Selective Use of Full-Text Databases

Gerard Salton*
J. Allan
C. Buckley

TR 92-1300
August 1992

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 89-15847.

Selective Use of Full-Text Databases

Gerard Salton*, J. Allan, and C. Buckley

Abstract

Large files of natural-language text are now available for automatic processing in machine readable form. Such text files may include documents of textbook size, medium-size newspaper articles, and short message and mail items, and the subject matter may be effectively unrestricted. Typically, the stored material is not meant to be read sequentially from beginning to end. Instead a selective, diagonal reading strategy may be preferred which skips among the text sections and paragraphs in accordance with individual user needs.

Methods are described in this study for analyzing text files covering arbitrary subject matter, constructing links among text segments of varying size in accordance with computed similarities between texts, and defining text traversal paths that are responsive to particular user needs. Such selective text traversal is useful in retrieving information from textbooks and instruction manuals, and in consulting dictionaries, encyclopedias, and other collections of text items.

Topics: Hypertext construction, automatic text linking, selective text utilization, diagonal text traversal.

*Department of Computer Science, Cornell University, Ithaca, NY 14853-7501. This study was supported in part by the National Science Foundation under grant IRI 89-15847.

1 Full Text Databases

Large masses of natural-language text are now available in machine-readable form. Such files may include the full text of dictionaries and encyclopedias, literary works, electronic mail and messages, textbooks, newspaper articles, and many other text items. Increasingly text databases are recorded on optical disks, and such disks are made available for electronic searching at little cost to a wide class of potential users. When large text files, often consisting of hundreds of megabytes of data, are recorded for use, selective access must be provided to the text items depending on text content and other user requirements. Because many text items are large, consisting not infrequently of hundreds of pages of text, the stored documents may need to be broken up by defining smaller text units to which access is separately provided on demand. This is true notably of textbooks and reference works that are not normally meant to be read sequentially from one end to the other. In such circumstances, it may be useful directly to access particular sections or paragraphs of text in accordance with stated user interests and requirements.

By characterizing individually the content of text sections and text paragraphs, new strategies of text utilization can be devised. Thus, text passages dealing with similar subject matter can be identified and appropriately linked in the text structure. A linked set of text passages then constitutes a selective, or diagonal, reading path which allows the reader to jump from a particular text segment to one or more related segments. Selectively chosen text segments may be more responsive to individual user needs than the full-text items normally processed in standard retrieval situations.

In the present study, methods are described for

- structuring text files by identifying related text excerpts that deal with similar subject matter, and linking such text excerpts with each other;
- identifying semantic text segments of appropriate scope that can function as homogeneous text units suitable for text linking and retrieval;
- providing access or reading paths that allow the reader to traverse the texts by following particular links between related text pieces in response to statements of user interests.

2 Analysis of Text Content

Before text excerpts of interest to information users can be properly identified, it is necessary to characterize the text content. The classical methods for tackling this problem are not available when large text files are processed in unrestricted topic areas. In particular, in an open subject environment it is not possible to build knowledge bases that identify all the main semantic entities of interest as well as all the main types of relationships between entities. Similarly, it is difficult in open topic areas to build thesauruses that specify synonym relationships between terms, or hierarchical term inclusion relations. Finally, the large comprehensive linguistic analysis program for syntactic and semantic text analysis do not normally prove effective when the topic coverage is unrestricted.

A more productive approach to the text analysis problem, applicable to large unrestricted text files, consists in using corpus-based methods where the characteristics of the available texts are used to derive appropriate representations of text content.[1,2] In particular, when the full text of large document collections is available for processing, the vocabulary environment can be studied and appropriate information concerning word utilization and text make-up can be derived. The "use theory" of text proposed by Wittgenstein and others can be invoked: this states that the meaning of a linguistic expression is a function of the manner and situation in which the expression is used by speakers and writers of the language. [3]

The question arises about how the use of words and expressions occurring in natural-language text can be properly characterized. One notes in this connection that a complex interaction exists between the meaning of individual text components and the meaning of the full text that contains them. Figure 1 gives a scheme for the derivation of text meaning: by assembling a number of ambiguous lower-level components — for example, a set of text words taken out of context — one obtains a disambiguated higher-level structure — for example, a sentence containing these words. Once the meaning of the sentence is at hand, the function of the individual text components becomes clearer. The same kind of interaction takes place between the meaning of individual component sentences and of the paragraph containing them, and between the meaning of individual paragraphs and that of the text section containing these paragraphs. These multiple interactions between higher- and lower-level structures are characterized in Figure 1.

