Automatic Text Structuring and Retrieval - Experiments in Automatic Encyclopedia Searching

Gerard Salton*
Chris Buckley

TR 91-1196
April 1991

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 Text Structuring and Retrieval – Experiments in Automatic Encyclopedia Searching

Gerard Salton* and Chris Buckley

Abstract

Many conventional approaches to text analysis and information retrieval prove ineffective when large text collections must be processed in heterogeneous subject areas. An alternative text manipulation system is outlined useful for the retrieval of large heterogeneous texts, and for the recognition of content similarities between text excerpts, based on flexible text matching procedures carried out in several contexts of different scope. The methods are illustrated by search experiments performed with the 29-volume Funk and Wagnalls encyclopedia.

1 Approaches to Text Retrieval

We are addressing the problem of text retrieval from large, heterogeneous text databases, where the vocabulary varies widely and the subject matter is unrestricted. This includes newspaper articles, newswire dispatches, textbooks, dictionaries and encyclopedias, manuals, magazine articles, and so on. Necessarily, the retrieval operations must be preceded by some form of text analysis, or text indexing operation, leading to the assignment to query and document texts of appropriate content identifications. When content identifiers are attached to text items, the text retrieval operations can be carried out by comparing the identifications assigned to the various text items with the query identifications, and retrieving texts that appear sufficiently similar to the corresponding queries.[1-2]

*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.

In discussing the text analysis operation, it is customary to contrast the so-called *keyword* approach with more sophisticated procedures based in part on linguistic and logical considerations. In the keyword approach, single-term concepts, known as keywords, index terms, or simply terms, are chosen manually or automatically for the content identification of the various texts. In some environments, the terms may be weighted in decreasing order of presumed importance for content identification purposes. A typical document D_i may then be represented by a *term vector* as $D_i = (w_{i1}, w_{i2}, ..., w_{it})$, where t terms in all are assumed available for content representation, and w_{ik} represents the weight term T_k in D_i. A zero-weight may be assumed for terms absent from particular documents, while positive weights are used for terms actually present in a particular vector.

It is widely believed that the keyword approach is not adequate for text content representation in information retrieval. By extension, the identification of text content by weighted term sets may also be unacceptable. Quoting from Blair:[3]

> "No number of brute linguistic facts (word statistics) can be added up to give us the meaning of a text, where the meaning of a text would include such things as its subject, intellectual content, context, use, purpose, or links to other documents."

Accordingly attempts are made to refine the text indexing process by using more refined linguistic and logical text analysis methods to obtain the needed text content representations. Among the possibilities that have been explored are the use of *syntactic analysis* systems designed to identify multi-word nominal constructs (noun phrases) for assignment to the texts replacing the standard single-term keywords. The available experience with the use of automatic syntactic analysis indicates that such are systems not sufficiently powerful to generate large numbers of correct content identifiers in the absence of substantial semantic know-how, and overall one finds that noun constructions are more reliably obtainable by using statistical methods based on detectable co-occurrences between phrase components in the available texts. [4,5]

The missing semantic component might in principle be supplied by consulting dictionaries and thesauruses, and extracting from them the needed semantic specifications.[6] Thesauruses of many kinds have been constructed, often tailored to particular topic areas, and designed to reveal semantic relationship between thesaurus entries. However, in practice, thesauruses and machine-readable dictionaries are not easily incorporated into operational retrieval situations. The construction and use of vocabulary specification tools is an art, and there is no guarantee that thesauruses tailored to particular collections can be usefully adapted outside their original environment. In any case, it has not been possible to obtain reliable improvements in retrieval effectiveness by using preconstructed vocabulary schedules in heterogeneous text environments.

A further extension in the sophistication of the text analysis is provided by the *knowledge-based* approaches that are popular in artificial intelligence. Many formalism have been proposed for the representation of knowledge, such as semantic nets, frames, scripts, and so on.[7-9] Network structures are often used, in which case the concepts of interest in a subject area are represented by network nodes, and the main relationships between concepts by network branches. In information retrieval, the knowledge bases are normally used with *inference rules* that specify how the concept network must be traversed in attempting to relate the document representations with the query identifiers. Among the techniques used to derive query responses from the available document identifications are spreading activation methods [10], Bayesian inference techniques [11], Dempster-Shafer belief theory [12], and others.

