Automatic Text Decomposition Using Text Segments and Text Themes

Gerard Salton*, Amit Singhal, Chris Buckley, and Mandar Mitra

Department of Computer Science, Cornell University
{singhal, chrisb, mitra}@cs.cornell.edu

ABSTRACT
With the widespread use of full-text information retrieval, passage-retrieval techniques are becoming increasingly popular. Larger texts can then be replaced by important text excerpts, thereby simplifying the retrieval task and improving retrieval effectiveness. Passage-level evidence about the use of words in local contexts is also useful for resolving language ambiguities and improving retrieval output.

Two main text decomposition strategies are introduced in this study, including a chronological decomposition into *text segments*, and semantic decomposition into *text themes*. The interaction between text segments and text themes is then used to characterize text structure, and to formulate specifications for information retrieval, text traversal, and text summarization.

KEYWORDS: Text structuring, text decomposition, segments, themes, information retrieval, passage retrieval, text summarization.

TEXT PASSAGES AND TEXT RELATIONSHIP MAPS

With the advent of full-text document processing, the interest in manipulating text passages rather than only full-text items has continued to grow. Retrieving large texts in answer to user queries tends to be inefficient because the user is then forced to cope with large masses of text, and ineffective because relevant text passages often provide better answers than complete document texts. In addition, passage-level evidence accounting for word usage in local text environments is often helpful in improving retrieval effectiveness, because the meaning of ambiguous terms becomes clear when the local context is properly specified. [1-6]

Since full texts are necessarily composed of individual text passages, a study of text passages is also important for determining overall text structure. A structural decomposition of texts into passages may then reveal information about the type of text under consideration, and knowledge of text type and text structure in turn affects many text handling operations, including retrieval, text reading and traversal, and text summarization.

The structure of individual texts, or sets of related texts, can be studied by using a text relationship map that exhibits the results of similarity measurements between pairs of texts, or text excerpts. Typically, each text, or text excerpt is represented by a vector of weighted terms of the form $D_i = (d_{i_1},\ d_{i_2}, ..., ..., d_{i_t})$ where d_{i_k} represents an importance weight for term T_k attached to document D_i. The terms attached to documents for content representation purposes may be words or phrases derived from the document texts by an automatic indexing procedure, and the term weights are computed by taking into account the occurrence characteristics of the terms in the individual documents and the document collection as a whole. [7]

Assuming that every text, or text excerpt is represented in vector form as a set of weighted terms, it is possible to compute pairwise similarity coefficients showing the similarity between pairs of texts based on coincidences in the term assignments to the respective items. Typically, the vector similarity might be computed as the inner product between corresponding vector elements, that is, $sim\ (D_i, D_j) = \sum_{k=1}^{t} d_{i_k}\ d_{j_k}$, and the similarity function might be normalized to lie between 0 for disjoint vectors and 1 for completely identical vectors.

Figure 1 shows a typical text relationship map for six texts included in the Funk and Wagnalls encyclopedia dealing with the general topic of Nuclear Energy. The documents appear as nodes (vertices) in the graph of Figure 1, and a link (branch) appears between two nodes when the similarity between two texts is sufficiently large. The similarity threshold used to build the map of Figure 1 is 0.01, that is, all branches representing a text similarity above 0.01 are shown on the map. Figure 1 shows that the similarity measure between documents 17012 and 17016 (Nuclear Energy and Nuclear Weapons) is a high 0.57, whereas no significant similarity exists between 8907 (Nuclear Fission) and 22387 (Thermonuclear Fusion).

*This study was supported in part by the National Science Foundation under grant IRI 9300124.

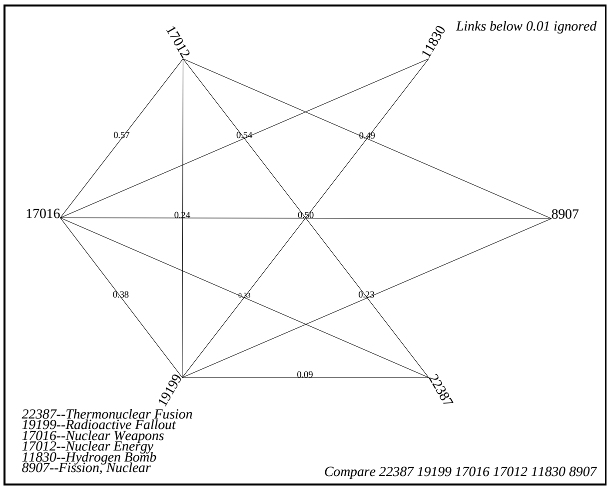

Figure 1: Text Relationship Map: Encyclopedia articles related to Nuclear Energy.

Graph structures have been used to represent relationships between text components that may be valid inside particular documents rather than only relationships between different documents. [8-11] In the present context, the text relationship maps can be refined by having the nodes represent shorter text excerpts, such as text sentences, or paragraphs, thereby forming a representation of the sentence or paragraph similarities for particular texts.

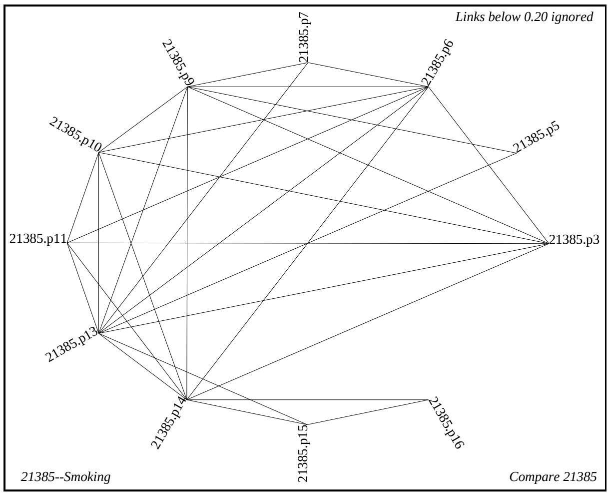

Figure 2: Well-Connected Map (Smoking – 16 paragraphs, 28 links above 0.20). Smoking: History, Health Effects, Cessation.

Figure 2 is a paragraph relationship map for encyclopedia article 21385 entitled "Smoking". The paragraph similarity threshold used to generate the map of Figure 2 is 0.20, that is, paragraph similarities smaller than 0.20 do not appear on the map. Various elements of text structure are immediately desirable from a paragraph-relationship map such as that of Figure 2. For example, the importance of a paragraph might be related to the number of incident branches of the corresponding node on the map. A central node might then be characterized as one with a large number of associated paragraphs. For article 21385, the most central paragraphs are then paragraphs 3, 6, 9, 13, and 14.

The paragraph relationship map also provides information about the homogeneity of the text under consideration. When the map is nicely convex with many cross-connections between paragraphs, and direct links between adjacent paragraphs, one expects a unified, homogeneous treatment of the topic. This is the case notably for the article on smoking and the relationship map of Figure 2. The corresponding encyclopedia article deals for the most part with the health effects of smoking, and the difficulties that arise when attempting to quit smoking.

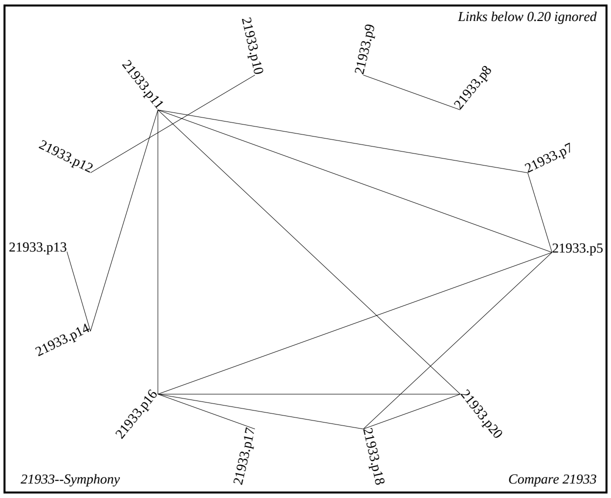

Figure 3: Poorly-Connected Map (Symphony – 20 paragraphs, 15 links above 0.20). Symphony: Italy, Germany and Austria, Haydn and Mozart, Beethoven, 19th Century, 20th Century.

