Automatic Structuring of Text Files*

Gerard Salton*
Chris Buckley
James Allan

TR 91-1241
October 1991

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 of Text Files

Gerard Salton,* Chris Buckley and James Allan.

Abstract

In many practical information retrieval situations, it is necessary to process heterogeneous text databases that vary greatly in scope and coverage, and deal with many different subjects. In such an environment it is important to provide flexible access to individual text pieces, and to structure the collection so that related text elements are identified and appropriately linked.

Methods are described in this study for the automatic structuring of heterogeneous text collections, and the construction of browsing tools and access procedures that facilitate collection use. The proposed methods are illustrated by performing searches with a large automated encyclopedia.

1 Access to Full Text Data Bases

In conventional information retrieval environments, documents are accessed by constructing large, so-called inverted, indexes containing all the distinct text words (except for some throw-away words), together with lists of document references that identify the documents, or document excerpts, in which the given text words occurs. Information is retrieved by formulating Boolean queries consisting of text words interrelated by Boolean operators, consulting the corresponding lists of document references in the index, and identifying all documents that contain the proper combination of query terms.

The retrieval technology using inverted term indexes together with Boolean query formulations has the considerable advantage that the identification of documents containing the required query term combination

*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.

is extremely rapid. In general the responses are available in a matter of seconds, even when the file contains several millions documents. Moreover, the retrieval effectiveness may be relatively high because the documents considered for retrieval, corresponding to entries in the appropriate reference lists in the index, are known in advance to contain at least one of the required search terms.

On the negative side, the conventional technology does not offer browsing capabilities, because the files are maintained in random rather than subject matter order. Occasionally users are turned off when errors in the Boolean query formulations produce questionable responses. For example, in answer to the query "Esau and Jacob" (the two biblical characters in the Old Testament), an automated encyclopedia may retrieve the encyclopedia article on "Brazilian literature" because a well-known Brazilian author wrote a text entitled "Esau and Jacob". In addition, in a conventional full-text retrieval system, a text is considered to be indivisible in the sense that relationships between text elements are not recognized, either within a text or between different texts, and that all documents are treated in the same way, regardless of scope, coverage, or text length.

In many practical retrieval situations, large differences exist in the scope of the individual text items, and in the manner in which the subject matter may be covered. For example, collections of electronic mail messages may include short reminders and brief answers to earlier questions; other messages may on the contrary provide long, discursive treatments of particular topics. In such circumstances, a system that treats all texts alike with no specification of internal or external structure may be of limited usefulness. In many cases, the longer documents must be broken down into individual pieces, relationships must be defined between text portions, and those relationships must be taken into account in the retrieval activities.

A number of related tasks are of interest in this connection: The text collection must be structured in such a way that text portions that belong together because of similarities in the subject area, or for other reasons, are properly recognized. This may be achieved by building hierarchical, or network arrangements of related text portions to be used for information browsing operations as well as for formal retrieval activities.

When a properly structured text collection is available that exhibits both internal and external links between related texts and text excerpts, browsing operations can be implemented by asking users to follow the linked structure and by providing automated tools to facilitate the browsing operations. Fast retrieval strategies are also important that use the collection structure to produce improved retrieval results. In the remainder of this study, collection structuring operations are outlined, and illustrations are given using a large automated encyclopedia for test purposes.

2 Document Structure

Two related aspects of document structure must be distinguished: the *abstract* structure specifies how different pieces of text fit together — for example, how the text is subdivided into sections, paragraphs, sentences, and so on, and how these elements are assembled into the full text; the *semantic* structure is concerned with text meaning, and with the manner in which the total meaning is built-up from the meaning of the individual pieces. For retrieval purposes, these two aspects are related, because the semantic document structure necessarily depends on a reasonable decomposition of a text into abstract text elements, and on a subdivision of topic coverage into introduction, development and conclusion.

The abstract structure of texts and documents has been studied by the *hypertext* community which concerns itself with the decomposition of texts into fragments, the relationships existing between text fragments, and the design of access structures to text fragments. [1-3] Two main types of abstract structure are distinguished: the *logical* structure defines the composition of document objects into successively larger text components, and the *layout* structure describes the composition of objects into successively larger physical units. In the hypertext environment, the logical structure is primarily of interest rather than the physical text representation. Distinctions may be introduced between primary, secondary, and auxiliary text structures, representing respectively, the primary text object in the actual documents; cross-references and relationships between the document elements and an index entry; and finally, external objects such bibliographies. [2]

The logical structure of texts, and the decomposition into main text elements is often specified by a mark-up language that identifies titles, chapter headings, section heads, paragraph separation, and so on. Such a logical decomposition may be useful in retrieval when access must be gained to text elements that may be directly identified by the mark-up. In addition to such a formal decomposition into distinct recognizable text elements, there is also a more subjective, informal decomposition of text into elements that are related by virtue of the fact that the subject matter treated by these text elements is similar. It is convenient to identify such subject-related text pieces by *linking* the corresponding text elements, thereby simplifying joint access when this is needed. Text links may also serve to join footnotes to related text fragments, to connect bibliographic references to the primary text, and to represent the cross-reference structure.

