83. M. David and A. C. Larner, *Science* **257**, 813 (1992).
84. K.-I. Igarashi, M. David, A. C. Larner, D. S. Finbloom, *Mol. Cell. Biol.* **13**, 3984 (1993).
85. R. Schreiber, personal communication.
86. E. H. Fischer, H. Charbonneau, N. K. Tonks, *Science* **253**, 401 (1991).
87. J. Mirkovitch and J. E. Darnell Jr., *Mol. Biol. Cell* **3**, 1085 (1992).
88. H. B. Sadowski, K. Shuai, J. E. Darnell Jr., M. Z. Gilman, *Science* **261**, 1739 (1993).
89. O. Silvennoinen, C. Schindler, J. Schlessinger, D. E. Levy, *ibid.*, p. 1736.
90. S. Ruff-Jamison, K. Chen, S. Cohen, *ibid.*, p. 1733.
91. A. C. Larner *et al.*, *ibid.*, p. 1730.
92. H. Kotanides and N. C. Reich, *ibid.* **262**, 1265 (1993).
93. A. Bonni, D. A. Frank, C. Schindler, M. E. Greenberg, *ibid.*, p. 1575.
94. R. Graham and M. Gilman, *ibid.* **251**, 189 (1991).
95. D. E. Levy, in *Interferon: Principles and Medical Applications*, S. Baron *et al.*, Eds. (University of Texas Medical Branch at Galveston, Galveston, TX, 1992), pp. 161–173.
96. S. Holland, G. R. Stark, I. M. Kerr, unpublished observations.
97. D. Watling, G. R. Stark, I. M. Kerr, unpublished observations.

Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts

Gerard Salton, James Allan, Chris Buckley, Amit Singhal

Vast amounts of text material are now available in machine-readable form for automatic processing. Here, approaches are outlined for manipulating and accessing texts in arbitrary subject areas in accordance with user needs. In particular, methods are given for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

Many kinds of texts are currently available in machine-readable form and are amenable to automatic processing. Because the available databases are large and cover many different subject areas, automatic aids must be provided to users interested in accessing the data. It has been suggested that links be placed between related pieces of text, connecting, for example, particular text paragraphs to other paragraphs covering related subject matter. Such a linked text structure, often called hypertext, makes it possible for the reader to start with particular text passages and use the linked structure to find related text elements (*1*). Unfortunately, until now, viable methods for automatically building large hypertext structures and for using such structures in a sophisticated way have not been available. Here we give methods for constructing text relation maps and for using text relations to access and use text databases. In particular, we outline procedures for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

Text Analysis and Retrieval: The Smart System

The Smart system is a sophisticated text retrieval tool, developed over the past 30 years, that is based on the vector space model of retrieval (*2*). In the vector space model, all information items—stored texts as well as information queries—are represented by sets, or vectors, of terms. A term is typically a word, a word stem, or a phrase associated with the text under consideration. In principle, the terms might be chosen from a controlled vocabulary list or a thesaurus, but because of the difficulties of constructing such controlled vocabularies for unrestricted topic areas, it is convenient to derive the terms directly from the texts under consideration. Collectively, the terms assigned to a particular text represent text content.

Because the terms are not equally useful for content representation, it is important to introduce a term-weighting system that assigns high weights to terms deemed important and lower weights to the less important terms. A powerful term-weighting system of this kind is the well-known equation $f_t \times 1/f_c$ (term frequency times inverse collection frequency), which favors terms with a high frequency (f_t) in particular documents but with a low frequency overall in the collection (f_c). Such terms distinguish the documents in which they occur from the remaining items.

When all texts or text queries are represented by weighted term vectors of the form $D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$, where d_{ik} is the weight assigned to term k in document D_i, a similarity measure can be computed between pairs of vectors that reflects text similarity. Thus, given document D_i and query Q_j (or sample document D_j), a similarity computation of the form $sim(D_i, Q_j) = \sum_{k=1}^{t} d_{ik}d_{jk}$ can produce a ranked list of documents in decreasing order of similarity with a query (or with a sample document). When ranked retrieval output is provided for the user, it is easy to use relevance feedback procedures to build improved queries on the basis of the relevance of previously retrieved materials.

In the Smart system, the terms used to identify the text items are entities extracted from the document texts after elimination of common words and removal of word suffixes. When the document vocabulary itself forms the basis for text content representation, distinct documents with large overlapping vocabularies may be difficult to distinguish. For example, the vectors covering biographies of John Fitzgerald Kennedy and Anthony M. Kennedy, the current Supreme Court justice, will show many similarities because both Kennedys attended Harvard University, were high officials of the government, and had close relationships with U.S. presidents. The global vector similarity function described earlier cannot cope with ambiguities of this kind by itself. An additional step designed to verify that the matching vocabulary occurs locally in similar contexts must therefore be introduced as part of the retrieval algorithm. This is accomplished by insisting on certain locally matching substructures, such as text sentences or text paragraphs, in addition to the global vector match, before accepting two texts as legitimately similar (*3*).