Consider in contrast the paragraph relationship map of Figure 3 for the encyclopedia article 21933 entitled "Symphony". The similarity threshold of Figure 3 is identical with that of Figure 2 (0.20). Nevertheless, the map of Figure 3 is much sparser than that of Figure 2. There are disconnected components such as 21933.p8 and p9, and 21933.p10 and p12, and the lack of convexity indicates that the subject treatment is much more heterogeneous than in the earlier example. In fact, the treatment in document 21933 is chronological and largely distinct for different time periods and different geographical areas. This accounts for the lack of connectivity in the sample of Figure 3.

The examples of Figures 2 and 3 indicate that quite different relationship maps can be generated for appar-

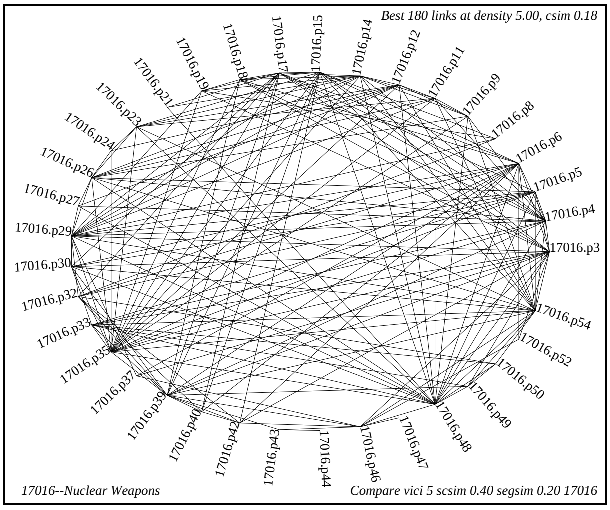

Figure 4: Text Relationship Map (17016 Nuclear Weapons).

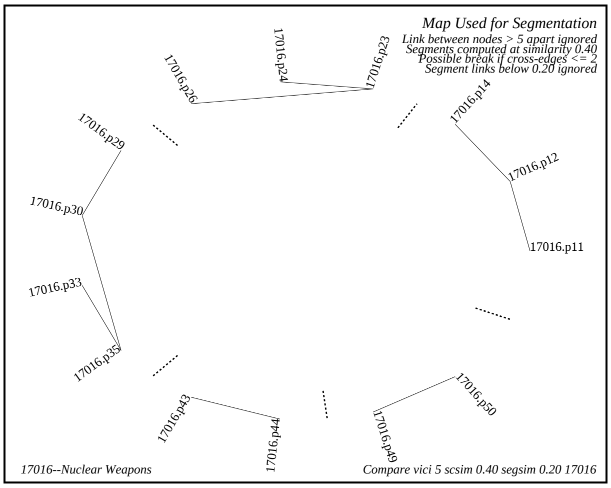

Figure 5: Formal Text Segmentation (sparse map: 5 segments).

ently similar texts, such as articles included in the same encyclopedia, apparently reflecting quite different topic treatments. This suggests that it may be useful to carry out a detailed study of text structure in the hope of generating a classification of text types leading to advanced text handling methods.

AUTOMATIC TEXT DECOMPOSITION

In studying the structure of written texts, we are interested in identifying text pieces exhibiting internal consistency that can be distinguished from the remainder of the surrounding text. The most immediate possibility consists in using the well-defined physical elements that are apparent for all types of running text, such as tables, figures, headings, titles, bibliographic references, sentences, paragraphs, sections, and so on.

In principle, a decomposition into recognizable physical units is easy to carry out. However, the resulting text units are often not particularly interesting, nor are they always distinguishable in terms of functionality from adjacent text pieces.

This suggests that a decomposition should involve more meaningful text units than the usual physical text elements:

1. The first possibility consists in generating functionally homogeneous text units, known as *text segments*. A text segment is a contiguous piece of text that is linked internally, but largely disconnected from the adjacent text. Typically, a segment might consist of introductory material, or cover the exposition and development of the text, or contain conclusions and results.

2. Another dual text decomposition could create semantically homogeneous text pieces where all components treat a common subject area. Semantically homogeneous text pieces, known as *text themes*, are represented by mutually similar (linked) text pieces that are not necessarily adjacent in the text. [12-14]

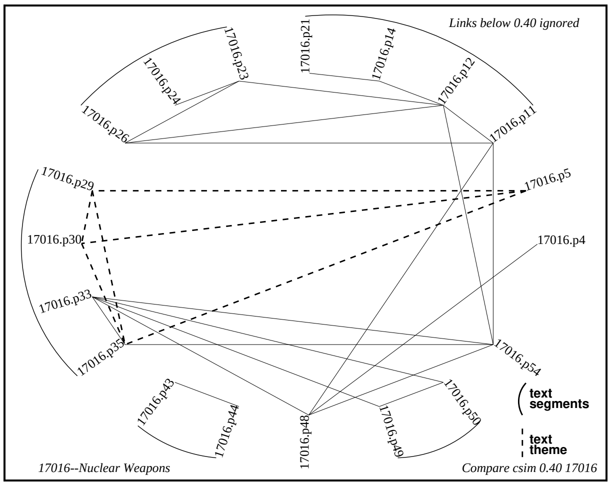

Figure 6: Text Segmentation and Text Theme Identification (encyclopedia article 17016 "Nuclear Weapons").

To obtain the text segments, it is necessary to find gaps in the connection pattern between adjacent paragraphs in the text relationship map. However, as the example of Figure 4 shows, it is often difficult to find obvious gaps in the connecting pattern. This suggests that the relationship map be simplified by considering only local connections between paragraphs located in close proximity to each other. When the long-distance links – those spanning more than five adjacent paragraphs – are eliminated from the map of Figure 4, the reduced map of Figure 5 is obtained showing an obvious break down into five disconnected segments: 170416.p11 to p14, 17016.p23 to p26, 17016.p29 to p35, 17016.p43 to

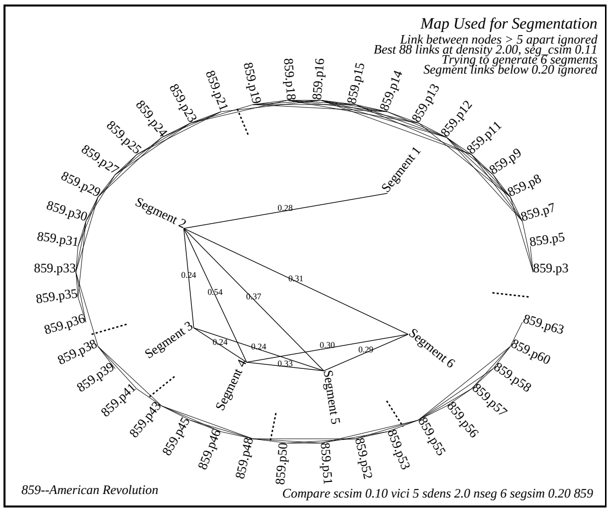

Figure 7: Segment Similarities Reveal Text Structure (semi-detached S1, closely related S2 through S6).

