Automatic Text Theme Generation and the Analysis of Text Structure

Gerard Salton*
Amit Singhal

TR 94-1438
July 1994

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

---

*Department of Computer Science, Cornell University, Ithaca, NY 14853-7501.
This study was supported in part by the National Science Foundation under grant IRI 9300124.

Automatic Text Theme Generation and the Analysis of Text Structure

Gerard Salton* and Amit Singhal

Abstract

Non-expository texts are not usually read from cover to cover. Readers are helped in such circumstances by providing selective access to text excerpts as needed. Text themes can be identified representing areas of importance in a text, and summaries can be constructed automatically. In this study, text theme generation and text summarization are related to text structure. It is shown that useful text derivatives are obtainable for texts with diverse structural characteristics.

1 Introduction

Much of the written information which circulates around the world is now available in machine-readable form, and can therefore be processed automatically in accordance with particular user requirements. Among the available texts of potential interest are non-expository writings, such as legal and instructional materials, as well as directives and regulatory data which are meant to be read selectively rather than from cover to cover. Readers of such data will want to concentrate on those texts passages which appear of special importance in particular circumstances. Automatic methods which allow readers to traverse texts selectively in accordance with individual

*Department of Computer Science, Cornell University, Ithaca, NY 14853-7501. This study was supported in part by the National Science Foundation under grant IRI 9300124.

requirements, and procedures providing text summaries of various kinds are therefore especially important.

A flexible text utilization system must be based on a careful identification of text content and text structure. Specifically, the main topics covered by a text must be isolated and the manner in which the subject matter is expressed and presented must be studied. In this report methods are given for the automatic determination of text themes, that is, subject areas which are emphasized in the text(s) under consideration. Assuming that each theme is represented by selected text excerpts, the themes can then be used to build text summaries by a proper choice of relevant text passages, and appropriate reading patterns and text traversal strategies can be generated.

In deriving the text themes, we make use of various capabilities incorporated into the Smart text retrieval system. [1,2] Among the most important is the ability to retrieve variable-length text passages rather than only full document texts; the possibility to obtain a measure of text similarity reflecting the degree to which two text excerpts cover the same subject matter; and finally the ability to use local context-checking methods to recognize various linguistic ambiguities. The assumption made in this last case is that words or expressions occurring in similar contexts have identical meanings, and that the reverse is likely to be true when the local contexts differ. Thus, when two highly matching texts are detected, indicating a substantial number of common terms, an attempt is made to find local substructures, for example text sentences, that also match. When that is the case, the corresponding texts are assumed to covered similar subject matter, and the text pair is accepted as related; otherwise the texts are assumed to be unrelated even when substantial vocabulary similarities exist. The basic text matching procedures have proved effective in various text retrieval environments. [3]

2 Text Theme Generation

The text theme generation strategy is illustrated in Table 1. The system is based on the computation of similarities between texts, or text excerpts. Specifically, let each text D_i be represented by a vector of the form $D_i = (d_{i1},\ d_{i2},\ \ldots, d_{it})$, where d_{ik} represents a weight, or importance factor, for term T_k assigned to document D_i. Thus given two texts D_i and D_j, a similarity measure between the texts can be computed based on coincidences between the respective term assignments, for example, as the inner product of the corresponding term vectors: sim $(D_i, D_j) = \sum_{k=1}^{t} d_{ik}d_{jk}$. An appropriate similarity threshold, known as the compare similarity value (csim value) is chosen, and text similarity values above the threshold are considered significant.

The computation of similarities between text pairs then leads to the generation of a text relationship map, where the vertices, or nodes, represent text excerpts, and the links, or branches, between pairs of nodes represent similarities between the corresponding text excerpts exceeding the state csim value. [4,5] The Figures attached to this report represent paragraph relationship maps where the nodes designate text paragraphs and the links are paragraph similarities.

The text theme generation program takes the information contained in the text relationship maps, and recognizes areas with a heavy concentration of links where a small number of text excerpts are closely interrelated. More specifically, all groups of three mutually related text excerpts are recognized first; such groups are represented by triangles in the text relationship map. (Triangles are used because the number of similar pairs is often very large, while the number of mutually related quadruples may be very small or zero.) For each triangle, a centroid vector is computed, representing the average vector for the three related text excerpts defining each triangle. Finally, triangles with sufficiently similar centroids are merged. A threshold, known as the theme similarity (tsim value) is used to control the triangle merging operation. When the centroid similarity exceeds the stated tsim value, the corresponding triangles are merged, and a new centroid value is computed for the merged entity. The merging operation continues until no further merge operations are possible, and the resulting sets of merged triangles, or higher-level entities, are identified with the *text themes.* [5,6]

A text theme can be characterized by the text excerpts included in a group of merged triangles. Alternatively, a theme vector may be generated for each set of merged triangles, defined as the centroid of all the vectors representing text excerpts included in the given set of grouped triangles. In general, a low csim value, such as 0.20, will insure that most significant relationships between texts are included in the theme computation. The tsim value, on the other hand, should be low enough to make sure that relatively similar, partly overlapping, themes are actually merged into a single theme; at the same time, the tsim value must be high enough so that distinct themes without overlapping text excerpts are kept separate. For the theme computations used in this study, a tsim value of 0.40 is used as a default when no other information about theme composition and overlap is known. By making judicious choices of the csim and tsim thresholds, it is possible to insure that the largest possible number of non-overlapping themes will actually be generated for each text sample.

3 Selective Text Summarization

Given a knowledge of text themes and of the text excerpts corresponding to each theme, a *tour* may be defined as an ordered sequence of text identifiers used for theme definition (that is, used to define triangles of a text relationship map) taken in chronological text order. [7] When text excerpts from more than one document are included in a text relationship map, the excerpt identifiers are listed in increasing document number order, while still maintaining the normal text order within each document. Normally, the document number order corresponds to a text ordering in increasing order of the date of publication, or data of acquisition of the documents. Hence the texts will be included in a tour in the expected order with the earliest texts listed first.

Each tour can be directly associated with a corresponding selective text traversal strategy which consists in traversing the documents in the given tour order while selecting only the text passages included in the tour. [8] Alternatively, the set of text excerpts defining the tour can be used for text summarization purposes. [9-11]. Because the tour construction is based on text excerpts that define the text themes, the tours, and hence the text summaries, include mostly important text excepts characterized by dense linking patterns in the text relationship maps.

By suitably varying the compare similarity (csim) parameter, it is possible to obtain longer or shorter tours, as appropriate for particular users and applications. Thus a high csim threshold produces sparse text relationship maps with few links, and few excerpts per theme, resulting in short tours or brief summaries. On the contrary, a low csim value results in dense maps where each theme includes many text elements. The resulting tours and summaries will then be longer and more elaborate. This suggests that a variable csim threshold be used to produce summaries of appropriate length. In practice, the user might give a size parameter for the desired text extract, for example, as a percentage of the total number of paragraphs included in the complete text. By varying the csim value, a tour can then be obtained that approaches the desired summary size as closely as possible. A program of this type of shown in Table 2.