3 Structural Text Decomposition

In order to use the structural text information for browsing or retrieval, it is necessary to convert the apparently linear text organization into structured text representations. The following steps may be useful in building a linked document structure from the existing source documents: [4]

a) An identification of the particular components of the text database that will be used to define the hypertext "nodes", that is, text elements that are individually linkable in the structured text organization.

b) The generation of links between text elements that actually belong together.

c) The optional creation of additional structured text elements providing indexes and access structures.

d) The possible reformatting of the text elements to tailor the structured database to output and display media.

e) The possible addition of introductory information providing title pages, prefaces, and explanations for the text structure.

The consensus in the hypertext community is that some steps in the database design are in fact mechanizable — for example, links can be automatically supplied between text segments and corresponding footnotes, or bibliographic references. However, the more complex tasks, such as the generation of informal text links between subject-related text excerpts are believed to require manual intervention:

> "Conversion of larger ... documents, such as scientific papers, seems to be a process that inherently requires manual assistance, even when considering only structural conversions, if for no other reason than to provide the initial partitioning of otherwise undivided text into hypertext units and to resolve ambiguous language use in determining links." [5]

In this study procedures are described for the placement of text links between semantically related text elements automatically without human intervention. When these semantic links are supplemented by the formal structural links obtained from the mark-up, the complete text structure is made available for browsing and information retrieval.

The introduction of text links between semantically related text excerpts must depend on a detailed content analysis of texts, and text excerpts, followed by a comparison of the content identifications attached to the texts, and the generation of links between text portions whose content identifications are sufficiently similar. The text analysis operations are briefly described in the next section.

4 Text Analysis and Text Representation

In situations where the available texts span a variety of different topic areas, the conceptually-based text analysis methods that are often advocated in the literature are not usable, because it is impossible to build the *knowledge bases* that specify the contents and structure of the subject areas of interest. Instead, the text themselves must form the basis for the text analysis operations: content terms attached to the texts must be obtained from the texts by an automatic indexing operation, and vocabulary schedules and preconstructed thesauruses that are difficult to supply must be used sparingly, or not at all. [6, 7]

Typically, a *term vector* is defined for each text, or text fragment, of the form $D_i = (w_{d_{i1}}, w_{d_{i2}}, ..., w_{d_{it}})$ where $w_{d_{ik}}$ represents to weight of term T_k assigned to document D_i. Each term might represent a word, or word stem, extracted from the document text, and the weights could be used to distinguish the important terms from the less important ones. Typical term weights range from 0 to 1, a zero being used for terms absent from a particular vector, while 1 represents a fully-weighted term. Intermediate weight between 0 and 1 apply to less important terms. [8]

Because information requests are much more easily generated as discursive natural-language queries than as Boolean formulations, the vector representation is used for queries as well as for documents. If a query Q_j is represented as $Q_j = (w_{q_{j1}}, w_{q_{j1}}, ..., w_{q_{jt}})$ where a zero weights once again designates a term not assigned to a text item, the usual inner product $\sum_k (w_{d_{ik}} \cdot w_{q_{ik}})$ may be used to reflect the similarity between query and document texts.

Standard text indexing systems can be used to generate the weighted term vectors. Typically the individual text words are isolated, common words are removed, and term weights are assigned to the remaining words or word phrases. It is known that terms that occur frequently in local text environments, such as particular text paragraphs, but that are relatively rare in the complete text collection are important for text retrieval purposes. This is reflected in the well-known $tf \times idf$ (term frequency times inverse collection frequency) term weighting function, which prefers terms with high local, but low global occurrence frequencies. A typical $tf \times idf$ term weighting formula for term T_k in D_i is the *ntn* weight defined as $w_{(atn)ik} = tf_{ik} \cdot logN/tf_k$ where $w_{(atn)ik}$ is the *atn* weight of term T_k in text D_i, tf_{ik} is the frequency of occurrence of the term in the document, n_k is the total number of documents with term T_k assigned, and N is the total number of texts in the database.

It should be noted that the term weight computation depends crucially on the text environment in which the indexing operation is performed. When text sentences are indexed, the term frequency factor, tf_{ik}, will normally be low because most sentences are short and contain few words. The collection size N (number of sentences) and the collection frequency n_k of term T_k may however be quite large. When text paragraphs or text sections are indexed, the collection size N (number of paragraphs or sections) and the collection frequency n_k are smaller, but the term frequency, tf_{ik}, may be larger.

When the inner product similarity formula is used to measure text similarities with *ntn* term weights, the similarity between two texts depends on the number and the weight of common terms in the two texts.

The *ntn* term weight thus favors the retrieval of longer texts that have more terms assigned and hence very likely more matching terms with other texts. Such a weight may be appropriate for text sentences to avoid the retrieval of short sentence fragments that may not be meaningful ("consider the following example", "Table X is an illustration").

