operations is evaluated by computing parameters such as recall and precision, defined as the proportion of the relevant items retrieved, and the proportion of retrieved items that are relevant, respectively. The concept of document relevance must be settled outside the retrieval environment. If texts are used both as search queries and as retrievable items in response to queries, then relevance assessments, or judgments of text relatedness, are needed for all pairs of texts. The encyclopedia used in the current experiments contains over 300 million pairs of articles, and over one billion pairs of text sections. Obviously, full relevance data can never be obtained for collections of that size, and hence complete recall-precision figures are not computable.

In the automated encyclopedia environment, fragmentary, objective relevance assessments are, however, available if the cross-references built into the text are used as relevance indicators. In particular, if article A carries a cross-reference to article B (for example, "Lee Harvey Oswald", see "John F. Kennedy"), then one can postulate that article B is relevant when A is submitted as a query (but not vice-versa). In the Funk and Wagnalls encyclopedia[7], the available cross-reference structure is sparse, consisting of only 80,000 references between article pairs, out of the over 300 million that are theoretically possible. This implies that very few relevant articles are objectively identified with respect to each query article. Articles not identified as relevant must be treated as nonrelevant, making it impossible to compute reasonable values for search precision and recall.

Tentative performance information is, however, obtainable by using relative performance differences between different search runs (instead of absolute performance measurements) to judge the effectiveness of the retrieval methodologies. Fig. 1 contains such data for seven restricted searches involving local paragraph comparisons, and six restricted searches using local sentence matches. Specifically, the graphs reflect the average improvement obtainable with the local paragraph or sentence matching restrictions at the indicated local similarity thresholds over a base run represented by an unrestricted search using a global vector comparison only. An overall evaluation parameter is used to obtain the percentage improvements of Fig. 1, consisting of the average search precision for 355 encyclopedia searches, evaluated at three recall points (0.20, 0.50 and 0.80). This measure represents a hybrid, single-value parameter that includes aspect of both recall and precision, and it has previously been used for global evaluation purposes.

The data of Fig. 1 are obtained by using as search requests the first 1,000 encyclopedia articles, generating the recall-precision information from the objective relevance data, averaging the per-query precision values valid at recall levels of 0.20, 0.50, and 0.80 over the 1,000 queries, and finally averaging the three average precisions to obtain a final, composite value. In practice, only 355 queries could be used for the computations, because no formally relevant items were defined by the available cross-reference structure in the encyclopedia for 645 out of the 1,000 query articles.

Fig. 1 shows that small improvements in overall performance are obtainable even with very low local paragraph or sentence matching thresholds of 30 or 40. Matching local structures (sentences or paragraphs) will be present at such a low threshold for most retrieved articles, including many of the nonrelevant ones. As the local matching threshold grows, the advantage of the local processing grows, reaching a maximum at thresholds of 75.0 to 100.0 for matching paragraph restrictions, and 50.0 to 75.0 for matching sentence restrictions. As the local matching requirement becomes even stricter, some of the relevant retrieved items are no longer able to meet the local similarity requirement, and consequently fall by the wayside. This depresses the recall and affects the overall evaluation parameter. This becomes especially obvious for the sentence comparisons where the performance decreases fast for sentence similarity thresholds greater than 100.0.

The information of Table 8 further clarifies the performance of the 355 sample encyclopedia searches. Here formal average recall and precision values are given to assess the performance after the retrieval of 15 documents per query. For some queries, it was impossible to find as many as 15 items with nonzero query similarity. In that case, a variable retrieval cut-off was used equal to the total number of encyclopedia articles exhibiting a nonzero similarity with the query. The recall values after 15 retrieved articles are generally high, indicating that the formally identified relevant articles are indeed retrieved among the top 15 items for most queries. The precision values are artificially low because most retrieved items are (falsely) declared as nonrelevant: formally only about 3 relevant articles exist on average for each query, but 15 are retrieved in each case.

The effect of the local search restrictions is most easily judged by considering the data in columns 2 and 3 of Table 8, giving the total number of retrieved items for the 355 search requests, and the total number of relevant retrieved items, respectively. As the table shows, the total number of retrieved items for all queries drops rapidly as increasingly severe local paragraph or sentence restrictions are applied. For example, when at least one matching paragraph pair is required above similarity threshold 75.0, only 4275 documents are retrieved in all, compared with 5155 for the unrestricted case, a decrease of 17 percent. However, the total number of relevant articles retrieved in that case decreases not at all but actually increases from 569 to 574. The same effect is noticeable for all restricted runs with a moderate local match requirement: the total number of items able to satisfy the local restriction is much lower, and a few of the items rejected in the restricted run are replaced by other relevant items not previously retrieved. This accounts for the fact that the search precision can increase substantially for the restricted runs while the recall improves moderately or stays the same. Properly applied search restrictions then act principally as precision devices, where they serve to reject large numbers of nonrelevant items, as illustrated earlier for the "John F. Kennedy" example of Tables 1-4.

6 Text Structuring and Retrieval Applications

A) Text Linking and Information Retrieval

When large text collections are processed in heterogeneous subject areas, the search and retrieval environment can be considerably enhanced by providing structured collection organizations where similar, or related texts are appropriately identified. Such an organization makes it possible to carry out browsing operations within particular subject areas, and to retrieve text items related to other relevant texts. Various types of relationships may prove useful in this connection, including formal relations holding between texts that share bibliographic citations, cross-references, and annotations, as well as the more elusive relationship between texts covering similar subject matter. In the present context, we assume that the formal text relationships are easily obtainable, either because the corresponding text elements are openly identified by appropriate text coding, or because citations, cross-references, and annotations are otherwise distinguishable from the remaining text. The discussion is then confined to the use of semantic text relations only.

Under the hypertext paradigm normally used for the display of structured text representations, a network of text nodes and text links is proposed, the nodes representing text elements, that is, pieces of text characterized by a variety of semantic specifications, and the links specifying relationships between the text elements. Quite detailed text specifications have been used in the literature with multiple types of text nodes characterized, for example, as arguments, claims, rebuttals, objections, and so on. The corresponding text links might then represent comparisons between text elements, replacements of one element by another, contributions of one element to another, references between elements, illustrations of one element by another, and so.[26-27]

In practice, such a detailed semantic specification cannot be supplied automatically, or even manually in many cases. The large multiplicity of semantic links and node types that might be useful in principle is therefore given up in favor of general semantic relationships indicating an unspecified general semantic affinity between corresponding text elements. The placement of a link between pairs of text nodes will indicate that users interested in text element A may also wish to consult text element B.

In the applications under consideration, a node represents a text excerpt of greater or lesser extent – typically an encyclopedia article, or a section of an article, or paragraph, or sentence included in an article. Two kinds of semantic links are used: first a hierarchical text link relating full texts to the corresponding text subdivisions, that is sections, paragraphs, and sentences; and second, a general semantic relationship link obtained by previously described vector similarity computations. Both link types indicate semantic affinity between corresponding text excerpts. In practice, they may be used separately or together, depending on particular requirements.

Many proposals have been made for using hypertext networks as components of normal information retrieval activities[28-33] Often a standard search process is used to specify a class of retrieved items, and the output is then modified in various ways by using the relations specified by the hypertext organization – for example, new items related in the hypertext may be added to the retrieved set of items, or the output ranking obtained in a standard search may be modified by using the links between items contained in the hypertext.

B) Topic Expansion in Information Retrieval

A common search tactic consists in starting with a general query formulation that approximately specifies an area of interest to a user, and then adding descriptions contained in some of the retrieved documents to build more specific, more narrowly-conceived formulations. A ranked retrieval strategy which presents the retrieved items in decreasing similarity order with the query, or in decreasing order of presumed relevance, is especially useful for query reformulation purposes. Facilities such as relevance feedback [34,35], and other methods of user-system interaction have been introduced to help the users in generating appropriate search formulations. When the texts are hierarchically decomposed into component parts, such as sections, paragraphs, sentences, or phrases, a narrower topic coverage may be automatically obtainable by using the component parts as retrieval units instead of full texts.