Substantial advantages have been claimed for the incorporation of knowledge representations and inference rules in various text processing applications, and it is likely that useful knowledge structures and reliable inference rules can in fact be generated that may be valid in well-circumscribed situations within limited subject domains. However, substantial doubt remains about the viability of techniques based on complex network representations when large text collections must be processed in unrestricted subject areas. Eco and others have argued in this connection that any artificially constructed knowledge structure necessarily provides only partial and inadequate representations of meanings: "Semantic models such as that of Quillian are already ..... a portion of the universe in which a system has intervened in order to establish attractions and repulsions (between concepts, thereby favoring some concept relationships at the expense of others)."[13]

In any case, complete theories of knowledge representation do not exist at the present time. As a result, it is not clear what entities must be included in a knowledge base, and what relationships between knowledge elements must be considered for particular applications. Attempts have been made to build very large knowledge bases covering unrestricted subject matter, but the applicability to large unrestricted document collections is unproven.[14]

Instead of using knowledge-based methods that are not well understood, or deep linguistic analysis systems of doubtful applicability in unrestricted text environments, a sophisticated text comparison system is considered that is carried out at various levels of detail in the text material to determine the similarities between text items.

2 Text Matching System

We approach the text analysis problem, by taking advice from Wittgenstein and others who suggest that text understanding must be based on a study of how text words are used in the language.[15]

> "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."

In determining how text words are used, we are not primarily interested in deriving detailed descriptions of text content, but instead an attempt is made to recognize text portions within which text meanings are homogeneous (that is, sufficiently similar to conclude that the texts are closely related). An identification of semantically homogeneous text excepts leads to the generation of text links between related text portions. Such links transform linear texts into structured text representations that provide selective text reading and traversal paths by following the available content links. In addition, a recognition of semantically related text portions also leads to the retrieval of relevant texts in answer to available search requests, because close similarities between document and query texts may be indicative of a relevance relationship between them.

In ascertaining the semantic homogeneity between texts in unrestricted text environments, we use the texts themselves as the principal basis for the text analysis. However, because of the well-known syntactic and semantic ambiguities inherent in natural language texts, a simple comparison between the words occurring in different texts will not be adequate in all cases. In the extensive paragraph matching experiments carried out for the 1140 paragraphs of a recent textbook in information science, substantial differences in meaning could be detected for certain paragraph pairs even when substantially overlapping vocabularies were present.[18] This occurred, for example, for paragraphs 209 and 930, both covering the use of scripts in text processing: in the first instance (paragraph 209 in chapter 4), the term "script" referred to a text editing language called Script, whereas in paragraph 930 of chapter 12, the same term was used to refer to a well-known knowledge-representation model.[19]

Remembering the injunction of Wittgenstein that a study of word use is necessary to understand text meaning, a detailed examination was made of the contexts in which terms such as "script" and "text" occurred in the sample paragraphs. This revealed substantial differences in the occurrence characteristics of the overlapping vocabularies in the sample texts. In paragraph 209, "script" was used in the context of "text editor", "formatting system" and "script commands", whereas the characterization in paragraph 930 included expressions such as "stored script", "automobile-accident script" and "script information".

A text analysis process is then suggested for the identification of semantically homogeneous text excerpts that makes use of several text comparison operations in environments of different scope:

1. Each available text (including query as well as document texts) is broken down into individual text units — for example, text sections, text paragraphs, and individual sentences.

2. A standard indexing system is used to assign to each text unit (that is, section, paragraph, sentence, etc.) a set of weighted terms to be used for content identification of the corresponding text fragment. These term vectors form the basis for the text comparison operations.

3. Similarities between particular text items (or between text items and information requests) are obtained by comparing the term vectors for pairs of text items at various levels of detail. When sufficient similarities are detected in both global as well as local contexts, the texts are assumed to be related.

In practice, two text sections might then be related when the similarity between the vectors describing the text sections exceeds a stated threshold, and in addition the sections also contain at least one paragraph pair with sufficiently large paragraph similarity. Analogously, paragraphs may be related when the global paragraph similarity is sufficiently large, and additionally sentence pairs are present with sufficiently large sentence similarities.

In using this text matching strategy, the assumption is made that the text meanings are indeed related when a sufficiently large overlap exists in the global vocabularies, and the overlapping terms occur in reasonably similar local environments. This conjecture was confirmed in the earlier study performed with textbook paragraphs, where the most favorable linking strategy based on paragraph and sentence similarity computations produced a linking accuracy of 98 percent; that is, a meaningful semantic relationship was detected for 98 percent of the linked paragraph pairs.[19] In retrieval environments where query texts must be linked to relevant document texts, this translates into a retrieval precision of 98 percent.