Consider, as an example, a typical search conducted in the 29-volume Funk and Wagnalls encyclopedia, using as a query the text of article 9667, entitled "William Lloyd Garrison" (Garrison was the best known of the American abolitionists, who opposed slavery in the early part of the 19th century) (*4*). The upper portion of Table 1 shows the top 10 items retrieved in response to a global vector comparison. The top retrieved item is article 9667 itself, with a perfect query similarity of 1.00, followed by additional articles dealing with abolitionism and the slavery issue, retrieved with lower similarity values.

The upper portion of Table 1 consists of relevant items only, with the exception of article 9628, entitled "Gar," retrieved in position eight on the ranked list. Gar is a type of fish, obviously unrelated to the slavery issue but erroneously retrieved because truncated terms were used in the text vectors, and the truncated form of "Garrison" matches "Gar." (Removal of "-ison" as part of the stemming process first reduced "Garrison" to "Garr," as in "comparison" and "compar"; removal of the duplicated consonant then reduced "Garr" to the final

The authors are in the Department of Computer Science, Cornell University, Ithaca, NY 14853–7501, USA.

"Gar.") The lower portion of Table 1 shows the results obtained with an additional local text comparison that required at least one matching text sentence between the query article and each retrieved document. There are no matching sentences in documents 9667 ("Garrison") and 9628 ("Gar"), because gar, meaning fish, and "Gar" derived from the name Garrison are obviously not used in similar contexts. Hence the offending document 9628 was removed from the retrieved list. Most linguistic ambiguities are similarly resolvable by this global-local vector-matching process. The lower portion of Table 1 also differs from the upper in that certain text passages are retrieved (labeled "c" for section and "p" for paragraph) in addition to certain full document texts. The passage retrieval issue is examined in more detail in the next section.

Text Decomposition and Structure

Practical retrieval searches deal with text items that are heterogeneous in both subject matter and text length. Thus, in the same text environment it may be necessary to cope with short e-mail messages as well as long book-sized texts. In an encyclopedia, three-word articles representing cross-references from one subject to another occur routinely, in addition to many long treatments such as the 175-page article entitled "United States of America." In a vector-processing environment, long articles that deal with diverse subject matter are difficult to retrieve in response to short, more specific queries, because the overall vector similarity is likely to be small for such items. Thus, the full article "United States of America" is not retrieved in the top 10 items in response to the query about William Lloyd Garrison, even though certain sections in the article specifically deal with abolitionism.

The rejection of long articles can reduce retrieval performance in some cases. More generally, long articles are difficult for users to handle even when retrieval is possible, because long texts cannot easily be absorbed and processed. This suggests that long texts be broken down into smaller text passages and that access be provided to shorter text excerpts in addition to full texts. Various attempts have been made in the past to implement passage retrieval capabilities, but flexible systems capable of handling text excerpts do not currently exist (5).

The Smart system can deal with text segments of varying length, including text sections, paragraphs, groups of adjacent sentences, and individual sentences. The lower portion of Table 1 thus shows the results of a mixed search in which text sections and paragraphs are retrieved instead of full texts whenever the query similarity for a shorter text passage exceeds the similarity for the full article. A number of new items are promoted into the top 10 list when text passages are retrievable, including section 33 of document 2974, "Blacks in the Americas," and section 97 of "United States of America." The text of document 2974.c33 (section 33 of document 2974) covers the founding of the American Anti-Slavery Society by William Lloyd Garrison in 1833. The relevance of this text to abolitionism and William Lloyd Garrison explains its good retrieval rank and high similarity coefficient of 0.50.

The available evidence indicates that when searching an encyclopedia, the use of the combined global and local similarity computations improves retrieval effectiveness by about 10% over the use of global vector similarity measurements alone. An additional 10% improvement is obtainable by use of the passage retrieval capability that identifies document excerpts in addition to full texts (6). The results obtained by extensive testing in the TREC (Text Retrieval Evaluation Conference) environment indicate that the Smart system produces consistently superior retrieval performance (7). Furthermore, response times are comparable to those obtainable in commercial retrieval environments. A Smart search of the TREC collections (700,000 full-text documents, or 2.4 gigabytes of text) has typical response times of 3 s for a 10-term query or 6 s for a 20-term query.

When text passages are available for processing and similarity measurements are easily computed between texts and text excerpts, text relation maps can be generated that show text similarities that exceed a particular threshold value. Figure 1 shows a relation map for four encyclopedia articles related to William Lloyd Garrison ("Slavery," "U.S. Civil War," "Abolitionists,"

Table 1. Text retrieval strategies. Query: article 9667, "William Lloyd Garrison." Section indicated by "c"; paragraph indicated by "p."

Document number	Query similarity	Title of retrieved item
Global text comparison only		
9667	1.00	Garrison, William Lloyd
18173	0.53	Phillips, Wendell
76	0.48	Abolitionists
21325	0.40	Slavery
827	0.36	American Anti-Slavery Society
21326	0.35	Slave Trade
8097	0.35	Emancipation Proclamation
9628	0.30	Gar
2883	0.27	Birney, James Gillespie
5584	0.27	Clay, Cassius Marcellus
Global-local text comparison and retrieval of text passages		
9667	1.00	Garrison, William Lloyd
18173	0.53	Phillips, Wendell
2974.c33*	0.50	Blacks in Americas
76	0.48	Abolitionists
21325.c8	0.42	Slavery
827	0.36	American Anti-Slavery Society
8097	0.35	Emancipation Proclamation
23173.c97*	0.31	United States of America
23545.p5*	0.29	Villard, Henry
5539.c28*	0.28	Civil War, American

*New article retrieved in restricted search.