When longer texts are analyzed of paragraph-length and beyond, a system favoring long texts over shorter ones may not be effective. In these circumstances, it is preferable to give each document an equal chance of being retrieved, regardless of document length. This is achieved by using a normalized term weighting system, such as the *ntc* term weight which carries a length-normalizing factor in the denominator:

$$w_{(ntc)ik} = \frac{tf_{ik} \cdot logN/n_k}{\sqrt{\sum_p (tf_{ip})^2 \cdot (logN/n_p)^2}} \tag{1}$$

In expression (1) the sum in the denominator runs over all terms in a particular document vector. When the inner product formula $\sum_k w_{d_{ik}} \cdot w_{d_{jk}}$ is used to compute the similarity between texts D_i and D_j (or between text D_i and query Q_j) for *ntc* weighted terms, the text similarity depends on the proportion, rather than the number of matching terms. This implies that longer texts are no longer retrieved in preference to shorter ones. [9-11] In the experiments described of this study, normalized term weights, such as those of expression (1) are used for paragraphs and longer texts, and unnormalized weights without the normalization factor in the denominator serve for text sentences and shorter texts.

5 Text Comparison Methods

Before text links can be placed between text excerpts based on similarities in text content, or texts can be retrieved in answer to user formulated queries, it is necessary to identify different texts with identical, or closely related subject matter. A simple comparison of the vocabularies included in the respective texts may not be adequate for this purpose in view of the well-known variability inherent in natural language, and the obvious ambiguities in meaning of many text words and expressions. In many areas of discourse, it is easy to select excerpts where similar word combinations carry quite different meanings, while quite distinct linguistic expressions may be used to render similar meanings. This explains the widespread use in text analysis and indexing of controlled indexing vocabularies where terms with unique and well-specified meanings are used to represent text content.

Manual indexing methods carried out by trained subject experts are, however, not usable when large text collections must be treated in arbitrary subject areas. In these circumstances, it is best to rely on the texts themselves for analysis purposes. The question arises how to eliminate unwanted variabilities in the indexing vocabulary when the text words themselves serve for indexing purposes. Countless examples can be cited to show that the meaning of linguistic expressions changes with the context. Thus "reaching a base" means different things when a baseball game is in progress and the phrase relates to a player on the offensive team, than in wartime when the phrase relates to particular troop movements. While phrases of this type an inherently ambiguous, the context within which such ambiguous phrases are used normally renders the meaning clear. This suggests that text meanings should be recognized by performing global vocabulary comparisons between different texts, based on similarity measurements between the corresponding term vectors, and in addition verifying that the local contexts in which the expressions are used are related.

The following strategy may then be followed to determine text similarities:

a) Each text is analyzed at various levels of detail, and indexing vectors are built for the complete text, as well as text sections, text paragraphs, and text sentences.

b) Vector comparison operation are carried out between complete texts as well as between lower-level text constructs.[1]

c) When two texts exhibit global text similarities exceeding a stated threshold, and in addition the texts include local passages, such as paragraphs or sentences, with sufficiently large local text similarities, the assumption is that the sample texts are semantically homogeneous, and cover similar subject areas. The subject similarity may be recognized by placing a text link between the corresponding text passages.

[1] Local paragraph and sentence matches have previously been used in information retrieval for text clustering and passage-retrieval purposes.[12,13]

Consequently, the global similarity computations that are based on a comparison of complete vocabulary patterns for different texts are used as initial filters to reject the large majority of texts pairs that do not exhibit a reasonable degree of global text similarity. For text pairs that pass the initial filter, a second more refined comparison operation is then used, based on local coincidences in the sentence or paragraph structures included in the texts.

In previous experiments, the proposed text linking strategy was found to operate with a high degree of accuracy. Substantially more than 90 percent of the generated links for text paragraphs included in a sample text book were found to be proper, even when the paragraph pair was drawn from distinct chapters of the book.[11, 14]

The available test results relate only to the precision of text linking, that is, to the proportion of correctly linked text pairs. The recall question is more difficult to deal with. The question arises whether complex vocabulary comparison methods are usable to recognize a substantial proportion of texts that are actually similar, or whether, on the contrary, closely related texts cannot be identified because the global and local vocabulary differences are too severe. A conclusive answer is not available. However, the recall performance of the text linking methods may also be relatively high, if one considers the difficulties that arise when an attempt is made to cover the same subject matter by using completely distinct terminologies and contexts. Ultimately, one must expect that many of the local contexts will be related when different texts are processed that cover similar subject matter. Such matching contexts may be detected, and the texts may be linkable even when the vocabulary differences are large.

6 The Encyclopedia Application

A. The Encyclopedia Environment

The recognition and utilization of document structure is of interest primarily for texts, or text collections, consisting of heterogeneous elements that are accessed and used selectively. This includes dictionaries and encyclopedia, textbooks, instruction manuals, and many large document collections. For present purposes, the nearly 25,000 articles included in the 29-volume Funk and Wagnalls New Encyclopedia are used as an experimental database.[2]

An encyclopedia is characterized by the large number of included articles and the wide subject coverage. There is also a wide disparity in the scope and the coverage of individual encyclopedia articles, ranging in the Funk and Wagnalls (FW) case from simple cross-references ("Abo, see Turku") to long discursive treatments of several hundred paragraphs ("United States of America"). Most encyclopedias also exhibit a substantial secondary structure, consisting of indexes, bibliographies, and cross-references between articles, designed to simplify the text access. The basic text statistics for the FW encyclopedia are included in Table 1.

