(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 November 2001 (15.11.2001)

PCT

(10) International Publication Number
WO 01/86490 A1

(51) International Patent Classification⁷: G06F 17/21, 17/28, 17/30

(21) International Application Number: PCT/US01/14635

(22) International Filing Date: 7 May 2001 (07.05.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/567,697 9 May 2000 (09.05.2000) US

(71) Applicant (*for all designated States except US*): **JUSTSYSTEM CORPORATION** [JP/JP]; Brains Park, Tokushima-shi, Tokushima 771-0189 (JP).

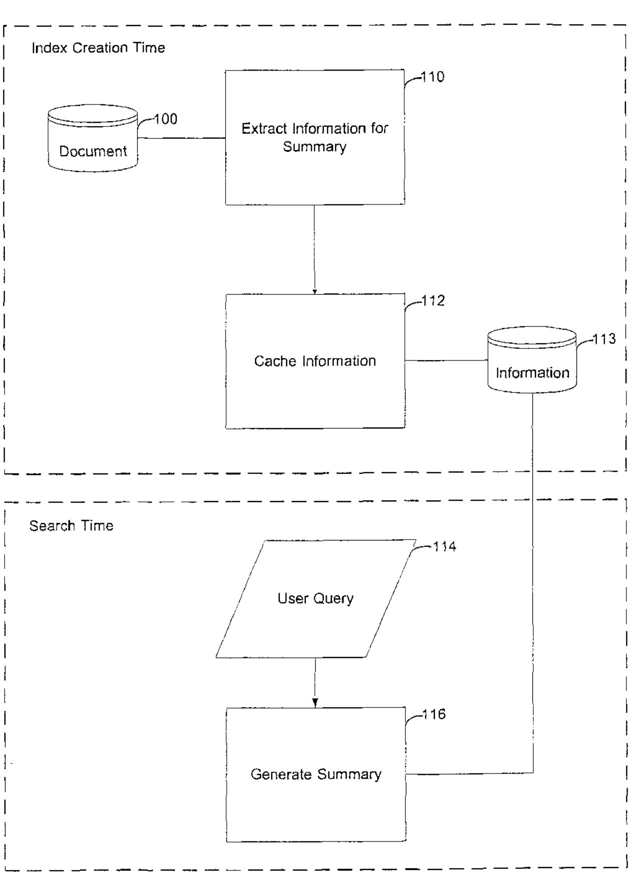
(72) Inventors; and
(75) Inventors/Applicants (*for US only*): **WITBROCK, Michael, J.** [NZ/US]; 249 Waverly Avenue, Newton, MA 02458 (US). **KANTROWITZ, Mark** [US/US]; 5503 Covode Street, Pittsburgh, PA 15217 (US). **MITTAL, Vibhu, O.** [IN/US]; 2483 Mount Royal Road, Pittsburgh, PA 15217 (US).

(74) Agents: **BYRNE, Richard, L.** et al.; Webb Ziesenheim Logsdon Orkin & Hanson, P.C., 700 Koppers Building, 436 Seventh Avenue, Pittsburgh, PA 15219-1818 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, CZ (utility model), DE, DE (utility model), DK, DK (utility model), DM, DZ, EE, EE (utility model), ES, FI, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

[Continued on next page]

(54) Title: METHOD AND APPARATUS FOR RAPIDLY PRODUCING DOCUMENT SUMMARIES AND DOCUMENT BROWSING AIDS



(57) Abstract: Disclosed is a computer-assisted method for generating a summary (420a) of or a browsing aid for a document (400). At an index creation time, information that is relevant to at least one dummy query (410a) and is necessary to compile at least one temporary summary for the summary or browsing aid is extracted from a document and cached for later use (415a). The information may be compiled into the summary and saved as such. At a search time (420a) or browsing aid is generated using the information (415a) that was cached at index creation time. An apparatus for performing this computer-assisted method is also disclosed.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

METHOD AND APPARATUS FOR RAPIDLY PRODUCING DOCUMENT SUMMARIES AND DOCUMENT BROWSING AIDS

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to computer-assisted information storage and retrieval and, more particularly, to producing document summaries and document browsing aids.