The recall performance is much more difficult to estimate because it is manifestly impossible to obtain accurate recall bases when hundreds of millions of text pairs are routinely processed. Nevertheless, it is not unreasonable to suppose that a particular subject cannot be discussed in depth without using vocabulary patterns pertinent to that area, and such patterns are detectable by using sufficiently sophisticated text matching operations. The examples often cited of totally dissimilar ways of expressing identical notions (manufacture of spaghetti, production of pasta, making of ravioli pasta, etc.) are applicable in the small. When a subject is covered in some detail, substantial overlap in coverage must be expected. In these circumstances, it is difficulty to believe that large numbers of obvious similarities in subject coverage will remain undetected when the proposed multi-level text matching system is used, and a reasonable recall performance should be expected in addition to superior search precision.

The similarity computations between text representation must depend on the types and weights of the coinciding terms in the respective term vectors. In retrieval, the most valuable terms for document content representation are those best able to distinguish particular texts from the remainder of the collection. This implies that the term weighting system should assign low weights to high-frequency terms that occur in many documents of a collection, and high weights to terms that are important in particular documents, but unimportant in the remainder of the collection. The weight of terms that occur rarely in a collection is of no consequence, because such terms contribute little to the text similarity.

A well-known term weighting system following that prescription assigns weight w_{ik} to term T_k in document D_i in proportion to the frequency of occurrence of a term in D_i, and in inverse proportion to the number of documents to which the term is assigned.[20] When the texts, are represented by term vectors, the similarity (sim) between two items D_i and D_j is conveniently obtained as the inner product between corresponding vectors, or $sim(D_i, D_j) = \sum_{k=1}^{t} w_{ik} \cdot w_{jk}$.

In practice it is found that the document length, and hence the number on non-zero term weights assigned to a document, varies widely. The same is true of the values of the occurrence frequencies, f_{ik}, of terms T_k in D_i. To give each text item an equal chance of being retrieved, it is then convenient to use an enhanced term weighting formula by normalizing for document length. This produces the so-called *atc* term weight:

$$w_{ik}\,(atc) = \frac{\left(0.5 + 0.5\,\frac{f_{ik}}{\max_p f_{ip}}\right)\left(\log \frac{N}{n_k}\right)}{\sqrt{\sum_{k=1}^{t} \left(0.5 + 0.5\,\frac{f_{ik}}{\max_p f_{ip}}\right)^2 \left(\log \frac{N}{n_k}\right)^2}} \qquad (1)$$

where w_{ik} is the weight of term T_k in document D_i, f_{ik} is the occurrence frequency of T_k in D_i, N represents the number of documents in the collection, n_k the number of documents with term T_k, and the maximization operation is extended over all terms in a given vector. The weights of expression (1) range from 0 to 1, and the values of the corresponding inner product similarity function between text items depends on the proportion of matching terms in the two vectors.

For short text excerpts, such as text sentences, a text similarity measure that depends on the proportion of matching terms is not indicative of coincidence in text meaning. For example fragments such as "consider figure x", or "equation y is an example", may be repeated many times in ordinary texts, but the vocabulary coincidences in such fragments are not meaningful. Short texts are therefore compared using an *atn* term weight, equivalent to the numerator of expression (1) without the length normalization of the denominator. The *atn* term weight ranges in size from 0 to *log N*, and the corresponding inner product text similarity depends on the number (rather than the proportion) of matching terms. The experimental output included in this study uses the insights gained in earlier text matching experiments in that *atn* similarity measures serve for comparisons involving full documents, and *atn* measures for paragraph and sentence comparisons.[19]

3 Encyclopedia Searching

An automated version of the 29-volume Funk and Wagnalls (FW) encyclopedia is used to perform the experiments.[21] This database consists of about 24,900 articles of text ranging in scope from simple cross-references ("Abo, see Turku"), to long articles that include dozens of subsections and hundreds of paragraphs ("United States of America"). An automated encyclopedia search system is implemented which uses particular encyclopedia articles as search requests, and retrieves related articles in decreasing order of presumed similarity with the request articles. A *single-level* search can be implemented by comparing the query article with all other articles in the encyclopedia and producing a single list of retrieved items in decreasing order of the query-document similarity. Alternatively, a *multi-level* search may be performed which retrieves only a few items with especially large query similarities in the initial search, and uses these newly retrieved items as secondary queries to retrieve additional output. When the initial query is short, as is the case for cross-reference articles, the multi-level search appears mandatory, because very few matching texts will be found when the original query statement is limited to a few words only.