For some years, the hypertext community has been interested in utilizing the structure of dictionaries and encyclopedia for text access purposes.[15-19] Two types of facilities are normally contemplated for encyclopedia texts:

a) A linked hypertext implementation which uses explicit links between articles, defined by the available cross-references and by listed bibliographic references, to connect related text excerpts.

b) A full-text Boolean query capability which provides access to articles containing a given combination of keywords in response to Boolean query formulations.

The Boolean, full-text query capability may produce useful output in some circumstances when users are able to generate effective query statement. Similarly, the browsing and menu tools that can be provided with the linked hypertext structure may also simplify the user's accessing capability. However, only formal text links are normally available in actual encyclopedia environments, and attempts to generalize the hypertext structure by adding, subjective content- related links have not been pursued in the past:

> "We expected that it would soon be possible to extract the implicit links (within documents) automatically with natural-language processing or clever indexing techniques, but we have been

[2]The writers are grateful to the Microsoft Corporation for making available a machine-readable version of the Funk and Wagnalls New Encyclopedia for experimental purposes.

disappointed ..... and we conclude that implicit intra-document links are best identified by the hypertext reader."[20]

The missing natural-language capability referred to in the previous quotation is supplied in this study, and implicit links between content-related text excerpts are automatically generated. In addition, various enhanced text accessing facilities are also implemented:[21]

a) Complete encyclopedia articles are accessed and retrieved by using for search purposes either discursive, English-language user formulated queries, or the text of sample encyclopedia articles. In either case, the system supplies ranked lists of related articles in decreasing order of the computed similarity with the query statement.

b) In addition, individual text sections, text paragraphs and text sentences are also separately retrievable in answer to available search requests.

c) The retrieved document output can be suitably controlled by insisting on matching lower-level structures between the query texts and the retrieved document output. For example, text sections may be retrieved only if matching paragraph or sentences are detected for the respective query-document texts.

d) Formal links supplied by the mark-up information may also be followed, including links to footnotes and to referenced bibliographic items.

e) Interactive query reformulation methods are available that generate improved query formulations based on user judgments about the relevance of previously retrieved items. The well-known *relevance-feedback* process has been shown to be effective in this context. [22] Alternatively, users can supply longer, more discursive query statements when the shorter formulations do not produce the expected results.

The capabilities of the encyclopedia search system are illustrated by sample searches in the remainder of this note.

B. Levels of Text Relations

It is well-known that the structure of running text is complex, and ordinary language serves to express a variety of different notions and relations. It is then not surprising that individual text excerpts are relatable in different ways, depending on the scope and breadth of the query statements, and the type of matching strategy used to compare the texts. In general, the narrower the query formulation, and the shorter the text excerpts used for query formulation purposes, the more direct will be the relationship between queries and retrieved texts, and the greater will be the importance of individual query terms. On the other hand, the more discursive the query texts, and the more diverse the ideas contained in the query formulations, the more elusive may be the relationship between queries and retrieved items.

Consider, as examples, various searches based on the encyclopedia article entitled "United States of America". This article is one of the longest in the encyclopedia, consisting of over 600 text paragraphs and section headings. Covered are many different topics relating to the United States, including the geology, agriculture, industry, literature, arts, politics and history of the country. When the full article is used as a search request, the corresponding document vector contains thousands of terms covering all the aforementioned topics. Hence the retrieved documents will all deal with the same variety of topics. Instead of retrieving information about the United States, the retrieved articles contain data about other countries, closely related to the United States. In the same way, long biographies of particular individuals can be used to obtain similar information about other related individuals. The top 10 documents (the 10 documents with the largest computed query similarity) obtained in response to the full article "United States of America" are shown in Table 2. As the Table shows, the retrieved documents with the largest query similarities cover countries such as Canada, the USSR, China, Great Britain, and so on, which are all large and diverse entities with close ties to the United States.

At the other extreme in the range of text linking possibilities are the responses obtainable with a query statement consisting only of the words "United States of America". Such a formulation approximates the Boolean query statements used conventionally in standard full-text systems: each query word takes on extraordinary importance, and the retrieved items are those containing large numbers of query terms. The responses generated by the title search "United States of America" will then include short cross-reference articles, containing only a few of the words "united", "state", and "america". The ranked output generated by the title search is shown in Table 3.

Because the title query is effectively equivalent to a single sentence, an unnormalized term weight is used to represent the query terms. As a result, the similarity values generated for the sentence-document output of Table 3 are quite different from those obtained for the document-document output of Table 2.

The results of Table 2 and 3 make it clear that not much useful information about the United States is obtainable by using either the full 600-paragraph text, or the simple title "United States of America". More meaningful results are obtainable when focused query statements are used that describe particular topics in detail. Consider, as an example, the text of section 40912 entitled "Literature" which appears as a section in the article "United States of America". This section is subdivided into 5 paragraphs (numbers 121322 to 121326) covering various periods of American literature. The text of section 40912 is reproduced in Fig. 1.

The output obtained by performing a section search with section 40912 as a search request is shown in Table 4. Here the output is much more specific, consisting of other text sections dealing with American literature. The top four sections retrieved in response to the text of Fig. 1 are excerpts from the article "Novel", entitled, respectively "The 20th Century – The U.S.", and "The 19th Century – The U.S." (sections 29275 and 29271). In addition two sections are retrieved from the article on "American Literature", entitled "The 20th Century – World War II and After: Fiction", and "The 19th Century – Early 19th Century".