2. Description of the Prior Art

As part of search results corresponding to a user query, for example, in an information retrieval system, a query-biased summary generation system provides a document summary that incorporates sentences, sentence fragments, or text spans that are relevant to the user query.. The full text of the document must be available in order to create the query-biased summary. Usually, the summary includes the sentences having the greatest number of user query terms that appear most frequently. The summary can also include sentences that are closely related to the query by incorporating synonyms of the query terms into the criteria for the selection of the included sentences. With the current state of the art, the generation of a query-biased summary requires significant processing time.

Current information retrieval systems and information management systems, such as web catalogs, search engines, and document indexes, do not use query-biased summaries. Topical document summaries that are relevant to the user query are not provided. Instead, for example, they present the first few sentences of a document as an indication of the content of that document. These first few sentences may be extracted from the document and stored as a summary of that document for later use in response to a user query. While this technique works well with news stories that use the inverted pyramid style of writing, where the most important facts are mentioned toward the beginning of the article, it does not work well with other text genres that typically do not use the inverted pyramid style.

As a result of the current state of the art, after results pages for a user query are displayed, the user may have to undertake the laborious process of visiting each website listed on the search results pages to determine whether the document listed is relevant. Many users do not have the time or patience to do this. Moreover, a user who leaves the web catalog to examine the mentioned documents for relevancy is more likely to be distracted and not return to the catalog.

A web catalog's revenue generation is primarily dependent on advertisements and, more specifically, on the number of advertisement exposures per second. Since the web catalog generates revenue by exposing a user to advertising, the web catalog generates more revenue when the user remains on the web catalog site for as long of a time as possible. Thus, when a user leaves the catalog to examine the mentioned documents for relevancy and does not return to the web catalog, potential advertising revenues are lost.

A study by Tombros and Sanderson recently showed that query-biased summaries allow users to decide whether a document is relevant without having to read the document. Tombros, Anastasios and Sanderson, Mark, Advantages of Query Biased Summaries in Information Retrieval, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28, 1998, pages 2-10. In this study, users of a typical web catalog output referred to the full text of a document 23.7% of the time. In contrast, with query-biased summaries, users referred to a document only 1.37% of the time. These results led to the conclusion that a query-biased summary provided users with enough clues to judge a document's relevance to the query without the need to read the document itself.

Current web search engines do not provide query-biased document summaries for several reasons. The main reason is that computation time is extremely limited. Since revenue generation is dependent on advertising exposures, response time to a query is critical.

Generating query-biased summaries as part of the retrieval process would add enough of a delay to decrease the revenue throughput of the web catalog. The added delay might additionally cause some users to switch to a competitor's faster web catalog. Moreover, web catalogs answer tens of millions of queries per day, and adding a second or two of computation time per query might necessitate the purchase of additional equipment to handle the increased demands on the system.

Since the state of the art for query-biased summarization requires that the full text of the document be available, legal restrictions may prevent web catalogs from producing query-relevant summaries at search time. Current copyright law may restrict the ability of web catalogs to maintain a copy of the full text of a document. Today, practitioners in the field generally believe that copyright law permits web catalogs to store only short excerpts of a document, not the entire document itself. It is also generally believed that web catalogs may retrieve a document's full text in order to index it, so long as the full text is discarded after generation of the index.

Lastly, the size of the web has been estimated to be three terabytes in late 1998 and to be growing at a rate of approximately 35% per year. Storing the full text of every web page so that query-biased summaries can be generated at search time would require a great deal of disk space and may be prohibitive in cost.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide a method and apparatus for rapidly producing document summaries and document browsing aids which do not require storing the full text of the documents.

Accordingly, we have developed a method and apparatus for rapidly producing document summaries and document browsing aids by, at index creation time, precomputing and caching query relevant information required for creating the summaries.