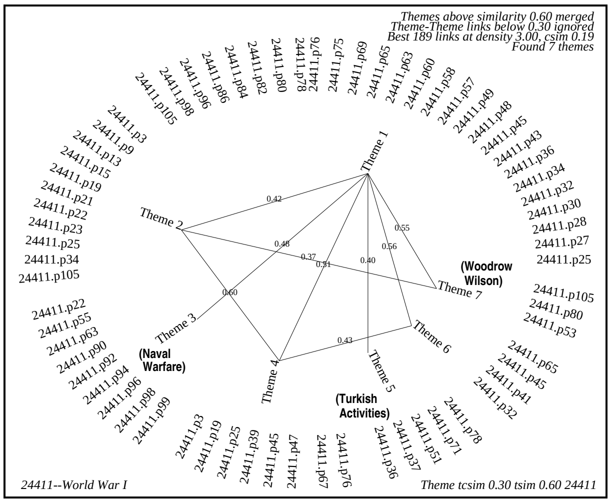

Figure 8: Theme Similarities Reveal Theme Centrality and Theme Specialization.

44, and 17016.p49 to p50.

When disconnected text segments are located as in Figure 5, one may expect that a minimal functional unity is maintained within each segment. However, the subject matter may not be treated in a linear fashion in many written texts. In other words, one would not necessarily expect that segments are always coextensive with text themes. [15] To obtain the text themes, attention must be paid to mutually linked text pieces that are not necessarily adjacent in the text. One possibility consists in locating all triangles in the full text relationship map, where a triangle is a set of three mutually related paragraphs. Each triangle can then be represented by a centroid vector defined as the average for the three vectors in the triangle; and triangles can be merged when the similarity between the corresponding centroid pair exceeds a given threshold. [15] The merging process is continued for higher-order structures until no further merging is possible. A typical theme is shown by dashed lines in Figure 6, consisting of paragraphs 17016.p5, p29, p30, and p35.

Given the set of segments and themes derived from a particular paragraph-relationship map, it is possible to compute segment-segment, theme-theme, and theme-segment similarities. As before, a threshold value must be chosen, and all segment and theme similarity values exceeding the threshold can then be displayed graphically. The segment-segment relationships provide information about the overall structure of the document under consideration. The segment similarities for encyclopedia article 859 entitled "American Revolution" are shown in the center of the graph of Figure 7. As the figure shows, the first segment dealing with the causes of the American Revolution is semi-detached, whereas substantial interconnections exist between segments 2 and 6 covering various military engagements in the Revolutionary War.

The theme similarities provide information about theme centrality and theme specialization. An example is shown in Figure 8 for article 24411 "World War 1". Since the themes are not necessarily represented by adjacent text pieces, the actual theme composition is listed along the periphery of the theme similarity map. Figure 8 shows that theme 1 is a central theme that covers large parts of the document and is related to all other text themes. On the other hand, a number of specialized themes are also included in the map of Figure 8, including theme 3 (Naval Warfare), theme 5 (Turkish activities in World War 1), and theme 7 (Woodrow Wilson).

SIMPLE AND COMPLEX TEXT STRUCTURES

The theme and segment text decompositions can be used to distinguish simple text structures from more complex ones. In particular, in many cases, the segment and theme decompositions are largely congruent: in that case each theme covers text excerpts occurring adjacently in the text, hence more or less corresponding to text segments. This is the case notably for homogeneous, single topic articles, and for multi-topic articles with a sequential topic treatment where the themes are then more or less isolated from each other. Consider the example of Figure 9 for encyclopedia article 78 entitled "Abortion". Two segments are apparent in Figure 9(a) dealing with the facts of abortion, and the legal implications of abortion, respectively. A single theme is obtained in the output of 9(b). The theme-segment similarities presented in the center of Figure 9(b) show that theme 1 (78.T1) is related to both segments 1 and 2 (78.S1 and 78.S2). Hence the single theme is coextensive with the two segments and therefore with the complete article.

A more complicated situation is presented in Figure 10. Here three disjoint segments are in evidence shown in

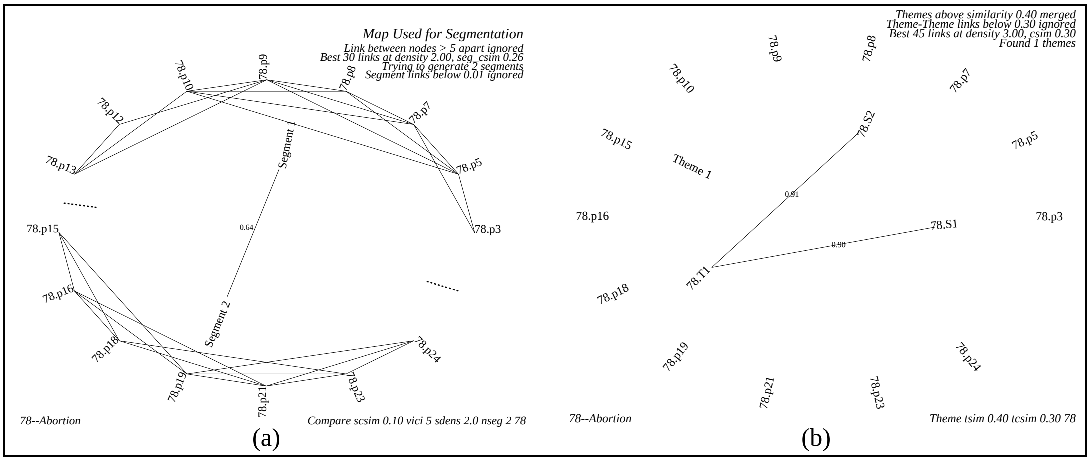

Figure 9: Simple Text Structure – Single Theme (Encyclopedia Article 78: Abortion). a) Segmentation, b) Theme and Theme-Segment Relations.

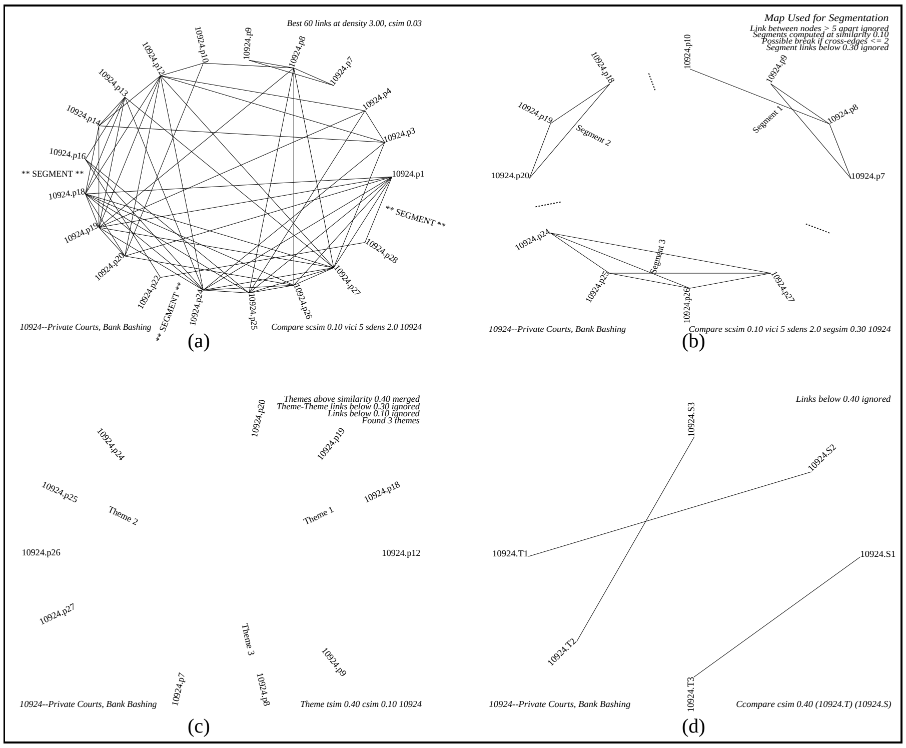

Figure 10: Simple Text Structure – Multiple Stories (Wall Street Journal 10924). a) Segments (all links), b) Segments (local links), c) Themes, and d) Theme-Segment Relations.