It may be noted that the previously mentioned theme merging operation is not needed for the tour construction and text extracting systems. Thus, the tsim threshold value does not play any role in this case. Typically, a 5 percent text extract might then be used to obtain a quick overview of text content, while a 20 percent extract would provide a reasonably thorough text summary.

Table 3 provides a summary of parameter characteristics used for tour construction and text summarization purposes for four documents included in the US Federal Register, a collection of some 46,000 documents (410 Mbytes) dealing with rules and regulations by regulatory and other government agencies. In addition to the four single texts (documents 2758, 7793, 44496, and 45858), one other example is included in Table 3 dealing with tour construction for three closely related items (items 9281, 23352, and 31262).

To generate the output of Table 3, attempts were made to obtain short (5%), medium-size (10%), and long (20%) tours for the sample documents. This turned out to be possible for most of the longer documents. For the short documents, the extracts corresponding to a single theme often cover more than the stipulated 5 percent of the full text, thus making it impossible to generate very short extracts.

Consider, as an example, document 2758 covering an airworthiness directive issued by the Federal Aviation Administration addressed to Mitsubishi Heavy Industries, a Japanese aircraft manufacturer. A single theme, roughly entitled, "vertical stabilizer front spar fuselage fitting" is generated using csim and tsim parameters of 0.66 and 0.40, respectively. Figure 1(a) shows that the six paragraphs defining the theme, corresponding to a 13 percent tour. The resulting text summary is shown in detail in Table 4. The first paragraph of the tour (paragraph 9 of document 2758) is a central paragraph with 9 links to other text paragraphs giving a summary of the complete document. The other paragraphs of the tour are added to provide details about the work to be accomplished in inspecting the vertical stabilizer fitting on various kinds of aircraft.

When the csim parameter is lowered to 0.46, one additional theme is identified, entitled, "Mitsubishi Heavy Industries", consisting of the cross-hatched triangle in Figure 1(b), and four new paragraphs are added to the summary of Table 4. The resulting longer 21 percent summary still includes the basic summary paragraph (2758.p9) followed by the text shown in Table 5. A comparison of Tables 4 and 5 shows that the provisions of the directive are treated more thoroughly in

Figure 5 than in Figure 4.

The output of Table 3 reveals similar characteristics for the theme and tour generation systems applied to other text examples. A single theme, corresponding to a 9 percent tour is obtained for document 2793 with csim and tsim parameters of 0.675 and 0.40, respectively. When the tsim threshold is raised to 0.50, the theme merging is rendered more difficult. The original single theme is then broken into two pieces. This same effect is also noticed for the larger 25 percent tour for item 2793 obtained with a csim threshold of 0.625 and tsim values of 0.40 and 0.50, respectively. In general, when the tsim threshold is lowered, the theme merging operation is encouraged and the number of themes decreases. The reverse is true when the tsim value is increased. Changes in tsim values will not affect the text extraction system producing tours and summaries.

4 Theme Generation and Text Structure

The text themes represent areas of subject concentration identified by a number of related text excerpts treating similar subject matter. When these related text excerpts are ordered, the resulting tour may be more or less useful for the intended user. The usability of a term for text summarization depends on the degree to which the main subject matter is covered in the summary (subject coverage), the degree to which all text subjects are included in the summary (comprehensiveness), and the readability of the summary (coherence and cohesion of the text).

The comprehensiveness and coherence factors tend to vary with text type and text structure. The following general rules apply to the tour construction and text summarization systems:

1. Short documents that often exhibit a single, central theme are normally easier to treat than longer documents covering a large variety of different topics. In the latter case, the text summary may lack comprehensiveness when one or another topic present in the complete

text is missing from the summary.

2. Documents with a generally convex text relationship map, such as for example the airworthiness directive treated in Figure 1 (document 2758), are much easier to traverse and read selectively without loss of crucial information than items with many disconnected outliers where the structure lacks convexity. The relationship map will be convex when most adjacent text excerpts are related, and hence linked in the map. When such related excerpts are included in a tour, the transitions from one excerpt to the next are generally smooth and the summary is appropriately coherent. In the tour of Figure 1(a), each tour paragraphs is related to the next one, and the summary of Table 4 is easy to read. When the text map lacks convexity, some adjacent tour paragraph may not be linked and the resulting summary may lack coherence.

3. The text summaries may be expected to be crisp when the text themes are disconnected, in the sense that the paragraphs defining the various themes do not overlap. On the other hand, when the themes overlap, the same information may be covered in several components of a tour, and the corresponding summary may contain redundant information.

4. When the text relationship maps include excerpts from several different documents, the text extraction system is most effective for multiple texts covering related subject areas. In that case, particular text excerpts of one document are linked to similar excerpts covering related matter in the other documents, leading to representative themes and easy text summarization.

The connection pattern of the text relationship maps can be characterized by various objective measures, such as the number of connected pieces (components) in the map; the number of dense areas on the map where many connections exist within a particular group of linked text excerpts, but the connections to be outside are sparse; and the number of areas on the map where groups of adjacent text excerpts are disconnected both from the immediately preceding and the immediately following excerpts. In general, the text processing task is simplified when no disconnections are present and the map consists of a single component in which all adjacent excerpts are properly linked, and the links cover the map uniformly. The tour construction and text summarization become more difficult as the degree of disconnection in the relationship map becomes greater.

The relation between text structure and theme or tour output is illustrated in the examples of Figures 1 to 5. The previously mentioned airworthiness directive of document 2758, described in Figure 1, represents a well-connected structure. Only one component exists, and for the most part, paragraphs that are adjacent on the map such as p39, p41, p6, p9, exhibit pairwise connections. One partially disconnected component exists in Figure 1, consisting of paragraphs p6, p11, p12, p17, p39, and p41. However, the number of internal links between the paragraphs (9) is not much larger than the number of links to outside nodes (6). In any case, only one main theme is apparent and the resulting tours and summaries are complete and coherent.

Document 7793 illustrated in Figure 2 represents a directive from the Commodity Credit Corporation of the US Department of Agriculture dealing with various aspects of the feed grain program, such as the acreage reduction program, and rules relating to marketing loans and loan deficiency payments. Because the document covers a number of unrelated topics, the text is decomposable into components that are not well-connected to their immediate neighbors, such as for example, the section between paragraphs 20 and 35, or between paragraphs 43 and 46. It is not surprising in these circumstances that several themes are recognizable. Figure 2(b) shows the four themes obtained with a tsim threshold of 0.50, leading to a 25 percent tour of 16 paragraphs that covers the main areas of subject concentration. When the tsim threshold is lowered to 0.40, the two themes shown at the bottom of Figure 2(b), (7793.p36-p37-p40, and 7793.p48-p49-p51-p56) covering the marketing loan program for feed grains are merged into one larger theme.