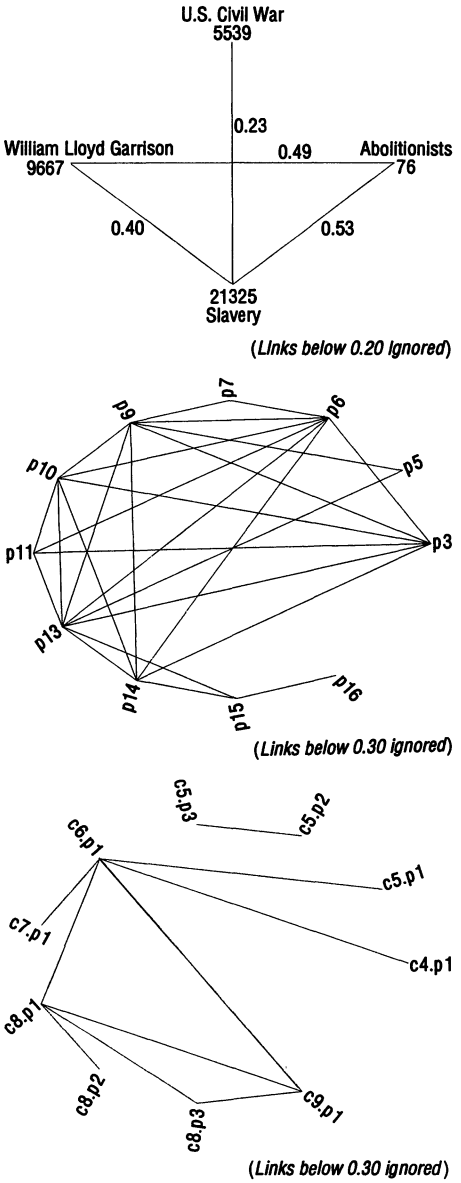


Fig. 1 (top). Basic text relation map. Vertices (nodes) represent texts; lines (links between nodes) represent text relations above a similarity threshold of 0.20. In all figures, "c" indicates section, "p" indicates paragraph. **Fig. 2 (middle).** Well-connected text relation map for paragraphs of article 21385, "Smoking." **Fig. 3 (bottom).** Poorly connected text relation map for paragraphs of article 21933, "Symphony."

and "Garrison"). The texts themselves are represented by nodes (vertices) of the map, and the pairwise text similarities are indicated by links (branches) between the corresponding node pairs. Figure 1 shows all similarities between full articles exceeding a similarity threshold of 0.20 (8). Text linking has been used in the past to build hypertext structures, but the links between related text pieces are normally assumed to be placed subjectively by individual text authors—a procedure manifestly impractical in environments where large masses of heterogeneous texts are stored for processing (9).

A study of various kinds of text relations between texts and text excerpts can reveal a good deal of information about the internal structure of individual texts, as well as the relations between different texts. Consider, as an example, the paragraph map for article 21385, "Smoking," shown in Fig. 2, which includes all pairwise paragraph similarities exceeding 0.30. In the corresponding graph, there are no disconnected components, and many similarities exist between adjacent paragraphs. The convex graph structure reflects a homogeneous treatment of the topic; in this case, the "Smoking" article emphasizes the health problems connected with smoking and the difficulties that arise when people attempt to quit smoking. For a homogeneous map such as this, it should be easy to determine the basic text content by looking at only a few carefully chosen paragraphs.

In contrast, consider the paragraph relation map in Fig. 3, which shows paragraph similarities for article 21933, "Symphony," and uses the same similarity threshold of 0.30. This map is much less dense; there are many outliers consisting of a single node only, and there is a disconnected component that includes paragraphs 2 and 3 of section 5. Clearly, the "Symphony" topic does not receive the same homogeneous treatment in the encyclopedia as "Smoking," and a determination of text content by selectively looking at particular text excerpts is much more problematic in this case. Attempts have been made in the past to relate certain manually linked hypertext structures to the corresponding text characteristics, but a detailed structural text analysis based on automatically linked structures at various levels of detail has not so far been undertaken (*10*).

In Figs. 1 to 3, the text nodes are equally spaced around the circumference of a circular structure. This makes it easy to recognize the links between individual text excerpts, but the actual link location in the running text is obscured. In particular, it is difficult to tell whether a link is placed at the beginning, in the middle, or at the end of a text. An alternative display format is shown in Fig. 4, in which the space assigned to each text along the circumference is proportional to the text length, and each text link is placed in its proper position within the texts. Figure 4 shows a paragraph map for four related articles ("Mohandas Gandhi," "Indira Gandhi," "Nehru," and "India") with the use of a similarity threshold of 0.30. It is obvious that the text of article 12017 ("India") is much longer than that of the other articles and that the coverage of Mohandas Gandhi (the Mahatma) is in turn more detailed than that of Indira Gandhi and Nehru.