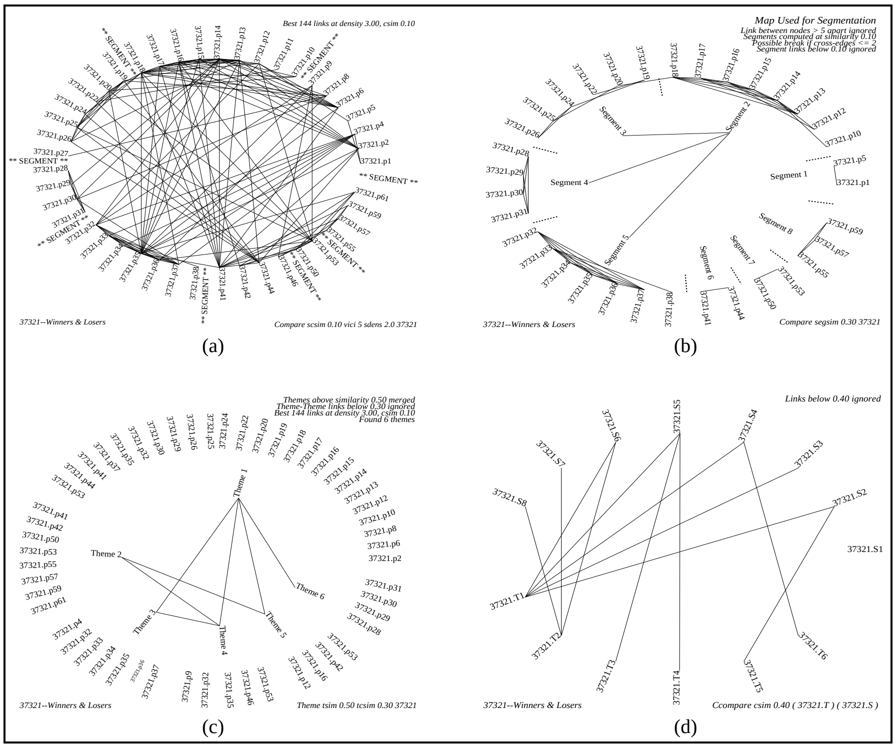

Figure 11: Simple Text Structure – Multiple Themes (Wall Street Journal 37321). a) Segments (all links), b) Segments (local links), c) Themes, and d) Theme-Segment Relations.

Figure 10(b). In addition three themes are obtained as shown in Figure 10(c). The theme-segment similarities of Figure 10(d) indicate that a one-to-one relationship exists between themes and segments. This is reflective of the fact that the article in question (Wall Street Journal article 10924) covers three disjoint topics that are largely unrelated to each other.

The output of Figure 11 provides a last example of simple text structure. Eight segments are shown in Figure 11(b), and 6 themes (see Figure 11(c)). The theme-segment similarities of Figure 11(d) demonstrate that each theme covers one or more adjacent text segments. Thus, theme 1 is related to segments 2 to 6; theme 2 to segments 6 to 8, and so on. The theme decomposition is therefore closely related to the text segmentation.

In many cases, segment and theme decompositions are not congruent. In that case, a particular topic may be raised, and then dropped, only to be restarted at some later time in the normal text order. A theme may then cover text excerpts from non-adjacent text segments. Two main examples arise in practice. When large, central themes are present, the theme often covers the main aspects of the text but auxiliary or more specialized text areas (segments) may be skipped as being unrelated to the principal theme. Figure 12 provides an example where the main theme (theme 1) is related to the large segments (segments 1 and 3), but no similarity exists with the small, auxiliary segment 2. The same information is contained in Figure 12(d) where a gap is visible in the connections between theme 1 (1022.T1) and segments 1 and 3 respectively (1022.S1 and 1022.S3).

Another gapping problem can arise for short themes. In many cases, a theme may be defined by only 3 or 4 paragraphs. In that case, the theme-segment relations are often confined to exactly those segments containing one of the theme paragraphs . Figure 13 provides an example where a large, central theme (theme 1) is supplemented by two small themes containing three paragraphs each. The theme-segment similarity map of Figure 13(d) shows that theme 1 is related to all four segments. However, theme 2 consisting of paragraphs p8, p16, and p32 is related only to segments 1, 2, and 4, while theme 3 (p7, p24, and p36) is related to segments 1, 3, and 4. In each case, the theme-segment similarities are due to the overlap between one theme paragraph and one segment paragraph.

RETRIEVAL STRATEGIES

The text classification system outlined in the previous section can be used as a basis for the generation of text retrieval and text traversal operations. Consider first the standard information retrieval environment. For texts with a simple topic outline where themes and segments are reasonably congruent, the standard passage-retrieval systems that are designed to retrieve the best adjacent text pieces should provide optimal retrieval output. A "mixed" retrieval strategy has been imple-

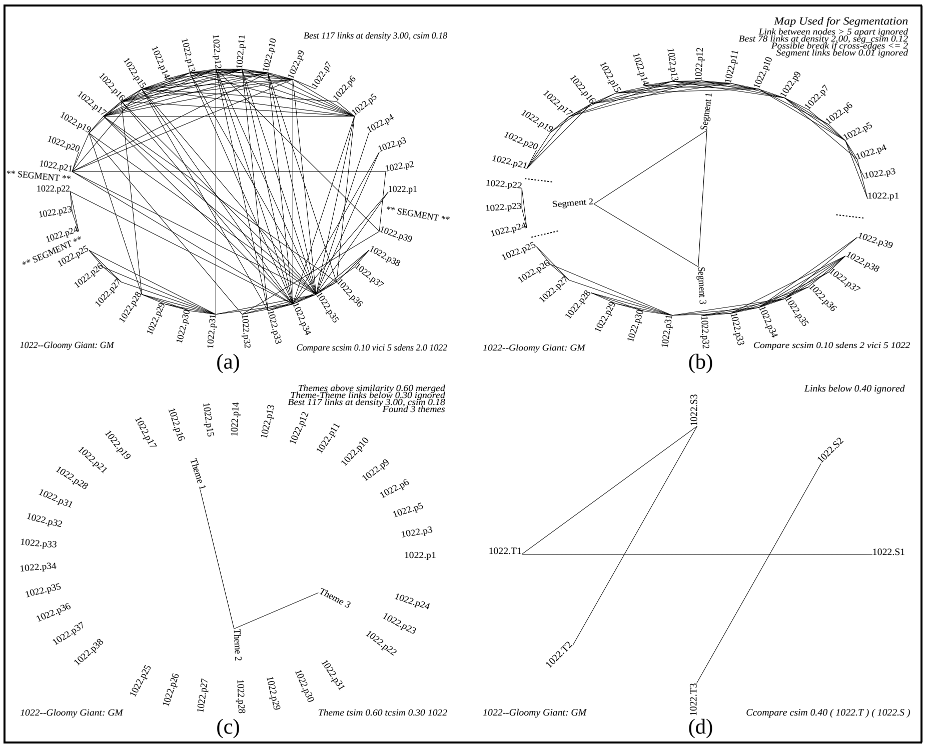

Figure 12: Complex Text Structure – Large Theme (Wall Street Journal 1022). a) Segments (all links), b) Segments (local links), c) Themes, and d) Theme-Segment Relations.