In the specification and claims, the words "cache" and "caching" mean to store data for reuse. For example, a disk cache is random-access memory that stores information retrieved from disk, keeping the most frequently accessed data in memory. Use of a disk cache saves time since it takes less time to retrieve information from memory than from disk. In this application, the word "cache" is used in a similar sense meaning that a precomputed summary or summaries are stored to avoid the need to compute them later when needed in response to a query.

In the specification and claims, the word "term" means single words, word n-grams, and/or phrases. An "n-gram" is a string of characters that may comprise all or part of a word.

The present invention avoids the problems in the current art by extracting topical information for each document at index time and caching either the key information required to generate the summary efficiently at search time or the topical summary. This substantially reduces the computation time and storage requirements and removes the necessity to retain entire documents for producing query-biased summaries. Thus, it becomes feasible for web catalogs and other information retrieval systems to provide topical summaries in the search results pages.

The present invention splits the summary generation process into two parts: one for index time and the other for search time. When the full text of a document is retrieved for indexing (or in an equivalent separate summarization process), the first part generates and stores query relevant information that will be used by the second part to produce or select summaries efficiently at search time.

Since computation time is not as critical at index creation time, the first part of the invention does not need to be particularly time efficient; however, the second part must be extremely time efficient. Caching the information for the summary at index creation time allows the web search engine to generate the summary at search time without requiring the full text of the document or a large amount of computation resources. Thus, the present invention shifts the least time efficient aspects of the summary generation process from the second part (at search time) to the first part (at index creation time) of the process. Shifting most of the computation time from search time to index creation time also reduces the total amount of computation time because the summary information need only be computed once instead of every time the document appears in search results for a user query.

Providing topical summaries will increase revenues by attracting more users to the catalog. Users will save time and have access to a more useful tool. Also, providing topical summaries will increase revenues by removing the need for current users to leave the catalog to examine a document in the search results for relevancy to the user query. Since users will remain at the web catalog instead of leaving it to visit other web pages, the number of advertising exposures per second is likely to increase.

The present invention can be adapted to any text summarization method involving text spans and to a variety of computation intensive document browsing aids. The invention significantly reduces the time it takes to produce a query-biased summary without substantially affecting the quality of the summary. Thus, the invention makes it feasible for web search engines and other information retrieval systems to display document summaries and other document browsing aids with the results of a search.

While the present invention is advantageous for use with web search engines, it also applies to other information retrieval systems where a short response time is critical. For example, the invention can apply to an application that indexes a user's personal files or archives of e-mail messages where it is more important for the system to run quickly when the user executes a search than when the documents are initially indexed. Likewise, any product that includes a document index, for example, an encyclopedia or an archive of news articles, can be shipped with the cached summary information. Thus, the consumer would be provided with a fast and efficient user interface.

The present invention also applies to document browsing aids, such as keyword gists, thumbnail images, clustering, and categories. A keyword gist is a shortened form of a document in which all but the keywords have been deleted. A thumbnail is a reduced image of the document (e.g., a photo reduction). Clustering involves grouping related documents together into a cluster. Categorizing is similar to clustering, but instead assigns a label to each document in which the label identifies which group that document belongs. By optimizing the search time generation of these aids through the precomputing and caching of information, the present invention makes these aids practical for real world applications, such as web catalogs and document indexes.

An apparatus of the present invention utilizes any computer-assisted system including hardware and/or software for processing data which is capable of storing and executing the methods according to the present invention. The apparatus also is capable of storing documents or accessing stored documents. The apparatus may include a client system, a server system, communications links, and a network. In such an apparatus, the communications links are utilized to establish communication channels across the network and between the client and server systems.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flow diagram of a first embodiment of a method according to the present invention;

Fig. 2 is a flow diagram of a second embodiment of a method according to the present invention;

Fig. 3 is a flow diagram of a third embodiment of a method according to the present invention;

Fig. 4 is a flow diagram of a fourth embodiment of a method according to the present invention;

