Approaches to Passage Retrieval in Full Text Information Systems

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

Abstract

Large collections of full-text documents are now commonly used in automated information retrieval. When the stored document texts are long, the retrieval of complete documents may not be in the users' best interest. In such circumstances, efficient and effective retrieval results may be obtained by using passage retrieval strategies designed to retrieve text excerpts of varying size in response to statements of user interest.

New approaches are described in this study for implementing selective passage retrieval systems, and identifying text passages responsive to particular user needs. An automated encyclopedia search system is used to evaluate the usefulness of the proposed methods.

Introduction

In operational retrieval environments, it is now possible to process the full text of all stored documents. Many long, book-sized documents are stored, often containing a mix of different topics covered in more or less detail. In these circumstances, it is not useful to maintain the integrity of complete undivided documents. Instead individual text passages should be identified that are more responsive to particular user needs than the full document texts.

Several advantages are apparent when individual text passages are independently accessible. First, the efficiency of text utilization may be improved because the users are no longer faced with large masses of retrieved materials, but can instead concentrate on the most immediately relevant text passages. Second, the effectiveness of the retrieval activities may also be enhanced because relevant short texts are generally more easily retrievable than longer ones. The longer items covering a wide diversity of different subjects may not closely resemble narrowly-formulated, specific user queries. As a result, many potentially relevant items may be rejected. When text excerpts are accessible, the query similarity is often higher for the text excerpts than for the corresponding full texts, leading to the retrieval of additional relevant material with corresponding improvements in recall and precision.

Most text items are naturally subdividable into recognizable units, such as text sections, paragraphs, and sentences. This leads to the notion of assembling text excerpts of varying size covering just the right amount of information to satisfy the user population. In the remainder of this study, effective methods are introduced for identifying relevant text excerpts in response to user interest statements, and assembling these excerpts into retrievable text passages responsive to individual user needs.

Classical Text Excerpting

Methods have been introduced in the past for the use of text passages in information retrieval and automatic abstracting. The standard approaches utilize bottom-up procedures based on the identification of important text sentences that are later assembled into retrievable text units. Typically, each text sentence is assigned a score, or weight, depending on its perceived importance in the texts under consideration. Retrievable text passages, or text abstracts, are then formed by grouping a number of important highly-weighted sentences into units that are then independently processed for particular purposes.

*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.

Different methods have been used for the sentence scoring process. Typically, weights are assigned to the individual text words, and the complete sentence scores are then based on the occurrence characteristics of highly-weighted terms in the respective sentences. In passage retrieval applications where the task consists in retrieving text excerpts that are similar to available user queries, the number and concentration of query words included in the individual sentences is used to generate the sentence score. Increasing concentration of query terms, measured by the closeness of these terms in the text sentences, leads to higher assigned sentence weights. In text abstracting, or text summarization applications where user queries are not necessarily available, the sentence score is similarly based on the number and concentration of text words thought to be important in representing text content.

In addition to using the occurrence characteristics of highly-weighted terms, the sentence scoring system may be influenced by a number of additional factors such as:

1. The location of the sentences in the texts under consideration, special importance being given to texts representing figure captions, titles, and section headings.

2. The inclusion in the sentences of special clue words and clue phrases that are thought to be important in the determination of topicality and sentence value.

3. The use of syntactic relationships detected between particular words and sentences in the text, indicating that the corresponding text units are related and ought to be jointly retrieved. [1-8]

Various procedures have also been suggested for assembling individual text sentences into meaningful larger units. For example, specific rules have been proposed for generating so-called connected, concentrated, and compound text passages. [9-11]

Unfortunately, it is difficult to produce readable text passages by using the low-level term-weighting approaches normally available for this purpose. Alternative strategies have therefore been advocated for use in text abstracting and summarization based on deep semantic analysis techniques, and the use of pre-constructed frames, or templates, that are appropriately filled by information extracted from the document texts. [12-13] It is possible that approaches based on deep knowledge of particular subject fields will be useful for restricted tasks, such as, for example, the construction of medical summaries of certain types. When unrestricted subject matter must be treated, as is often the case in practice, the passage retrieval and text summarization methods proposed heretofore have not proven equal to the need.

