Approaches to Global Text Analysis

Gerard Salton*
Chris Buckley*

TR 90-1113
April 1990

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

*Department of Computer Science, Cornell University, Ithaca, NY 14853. This study was supported in part by the National Science Foundation under grant IRI-87-02735.

Approaches to Global Text Analysis

Gerard Salton and Chris Buckley*

April 12, 1990

Abstract

The current approaches to the analysis of natural language text are not viable for documents of unrestricted scope. A global text analysis system is proposed designed to identify homogeneous text environments in which the meaning of text words and phrases remains unambiguous, and useful term relationships may be automatically determined. The proposed methods include document clustering methods, as well as comparisons of local document excerpts in specified global contexts, leading to structured text representations in which similar texts, or text excerpts, are appropriately linked.

1 Language Understanding

A conventional view of language structure, that has become popular with the increasing use of computers for language processing, maintains that word meanings are reasonably standardized, and that well-defined structural and semantic rules control language generation and analysis. Under such circumstances, substantial resources must be devoted to the design of large word or phrase dictionaries specifying the syntactic as well as the semantic properties of text units, and to the construction of rule systems capable of explicating sentence structure and language use in various contexts.

A somewhat different view of language characteristics holds that the dictionary and rule-based approaches remain inadequate in most cases, because the meaning of words is often elusive, and the interpretation of language depends on context, circumstances, and background of writers and readers. Pitkin expresses such feelings in the following terms:

> "Meaning is compounded out of cases of a word's use, and what characterizes those cases is often the speech situation, not the pres-

*Department of Computer Science, Cornell University, Ithaca, NY 14853. This study was supported in part by the National Science Foundation under grant IRI 87-02735.

> ence of something referred to. As a consequence, the significance for meaning of situation, of circumstances, of context, is much greater than we might suppose." [1, p.71]

Such a pragmatic view of language structure leads to the realization that even the most complete dictionaries and the best linguistic rule systems will fail because individual words and expressions carry too many different meanings to be easily characterized by simple dictionary entries, and because the rules explaining word use and sentence formation are too complex to be formulated in an operational way.

Recognizing these difficulties, many artificial intelligence experts have recommended that automatic text understanding systems be based on the use of elaborate knowledge bases, instead of simple dictionaries, in which an attempt is made to represent much of the complexity of the environment in particular subject areas. A knowledge base normally exhibits the main concepts of interest in a given circumstance, and describes the use of these concepts and the applicable concept relationships.[2-4] To understand the meaning and function of a text, the text units must then be related to the contexts of this preconstructed knowledge base.

While knowledge constructs, such as frames and scripts, provide a richer linguistic framework than dictionaries and word lists, it is clear that no fixed set of knowledge bases can provide the infinite number of functions required for language interpretation when arbitrary documents in arbitrary subject areas must be processed.

The following quotation by Weizenbaum makes this point very clearly:

> "it is hard to see .... how Schank's schemes could possibly understand (the sentence "will you come to dinner with me this evening?") to mean a shy young man's desperate longing for love." [5, p. 200]

More generally, one must expect that a formal analysis procedure based on available knowledge structures automatically restricts the application to specific cases that have been specially provided, while leaving aside many other possibilities that had not been foreseen when the knowledge structure was built. Winograd and Flores describe the phenomenon of selective "blindness" as follows:

> "A program is forever limited to working within the world determined by the programmer's explicit articulation of possible objects, properties, and relations among them. It therefore embodies the blindness that goes with them" [6, p.97]

The notion of blindness implies that in an open-ended text environment, a computer that lacks the sophisticated background of human beings, is unlikely to operate satisfactorily in many unforeseeable situations. Blindness does not,

however, prevent a program from reaching interesting, and valid interpretations when the environment is suitably restricted, and the variety of objects and relationships of interest is limited.

2 Automatic Text Analysis

When the artificial constructs needed for text analysis are not available, an alternative approach may consist in analyzing large bodies of text, and attempting to identify term meaning and term relation factors by studying the contexts in which individual terms and expressions occur. In so doing, one can make the following not unreasonable assumptions:

- Much of what is written today can be captured in machine-readable form and made accessible to analysis by computer.
- By extension, much of what is read today is similarly available for computer analysis.
- The author of a text necessarily includes a great deal of human background in a piece of writing; this background information may be implicit, but by analyzing large amounts of text, one may hope that some elements of the implicit background may be automatically determined.
- A writer may also include elements of a reader model in the writing; in most cases, the formulations may be tailored to a particular intended readership, and elements of such a reader model may then be recoverable in an automatic text analysis systems.