Various kinds of topic relationships can be distinguished in Fig. 4, depending on the particular linking pattern between text elements. For example, when multiple links relate a particular (shorter) document such as "Indira Gandhi" (9619) and a subsection of a longer document such as "India" (12017), a narrower-broader text relation normally exists. Similarly, when a particular section of one document has multiple links to a particular section of another document, the two text items usually share a common subtopic. One can thus conclude that "Nehru" (16579) and the two "Gandhis" (9619 and 9620) represent subtopics of "India" (12017). Similarly, "Mohandas Gandhi" and "Nehru," and "Indira Gandhi" and "Nehru," are pairs of related documents that share common subtopics. Finally, the lives of the two Gandhis appear to be largely unrelated—a single linked paragraph pair exists that refers to unrest in India, a condition that plagued both politicians. The relation between Mohandas and Indira Gandhi is entirely through Nehru, who was a disciple of the Mahatma and also the father of Indira.

This type of analysis gives an objective view of the topic coverage in individual texts and of the information shared among sets of related texts. In the rest of this article, we examine three kinds of text analysis systems in more detail, which leads to the identification of text themes, the selective traversal of texts, and the summarization of text content by extraction of important text excerpts.

Text Theme Identification

A text theme can be defined as a specific subject that is discussed in some depth in a particular text or in a number of related texts. Themes represent centers of attention and cover subjects of principal interest to text authors and presumably also to text readers. The identification of text themes is useful for many purposes—for example, to obtain a snapshot of text content and as an aid in deciding whether actually to read a text.

Various approaches based on linguistic text analysis methods suggest themselves for the identification of text themes (*11*). In the present context, the text relation maps are used as inputs to a clustering process that is designed to identify groups of text excerpts that are closely related to each other but also relatively disconnected from the rest of the text (*12*). The following simple process leads to text theme identification: First, the triangles in the relation map are recognized (a triangle is a group of three text excerpts, each of which is related to the other two to a degree that is above the stated similarity threshold). A centroid vector is then constructed for each triangle, as the average vector for the group of three related items. Finally, triangles are merged into a common group (theme) whenever the corresponding centroids are sufficiently similar (that is, when the pairwise centroid similarity exceeds a stated threshold). Each theme may be represented by a global centroid vector that is constructed as the average vector of all text excerpts included in the theme.

Figure 5 shows the four themes derived by this method for the Gandhi-India subject area shown in Fig. 4. The following themes are apparent: (i) the single solid triangle consisting of paragraphs 9619.p5, 16579.p4, and 16579.p5 on the right-hand edge of Fig. 5 (main subject: Nehru); (ii) the single hashed triangle consisting of paragraphs 9619.p3, 12017.p219, and 12017.p220 (main subject: Sikhs, Punjab); (iii) the group of dark triangles consisting of paragraphs 9619.p7, 12017.p211, 12017.p216, 12017.p218, and 12017.p222 (main subject: Indira Gandhi); (iv) the group of light triangles consisting of paragraphs 9620.p3, 9620.p6, 9620.p8, 9620.p11, 9620.p14, 9620.p15, 9620.p18, 12017.p148, and 16579.p4 (main subject: Mohandas Gandhi). The clear separation between the two Gandhis already noted in the map of Fig. 4 is present also in the theme map of Fig. 5, in which no overlap exists between the dark and light triangle groupings.

An alternative, less onerous but also less refined theme generation method is to build a text relation map with the use of a high similarity threshold (where the number of linked text excerpts is small). Each disconnected component of the map, consisting of groups of highly related text excerpts, is then identified with a particular theme. The graph obtained by use of a text similarity threshold of 0.50 for the Gandhi-India subject area is shown in Fig. 6. The high similarity threshold reduces the similarity map to three areas, identified as Mohandas Gandhi (top theme), Indira Gandhi (middle), and Nehru (bottom). These themes duplicate those of Fig. 5, but the second theme in Fig. 5, which covers Indira Gandhi's problems with the Sikhs in Pun-

jab, is no longer recognized as a separate subject.

When text relation maps are used as the main input, themes can be generated at various levels of detail. The larger the text excerpts used for text grouping purposes, the wider in general is the scope of the corresponding themes. Contrariwise, when sentences and other short excerpts are used in the grouping process, the theme coverage is normally narrow. Thus, when themes are derived from the texts of documents 9667 and 76 ("William Lloyd Garrison" and "Abolitionists," respectively), a theme derived from paragraph relations might cover the "beginnings of U.S. abolitionism"; a more detailed theme derived from sentence relations might cover the "founding of the newspaper *Liberator*," which was a milestone in the early years of the abolitionist movement. By suitable variation of the scope of the theme generation process, it is thus possible to derive a smaller number of broader themes or a larger number of narrower themes.

Selective Text Traversal

When large text collections are in use, flexible methods should be available that will skim the texts while concentrating on text passages that may be of immediate interest. Such a skimming operation can then be used both for selective text traversal, in which only text passages deemed of special importance are actually retrieved or read, and for text summarization, in which summaries are constructed by extraction of selected text excerpts.