Consider a sample search carried out with the query article "Greece" (document 10300) illustrated in Table 9. Tables 9(a) and 9(b) contain output produced by unrestricted as well as sentence-restricted searches for query article "Greece". Table 9(c) indicates that the search restriction eliminates two initially retrieved cross-reference articles which are replaced by longer articles dealing with various Greek topics. Assuming that the user is more specifically interested in the "Delian League" (document 6944, ranked sixth on the original retrieved list and fourth in the restricted search), it is of course possible to use the text of article 6944 as a search request. The corresponding output was presented earlier in Table 5.

Alternatively, a relevance feedback step may serve to obtain a new query statement as a combination of the initial query text and the text of additional items designated as "relevant" by the user. Table 9(d) shows the results of such an automated feedback search. The top of

Table 9(d) contains the initial output obtained with the original query "Greece". The "yes" marker next to item 6944 indicates that the user has designated that item as relevant. The new search results obtained with the combined automatically formulated query for documents 10300 and 6944 are shown at the bottom of Table 9(d). A comparison of Tables 5(b) and 9(d) shows substantial similarities, but also some differences especially at the tail-end of the output lists.

An alternative way of reaching narrower topic classes consists in following the hierarchical structure and using text subdivisions as query formulations. A search of this kind is illustrated in Table 10. A user interested in Greek history could start the search by using section 76613 "Greece/History" as a query. The listing of Table 10(a) indicates that many of the retrieved sections cover particular aspects of Greek history. One such retrieved subsection, number 76615 "Greece/History/Ancient Greece", may be used next to narrow the search coverage to aspects of early Greek history. The corresponding output is given in Table 10(b). An additional refinement to Athenian topics within early Greek history is made by using as a query sub-subsection 76620 "Greece/History/Ancient Greece/The Ascendancy of Athens". The results are given in Table 10(c). The third retrieved item, section 51079, corresponds to article 6944 "Delian League". (Article 6944 consists of only one text section, and is therefore reachable by searching full articles as in Tables 5 and 9, or by a section search as in Table 10.)

A comparison of the output of Table 10(c) with the earlier searches of Tables 5 and 9 shows that some of the articles retrieved jointly with "Delian League" in response to subsection 76620 "Greece/...../The Ascendancy of Athens" were also obtained by the full document searches of Tables 5 and 9. Refined output can therefore be generated in many ways, by starting with specific query formulations such as "Delian League" (document 6944), by iterative feedback searches starting with broad topic descriptions such as "Greece" (document 10300), or finally by successive refinement based on hierarchical text decomposition as in the illustrations of Table 10. This last strategy is reminiscent of searches conducted with structured Tables of Content as proposed in the "Superbook" project.[36]

In situations where heterogeneous texts of widely differing scope and extent are present, mixed output might be provided consisting of the full texts of short and specific items, plus additionally short excerpts – for example, individual sections or paragraphs – of the longer more discursive items. In an encyclopedia environment, some articles consist only of short definitions, or cross-references to other articles ("Pallas Athena, Greek goddess, see Athena"), while other articles may cover a wide selection of topics under general headings such as "United States of America". Providing references to long, undivided text without additional location information will not help the user in finding what is wanted. A mixed strategy may then be attractive which treats single-section texts as complete entities that are retrievable directly, while using text excerpts, such as sections or paragraphs, as retrieval objects for longer texts.

The following strategy was used to produce the mixed output for the searches of Table 11:

a) Documents consisting of single text sections are treated in the conventional way by using normal vector comparisons between document and query vectors.

b) Documents retrieved by the global vector comparison consisting of more than one text section are subdivided, and each section or subsection is individually compared with the query.

c) If the query-section similarity for all available sections falls below a stated threshold (0.35 for the output of Table 11), no change appears in the output listing; that is, the references to the complete articles are maintained as before.

d) If, however, a substantial query-section similarity exists above the stated threshold for at least one of the text sections, then the text section with the highest query-section similarity replaces the reference to the full text on the list of retrieved items.

Consider the output of Table 11(a) for query 9894 "Giotto" (the fourteenth century Italian painter). The retrieval list includes references to a number of other artists closely allied with Giotto (Gaddi, Cimabue, Pisano, Masaccio, etc.). Also included are long essays of several dozen pages on "Painting" (document 17379) and "Fresco" (document 9318). A separate section comparison for these texts identifies section 17 of document 17379, and section 7 of document 9318 as the ones most closely related to the query topic. This is not surprising because section 17379.17 carries section titles "Painting/Medieval Painting/Giotto", and section 9318.7 is entitled "Fresco/History". In both cases, the connection with the query topic is immediate and far more direct than the original references to the full article texts.

The same effect is noted for the search of Table 11(b) for query 9807 "Battle of Gettysburg". The retrieval list includes a variety of specific references to persons and events of the American Civil War. Also included is article 5496 entitled "Civil War, American" which covers the subject in its entirety in several dozen pages of text. A separate section match identifies section 21 of item 5496 as the one most closely related to the query. Its section headings are "Civil War, American/ Hostilities/ Gettysburg".

The possibility of breaking up long texts into component pieces opens up a variety of opportunities for tailoring the retrieval output to the requirements and interest of particular users.

C) Structured Text Output

The text relationships illustrated in the previous section are used mainly to improve the usefulness and accuracy of query-document comparisons in information retrieval. The study of arbitrary relationships that may be valid in large text collections may prove useful even when information retrieval is not specifically involved, and when specific user queries are not present. The display of such multiple text relationships typically takes the form of maps, consisting of nodes representing documents or document excerpts, and branches representing relations between corresponding text pairs. The question arises whether such text maps can be generated automatically, what form such a display might take, and how the maps are to be used. The discussion here is restricted to a few suggestions involving general semantic text relationships.

The first possibility for obtaining a network of related text nodes consists in using a multi-stage search: An initial query is used first to retrieve m particular items. Each of these retrieved items is used next as a query for a further search also retrieving m items each. When n different search stages of this type are used, a breath m-depth n search will have been carried out. The results of such a composite search can be represented in tree form as shown in Table 12. Here the article "Delian League" (document 6944) is used as an initial search query. Three documents are retrieved at stage 1 (numbers 6957, 1655 and 10300); these are used in the second stage to retrieve 3 additional items each. In the example of Table 12, the search is repeated a third time and a breadth 3-depth 3 search tree is obtained. In principle, the total number of retrieved articles grows exponentially at each stage. In practice, some of the documents already retrieved in earlier stages may be obtained repeatedly. In that case, the tree is pruned by removing duplicate subtrees.

In the example of Table 12, three duplicate items are found among the nine items retrieved in the second stage. Two of these repeat the original query, and the third (1655 "Athens") is a duplicate of a level-one item. Because the searches are not pursued for the duplicated items on level two, only 18 items are retrieved on level three of the tree, instead of a possible 27. Of these, seven are repetitions of items seen earlier (labelled 1655 "Athens", 6957 "Confederation of Delos", 6944 "Delian League", 10300 "Greece", and 17824 "Pericles"). The total number of distinct items obtained in the sample search (exclusive of the query) is therefore 3 + 6 + 11 = 20 instead of maximum possible number of 3 + 9 + 27 = 39. A comparison of the effectiveness of multilevel tree searching with the standard single-stage searches must take into account the possibility of duplicate retrievals and of variable numbers of retrieved items.