When the initial query is long, possibly consisting of several subsections and many pages of text, the standard single-level or multi-level searches will not produce the expected results. Long texts are not normally concentrated in narrow subject areas, but cover many different aspects of particular topic areas. In such cases, the global text matching system often retrieves a heterogeneous collection of output texts, each related to a different aspect of the input. Furthermore, the *atc* similarity measure which depends on the proportion of matching terms between texts will necessarily be small when long texts are characterized by large numbers of non-zero terms. Larger texts are therefore most easily processed by subdividing them into shorter units before the text comparison system is used. One possibility is to define subdocuments consisting of the various subsections of text into which these long encyclopedia articles are subdivided. Each subdocument is then used as a separate query and the outputs obtained with the several subdocuments are appropriately combined.

The use of an encyclopedia as a search tool raises the usual problems relating to search evaluation. In principle, the retrieved items can be individually examined for relevance to each search query. Experience indicates that the overwhelming number of retrieved items will then be found to be germane to the respective query articles, assuming that the query-document similarities are sufficiently large. Such a subjective precision computation is however tedious, and its accuracy is subject to challenge. It is certainly more convenient in actual practice to use objective relevance assessments between queries and documents, that provide more credible retrieval precision computations.

For the encyclopedia implementation, a possible solution to the document relevance problem consists in using the available cross-references between articles as indicators of relevance. Three kinds of references between articles are available in the encyclopedia, consisting of "see", "see also", and "qv" (quod vide) references. The reasonable assumption can be made that when one of these references is present citing article B within, or after, the text of article A, then B is assumed to be relevant to query article A. Two problems arise in using the cross-references as the main relevance criteria: many of the articles that are cited as references are long articles, covering broad subject matter, and these articles are difficult to retrieve using standard *atc* similarity measures, as previously mentioned. More seriously, the FW encyclopedia contains very few references. Among the more than 310 million article-pairs, only 80,000 cross-references are actually included in the text, and FW editors are specifically warned to be selective in using the cross-reference tool in order not to clutter up the printed material. This results in an unfortunate underestimation of the power of the automatic encyclopedia search system when objective relevance assessments are used, because dozens of relevant articles are likely to exist for each search article of which only very few are cross-referenced in the published text.

A standard encyclopedia search for a one-paragraph query (document 114, Acacia) is illustrated in Table 1. A multilevel search strategy is used where all articles with a global query similarity exceeding 0.20 are retrieved initially. The retrieved items (documents 15552, Mimosa and 11949, Indigo Plant) are used as queries for second-level searches with an increased retrieval threshold of 0.25. Query 15552 (Mimosa) retrieves additional documents 12666 (Judas Tree) and 20683 (Sensitive Plant) on level 2, and query 11949 (Indigo Plant) retrieves documents 20679 (Senna) and 3487 (Brazilwood). When the four second-level documents are used in turn for third level searches with a retrieval threshold of 0.30, no additional items are retrievable on level 3. The six retrieved items are all plants in the legume family, related to the original query (114, Acacia). Presumably, these documents would be of interest to a user submitting the Acacian query. However, no formal cross-references exist between document 114 (Acacia) and any of the retrieval items.

The multi-level search illustrated at the top of Table 1 may be compared with a single-level search that retrieves six items in a single search operation carried out with the original query alone. The output shown in the lower part of Table 1 shows that the single-level search actually retrieves the only formally relevant document (13736, legume) with an *atc* similarity of 0.1969. (The asterisk in Table 1 confirms the existence of a formal cross-reference in the encyclopedia between documents 114 (Acacia) and 13736 (legume)). Because the legume article is long (six paragraphs of text), the *atc* similarity with the query falls below the retrieval threshold of 0.20 used for the multi-level search.

Table 2 presents the search results for a short cross-reference query (document 75, Abo, see Turku). Five items are retrieved using a multi-level search with thresholds of 0.20, 0.25, and 0.30 for the first three search levels. The retrieved items are all Finnish cities related to Abo. (Abo is the Swedish name of the Finnish city Turku – hence the formal cross-reference between the query article 75 and document 23149, Turku). Document 23149 is retrieved at the top of the ranked list with a query similarity of 0.5058. The other retrieved items are also relevant, but they lack the formal cross-references.