In selective text traversal (*13*), starting with a text relation map and a particular text excerpt of special interest, a user may follow three different traversal strategies: (i) The path may cover many of the central nodes, which are defined as nodes with a large number of links to other nodes of the map. (ii) The path may use text excerpts located in strategic positions within the corresponding documents—for example, the first paragraphs in each text section or the first sentences in each paragraph. (iii) The path may use the link weight as the main path generation criterion by starting with the desired initial node and choosing as the next node the one with maximum similarity to the current node. This last strategy is known as a depth-first search.

When individual text excerpts are selected for path formation or summarization, a number of factors must receive special attention; among these are the coherence of the resulting text, that is, the ease with which the text can be read and understood; the exhaustivity of coverage of the final text, that is, the degree to which all the main subject areas are covered; the text chronology, that is, the accuracy with which timing factors are recognized; and finally, the amount of repetition in the selected text excerpts. Some of these factors are handled relatively easily; for example, text chronology is often maintained by the use of only forward-pointing paths and backtracking is not allowed (if a particular paragraph is included in a path, no other text excerpt appearing earlier in the same document can appear in the same path).

In the present context, text coherence is used as the main criterion, and forward depth-first paths are used in which each chosen text excerpt is linked to the most similar text excerpt not yet seen at this point. In a depth-first path, each chosen excerpt is closely related to the next one, and the chance of poor transitions between selected paragraphs is minimized. Consider, as an example, the paragraph map in Fig. 7, which is based on six documents related to the Greek god Zeus (article 24674). The assumption is that the path starts with the initial text paragraph of "Zeus" (24674.p3). A short depth-first path may be defined as a single-link path that includes only the initial text excerpt plus the next most similar excerpt. In Fig. 7, this defines path 24674.p3 to 17232.p4 (paragraph 4 of the

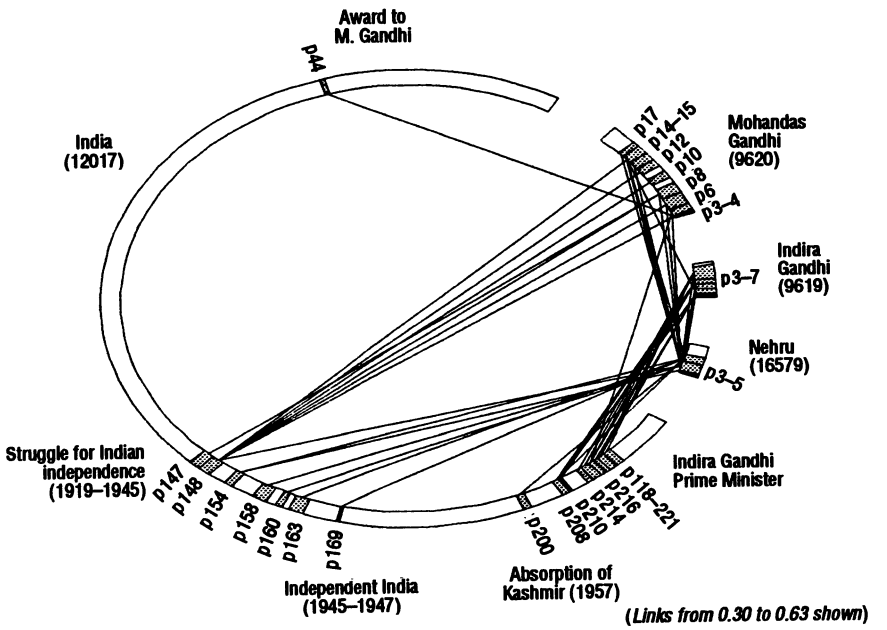


Fig. 4. Paragraph similarity map for articles related to "India" (12017). Length of curved segments is proportional to text length; links are placed in correct relative position within each text.

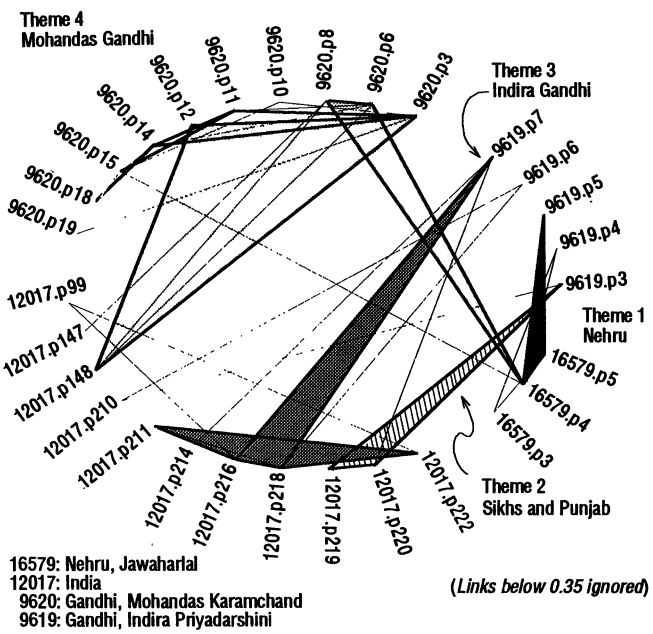


Fig. 5. Text themes derived by merging of triangles for four articles related to "India" (12017).