Fig. 5 is a flow diagram of a fifth embodiment of a method according to the present invention;

Fig. 6 is a flow diagram of a sixth embodiment of a method according to the present invention;

Fig. 7 is a flow diagram of a seventh embodiment of a method according to the present invention; and

Fig. 8 is an illustration of an apparatus according to the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring to Fig. 1, a first embodiment of a method according to the present invention begins at index creation time with step 110 by extracting information from a document 100. The extracted information is information that is relevant to at least one dummy query and is necessary to compile at least one temporary summary for the document 100. In step 112, information 113 is cached for later use in creating a summary of the document 100. At search time, a user query is entered at step 114, and step 116 generates the summary for the document 100 by utilizing the cached information 113.

Dummy queries are constructed to resemble the most probable user queries that would be relevant to the document being summarized. The dummy query may include terms most likely to be contained in a user query searching for the document 100. The dummy query may include terms that frequently appear in the document 100 after stop words are removed. The frequency of the terms may be obtained from document frequency statistics. Likewise, the dummy query may include terms in a term vector or a passage term vector (i.e., a term vector for a passage of a document) of the document 100. The dummy query may also include terms that are synonyms of the terms in the term vector. The dummy query may contain terms that appear frequently in documents in a top ranked documents list resulting from a user query containing the terms of the term vector. Likewise, the dummy query may include terms that are in both the document 100 and prior user queries. The prior user queries may be those in which the document 100 appeared in a top ranked documents list.

Step 110 may utilize an information extraction algorithm that extracts specific information. For example, the information may be the title of the document, the author's name, the time and date of an event, certain sentences, or a summary or an abstract coded into the document 100.

Whole sentences may be extracted. In order to aid in the decision of which sentences to extract, a score may be assigned to each sentence of the document 100. For example, step 110 may extract at least the highest scoring sentence, or may extract a pre-defined number of the highest scoring sentences. Step 110 may extract the highest scoring sentences that have a score greater than a threshold score. Up to a pre-defined number of the highest scoring sentences having a score greater than a threshold may be extracted. Likewise, step 110 may extract all of the sentences of a paragraph that contain a number of the highest scoring sentences. Furthermore, each score may be assigned based upon the similarity of the sentence to sentences of documents in a results document list created by the dummy query.

Step 112 may cache the extracted information 113 in an inverted document index with a document pointer and document attributes. Likewise, the information 113 may be cached in a query table.

At search time, if a paragraph was extracted in step 110, the summary generated in step 116 may be the paragraph that was extracted.

A second embodiment of a method of the present invention is shown in Fig. 2. At index creation time, information from a document 200 which is relevant to at least one dummy query and is necessary to compile at least one temporary summary is extracted in step 210. Step 212 caches information 213. Step 214 compiles a summary of the document 200. Summary 217 is cached in step 216. At search time, a user query is entered at step 218. The cached summary 217 is utilized as the generated summary in step 220.

In a third embodiment of a method of the present invention, shown in Fig. 3, information from a document 300 that is relevant to at least one dummy query and is necessary to compile at least one temporary summary is extracted in step 310. Step 312 caches information 313. Step 314 generates links from the information 313 to the corresponding locations in the document 300 at which the information 313 is found. Step 316 caches links 317 that were generated in step 314. At search time, a user query is entered at step 318 and a summary of the document 300 is generated in step 318 using the information 313 cached in step 312. The links 317 may be provided with the summary to the user such that the user can jump directly to the relevant portions of the document 300.

Fig. 4 illustrates a fourth embodiment of a method of the present invention. At index creation time, at least two dummy queries are entered in steps 410a and 410b for use in steps 412a and 412b to extract information from a document 400 that is relevant to at least one dummy query and is necessary to compile at least one temporary summary. Steps 414a and 414b separately cache the extracted information 415a and 415b, respectively. At search time, a user query is entered in step 416. Step 418 determines which dummy query best matches the user query. The dummy query that best matches the user query, as determined in step 418, determines whether step 420a or 420b generates the summary of the document 400 using the information 415a or 415b, respectively, cached in step 414a or 414b, respectively.