One suggestion that may represent a step in the right direction is based on the use of text paragraphs, rather than sentences for the construction of text passages. In that case, relationships are computed between individual text paragraphs, based on the number of common text components in the respective paragraphs. Certain paragraphs are then chosen for abstracting purposes, replacing the originally available texts. [14] In the present study, the use of complete paragraphs is generalized to include text passages of varying length, covering the subject at varying levels of detail, and responding to varying kinds of user needs. A top-down approach is used whereby large text excerpts are chosen first, that are successively broken down into smaller pieces covering increasingly specific user needs. This makes it possible to retrieve full texts, text sections, text paragraphs, or sets of adjacent sentences depending on particular user requirements.

Global-Local Processing in Text Analysis and Retrieval

The Smart retrieval strategies are based on the vector processing model, where document and query texts are represented by sets, or vectors, of weighted terms. A term may be a word stem, or phrase, included in a particular text, and the term weights are chosen so as to favor terms that occur with high frequency inside particular documents, while being relatively rare in the collection as a whole. [15] To determine the similarity between a query and a stored document, or between two stored documents, the corresponding term vectors are compared, and coefficients of similarity are computed based on the number and the weight of common terms included in a vector pair. This makes it possible to rank the documents at the output in decreasing order of the computed query similarity.

The global vector similarity reflects global coincidence between query and document texts. The assumption is that when two vectors do not exhibit a given minimal global similarity, the corresponding texts are not related. Documents whose query similarity falls below a stated threshold are therefore rejected. The reverse

Query: [9562] Gall (Parasitic growth in plants)

	Retrieved Document Number	Query Similarity	Title	Unrestricted Output	Simple Sentence Match	Restricted Sentence Match
1.	9567	0.30	Galle (city in Sri Lanka)	•	× N	
2.	1494	0.28	Art Gallery	•	× N	
3.	9563	0.27	Galla (African people)	•	× N	
4.	12134	0.27	Insect	•	•	•
5.	17675	0.24	Parasite	•	•	•
6.	23008	0.24	Turner, J.M.W. (painter)	•	•	× N
7.	20402	0.24	St. Gallen (city)	•	•	× N
8.	11847	0.24	Hymenoptera (insects)	•	•	•
9.	17061	0.23	Oak	•	•	•
10.	22077	0.23	Tannins	•	•	•
11.	18414	0.23	Plant	•	•	•
12.	22034	0.23	Tamarisk	•	•	•
13.	170	0.22	Acorn (cross-reference)	•	× N	
14.	9578	0.21	Gallium (metal)	•	× N	
15.	22075	0.21	Tannic Acid	•	•	•
16.	11913	0.20	Ichneumon Fly	•	•	•
17.	4953	0.20	Chalcid (parasite)	•	•	•
18.	18228	0.20	Phylloxera (insect)	•	× R	
19.	7304	0.19	Diseases of Plants			
20.	23827	0.19	Wasp			

Table 1: Effect of Local Sentence Match Restrictions for Query [9562] Gall (removal of 7 nonrelevant (N) items and 1 relevant (R) item)

assumption may or may not hold, that is, when the global similarity between two vectors is sufficiently large, the texts may nevertheless be unrelated because the common vocabulary may be used in different senses in the respective texts. Thus, based on a global vector similarity computation, documents about table salt might be retrieved in answer to queries about the SALT treaties.

Problems caused by language ambiguity can be largely solved by making appeal to the "use theory" of meaning proposed by Wittgenstein and others, which states that the meaning of' words and expressions depends on the use of these words in the language. [16] This suggests that linguistic ambiguities and multiple meanings can be eliminated by checking the local context in which text words and expressions occur. The linguistic context of a term such as "salt" is likely to be the same for many different texts dealing with the SALT treaties. The same is true for texts discussing table salt. On the other hand, the contexts differ when the query texts are related to arms control but the documents deal with food consumption.

In practice, a dual text comparison system can be used, designed to verify both the global similarity between query and document texts, as well as the coincidence in the respective local contexts: [17,18]

1. The global text similarity is computed first by comparing the respective term vectors as previously explained. Text pairs without sufficient global similarity are not further considered.