When a single-level search is used to retrieve the top five documents in response to query 75 (Abo), several of the retrieved items exhibit relatively low query similarities (0.1757 for the document on Fredrika Bremer and 0.1548 for item 19809 on Johan Ludvig Runeberg). The two personages are writers with close ties to Abo/Turku, but the relevance to the query article Abo may be considered to be somewhat remote.

A search conducted with a long query (article 8709, Feudalism) containing 14 paragraphs of text is illustrated in Table 3. The single-level output presented in column one of the table consists in part of articles covering geographic regions with some role during the feudal period (England, France, Germany, Scotland, Europe). While a relationship exists with document 8709 (Feudalism), the main thrust of these geographic articles lies elsewhere. Documents covering Feudalism more directly can be retrieved by breaking up the long query-article into shorter subqueries. Columns 2 and 3 of Table 3 cover the search output obtained by using the first two paragraphs (numbers 44055 and 44057) of the query article 8709 for search purposes. The items retrieved in response to query-paragraph 44055 are all directly connected with the feudal system (fief, homage, medieval period, seignorialism, manor, serfdom) but a formal cross-reference from item 8709 exists only for document 20632 (seignorialism).

An illustration of the usefulness of sentence matching within matching documents appears in Table 4. The search query is a one-paragraph document (document 89, Absalom). Four documents are retrieved with a similarity exceeding 0.20. Of these, two are immediately relevant (423, Ahitophel and 65, Abiathar) — the two corresponding biblical characters are closely related to Absalom, a son of king David of Israel. The relevance of the two other retrieved items (73, Abner and 12616, Jonathan) is more remote, the connection to Absalom being established via King David.

When a sentence-comparison system is used between the sentences of the query and the four retrieved items, the number of matching sentence pairs above an *atn* threshold of 100.0 is 4, 1, 3, and 0 for documents 423, 73, 65, and 12616, respectively. When at least one matching sentence-pair is required for retrieval, one of the two items of questionable relevance (12616, Jonathan) can be rejected. When two matching sentence pairs are needed, the other questionable items (73, Abner) is also rejectable, leaving the two directly relevant items as the only ones that are retrieved.

It may happen that a suitable encyclopedia article expressing a user's interests is not initially available. In that case, an encyclopedia search may be conducted in two iterations, by first entering a free-form English query that is compared with all encyclopedia articles. When the corresponding output is available, one or more of the previously retrieved encyclopedia articles can then in turn serve as a new query for subsequent searches. Alternatively, a relevance feedback technique can be used to produce a new query automatically, based on relevance judgments provided to the search system about one or more of the previously retrieved items.[22]

An illustration of a search performed with a free-form query is shown on Table 5, where the original search retrieves 7 items, including the main article on "Richelieu, Armand Jean du Plessis, Duc de". When the latter (article 19437) is used for search purposes, additional relevant materials not previously seen are retrieved, such as document 14902 (Marie de Medicis).

The search output of Tables 1 to 5 is illustrative of the operation of the encyclopedia search system. Several partial multi-level searches are best used for short queries consisting of a simple cross-reference to other articles. Long query articles consisting of many text paragraphs are best broken down into more focused parts by using the individual paragraphs as subqueries, and combining the respective search results. Conventional single-level and multi-level searches can be used for standard query articles covering a topic in one or two text paragraphs. In all cases, substantially enhanced search precision is obtainable by using searches of restricted scope (paragraphs within matching documents, or sentences within matching paragraphs) to restrict the output produced by the original wider-scope searches to items that may be more immediately useful.

The behavior of the local paragraph and sentence matching system used in addition to the normal comparison between complete encyclopedia articles is evaluated by using the first 1,000 encyclopedia articles as search requests, and exhibiting the corresponding average recall and precision results. In each case, the relevance of documents with respect to the queries is objectively determined by using as relevance indicators the available cross-references between encyclopedia articles. As explained earlier, very few cross-references are included in the encyclopedia, so that very few retrieved items are rated as relevant. The computed precision values are thus necessarily vastly understated.