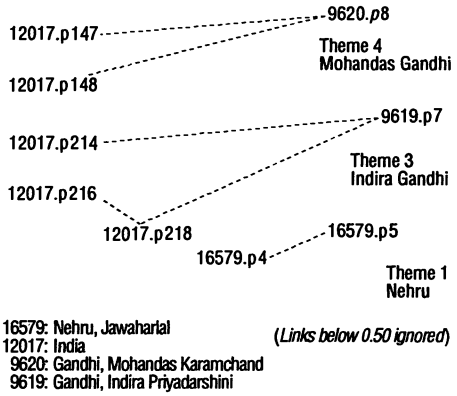
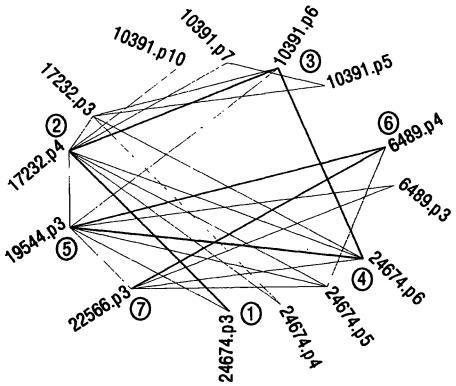


Fig. 6. Simplified text themes derived from high-threshold (disconnected) text relation map for articles related to "India" (12017).

article "Mount Olympus"). The corresponding paragraphs introduce Zeus as the god of the sky and the ruler of the Olympian gods and then proceed by identifying the 12 major Olympian deities, including Zeus, his wife Hera, and his siblings and children.

A more complete forward depth-first path proceeds from item 17232.p4 to include 10391.p6 (paragraph 6 of 10391, "Greek Religion and Mythology") and four additional paragraphs presented in detail in Fig. 7. The complete forward depth-first path includes information about Rhea, Zeus' mother; Cronus, Zeus' father; and the Titans, a race of giants that included Rhea and Cronus, among other gods.

Instead of initiating the text traversal at the beginning of a text, it is also possible for a searcher to use context-dependent text-traversal strategies that start with a special text excerpt of immediate interest. For example, someone interested in the foreign policy of President Nixon might locate paragraph 622 of article 23173 ("United States of America") by using a standard text search. A depth-first path starting at 23173.p622 can then be used to obtain further information. Such a path also includes paragraphs 9086.p13 (paragraph 13 of article 9086, "Gerald R. Ford") and 16855.p11 (paragraph 11 of 16855, "Richard M. Nixon"). The corresponding texts deal with the exchange of visits between President Nixon and Leonid Brezhnev; the continuation of detente between the United States and the Soviet Union that was pursued by President Ford and Secretary of State Kissinger; and finally, Nixon's approach to the People's Republic of China. A completely different topic will be covered by a depth-first path starting with paragraph 23173.p624, describing the Watergate break-in. The corresponding coverage includes Nixon's presumed implication in the Watergate burglary (23173.p624), Vice President Ford's staunch defense of Nixon during his term as vice president (9086.p8),



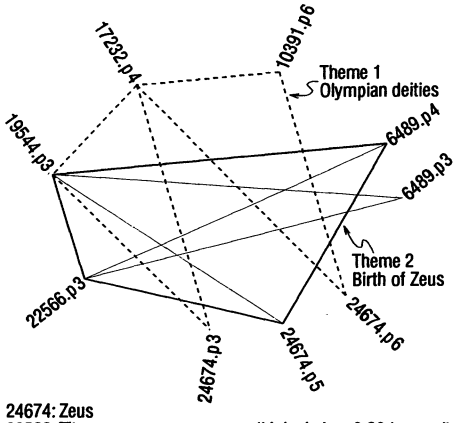


Fig. 7 (top). Depth-first paragraph-traversal order for six articles related to "Zeus" (24674). Path starts with initial paragraph of "Zeus" (24674.p3) and covers, in order, 17232.p4 ("Mount Olympus"), 10391.p6 ("Greek Religion and Mythology"), 24674.p6 ("Zeus"), 19544.p3 ("Rhea"), 6489.p4 ("Cronus"), and 22566.p3 ("Titans"). **Fig. 8 (bottom).** Two themes, "Olympian deities" (dashed triangles) and "birth of Zeus" (heavy lines), for articles related to "Zeus" (24674).

and finally, Nixon's resignation on 9 August 1974 and Ford's pardon (23848.p19).

Current experience indicates that a depth-first path provides a reasonably coherent body of information in practically every subject environment. The resulting paths may, however, be flawed in some ways. For example, there may be repetition of subject coverage in two or more excerpts in a given path; in the previous example, Nixon's resignation is mentioned in paragraphs 9086.p8 and 23848.p19. Repeated text passages may be eliminated by a sentence-sentence comparison, followed by the removal of duplicate occurrences of sufficiently similar sentences. Alternatively, a larger text excerpt in a path can sometimes be replaced by a shorter excerpt whose similarity to the previous text element is large. In the depth-first path of Fig. 7, the long paragraph 10391.p6 that deals with the divine hierarchy on Mount Olympus (node 3 in the figure) may be replaced by a group of three adjacent sentences (10391.g17) consisting of the first three sentences of the paragraph. Similarly, paragraph 22566.p13 is replaceable by sentence group 22566.g7, which includes only the last three sentences of the paragraph.