In the search strategy of Table 12, each separate query is expected to retrieve a fixed number of m items. The number retrieved at each point can be made to vary by using a retrieval threshold based on the query-document similarity. In particular by increasing the similarity threshold needed for retrieval from one search stage to the next, fewer and fewer items will emerge in the later search stages. A closed output map is then generated containing items that are increasingly similar to each other as one moves further away from the original query.

Fig. 2 shows such an automatically produced map starting with the original query text "Giotto" (document 9874). The retrieval cut-off is set at 0.35 for documents retrieved at stage one; this increases to 0.40 for stage-two searches, and to 0.45 for stage three. As the figure shows, four items are retrieved at stage one including a teacher of Giotto's, a pupil, an assistant, and another artist directly influenced by Giotto's work. The similarities with the search item are listed along the respective branches. Only three new items are obtained on level two, because of the high retrieval cut-off of 0.40, and only one more on level three. Document 983 provides an alternative name for Andrea Pisano (document 18198), and the three other items retrieved in the later stages cover aspects of early Renaissance painting in Italy that continue the development started by Giotto.

Linked hypertext maps such as that of Fig. 2 represent self-contained topic areas, and are directly usable for collection browsing. Such output maps might also be tailored to user interests by using variable retrieval thresholds that are lowered for topic areas of special interest, and increased for more marginal topic classes. This should help in including in the maps items likely to be of interest to the user, while rejecting the others.[37]

When items are retrieved in answer to particular queries, a substantial similarity is normally present between the query and each of the retrieved items. This does not, however, imply that the retrieved items are especially similar to each other. In a retrieval situation, the search precision may be improvable by making sure that all retrieved items are similar not only to the common query, but also to each other. The likelihood of retrieving non-relevant items is thus reduced.

This suggest that a clustering scheme capable of grouping mutually related items into affinity classes, or clusters, be added to the retrieval environment.[38-41] In the hypertext context, aggregates of similar nodes have been introduced as a way of simplifying overly complex structures by creating super-nodes each of them representing a whole set of highly similar nodes.[42-44] Methods have also been proposed for using clustering systems in hypertext environments.[45,46]

One simple method for implementing a cluster-based text linking and retrieval system consists in carrying out search and retrieval operations normally; and then modifying the ranked output to take into account the available clustering information. Fig 3(a) contains a set of complete-link clusters all of them including document 17824 "Pericles". The overall clustering similarity, representing the minimum similarity between any pair of documents included in a common cluster, appears next to each defined cluster. Thus, the cluster similarity for the three-element cluster consisting of 17824 "Pericles", 14851 "Marathon", and 17416 "Pallas Athena" is a relatively high 0.39 (on a scale from 0 to 1). This indicates that the three items are closely related.

The list of retrieved items obtained by an unrestricted search with query 6944 "Delian League" appears in Fig 3(b), with added clustering information. Assuming that a clustering threshold of 0.30 is considered significant, the retrieval of item 1655 "Athens" might immediately lead to four additional items grouped with "Athens" in a common cluster, including 17416 "Pallas Athena", 17824 "Pericles", 14851 "Marathon", and 1651 "Athena". When a cluster-based search strategy is used, these items could be retrieved ahead of the unclustered items that immediately follow "Athens" on the original retrieval list.

The use of clustering organizations in information retrieval environments has been limited in practice, because of the expense of generating the cluster structures for large collections of documents. In the proposed implementation, an overall clustering system is unnecessary. Instead the clustering operation can be applied locally to items jointly retrieved in response to available queries. In such circumstances, the clusters can be generated efficiently. The generated classes might also be adapted to user interests when that type of information is available.

7 Conclusion

A text manipulation system is introduced in this study which uses global and local text matching operations to identify stored texts likely to be of interest to system users. The text analysis and retrieval strategies described here are applicable to large collections of texts without subject matter restrictions, and to heterogeneous documents that vary widely in scope and extent. No large preconstructed knowledge sources are needed; however such tools can be added whenever they become available for use in unrestricted topic environments.

Retrieval strategies are also described that make use of hierarchical text decomposition to successively refine the query statements and hence the coverage of the retrieved items. In addition, several types of automatically constructed linked text structures are introduced. Such hypertext products should make it possible to carry out flexible browsing operations among sets of related items, and simplify collection utilization by the user population.

8 Acknowledgement

The writers are grateful to the Microsoft Corporation for making available an electronic version of the Funk and Wagnalls encyclopedia for experimental purposes. The first author is also greatly indebted to the Alexander von Humboldt Foundation for supporting a research stay in Germany, and to Professor Rainer Kuhlen and the information science group at the University of Konstanz for providing the attractive working environment where this study was performed.

9 References

1. Y. Bar-Hillel, Theoretical Aspects of the Mechanization of Literature Searching, in Digital Information Processors, W. Hoffmann, ed., Interscience Publishers, New York, 1962, 406-443.

2. D.C. Blair and M.E. Maron, Full-Text Information Retrieval: Further Analysis and Clarification, Information Processing and Management, 26:3, 1990, 437-447.

3. D.C. Blair, Language and Representation in Information Retrieval, Elsevier Science Publishers, New York, 1990.

4. R. Furuta, C. Plaisant, and B. Shneiderman, Automatically Transforming Regularly Structured Linear Documents into Hypertext, Electronic Publishing, 2:4, December 1989, 211-229.

5. R.J. Glushko, Design Issues for Multi-Document Hypertext, Proceedings Hypertext-89, Association for Computing Machinery, New York, November 1989, 51-60.

6. C. Stanfill and D. Waltz, Toward Memory-Based Reasoning, Communications of the ACM, 29:12, December 1986, 1213-1228.

7. Funk and Wagnalls New Encyclopedia, Funk and Wagnalls, New York, 1979, 29 volumes.

8. G. Salton, Automatic Text Processing – The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing Co., Reading, MA 1989.

9. G. Salton, Developments in Automatic Text Retrieval, Science, 253, 30 August 1991, 974-980.

10. G. Salton, C.S. Yang and A. Wong, A Vector Space Model for Automatic Indexing, Communications of the ACM, 18:11, Nov. 1975, 613-620.

11. G. Salton, ed., The Smart Retrieval System – Experiments in Automatic Document Processing, Prentice Hall, Inc., Englewood Cliffs, NJ. 1971.

12. G. Salton, A Theory of Indexing, Regional Conference Series in Applied Mathematics No. 18, Society for Industrial and Applied Mathematics, Philadelphia, PA 1975.

13. G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, Journal of the Am. Society for Information Science, 26:1, January 1975, 33-44.

14. G. Salton and C. Buckley, Term Weighting Approaches in Automatic Text Retrieval, Info. Proc. and Management, 24:5, 1988, 513-523.

15. F. Armengaud, La Pragmatique, Presses Universitaires de France, Paris, 1985.

16. G.N. Leech, Principles of Pragmatics, Longman, London, 1983.

17. J.R. Searle, The Background of Meaning, in Speech Act Theory and Pragmatics, J.R. Searle, F. Kiefer, and M. Bierwich, editors, Reidel Publishing Co., Dordvecht, Holland, 1980, 221-237.

18. W.T. Alston, Meaning and Use, in J.F. Rosenberg and C. Travis, Readings on the Philosophy of Language, Prentice Hall, Englewood Cliffs, N.J., 1971.

19. L. Wittgenstein, Philosophical Investigations, Basil Blackwell and Mott Ltd., Oxford, England, 1953.

20. J.L. Austin, How to do Things with Words, Clarendon Press, Oxford, England, 1962.

21. G. Salton and C. Buckley, Global Text Matching for Information Retrieval, Science 253, 5023, 30 August 1991, 1012-1015.