The section search of Table 4 also retrieves information about specific American writers (section 42729, Walt Whitman, and the complete article on Eudora Welty). However, the query similarities are relatively low for these articles – query similarities that fall below 0.20 for the normalized term-weight comparisons used for document-document and section-section searches are normally considered to be marginal; analogously, query similarities below 5.0 for unnormalized term weights comparisons might also be disregarded.

More specific document relations than those obtained with the preceding queries are generated when more specific query formulations are used. The first possibility consists in using one or more of the text paragraphs in the "Literature" section for search purposes, and comparing the query paragraph with all other paragraphs in the encyclopedia. Table 5 contains the output obtained by performing paragraph searches using the third and fourth paragraphs of Fig. 1 as search requests (the paragraphs starting with the words "Theodore Dreiser" and "Hemingway and Faulkner", respectively). As the Table shows, the retrieved articles relate to specific American writers. In addition, the long articles entitled "Novel" and "American Literature" are also recovered by the paragraph queries above the stated similarity threshold of 5.0.

When specific topics are described in some depth in the query statements, the retrieved output normally consists of material closely related to the original topic specification. Occasionally, a query article covering a well-defined topic contains nonspecific words, or formulations that do not lend themselves to a properly focused search operations. For example, the use of a cross-reference article such as "Church of Christ Scientist, see Christian Science Church" may produce little information about "Christian Science", because terms such as "church", "Christ", and "science" are common encyclopedia words with meanings unrelated to the Christian Science religion. In such cases, a user-formulated, discursive, natural-language query produces better output.

Consider the example of Table 6. Here the user-formulated query retrieves first a short cross-reference article. (The large query-similarity is due in that case to the large proportion of matching terms between the query formulation and the retrieved article). However, the main six-paragraph article covering Christian Science is also retrieved (document 5344), together with articles about well-known Christian Scientists (Mary Baker Eddy and Phineas Parkhurst Quimby), as well as two cities in Massachusetts closely tied to Mary Baker Eddy (Lynn and Concord).

The output shown in Tables 2-6 was obtained by performing global text matches between full documents, text sections, text paragraphs, and document or section titles. Marginal items are easier to reject when lower-level matching constructs are detected within the globally matching texts. Fig. 2 shows a sample sentence match for two text excerpts consisting, respectively, of the query section "Prohibition" included in the article on "United States of America", and the retrieved document "George Woodward Wickersham". In the global text comparison between these two texts, the document on G.W. Wickersham (doc. 24288) was retrieved first with a large query-document similarity of 8.47. When large similarities exist between query texts and retrieved document excerpts, it is perhaps not necessary to insist on lower-level structural similarities. For the two sample texts, significant local sentence similarities are, however, easy to find between the high-lighted sentences of Fig. 2 (sentences 383651 and 398807). When lower-level text matches are detected in addition to the global text similarities, the retrieval effectiveness is expected to be very high.[23]

C) Automatic Text Linking

The retrieval procedures illustrated in the previous section are used to produce ranked output in decreasing order of the query-document similarity. The same methods also serve directly for the generation of browsing maps and linked text structures that may help users in traversing the text structure. Specifically, an initial query statement can be used first to obtain closely related documents; these retrieved items can in turn serve as search requests that locate additional related items. Fig. 3 shows a three-level map for the base article "Integrated Circuits". A similarity threshold of 0.20 is used initially to obtain six related items shown on level 1 of the Figure. A higher similarity threshold of 0.25 is used for the second-stage searches, producing 5 additional articles on level 2, plus a number of links between the previously retrieved items on level 1. A third search stage with an increased similarity threshold of 0.30 finally retrieves two more items shown on the lowest level of Fig. 3.

By varying the similarity thresholds used to perform the various search stages, the generated reading maps will become more or less complete. Procedures such of those used to generate the output of Fig. 3 are useful to structure large texts, thereby providing structured representations at various levels of details, with different types of text links representing the multiple relationship inherent in natural-language texts.[23, 24] By using flexible text matching procedures, such as those described in this study, it may well be possible to generate both the abstract (logical) text structure as well as the semantic structure of content relations fully automatically, replacing the elaborate, manual linkage-generation procedures currently in use. This would vastly increase the usefulness of hypertext linking for text browsing and retrieval.

References

1. B. Shneiderman, Reflections on Authoring, Editing, and Managing Hypertext, in *The Society of Text*, E. Barrett, editor, MIT Press, Cambridge 1989.

2. R. Furuta, An Object-Based Taxonomy for Abstract Structure in Document Models, *Computer Journal*, 32:6, 1989, 492-504.

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

4. R. Furuta, C. Plaisant, and B. Shneiderman, A Spectrum of Automatic Hypertext Constructions, Hypermedia, 1:2, 1989, 179-195.

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

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

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

8. G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis, *Journal of the ASIS*, 26:1, January 1975, 33-44.

9. G. Salton and C. Buckley, An Evaluation of Text Matching Systems for Text Excerpts of Varying Scope, Technical Report 90-1134, Dept. of Computer Science, Cornell University, June 1990.

10. G. Salton and C. Buckley, A Note on Term Weighting and Text Matching, Technical Report TR 90-1166, Dept. of Computer Science, Cornell University, October 1990.

