**Selective Text Utilization and
Text Traversal**

Gerard Salton*
James Allan

TR 93-1366
July 1993

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.

Selective Text Utilization and Text Traversal

Gerard Salton*

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501
Telephone: 607-255-4117
Fax: 607-255-4428
gs@cs.cornell.edu

James Allan

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501
allan@cs.cornell.edu

ABSTRACT

Many large collections of full-text documents are currently stored in machine-readable form and processed automatically in various ways. These collections may include different types of documents, such as messages, research articles, and books, and the subject matter may vary widely. To process such collections, robust text analysis methods must be used, capable of handling materials in arbitrary subject areas, and flexible access must be provided to texts and text excerpts of varying size.

In this study, global text comparison methods are used to identify similarities between text elements, followed by local context-checking operations that resolve ambiguities and distinguish superficially similar texts from texts that actually cover identical topics. A linked text structure is then created that relates similar texts at various levels of detail. In particular, text links are available for full texts, as well as text sections, paragraphs, and sentence groups. The linked structures are usable to identify important text passages, to traverse texts selectively both within particular documents and between documents, and to provide flexible text access to large text collections in response to various kinds of user needs. An automated 29-volume encyclopedia is used as an example to illustrate the text accessing and traversal operations.

KEYWORDS

full-text access, information retrieval, passage retrieval, text analysis, global text comparisons, local context checking, automatic text linking, selective text reading, text summarization.

INTRODUCTION

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

The conventional wisdom is that sophisticated conceptual text representations are needed for information retrieval purposes, including the use of thesauruses (synonym dictionaries) tailored to particular subject areas, and of preconstructed knowledge structures that classify the main entities of interest in a given subject, and specify the relationships that may hold between the entities in particular areas of application. However, many conceptual and

* This study was supported in part by the National Science Foundation under grant IRI 89-15847.

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

Fortunately, when the text collections are available in machine-readable form, alternative *corpus-based* text handling systems may be implemented that use the text collections themselves to derive the information necessary for analysis and text characterization. In particular, by examining the local contexts in which words and expressions are used, it is possible to recognize linguistic ambiguities of many kinds, and hence to avoid most false retrievals. Furthermore, by providing methods for accessing individual text components, relevant text passages are often identifiable that can be extracted from the longer, more heterogeneous texts in which they are embedded.

The problem of dealing with large literatures and determining text structure has been treated before [Bern91,Bota92], and proposals have been made for automatically generating links between related text pieces and constructing paths through such a linked structure [Bern90,Guin92]. The present study extends this earlier work, first by using a robust text analysis system that may be expected to produce high-quality links between text excerpts, and then by making the linked text structure available at various levels of detail. The resulting hypertext offers flexible methods for selective text utilization and traversal by making it possible to access text excerpts of appropriate length in accordance with particular user needs.

TEXT ANALYSIS AND TEXT MATCHING

In conventional information retrieval environments, keywords, or terms, are manually or automatically assigned to the information items, and queries are formulated by using terms interconnected by Boolean operators. Although widely used, the Boolean retrieval model is not ideally suited to the retrieval task: users find it hard to generate effective Boolean queries that will retrieve just the right type and amount of information; the retrieved items are presented to the users in a random order that does not correspond to any presumed order of relevance or usefulness; and term weights reflecting term importance are awkward to incorporate into Boolean systems in a consistent way. Most importantly, the operations of Boolean logic are unforgiving and inflexible, and the retrieval results are often inadequate. [Salt83,Salt91c]

The vector-processing model represents an alternative possibility for handling information retrieval operations. In that case, both the stored documents as well as the search requests are represented by sets of terms (term vectors) without Boolean operators. Different vectors can be compared with each other and vector similarity coefficients are obtainable reflecting similarity in the term assignments for different vectors. In the vector processing model of retrieval, the same methods are usable for collection structuring (by comparing pairs of document vectors with each other and identifying document pairs found to be sufficiently similar), and for information retrieval (by comparing query vectors with the vectors representing the stored items and retrieving items found to be similar to the queries). The results of a similarity computation between a query vector and the stored document vectors can be ranked in decreasing order of the computed query similarity. This makes it possible to retrieve the most important items (those most similar to the user queries) first. Furthermore, term weights are easily accommodated because vectors of weighted terms are manipulated almost as easily as binary term vectors (where weights are restricted to 1 for assigned terms and 0 for missing terms). [Salt75,Salt71]

A high performance term weighting system assigns large weights to terms that occur frequently in particular documents, but rarely on the outside, because such terms are able to distinguish the items in which they occur from the remainder of the collection. A typical term weight of this type, known as a *tf x idf* weight (term frequency times inverse document frequency), may be defined as