A still greater degree of disconnection is shown in Figure 3 for document 44496 covering an announcement of the availability of competitive research grants from the US Department of Agriculture. Here certain completely disconnected components exist, such as the linked pair 44496.p32 and p34. Furthermore a large number of outliers are present (that is, nodes with a single link) that by definition can never be included in any theme, since the theme construction method is based on triangles of nodes carrying at least two links. Examples of such outliers are 44496.p24, p49, p64, p65, p69, p70, p71, and so on. Because the research grants program described in document 44496 covers a number of different, unrelated topics, it is not surprising that the paragraph connection pattern does not exhibit the convexity properties noticed earlier in Figure 1. As a consequence the summaries constructed from the tours may lack some comprehensiveness, and the coherence between certain adjacent paragraphs in the summary may be imperfect. The example of Figure 3 makes it clear that the tour or summary coverage becomes increasingly more complete as the csim parameter is lowered (from 0.51 in Figure 3(a) to 0.36 in Figure 3(b)).

The map of Figure 4 represents paragraph relations for document 45858, covering amendments of certain sections of the revenue act for pipe tobacco. This text contains 298 very short, partly repetitive paragraphs giving the amended wording compared with the previously valid regulations. Because of the repetitions in the covered material, a large number of links are evident in Figure 4 even at a high csim threshold of 0.70. Nevertheless, the connection pattern is imperfect, as demonstrated by the disconnected part between paragraphs p30 and p34.

When repetitive material is processed, it may be convenient to eliminate the links between identical text excerpts thereby reducing the chances of redundancy in the tour or summary outputs. This can be achieved by using a maximum similarity (msim) threshold above which links will not be recognized. For the output of Figure 4, the maximum allowed similarity between excerpts is 0.90. Furthermore, when the text excerpts are very short, and consist of only a few words as in the case of document 45858, the number of significant paragraph similarities may become large. Since a close similarity between short paragraph pairs may not be indicative of important subject relationship, it is useful to eliminate the very short paragraphs from the theme generation and text summarization processes. This is done by specifying a minimum vector size (vsize) for the text excerpts under consideration. Effectively, a vsize specification eliminates those text excerpts for which the number of terms in the analyzed term vector falls below the specified vsize threshold. The text relationship map of Figure 4 is obtained with msim and vsize parameters of 0.90 and 5, respectively.

A short, 7 percent tour for document 45858 is shown in Figure 4, corresponding to the four themes labeled 1 to 4 in the Figure. Forty-one partly overlapping themes are found for the 20 percent tour for document 45858 using csim and tsim thresholds of 0.66 and 0.80, respectively. These 41 themes merge into only 7 different themes when the tsim parameter is lowered to 0.70. The large 20 percent tour includes 62 paragraphs covering most of the dense areas in the relationship map.

The last example of Figure 5 deals with multi-document maps for three parallel documents representing different meeting announcements that concern the Integrated Services Digital Network (ISDN). The parallel nature of the documents is evident from the connection pattern in the map which reveals close connections between 9281.p3-p6-p7, 23352.p4-p5, and 31262.p3-p6-p7-p8. The short 6 percent tour of Figure 5(a) contains excerpts from document 9281 only. The corresponding 3-paragraph summary is shown in Table 6. Note that the essence of the full text of all three items is contained in only three paragraphs of one document as shown in Table 6. The long 10-paragraph (20 percent) summary of Table 7 adds information about the other two meetings as well. For groups of parallel documents, such as documents 9281, 23352, 31262, the output is comparable in usefulness to that obtained for single convex documents. The subject area is properly represented, and the derived text products are comprehensive and coherent.

5 Evaluation of Theme Construction and Text Summarization

No simple methods exist for evaluating text extracting or text summarization systems. Some global evaluation data derived from an examination of 27 text relationship maps, involving excerpts from 43 different documents, are included in Table 8. The evaluation data are based on an examination of the 76 derived text themes, and the 27 summaries corresponding to 20 percent tours. All the derived text themes appear to be correctly represented areas of emphasis in the subject matter treated in the documents. Correspondingly, the subject coverage of the summaries appears acceptable in every case. On the other hand, some lack of coherence is evident in 8 of the 27 derived summaries, and the summaries are not always comprehensive in the sense that certain topic areas mentioned in the full text might not be included in a short summary.

The coherence suffers when an included text excerpt starts with a transition term, such as "accordingly", "in view of this", etc., and the preceding excerpt does not include the correct referent. Alternatively, the "glue" may be missing between two adjacent summary excerpts that are in fact not adjacent in the full text. Analogously, comprehensiveness may suffer when the subject treatment is disconnected, as in texts representing catalogs of different subjects or products, inventory lists, repair manuals for different components of a machine, and so on.

In general, the proposed theme generation and text summarizations systems provide usable output for a wide variety of text collections and subject matter, including many classes of non-expository text that are difficult to handle by alternative methods.

References

1. G. Salton, editor, The Smart Retrieval System — Experiments in Automatic Document Processing, Prentice Hall, Inc., Englewood Cliffs, N.J., 1971.

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

3. G. Salton, C. Buckley and J. Allan, Automatic Structuring and Retrieval of Large Text Files, *Communications of the ACM* 37:2, February 1994, 97-108.

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

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

6. M.R. Hearst and C. Plaunt, Subtopic Structuring for Full-Length Document Access, Proc. SIGIR-93 16th Annual Int. ACM-SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, June 1993, 59-68.

7. C. Guinan and A.F. Smeaton, Retrieval from Hypertext Using Dynamically Planned Guided Tours, Proc. ECHT '92, European Conference on Hypertext, Association for Computing Machinery, New York, December 1992, 122-130.

8. J. O'Connor, Retrieval of Answer Sentences and Answer Figures by Text Searching, Information Processing and Management, 11:5/7, 1975, 155-239.

9. H.P. Luhn, The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development 2:2, April 1958, 159-165.

10. H.P. Edmundson and R.E. Wyllys, Automatic Abstracting and Indexing - Survey and Recommendations, Comm. of the ACM, 4:5, 226-234, May 1961.

11. 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 Am. Soc. of Info. Science, 22:4, July-August 1964, 260-274.

1. Generate text relationship map exhibiting similarities between text passages above a particular similarity threshold (csim value)

2. Identify groups of three mutually related text pieces (triangles in text relationship map)

3. Compute central vector (centroid) for each text group.

4. Merge text groups into larger similarity classes whenever the corresponding centroid similarity exceeds a stated similarity threshold (tsim value)

5. Repeat steps 3 and 4 until no further merging is possible. Identify each of the final text groupings as a text theme.

Table 1. Generation of Text Themes

1. Choose desired extract size (as percentage of total number of paragraphs in text)

2. Use variable threshold for text similarity (csim) parameter to generate groups of three mutually related text passages (triangles) in such a way that total number of paragraphs in the groups of three is close to desired extract percentage.

3. List grouped paragraphs in chronological text order and use for text summarization and for selective text traversal.

4. For extracts belonging to several documents, use chronological document order, and within each document use correct text order.