It is then possible to ask whether an automatic text-based analysis system that uses enough context, and discovers enough background information, might not after all furnish reasonable interpretations of statements such as "will you come to dinner with me this evening?".

In an information retrieval setting, a text analysis system is designed to distinguish various texts from each other, leading to the retrieval of useful texts on demand, and to the rejection of materials that appear extraneous. The following components appear important in a text analysis system designed for information retrieval:

- A system that identifies or constructs text units used for the representation of text content; various types of content identifying units may be of interest, including text words, word stems, phrases, thesaurus entries, and so on.
- The determination of a local context that constitutes the individual retrieval units; depending on the circumstances, the local context can be

defined as individually retrievable text sentences, or text paragraphs, or document abstracts, sections, chapters, or even complete books.

- The determination of a global context within which the individual retrieval units are embedded; the global context is made up of collections of local items.
- A system that assigns term importance indicators, or weights, to the text identifying units, based on the term importance in the local and global retrieval environments.

A summary of the various retrieval contexts appears in Table 1.

When the local retrieval units consist of at least abstract-length text excerpts, a term weight that depends in part of the term frequency (tf) of a term in the local context, and in part on an inverse function of the collection frequency (or inverse document frequency, idf) has been shown to be useful for text retrieval purposes. [7,8] Such a function assigns high values to terms that occur frequently in individual local environments, and at the same time occur relatively rarely in the collection as a whole.

When the local retrieval units differ substantially in length, it is also useful to normalize the term weight by division by the document length. In this way, each local text unit is given an equal chance for retrieval.

A typical term weight, w_{ik}, for term k assigned to local document D_i is then given by

$$w_{ik} = \frac{tf_{ik} \cdot \log\ (N / n_k)}{\sqrt{\sum_{\substack{\text{all terms} \\ k \text{ in } D_i}} (tf_{ik} \cdot \log N / n_k)^2}} \qquad (1)$$

where tf_{ik} represents the frequency of term k in D_i, n_k is the number of local documents with term k assigned, $\mathcal{N}$ is the collection size, or total number of local documents.

It is clear that when the definition of the local and global text units used in retrieval changes, the corresponding term weights change. Thus, when a particular term is used to identify a set of text sentences to be retrieved from particular document texts, a quite different term weight is assigned than when the same term identifies complete documents embedded in a larger document collection.

Given two documents D_i and D_j, indexed by a set of t different weighted index terms $D_i = (w_{i1}, w_{i2}, ..., w_{it})$ and $D_j = (w_{j1}, w_{j2}, ..., w_{jt})$, a pairwise document similarity measure, sim $(D_i\ ,\ D_j\ )$, can be computed as the inner product of matching terms as follows: [9]

$$Sim(D_i, D_j) = \sum_{k=1}^{t} w_{ik} \cdot w_{jk} \qquad (2)$$

In some environments, the term weighting function of expression (1) is unnecessarily complicated. In particular, the length normalization factor in the denominator of (1) is not needed when the local documents are homogeneous and of equal length. To compare short texts, such as individual text sentences, it may be sufficient to weight the terms by the term frequency factor alone. In that case, two sentences may be represented as $S_i = (tf_{i1}, tf_{i2}, ..., tf_{it})$ and $S_j = (tf_{j1}, tf_{j2}, ..., tf_{jt})$, and the sentence similarity function becomes

$$Sim (S_i, S_j) = \sum_{k=1}^{t} tf_{ik} \cdot tf_{jk} \qquad (3)$$

In a retrieval environment, the assumption is that query texts are furnished by a user population, and that the same analysis system is used for both query and document texts. The document and sentence similarity measures of expressions (2) and (3) can then be used equally to obtain similarity indications between queries and stored texts.

In an automatic text processing enviroment, the local and global text environments can be freely chosen, depending on requirements. In some situations, complete books must be retrieved from a book collection. Alternatively, individual book paragraph may be treated as retrieval units, or individual text sentences. In hypertext applications, it may be useful to identify and retrieve chains of linked local documents – for example, chains of paragraphs, or chains of text sentences. In that case, the similarity measures described earlier are used to link similar local items with each other. [10]

In an environment where texts can be freely decomposed into local units of varying extent, and where local items are comparable within differing global environments, homogeneous text environments may be definable in which the meanings of the text units are stable and unambiguous, and where term relationships are straightforward and relatively easy to use. This question is discussed further in the remainder of this note.