22. G. Salton and C. Buckley, Automatic Text Structuring and Retrieval - Experiments in Automatic Encyclopedia Searching, Proc. Fourteenth Int. ACM/SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 1991, 21-30.

23. D.R. Swanson, Complementary Structures in Disjoint Science Literatures, Proc. Fourteenth Annual Int. ACM/SIGIR Conference on Research and Development in Infor-

mation Retrieval, Association for Computing Machinery, New York, 1991, 280-289.

24. M. Bernstein, An Apprentice That Discovers Hypertext Links, in Hypertext: Concepts, Systems and Applications, A. Rizk, N. Streitz, and J. André (editors), Cambridge University Press, 1990, 212-223.

25. M. Bernstein, J.D. Bolter, M. Joyce, E. Mylonas, Architectures for Volatile Hypertext, Proc. Hypertext 91, Association for Computing Machinery, New York, 1991, 243-260.

26. J. Nanard and M. Nanard, Using Structured Types to Incorporate Knowledge in Hypertext, Proc. Hypertext-91, Association for Computing Machinery, New York, 1991, 329-343.

27. W. Schuler and J.B. Smith, Argumentative Texts, in Hypertext: Concepts, Systems, and Applications, in A. Rizk, N. Streitz, and J. André (editors), Cambridge University Press, 1990, 137-151.

28. N. Fuhr, Hypertext and Information Retrieval, in P. Gloor and N. Streitz, editors, Hypertext and Hypermedia, Springer Verlag, Berlin, 1990, 101-111.

29. W.B. Croft and H. Turtle, A Retrieval Model Incorporating Hypertext Links, Proc. Hypertext 89, Association for Computing Machinery, New York, 1989, 213-223.

30. P. Clitherow, D. Riecken, and M. Muller, VISAR: A System for Inference and Navigation in Hypertext, Proc. Hypertext 89, Association for Computing Machinery. New York, Nov. 89, 293-304.

31. D. Lucarella, Systems and Applications, in A. Rizk, N. Streitz, and J. André (editors) Cambridge University Press, 1990, 212-223.

32. D.R. Raymond and F.W. Tompa, Hypertext and the Oxford English Dictionary, Communications of the ACM, 31, 1988, 871-877.

33. M.E. Frisse and S.B. Cousins, Guides for Hypertext: An Overview, Artificial Intelligence in Medicine, 2, 1990, 303-314.

34. J.J. Rocchio, Relevance Feedback in Information Retrieval, in The Smart System – Experiments in Automatic Document Processing, G. Salton, editor, Prentice Hall Inc., Englewood Cliffs, N.J. 1971, 313-323.

35. G. Salton and C. Buckley, Improving Retrieval Performance by Relevance Feedback, Journal of the Am. Soc. for Information Science, 41:4, 1990, 288-297.

36. E. Egan, M.E. Lesk, R.D. Ketchum, C.C. Lochbaum, J.R. Remde, M. Littman, T.R. Landauer, Hypertext for the Electronic Library? CORE Sample Results, Proc. Hypertext 91, Association for Computing Machinery, New York, December 1991, 299-312.

37. J.H. Coombs, Hypertext, Full-Text, and Automatic Linking, Proc. Thirteenth Int. ACM/SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 1990, 83-98.

38. N. Jardine and C.J. van Rijsbergen, The Use of Hierarchic Clustering in Information Retrieval, Information Storage and Retrieval, 7:5, December 1971, 217-240.

39. G. Salton and A. Wong, Generation and Search of Clustered Files, ACM Trans. on Database Systems, 3:4, December 1978, 321-346.

40. F. Murtagh, A Survey of Recent Advances in Hierarchical Clustering Algorithms, The Computer Journal, 26:4, 1982, 354-360.

41. W.B. Croft, Clustering Large Files of Documents Using the Single Link Method, Journal of the ASIS, 28:6, November 1977, 341-344.

42. H. Van Dyke Tarunak, Don't Link Me In: Set Based Hypermedia for Taxonomic Reasoning, Proc. Hypertext 91, Association for Computing Machinery, New York, December 1991, 233-242.

43. P.A. Gloor, CYBERMAP: Yet Another Way of Navigation in Hyperspace, Proc. Hypertext 91, Association for Computing Machinery, New York, December 1991, 107-121.

44. R.A. Botafogo and B. Shneiderman, Identifying Aggregates in Hypertext Structures, Proc. Hypertext 91, Association for Computing Machinery, New York, Dec. 1991, 63-74.

45. D.B. Crouch, C.J. Crouch, and G. Andreas, The Use of Cluster Hierarchies in Hypertext Information Retrieval, Proc. Hypertext 89, Association for Computing Machinery, New York, Nov. 1989, 225-237.

46. W.B. Croft and R.H. Thompson, I3R: A New Approach to the Design of Document Retrieval Systems, Journal of the Am. Society for Information Science, 33:6, 1987, 389-404.

Query: Text of encyclopedia article (12941) "John F. Kennedy"		
Document Number	Query Similarity	Title of Retrieved Item
12941	1.00	Kennedy, John Fitzgerald
12943	0.58	Kennedy, Robert F(rancis)
12942	0.53	Kennedy, Joseph P(atrick)
17274	0.42	Oswald, Lee Harvey
23948	0.41	Warren Report
12939	0.40	Kennedy, Anthony M.
12940	0.40	Kennedy, Edward M(oore)
20456	0.38	Schlesinger, Arthur Meier, Jr.
12589	0.35	Johnson, Lyndon Baines
23320	0.34	United States of America
4341	0.31	Cape Canaveral
2338	0.29	Bay of Pigs Invasion
12944	0.26	Kennedy Center
12575	0.26	John F. Kennedy Center for the Performing Arts
18612	0.25	President of the United States

Table 1: Retrieval Output for Query Text 12941 "John F. Kennedy". Global text comparison only.

Common Terms (and term weights)				
Vec1:	article 12941 (John F. Kennedy)			
Vec2:	article 12939 (Anthony M. Kennedy)			
Concept Number	Term Weight Vec1	Vec2	Combined	Corresponding Word Stem
9522	0.1128	0.0861	0.0097	u.s.
13391	0.0141	0.2262	0.0032	court
15637	0.0378	0.1010	0.0038	senat
25092	0.0220	0.1178	0.0026	harvard
29496	0.0372	0.1324	0.0049	school
42502	0.0210	0.1121	0.0024	graduat
61521	0.1577	0.1404	0.0221	presid
84149	0.0587	0.1045	0.0061	justic
84404	0.0099	0.0530	0.0005	educat
86140	0.0176	0.1881	0.0033	suprem
89788	0.7082	0.4539	0.3214	kenned
100530	0.0070	0.0373	0.0003	born
109419	0.0243	0.0650	0.0016	consid
114184	0.0178	0.0950	0.0017	univers
131307	0.0157	0.0419	0.0007	year
134333	0.0207	0.1107	0.0023	appeal
163578	0.0119	0.2537	0.0030	law
182978	0.0489	0.1306	0.0064	nominat
196494	0.0230	0.1227	0.0028	confirm

Table 2: Matching Vocabulary for Articles 12939 and 12941 (Anthony M. Kennedy and John F. Kennedy)

12939	Kennedy, Anthony M.
	Kennedy, private law practice
	Kennedy, conservative
	Kennedy, nominated to Supreme Court
	Kennedy, nominated by President Reagan
12941	Kennedy, John F.
	Kennedy, born in Brookline, Mass.
	Kennedy, completed Profiles in Courage
	Kennedy, began plan for presidental election
	Kennedy, approved invasion of Cuba
	Kennedy, chilled by Krushchev's warning
	Kennedy, sent 1500 troops over land to Berlin
	Kennedy, wit and charm
	Kennedy, shot in the back
	Kennedy, state funeral watched on TV