Table 2. Variable Size Text Extracting

Document Numbers	Topic Characteristics	Parameters	Number of Themes	Percent Tour	Paragraphs in Tour
2758	airworthiness	csim 0.66	1	13%	6
	directive from FAA	csim 0.46	2	21%	10
	(46 paragraphs)	tsim 0.40			
7793	directive regarding	csim 0.675			
	feed grains and	tsim 0.40	1	9%	6
	crops from CCC	tsim 0.50	2	9%	6
	(64 paragraphs)	csim 0.625			
	(some disconnected	tsim 0.40	3	25%	16
	pieces)	tsim 0.50	4	25%	16
44496	competitive research	csim 0.51	2	6%	6
	grants program from	csim 0.41	2	11%	10
	DA (disconnected)	csim 0.36	4	20%	18
	(80 paragraphs)	tsim 0.40			
		tsim 0.35	3	20%	18
45858	regulations regarding	csim 0.75	4	7%	23
	taxation of pipe	csim 0.78	3	10%	23
	tobacco from BATF	csim 0.66	41	20%	62
	msim 0.90, vsize 5	tsim 0.80			
	(298 paragraphs)	tsim 0.70	7	20%	62
9281	meeting notices	csim 0.63	1	6%	3
23352	regarding ISDN	csim 0.58	1	8%	4
31262	from NBS	csim 0.38	1	20%	10
	(48 paragraphs)	tsim 0.40			

Table 3. Characteristics of Themes and Tours

- - - - - - - - - - - - - - -
Dtext 2758.p9 p.18-19 .p22 .p34 .p38
- - - -

- - - - - Paragraph 2758.p9 - - - - -
SUMMARY: This notice proposes to supersede an existing airworthiness directive (AD), applicable to the Mitsubishi Heavy Industries, Limited (MHI), Model YS-11/-11A series airplanes, which currently requires replacement of the vertical stabilizer front spar fitting attachment bolts. This proposal would add a requirement to replace the attaching washers and nuts, inspect certain vertical stabilizer-to-fuselage attachment fittings for cracks and corrosion, and accomplish corrosion preventative treatment on certain parts of the fitting attachment assembly. This action is prompted by a report of a cracked lug and corrosion found in the vertical stabilizer front spar fuselage side fitting. Failure of the attachment fittings could lead to the structural failure of the vertical stabilizer and loss of control of the airplane.

- - - - - Paragraph 2758.p18 - - - - -

Discussion: On January 29, 1986, FAA issued AD 86-03-05, Amendment 39-5233 (51 FR 4304; February 4, 1986), to require replacement of the vertical stabilizer front spar fitting attachment bolts on Mitsubishi Heavy Industries (MHI) [formerly Nihon Aeroplane Manufacturing Company (NAMC)] Model YS-11/-11A series airplanes. That action was prompted by a report of a failure of a vertical stabilizer front spar fuselage side fitting attachment bolt due to stress corrosion. Failure of the attachment fitting could lead to the structural failure of the vertical stabilizer and loss of control of the airplane.

- - - - - Paragraph 2758.p19 - - - - -
Since issuance of that AD, MHI received a report that, during a routine periodic inspection, a cracked lug and corrosion were found in the vertical stabilizer front spar fuselage side fitting, and corrosion was found on the front spar stabilizer side fitting and the tapered joining bolt. Failure of this joining bolt could contribute to the structural failure of the vertical stabilizer and consequent loss of control of the airplane.

- - - - - Paragraph 2758.p22 - - - - -
Since this condition is likely to exist or develop on other airplanes of the same type design registered in the U.S., an AD is proposed which would require replacement of the bolt, washer, and nut installed in the vertical stabilizer front spar fuselage side fittings; replacement of certain other parts of this assembly, if conditions warrant; and inspection of the vertical stabilizer and fuselage fittings for cracks and corrosion, in accordance with the service bulletins previously mentioned.

- - - - - Paragraph 2758.p34 - - - - -
To prevent failure of the vertical stabilizer front spar to fuselage fittings, accomplish the following:

- - - - - Paragraph 2758.p38 - - - - -
D. The repetitive inspections required by Paragraph A., above, may be terminated if the vertical stabilizer front spar fuselage side fitting P/N 01-381010-11/-12 had been given corrosion preventive treatment after September 1, 1985, or once it has been replaced by fitting P/N 01-38101-21/-22.

Table 4. Typical Medium-Size (13%) Summary (Document 2758 Airworthiness Directive Issued by Federal Aviation Administration)

- - - - - - - - - - - - - - -
Dtext 2758.p9 p.18-20 .p22 .p34-38
- - - -

- - - - - Paragraph 2758.p18 - - - - -

Discussion: On January 29, 1986, FAA issued AD 86-03-05, Amendment 39-5233 (51 FR 4304; February 4, 1986), to require replacement of the vertical stabilizer front spar fitting attachment bolts on Mitsubishi Heavy Industries (MHI) [formerly Nihon Aeroplane Manufacturing Company (NAMC)] Model YS-11/-11A series airplanes. That action was prompted by a report of a failure of a vertical stabilizer front spar fuselage side fitting attachment bolt due to stress corrosion. Failure of the attachment fitting could lead to the structural failure of the vertical stabilizer and loss of control of the airplane.

- - - - - Paragraph 2758.p19 - - - - -
Since issuance of that AD, MHI received a report that, during a routine periodic inspection, a cracked lug and corrosion were found in the vertical stabilizer front spar fuselage side fitting, and corrosion was found on the front spar stabilizer side fitting and the tapered joining bolt. Failure of this joining bolt could contribute to the structural failure of the vertical stabilizer and consequent loss of control of the airplane.

- - - - - Paragraph 2758.p20 - - - - -
MHI issued NAMC YS-11 Service Bulletin 53-70 and Alert Service Bulletin A53-71, both dated May 23, 1986, which provide instructions for replacement of certain attachment bolts, washers, and nuts; and procedures for inspection and corrosion preventive treatment of the vertical stabilizer front spar fuselage side fittings and the vertical stabilizer front spar stabilizer side fittings. The Japanese Civil Aviation Bureau (JCAB) issued Japanese Airworthiness Directive No. TCD-2614-86, dated June 20, 1986, making NAMC YS-11 Service Bulletins 53-70 and A53-71 mandatory on all NAMC Model YS-11/-11A airplanes under Japanese registry.

- - - - - Paragraph 2758.p22 - - - - -
Since this condition is likely to exist or develop on other airplanes of the same type design registered in the U.S., an AD is proposed which would require replacement of the bolt, washer, and nut installed in the vertical stabilizer front spar fuselage side fittings; replacement of certain other parts of this assembly, if conditions warrant; and inspection of the vertical stabilizer and fuselage fittings for cracks and corrosion, in accordance with the service bulletins previously mentioned.

- - - - - Paragraph 2758.p34 - - - - -
To prevent failure of the vertical stabilizer front spar to fuselage fittings, accomplish the following:

- - - - - Paragraph 2758.p35 - - - - -
A. Within 600 hours time-in-service after the effective date of this AD or within 4 months after the effective date of this AD, whichever occurs first, visually inspect the vertical stabilizer front spar fuselage side fittings, Part Number (P/N) 01-38101-11/-12, for cracked lugs, in accordance with Paragraph 2, "Instructions," of NMAC YS-11 Alert Service Bulletin (SB) A53-71, dated May 23, 1986. Repeat this inspection at intervals not to exceed 1,000 hours time-in-service.