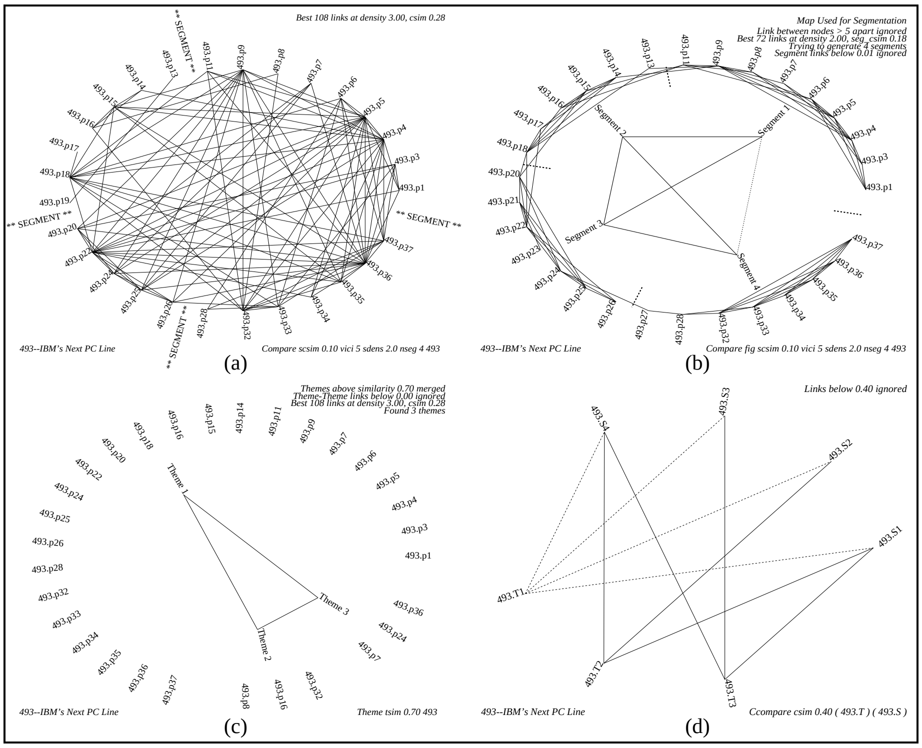

Figure 13: Complex Text Structure – Small Themes (Wall Street Journal 493). a) Segments (all links), b) Segments (local links), c) Themes, and d) Theme-Segment Relations.

Document: 859 American Revolution **Query**: Revolutionary War, Stamp Act, Navigation Act, Lexington, Massachusetts	
Baseline (full text 859)	0.1337
Paragraph Retrieval	
859.p.7	0.3338
859.p.9	0.3748
859.p.64	0.2812
Section Retrieval: 859.c5 (p.7,8,9)	**0.3953**
Segment Retrieval: 859.S1 (p.3,7,8,9)	0.3484
Theme Retrieval: 859.T2 (p.3,7,8,9,12–14,16,18,19,21,48,64)	0.2685

Table 1: Query Processing for Simple Document Structure: Encyclopedia Article 859 American Revolution.

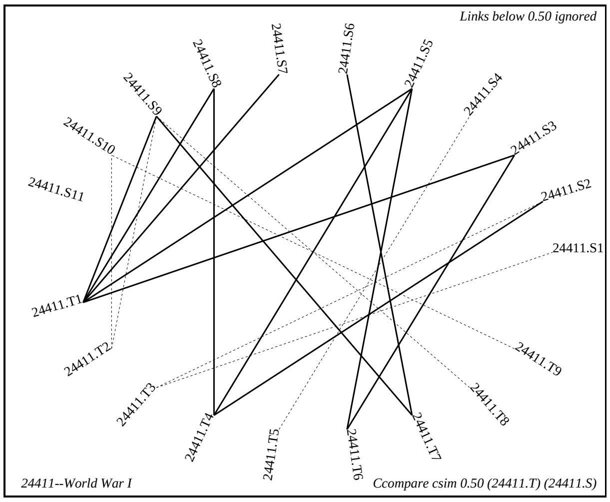

Figure 14: Segment-Theme Relations (24411 World War 1).

mented with the Smart system for many years, which retrieves full text, or various subdivisions such as paragraphs or sections, depending on which text passage provides the highest similarity with the query. When the themes cover adjacent text pieces, such a mixed retrieval strategy provides a useful approach to text retrieval. [16]

Consider as an example a query dealing with the revolutionary war processed against encyclopedia article 859 (American Revolution). Table 1 contains the size of the query similarities for various passages of document 859. The least attractive possibility consists in retrieving the full text, and the best is provided by the retrieval of section 5 (859.c5) which includes paragraphs 7, 8, and 9. Table 1 shows that the query similarity decreases when best segment (859.S1) or best theme (859.T2) are considered.

Document: 24411 World War 1 **Query**: Turkish Activities in World War 1. Turkish Role Against British in Middle East	
Baseline (full text 24411)	0.1718
Paragraph Retrieval: 24411.p36	0.3627
Section Retrieval: 24411.c16 (p.36,37)	0.3751
Segment Retrieval: 24411.S4 (p.36,37,39)	0.3240
Theme Retrieval: 24411.T5 (p.36,37,51,71,78)	**0.4287**

24411.p36: Turkey entered the war on October 29, 1914, when Turkish warships cooperated with German warships in a naval bombardment of Russian Black Sea ports; Russia formally declared war on Turkey on November 2, and Great Britain and France followed suit on November 5. . . .

24411.p37: In the Mesopotamian Valley, meanwhile, British forces from India defeated the Turks in several battles during 1914-15, particularly that of Kut-al-Imara; but in the Battle of Ctesiphon, . . .

24411.p51: Considerable military activity took place in 1916 in three parts of Turkey: Mesopotamia, Arabia, and Palestine. In Mesopotamia the besieged town of Kut-al-Imara fell to the Turks on April 29, 1916. In December of the year strong British forces began . . .

24411.p71: In Palestine during 1917 the British made two unsuccessful attempts (March and April) to take the city of Gaza. Under a new commander, General (later Field Marshal) Sir Edmund Allenby, the British broke through the Turkish lines at Beersheba . . .

24411.p78: During 1918 the Allies also brought the campaigning in Palestine to a successful conclusion. In September the British forces broke through the Turkish lines at Megiddo and routed the Turkish army and the German corps that was assisting it; . . .

Table 2: Query Processing for Complex Document Structure: Encyclopedia Article 24411 World War 1.

The situation is different when the topic arrangement is complex. Consider, for example, the theme-segment relations for encyclopedia article 24411 (World War 1). Figure 14 shows that segmentation gaps exist for several themes, including theme 1 (related to segments 3, 5, 7, 8, and 9 but not to segments 4, and 6), theme 4 (related to segments 2, 5, and 8, but not 3, 4, 6, and 7), theme 6 (related to segments 3, and 5, but not 4), and theme 7 (related to segments 6, and 9, but not 7, and 8).

When a query is processed that covers the subject matter of one of the disconnected themes, a standard passage retrieval method that locates the best stretch of adjacent text, is not likely to be optimal. Instead, the best output should be provided by a theme, or a set of disconnected segments. Consider as an example, the output of Tables 2 and 3 in which queries corresponding to themes 5 and 7, respectively, are processed against the text of document 24411 (World War 1). In both cases, the corresponding theme output is better than any conventional passage of adjacent text.

Document: 24411 World War 1 **Query**: President Woodrow Wilson, Allies American Troops, War Aims, Peace Negotiations	
Baseline (full text 24411)	0.1908
Paragraph Retrieval	
24411.p.53	0.4834
24411.p.105	0.3757
24411.p.80	0.2628
Section Retrieval	
24411.c23 (p.53)	0.4834
24411.c45 (p.105,106,107)	0.3395
Segment Retrieval: 24411.S6 (p.53,55)	0.4090
Theme Retrieval: 24411.T7 (p.53,80,105)	**0.5252**

Table 3: Similarity Computation Between Query and Text Excerpts.