2. The local text environments are considered next for texts with sufficient global similarity, and local structures included in the texts such as sections, paragraphs and sentences are compared. When globally similar text pairs also contain a sufficient number of locally similar substructures, the texts are assumed to be related.

The effect of the global/local text comparison system is illustrated in the example of Table 1. Here an article included in the Funk and Wagnalls encyclopedia is used as a search request, and other related encyclopedia articles are retrieved in response to the query articles. [19] Table 1 shows the 20 items exhibiting the highest similarity with the query article "Gall" (article number [9562]) based on the global vector similarity between query and retrieved article texts. Normally a minimum threshold of 0.20 applies for retrieval purposes, and documents with smaller global similarities are rejected.

The list of retrieved articles shown in Table 1 includes for the most part relevant items dealing with parasitic insects that cause the formation of galls in plants. Unfortunately, a number of extraneous documents are included on the list of items in Table 1, such as "Galle", a city in Sri Lanka, and "J.M.W. Turner", a British painter. The retrieval errors are produced by the term truncation used in constructing the term vectors: normally word stems, rather than complete words, are used in the query and document vectors, so that terms such as "gallery", "Galla", "gallium", and so on, are all reduced to *gall*, matching the term "gall" in the query.

The left-most column of bullets in Table 1, entitled "unrestricted output", identifies the top 18 items that are retrieved in answer to [9562] Gall by the simple global vector similarity computation. The next column to the right entitled "simple sentence match" represents a local retrieval strategy where at least one matching pair of sentences must be found in query and retrieved document texts above a sentence matching threshold of 75.0 in addition to a sufficiently large global similarity. [17,18] As the middle column of bullets shows, six of the originally retrieved items do not fulfill the simple local sentence matching requirement. This includes five obviously nonrelevant items (article [170] Acorn is not relevant because the whole article consists of only a cross-reference to another article), plus a possibly relevant item, [18228] Phylloxera, that appears at the bottom of the list.

The sentence matching strategy used to carry out the simple sentence match is based on a straightforward computation of pair-wise sentence similarity between all sentence pairs in the respective query and document texts. When a term is highly weighted, as is the case for "gall" in most of the retrieved articles, the required sentence similarity threshold of 75.0 may be reached even though a sentence pair has only that single term in common without additional matching context. To avoid that possibility, a restricted sentence match requirement can be used specifying that at least two matching terms must be present in a matching sentence pair, with no single term contributing more than 90 percent of the total sentence similarity. The right-most column of Table 1 shows that the two remaining nonrelevant items ([23008] J.M.W. Turner and [20402] St. Gallen) are now rejected because of the restricted local sentence match requirement.

An evaluation of the global/local text matching system is presented in the first 3 columns of figures of Table 2(a). The results of Table 2(a) are average recall and precision figures obtained by using as queries all encyclopedia articles starting with the letter G, and comparing them with all other encyclopedia articles. A global similarity threshold of 0.20 is used, and all query articles are included in the evaluation for which the relevance assessors found at least one relevant document. This represents a total of 982 "G" queries. Full relevance data were obtained from the owners of the encyclopedia for all articles retrieved by the simple sentence match (run 2 of Table 2(a), but not for the extra 5,000 documents retrieved by the unrestricted run 1 that were rejected by the local sentence-matching process of run 2. All these items are treated as nonrelevant in the evaluation of Table 2(a). The detailed example of Table 1 shows that items rejected by the local matching process are indeed generally nonrelevant. This is also confirmed by the detailed recall and precision figures included in Table 2(a) which both increase sharply between runs 1 and 2.

The precision figures included in Table 2 are exact; the recall figures must be interpreted as relative recall representing the total number of retrieved relevant items divided by the total number of known relevant for each query. Overall, the average precision improves by 9 percent over the unrestricted global text match of run 1 when the simple local sentence match is used, and by an additional 2 percent for the restricted sentence matching system.

Retrieval of Text Excerpts

The global/local text matching strategy described in the previous section is capable of retrieving relevant documents with a high degree of accuracy. However, restricting the retrieval to full document texts presents two main problems. Most obviously, the users will be overloaded rapidly when large document texts are involved. Second, a potential loss in retrieval effectiveness (recall) occurs because the global query similarity of the long, discursive documents that cover a number of different topics will be low, implying that many long documents will be rejected, even when they contain relevant passages.