- - - - - Paragraph 2758.p36 - - - - -
B. If any crack is found in fitting P/N 01-38101-11/-12 during the inspections required by paragraph A., above: prior to further flight, remove that fitting from the airplane and accomplish the inspections, corrosion treatment, and replacement of parts, as necessary, in accordance with Paragraph 2, "Instruction," of NAMC YS-11 Service Bulletin 53-70, dated May 23, 1986. Once this has been accomplished, the required repetitive inspections may be discontinued.

- - - - - Paragraph 2758.p37 - - - - -
C. If no cracking is found in fitting P/N 01-38101-11/-12 during the inspections required by Paragraph A., above: within 6,000 hours time-in-service after the effective date of this AD, or by January 1, 1990, whichever occurs first, accomplish the inspections, corrosion treatment, and replacement of parts, as necessary, in accordance with Paragraph 2, "Instructions," of NAMC YS-11 Service Bulletin 53-70, dated May 23, 1986. Once this has been accomplished, the required repetitive inspections may be discontinued.

- - - - - Paragraph 2758.p38 - - - - -
D. The repetitive inspections required by Paragraph A., above, may be terminated if the vertical stabilizer front spar fuselage side fitting P/N 01-381010-11/-12 had been given corrosion preventive treatment after September 1, 1985, or once it has been replaced by fitting P/N 01-38101-21/-22.

Table 5. Typical Long (21%) Summary Document 2758 Airworthiness Directive Issued by Federal Aviation Administration)

- - - - - - - - - - - - - - -
Dtext 9281.p3 p.6 .p7
- - - -

- - - - - Paragraph 9281.p3 - - - - -
Announcement of Workshops for Users and Implementors of Integrated Services Digital Network (ISDN)

- - - - - Paragraph 9281.p6 - - - - -
SUMMARY: The Institute for Computer Sciences and Technology at the National Bureau of Standards (NBS) announces another workshop of a continuing workshop series to discuss issues related to the use and implementation of Integrated Services Digital Network (ISDN) technology. These workshops are part of the North American ISDN Users' Forum (NIU-FORUM) which was formed recently under the auspices of NBS to create a strong user voice in the implementation of ISDN and ISDN applications and to ensure that the emerging ISDN meets users' application needs.

- - - - - Paragraph 9281.p7 - - - - -
DATES: A Joint ISDN Users' and Implementors' Workshop will be held in St. Louis, Missouri on September 27, 28, and 29, 1988. The Users' Workshop will continue to identify, define and prioritize applications of ISDN and the Implementators' Workshop will define implementation agreements for ISDN and sponsor multivendor trials and demonstrations.

Table 6. Typical Short Multi-Document (6%) Summary (Documents 9281-23352-31262 covering meeting announcements)

- - - - - - - - - - - - - - -
Dtext 9281.p3 p.6-8 23352.p4-5 31262.p6-8
- - - -

- - - - - Paragraph 9281.p3 - - - - -

Announcement of Workshops for Users and Implementors of Integrated Services Digital Network (ISDN)

- - - - - Paragraph 9281.p16 - - - - -
SUMMARY: The Institute for Computer Sciences and Technology at the National Bureau of Standards (NBS) announces another workshop of a continuing workshop series to discuss issues related to the use and implementation of Integrated Services Digital Network (ISDN) technology. These workshops are part of the North American ISDN Users' Forum (NIU-FORUM) which was formed recently under the auspices of NBS to create a strong user voice in the implementation of ISDN and ISDN applications and to ensure that the emerging ISDN meets users' application needs.

- - - - - Paragraph 9281.p7 - - - - -
DATES: A Joint ISDN Users' and Implementors' Workshop will be held in St. Louis, Missouri on September 27, 28, and 29, 1988. The Users' Workshop will continue to identify, define and prioritize applications of ISDN and the Implementators' Workshop will define implementation agreements for ISDN and sponsor multivendor trials and demonstrations.

- - - - - Paragraph 9281.p8 - - - - -
ADDRESS: To obtain registration forms for the North American ISDN Users' Forum, contact: ISDN Workshop Series, Attn: Trudy Johnson, National Bureau of Standards, Building 225, Room A224, Gaithersburg, MD 20899, Telephone: (301) 975-2985.

- - - - - Paragraph 23352.p4 - - - - -
U.S. Organization for International Telegraph @ Telephone Consultative Committee (CCITT) Integrated Services Digital Network (ISDN) Joint Working Party and Study Group C; Meeting

- - - - - Paragraph 23352.p5 - - - - -
The Department of State announces that the Integrated Service Digital Network (ISDN) Joint Working Party, and Study Group C of the U.S. Organization for the International Telegraph and Telephone Consultative Committee (CCITT) will meet on March 15, 1989 at 9:30 a.m. in Room 1205, Department of State, 2201 C Street NW., Washington, DC.

- - - - - Paragraph 31262.p5 - - - - -
The Department of State announces that the Integrated Service Digital Network (ISDN) Joint Working Party, and Study Group C of the U.S. Organization for the International Telegraph and Telephone Consultative Committee (CCITT) will meet on March 15, 1989 at 9:30 a.m. in Room 1205, Department of State, 2201 C Street NW., Washington, DC.

- - - - - Paragraph 31262.p6 - - - - -
Summary: The National Computer Systems Laboratory (NCSL) at the National Institute of Standards and Technology (NIST) announces the Sixth North American ISDN Users' Forum (NIU-FORUM). The NIU-FORUM will be sponsored and held in conjunction with The Association of Data Communications Users (ADCU) National Conference. Issues related to the use and implementation of Integrated Services Digital Network (ISDN) technology will be discussed. The NIU-FORUM was formed in 1988 under the auspices of NIST to create a strong user voice in the implementation of ISDN and to ensure that the emerging ISDN services meet users' application needs.