– Path Length 3 –
– Path Summary –
78.p3 78.p5 78.p9 : S1

—– Paragraph 78.p3 —–
Abortion, termination of pregnancy before the fetus is capable of independent life. When the expulsion from the womb occurs after the fetus becomes viable (capable of independent life), usually at the end of six months of pregnancy, it is technically a premature birth.

—– Paragraph 78.p5 —–
Abortion may be spontaneous or induced. Expelled fetuses weighing less than 0.45 kg (16 oz) or of less than 20 weeks' gestation are usually considered abortions.

—– Paragraph 78.p9 —–
The most common symptom of threatened abortion is vaginal bleeding, with or without intermittent pain. About one-fourth of all pregnant women bleed at some time during early pregnancy, however, and up to 75 percent of these women carry the fetus for the full term. Treatment for threatened abortion usually consists of bed rest. Almost continuous bed rest throughout pregnancy is required in some cases of repeated abortion, and vitamin and hormone therapy also may be given. In addition, surgical correction of uterine abnormalities may be indicated in certain cases of repeated abortion.

Table 4: Global Text Traversal (20% Global Path): Encyclopedia Article 78, Abortion: p3–5–9 segment-1. Single Theme Article.

The conclusion is that for simply structured texts, with a topic arrangement corresponding to the linear text structure, the normal passage retrieval techniques are useful. When the topic arrangement is more complex, the best text passage is likely to be disconnected, and a text theme computation must precede the passage retrieval operation.

TEXT TRAVERSAL AND TEXT SUMMARIZATION

In addition to conventional information retrieval, it may also be useful to provide various text extracts on demand. Text traversal implies a type of speed reading where an area of interest is specified, and the best text passages representing that area are chosen in response. Text summarization is a related problem where a set of text passages is used that collectively represents the full text. [16]

In dealing with text traversal, it is necessary to distinguish the so-called global paths, that operate on a complete text, from paths restricted to some substructure, such as paths within themes or within segments. In either case, many different traversal orders can be considered, the most important being:

1. Central node paths of a particular length that include the n nodes with the largest number of links to other nodes in a relationship map. The desired path length may be stated as a percentage of the total number of nodes appearing in a map. Normally, central node paths are more or less comprehensive depending on path length, but they may not be coherent because there is no guarantee that adjacent path nodes are in fact connected and hence similar in subject matter.
2. Depth-first paths are paths that start with an important node – for example, with a highly-connected central node – and then use the next most similar node at each point. In a depth-first path, coherence is normally assured because all pairs of adjacent nodes are properly linked. However, depth-first paths are generally not comprehensive because the subject matter covered by the path is largely controlled by that of the initial path node.

Given the earlier text classification, various kinds of traversal strategies can be formulated: for homogeneous, single-theme articles, a global traversal strategy is best used in increasing text order, specified either as a central or a depth-first path. For multi-topic articles with a sequential topic structure, the relevant adjacent text segments can be isolated, and a traversal path can then be used that operates within the relevant text segments. Finally for complex texts where the themes cover non-adjacent text pieces, a traversal path is defined within text themes. Traversal within text themes is especially interesting for large, comprehensive themes. A twenty percent traversal path is generally sufficiently long to insure that the more important aspects of the subject are actually covered.

Consider, as an example, the 20 percent global text traversal path for encyclopedia article 78 (Abortion) shown in Table 4. The segmentation map of Figure 9(a) contains 15 nodes. Hence, the 20 percent paths will include 3 paragraphs. The extract shows that the main aspects of the topic are indeed covered by the traversal path; however, all the paragraphs included in the global extract appear in segment 1 where the facts of abortion are discussed. An alternative traversal path, also consisting of three paragraphs, is presented in Table 5. Here an additional requirement is added that some paragraph must be chosen from each segment. This leads

– Total Path Length (over all segments) 3 –
– Path Summary –
78.p5 78.p9 : S1
78.p21 : S2

—– 15% Bushy path in segment 1 (degree >= 14) —–
—– Paragraph 78.p5 —–
Abortion may be spontaneous or induced. Expelled fetuses weighing less than 0.45 kg (16 oz) or of less than 20 weeks' gestation are usually considered abortions.

—– Paragraph 78.p9 —–
The most common symptom of threatened abortion is vaginal bleeding, with or without intermittent pain. About one-fourth of all pregnant women bleed at some time during early pregnancy, however, and up to 75 percent of these women carry the fetus for the full term. Treatment for threatened abortion usually consists of bed rest. Almost continuous bed rest throughout pregnancy is required in some cases of repeated abortion, and vitamin and hormone therapy also may be given. In addition, surgical correction of uterine abnormalities may be indicated in certain cases of repeated abortion.

—– 15% Bushy path in segment 2 (degree >= 12) —–
—– Paragraph 78.p21 —–
Opponents of the 1973 Supreme Court ruling, arguing that a fetus is entitled as a "person" to constitutional protection, attacked the decision on a variety of fronts. State legislative bodies were lobbied for statutes narrowing the implications of the decision and circumscribing in several ways the mother's ability to obtain an abortion. A nationwide campaign was instituted to amend the Constitution to prohibit or severely restrict abortion. "Right-to-life" groups also engaged in grassroots political activity designed to defeat abortion proponents and elect abortion opponents. Abortion became, rather than simply a legal and constitutional issue, one of the major political and social controversies of the late 1970s and the '80s. Many state legislatures responded with a succession of statutes imposing additional procedural burdens on women who sought abortions; federal court decisions holding these new statutes unconstitutional usually followed each legislative initiative.

Table 5: Equivalent Traversal Within Segments (15% Path Within Every Segment): Encyclopedia Article 78, Abortion: p5–9 segment-1, p21 segment 2. Single Theme Article.

to the addition of paragraph 21 from segment 2, dealing with certain legal aspects of the abortion problem. Such a requirement broadens the coverage of the traversal path, but may also introduce some lack of coherence in the final text traversal. In general, as the focus of discourse changes from one segment to the next, a paragraph traversal path may never be totally smooth.

Possible solutions consist in actually supplying some transition material between different segments taken from the available text material to provide a more unified treatment for the text content. A summary of traversal path properties appears in Table 6. As the Table indicates, within each segment, the depth-first path provides a maximum of coherence because pairs of adjacent paragraphs exhibit the required minimal similarity. Global central paths, on the other hand, tend to be more comprehensive than depth-first paths. This suggests that within each segment, the traversal order be tailored to depth-first or global central traversal orders. Since the emphasis in the discourse may change between adjacent segments, greater coherence will be obtained by supplying transition materials to introduce new material covered by the added segments. This leads to the following text traversal prescription:

	Importance of Initial Paragraph	Coherence Comprehensiveness
Global Central Path	Usually starts with important early paragraph	Not coherent because adjacent paragraphs may be unrelated
Central Path Within Segments	May lose important first paragraph because of need to include material from other segments	Not coherent but more comprehensive than global central path
Depth-First Path	Starts with important first paragraph	Not comprehensive but more coherent than central paths, may be specialized to important subtopic

Table 6: Important Properties for Different Traversal Paths

1. maintain important first paragraphs from global central path;
2. add other paragraphs in depth-first or global central orders within each text segment;
3. supply more comprehensiveness by adding materials from additional segments (central paths within segments);
4. supply more coherence by using transition materials introducing the topics covered by the various segments.

Good transition paragraphs are normally those exhibiting high similarities with the initial significant paragraphs in a segment. For the example of article 78 (Abortion), the first significant paragraph in segment 2 is 78.p21 (see Table 5). A good transition paragraph is therefore an early paragraph of segment 2 that relates highly with paragraph 21.

For the example of document 78 (Abortion), it turns out that paragraph 18 has much higher similarities with paragraph 21 than any other early paragraphs in segment 2 of document 78 (similarity(78.p18, 78,p21) = 0.4880). This suggests that paragraph 78.p18 be added to the standard traversal order ahead of paragraph 78.p21.