Text meaning thus depends only to a minor extent on the meaning of individual text words. In addition it is necessary to take into account the complex linguistic context in which words are embedded, the pragmatic context (outside knowledge, social environment, and associated material) in which the text is placed, and finally the particular role performed by the texts in question. If the emphasis is placed on the occurrences of individual text words in large texts, as is done in many of the keyword-type information retrieval systems where texts are identified by sets of disconnected keywords, the retrieval performance may be mediocre because of the ambiguities inherent in conventional keyword systems. The solution consists in taking into account the context needed to characterize text meaning more precisely.

An approach to the content analysis problem for full-text databases presents itself if one remembers that in text retrieval and text linking applications it is not normally necessary to specify the exact meaning of individual text words and expressions. Instead, it is generally sufficient to find out whether the meaning of the vocabulary in two different text excerpts, or in user queries and text items considered for retrieval, is congruent or not. If the vocabulary meanings are congruent, the corresponding texts should be linked to make possible joint retrieval in response to an appropriate user query. Alternatively, a stored text should be retrieved in response to a user query when the corresponding texts carry related meanings.

Concretely, the assumption can be made that when the common vocabulary in two different texts (or in query and document texts) occurs in similar local contexts, then text meaning is congruent and text linking or text retrieval are in order. On the other hand, if the common vocabulary occurs in different local contexts, the word meanings are likely to be distinct and the texts will be unrelated. For example, when a term such as "bank" occurs in a local environment with "finance", "money", etc., in some text, while the context is "river", "water", etc. in another text, the texts are likely to be distinct, even when substantial vocabulary overlap exists between the texts.

Assuming that query formulations are available as natural-language statements of user need, the following two-step process is then suggested as a realistic way for determining congruence of meaning for different text excerpts: [4,5]:

1. Two text excerpts, or alternatively a query text and the text of a stored document, are relatable when there is a sufficient global overlap between the vocabulary of the corresponding texts. This vocabulary overlap is measurable by global text comparison operations that take into account the proportion of matching vocabulary terms, as well as the weights of the matching terms.

2. In addition, local text similarities are measured by comparing local text constructs — for example, text sentences and text phrases — occurring in the respective texts. When sufficiently large local text similarities are found in the form of matching local constructs, such as text sentences, then the vocabulary is assumed to be used in a similar sense, and the texts are relatable.

Note that little of importance can be concluded when only the local similarities of step 2 above are present, but the global text similarity does not exist, because the local similarities may then be due to the fortuitous matching of short text fragments that are often not indicative of text meaning ("consider, for example, the illustrations of Figure X"). At the same time, the existence of global text similarities (step 1) without corresponding local relationships may be due to common terms used in different environments with distinct meanings.

The actual method of computing text similarities must depend on the text size. To obtain the global text similarities (step 1 above), a measure is preferred that increases as the proportion of matching terms in the two texts increases. The local text similarities (step 2 above) on the other hand are made to depend on the number (rather than the proportion) of matching terms, because for short texts the proportion of matching terms could be high even when only few term matches are in evidence. Concretely, a standard automatic indexing package is used to reduce each document and query text into vector form, such as $D_i = (w_{i1}, w_{i2}, \ldots, w_{it})$, where w_{ik} is the weight of term T_k assigned to document D_i. [6]. The term weights are so chosen that the weights will be high when the term frequency inside a particular document is large but the overall collection frequency for the term is small. This insures that high weights are assigned to terms that are concentrated in only a small number of documents in a collection.

When normalized terms weights are used in the term vectors (that is when a length normalization factor is used with the basic term weights w_{ik} of the form $w_{ik}/\sqrt{\sum_{vector}(w_{ip})^2}$ then a standard inner product vector similarity function, $\text{sim}(D_i D_j) = \sum_{k=1}^{t} w_{ik} w_{jk}$ produces similarity coefficients between 0 and 1 that depend on the proportion and the weight of matching terms in the two texts D_i and D_j. When the term weights are not normalized for vector length, the vector similarity is normally larger than 1 and depends on the number and weight of matching terms. [4-6]

3 Automatic Text Linking and Retrieval

When choosing a strategy for retrieving text excerpts in response to user interest statements, or specifying a text traversal strategy, a decision must be made about the type of text unit used for retrieval or text linking purposes. Ideally, a retrieved text segment should be a self-contained semantic unit that tells a recognizable story, or represents a rounded argument or discussion. When a number of such semantic units are linked together, the whole set would then form a complete story, report, discussion, or argment. In these circumstances, individual text words are not ideal retrieval units because of the inherent ambiguities, and the lack of necessary context. Individual text sentences are also normally inadequate, because sentences are often fragmentary entities and may not represent complete thoughts. (For example, sentences may contain pronouns and other anaphoric expressions that are interpretable only by understanding the surrounding context.) On the other hand, full text items are often long and unfocused, often covering several hundred pages of text, and such texts may deal with many different topics.