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

where tf_{ik} is the frequency of occurrence of term T_k in document D_i , $tf_{ik} = 0$ for terms not assigned to D_i , N is the size of the document collection, and n_k represents the number of documents in the collection with term T_k . The summation in the denominator, taken over all terms in a particular vector, is used for length normalization purposes to insure that all documents have equal chance of being retrieved. (Without length normalization, the longer

documents with more assigned terms and higher term frequencies would generate higher document similarities, and exhibit higher retrieval potential than the shorter items.) [Salt88]

Given query and document vectors $Q_j = (w_{j1}, w_{j2}, \dots, w_{jt})$ and $D_i = (w_{i1}, w_{i2}, \dots, w_{it})$, respectively, or alternatively, given two particular document vectors, a vector similarity function of the form:

$$\text{sim}(Q_i, D_j) = \sum_{k=1}^t w_{jk} w_{ik} \quad (2)$$

may be used reflecting the similarity between the corresponding term vectors. When normalized term weights are used, such as those of expression (1), the vector similarity lies between 0 and 1 and depends on the proportion and the weight of matching terms in the vectors. In retrieval systems such as the Smart system [Salt71], the terms included in the document vectors are typically word stems, or phrases, extracted from the corresponding texts. Normally, the individual text words are recognized, common words included on a special "stop list" are removed, and word stems are generated by a suffix deletion process. The remaining word stems are then weighted using a formula such as that of expression (1). Analogously, the query terms may be word stems extracted from a user's natural-language query formulation, or stems included in the texts of documents that are used for query formulation purposes.

Two main difficulties must be faced when text similarities are obtained by simple vector comparison methods: word or word stem comparisons are inherently inadequate because a given term or expression may be used in many different environments, and may take on many different meanings; furthermore, maintaining the integrity of complete documents is unreasonable in situations where specific queries are processed against collections that often contain long, book-size documents. (Long documents are difficult to retrieve when normalized term vectors such as those obtained from expressions (1) and (2) are used, because the proportion of matching terms will be small for documents containing many terms. Furthermore, providing the user with a long document in response to a specific information request hardly ever responds to a real user need.)

To handle the text ambiguity problem, appeal is made to the "use theory" of meaning, which states that the meaning of words and expressions depends on their use in the language. [Witt53] Operationally, this implies that the local contexts must be examined in which text words are used, and that similar word forms should be accepted as related only when their local contexts are sufficiently similar. In the Smart system environment, a restricted retrieval strategy is usable, where document texts are retrieved only when there is a sufficiently high global vector similarity with the query text, and when document and query texts also contain locally matching text components, such as pairs of similar sentences. Local context similarities can be derived by vector matching methods similar to those described earlier for full texts, except that the term vectors being compared now represent small document components rather than complete documents. [Salt91a, Salt91b]

The need for text passage retrieval as an aid in controlling text heterogeneity has been recognized for a long time. [OCon75, OCon80] In the Smart system, a "mixed" retrieval strategy has been implemented which rejects texts or text excerpts that do not satisfy the local context check, and additionally breaks down the longer texts into text sections and paragraphs. The retrieved material then consists variously of full texts, text sections, or text paragraphs, depending on which of these text elements exhibits the highest similarity with the query text. The mixed output is obtained by comparing various text components with the query texts and choosing a text component for retrieval when the corresponding query similarity is higher than the query similarity for the full text. [Salt93]

The operations of the global/local vector comparison system may be illustrated by using searches conducted with an automated version of the 29-volume (65 Mbytes) Funk and Wagnalls New Encyclopedia [Funk79]. The output of Figure 1 shows retrieval results for both the unrestricted, global vector comparison system, and the restricted, mixed retrieval strategy, using as a query the text of encyclopedia article 9667 dealing with William Lloyd Garrison (the American abolitionist). The output consists in each case of the document number, the query-document similarity, and the title for each retrieved item. In Figure 1, the top 20 retrieved items are shown for each search in decreasing order of the query-document similarity.

The results in Figure 1(a) are obtained by an unrestricted search using only the global vector comparisons without local context checking, and retrieving full documents only. With two exceptions, the retrieved output consists correctly of articles covering leaders of the abolitionist movement related to William Lloyd Garrison, or dealing with the anti-slavery movement. The erroneously retrieved items are 9628 Gar (a type of fish), and 10277 Grand Army of the Republic (also abbreviated GAR). These items are obtained in an unrestricted global search because Garrison matches Gar after suffix removal, and the corresponding texts contain a number of other matching terms. (The suffix

ison is first deleted from Garrison as in *comparison/compar*, followed by the removal of the duplicated consonant in *Garr* as is necessary to reduce *hotter* to *hot*, or *stopping* to *stop*.)