Table 7 shows a final text traversal order for the example of article 78 (Abortion), starting with paragraph 3 (from the global central path), and continuing with paragraphs 7 and 9 (depth-first path in segment 1). This

—– Paragraph 78.p3 —–
Abortion, termination of pregnancy before the fetus is capable of independent life. When the expulsion from the womb occurs after the fetus becomes viable (capable of independent life), usually at the end of six months of pregnancy, it is technically a premature birth.

—– Paragraph 78.p7 —–
It is estimated that some 25 percent of all human pregnancies terminate spontaneously in abortion, with three out of four abortions occurring during the first three months of pregnancy. Some women apparently have a tendency to abort, and recurrent abortion decreases the probability of subsequent successful childbirth.

—– Paragraph 78.p9 —–
The most common symptom of threatened abortion is vaginal bleeding, with or without intermittent pain. About one-fourth of all pregnant women bleed at some time during early pregnancy, however, and up to 75 percent of these women carry the fetus for the full term. Treatment for threatened abortion usually consists of bed rest. Almost continuous bed rest throughout pregnancy is required in some cases of repeated abortion, and vitamin and hormone therapy also may be given. In addition, surgical correction of uterine abnormalities may be indicated in certain cases of repeated abortion.

a) Paragraph 78.p3 from global central path. Paragraphs 78.p7, p9 from depth–first path in segment 1.

—– Paragraph 78.p18 —–
In the U.S., legislation followed the world trend. Fourteen states adopted the moderately restrictive type of abortion law between 1967 and 1972. Alaska, Hawaii, New York, and Washington legislated abortion on request with few restrictions. In 1973, the United States Supreme Court declared unconstitutional all state statutes but the least restrictive type. Noting that induced early abortions had become safer than childbirth and holding that the word person in the United States Constitution "does not include the unborn," the Court defined, within each of the three stages of pregnancy, the reciprocal limits of state power and individual freedom:

b) Transition material to introduce section 2.

—– Paragraph 78.p21 —–
Opponents of the 1973 Supreme Court ruling, arguing that a fetus is entitled as a "person" to constitutional protection, attacked the decision on a variety of fronts. State legislative bodies were lobbied for statutes narrowing the implications of the decision and circumscribing in several ways the mother's ability to obtain an abortion. A nationwide campaign was instituted to amend the Constitution to prohibit or severely restrict abortion. "Right-to-life" groups also engaged in grass-roots political activity designed to defeat abortion proponents and elect abortion opponents. Abortion became, rather than simply a legal and constitutional issue, one of the major political and social controversies of the late 1970s and the '80s. Many state legislatures responded with a succession of statutes imposing additional procedural burdens on women who sought abortions; federal court decisions holding these new statutes unconstitutional usually followed each legislative initiative.

c) Paragraph 78.p21 from central path in segment 2.

Table 7: Final Modified Traversal Path for Article 78 (Abortion).

is followed by the transition material from paragraph 18 introducing the 1973 “Roe vs. Wade” decision by the U.S. Supreme Court, and finally ending with paragraph 21 which is the representative from segment 2 in the path within segments of Table 5.

In standard selective text traversal, texts can be traversed entirely, or alternatively, it is possible to concentrate on particular segments that may cover relevant subject matter. The text summarization question can be treated by methods similar to those used for text traversal, except that more emphasis is placed on comprehensiveness of the resulting extract. In general, a 20 percent global path should be adequate for most purposes; alternatively, an equivalent path within segments, with extracts being concatenated in the normally text (or segment) order, should provide a comprehensive picture. Finally, for large central themes, a global path within such a theme may provide useful output.

Summaries, consisting of selected paragraph excerpts, plus appropriate transition materials between segments, should give high performance for the vast bulk of simply structured documents. However, even when the text structure is complex, the experimental evidence suggests that useful, readable output is obtainable. Consider, for example, the encyclopedia article 24411 (World War 1). Here, the various topics are intermeshed, covering variously the Western Front (Belgium, France), the Eastern Front (Russia), and the Southern Front (Serbia, Turkey). In addition, auxiliary topics are covered by specialized segments.

The sample traversal order for segment 1 (beginning of World War 1) is shown in Table 8. The significant paragraphs consist of paragraphs 22, 32, and 36, preceded by the transition paragraphs 3 and 21. As the example indicates, this excerpt provides a nearly perfect summary of the situation at the beginning of the War.

Segment 3 is a specialized segment, covering the activities of U.S. President Woodrow Wilson prior to the entry of the United States in the war. The significant paragraph in this case, is paragraph 55 preceded by the transition material of paragraph 53. The complete summary is once again well-constructed as shown in the sample of Table 9.

In conclusion, central node paths within segments appear to offer reasonable solutions to the text summarization problem. Greater coherence can be supplied by adding relevant transition materials at the beginning of the various segments. Ultimately, completely smooth and comprehensive summaries will be difficult to build by pure text extracting methods. However, from a practical viewpoint, the text extraction system appears to provide rapid text reading, and easily comprehensible text extracts.

24411: World War I (107 Paragraphs), Segment 1 **(24411.p3–24411.p39)**

Global central path in segment 1: p22, p32, p36

Transition material between the initial paragraph of Segment 1 (24411.p3) and the initial paragraph of central path in Segment 1 (24411.p22): 24411.p21. Sim(24411.p3, 24411.p21) = 0.2742, Sim(24411.p21, 24411.p22) = 0.2746

—– Paragraph 24411.p3 —–
World War I, military conflict, from 1914 to 1918, that began as a local European war between Austria-Hungary and Serbia on July 28, 1914; was transformed into a general European struggle by declaration of war against Russia on August 1, 1914; and eventually became a global war involving 32 nations. Twenty-eight of these nations, known as the Allies and the Associated Powers, and including Great Britain, France, Russia, Italy, and the United States, opposed the coalition known as the Central Powers, consisting of Germany, Austria-Hungary, Turkey, and Bulgaria. The immediate cause of the war between Austria-Hungary and Serbia was the assassination on June 28, 1914, at Sarajevo in Bosnia (then part of the Austro-Hungarian Empire; now in Bosnia and Hercegovina), of Archduke Francis Ferdinand, heir-presumptive to the Austrian throne, by Gavrilo Princip (1893-1918), a Serb nationalist. The fundamental causes of the conflict, however, were rooted deeply in the European history of the previous century, particularly in the political and economic policies that prevailed on the Continent after 1871, the year that marked the emergence of Germany as a great world power.

—– Paragraph 24411.p21 —–
On July 28 Austria declared war against Serbia, either because it felt Russia would not actually fight for Serbia, or because it was prepared to risk a general European conflict in order to put an end to the Greater Serbia movement. Russia responded by partially mobilizing against Austria. Germany warned Russia that continued mobilization would entail war with Germany, and it made Austria agree to discuss with Russia possible modification of the ultimatum to Serbia. Germany insisted, however, that Russia immediately demobilize. Russia declined to do so, and on August 1 Germany declared war on Russia.

a) Transition Material between Beginning of Segment and Initial Paragraph of Central Path.