Table 6 contains the evaluation results for a retrieval threshold of 15 retrieved documents per query. Analogous results are given in Table 7 for 5 retrieved items per query. Of the 1,000 articles used as queries, 489 did not carry any attached cross-references; for these items no formally relevant documents are included in the collection. The averages for the standard run without extra paragraph or sentence matches are therefore computed for the 511 queries with relevant documents. The second column of Tables 6 and 7 shows that somewhat fewer than 511 queries are used for the complex runs that include matching paragraphs or sentences, because for some short cross-reference queries, no paragraph or sentence matches could be identified.

The output for the base run of Table 6 (0.1235 in precision and 0.6710 in recall) reveals a high degree of effectiveness even without additional paragraph or sentence matching. For the 511 queries under consideration, a precision ceiling of 0.2386 exists because over 7,000 documents must be retrieved overall with the 15 document threshold, while only 1,852 relevant documents in all are defined for these queries. The entries in columns 3 and 4 of Tables 6 and 7 illustrate the effect of the additional local paragraph and sentence matching: the total number of retrieved documents necessarily decreases when additional matching requirements are imposed, because a local paragraph or sentence match will not occur for some of the retrieved documents; however, the number of relevant retrieved items remains nearly constant. When the number of retrieved relevant items remains nearly constant while the total number of retrieved items decreases, the precision will increase, and so may the recall. Requiring at least two paragraphs matches between queries and retrieved documents above an *atn* threshold of 75.0 increases the precision by 88.7 percent and the recall by 6.9 percent for the output of Table 6. For two sentence matches above *atn* threshold of 100.0, the performance improvement is 53% in precision and 7.1 percent in recall.

When *atn* (or *atc*) term weights are used, the sentence similarity coefficients are normally larger than the corresponding paragraph similarity values. The reason is the much larger number of sentences (409,572) present in the collection, compared with the number of paragraphs (130,038). The inverse document frequency component of the term weight ($logN/n_k$ in expression (1)) is therefore larger for sentences than for paragraphs. This accounts for the fact that a sentence similarity threshold of 100.0 is used for the computations of Tables 6 and 7, while a smaller similarity threshold of 75.0 serves for the paragraph matches. These two thresholds are chosen so as to equalize (approximately) both the number of queries used in the retrieval operations, and the number of relevant retrieved documents.

A study of the output of Tables 6 and 7 reveals that the paragraph match is relatively more powerful in improving retrieval performance than the sentence match, although the latter operates quite well also.

Furthermore, the increase in performance is relatively higher when two paragraph matches are required instead of one, than it is when two sentence matches are used instead of one. The paragraph and sentence matches are carried out independently in the output of Tables 6 and 7. It remains to be seen whether additional improvements are possible when sentence matches are used within matching paragraph.

A comparison of Tables 6 and 7 for retrieval thresholds of 15 and 5 documents per query, respectively, shows that, as expected, higher recall and lower precision are obtained as more items are retrieved. It may be noted that even at the higher retrieval threshold used in Table 6, not all relevant items included in the collection are retrievable. The reason may be the previously mentioned inordinate length of some of the relevant articles. Long articles are not retrievable in the top 15 ranks when *atc* term weights are used; hence these articles are never subjected to local paragraph or sentence matches. This problem is likely to be solvable by subdividing the long articles, as previously shown in the example of Table 3.

In the experiments carried out in this study, the same text processing method was used for all encyclopedia articles. In an operational encyclopedia retrieval service, distinctions would need to be introduced between different article types. A text matching system useful for simple cross-reference entries will necessarily differ somewhat from a procedure used for long texts describing heterogeneous subjects. Nevertheless, the present experiments demonstrate that the flexible text matching systems operating at various levels of text granularity provide powerful tools for the identification of text relationships, and for the retrieval of relevant texts in answer to user queries. No equally effective approach appears on the horizon for handling heterogeneous texts covering arbitrary subject areas.

4 Acknowledgment

The writers are grateful to the Microsoft Corporation for making available a machine-readable version of the Funk and Wagnalls New Encyclopedia for experimental use at Cornell University.

References and Notes

[1] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill Book Co., New York, 1983.

[2] C.J. Van Rijsbergen, *Information Retrieval*, Second Edition, Butterworths, London, 1979.

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

[4] J. Fagan, The Effectiveness of a Nonsyntactic Approach to Automatic Phrase Indexing for Document Retrieval, *Journal of the ASIS*, 40:2, March 1989, 115-132.