Table 3: Local Context of Term 'Kennedy' in Documents 12939 and 12941

Restricted Search Query: Text of encyclopedia article (12941) "John F. Kennedy" Restriction: At least one matching pair of sentences with similarity of 75.0 or more		
Document Number	Query Similarity	Title of Retrieved Item
12941	1.00	Kennedy, John Fitzgerald
12943	0.58	Kennedy, Robert F(rancis)
12942	0.53	Kennedy, Joseph P(atrick)
17274	0.42	Oswald, Lee Harvey
23948	0.41	Warren Report
12940	0.40	Kennedy, Edward M(oore)
20456	0.38	Schlesinger, Arthur Meier, Jr.
12589	0.35	Johnson, Lyndon Baines
23320	0.34	United States of America
2338	0.29	Bay of Pigs Invasion
18612	0.25	President of the United States
1410	0.23	Arlington National Cemetery
6489	0.22	Cuba
13032	0.21	Khrushchev, Nikita S(ergeyevich)
23639	0.21	Vice President of the United States
Items rejected because of search restriction:		
12939	A.M. Kennedy	
4341	Cape Canaveral	
12944	Kennedy Center	
12575	J.F. Kennedy Center for the Performing Arts	

Table 4: Results of Local Retrieval Strategy for Query Document 12941 (Requiring at Least One Matching Sentence Pair)

Document Number	Similarity Measure	Title
6944	1.00	Delian League.
1653*	0.66	Athenian League.
6957	0.48	Delos, Confederation of.
1655	0.43	Athens.
13673*	0.43	League of Nations.
10300	0.41	Greece.
17416*	0.40	Pallas Athena.
17824	0.39	Pericles.
6956	0.38	Delos.
1100	0.36	Antalcidas.
1386	0.35	Aristides.
14851	0.34	Marathon.
5436	0.33	Cimon.
1651	0.33	Athena.
23373*	0.33	Urban League.

a) Unrestricted Output

Document Number	Similarity Measure	Title
6944	1.00	Delian League.
6957	0.48	Delos, Confederation of.
1655	0.43	Athens.
10300	0.41	Greece.
17824	0.39	Pericles.
6956	0.38	Delos.
1100	0.36	Antalcidas.
1386	0.35	Aristides.
14851	0.34	Marathon.
5436	0.33	Cimon.
1651	0.33	Athena.
850	0.33	Alcibiades.
907	0.32	Amphictyonic League.
22619	0.32	Thucydides.
20144	0.31	Samos.

b) Restricted Output (at Least One Sentence Match with
Similarity Threshold 75.0)

Table 5: Search Results for Query Article 6944
"Delian League"

Document Number	Title	Characterization
1653	Athenian League	A cross-reference only
13673	League of Nations	Extraneous topic
17416	Pallas Athena	A cross-reference only
23373	Urban League	Extraneous topic

c) Originally Retrieved Items Deleted in Restricted Search

Document Number	Title	Characterization
580	Alcibiades	Famous Athenian Statesman
901	Amphictyonic League	League of ancient Greek tribes
22619	Thucydides	Greek historian living around 450 BC
20144	Samos	Greek island

d) Items Added in Restricted Search (to replace deleted items)

Table 5 (cont.): Search Results for Query Article 6944 "Delian League"

Section Number	Document Number	Similarity Measure	Title of Retrieved Section
76585	10300.13	1.00	Greece./Population./Political.Divisions.
76577	10300.5	0.36	Greece.
91581*	12261.13	0.34	Italy./Population./Political Divisions.
76584	10300.12	0.34	Greece./Population./Population Characteristics.
107621	14492.7	0.33	Macedonia./Greek Macedonia.
76578	10300.6	0.33	Greece./Land and Resources.
132547	17748.5	0.32	Peloponnisos. .
76611	10300.39	0.32	Greece./Government./Local Government.
60485	8173.5	0.31	Epirus.
76586	10300.14	0.30	Greece./Population./Principal Cities.
40171	5476.5	0.29	City-State.
97048	13007.5	0.28	Khalkis.
12516	1660.5	0.26	Athos.
119642	16088.7	0.26	Municipal Government./Weak-Mayor-Council Plan.
76627	10300.55	0.26	Greece./History./Roman and Medieval Greece.

a) Unrestricted Search Output

Section Number	Title	Characterization
91581	Italy./Population./Political Divisions.	Related topic; not a specific answer
132547	Peloponnisos.	A cross-reference only
76611	Greece./Government./Local Govt.	A cross-reference only
40171	City-State.	A cross-reference
119642	Municipal Govt./Weak-Mayor-Council Plan.	Extraneous topic
76627	Greece./History./Roman and Medieval Greece.	Related topic; not a specific answer

b) Originally Retrieved Items Deleted in Restricted Search (at least one matching sentence with similarity threshold 100)

Section Number	Document Number	Similarity Measure	Title
2102	296.5	0.25	Aegean Islands.
76583	10300.11	0.24	Greece./Population.
76635	10300.63	0.24	Greece./History./Modern Greece./Struggle for Territory.
76614	10300.42	0.24	Greece./History./Prehistoric Period.
132481	17738.5	0.24	Pelasgians (early inhabitants of Greece)
40145	5474.5	0.23	City (refers to Greek city-state)

c) Items Added in Restricted Search (to replace deleted items)

Table 6: Search Results for Query Section 76585 "Greece/Population/ Political Divsions"

Problem	Illustration
Highly weighted ambiguous term (wheel)	Sentence 77, Document 9679 (Gemstones) The stone to be shaped is cemented to the end of a wooden stick called a dop and is held against the revolving wheel or lap with the end of a supporting block placed adjacent to the wheel. Sentence 124, Document 454 (Airplane) A tricycle gear consists of two large wheels behind the center of gravity and a third wheel, called the nosewheel, in front of the two main wheels.
Ambiguous proper noun (Bourbon)	Sentence 218, Document 9754 (Georgia) Poor agricultural conditions created widespread support for the Populists, who challenged the Bourbons for political power in the 1890's but quickly faded thereafter. Sentence 9, Document 3334 (Bourbons) The earliest documented member of the Bourbon family was a French feudal lord, Almar or Adhemar, who became baron of Bourbon in the late 9th century.
Term truncation problem	Sentence 9, Document 9698 (Genetics) Emergence of Genetics: The science of genetics began in 1900, when several plant breeders independently discovered the work of the Austrian monk Gregor Mendel. Sentence 6, Document 9693 (Genet) Genets inhabit forests and dense grasslands throughout Africa; one species, G. genetta, is also found in southwestern Europe.
Matching events in different time frame	Sentence 366, Document 10300 (Greece/History/Roman and Medieval Greece) Roman legions under Lucius Cornelius Sulla forced Mithridates out of Greece and crushed the rebellion, sacking Athens in 86 BC and Thebes a year later. Sentence 10, Document 18272 (Plataea) The allegiance of the Plataeans to Athens angered the Thebans, and in 429 BC, the third year of the Peloponnesian War, the city was attacked by a combined force of Thebans and Spartans and razed to the ground.