—– 20% Bushy path in segment 1 (degree >= 18) —–

—– Paragraph 24411.p22 —–
The French began to mobilize on the same day; on August 2 German troops traversed Luxembourg and on August 3 Germany declared war on France. On August 2 the German government informed the government of Belgium of its intention to march on France through Belgium in order, as it claimed, to forestall an attack on Germany by French troops marching through Belgium. The Belgian government refused to permit the passage of German troops and called on the signatories of the Treaty of 1839, which guaranteed the neutrality of Belgium in case of a conflict in which Great Britain, France, and Germany were involved, to observe their guarantee. Great Britain, one of the signatories, on August 4 sent an ultimatum to Germany demanding that Belgian neutrality be respected; when Germany refused, Britain declared war on it the same day. Italy remained neutral until May 23, 1915, when, to satisfy its claims against Austria, it broke with the Triple Alliance and declared war on Austria-Hungary. In September 1914 Allied unity was made stronger by the Pact of London, signed by France, Great Britain, and Russia. As the war progressed, other countries, including Turkey, Japan, the U.S., and other nations of the western hemisphere, were drawn into the conflict. Japan, which had made an alliance with Great Britain in 1902, declared war on Germany on August 23, 1914. The United States declared war on Germany on April 6, 1917.

—– Paragraph 24411.p32 —–
On the eastern front, in accordance with the plans of the Allies, the Russians assumed the offensive at the very beginning of the war. In August 1914 two Russian armies advanced into East Prussia, and four Russian armies invaded the Austrian province of Galicia. In East Prussia a series of Russian victories against numerically inferior German forces had made the evacuation of that region by the Germans imminent, when a reinforced German army commanded by General Paul von Hindenburg decisively defeated the Russians in the Battle of Tannenberg, fought on August 26-30, 1914. The four Russian armies invading Austria advanced steadily through Galicia; they took Przemysl and Bukovina, and by the end of March 1915 were in a position to move into Hungary. In April, however, a combined German and Austrian army drove the Russians back from the Carpathians. In May the Austro-German armies began a great offensive in central Poland, and by September 1915 had driven the Russians out of Poland, Lithuania, and Courland, and had also taken possession of all the frontier fortresses of Russia. To meet this offensive the Russians withdrew their forces from Galicia. The Russian lines, when the German drive had ceased, lay behind the Dvina River between Riga and Dvinsk (Daugavpils), and then ran south to the Dnestr River. Although the Central Powers did not force a decision on the eastern front in 1914-15, the Russians lost so many men and such large quantities of supplies that they were subsequently unable to play any decisive role in the war. In addition to the Battle of Tannenberg, notable battles on this front during 1914-15 were the First Battle of the Masurian Lakes (September 7-14, 1914), and the Second Battle of the Masurian Lakes (February 7-21, 1915), both German victories.

—– Paragraph 24411.p36 —–
Turkey entered the war on October 29, 1914, when Turkish warships cooperated with German warships in a naval bombardment of Russian Black Sea ports; Russia formally declared war on Turkey on November 2, and Great Britain and France followed suit on November 5. In December the Turks began an invasion of the Russian Caucasus region. The invasion was successful at its inception, but by August 1915 the hold that Turkish forces had gained had been considerably reduced. Turkish pressure in the area, however, impelled the Russian government early in 1915 to demand a diversionary attack by Great Britain on Turkey. In response, British naval forces under the command of General Sir Ian Hamilton bombarded the Turkish forts at the Dardanelles in February 1915, and between April and August, two landings of Allied troops took place on the Gallipoli Peninsula, one of British, Australian, and French troops in April, and one of several additional British divisions in August. The Allied purpose was to take the Dardanelles; however, strong resistance by Turkish troops and bad generalship on the part of the Allied command resulted in complete failure. The Allied troops were withdrawn in December 1915 and January 1916 (see Gallipoli Campaign).

b) Central Path in Segment 1.

Table 8: Text Summary for Segment 1 (24411 World War 1).

24411: World War I (107 Paragraphs)

Segment 3: **3 paragraphs**: 24411.51, p53, p55
(Specialized Segment: President Woodrow Wilson).

—– Paragraph 24411.p53 —–
In 1916 President Woodrow Wilson of the U.S., at that time a neutral nation, attempted to bring about negotiations between the belligerent groups of powers that would in his own words bring "peace without victory." As a result of his efforts, and particularly of the conferences held in Europe during the year by Wilson's confidential adviser, Colonel Edward M. House, with leading European statesmen, some progress was at first apparently made toward bringing an end to the war. In December the German government informed the U.S. that the Central Powers were prepared to undertake peace negotiations. When the U.S. informed the Allies, Great Britain rejected the German advances for two reasons: Germany had not laid down any specific terms for peace; and the military situation at the time (Romania had just been conquered by the Central Powers) was so favorable to the Central Powers that no acceptable terms could reasonably be expected from them. Wilson continued his mediatory efforts, calling on the belligerents to specify the terms on which they would make peace. He finally succeeded in eliciting concrete terms from each group, but they proved irreconcilable.

a) Transition Material to Segment 3
(Sim(24411.p53, 24411.p55) = 0.2956)

—– Paragraph 24411.p55 —– Wilson still attempted to find some basis of agreement between the two belligerent groups until a change in German war policy in January 1917 completely altered his point of view toward the war. In that month Germany announced that, beginning on February 1, it would resort to unrestricted submarine warfare against the shipping of Great Britain and all shipping to Great Britain. German military and civil experts had calculated that such warfare would bring about the defeat of Great Britain in six months. Because the U.S. had already expressed its strong opposition to unrestricted submarine warfare, which, it claimed, violated its rights as a neutral, and had even threatened to break relations with Germany over the issue, Wilson dropped his peacemaking efforts. On February 3, the U.S. broke diplomatic relations with Germany and at Wilson's request a number of Latin American nations, including Peru, Bolivia, and Brazil, also did so. On April 6 the United States declared war on Germany.

b) Central Path in Segment 3

Table 9: Text Summary for Specialized Segment 3 (24411 World War 1).

REFERENCES

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

2. J.P. Callan, Passage-Level Evidence in Document Retrieval, Proc. SIGIR '94, Springer Verlag, Berlin, 301-310.

3. R. Wilkinson, Effective Retrieval of Structured Documents, Proc. SIGIR '94, Springer Verlag, Berlin, 311-317.

4. D. Knaus, E. Mittendorf and P. Schauble, Improving a Basic Retrieval Method by Links and Passage Level Evidence, Text Retrieval Conference, Washington, D.C., November 1994.

5. S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu and M. Gatford, Okapi at TREC 3, Text Retrieval Conference, Washington, D.C., November 1994.

6. G. Salton, J. Allan and C. Buckley, Approaches to Passage Retrieval in Full Text Information Systems, Proc. SIGIR-93, *Association for Computing Machinery*, New York, June 1993, 49-58.

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

8. R.H. Trigg and M. Weiser, Textnet: A Network-Based Approach to Text Handling, Transactions on Office Information Systems, 4:1, January 1986, 1-23.

9. P. Gloor, Cybermap: Yet Another Way of Navigating Hyperspace, Proc. Hypertext '91, *Association for Computing Machinery*, New York, December 1991, 107-121.

10. R.S. Gilyarevskii and M.M. Subbotin, Russian Experience in Hypertext, Automatic Computing of Coherent Texts, Journal of the Am. Soc. for Information Science, 44:4, 1993, 185-193.

11. 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.

12. M.A. Hearst and C. Plaunt, Subtopic Structuring for Full-Length Document Access, Proc. SIGIR '93, *Association for Computing Machinery*, New York, 59-68.

13. G. Salton and J. Allan, Automatic Text Theme Decomposition and Structuring, Proc. RIAO-94, Centre d'Information Documentaire, Paris, November 1994, 6-20.

14. G. Salton and A. Singhal, Automatic Text Theme Generation and the Analysis of Text Structure, Technical Report TR 94-1438, Computer Science Department, Cornell University, Ithaca, NY, June 1994.

15. G. Salton, J. Allan, C. Buckley and A. Singhal, Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts, *Science* 264, 1421-1426, 3 June 1994.

16. G. Salton and J. Allan, Selective Text Utilization and Text Traversal, Proc. Hypertext-93, *Association for Computing Machinery*, New York, November 1993, 131-144.