This suggests that a mixed retrieval, or text traversal, strategy be used which extracts text excerpts consisting of some full documents, as well as individual text sections and paragraphs that are identifiable as complete semantic units. In each case, a shorter text identified as a self-contained semantic unit by the global and local text comparison system is preferred over a corresponding longer segment. The following text linking and retrieval strategy may be used:

1. Choose as an anchor, or base document an available natural-language query formulation, or the text of a document known to be relevant to a user's information need.
2. Use this base document as a query and perform a standard retrieval operation that displays the output documents in decreasing order of the global similarity with the user query. [6]
3. Consider the top 100 retrieval items, and discard any retrieved document that does not exhibit at least one local sentence match between query text and retrieved document

text. (Two sentences are considered to match when their local vector similarity reaches at least 75.0 and there are at least two matching terms).

4. Break up the remaining retrieved document texts into text sections and text paragraphs and construct a term vector for each component text.

5. Discard any text component whose global similarity with the query text is smaller than the similarity with the complete document. (This insures that shorter text excerpts will be condensed for retrieval only when they appear to be more similar to the query than the corresponding full text).

6. Of the remaining text components (sections and paragraphs) corresponding to a particular retrieved document, choose that excerpt whose global query similarity is maximum, while also exhibiting a valid local sentence-pair match with the query (see step 3).

7. If the global query similarity between the text excerpt identified in step 6 exceeds 0.20, replace the full document text on the output list by the corresponding section or paragraph output.

8. Rank the remaining set of full documents, document sections, and document paragraphs in decreasing order of the computed global query similarity, and retrieve an array consisting of the top n items (alternatively retrieve all items whose global similarity with the query exceeds 0.20).

An illustration of the multi-step (global and local) text comparison system is provided in Figure 2. A typical search conducted in the 29-volume Funk and Wagnalls New Encyclopedia is used as an example. This encyclopedia contains more than 25,000 articles ranging in length from 1 line to over 150 pages of text. Individual encyclopedia articles are used as search queries, and text excerpts from semantically related articles are obtained as retrieval output. The encyclopedia articles are assigned a serial number for identification. Suffixes c_i, p_j, and s_k refer to section i, paragraph j, and sentence k of the corresponding article.

The text of a typical query article ([9573]) Gallieni) is shown in Figure 2(a). This article consists of a single text section not further broken down into paragraphs (a section is a text segment appearing between adjacent text heads). A typical article retrieved among the top 100 is [15074] entitled "Battle of the Marne". The article consists of three sections (labeled [15074.c3] to [15074.c5]) without further break-down into paragraphs. The text and vector similarities as well as the maximum sentence similarities between query and document excerpts are shown in Figure 2(b). In this case, two sections have a higher global query similarity than the full article (sections [15074.c3] and [15074.c4]). However, the first of these does not contain any sentence with the required sentence similarity of at least 75.0 with the query. Hence, sections 3 and 5 are discarded, and section 4 is used for retrieval purposes in this case. The text of section [15074.c4], entitled, "Battle of the Marne/First Battle of the Marne" is shown in Figure 2(c).

The common vocabulary between query text [9573] and the retrieved section text is shown in Figure 2(d). The columns in that figure represent, respectively, an assigned term number (con), the term weight in document [9573] (Vec1), the weight in document [15074.c4] (Vec 2), the product of the term weights in the two vectors (product), and finally the actual matching word stem occurring in the two texts.

The output of Figure 2(d) confirms that a substantial overlap exists between the vocabularies of the two documents. The highest sentence-pair similarity for the two documents is 224.45. The two corresponding sentences are reproduced in Figure 2(e), and the matching sentence vocabulary is listed in Figure 2(f). Figure 2 shows that the matching sentences are semantically very close, and the retrieved text vector [15074.c4] is correspondingly closely related to the text of article [9573].

4 Experimental Output

Figure 3 contains sample output for a number of test queries. In each case, the title of the query is given with the article serial number, followed by the list of retrieved items in decreasing query similarity order. The retrieval threshold for the output of Figure 3 is set at 0.20, and up to 15 retrieved items appear on the output list for each query. An examination of the output obtained for 100 queries indicates that approximately 8 items are retrieved for each query using the retrieval strategy outlined in the previous section. Approximately 17 percent of the output items consist of full articles, 32 percent are article sections, and the remaining 52 percent are text paragraphs. The output of Figure 3 shows that few problems arise in resolving the normal linguistic ambiguities when the multi-step text comparison system is used. For example, the distinction between Galicia 1 (article [9557]), the former Spanish province, and Galicia 2 (article [9558]), the region in Poland, was made flawlessly as the output demonstrates.

The retrieved output for Galileo [9561] is reproduced in more detail in Figure 4. A full section and/or paragraph title is given for all retrieved text excerpts that are not full articles. For example, paragraph 17 of article 1640 [1640.p17] is included in a section entitled "Copernican theory" in the astronomy article. When the query text consists of a single text paragraph, the text excerpts corresponding to the ranked output such as that included in Figures 3 and 4 can be presented to the user in the normal retrieval order. On the other hand, when a longer text is used as a query statement, then the retrieved text excerpts can be grouped according to similarities with particular query portions. This tends to collect the output into subgroups according to affinities with particular aspects of a query statement.