The user overload can be reduced and the retrieval effectiveness enhanced by making it possible to retrieve text passages instead of full documents only, whenever the query similarity of a text excerpt is larger than the similarity of the complete document. This suggests that a hierarchical text decomposition

Evaluation Parameters	1. Global Text Match (un-restricted)	2. Simple Sentence Match	3. Restricted Sentence Match	4. Output With Text Sections	5. Full Text Sections or Paragraphs	
Retrieved Documents all queries	17,328	12,491	11,296	12,285	13,669	
Recall after 5 docs	0.5366	0.5820	0.5915	0.6316	0.6164	
Recall after 15 docs	0.7705	0.8027	0.7978	0.8702	0.8631	
Precision after 5 docs	0.6051	0.6629	0.6887	0.7097	0.6938	
Precision after 15 docs	0.3862	0.4167	0.4214	0.4466	0.4391	
11-point Average Precision	0.6988	0.7623 +9.1%	0.7779 +11.3%	0.8569 +22.5%	0.8340 +19.3%	(0.9280) (+24.7%)

a) Retrieval Evaluation for 982 Query Articles (letter "G")
(partial relevance assessments based on run 4)

Evaluation Parameters	3. Restricted Sentence Match	4. Output With Text Sections	5. Full Text Sections or Paragraphs
Retrieved Documents all queries	1106	1204	1345
Recall after 5 docs	0.4583	0.4834	0.5043
Recall after 15 docs	0.6667	0.7141	0.7631
Precision after 5 docs	0.7689	0.7867	0.8067
Precision after 15 docs	0.5022	0.5274	0.5519
11-point Average Precision	0.6920	0.7520 +9.7%	0.8198 +18.5%

b) Retrieval Evaluation for 90 "G" Query
(full relevance assessments for all retrieved items)

Table 2: Evaluation of Global/Local Text Comparison and Passage Retrieval

system be used which successively considers for retrieval text sections, text paragraphs, and sets of adjacent text sentences. In each case, the text excerpt with the highest query similarity may be presented to the user first, while providing options for obtaining larger or smaller text pieces. Because of the local context checking requirement used for retrieval purposes, text excerpts such as paragraphs and sentences are already individually accessible, and the additional resources needed for passage retrieval purposes are relatively modest.

The basic operations of a section and paragraph retrieval capability are illustrated in Table 3 for the query [9562] Gall, previously used in the illustration of Table 1. The left-most column of bullets in Table 3 corresponds to the restricted sentence match retrieving full documents only (equivalent to the right-most column of bullets in Table 1). In the middle column of Table 3, text sections are retrievable in addition to full documents, with the right-most column covering paragraph retrieval as well as sections and full text items.

The move to section retrieval promotes a number of retrieved items. Thus, the full article on Insects [12134] which contains 43 text paragraphs is replaced by section 10 of that article describing insect reproduction, and particularly the reproduction of gall wasps. Furthermore, the query similarity for vector [12134.c10] is 0.32, instead of 0.27 for the full document. This lifts the material on insects from retrieval rank 4 to rank 1. The section retrieval system also raises a previously unretrieved relevant item above the retrieval threshold of 0.20. Table 3 shows that the article on Wasps [23827] was originally rejected because of the low query similarity of 0.19. The query similarity for Section 4 of that article ([23827.c4]) reaches 0.20, leading to the retrieval of that relevant item.

Additional relevant items are promoted when paragraphs are added to the retrieval process as shown in the right-most column of Table 3. Thus, [11913] Ichneumon Fly is lifted from rank 14 to rank 12, [11847] Hymenoptera is raised from rank 7 to rank 2, and section 5 on Parasitic Plants ([17675.c5]) is replaced by paragraph 6 of that document ([17675.p6]) and raised from rank 5 to rank 3.