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

12. J. O'Connor, Answer Passage Retrieval by Text Searching, *Journal of the ASIS*, 32:4, July 1980, 227-239.

13. S. Al-Hawamdeh and P. Willett, Paragraph-Based Near-Neighbor Searching in Full Text Documents, *Electronic Publishing*, 2:4, December 1989, 179-192.

14. G. Salton, C. Buckley, Flexible Text Matching for Information Retrieval, Technical Report TR 90-1158, Dept. of Computer Science, Cornell University, September 1990.

15. S.A. Weyer and A.H. Borning, A Prototype Electronic Encyclopedia, *ACM Transactions on Office Information Systems*, 3:1, January 1985, 63-88.

16. D.R. Raymond and F.W. Tompa, Hypertext and the Oxford English Dictionary, *Communications of the ACM*, 31:7, July 1988, 871-878.

17. G. Marchionini and B. Shneiderman, Finding Facts vs. Browsing Knowledge in Hypertext Systems, *Computer*, 21:1, Jan. 1988, 70-80.

18. R.J. Glushko, Transforming Text into Hypertext for a Compact Disc Encyclopedia, Proc. CHI 89 Conference, *Human Factors in Computing Systems*, Assoc. for Computing Machinery (ACM), New York, 1989, 293-298.

19. P. Kahn, Linking Together Books: Experiments in Adapting Published Material into Intermedia Documents, *Hypermedia*, 1:2, 1989, 111-145.

20. R.J. Glushko, Design Issues for Multi-Document Hypertexts, Hypertext 89 Proceedings, Assoc. for Comp. Machinery (ACM), New York, Nov 1989, 51-60.

21. G. Salton and C. Buckley, Automatic Text Structuring and Retrieval — Experiments in Automatic Encyclopedia Searching, Technical Report TR 91-1196, Dept. of Computer Science, Cornell University, April 1991, to be presented at the 14th Annual ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, October 1991.

22. J.J. Rocchio, Jr., Relevance Feedback in Information Retrieval, in *The Smart System — Experiments in Automatic Document Processing*, G. Salton, editor, Prentice Hall Inc., Englewood Cliffs, NJ, 1971, 313-323.

23. G. Salton and C. Buckley, Approaches to Text Retrieval for Structured Documents, Technical Report 90-1083, Dept. of Computer Science, Cornell University, January 1990.

24. G. Salton, C. Buckley, and Z. Zhao, Text Linking and Retrieval Experiments for Textbook Components, Technical Report 90-1125, Dept. of Computer Science, Cornell University, May 1990.

Text Construct	Number of Entries
Number of full articles	24,900
Number of text sections	44,000
Number of text paragraphs and section titles	130,000
Number of text sentences	410,000

Table 1: Approximate Number of Text Entries in Funk and Wagnalls Encyclopedia

Query: United States of America (full text)			
Retrieval Rank	Document Number	Document Title	Similarity Measure
1	4252	Canada	0.3439
2	23288	USSR	0.3371
3	5232	China	0.3229
4	10268	Great Britain	0.3202
5	9163	France	0.3046
6	8349	Europe	0.2976
7	12381	Japan	0.2950
8	12268	Italy	0.2949
9	2950	Blacks in the Americas	0.2957
10	16762	North America	0.2912

Table 2: Output for Full Article "United States of America" (document-document comparison)

Query: United States of America (title words only)			
Retrieval Rank	Document Number	Document Title	Similarity with Query
1	1438	Army, U.S.	4.24
2	449	Air Force, U.S.	3.66
3	20668	Senate, U.S.	3.64
4	16404	Navy, U.S.	3.41
5	15508	Military Accademy, U.S.	3.33
6	19308	Representatives, House of	3.18
7	11539	House of Representatives	3.18
8	11905	Inauguration, Presidential	3.02
9	23301	United Fund	3.02
10	450	Air Force Academy, U.S.	3.01

Table 3: Output for Title "United States of America" (title-document comparison)

Query: Section 40912 entitled "Literature" under "United States of America" (see Fig. 1)

Retrieval Rank	Section Number	Section Structure	Similarity with Query
1	29275	Novel The 20th Century The U.S.	0.3350
2	1519	American Literature The 20th Century World War II and After: Fiction	0.2264
3	29271	Novel The 19th Century The U.S.	0.2177
4	1910	American Literature The 19th Century Early 19th Century	0.1754
5	42729	Whitman, Walt Whitman's Reputation	0.1716
6	42522	Welty, Eudora	0.1659
7	36431	Short Story The 20th Century The U.S.	0.1651

Table 4: Output for Section Search Using Query Section on "Literature" under "United States of America"

Retrieval Rank	Document Number	Document Title	Similarity Value (threshold 5.0)
Query:	Paragraph 121324 "Theodore Dreiser" from "Literature" Section under "United State of America"		
1	24211	Edith Newhold Wharton	6.81
2	7474	Theodore Dreiser	5.80
3	13875	Harry Sinclair Lewis	5.61
4	16851	Novel	5.08
Query:	Paragraph 121325 "Hemingway and Faulkner" from "Literature" Section under "United State of America"		
1	24129	Eudora Welty	9.43
2	2516	Saul Bellow	7.80
3	16851	Novel	7.08
4	23806	Kurt Vonnegut, Jr.	6.83
5	847	American Literature	6.60
6	14635	Norman Mailer	5.85
7	16210	Vladimir Nobokov	5.48
8	11905	Edward Albee	5.36
9	21664	John Steinbeck	5.31