Consider as an example, article [9561] Galileo. The query text consists of one text section broken down into 11 text paragraphs. Each of these paragraphs treats a different aspect of the life of Galileo. By comparing each retrieved text excerpt with the full text of article [9561], and also with each individual paragraph of the article, it is possible to select the query segment exhibiting the largest similarity with each given retrieved item. The right-most column of Figure 4 identifies the closest query text excerpt for each retrieved text item. For three of the retrieved items, the full query text turned out to represent the best match. For the other 12 items, some specific query paragraph came closest. If each query segment is briefly identified by using a short characteristic phrase (as shown for some sample paragraphs of "Galileo" at the bottom of Figure 4), a tailored output can be produced for each user consisting of retrieved text excerpts that most closely match the user's particular interests. For example, a user interested in Galileo's work with falling bodies and the Copernican theory (query segment [9561.p3] would obtain excerpts from the articles on astronomy [1640], Copernicus [6165], Cosmology [6284], and the Copernican System [6164].

Figure 5 shows text excerpts obtained by a user interested in the general contributions of

Galileo [9561.p3]. They cover some aspects of Physics, Space Exploration, and the new understanding of aspects of the Creation. Two additional text excerpts dealing with scientific advances during the Renaissance [19463.p12] and the history of the Telescope [22223.p4] are not reproduced in Figure 5, although they also respond directly to query segment [9561.p3]. It is clear that the display of linked lists of text excerpts in response to user interest statements can lead to many interesting applications in the study and utilization of text collections. New methods can be devised for the use of textbook materials, the study of legal texts, the rapid reading of newspaper materials, and other similar text processing applications.

5 Evaluation

The evaluation of retrieval operations covering large text files raises many difficult problems. In conventional retrieval environments, it is not possible to generate the relevance assessments which would make it possible to determine that a particular stored text is, or is not, relevant to a particular user query. This makes it impossible to compute the formal recall and precision parameters that are normally used to evaluate retrieval performance. [6] Even under the assumption that only full encyclopedia articles are processed, rather than encyclopedia excerpts, over 300 million assessments are needed to establish the relevance of each retrieved encyclopedia article in response to each possible query article. When text sections and paragraphs are used independently, as outlined earlier, the number of needed judgments reaches many billions. In practice, it is necessary to use small samples, and more often than not, to substitute an investigator's subjective judgment for the preferred user assessments. This relatively undesirable approach was used to obtain the sample evaluation output of

Figure 6.

The evaluation of Figure 6 covers 100 sample encyclopedia searches, collectively retrieving a total of 802 text excerpts. In each case, the output was limited to a maximum of 15 retrieved documents per query, of which one half were text paragraphs, and about one third were text sections. Each of the retrieved items was examined for relevance with respect to the corresponding query text, making it possible to compute the precision data shown in Figure 6. (The retrieval precision is defined as the proportion of retrieved items that are judged relevant). As the figure indicates, nearly 95 percent of the retrieved items were assessed as relevant, and somewhat fewer than 5 percent were judged to be extraneous. This represents a very high standard of precision performance. Unfortunately, nothing concrete can be said about the recall, defined as the proportion of relevant material that is retrieved, because it is plainly impossible in realistic text environments to determine the possible relevance of all nonretrieved items, in addition to assessing the retrieved items that are needed to obtain the precision.

The sample output on Figure 3 illustrates the overall accuracy of the retrieval operations. It may be useful to examine some borderline cases. Figure 7(a) presents the output obtained in response to query [9621] Ganges River. As expected, the retrieved items at the top of the list are all rivers and regions in India that are closely tied to the Ganges. As the query similarity decreases, documents are retrieved that relate to other rivers, including the Wabash in North America and the Orinoco in South America. Such items might be considered to be extraneous. In an encyclopedia search such retrievals are, however, unavoidable because of the stereotyped way in which entities such as rivers, cities, mountain, lakes, etc., are described.

An example of erroneous retrieval is shown on Figure 7(b) where items dealing with the banking system are retrieved in response to [9561] Galley. The problem here is the high weight of ambiguous terms such as "bank", and the difficulty of distinguishing the placement of banks of rowers in galleys, with the placement of banks in the financial world. Problems such as those illustrated in Figure 7(b) are overcome easily, for example by raising the global retrieval threshold from 0.20 to 0.30, or by lowering the assigned weight of terms that are known to be unreliable because of the multiplicity of possible meanings (bank, base, etc.).