[5] G. Salton, C. Buckley and M. Smith, On the Application of Syntactic Methodologies in Automatic Text Analysis, *Information Processing and Management*, 26:1, 1990, 73-92.

[6] R.A. Amsler, Machine-Readable Dictionaries, in *Annual Review of Information Science and Technology*, 19, M.E. Williams, editor, Knowledge Industry Publications, Inc. White Plains, NY, 1984, 161-209.

[7] M.R. Quillian, Semantic Memory, in *Semantic Information Processing*, M. Minsky, editor, MIT Press, Cambridge, MA, 1968, 216-270.

[8] N. Findler, editor, *Association Networks: The Representation and Use of Knowledge by Computers*, Academic Press, New York, 1979.

[9] R. Brachman and H. Levesque, *Readings in Knowledge Representations*, Morgan-Kaufmann, Los Altos, CA, 1985.

[10] P.R. Cohen and R. Kjeldsen, Information Retrieval by Constrained Spreading Activation in Semantic Networks, *Information Processing and Management*, 23:4, 1987, 255-268.

[11] H. Turtle and W.B. Croft, Inference Networks for Document Retrieval, *Proc. Thirteenth Annual ACM/SIGIR Conference for Research and Development in Information Retrieval*, Brussels, Sept. 1990, 1-24.

[12] G. Bismas, J.C. Bezdek, V. Subramanian, and M. Marques, Knowledge Assisted Document Retrieval, *Journal of the ASIS*, 38:2, March 1987, 83-110.

[13] U. Eco, *A Theory of Semiotics*, Indiana Univ. Press, Bloomington, IN, 1976, p. 125.

[14] D. Lenat, M. Prakash, and M. Shepherd, CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks, *AI Magazine*, 6, Winter 1986, 65-84.

[15] L. Wittgenstein, *Philosophical Investigations*, Basil Blackwell, Oxford Univ. Press, 1953, Paragraph 43.

[16] J. Conklin, Hypertext: An Introduction and Survey, *Computer*, 20:9, September 1987, 17-41.

[17] M.E. Frisse, Searching for Information in a Hypertext Medical Handbook, *Communications of the ACM*, 31:7, July 1988, 880-886.

[18] G. Salton, *Automatic Text Processing — The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley Publishing Co., Reading MA, 1989.

[19] G. Salton and C. Buckley, Flexible Text Matching in Information Retrieval, Technical Report 90-1158, Dept. of Computer Science, Cornell University, September 1990.

[20] G. Salton and C. Buckley, Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24:5, 1988, 513-523.

[21] *Funk and Wagnalls New Encyclopedia*, Funk and Wagnalls, New York, 1979, 29 volumes.

[22] G. Salton and C. Buckley, Improving Retrieval Performance by Relevance Feedback, *Journal of the ASIS*, 41:4, 1990, 288-297.

One-Paragraph Query : Doc 114 Acacia				
Retrieved Items Level 1 (threshold 0.20)		Retrieved Items Level 2 (threshold 0.25)		Retrieved Items Level 3 (threshold 0.30)
15552	Mimosa (0.2517)	12666	Judas Tree (0.2544)	—
		20683	Sensitive Plant (0.2540)	—
11949	Indigo Plant (0.2152)	20679	Senna (0.2575)	—
		3487	Brazilwood (0.2543)	—
13736	Legume (0.1969 R)*			
13831	Lespedeza (0.1901)			
3487	Brazilwood (0.1844)			
17681	Pea (0.1787)			

Table 1: Multi-Level and Single-Level Search for One-Paragraph Query

Cross-Reference Query : Doc 75 Abo (Turku)				
	Retrieved Items Level 1 (threshold 0.20)		Retrieved Items Level 2 (threshold 0.25)	Retrieved Items Level 3 (threshold 0.30)
23149	Turku (0.5058 R)*	11034	Helsinki (0.2968)	—
18475	Pori (0.2174)	2898	Bjorneborg (Pori) (0.5172)	—
		17304	Oulu (0.2678)	—
3514	Fredrika Bremer (0.1757)			
19809	John Ludvig Runeberg (0.1578)			
11034	Helsinki (0.1238)			