Table 5: Sample Searches for Paragraph Queries using Two Paragraphs on "Literature" under "United States of America"

Query: User-formulated, discursive query statement "I am interested in the Christian Science movement, the religion founded by Mary Baker Eddy in Boston, and the foundation of the First Church of Christ Scientist"

Retrieval Rank	Document Number	RankDocument Title	Similarity with Query
1	5390	Church of Christ Scientist (cross-reference only)	0.4658
2	18966	Quimby, Phineas Parkhurst	0.3129
3	14442	Lynn, Masssachusetts	0.3021
4	7779	Eddy, Mary Baker	0.2917
5	5344	Christian Science (6 paragraphs)	0.2363
6	5935	Concord, Massachusetts	0.2318

Table 6: Output for User-Formulated Natural-Language Query

Literature

The first major American novelist was james Fenimore Cooper, with The Last of the Mohicans (1826), The Prairie (1827), and other works about the frontier. The romantic period of American literature, from about 1830 to 1865, introduced important novelists such as Nathaniel Hawthorne, author of The Scarlet Letter (1850) and The House of the Seven Gables (1851), in which he probed New England's Puritan heritage; and Herman Mellville, author of Moby-Dick (1851), a complex and poetic novel of the sea.

Realism in American literature, prominent from the close of the Civil War until about the beginning of the 20th century, was the product of a new mass audience and the experience of industrialization. Major figures of this time included writers as diverse as the humorist Mark Twain, with his classic tales of boyhood Tom Sawyer (1876) and Huckleberry Finn (1884); and Henry James, a stylistic innovator whose works, such as The Portrait of a Lady (1881) and The Ambassadors (1903), were landmarks in the development of the novel.

Theodore Dreiser, whose Sister Carrie (1900) and An American Tragedy (1925) describe in an awkward but compelling style how spiritually empty industrial America had become, marked the new age of naturalism, which ran until about 1930. This was a rich period of American letters; important novelists during this period included Edith Wharton (Ethan Frome, 1911; The Age of Innocence, 1920); F. Scott Fitzgerald (The Great Gatsby, 1925); Sinclair Lewis (Main Street, 1920; Babbitt, 1922), the first American Nobel Prize winner in literature; Ernest Hemingway, also a Nobel Prize winner, perhaps most highly praised today for his terse, carefully crafted short stories collected in In Our Time (1924), Men Without Women (1927), and Winner Take Nothing (1933); and William Faulkner, whose innovative techniques and thoughtful characterizations in such novels as The Sound and the Fury (1929), Light in August (1932), and Absalom, Absalom! (1936) won him the Nobel Prize in 1949.

Hemingway and Faulkner remained leading writers into the 1950s; they were joined by John Steinbeck (The Grapes of Wrath, 1939; Nobel Prize, 1962), Robert Penn Warren (All the King's Men, 1946), James Jones (From Here to Eternity, 1951), James Baldwin (Go Tell It on the Mountain, 1953), Norman Mailer (The Naked and the Dead, 1948; The Executioner's Song, 1979), and Vladimir Nabokov (Lolita, 1955; Pale Fire, 1962). Novelists of contemporary note include Eudora Welty (The Ponder Heart, 1954; The Optimist's Daughter, 1969), well known also for her short stories; Saul Bellow (The Adventures of Augie March, 2953; Humboldt's Gift, 1976; Nobel Prize, 1976); Kurt Vonnegut, Jr. (Slaughterhouse-Five, 1969); and John Updike (Rabbit, Run, 1960; The Coup, 1978).

Distinctive American poetry first appeared in the 19th century, with the musical and highly rhythmic works of Edgar Allan Poe, the experimental democratic chant of Walt Whitman (Leaves of Grass, 1855), and the tightly wrought lyrical verse of Emily Dickinson. Modern American poetry began in the early 20th century with the lyrics and dramatic poems of Robert Frost; the Cantos of Ezra Pound, the founder of imagism; and T. S. Eliot's revolutionary long poem, The Waste Land (1922). Often considered the most innovative poetry in the English language, modern American poetry has continued to be enriched by such gifted poets as Wallace Stevens, William Carlos Williams, Elizabeth Bishop, Robert Lowell, Allen Ginsberg, Howard Nemerov, Richard Wilbur, and Adrienne Rich (1929-). For more information on American prose and poetry, American Literature.

Fig. 1. Text of Section 40912 ("Literature") under "United States of America".