Any such measure designed to improve the retrieval precision (that is, to reduce the number of extraneous retrievals) will cause some loss of recall (that, is, it will also reduce the number of correct retrievals). In a practical retrieval environment it is then necessary to choose some precision-enhancing method that will minimize the recall loss.

Overall, the proposed multi-step text comparison system outlined in this study exhibits a very high standard of performance and it is immediately usable for the selective traversal and utilization of large text databases. The use of text components, rather than full texts, and the possibility of linking related text excerpts included in different texts, leads to flexible information browsing operations and improves the efficiency of the retrieval operations and the search effectiveness. The proposed methods are usable in arbitrary collection environments and could improve text access in many areas of application in the foreseeable future.

6 References

1. C. Stanfill and D. Waltz, Toward Memory-Based Reasoning, Communications of the ACM, 29:12, December 1986, 1213-1228.

2. U. Zernik, Corpus-Based Thematic Analysis, in Text-Based Intelligent Systems, P.S. Jacobs, editor, Lawrence Erlbaum Associates, Hillsdale, NJ, 1992, 101-122.

3. L. Wittgenstein, Philosophical Investigations, Basil Blackwell and Mott Ltd., Oxford University Press, Oxford, England, 1953.

4. G. Salton, and C. Buckley, Global Text Matching for Information Retrieval, Science, 253:5023, 30 August 1991, 1012-1015.

5. G. Salton and C. Buckley, Automatic Text Structuring and Retrieval — Experiments in Automatic Encyclopedia Searching, Proc. 14th ACM-SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 1991, 21-30.

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

7. Funk and Wagnalls New Encyclopedia, Funk and Wagnalls, New York, 1979. 29 volumes.

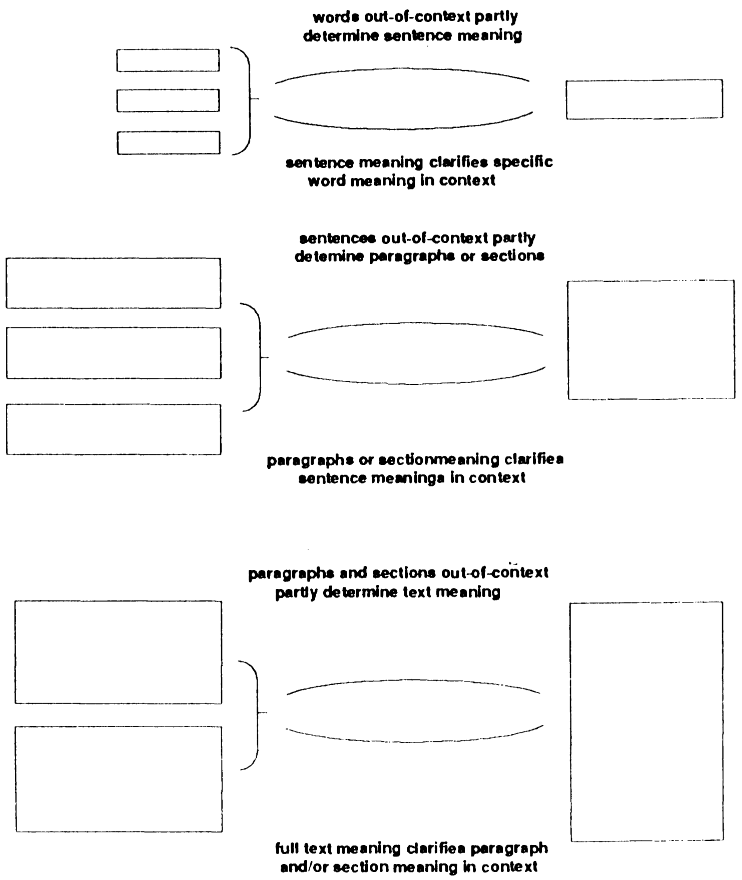

Figure 1. Schema for Derivation of Text Meaning

Dtext 9573

Gallieni, Joseph Simon.

(1849-1916), French soldier and colonial administrator, born in Saint-Beat, and educated at the military academy of Saint-Cyr. From 1877 until 1881 he participated in the explorations and military campaigns in the upper Niger River region that resulted in the extension of French influence in western Africa. In 1886, after three years in Martinique, he became governor of Upper Senegal. In 1896, when Madagascar became a French colony, Gallieni was military commander. He was then appointed governor general of the island, and retained that post until 1905. He established firm French control of Madagascar and instituted a program of economic development. On his return to France he was made general of a division and in 1906 was named military governor of Lyon. At the beginning of World War I, Gallieni was chosen to head the military government of Paris. He is credited with persuading General Joseph Jacques Cesaire Joffre to attack the Germans on the line of the Ourcq River in the first Battle of the Marne. During that battle Gallieni dispatched several thousand troops from Paris, using every available means of transportation, including taxicabs, to reinforce the army of General Michel Joseph Maunoury (1847-1923). For this action, which resulted in the repulse of the German right flank under General Alexander von Kluck (1846-1934), Gallieni was called the savior of Paris. In 1915 he was made minister of war in the cabinet of Prime Minister Aristide Briand, but ill health caused him to resign shortly before his death in 1916. The title of marshal was awarded posthumously to Gallieni in 1921.