Num	Sim	Title
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
10277	0.27	Grand Army of the Republic
14104	0.26	Lincoln, Abraham
2974	0.25	Blacks in the Americas
9454	0.24	Fugitive Slave Laws
24162	0.24	Wilberforce, William
3713	0.24	Brown, John
7476	0.21	Douglass, Frederick
5949	0.21	Compromise Measures of 1850
6435	0.21	Creole Case
3963	0.20	Burns, Anthony

Figure 1a. Unrestricted Search (Query 9667)
(full document output)

Num	Sim	Title
9667	1.00	Garrison, William Lloyd
18173	0.53	Phillips, Wendell
2974.c33	0.50	Blacks in the Americas
76	0.48	Abolitionists
21325.c8	0.42	Slavery
827	0.36	American Anti-Slavery Society
8097	0.35	Emancipation Proclamation
new 23173.c97	0.31	United States of America
new 23545.p5	0.29	Villard, Henry
new 5539.c28	0.28	Civil War, American
24162.p4	0.28	Wilberforce, William
2883	0.27	Birney, James Gillespie
5584	0.27	Clay, Cassius Marcellus
14104.c10	0.27	Lincoln, Abraham
new 1120.p6	0.25	Anthony, Susan B(rownell)
9454	0.25	Fugitive Slave Laws
7476.p5	0.24	Douglass, Frederick
new 24134.p3	0.23	Whittier, John Greenleaf
5949	0.21	Compromise Measures of 1850
new 22914	0.18	Truth, Sojourner

Figure 1b. Restricted Search (Query 9667)
(mixed output: ci section i; pj paragraph j)

Without performing the local context checks it is difficult to avoid errors of the type illustrated in Figure 1(a) when only simple vector matches are used. Such errors are however absent from the restricted output of Figure 1(b), that includes the local context checking feature. The output of Figure 1(b) consists of certain full documents (identified by the lack of a suffix following the document number), some document sections (suffix *ci*), and some text paragraphs (suffix *pj*). It is easy to see that the mixed output contains many partial documents not contained in the original list of retrieved items. Thus, the full item *United States of America* consisting of 209 text sections was correctly not retrieved in the unrestricted search, but section 97 of that article entitled *USA/History/The Crisis of Disunion* (23173.c97) is properly obtained by the mixed retrieval strategy. Furthermore, some originally retrieved items receive a much higher query similarity in the mixed output when a particularly relevant text excerpt is located. Thus, the full text of *Blacks in the Americas* (2974) consisting of 60 text sections has a global similarity of 0.25 with the query text on William Lloyd Garrison (9667), but section 33 of that article, entitled *Blacks in the*

Americas/North America/The Slave Era/Abolitionist Movement , reaches a higher query similarity of 0.50. (A slash (/) is used to separate section and subsection titles). Both output lists of Figure 1 contain article 9667 used as a query at the head of the retrieved list with a perfect query similarity of 1.00.

SELECTIVE TEXT UTILIZATION

It is well-known that most texts currently processed by computer are not meant to be read sequentially from one end to the other. A preferred text utilization strategy consists instead in finding a useful text excerpt and proceeding from there to additional interesting excerpts. Since an automated text accessing system must serve a diverse user population with different interests and background, a variety of different accessing methods must be provided to satisfy the varying user needs, thereby creating a democratic situation where wide audiences with diverse needs can be adequately serviced. [Bolt90,Dela91] The procedures outlined in the previous section can be used to build flexible text utilization systems where particular text elements can be isolated, and texts can be read or traversed selectively depending on the context and on user interests.

Consider first the important problem of text summarization. It is well-known that the automatic construction of text summaries and text abstracts is exceedingly complex. The known simple abstracting methods based on sentence extraction often produce incoherent text [Luhn58,Edmu64,Rush64,Earl70,Baxe58,Paic90], and more complex knowledge-based methods are unusable with general-purpose unrestricted text collections. An alternative approach to text summarization may depend on the placement of particular texts in a wider context, that is, in relation to other texts covering related topics. A reasonable summary of the original text might then be obtained by choosing text passages that exhibit a large amount of overlap, or a large number of relationships, within the chosen context.

Consider a user interested in the influence of Garibaldi, the Italian revolutionary, on the history of Italy. The original encyclopedia article on Garibaldi contains 5 text sections subdivided into 10 paragraphs. When this article (number 9652) is used as a search request, four closely related items are located, including 23502 Victor Emmanuel II, 12364 Italy, 9248 Francis II, and 4425 Caprera. Figure 2(a) is a graphic representation of all section similarities exceeding 0.30 for the 5 documents under discussion. Each vertex of the graph represents a text section, and an arc between two vertices represents a section similarity exceeding the stated threshold between the corresponding pair of vertices. For the Garibaldi article, section 9652.c5 exhibits the largest number of outside connections to other sections in related documents (5 in all), and the sum of the link weights is maximized. This section is likely to be of central importance in covering the subject matter under discussion.