3 Determination of Stable Text Environments

Various possibilities exist for the identification of homogeneous text environments in which text interpretation may be simplified. The most obvious consists in clustering the document collection by collecting into common classes items that appear sufficiently similar. Such grouped items may then in fact represent related subject matter and unambiguous language environments. Many

clustering strategies are possible; in information retrieval, a complete-link clustering system with its relatively stringent clustering criteria often produces a high order of retrieval effectiveness. [11-13]

Fig. 1 shows a typical upper-level cluster obtained by a complete-link clustering system for the 1137 paragraphs of the text of a sample textbook.[9] For that example, the text is subdivided into 71 upper-level clusters (clusters that cannot themselves be grouped into larger classes because the pairwise cluster similarity is equal to 0). The average cluster size of the upper-level clusters consists of 16 local documents (book paragraphs). The topic description of Table 2 shows that the sample cluster of Fig. 1 covers the general area of office automation. Of the 13 documents in the sample cluster, 12 are taken from chapter 3 of the text, where the automated office environment is discussed, and one document (number 1040) comes from chapter 13 covering automatic mail and message systems.

In a complete-link clustering system, the overall cluster similarity is defined as the smallest pairwise similarity between any pair of clustered items. The cluster similarity measures are entered at the roots of the various sub-trees in the illustration of Fig. 1. Thus the two documents 111 and 1040 shown at the lower left end of the Figure have a high pairwise similarity of 0.526. The similarity computations of Fig. 1 use the term weighting system of expression (1) and the similarity measure of expression (2). Because of the term weight normalization, the pairwise similarity coefficients appear in the range between 0 and 1. When item 176 is added to the previous 2-item cluster, the overall similarity drops to 0.425, indicating that in the 3-item group (111, 176, 1040), the smallest pairwise similarity is 0.425.

Because the most highly-matching document pairs are clustered first, and appear at the bottom of the cluster tree, the cluster similarities decrease as additional elements are added to the cluster. The overall cluster similarity for the complete sample cluster of 13 items is only 0.084. Such a small similarity level indicates that the subject relatedness for a reasonably-sized cluster is likely to be too low to be useful for text disambiguation. A tighter text environment must then be defined if stable text environments are wanted.

One possibility consists in taking sets of clustered items, and computing the similarity for each pair of clustered items (for example, between each pair of clustered text paragraphs, or each pair of sentences within clustered paragraphs). Pairs of items with a sufficiently high pairwise similarity can then be linked to signify subject relatedness, and the linked structures can define the desired homogeneous text environment. Using a pairwise similarity threshold of 0.35, the linked structure of Fig. 2 is obtained for the paragraphs in the sample cluster of Fig. 1.

In Fig. 2, a link is placed between two nodes, whenever the similarity between the corresponding paragraph pair exceeds 0.35. As the example shows, all clustered paragraphs are linked except for items 113, 115, and 188, whose similarity with the rest of the structure is too small. Fig. 2 also shows that

some links appear more important than others. The heavy links are used to relate matching paragraphs which contain at least one pair of matching sentences with a sentence similarity coefficient exceeds 5. (The similarity measure of expression (3) is used to measure sentence similarity). The paragraphs with the heavy links thus carry a pairwise global similarity of at least 0.35, and in addition contain a pair of locally linked sentences. In the example of Fig. 2, several sets of documents appear to be tightly linked – for example, the set (160, 170, 176) consists of pairs of matching paragraphs, each of which contains pairs of matching sentences.

The paragraph and sentence matching systems are illustrated in the examples of Figs. 3 and 4. Two typical paragraph texts are shown in Fig. 3. The overall paragraph similarity is 0.526 (see Fig. 1). The list of matching terms appearing on the left side of Table 3 shows that a great deal of overlap exists between the vocabularies of the two documents, although these paragraph appear far apart in the text of reference [9]. For example, there are 5 different matches of the term "office", and 4 of "process".

A decomposition into sentences after deletion of common function words, and elimination of word suffixes, is shown in Fig. 4 for the text of Fig. 3. Fig. 4 shows that the matching term list for sentences 11106 (sentence 06 of document 111) and 104001 includes "edit", "includ", offic", "process" and "task". In addition to the single term matches, the sentence matching system may also take into account matching phrases, consisting of adjacent pairs of matching single terms. "Task includ" represents such a phrase in the illustration of Fig. 4 and Table 3.