The effects of the passage retrieval capability are further detailed in the example of Table 4,

Query: [9562] Gall

	Retrieved Document Number	Query Similarity	Title	Restricted Sentence Match	Output With Sections	Section and Paragraph Output
1.	12134.c10	0.32	Insect/Reproduction		•	•
2.	11847.p6	0.32	Hymenoptera			•
3.	17675.p6	0.28	Parasite			•
4.	12134	0.27	Insect	•		
5.	17675.c5	0.25	Parasite/Parasitic Plants		•	
6.	17675	0.24	Parasite	•		
7.	11847	0.24	Hymenoptera	•	•	
8.	17061	0.23	Oak	•	•	•
9.	22077	0.23	Tannins	•	•	•
10.	18414	0.23	Plant	•	•	•
11.	22034	0.23	Tamarisk	•	•	•
12.	11913.p3	0.22	Ichneumon Fly			•
13.	22075	0.21	Tannic Acid	•	•	•
14.	11913	0.20	Ichneumon Fly	•	•	
15.	4953	0.20	Chalcid	•	•	•
16.	23827.c4	0.20	Wasp/Characteristics		•	•
...	...	...	...	...	...	...
17.	23827	0.19	Wasp			

Table 3: Effect of Retrieval of Text Excerpts (Sections and Paragraphs) for Query [9562] Gall (c_i = section i; p_j paragraph j)

covering a retrieval run for query [9561] on Galileo. Here all items retrieved by the restricted sentence match are relevant to the query, but the section and paragraph retrieval strategies are able to locate large numbers of previously unretrieved relevant items. A comparison of the three columns of bullets in Table 4 shows that the six originally retrieved full-text items are replaced by four full text items plus 8 text sections (see the middle columns representing section retrieval), and by one full-text item, two text sections, and 17 text paragraphs in the right-most column covering paragraph and section retrieval in addition to full texts.

All the long, originally retrieved items, such as [1640] Astronomy (39 paragraphs), and [20683] Science (25 paragraphs), are subdivided in the passage retrieval mode, and the retrieved sections and paragraphs are much more closely related to the query topic [9561] Galileo than the original full articles. Furthermore, a large number of long, originally rejected articles are retrieved in the passage retrieval mode, including very long items such as the documents on Physics [18234] containing 146 paragraphs of text, and on Philosophy [18179] containing 117 paragraphs. Many shorter relevant items are also obtained for the first time in the passage retrieval mode, including [6284] Cosmology (30 paragraphs), [12132] Inquisition (21 paragraphs), and [19463] Renaissance (27 paragraphs).

The efficiency improvements provided by the section and paragraph retrieval modes are illustrated at the bottom of Table 4. When full documents are retrieved, the average number of paragraphs contained in each retrieved item is nearly 16 for the Galileo query. This drops to only a little over 5 paragraphs per retrieved item in the section mode, and to only 1.3 paragraphs on average for the combined section plus paragraph mode. For the 982 query articles that start with the letter G, the average reduction in retrieved text length is 43 percent for the section retrieval, and 54 percent for paragraph retrieval. The users necessarily benefit when shorter text excerpts are obtained that are immediately related to the query topic instead of only full-text items.

The effectiveness of the section and paragraph retrieval is reflected by the evaluation data appearing in the two right-most columns of Table 2(a), covering section, and section plus paragraph retrieval, respectively. The recall and precision figures show that the passage retrieval techniques furnish substantially better output than the standard restricted sentence match for full documents alone. The average precision data of Table 2(a) show that the section output (run 4) is about 9 percent better than the restricted sentence match of run 3. This indicates that the vast majority of the 1,000 or so additional documents obtainable through the section output are indeed relevant, as shown earlier in the examples of Tables 3 and 4.