Figure 2a. Section Similarities with Central Text Excerpt (9652.c5)

The text of section 9652.c5 is reproduced in Figure 2(b) with the connections to the outside context noted in the margin. It is clear from the text that section 5 of article 9652 indeed covers the essence of Garibaldi's importance in the struggle for Italian unification. One or more central text elements may serve for text summarization purposes, as well as for text traversal as will be seen.

Garibaldi, Giuseppe
Struggle for Unification.

Garibaldi was deeply involved in the complicated military and political struggles that took place in the following years. In 1859 he led a successful expedition against the Austrian forces in the Alps; in 1860 he led a force of 1000 men from Genoa to Sicily, then ruled by the king of Naples. Distinctively clad in bright red shirts, Garibaldi's men became known as the Red Shirts and as The Thousand. Between May and August 1860, Garibaldi conquered Sicily and set up a provisional insular government. He then crossed to the Italian mainland; took Naples; defeated the Neapolitans in a decisive engagement on the banks of the Volturno River on October 26, 1860; and besieged the fortress of Gaeta, which fell in February 1861.

Later that year, the kingdom of Italy was established with Victor Emmanuel as king; Rome, a papal possession garrisoned by French troops, was not included in the new kingdom, nor were areas in the north of the peninsula held by the Austrians. Garibaldi, declining all honors and positions in the new kingdom, retired to his island home on Caprera. In the following year, however, he organized the Society for the Emancipation of Italy and visited Sicily, where he raised a force of volunteers with the object of capturing Rome and including it in a unified Italian state. He was opposed by Victor Emmanuel, who defeated him at the Battle of Aspromonte, August 29, 1862. Garibaldi was wounded and captured in that battle but was soon pardoned and released.

In 1866, despite the opposition of the Italian government, Garibaldi again raised a volunteer force with the aim of annexing the Papal States to the kingdom of Italy. After a number of initial engagements, he was defeated by combined papal and French forces at the Battle of Mentana on November 3, 1867. He was taken prisoner but was held only a short time. For about two years thereafter Garibaldi lived the life of a farmer on Caprera. In 1870 he offered his services to the French government and fought with his two sons in the Franco-Prussian War. Rome was annexed to Italy in October 1870, and Garibaldi was elected a member of the Italian parliament in 1874. In his last years he sympathized with the developing socialist movement in Italy and other countries. Garibaldi died on Caprera on June 2, 1882. His autobiography was published in 1887.

**12364.c49 Italy/
Garibaldi and Cavour**

9248 Francis II

**23502 Victor
Emmanuel II**

4425 Caprera

**12364.c50 Italy/The
Kingdom of Italy**

4425 Caprera

Figure 2b. Text of Central Text Excerpt (9652.c5)

When text excerpts are individually accessible, it is obvious that the text content can be utilized at various levels of detail. For the longer documents, it may then be especially useful to offer a sequence of accessing mechanisms that provides increasingly more refined contexts. Consider again article 9667 on William Lloyd Garrison previously used in the searches of Figure 1. This text is related to a number of other long documents, including 21325 Slavery (30 paragraphs of text) and 76 Abolitionists (12 paragraphs of text). The three documents (76, 9667, and 21325) form a cluster of related articles for which the pairwise similarity between any two of the articles exceeds 0.30 in all cases. Although, these items are obviously mutually relevant, it may not be useful to retrieve or jointly to display the long texts of these three items.

Figure 3a. Section Similarities Exceeding 0.30 (76.c4, 9667.c3, 21324.c8)

Increasingly more refined contexts are displayed for this cluster of related items in Figure 3. Figure 3(a) shows all section similarities exceeding a similarity threshold of 0.30. The corresponding paragraph similarities are similarly included in Figure 3(b), and similarities between groups of three adjacent sentences (sentence groups) are included in Figure 3(c). The similarity thresholds can be chosen automatically in such a way that the number of links between nodes is not much larger than the number of nodes. When too many links are defined, the user rapidly becomes overburdened. In general, this implies that the threshold increases as the text excerpts represented by the nodes become smaller.

Figure 3b. Paragraph Similarities Exceeding 0.25 (76.p6, 9667.p8, 21325.p24)

The darkened triangles visible in Figure 3 represent related sets of excerpts included in the next higher hierarchical level. That is, the three related sentence groups identified in Figure 3(c) (sentences 12-14 of article 76, 25-27 of 9667, and 70-72 of article 21325), are included in the respective paragraphs of Figure 3(b) (76.p6, 9667.p8, and 21325.p24), which are in turn included in the related sections 76.c4, 9667.c3, and 21325.c8 of Figure 3(a). When a sequence of hierarchically arranged narrower or wider contexts is provided, users can easily move from one context to another. For example, if paragraph relationships are used initially, broader contexts can be reached by using the section, or full document relations that contain the related paragraphs that were originally identified. Alternatively, narrower contexts are defined by using sentence groups or individual sentences included in the originally available paragraphs. This may make it possible to furnish just the right amount and the right depth of material satisfying particular user needs.