Table 4 shows the paragraph and sentence statistics for the sample textbook of reference [9]. For the 1137 paragraphs of text, about 2500 paragraph pairs exist with a pairwise similarity exceeding 0.350, and about 6,500 sentence pairs included in these matching paragraph pairs with a sentence similarity not smaller than 5.

The following sample process may be usable to define stable text environments in which word meanings are unambiguous and well-defined:

a) Clustering of the local documents into affinity groups, as shown for the sample database of book paragraphs in the illustration of Fig. 1. For the 1137 book paragraphs, this produces 71 high-level clusters using a complete-link clustering process.

b) Linking of clustered document pairs that exhibit a sufficiently high global document similarity. For the sample database, 1818 pairs a paragraphs located in a common cluster exhibit a global document similarity exceeding 0.350.

c) Optional elimination of linked paragraph pairs that do not include at least one pair of sentences with a sufficiently high sentence similarity. For the sample database, 344 of the 1818 paragraph pairs do not

include sentence pairs with a pairwise similarity exceeding 5.0. This leaves 1474 highly linked paragraph pairs with a global similarity exceeding 0.350 and a sentence similarity in excess of 5.0.

Table 5 contains an evaluation of the paragraph and sentence linking process for the sample textbook.[9] The table shows that the vast majority of the text links relate texts that are homogeneous in nature. For the global document linking, only 90 paragraph pairs out of 1818 are of a questionable nature. This shrinks to only 56 paragraph pairs out of 1474 when the additional sentence links are taken into account.

A sample incorrectly linked text pair is excerpted in Fig. 5, consisting of document pairs 168 and 325. These documents cover serializability and locking protocols of the kind used in database security on the one hand, and cryptographic transformations on the other. These topics are somewhat related, but not identical, and the corresponding text link may be considered questionable. Examples of the kind shown in Fig. 5 are rare for the database under study.

It remains to be seen whether stable, well-defined text environments can also be generated by global text-based analysis systems for more heterogeneous text materials than the sample database used in this study.

REFERENCES

1. H.F. Pitkin, *Wittgenstein and Justice*, University of California Press, Berkeley CA, 1972.

2. M. Minsky, A Framework for Representing Knowledge, in the *Psychology of Computer Vision*, P. Winston, editor, McGraw Hill Book Co., New York 1975.

3. R.C. Schank and R.P. Abelson, *Scripts, Plans, Goals and Understanding*, Lawrence Erlbaum Associates, Hillsdale, NY 1977.

4. R.J. Brachman, *The Epistemological Status of Semantic Networks*, Academic Press, New York, 1979.

5. J. Weizenbaum, *Computer Power and Human Reason*, W.H. Freeman and Company, San Francisso, 1976.

6. T. Winograd and F. Flores, *Understanding Computers and Cognition*, Ablex Publishing Corporation, Norwood, NY 1986.

7. G. Salton and C.S. Yang, On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, 29:4, December 1973, 351-372.

8. K. Sparck Jones, A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, 28:1, March 1972, 11-21.

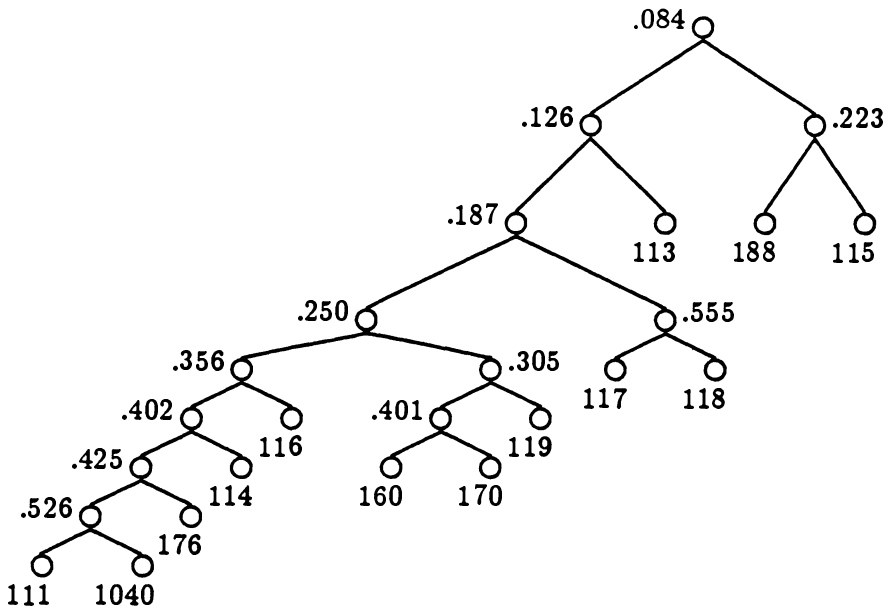
9. G. Salton, *Automatic Text Processing*, Addison Wesley Publishing Co., Reading, MA, 1989.