Table 2: Multi-Level Search for Cross-Reference Query

Multi-Paragraph Query : Doc 8709 Feudalism (14 paragraphs)					
Retrieved Items Single-Level (complete document)		Retrieved Items First Paragraph 44055 (threshold 5.0)		Retrieved Items Second Paragraph 44057 (Feudalism - Origins)	
20632	Seignorialism (0.2120 R)*	8709	Feudalism (10.02 R)	8709	Feudalism (10.01 R)
8108	England (0.2059)	8729	Fief (8.23 R)	14993	Charles Martel (6.62)
13911	Liberty (0.1974)	11399	Homage (8.19 R)	11087	Henri I (5.86)
16717	Nobility (0.1928 R)	15222	Medieval Period (5.66 R)	5280	Chivalry (5.23 R)*
1436	Army (0.1913)	20632	Seignorialism (5.48 R)*		
9183	France (0.1865)	14796	Manor (5.45 R)		
23933	Warfare (0.1849)	20709	Serfdom (5.00 R)		
9786	Germany (0.1826)				
20919	Scotland (0.1818)				
9349	Europe (0.1803)				

Table 3: Illustration of Paragraph-Search for Long Documents

One Paragraph Query : Doc 89 Absalom				
Retrieved Items (threshold 0.20)		Matching Sentences between Query and Retrieved Document (threshold 100.0)		
		Sentence Pair	Matchings Terms	Sentence Similarity
423	Ahitophel (0.3580 R)	818-6655	absalom	110.795
		818-6656	absalom	110.795
		820-6656	absalom kill sam	228.542
73	Abner (0.2503 N)	822-589	joab	149.589
65	Abiathan (0.2450 R)	820-526	absalom sam	142.970
		821-527	david king rebel	117.206
		823-526	absalom david son	113.866
12616	Jonathan (0.2427 N)	—	—	—

Table 4: Illustration of Sentence-Matching within Matching Documents

Original free form query: What is known about Cardinal Richelieu, his role in French history, his relationship with French kings, and his role in European wars such as the 30 years' war.			
	Retrieved Items (threshold 0.20)		Using Encyclopedia Item 19437 as query (threshold 0.20)
14275	Louis XIII (0.2520)	14275	Louis XIII (0.3142)
19437	Richelieu, Armand Jean du Plessis. Duc de (0.2271)	14902	Marie de Medicis (0.2805)
		13567	La Rochelle (0.2460)
12639	Joseph, Father (0.2170)	19436	Richelieu (0.2387)
17202	Orleans, Gaston, Duc d'	11619	Huguenots (0.2298)
20001	St-Jean-sur Richelieu (0.2084)	11070	Maria Henrietta (0.2236)
19436	Richelieu (0.2077)		
1078	Anne of Austria (0.2000)		

Table 5: Illustration of Encyclopedia Search with Free-Form Query

	Number of Queries Retrieving Documents	Total Number of Retrieved Documents	Total Number of Relevant Retrieved	Average Precision	Average Recall
Standard Query - Document Comparison	511	7485	905	0.1235	0.6710
Query Document Match incl. One Paragraph Match (75.0 threshold)	475	5941	910	0.2079 (+68.3%)	0.7074 (+5.4%)
Query - Document Match incl. Two Paragraph Matches	454	5012	913	0.2330 (+88.7%)	0.7174 (+6.9%)
Query - Document Match incl. One Sentence Match (100.0 threshold)	482	6261	916	0.1769 (+43.2%)	0.7094 (+5.4%)
Query - Document Match incl. Two Sentence Matches	462	5722	921	0.1889 (+53.0%)	0.7185 (+7.1%)

Table 6: Recall - Precision Results for Encyclopedia Searches
(15 retrieved documents per query, about 500 queries)

	Number of Queries Retrieving Documents	Total Number of Retrieved Documents	Total Number of Relevant Retrieved	Average Precision	Average Recall
Standard Query - Document Comparison	511	2542	597	0.2357	0.5200
Query Document Match incl. One Paragraph Match (75.0 threshold)	471	2151	600	0.3087 (+31.0%)	0.5633 (+8.3%)
Query - Document Match incl. Two Paragraph Matches	450	1974	600	0.3326 (+41.1%)	0.5743 (+10.4%)
Query - Document Match incl. One Sentence Match (100.0 threshold)	481	2256	599	0.2821 (+19.7%)	0.5542 (+6.6%)
Query - Document Match incl. Two Sentence Matches	461	2136	585	0.2888 (+22.5%)	0.5468 (+5.2%)

Table 7: Recall - Precision Results for Encyclopedia Searches
(5 retrieved documents per query, about 500 queries)