Figure 3c. Sentence Group Similarities Exceeding 0.27 (76.s12, 9667.s25, 21325.s70)

Of greatest interest in a text retrieval setting is the strategy used to traverse or use a given text, or series of related texts. Assuming once again that relationships are identified between texts and text excerpts, a number of different strategies suggest themselves for selective text traversal. Consider first the problem of intra-document traversal, where a single text needs to be accessed selectively. The following traversal orders may be considered among others:

- Choose the longest possible path, that is, the sequence with the largest number of excerpts, such that each excerpt is related above some threshold to the immediately following excerpt. Excerpts are included in the path in chronological order of appearance in the full text.
- Alternatively, choose the longest path as defined under (a) that also contains a central text element (an excerpt with the largest number of related excerpts).
- As a third possibility choose the longest path that also includes the initial text excerpt, that is the excerpt covering the beginning of the text under consideration.
- Finally, instead of using path length as a principal criterion for defining the traversal order, it may be useful to consider the branch similarities by choosing the path that maximizes the average text similarity between adjacent text excerpts.

In practice, the given criteria may not all lead to distinct paths. In particular criteria (a) to (c) often define one and the same path. Consider as an example, the set of related paragraphs of article 9620 Mohandas Gandhi shown in Figure 4(a), using a similarity threshold of 0.30. A traversal path chosen according to criterion (c) above is shown in Figure 4(a) including paragraphs 3, 11, 15, and 18 in order. The actual text of the chosen excerpts is given in Figure 4(b). As the example shows, the chosen traversal provides a reasonable picture of the main stages of Gandhi's life: covered successively are the main activities and influence of Gandhi (paragraph 3), Gandhi's early work in the 1920s (paragraph 11), his political activities during the struggle for Indian independence (paragraph 15), and finally his activities after independence (paragraph 18). The traversal order of Figure 4(a) reduces the 19 text paragraphs contained in the full article 9620 to the chosen four included in the path.

The methods used for selective intra-document traversal are applicable also for inter-document traversal when excerpts must be chosen from a variety of related, but different texts. Figure 5 shows a path of related paragraphs for texts connected to article 9562 Garibaldi. The chosen path contains paragraphs from the largest possible number of outside documents that also includes the central paragraph for Garibaldi (9652.p7). This path covers paragraphs from articles 4425 Caprera (where Garibaldi lived), 9248 Francis II King of Naples (who was deposed by Garibaldi), 23502 Victor Emmanuel II (whom Garibaldi supported), and of course 12364 Italy.

Figure 4a. Intradocument Paragraph Traversal (9620 Gandhi)
(9620.p3-p11-p15-p18)

----- Paragraph 9620.p3 -----

Gandhi, Mohandas Karamchand, called Mahatma Gandhi (1869-1948), Indian nationalist leader, who established his country's freedom through a nonviolent revolution.

----- Paragraph 9620.p11 -----

The Mahatma's political and spiritual hold on India was so great that the British authorities dared not interfere with him. In 1921 the Indian National Congress, the group that spearheaded the movement for nationhood, gave Gandhi complete executive authority, with the right of naming his own successor. The Indian population, however, could not fully comprehend the unworldly ahimsa. A series of armed revolts against Great Britain broke out, culminating in such violence that Gandhi confessed the failure of the civil-disobedience campaign he had called, and ended it. The British government again seized and imprisoned him in 1922.

----- Paragraph 9620.p15 -----

In 1934 Gandhi formally resigned from politics, being replaced as leader of the Congress party by Jawaharlal Nehru. Gandhi traveled through India, teaching ahimsa and demanding eradication of "untouchability." The esteem in which he was held was the measure of his political power. So great was this power that the limited home rule granted by the British in 1935 could not be implemented until Gandhi approved it. A few years later, in 1939, he again returned to active political life because of the pending federation of Indian principalities with the rest of India. His first act was a fast, designed to force the ruler of the state of Rajkot to modify his autocratic rule. Public unrest caused by the fast was so great that the colonial government intervened; the demands were granted. The Mahatma again became the most important political figure in India.

----- Paragraph 9620.p18 -----

By 1944 the Indian struggle for independence was in its final stages, the British government having agreed to independence on condition that the two contending nationalist groups, the Muslim League and the Congress party, should resolve their differences. Gandhi stood steadfastly against the partition of India but ultimately had to agree, in the hope that internal peace would be achieved after the Muslim demand for separation had been satisfied. India and Pakistan became separate states when the British granted India its independence in 1947. During the riots that followed the partition of India, Gandhi pleaded with Hindus and Muslims to live together peacefully. Riots engulfed Calcutta, one of the largest cities in India, and the Mahatma fasted until disturbances ceased. On January 13, 1948, he undertook another successful fast in New Delhi to bring about peace, but on January 30, 12 days after the termination of that fast, as he was on his way to his evening prayer meeting, he was assassinated by a fanatic Hindu.