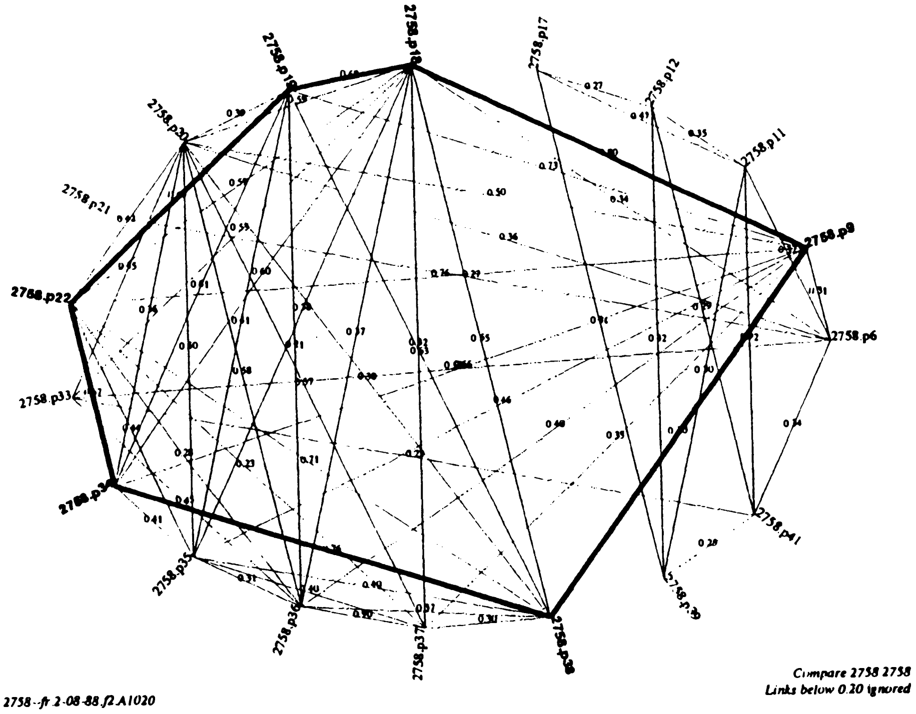
- - - - - Paragraph 31262.p7 - - - - -
Dates: The Sixth North American ISDN Users' Forum (NIU-FORUM) will be held at The Boston Marriott Copley Place, Boston, Massachusetts, June 14-16, 1989. An ISDN tutorial will be conducted for the Users the afternoon of June 13. This FORUM will consist of joint workshops for the Users (IUW) and Implementors (IIW). The IUW will continue work identifying, defining, and prioritizing user applications of ISDN. The IIW will continue defining implementation agreements for ISDN and will sponsor multivendor demonstrations and trials. Manufacturuers and service providers are invited to participate in this workshop.

- - - - - Paragraph 31262.p8 - - - - -
Address: To obtain registration forms for the workshops, companies may contact: ISDN Workshops, Attn: Kim Brashears, National Institute of Standards and Technology, Building 223, Room B364, Gaithersburg, MD 20899, Telephone: (301) 975-4853.

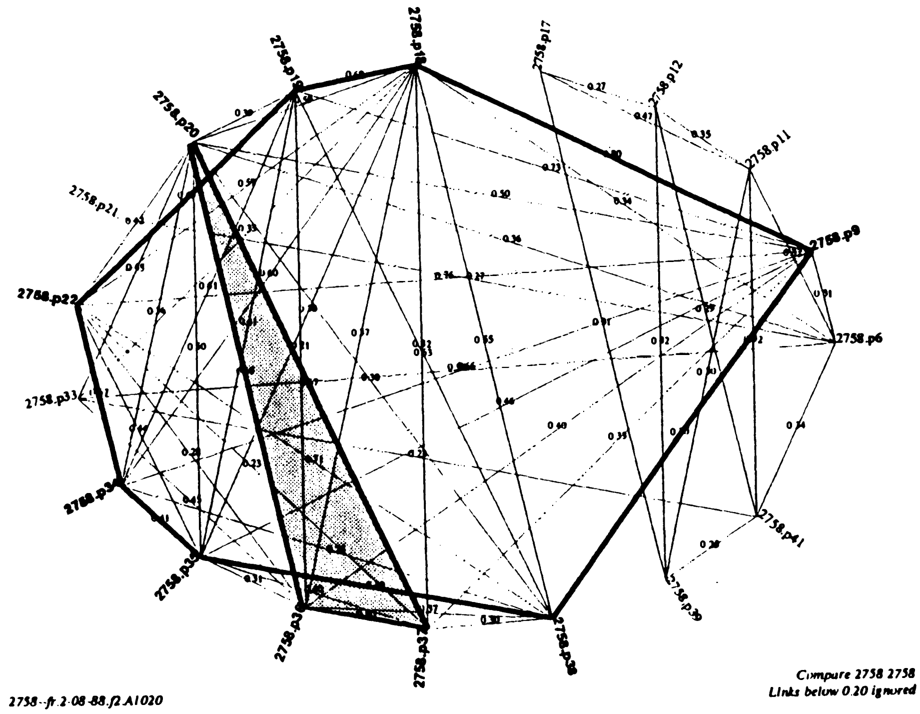
Table 7. Typical Long Multi-Document (20%) Summary (Documents 9281-23352-31262 covering meeting announcements)

Maps	
Number of single document maps	17
Number of multiple document maps	10
Number of documents included in maps	43
Themes	
Number of themes included in text	76
Number of correct themes	76
Summaries and Tours	
Number of 20% Tours	27
Subject coverage of summaries (27 excellent)	100%
Coherence of summaries (19 satisfactory)	70%
Comprehensiveness of summaries (17 satisfactory)	63%