10. G. Salton and C. Buckley, Approaches to Text Retrieval for Structured Documents, Tech. Report 90-1083, Computer Science Department, Cornell University, Ithaca, NY, January 1990.

11. N. Jardine and C.J. van Rijsbergen, The Use of Hierarchic Clustering in Information Retrieval, *Information Storage and Retrieval*, 7:5, December 1971, 217-240.

12. G. Salton and A. Wong, Generation and Search of Clustered Files, *ACM Transactions on Database Systems*, 3:4, December 1978, 321-346.

13. F. Murtagh, A Survey of Recent Advances in Hierarchical Clustering Algorithms, *The Computer Journal*, 26:4, 1982, 354-360.

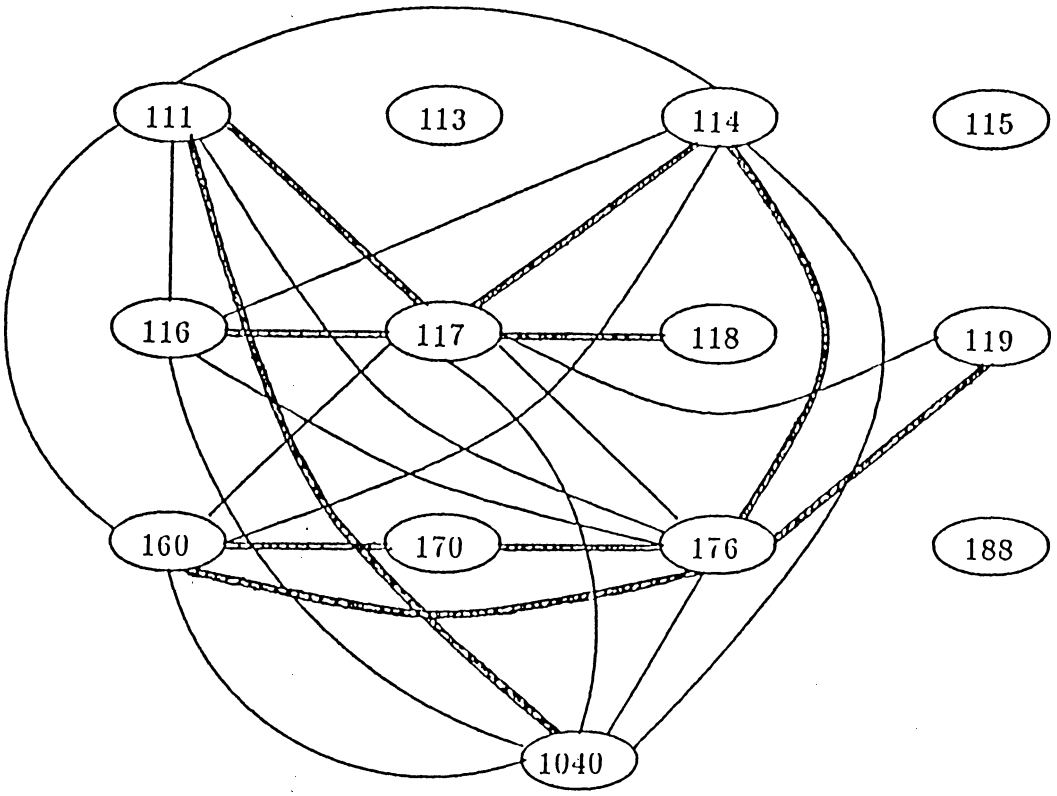


Complete Link Cluster

(overall cluster similarity is 0.084)

Cluster Elements (Paragraphs)	Chapter in Book	Principal Topic Area
111-115	3	automated office
160	3	data security, integrity
170	3	office display systems
176	3	office rehearsal
188	3	word processing approaches
1040	13	office mail handling

Figure 1



Paragraphs and Sentence Links in Sample Cluster

Single link is paragraph similarity
(threshold .350)

Double link is sentence similarity
within similar paragraph pairs
(threshold 5.00)