Figure 4b. Typical Intradocument Paragraph Selection (9620 Gandhi)

Figure 5. Typical Interdocument Text Traversal (Paragraph Similarities)
(9428.p3 -- 9652.p3 -- 9652.p7 -- 23502.p3 -- 12364.p112)

Different paths or traversal orders are obtainable by varying the context in accordance with expressed user interests. Figure 6 shows relationships for groups of three adjacent sentences included in five related articles dealing with Gandhi and India. Only sentence groups with a high pairwise similarity of 0.45 are shown in Figure 6. Included in the figure are the two articles 9619 Indira Gandhi and 9620 Mohandas Gandhi. Although the two Gandhis are not directly related, the corresponding texts are formally connected by a third document 16579 Nehru. (Nehru was Indira Gandhi's father, and also a close associate of Mohandas Gandhi during the struggle for Indian independence.)

Figure 6. Context-Dependent Text Access
(12017 India, 9619 I. Gandhi, 9620 M. Gandhi)

While the vocabulary similarities between the two Gandhi articles are relatively high -- both personages were political leaders in 20th century India, and both are named Gandhi -- the sentence group similarities of Figure 6 clearly reveal the separation of the contexts. The similarities between the long article on India (number 12017) and

Mohandas Gandhi lie in the vicinity of sentence groups 509 to 519. These sentences are included in a section entitled *India/History/Gandhi's Protest Movement*. The Mohandas Gandhi article (9620) is also the only one with links to 5531 Civil Disobedience. On the other hand, when India is accessed in the context of document 9619 Indira Gandhi, the related sentence groups 765 to 776 occur much later in the India document, being located in a section entitled *India/History/Gandhi Returns* (referring to the return to power of Indira Gandhi in the late 1970s after her party had been defeated in elections two years earlier).

The text similarity analysis illustrated in the examples of Figures 2 to 6 thus offers accurate text access at different levels of detail to users with a wide variety of different interests and needs.

EVALUATION OF SELECTIVE TEXT ACCESS

The usefulness of a selective text traversal strategy can be evaluated by using the collection of text excerpts constituting each given traversal path to query the available document collection, and judging for relevance each text, or text excerpt retrieved in answer to such a path query. This makes it possible to compute the relative recall and precision figures that are normally used to assess information retrieval performance. In the encyclopedia environment a correct recall-precision computation must be based on relevance information of each retrieved text excerpt with respect to each text passage included in a traversal path. Unfortunately, such detailed relevance information between text excerpts of various kinds is not normally available in any text searching environment, and particularly not for the encyclopedia searches.

However, full encyclopedia searches were made at an earlier time using each full encyclopedia article for query purposes and retrieving other related full encyclopedia articles. Complete relevance information is now available concerning the appropriateness of each retrieved full article with respect to these full-text searches.¹ The available relevance assessments for the retrieval runs using full articles may be applied to the selective text traversal situation by assuming that a document excerpt is relevant to a query excerpt whenever the respective full-text items are relevant to each other. Such an assumption may be good enough to obtain reasonable ballpark figures for various selective text traversal strategies. Table 1 contains an evaluation of intra-document text traversal methods using linked paragraph structures for query formulation purposes. The table contains average relative recall and precision figures obtained for about 100 encyclopedia queries chosen in such a way that each query article contains at least two paragraphs with a pairwise paragraph similarity exceeding 0.30.² (The number of queries used for the different runs of Table 1 varies slightly because some query formulations did not lead to any retrieved text items.)

The output of Table 1 covers intra-document traversal where the query formulation consists of one or more paragraphs included in a single encyclopedia article. The first two columns contain data for single-paragraph queries (either the first paragraph of an article, or the central paragraph with the greatest number of significant links to other paragraphs in the article). The next three columns (runs 3, 4, and 5) cover queries formed from selective traversal paths (either the longest path through the paragraph structure, or the longest path that includes a central paragraph, or the path containing at least four paragraphs with a maximum average similarity between adjacent paragraphs). Finally, column 6 contains results for full document queries for which accurate relevance data are available.

The overall performance is reflected in the 11-point recall-precision average which gives average precision figures computed over the query set for 11 different recall levels ranging from 0 to 1 in steps of 0.1. Table 1 shows that much better performance is obtainable with the selective path queries than with a query consisting of a single paragraph only --- the advantage is in the range from 40 to 45 percent in the 11 point average. This demonstrates that the paragraph relationships used to define the traversal paths are in fact useful in specifying related relevant information. The last two lines of Table 1 also show that the deterioration in the average retrieval performance of the selected query paragraphs is only about 25 percent compared with the performance of queries consisting of the full article texts, although the average query length for the paragraph paths is only about 30 percent of the length of the full articles. (The average paragraph path length is about 3300 text characters compared with an average of about

¹ The writers are grateful to Microsoft Corporation for furnishing relevance data for the full-text encyclopedia searches.