Query: [9561] Galileo

	Retrieved Document	Query Similarity	Title	Number of Paragraphs	Restricted Sentence Match	Output With Sections	Section and Paragraph Output
1.	18179.p72	0.58	Philosophy	1			•
2.	12083.p6	0.49	Inertia	1			•
3.	2501.p5	0.45	Bellarmine, St. Robert	1			•
4.	1640.p22	0.45	Astronomy	1			•
5.	18234.c8	0.44	Physics/History	2		•	•
6.	1640.c8	0.43	Astronomy	3		•	
7.	6165.p16	0.39	Copernicus N.	1			•
8.	18179.c25	0.36	Philosophy	4		•	
9.	20683.p16	0.35	Science	1			•
10.	6165.p17	0.32	Copernicus, N.	1			•
11.	6165.c7	0.31	Copernicus, N.	4		•	
12.	6284.p5	0.31	Cosmology	1			•
13.	18353.p6	0.30	Pisa	1			•
14.	6419.p10	0.30	Creation	1			•
15.	6284.c4	0.30	Cosmology	2		•	
16.	20683.p13	0.28	Science	1			•
17.	6165	0.28	Copernicus, N.	13	•		
18.	20683	0.27	Science	25	•	•	
19.	6164	0.27	Copernican System	4	•	•	•
20.	2501	0.27	Bellarmine, St. Robert	6	•	•	
21.	18234.p19	0.26	Physics	1			•
22.	6419.c5	0.26	Creation	3		•	
23.	18234.c9	0.26	Physics	2		•	
24.	20686.p6	0.24	Scientific Method	1			•
25.	19463.p17	0.24	Renaissance	1			•
26.	22223.p5	0.23	Telescope	1			•
27.	1640	0.23	Astronomy	39	•		
28.	20686	0.22	Scientific Method	8	•	•	
29.	1396.c7	0.21	Aristotle	2		•	•
30.	1636.p6	0.20	Astrology	1			•
31.	12132.p16	0.20	Inquisition	1			•
	Total Number of Retrieved Items				6	12	20
	Total Retrieved Paragraphs				95	65	25
	Avg. Number of Paragraphs Per Item				15.8	5.4	1.3

Table 4: Effect of Retrieval of Text Excerpts for Query [9561] Galileo
(c_i = section i ; p_j = paragraph j)

Unfortunately, the actual improvements obtained by the full section and paragraph retrieval system (run 5) are understated in Table 2(a). Full relevance information was available only for the section retrieval, but not for section plus paragraph output. The improvement in retrieval effectiveness shown in the right-most column of Table 2(a) is then due entirely to changes in the retrieval ranks of originally retrieved relevant items that are promoted by the section and paragraph retrieval system. The newly retrieved, originally rejected, items obtained in the paragraph mode, exemplified by the retrieval of [12083.p6] Inertia and [18353.p6] Pisa in Table 4, are all treated as nonrelevant in the evaluation since no relevance data were available for these items. This explains the apparent deterioration in performance between section and paragraph outputs (runs 4 and 5 of Table 2(a).

Table 2(b) contains complete evaluation data for 90 of the "G" queries previously included in Table 2(a) for which full relevance data were available. Table 2(b) shows that the actual improvement to be expected with the section output is about 10 percent over the restricted sentence match, with an additional 8 percent improvement for the paragraph output. If that additional 8 percent improvement were applied to the output of Table 2(a) for the complete query set, the average precision for run 5 would increase to 0.9280, giving an advantage of nearly 25 percent over the base run 1 for the complete set of 982 queries. These extrapolated figures are shown in parentheses to the right of column 5 in Table 2(a). The Galileo example of Table 4 shows that sections and paragraphs retrieved by the passage retrieval system are indeed relevant to the query topic, confirming to some extent the accuracy of the extrapolated performance figures.

The performance of the global/local text matching and text passage retrieval systems can be assessed by noting the variations in the figures given in the first and last rows of Table 2. The unrestricted global text match first retrieves a large number of items, many of which are deleted by the local sentence-matching process. The local text match thus acts as a precision device that rejects many previously retrieved nonrelevant items reducing the total retrieved from 17,328 to 11,296 overall. The passage retrieval system then acts as a recall device by adding text excerpts from relevant items that were not obtained earlier, and increasing the total number of retrieved items from about 11,300 to about 13,200. The combined effect of these procedures appears to improve retrieval effectiveness by about 25 percent over the unrestricted global text match.