An alternative way of reducing the path size is to use theme generation methods to obtain text excerpts covering the desired subject area. Figure 8 shows a theme generation map obtained by triangle merging for the Zeus subject area used in Fig. 7. Two themes are distinguished, "Olympian deities" and "birth of Zeus." The two text excerpts that are most similar to the respective theme centroids are 17232.p4 ("The 12 major Olympian deities were Zeus and his wife Hera. . .") and 24674.p5 ("Zeus was the youngest son of the Titans Cronus and Rhea. . ."). An appropriate short path covering Zeus can then be obtained as 24674.p3 (the initial paragraph of the Zeus article), followed by 17232.p4 and 24674.p5, representing the most important paragraphs in the two themes, respectively.

Text Summarization

In the absence of deep linguistic analysis methods that are applicable to unrestricted subject areas, it is not possible to build intellectually satisfactory text summaries (*14*). However, by judicious text extraction methods, collections of text passages can be identified that provide adequate coverage of the subject areas of interest. For example, when homogeneous text relation maps are available, a good summary is normally obtainable by use of one of the longer text-traversal paths in chronological (forward) text order.

Consider, as an example, the paragraph map for document 16585 ("Horatio, Viscount Nelson") (Fig. 9). Two paths are shown that start with the initial text paragraph 16585.p3. The dashed path traverses all "bushy" nodes—that is, nodes in which the number of incident links is large (≥ 6). The path marked by a heavy line is a complex depth-first path—that is, a depth-first path obtained by starting at each of the bushy nodes, proceeding in depth-first order, and assembling the resulting excerpts into a single path in forward text order. The shorter, dashed path covers the highlights of Nelson's life in paragraphs 16585.p3, p6, p9, and p11, which deal, respectively, with a summary of Nelson's achievements as a British naval commander, his role in the battle of Copenhagen in 1801 after he had

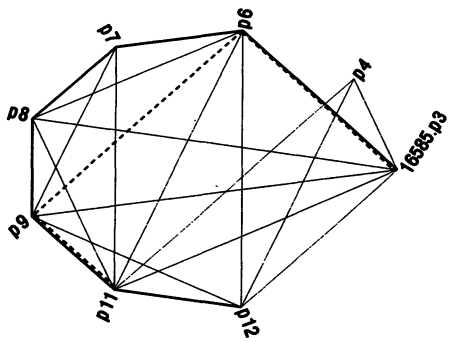


Fig. 9. Complex paths used for text summarization. Dashed path 16585.p3-p6-p9-p11 includes all bushy nodes; solid path 16585.p3-p6-p7-p8-p9-p11-p12 is a depth-first path.

become a vice admiral, and the crucial defeat of Napoleon during the battle of Trafalgar in 1805. The longer, solid path adds paragraphs p7, p8, and p12 of document 16585 to the paragraphs already present in the dashed path. This adds information about the battle of the Nile in 1798, plus a wrap-up paragraph covering Nelson's burial in St. Paul's Cathedral.

When the text relation map is substantially disconnected, the text-traversal process will not produce comprehensive summaries. In that case, adequate subject coverage is generally obtained by taking the initial paragraph of the main document under consideration, followed by the best paragraph for each text theme, as explained earlier. For the Zeus subject matter in Figs. 7 and 8, the resulting summary consists of paragraphs 24674.p3, 6489.p4, and 17232.p4. The corresponding summary introduces Zeus, the ruler of the Olympian Gods (24674.p3), mentions the story of the birth of Zeus as the sixth child of Cronus and Rhea (6489.p4), and terminates with an introduction to the 12 major Olympian deities (17232.p4).

When paths and themes are used for text summarization, longer summaries will be obtained when the text relation map is generated with low similarity thresholds. This produces denser maps with large numbers of text links. The themes may then partly overlap, and the summaries obtained by text extraction will be discursive. Contrariwise, when high similarity thresholds are used, the maps and themes tend to be disconnected and the summaries become sparser.

Conclusion

Formal evaluation data on the effectiveness of the methods introduced here are difficult to produce in the absence of detailed relevance information relating the content of many kinds of text excerpts to large numbers of subject queries. The experience accumulated with the wide-ranging subject matter in the Funk and Wagnalls encyclopedia indicates that useful output products are obtained in most cases. Because the approaches described here are robust and generally applicable to a wide variety of texts in many different environments, one may anticipate that such text-processing and knowledge-extraction capabilities will soon be widely used.