a) Query Article [9573] "Gallieni"

	Global Text Similarity	Maximum Local Sentence Similarity
Sim(9573, 15074)	0.3206	224.45
Sim(9573, 15074.c3)	0.6404	59.64
Sim(9573, 15074.c4)	0.4296	224.45
Sim(9573, 15074.c5)	0.1034	56.05

b) Global and Local Text Similarities for Documents [9573] and [15074]

Dtext 15074.c4

----- Section 4 -----
Marne, Battle of the.
First Battle of the Marne.

(September 6-9, 1914), a decisive battle that halted the German advance near the Marne River, less than 48 km (30 mi) from Paris. The German forces had been encountering little resistance in their march on Paris. Then, supposedly because of an error in decoding an order, they wheeled to the southeast. Joseph Simon Gallieni, the military governor of Paris, persuaded the French commander in chief, Joseph Jacques Cesaire Joffre, to attack the flank thus exposed. Under Joffre's orders troops were rushed to the front by all available means, including taxicabs, and the Allied attack was begun on September 6. By September 9 the German armies had retreated, and the threat to Paris was ended.

c) Retrieved Article Section [15074.c4] (global similarity 0.4296)

```
-------------------------
Dmatch 9573 15074.c4
----
```

Ctype	Con	Vec1	Vec2	Product	Token
0	189	0.0162	0.0269	0.0004	includ
0	3755	0.0421	0.0699	0.0029	command
0	6609	0.0742	0.1234	0.0092	flank
0	13815	0.1161	0.0483	0.0056	french
0	13832	0.0849	0.2117	0.0180	batl
0	28437	0.1382	0.0766	0.0106	governor
0	32256	0.0262	0.0436	0.0011	begin
0	45281	0.0655	0.1090	0.0071	persuad
0	56697	0.0464	0.0771	0.0036	troop
0	58314	0.7037	0.1950	0.1372	gallien
0	66027	0.1119	0.1860	0.0208	cesair
0	78596	0.0678	0.1128	0.0077	simon
0	90517	0.0383	0.0637	0.0024	mean
0	96040	0.1119	0.3720	0.0416	joffr
0	96217	0.1218	0.2700	0.0329	pari
0	108081	0.0507	0.0422	0.0021	rive
0	119128	0.0908	0.4529	0.0411	marn
0	120322	0.1843	0.0613	0.0113	milit
0	122838	0.0667	0.1663	0.0111	germ
0	126369	0.1490	0.1651	0.0246	joseph
0	138160	0.0640	0.1065	0.0068	jacqu
0	159072	0.0431	0.1435	0.0062	attack
0	217504	0.0376	0.0625	0.0023	army
0	237700	0.1173	0.1950	0.0229	taxicab

d) Common Terms Between Texts [9573] and [15074.c4]

```
-------------------------
Dtext 9573.s11
----

----- Sentence 11 -----
He is
credited with  persuading General Joseph Jacques Cesaire Joffre to
attack the Germans  on the line of the Ourcq River in the first
Battle of the Marne.
```

```
-------------------------
Dtext 15074.c4.s5
----

----- Sentence 5 -----
Joseph Simon Gallieni, the
military  governor of Paris, persuaded the French commander in chief,
Joseph Jacques  Cesaire Joffre, to attack the flank thus exposed.
```

e) Typical Matching Sentence Pair [9573.s11] and [15074.c4.s5]
(sentence similarity 224.4535)

```
Dmatch 9573.s11 15074.c4.s5
-------------------------
```

Ctype	Con	Vec1	Vec2	Product	Token
0	45281	0.2419	0.2307	0.0558	persuad
0	66027	0.4128	0.3938	0.1626	cesair
0	96040	0.4128	0.3938	0.1626	joffr
0	126369	0.1832	0.3496	0.0641	joseph
0	138160	0.2363	0.2254	0.0532	jacqu
0	159072	0.1592	0.1518	0.0242	attack

f) Common Terms for Sentences [9573.s11] and [15074.c4.s5]

Figure 2. Illustrations of Global and Local Text Matching

ID=GALICIA1 [9557]

0.34	13900	LEON1
0.33	21610.p55	SPAIN
0.31	8726.c2	FERDINAND.I4
0.30	17323	ORENSE
0.28	8729.c2	FERDINAND.III1
0.28	4672.c3	CASTILE
0.28	13903	LEON4
0.25	13500	LA.CORUNA
0.23	12312.c3	ISABELLA.I
0.22	12679.p4	JOHN.OF.GAUNT.DUKE.OF.LANCASTE
0.21	18675.c31	PORTUGAL
0.21	645.c3	ALFONSO.I2