Table 7: Examples of Sentence Matching Problems

Type of Search	Total Number of Retrieved Items	Total Number of Relevant Retrieved Items	Recall after 15 retrieved Items per Query	Precision after 15 retrieved Items per Query	Overall Rating
Unrestricted Search	5155	569	0.8198	0.1147	--
1 para 30	5143	576	0.8367	0.1162	+ 5.8%
1 para 50	4860	582	0.8456	0.1269	+ 11.8%
1 para 75	4257	574	0.8173	0.1661	+ 13.8%
1 para 100	3844	565	0.8052	0.2133	+ 13.2%
1 para 125	3584	549	0.7827	0.2371	+ 11.1%
1 para 150	3418	524	0.7529	0.2354	+ 6.7%
1 para 200	3229	492	0.7283	0.2424	+ 0.8%
1 sent 30	4906	580	0.8436	0.1245	+ 7.0%
1 sent 40	4712	573	0.8293	0.1303	+ 7.5%
1 sent 50	4268	578	0.8352	0.1663	+ 13.4%
1 sent 75	3459	535	0.7666	0.2335	+ 10.5%
1 sent 100	2877	449	0.6622	0.2187	- 4.8%
1 sent 150	1758	323	0.4771	0.2149	- 27.4%

Table 8: Evaluation of Local Sentence and Paragraph Matching (global statistics for 355 search sequests)

Document Number	Query Similarity	Title of Retrieved Item
10300	1.00	Greece.
1655	0.46	Athens.
5476	0.45	City-State.
15390	0.42	Metaxas, Ioannes.
17883	0.41	Persian Wars.
6944*	0.41	Delian League.
23566	0.38	Venizelos, Eleutherios.
18190	0.38	Piraievs.
12824	0.38	Karamanlis, Constantine.
6398	0.37	Crete.
8349	0.37	Europe.
11016	0.36	Hellas.
10305	0.35	Greek Language.
296	0.35	Aegean Language.
22512	0.35	Thebes.

a) Unrestricted Search Output for Query 10300 (Greece)

Document Number	Query Similarity	Title of Retrieved Item
10300	1.00	Greece.
1655	0.46	Athens.
15390	0.42	Metaxas, Ioannes.
6944*	0.41	Delian League.
23566	0.38	Venizelos, Eleutherios.
18190	0.38	Piraievs.
12824	0.38	Karamanlis, Constantine.
6398	0.37	Crete.
8349	0.37	Europe.
11016	0.36	Hellas.
10305	0.35	Greek Language.
296	0.35	Aegean Language.
22512	0.35	Thebes.
10307	0.35	Greek Literature.
1694	0.35	Attica.

b) Restricted Search for Query Greece (10300) (at least one sentence match at 75.0)

Items deleted in restricted search:		
5476	City-State	(a cross-reference)
17883	Persian Wars	(a cross-reference)
Items added in restricted search:		
10307	Greek Literature	(long article)
1694	Attica	(Greek region)

c) Effect of Search Restriction

Table 9: Typical Search Example for Query 10300 "Greece"

Previously retrieved docs from iteration 0			
10300		1.00	Greece.
1655		0.46	Athens.
15390		0.42	Metaxas, Ioannes.
6944	yes	0.41	Delian League.
23566		0.38	Venizelos, Eleutherios.
18190		0.38	Piraievs.
12824		0.38	Karamanlis, Constantine.
6398		0.37	Crete.
8349		0.37	Europe.
11016		0.36	Hellas.
10305		0.35	Greek Language.
296		0.35	Aegean Islands.
22512		0.35	Thebes.
10307		0.35	Greek Literature.
1694		0.35	Attica.

Feedback Search 6944		
1653	0.58	Athenian League.
17824	0.42	Pericles.
17416	0.40	Pallas Athena.
14851	0.38	Marathon.
5436	0.37	Cimon.
6956	0.37	Delos.
1651	0.37	Athena.
20144	0.36	Samos.
580	0.36	Alcibiades.
22619	0.35	Thucydides.
148	0.35	Achaean League.
1656	0.34	Athens.
4779	0.34	Cecrops.
14453	0.34	Lysander.
3467	0.33	Brasidas.

d) Relevance Feedback Search (initial query "Greece", feedback item "Delian League"). Restricted Search: at least one sentence match above threshold 75.0.

Table 9 (cont.): Typical Search Example for Query 10300 "Greece"

Section Number	Document Number	Query Similarity	Title
76613	10300.41	1.00	Greece./History.
76622	10300.50	0.32	Greece./.../Hellenic Period/Shifting Alliances.
76577	10300.5	0.29	Greece.
76628	10300.56	0.27	Greece./History./.../Greek Renaissance.
76620	10300.48	0.26	Greece./History./.../Ascendancy of Athens.
76621	10300.49	0.26	Greece./History./.../The Peloponnesian War.
76615	10300.43	0.25	Greece./History./Ancient Greece.
76627	10300.55	0.25	Greece./History./Roman and Medieval Greece.
2102	296.5	0.25	Aegean Islands.
76578	10300.6	0.25	Greece./Land and Resources.
61802	8349.31	0.24	Europe./History./.../Supremacy of Greece.
169034	22512.5	0.24	Thebes.
45329	6159.5	0.24	Corinth.
76592	10300.20	0.23	Greece./Population./Culture.
12776	1694.5	0.22	Attica.

a) Search Results for Section 76613 "Greece/History" (at least one matching sentence required at threshold 75.0)

Section Number	Document Number	Query Similarity	Title
76615	10300.43	1.00	Greece./History./Ancient Greece.
54437	7394.5	0.48	Dorians.
76614	10300.42	0.39	Greece./History./Prehistoric Period.
1034	149.5	0.36	Achaeans.
1026	148.5	0.35	Achaean League.
90290	12130.5	0.32	Ionians.
61802	8349.31	0.31	Europe./History./.../Supremacy of Greece.
47078	6398.8	0.31	Crete./History.
2102	296.5	0.31	Aegean Islands.
76577	10300.5	0.29	Greece.
76625	10300.53	0.29	Greece./History./Hellenistic Period./ The Diadochi.
61800	8349.29	0.28	Europe./History./.../Arrival of Indo-Europeans.
2083	295.5	0.28	Aegean Civilization.
90278	12128.5	0.28	Ionia.
2212	313.5	0.27	Aeolians.

b) Search Result for Section 76615 "Greece/History/Ancient Greece" (at least one matching sentence required at threshold 75.0)

Table 10: Examples of Hierarchical Expansion Using Narrower Contexts

Section Number	Document Number	Query Similarity	Title
76620	10300.48	1.00	Greece./History./.../The Ascendancy of Athens.
133112	17824.5	0.62	Pericles.
51079	6944.5	0.59	Delian League.
12476	1655.9	0.57	Athens./History./The Classical Period.
76621	10300.49	0.51	Greece./History./.../ The Peloponnesian War.
76618	10300.46	0.51	Greece./History./.../From Monarchy to Democracy.
130121	17416.5	0.51	Pallas Athena.
12477	1655.10	0.49	Athens./History./Foreign Domination.
110252	14851.5	0.47	Marathon.
76619	10300.47	0.45	Greece./.../The Hellenic Period./The Persian Wars.
12446	1651.5	0.44	Athena.
39860	5436.5	0.44	Cimon.
12458	1653.5	0.44	Athenian League.
133228	17839.5	0.42	Peristeri.
169805	22619.5	0.41	Thucydides.

c) Search Result for Section 76620 "Greece/History/Ancient Greece/The Ascendancy of Athens" (at least one matching sentence at 75.0)

Table 10 (cont.): Examples of Hierarchical Expansion Using Narrower Contexts

Document Number	Title of Retrieved Item	Query Similarity
9894	Giotto	1.00
9451	Gaddi, Taddeo	0.47
5431	Cimabue	0.46
18198	Pisano, Andrea	0.42
15028	Masaccio	0.37
19289	Renaissance/Art and Architecture	0.33
4744	Cavallini, Pietro	0.32
17379.17	Painting	0.32
14044	Lippi, Fra Filippo	0.30
7349	Domenico, Veneziano	0.28
9822.6	Ghirlandaio	0.26
9318.7	Fresco	0.25
17153	Orcagna, Andrea	0.25
985	Andrea del Sarto	0.25
9717	Gentile da Fabriano	0.24

a) Output for Query "Giotto" (document 9894)
[section output: 17379.17: Painting/Medieval Painting/Giotto
9822.6 : Ghirlandaio/Domenico di Tommaso Bigordi Ghirlandaio
9318.7 : Fresco/History]