² Relative recall is the proportion of all relevant items retrieved as a percentage of all known relevant in the collection. Precision is the proportion of retrieved items that are relevant. [Salt83,Salt89]

11,600 characters for the full text queries.) Substantially improved overall evaluation figures and much better discrimination among different traversal paths could be obtained if accurate relevance data were available for partial text queries.

Evaluation Parameters	1. First Query Paragraph	2. One Central Query Paragraph	3. Longest Paragraph Path	4. Longest Path with One Central Node	5. Path with Max Average Para Sim	Full Document Query
Number of Queries	97	96	101	100	101	102
Total retrieved	782	779	1045	1043	1022	1298
Total relevant	1759	1868	1878	1863	1878	1910
Total relevant retrieved	452	548	765	761	737	1081
Relative recall after 5 docs	0.2634	0.2515	0.3260	0.3288	0.3169	0.3898
after 15 docs	0.3527	0.3461	0.4832	0.4850	0.4708	0.6260
Relative precision after 5 docs	0.5528	0.6125	0.6931	0.6990	0.6832	0.8020
after 15 docs	0.3508	0.3549	0.4554	0.4554	0.4442	0.5954
11-point recall-precision average	0.3261	0.3459 (+6.1%)	0.4714 (+44.5%)	0.4714 (+44.5%)	0.4541 (+39.3%)	0.8020 (+90.5%)
Query length as percentage of full query article	5.2%	7.0%	29.1%	30.4%	24.3%	100%
Performance deterioration as percentage of full query performance	-47.5%	-44.3%	-24.1%	-24.1%	-27.1%	0%

Table 1. Performance Evaluation of Selective Intra-document Paragraph Traversal (about 100 encyclopedia searches)

Table 2 shows evaluation data for inter-document paragraph traversal where the queries are constructed by taking related paragraphs from different encyclopedia articles. An initial query article is first used to select a number of other highly-similar encyclopedia articles. Each group of related articles is then used to build query paths using the path generation strategies described in the previous section. The first two columns of Table 2 are identical with the corresponding columns of Table 1, showing the effect of using the starting and central paragraphs of the original query article. Runs 3, 4, and 5 correspond to those of Table 1, but with the path extracted from the complete cluster of documents related to the query. Run 6 is the same as in Table 1, giving the search results for the full original query article.

The results shown in Table 2 are largely similar to those of Table 1. Columns 3 and 5 show the potential adverse effects of including paragraphs from related texts: in each case, the amount of text in the path increases, but the performance fails to increase correspondingly. Column 4, however, representing paths through a central node in the initial query article, shows a marked improvement over Columns 3 and 5, and even a modest improvement over its analog in Table 1. Forcing the query path to remain central to the initial article, but allowing it to include text from other highly-related articles, provides superior evaluation results. Without the central paragraph from the initial article, an inter-article path drifts from the topic of its source. The results of Tables 1 and 2 clearly show that certain constraints imposed on the construction of reading paths will be beneficial for the user. Additional experimentation is needed to make detailed distinctions between particular types of text traversal strategies.

CONCLUSION

Viable procedures are covered in this study for linking related texts in a large collection of unrestricted subject matter. Many questionable text links are discarded by using global/local context restrictions. Furthermore, by focusing attention on small portions of text, the quality of the text links is substantially improved. In particular, long heterogeneous texts are easily broken down into subject-specific text fragments.

The automatic text linking system is usable for the identification of central text portions that could be assembled into summaries of full documents. The linked text fragments can also serve for the construction of selective text traversal paths, both within particular documents and between related documents. A preliminary evaluation shows that the use of related text excerpts for query formulation purposes will retrieve a large proportion of the information relevant to stated user needs.

Evaluation Parameters	1. First Query Paragraph	2. One Central Query Paragraph	3. Longest Paragraph Path	4. Longest Path with One Central Node	5. Path with Max Average Para Sim	Full Document Query
Number of Queries	97	96	101	100	101	102
Total retrieved	782	779	1164	1164	1118	1298
Total relevant	1759	1868	1878	1868	1878	1910
Total relevant retrieved	452	548	679	766	667	1081
Relative recall after 5 docs	0.2634	0.2515	0.3089	0.3246	0.3175	0.3898
after 15 docs	0.3527	0.3461	0.4723	0.5061	0.4639	0.6260
Relative precision after 5 docs	0.5528	0.6125	0.6376	0.6800	0.6396	0.8020
after 15 docs	0.3058	0.3549	0.4099	0.4607	0.4007	0.5954
11-point recall-precision average	0.3261	0.3459 (+6.1%)	0.4407 (+35.1%)	0.4795 (+47.0%)	0.4454 (+36.6%)	0.6214 (+90.5%)
Query length as percentage of full query article	5.2%	7.0%	37.7%	41.2%	25.6%	100%
Performance deterioration as percentage of full query performance	-47.5%	-44.3%	-34.1%	-22.8%	-33.3%	0%