Toward Flexible Automatic Passage Retrieval

The section and paragraph retrieval strategies outlined earlier are used to retrieve the most closely matching text sections and text paragraphs in answer to available query articles. In the earlier discussion, relatively general query statements were used to specify the retrieval context, covering, for example, "the life and works of Galileo", or "characteristics and causes of galls in plants". In many practical environments, it is desirable to extend the retrieval capability to much more specific contexts, for example, the "effect of Galileo's discoveries on the philosophical thought of the times", or "the effect of gall wasps on the health of oak trees". In these circumstances the search context must be substantially narrowed, and methods must be introduced for providing variable-length retrieval output at various levels of detail depending on particular user requirements.

One possibility consists in using search strategies capable of isolating variable-length text fragments from a variety of related documents, to be assembled into retrieval units covering the desired subject areas in detail. Consider, as an example, a request dealing with the philosophical influence of Galileo. It seems reasonable to carry out a dual search first of the Philosophy article (document [18179]) in the Galileo context, that is, using the Galileo article (document [9561]) as a query, and then of the Galileo article in the context of Philosophy.

By using a "moving window" containing n sentences at a time, it is possible to isolate n-sentence fragments that closely match a given query context. Table 5(a) shows some typical three-sentence fragments from article [9561] Galileo whose global similarity with [18179] Philosophy exceeds 0.10.

Table 5(b) similarly contains three-sentence fragments from [18179] Philosophy obtained in response to query [9561] Galileo.

To generate answers in response to specific queries about the philosophical impact of Galileo's work, it may be useful to combine related fragments obtained from various sources (articles [9561] Galileo and [18179] Philosophy in the present case). Sample similarity measurements between the three-sentence fragments of Table 5 are shown in Table 6(a).

An examination of the fragment texts reveals close relationships in many cases. For example, fragments s16-18 from article [9561] and s234-236 from [18179] both deal with aspects of the Copernican theory for the movement of planets around the sun. These fragments would therefore be combined in answer to an appropriate query. Similarly fragments s49-51 from [9561] Galileo and s251-253 from [18179] Philosophy specifically deal with the role of Galileo in shaping the philosophical thought of his time by introducing experimental approaches in physics. The text passage constructed from the text fragments used in the last example is reproduced in Table 6(b). Such a passage may be generated automatically in answer to queries about the philosophical contributions of Galileo.

Detailed experiments with text passage construction of this type remain to be carried out.

Sentence Numbers	Adjacent Sentence Fragments (3 Sentences)	Similarity Coefficient With Query
s16-18	At Padua, Galileo invented a calculating "compass" for the practical solution of mathematical problems. He turned from speculative physics to careful measurements, discovered the law of falling bodies and of the parabolic path of projectiles, studied the motions of pendulums, and investigated mechanics and the strength of materials. He showed little interest in astronomy, although beginning in 1595 he preferred the Copernican theory (see Astronomy: The Copernican Theory)–that the earth revolves around the sun–to the Aristotelian and Ptolemaic assumption that planets circle a fixed earth.	0.1012
s28-30	Professors of philosophy scorned Galileo's discoveries because Aristotle had held that only perfectly spherical bodies could exist in the heavens and that nothing new could ever appear there. Galileo also disputed with professors at Florence and Pisa over hydrostatics, and he published a book on floating bodies in 1612. Four printed attacks on this book followed, rejecting Galileo's physics.	0.1958
s49-51	Galileo's most valuable scientific contribution was his founding of physics on precise measurements rather than on metaphysical principles and formal logic. More widely influential, however, were The Starry Messenger and the Dialogue, which opened new vistas in astronomy. Galileo's lifelong struggle to free scientific inquiry from restriction by philosophical and theological interference stands beyond science.	0.1912

a) Typical Three-Sentence Fragments from [9561] Galileo
Obtained in Response to [18179] Philosophy

s234-236	In the 15th and 16th centuries a revival of scientific interest in nature was accompanied by a tendency toward pantheistic mysticism. The Roman Catholic prelate Nicholas of Cusa anticipated the work of the Polish astronomer Nicolaus Copernicus in his suggestion that the earth moved around the sun, thus displacing humanity from the center of the universe; he also conceived of the universe as infinite and identical with God. The Italian philosopher Giordano Bruno, who similarly identified the universe with God, developed the philosophical implications of the Copernican theory.	0.1870
s251-253	The work of Galileo was of even greater importance in the development of a new world view. Galileo brought attention to the importance of applying mathematics to the formulation of scientific laws. This he accomplished by creating the science of mechanics, which applied the principles of geometry to the motions of bodies.	0.5800

b) Typical Three-Sentence Fragments from [18179] Philosophy
Obtained in Response to [9561] Galileo