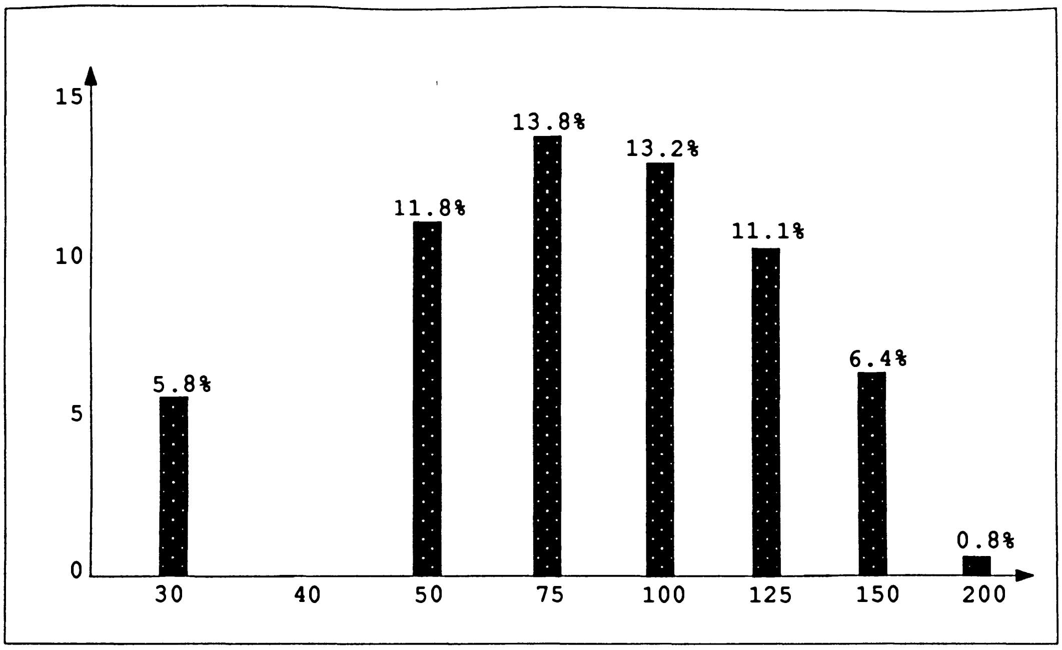
Document Number	Title of Retrieved Item	Query Similarity
9807	Gettysburg, Battle of	1.00
24316	Wilderness, Battle of the	0.48
14213	Longstreet, James	0.42
5496.21	Civil War, America	0.40
18103	Pickett, George Edward	0.39
3856	Bull Run, Battle of	0.36
8408	Ewell, Richard Stoddert	0.35
4948	Chancellorsville, Battle of	0.34
11258	Hill, Ambrose Powell	0.34
15179	Meade, George Gordon	0.32
5736	Cold Harbor, Battle of	0.31
13709	Lee, Robert Edward	0.31
1129	Antietam, Battle of	0.29
9806	Gettysburg	0.29
21838	Stuart, James Ewell Brown	0.27

b) Output for Query "Gettysburg, Battle of" (document 9807)
[section output: 5496.21 "Civil War, American/Hostilities/Gettysburg]

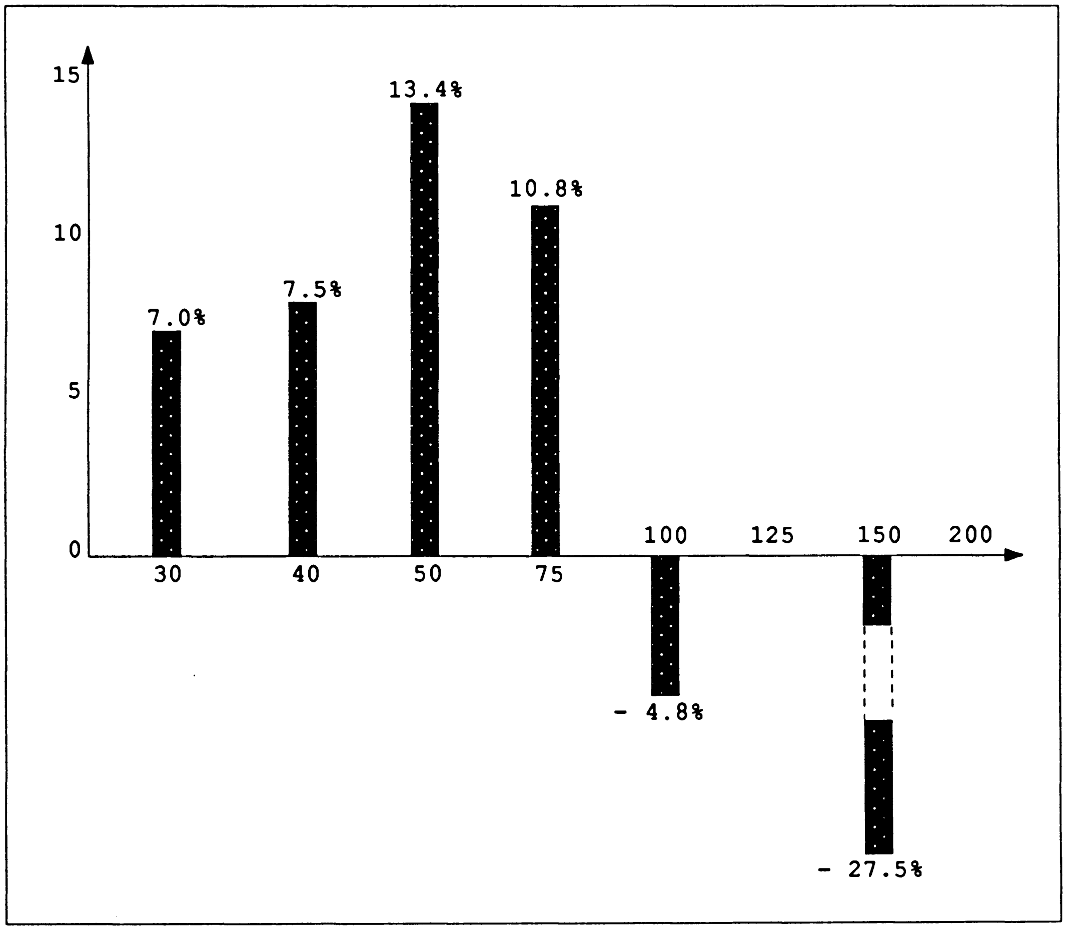
Table 11: Mixed Output Consisting of Full Documents and Document Sections

```
Web starting from 6944 Delian League
|
|
+---6957         0.48 Delos, Confederation of.
|    +---6956         0.59 Delos.
|    |   +---6957          0.59 Delos, Confederation of.
|    |   +---6944          0.38 Delian League.
|    |   +---6587          0.37 Cyclades.
|    +---6944         0.48 Delian League.
|    +---5965         0.35 Confederation.
|            +---16786      0.57 North German Confederation.
|        +---8623          0.46 Federal Government.
|         +---5967          0.43 Confederation of the Rhine.
|
+---1655         0.43 Athens.
|   +---17824        0.55 Pericles.
|   |   +---1575          0.61 Aspasia.
|   |   +---1655          0.55 Athens.
|   |   +---17983         0.50 Phidias.
|   +---18200        0.47 Pisistratus.
|   |   +---1655          0.47 Athens.
|   |   +---11290         0.45 Hippias.
|   |   +---17824         0.33 Pericles.
|   +---1651         0.47 Athena.
|       +---17591         0.50 Parthenon.
|       +---1655          0.47 Athens.
|       +---4779          0.44 Crecrops.
|
+ ---10300       0.41 Greece.
    +---1655         0.46 Athens.
    +---15390        0.42 Metaxas, Ioannes.
    |   +---10300         0.42 Greece.
    |   +---9743          0.37 George II.
    |   +---12266         0.18 Ithaki.
    +---6944         0.41 Delian League.
```