Table 2. Performance Evaluation of Selective Inter-document Paragraph Traversal
(about 100 encyclopedia searches)

REFERENCES

- [Baxe58] P.B. Baxendale, Man-Made Index for Technical Literature --- An Experiment, IBM Journal of Research and Development 2:4, 1958, 354-361.
- [Bern90] M. Bernstein, An Apprentice that Discovers Hypertext Links, in Hypertexts: Concepts, Systems and Applications, A. Rizk, N. Streitz, and J. Andre, editors, Proceedings European Conference on Hypertext, Cambridge University Press, November 1990, 212-223.
- [Bern91] M. Bernstein, J.D. Bolter, M. Joyce, E. Mylonas, Architectures for Volatile Hypertext, Proceedings Hypertext 91, Association for Computing Machinery, New York, December 1991, 243-260.
- [Bolt90] Jay D. Bolter, The Writing Space: The Computer, Hypertext, and the History of Writing, Lawrence Erlbaum, Hillsdale, N.J., 1990.
- [Bota92] R.A. Botafogo, E. Rivlin, and B. Shneiderman, Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics, ACM Transactions on Information Systems, 10:2, April 1992, 142-180.
- [Dela91] P. Delaney and G.P. Landow, eds., Hypermedia and Literary Studies, MIT Press, Cambridge MA, 1991.
- [Earl70] L.L. Earl, Experiments in Automatic Extracting and Indexing, Information Storage and Retrieval 6:4, October 1970, 313-334.

- [Edmu64] H.P. Edmundson, Problems in Automatic Abstracting, Communications of the ACM 7:4, April 1964, 259-263.
- [Funk79] Funk and Wagnalls New Encyclopedia, Funk and Wagnalls, New York, 1979, 29 volumes, 25,000 encyclopedia articles.
- [Guin92] C. Guinan and A.F. Smeaton, Information Retrieval from Hypertext Using Dynamically Planned Guided Tours, Proceedings European Conference on Hypertext, ECHT 92, Association for Computing Machinery, New York, December 1992, 122-130.
- [Luhn58] H.P. Luhn, The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development 2:2, April 1958, 159-165.
- [OCon75] J. O'Connor, Retrieval of Answer Sentences and Answer Figures by Text Searching, Information Processing and Management, 11:5/7, 1975, 155-164.
- [OCon80] J. O'Connor, Answer Passage Retrieval by Text Searching, Journal of the American Society for Information Science, 32:4, July 1980, 227-239.
- [Paic90] C.D. Paice, Constructing Literature Abstracts by Computer: Techniques and Prospects, Information Processing and Management, 26:1, 1990, 171-186.
- [Reim87] U. Reimer and U. Hahn, Text Condensation as Knowledge Base Abstraction, Report MIP-8723, University of Passau, Germany, December 1987.
- [Rush64] J.E. Rush, R. Salvador, and A. Zamora, Automatic Abstracting and Indexing --- Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria, Journal of the ASIS 22:4, July-August 1964, 260-274.
- [Salt71] G. Salton, ed., The Smart Retrieval System --- Experiments in Automatic Document Processing, Prentice Hall, Inc., Englewood Cliffs, N.J., 1971.
- [Salt75] G. Salton, C.S. Yang and A. Wong, A Vector Space Model for Automatic Indexing, Communications of the ACM 18:11, November 1975, 613-620.
- [Salt83] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill Book Co., New York, 1983.
- [Salt88] G. Salton and C. Buckley, Term Weighting Approaches in Automatic Text Retrieval, Information Processing and Management 24:5, 1988, 513-523.
- [Salt89] G. Salton, Automatic Text Processing -- The Transformation Analysis and Retrieval of Information by Computer, Addison Wesley Publishing Company, Reading, MA, 1989.
- [Salt91a] G. Salton and C. Buckley, Global Text Matching for Information Retrieval, Science 253:5023, 30 August 1991, 1012-1015.
- [Salt91b] G. Salton and C. Buckley, Automatic Text Structuring and Retrieval--Experiments in Automatic Encyclopedia Searching, Proc. Fourteenth Int. ACM/SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 1991, 21-30.
- [Salt91c] G. Salton, Developments in Automatic Text Retrieval, Science, 253, 30 August 1991, 974-980.
- [Salt93] G. Salton, J. Allan, and C. Buckley, Approaches to Passage Retrieval in Full Text Information Systems, Proceedings of the 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, June 1993, 49-58.
- [Witt53] L. Wittgenstein, Philosophical Investigations, Basil Blackwell and Mott Ltd., Oxford, England, 1953.