****

doc 41022
United States of America
History
The Roaring Twenties: Boom and Crash
Prohibition

The most violent controversial issue of the period 1920-32 was Probibition. The movement to prohibit the manufacture and sale of intoxicating beverages in the U.S. originated in the early part of the 19th century and culminated with the ratification, in January 1919, of the 18th Amendment to the Constitution. In 1929 a commission appointed by President Hoover and headed by former U.S. Attorney General George W. Wickersham reported that federal enforcement of the liquor laws was a failure. Public sentiment, meanwhile, had steadily been growing for repeal of the 18th Amendment, and in February 1933, Congress passed and submitted to the states the 21st Amendment to the Constitution, which gave the control of the liquor traffic back to the individual states; by December of that year, 36 states had ratified the 21st Amendment and it was declared part of the Constitution.

a) Query Section 41022 "Prohibition"
in article "United States of America",
Section "History — The Roaring Twenties"

****

doc 24288
Wickersham, George Woodward

(1858-1936), American lawyer and public official, born in Pittsburgh, PA, and educated at Lehigh University and the University of Pennsylvania. He was attorney general in the cabinet of President William Howard Taft from 1909 to 1913. From 1923 until his death he was president of the American Law Institute, and from 1924 to 1929, a member of the commission on progressive codification of international law, appointed by the Council of the League of Nations. In 1929 President Herbert Hoover appointed Wickersham chairman of the National Law Enforcement Commission, which became known as the Wickersham Commission, and was most famous for its negative report, released in 1930, on the status of Prohibition.

b) Retrieved Document 24288 "G.W. Wickersham"

Fig. 2. Sentence Match between Relevant Text Excerpts (sentences 383651 – 398807, similarity 138.25)

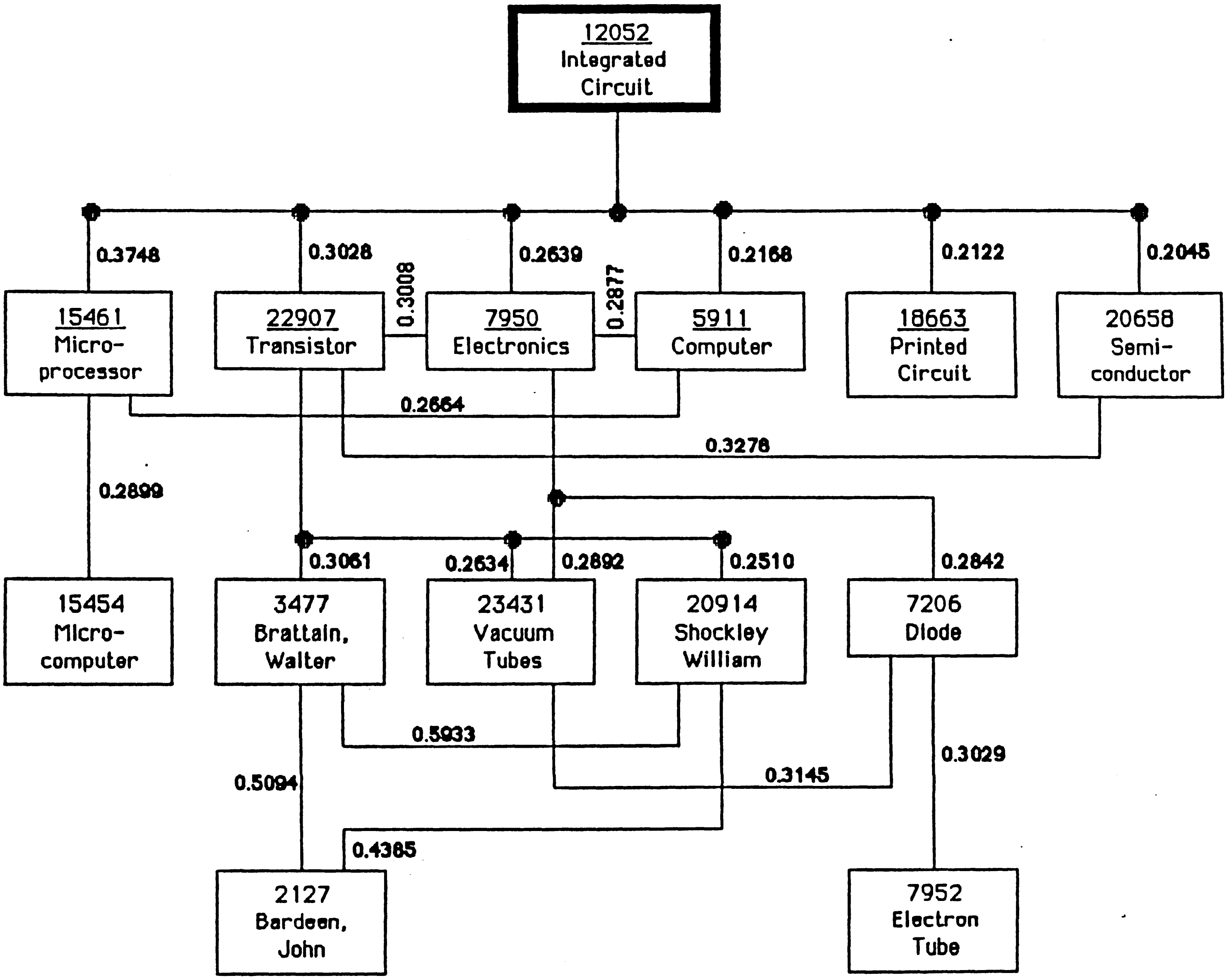

Fig. 3. Linked Structure of Topics related to "Integrated Circuits" (document numbers inside boxes; document similarities along branches between documents – Map by Ian Salkind)