Table 5: Typical Three-Sentence Fragments Extracted from Related Documents (s = sentence)

Three-Sentence Fragment From [9561] Galileo	Three-Sentence Fragment From [18179] Philosophy	Pairwise Fragment Similarity
9561.s28-30	18179.s251-253	0.4484
9561.s49-51	18179.s251-253	0.4040
9561.s16-18	18179.s234-236	0.2002
9561.s16-18	18179.s251-253	0.1816

a) Typical Fragment Similarity Measurements for Fragments of Table 5

18179.s251–253	The work of Galileo was of even greater importance in the development of a new world view. Galileo brought attention to the importance of applying mathematics to the formulation of scientific laws. This he accomplished by creating the science of mechanics, which applied the principles of geometry to the motions of bodies.
9561.s49–51	Galileo's most valuable scientific contribution was his founding of physics on precise measurements rather than on metaphysical principles and formal logic. More widely influential, however, were The Starry Messenger and the Dialogue, which opened new vistas in astronomy. Galileo's lifelong struggle to free scientific inquiry from restriction by philosophical and theological interference stands beyond science.

b) Extract Covering "Galileo's Role in Shaping Philosophical Thought"

Table 6: Text Passage Formed by Fragment Juxtaposition

References

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

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

3. H.P. Edmundson, Problems in Automatic Abstracting, *Communications of the ACM*, 7:4, April 1964, 259-263.

4. J.E. Rush, R. Salvador, and A. Zamora, Automatic Abstracting and Indexing – Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria, *Journal of the ASIS*, 22:4, July-August 1964, 260-274.

5. L.L. Earl, Experiments in Automatic Extracting and Indexing, *Information Storage and Retrieval*, 6:4, October 1970, 313-334.

6. P.B. Baxendale, Man-Made Index for Technical Literature – An Experiment, *IBM Journal of Research and Development*, 2:4, 1958, 354-361.

7. C.D. Paice, Automatic Generation of Literature Abstracts – An Approach Based on the Identification of Self Indicating Phrases, in *Information Retrieval Research*, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen and P.W. Williams, editors, Butterworths, London, 1981, 172-191.

8. C.D. Paice, Constructing Literature Abstracts by Computer: Techniques and Prospects, *Information Processing and Management*, 26:1, 1990, 171-186.

9. J. O'Connor, Retrieval of Answer Sentences and Answer Figures by Text Searching, *Information Processing and Management*, 11:5/7, 1975, 155-164.

10. J. O'Connor, Data Retrieval by Text Searching, *Journal of Chemical Information and Computer Sciences*, 17, 1977, 181-186.

11. J. O'Connor, Answer Passage Retrieval by Text Searching, *Journal of the American Society for Information Science*, 32:4, July 1980, 227-239.

12. U. Hahn and U. Reimer, Entwurfsprinzipien und Architektur des Textkondensierungssystems TOPIC, Report TOPIC 14/85, University of Konstanz, Germany, April 1985.

13. U. Reimer and U. Hahn, Text Condensation as Knowledge Base Abstraction, Report MIP-8723, University of Passau, Germany, December 1987.

14. J.W. McInroy, A Concept Vector Representation of the Paragraphs in a Document Applied to Automatic Extracting, Report TR78-001, Computer Science Department, University of North Carolina, Chapel Hill, NC, 1978.

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

16. L. Wittgenstein, Philosophical Investigations, Basil Blackwell and Mott Ltd., Oxford, England, 1953.

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

18. G. Salton and C. Buckley, Automatic Text Structuring and Retrieval – Experiments in Automatic Encyclopedia Searching, Proc. Fourteenth Int. ACM/SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, 1991, 21-30.

19. Funk and Wagnalls New Encyclopedia, Funk and Wagnalls, New York, 1979, 29 volumes, 25,000 encyclopedia articles, ranging in length from one line for cross-reference articles to 150 pages of text for the article entitled "United States of America".