ID=GALICIA2 [9558]

0.40	23092.p7	UKRAINIAN.SOVIET.SOCIALIST.REP
0.32	14526.p4	LVOV
0.30	18518	POLAND
0.27	20029.c3	RUTHENIA

ID=GALILEE [9559]

0.68	15054.p9	MARK.GOSPEL.ACCORDING.TO
0.56	12351.p5	ISRAEL2
0.31	2728.c3	BETHSAIDA
0.29	4403.c3	CAPERNAUM
0.29	22490.p3	TIBERIAS.LAKE
0.27	12735.c4	JORDAN.HASHEMITE.KINGDOM.OF
0.27	17560.c4	PALESTINE
0.24	8305.c3	ESDRAELON.PLAIN.OF
0.23	24657.c3	ZEFAT
0.22	12746.p4	JOSEPHUS.FLAVIUS
0.21	6867.p3	DAYAN.MOSHE
0.21	15589.p19	MIDDLE.EAST
0.21	22489	TIBERIAS
0.21	24005.c3	WEST.BANK

ID=GALILEO [9561]

0.58	18179.p50	PHILOSOPHY
0.48	12083.p6	INERTIA
0.46	1640.p17	ASTRONOMY
0.44	2501.p5	BELLARMINE.SAINT.ROBERT
0.44	18234.c8	PHYSICS
0.41	6165.p12	COPERNICUS.NICOLAUS
0.35	20683.p12	SCIENCE
0.32	21607.p14	SPACE.EXPLORATION
0.32	6284.c4	COSMOLOGY
0.30	6419.p8	CREATION
0.30	18353.p6	PISA
0.27	6164	COPERNICAN.SYSTEM
0.24	19463.p12	RENAISSANCE
0.24	20686.p6	SCIENTIFIC.METHOD
0.23	22223.p4	TELESCOPE

Figure 3. Mixed Text Output Obtained for Four Test Queries

ID=Galileo [9561]

0.58	18179.p50	Philosophy/Modern Philosophy/Mechanism	9561.p11
0.48	12083.p6	Inertia/	9561
0.46	1640.p17	Astronomy/Copernican Theory	9561.p5
0.44	2501.p5	Bellarmine.Saint.Robert	9561.p8
0.44	18234.c8	Physics/Early History/16th and 17th Centuries	9561.p3
0.41	6165.p12	Copernicus.Nicolaus/Copernican System	9561.p5
0.35	20683.p12	Science/Modern Science	9561
0.32	21607.p14	Space.Exploration/History/Scientific Discoveries	9561.p3
0.32	6284.c4	Cosmology/Early Cosmological Theories	9561.p5
0.30	6419.p8	Creation/Scientific Controversy	9561.p3
0.30	18353.p6	Pisa/	9561.p7
0.27	6164	Copernican.System	9561.p5
0.24	19463.p12	Renaissance/Characteristics/Science Technology	9561.p3
0.24	20686.p6	Scientific.Method	9561
0.23	22223.p4	Telescope/History	9561.p3

Figure 4. Output for [9561] Galileo with Closest Related Text Elements Specified

([9561] complete article; [9561.p3] main contributions; [9561.p5] falling bodies; Copernican theory; [9561.p7] dispute with Pisan professors; [9561.p8] censorship; [9561.p11] experimental science)

Dtext 9561.p3

----- Paragraph 3 -----
Galileo (contributions)

full name Galileo Galilei (1564-1642), Italian physicist and astronomer, who, with the German astronomer Johannes (or Johann) Kepler, initiated the scientific revolution that flowered in the work of the English physicist Sir Isaac Newton. Galileo's main contributions were, in astronomy, the use of the telescope in observation and the discovery of sunspots, lunar mountains and valleys, the four largest satellites of Jupiter, and the phases of Venus. In physics, he discovered the laws of falling bodies and the motions of projectiles. In the history of culture, Galileo stands as a symbol of the battle against authority for freedom of inquiry.

.44

Dtext 18234.c8

----- Section 8 -----
Physics.
Early History of Physics.
16th and 17th Centuries.