Table 8. Evaluation of Theme Generation and Text Summarization

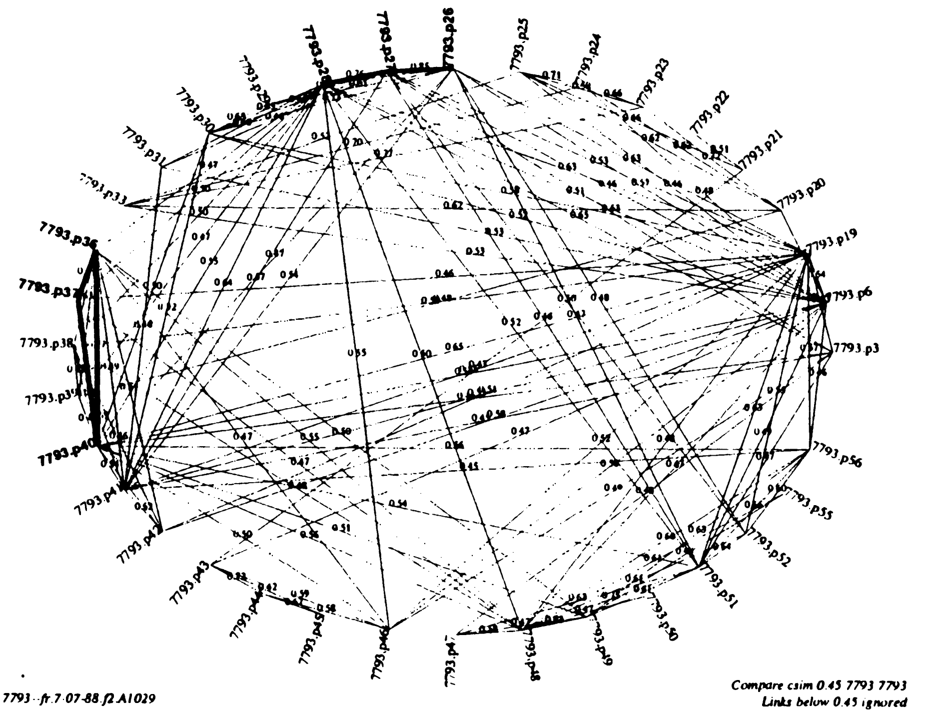


a) Medium Size (13%) Tour - Single Theme

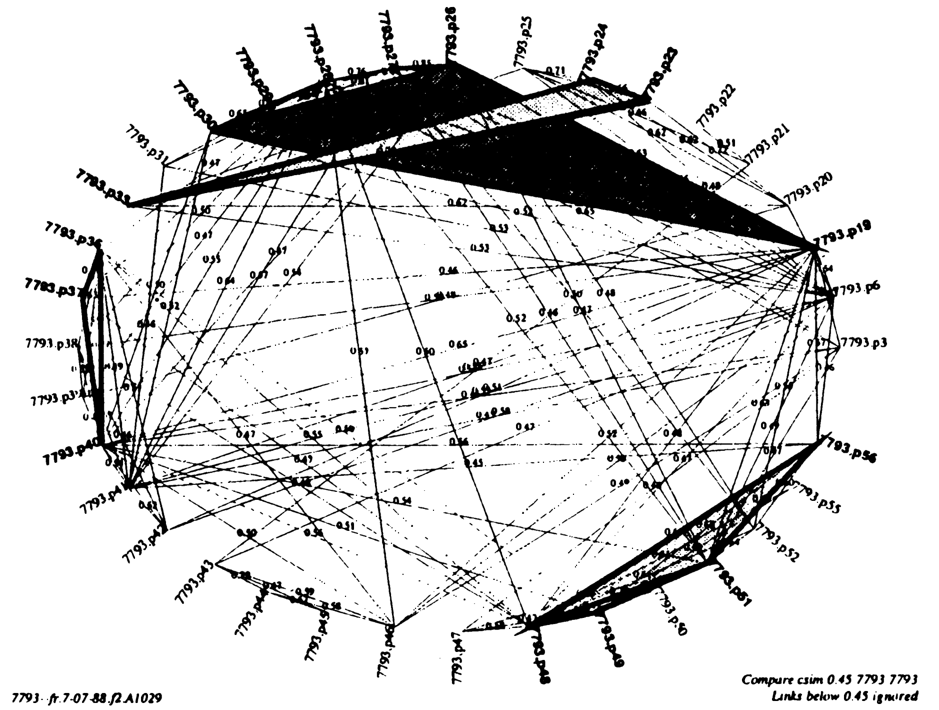


b) Long (21%) Tour - Two Themes

Figure 1. Tours and Themes for Homogenous Text (document 2758)

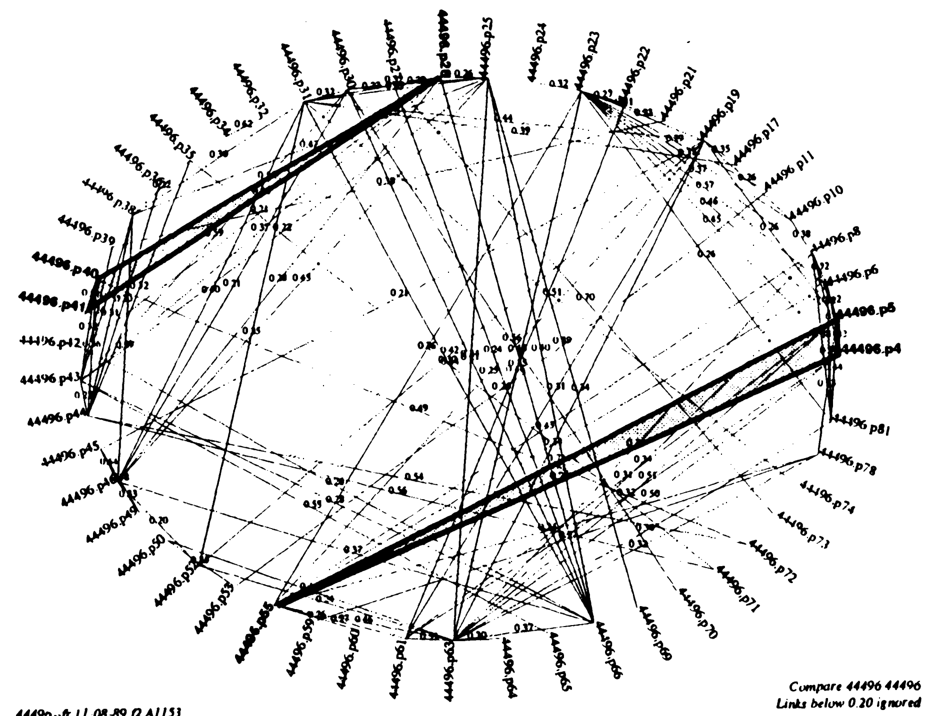


a) Short (9%) Tour - Two Distinct Themes Shown
(two themes at tsim = 0.50, single theme tsim = 0.40)

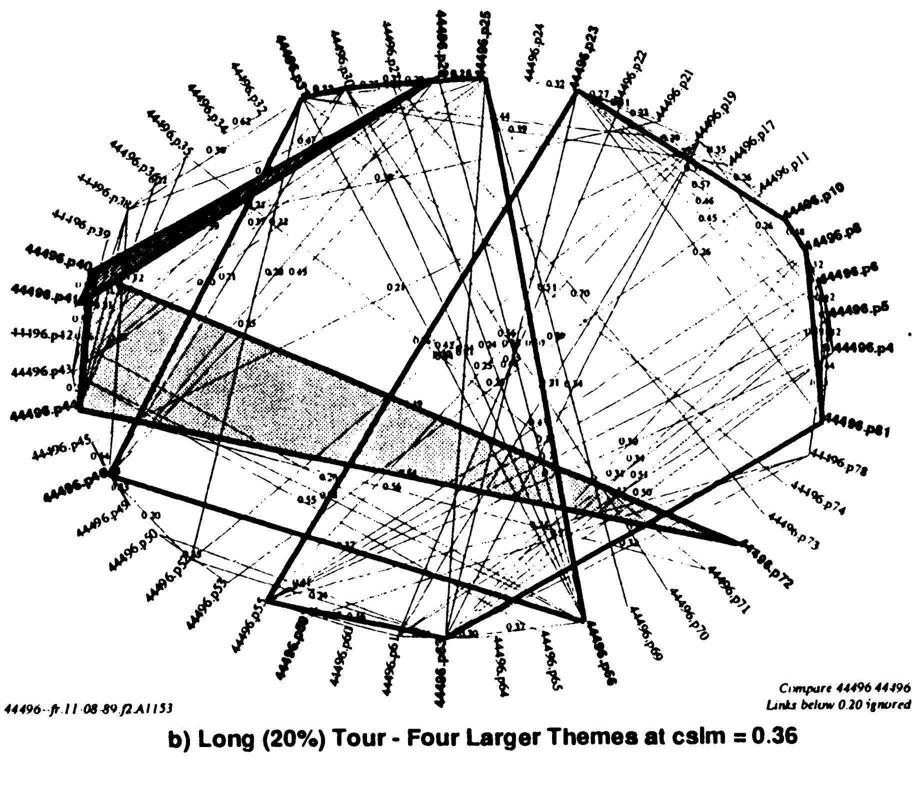


b) Long (25%) Tour - Four Themes at sim = 0.50
(two bottom themes merge at tsim = 0.40)