REFERENCES AND NOTES

1. M. Bernstein, J. D. Bolter, M. Joyce, E. Mylonas, in *Proceedings of Hypertext-91*, Association for Computing Machinery, San Antonio, TX, 15 to 18 December 1991 (ACM Press, Baltimore, MD), pp. 246–260; G. P. Landow, *Comput. Humanities* **23**, 173 (1989); J. D. Bolter, *Writing Space—The Computer, Hypertext, and the History of Writing* (Erlbaum, Hillsdale, NJ, 1991); P. Delaney and G. P. Landow, Eds., *Hypermedia and Literary Studies* (MIT Press, Cambridge, MA, 1991).
2. G. Salton, Ed., *The Smart Retrieval System—Experiments in Automatic Document Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1971); ———, C. S. Yang, A. Wong, *Commun. ACM* **18**, 613 (1975); G. Salton, *Automatic Text Processing—The Transformation, Analysis, and Retrieval of Information by Computer* (Addison-Wesley, Reading, MA, 1989); *Science* **253**, 974 (1991).
3. G. Salton and C. Buckley, *Science* **253**, 1012 (1991); ———, in *Proceedings of SIGIR-91—Fourteenth International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Chicago, IL, 13 to 16 October 1991, A. Bookstein, V. Chiaramella, G. Salton, V. V. Raghavan, Eds. (Association for Computing Machinery, New York, 1991), pp. 21–30.
4. An electronic version of the Funk and Wagnalls encyclopedia containing approximately 26,000 articles of text was used as a sample database in this study.
5. J. O'Connor, *Inf. Process. Manage.* 11, 155 (1975); *J. Am. Soc. Inf. Sci.* **32**, 227 (1980); S. Al-Hawamdeh and P. Willett, *Electron. Publ.* **2**, 179 (1989).
6. G. Salton, C. Buckley, J. Allan, *Electron. Publ.* **5**, 1 (1992); G. Salton, J. Allan, C. Buckley, in *Proceedings of SIGIR-93—Sixteenth International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, 27 June to 1 July 1993, R. Karfhage, E. Rasmussen, P. Willet, Eds. (Association for Computing Machinery, New York, 1993), pp. 49–58.
7. C. Buckley, G. Salton, J. Allan, in *The First Text Retrieval Conference*, D. K. Harman, Ed. (NIST Special Publication 500-207, Government Printing Office, Washington, DC, 1993), pp. 59–72; C. Buckley, J. Allan, G. Salton, in *The Second Text Retrieval Conference*, D. K. Harman, Ed. (NIST Special Publication, Government Printing Office, Washington, DC, in press).
8. All text relation maps in this study are based on global text similarity as well as local context check restrictions. The similarity thresholds used to construct the text relation maps can be chosen so that the number of links does not greatly exceed the number of nodes appearing in the maps.
9. M. H. Anderson, J. Nielsen, H. Rasmussen, *Hypermedia* 1, 255 (1989); M. Bernstein, in *Proceedings of the European Conference on Hypertext*, Versailles, France, November 1990, A. Rizk, N. Streitz, J. Andre, Eds. (Cambridge Univ. Press, New York, 1990), pp. 212–223; M. H. Chignell, B. Nordhausen, J. F. Valdez, J. A. Waterworth, *Hypermedia* **3**, 187 (1991); R. Furuta, C. Pleasant, B. Shneiderman, *ibid.* 1, 179 (1989); P. Gloor, in *Proceedings of Hypertext-91—Third ACM Conference on Hypertext*, San Antonio, TX, 15 to 18 December 1991 (ACM Press, Baltimore, MD, 1991), pp. 107–121; T. C. Rearick, in *Hypertext/Hypermedia Handbook*, J. Devlin and E. Berk, Eds. (McGraw-Hill, New York, 1991), pp. 113–140.
10. R. A. Botafogo, E. Rivlin, B. Shneiderman, *ACM Trans. Inf. Sys.* **10**, 142 (1992); R. S. Gilyarevskii and M. M. Subbotin, *J. Am. Soc. Inf. Sci.* **44**, 185 (1993); C. Guinan and A. F. Smeaton, in *Proceedings of ECHT-92—ACM-ECHT Conference*, Milan, Italy, 30 November to 4 December 1992 (ACM Press, Baltimore, MD, 1992), pp. 122–130.
11. M. A. Hearst and C. Plaunt, in *Proceedings of SIGIR-93—Sixteenth International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, 27 June to 1 July 1993, R. Korfhage, E. Rasmussen, P. Willet, Eds. (Association for Computing Machinery, New York, 1993), pp. 55–68.
12. F. Murtagh, *Comput. J.* **26**, 354 (1982); W. B. Croft, *J. Am. Soc. Inf. Sci.* **28**, 341 (1977); G. Salton and A. Wong, *ACM Trans. Database Syst.* **3**, 321 (1978).
13. G. de Jong, in *Strategies for Natural Language Processing*, W. G. Lehnert and M. H. Ringle, Eds. (Erlbaum, Hillsdale, NJ, 1982), pp. 149–176.
14. H. P. Luhn, *IBM J. Res. Dev.* **2**, 159 (1958); H. P. Edmundson and R. E. Wyllys, *Commun. ACM* **4**, 226 (1961); C. D. Paice, *Inf. Process. Manage.* **26**, 171 (1990); J. E. Rush, R. Salvador, A. Zamora, *J. Am. Soc. Inf. Sci.* **22**, 260 (1964).

15. The authors are grateful to the Microsoft Corporation for making the Funk and Wagnalls encyclopedia available in machine-readable form. Supported in part by NSF grant IRI 93-00124.