Table 12: Linked Hypertext Web for Document 6944 "Delian League" (depth 3 - breadth 3 search; at least one matching sentence pair required at threshold 75.0)



a) Paragraph Matches at Indicated Thresholds (improvement in average precision at three recall points)



b) Sentence Matches at Indicated Thresholds (improvement/deterioration in average precision at three recall points)

Figure 1: Evaluation of Local Sentence and Paragraph Matching in Terms of Average Precision at Recall 0.20, 0.50, 0.80 for 355 Queries

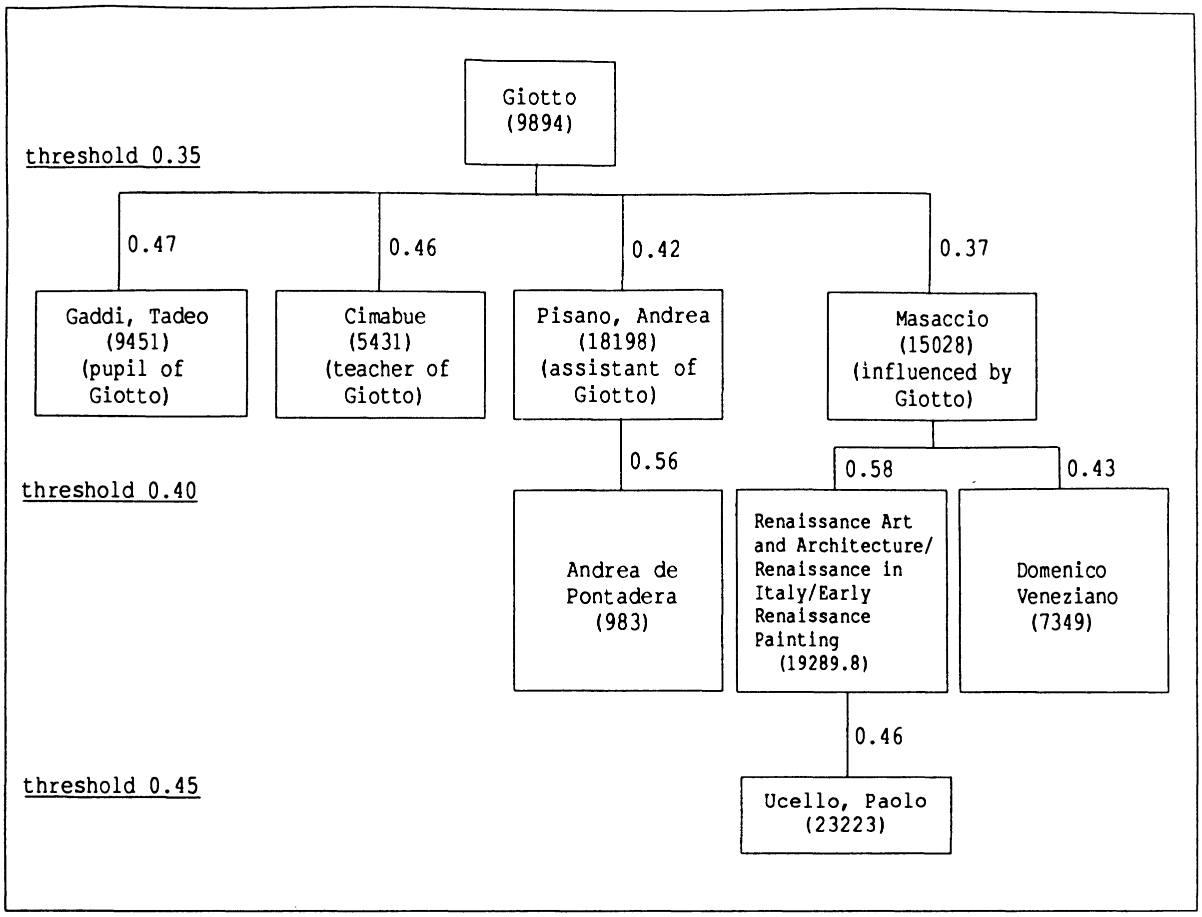
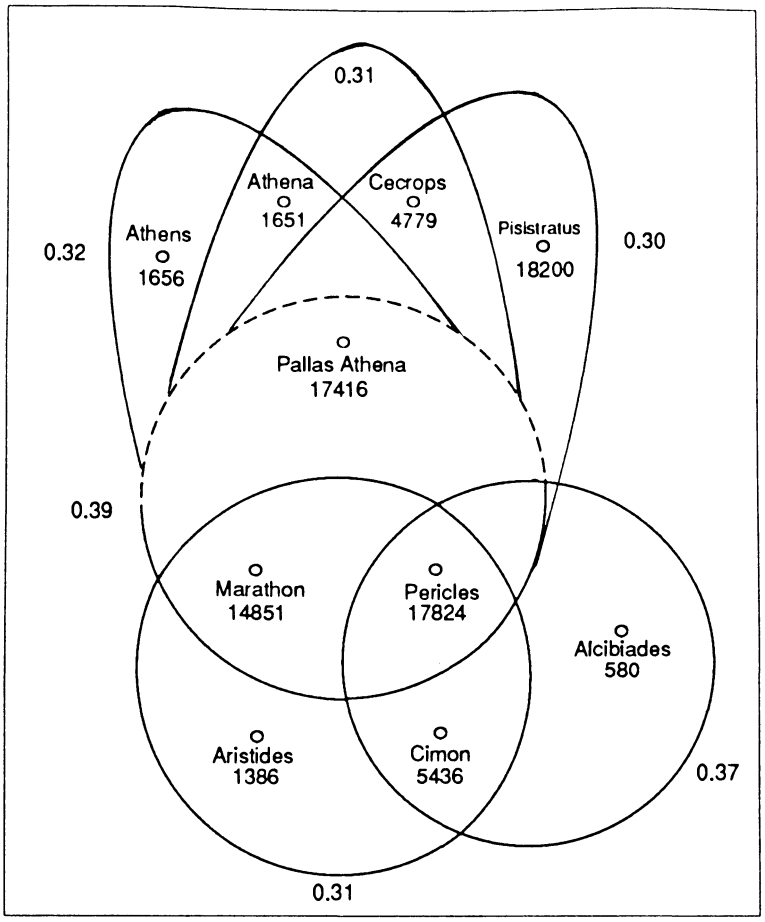


Figure 2: Automatic Map of Related Items for Document 9894 "Giotto"



a) Complete Link Clusters Including Document 17824 "Pericles"

	Documer Number	Similarity	Title	
	6944	1.00	Delian League.	
	1653	0.66	Athenian League.	
	6957	0.48	Delos, Confederation of.	
	1655	0.43	Athens.	
	13673	0.43	League of Nations.	
	10300	0.41	Greece.	
	17416	0.40	Pallas Athena.	
	17824	0.39	Pericles.	
	6956	0.38	Delos.	0.32
0.31	1100	0.36	Antalcidas.	0.39
	1386	0.35	Aristides.	
	14851	0.34	Marathon.	
	5436	0.33	Cimon.	
	1651	0.33	Athena.	
	23373	0.33	Urban League.	

b) Unrestricted Search Output for Query 6944 "Delian League" with some identified cluster structures

Figure 3: Search Results with Identified Cluster Structures