Figure 2. Tours and Themes for Text with Partially Disconnected Elements (document 7793) - Effect of tsim Parameter



a) Short (6%) Tour - Two Small Themes at csim = 0.51



b) Long (20%) Tour - Four Larger Themes at csim = 0.36

Figure 3. Tours and Themes for Poorly Connected Text (document 44496) - Effect of csim Parameter

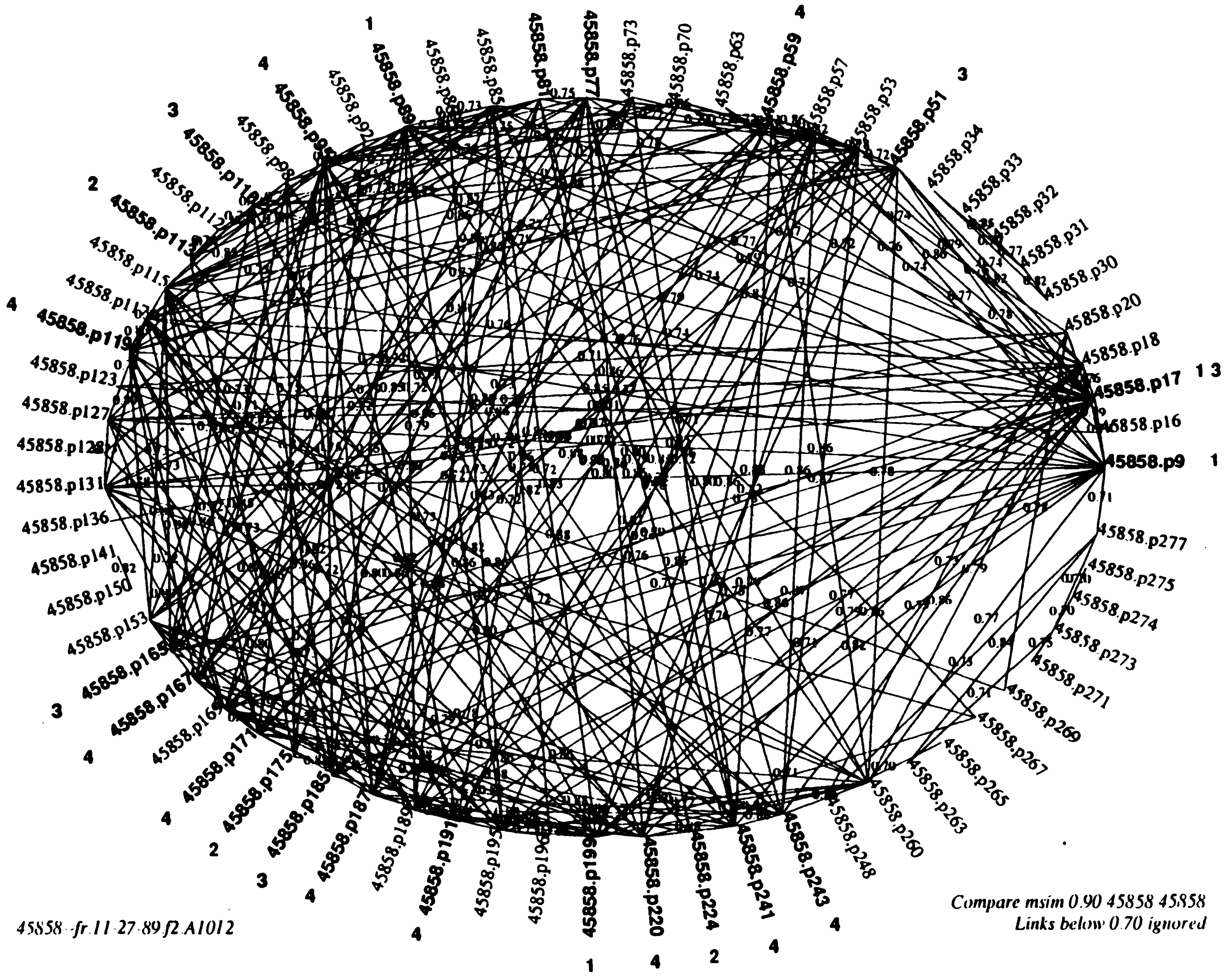
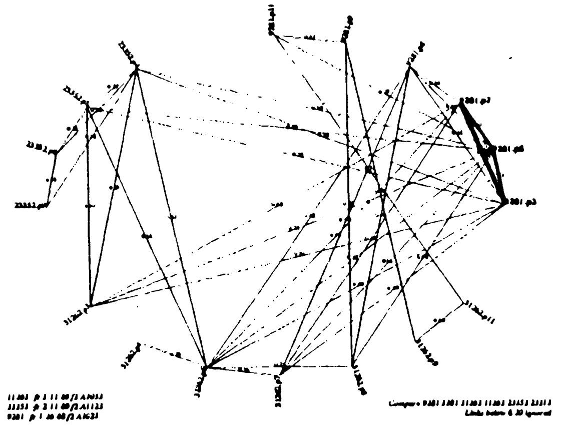
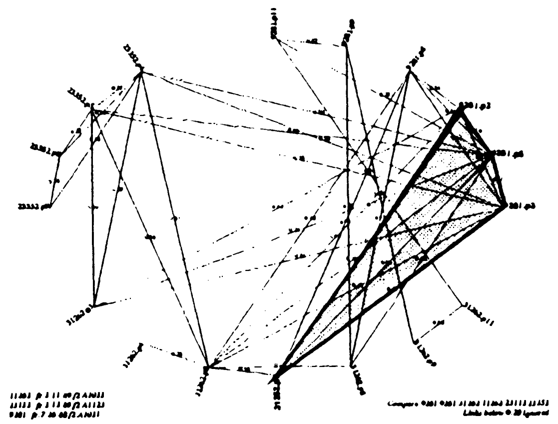


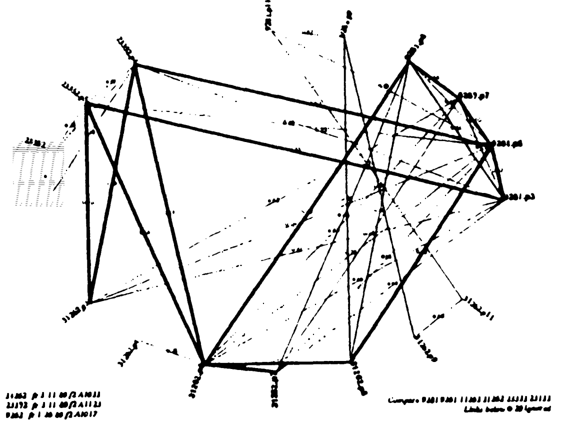
Figure 4. Short (7%) Tour and Four Themes (labeled 1-4)
(csim 0.79, tsim 0.80, msim 0.90, vsize 10)



a) Short 6% Tour



b) Medium Size (8%) Tour



c) Long (20%) Tour

Figure 5. Three Tours and Single Theme for Three Parallel Texts (documents 9281, 23352, 31262)