Figure 2

.I 111
Chapter 3 The Automated Office
The Office Environment
The new computing environment has substantially changed the way in which many information-processing problems are solved. In particular, hands-on computing now makes it possible for users to directly control complex operations in a dynamic environment. As a result, new approaches are being taken not only in many scientific and technological applications, but also in most commercial and data-processing tasks. The conventional business office provides a particularly good example of the changed information processing situation – an office is a complex environment for processing many diverse objects, involving interactions and communications among different classes of participants. **In the modern office, all the well-known processing tasks, including dictating, typing, editing, mailing, and filing, have been eliminated, or at least substantially altered, by new systems and procedures.**

a) Document Text 111 (Chapter 3)

.I 1040
The modern electronic office constitutes a principal application for electronic mail services. **In an office, the main tasks include creating and editing office documents, storing and retrieving items, sending and receiving documents, and handling various form-based office applications such as billing, inventory processing, and sales processing.** The mail system can serve as the principal communications path for office personnel. [56] Each office can be served by an office processor that controls the user mailboxes, and also provides access to the communications lines used to contact local or remote offices. Individual workstations may be attached to the local office processors, or can be attached directly to the communications lines. A typical system configuration of this kind is shown in Fig. 13.16.

b) Document Text 1040 (Chapter 13)

Document Texts 111-1040 (Global Similarity 0.526)

Figure 3

11100	chapt automat offic
11101	offic environ
11102	comput environ subst chang inform process problem solut
11103	hand comput make use direct control complex operat dynam environ
11104	result approach scientif techn applic commerc data process task
11105	convent busi offic good chang inform process situat offic complex environ process divers object involut interact commun class particip
11106	modern **offic know process task includ** dictat typ **edit** mail file elimin subst alter system procedur

a) Document 111 (Sentence Indexing)

104000	modern electron offic constitut princip applic electron mail servic
104001	**offic** main **task includ** creat **edit** offic docu stor retrief item send receiv docu handl form base offic applic bill invent **process** sale process
104002	mail system serv princip commun path offic personnel
104003	offic serv offic process control use mailbox access commun line contact local remot offic
104004	individual workst attach local offic process attach direct commun line
14005	typic system configur kind shown fig

b) Document 1040 (Sentence Indexing)

Sentence Pairs 111-06 — 1040-01 (Sentence Similarity 6)

Figure 4

Serializability is obtained by using locking protocols that force users to lock files before updating them. Locked files are generally accessible only to users who actually hold the file lock. Several file locking protocols have been developed.

a) Excerpt from Document 168 (Chapter 3)

Cryptographic transformations are related to compression because they attempt to even out the occurrence characteristics of the components of the encrypted text, making unauthorized decryption more difficult.....
The basic aim is to safeguard the confidentiality of the data.....
In the case of cryptography, the reverse transformation cannot be executed without access to the normally secret key information, just as a particular combination lock cannot be opened without knowing the combination for that lock.

b) Excerpt from Document 325 (Chapter 6)

Incorrectly Linked Document Pair

Figure 5

Context	Description
individual content identifiers	individual words, word stems, thesaurus entries, word phrases
individual retrieval units (local contexts)	sentences, paragraphs, text sections, book chapters, individual documents
retrieval collections (global contexts)	complete paragraphs (retrieval of individual sentences), text sections, book chapters, collections of books

Retrieval Contexts

Table 1

Cluster Elements (Paragraphs)	Chapter in Book	Principal Topic Area
111-119	3	automated office
160	3	data security, integrity
170	3	office display systems
176	3	office retrieval
188	3	word processing approaches
1040	13	office mail handling

Topic Description for Cluster Elements of Fig. 1

Table 2

Matches between Documents 111 - 1040		Matches in Sentences 11106 - 104001
applic	mail	edit
commun	moden	includ
control	office (5)	office
[illegible]t	process (4)	process
edit	task (2)	task
includ	task includ	task includ

Matching Terms for Documents 111 - 1040

Table 3

Number of local documents (paragraphs)	1,137
Number of sentences	~ 4,500
High-level complete-link document clusters	71
Distinct paragraph pairs	~ 650,000
Distinct paragraph pairs with similarity threshold > 0.350	~ 2,500
Distinct sentence pairs	~ 10,125,000
Distinct sentence pairs with s[illegible]arity threshold > 5	~ 6,500

Paragraph and Sentence Statistics for Sample Textbook [9]

Table 4