The advent of modern science followed the Renaissance and was ushered in by the highly successful attempt by four outstanding individuals to interpret the behavior of the heavenly bodies during the 16th and early 17th centuries. The Polish natural philosopher Nicolaus Copernicus propounded the heliocentric system that the planets move around the sun. He was convinced, however, that the planetary orbits were circular, and therefore his system required almost as many complicated elaborations as the Ptolemaic system it was intended to replace (see Copernican System). The Danish astronomer Tycho Brahe, believing in the Ptolemaic system, tried to confirm it by a series of remarkably accurate measurements. These provided his assistant, the German astronomer Johannes Kepler, with the data to overthrow the Ptolemaic system and led to the enunciation of three laws that conformed with a modified heliocentric theory. Galileo, having heard of the invention of the telescope, constructed one of his own and, starting in 1609, was able to confirm the heliocentric system by observing the phases of the planet Venus. He also discovered the surface irregularities of the moon, the four brightest satellites of Jupiter, sunspots, and many stars in the Milky Way. Galileo's interests were not limited to astronomy; by using inclined planes and an improved water clock, he had earlier demonstrated that bodies of different weight fall at the same rate (thus overturning Aristotle's dictums), and that their speed increases uniformly with the time of fall. Galileo's astronomical discoveries and his work in mechanics foreshadowed the work of the 17th-century English mathematician and physicist Sir Isaac Newton, one of the greatest scientists who ever lived.

.30

Dtext 6419.p8

----- Paragraph 8 -----
Creation.
Scientific Controversy.

The medieval Christian church accepted Genesis as the complete story of creation. The story of Noah and the flood accounted for the existence of the different human races and the animals and plants found in the familiar world. As formal science developed, however, and the thought of the Greeks, particularly that of Aristotle (see Greek Philosophy), was recovered in the West about 1200, questions arose concerning the evidence of personal observation. Humanity as the center of the universe, for example, could not be accepted if the earth revolved about the sun, as proposed by the Polish astronomer Nicolaus Copernicus in the 16th century and refined by the German and Italian astronomers, respectively, Johannes Kepler and Galileo. Galileo was judged a heretic, but his observations could not be ignored. By the 17th century Western philosophers such as the Deists and Rene Descartes of France had laid the basis for what may be called the argument from design for the existence and nature of God. In simple form, this argument could be phrased in an analogy between the world and a clock. Even if one disbelieved the biblical account of creation, the intricacy of the operations of the world seemed to indicate the necessity for a supreme designer, something like a watchmaker, who set the mechanism going and regulated it if necessary. Explanations of the mechanics of the physical universe by the English mathematician and physicist Isaac Newton and others were accepted more or less readily in the 18th century. But discoveries in geology, and the growing possibility that the earth might be older than the 6000 years postulated by the Irish archbishop James Ussher (see Chronology) in the 17th century, disturbed traditionalists. Even more disturbing was the speculation that led to the evolutionary theories of Charles Darwin (see Evolution). The physical world, animal life, and even human beings were, according to this thesis, the product of gradual development, and specific creation was at least implicitly denied.

.32

.46

Dtext 21607.p14

----- Paragraph 14 -----
Space.
Exploration.
History.
Scientific Discoveries.

Not until some 1400 years later did the Polish astronomer Nicolaus Copernicus systematically explain that the planets, including the earth, revolve about the sun (see Copernican System). Later in the 16th century the observations of the Danish astronomer Tycho Brahe greatly influenced the laws of planetary motion set forth by Kepler (see Orbit). Galileo, Edmund Halley, Sir William Herschel, and Sir James Jeans were other astronomers who made contributions pertinent to astronautics.

Figure 5. Typical Text Excerpts Retrieved in Response to [9561.p3] "Galileo, contributions"

Total Number of Retrieved Text Items		802
Number of full documents	136	
Number of document sections	256	
Number of document paragraphs	410	

Total number of correct items retrieved	756 (94.3%)
Total number of marginal items	10
Total number of wrong items	36 (4.5%)

Figure 6. Retrieval Evaluation for 100 Test Queries
(maximum 15 retrieved items per query)

ID=GANGES [9621]

0.43	11534.c3	HOOGHLY
0.42	24518.c3	YAMUNA
0.34	2570.p5	BENGAL
0.33	23561	VINDHYA.RANGE
0.33	3454.c3	BRAHMAPUTRA
0.32	2084.c6	BANGLADESH
0.32	9882.c3	GHAGHARA
0.31	696	ALLAHABAD
0.30	10625	GUMTI
0.29	4772.c3	CAUVERY
0.28	23686.c3	WABASH
0.28	10078.c3	GODAVARI
0.27	7016.c3	DELTA
0.26	17338.p4	ORINOCO
0.24	23365	VARANASI

a) Retrieval Output with Some Marginal Items
([23686] Wabash; [17338] Orinoco)

ID=GALLEY [9569]

0.62	21087.p16	SHIPS.AND.SHIPBUILDING
0.31	19930	ROWING
0.30	16539.c4	NAVY
0.30	2094.c9	BANKING
0.24	8695.p8	FEDERAL.RESERVE.SYSTEM
0.20	8683.p5	FEDERAL.DEPOSIT.INSURANCE.CORP

b) Retrieval Output With Some Wrong Items
([2094] Banking; [8683] FDIC; [8695] Federal Reserve)

Figure 7. Examples of Questionable Output