The information extracted in steps 412a and 412b may be a summary or abstract coded in the document or a set of results pages for each dummy query. The information may be extracted by assigning at least one score to each sentence in the document according to the relevance of the sentence to the at least one dummy query and extracting at least the highest scoring sentence.

Steps 414a and 414b may also cache a label with the information 615a and 615b. Each label may consist of the terms of the corresponding dummy query. At search time, step 418 determines which dummy query label best matches the user query. Steps 420a or 420b generates the summary using the information 415a or 415b, respectively, depending upon which label was deemed to best match the user query in step 418. Steps 412a and 412b may also generate a summary for each dummy query from the information extracted. Each summary is separately cached with the extracted information in steps 414a and 414b. At search time, step 420a or 420b generates the summary using the summary cached in 414a or 414b, respectively.

Fig. 5 shows a fifth embodiment of the present invention wherein dummy queries are utilized as in Fig. 4. At index creation time, dummy queries are entered in steps 510a and 510b for use in steps 512a and 512b to extract information from a document 500 that is relevant to at least one dummy query and necessary for compiling at least one temporary summary. Steps 514a and 514b separately cache extracted information 515a and 515b. However, once the information 515a and 515b is cached in steps 514a and 514b, steps 516a and 516b generate links from the query terms to the locations in the document 500 at which the query terms are found. Steps 518a and 518b cache links 519a and 519b generated in steps 516a and 516b, respectively. At search time, a user query is entered in step 520. Step 522 determines which dummy query best matches the user query. The dummy query that best matches the user query determines whether step 524a or 524b generates the summary of the document 500 using the information 515a or 515b, respectively, cached in step 514a or 514b, respectively. The links 519a and 519b may be provided to the user such that the user can jump directly to the relevant portions of the document 500.

Fig. 6 shows a sixth embodiment of a method of the present invention. At index creation time, for one or more documents of a collection of documents 600, step 610 extracts information that is relevant to at least one dummy query and is necessary to compile one or more temporary summaries for each of the one or more documents. Step 612 compiles the one or more temporary summaries, and step 614 caches the one or more summaries. At search time, a user query is entered in step 618. Step 620 generates a summary for one or all of the documents in the collection 600 using the cached one or more temporary summaries 616.

Fig. 7 shows a seventh embodiment of a method of the present invention. At index creation time step 710 extracts information from a document 700 that is relevant to at least one dummy query and is necessary to compile one or more temporary summaries in order for a document browsing aid to be generated at search time. Information 713 is cached in step 712. At search time, a user query is entered in step 714. Step 716 generates a document browsing aid using the cached information 713 in step 712.

Fig. 8 shows an embodiment of an apparatus 800 capable of enabling the methods according to the present invention. The apparatus 800 includes a client system 810, communications links 812 and 814, a network 816, and a server system 818. The client system 810 may be any apparatus including hardware and/or software for processing data. The server system 818 may be any apparatus including hardware and/or software for processing data. The communications links 812 and 814 may be any channels of communication. The network 816 may be any network, for example, the Internet.

The client system 810 is coupled to the network 816 by the communications link 812. The server system 818 is coupled to the network 816 by the communications link 814.

In use, a document is stored on the server system 818. The document may be stored at a location other than the server system 818 as long as the server system 818 has the ability to access and retrieve information from the document. At index time, a method according to the present invention executes on the server system 818 to precompute and cache the information that is relevant to at least one dummy query and is required for creating a summary of or a document browsing aid for the document at search time. At search time, a user enters a user query on the client system 810. The communications links 812 and 814 relay this query across the network 816 to the server system 818. The method then generates the summary or document browsing aid in response to the user query and sends the results back to the client system 810 over the network 816 using the communications links 812 and 814.

While a client/server system is described, any computer-assisted apparatus may be utilized as long as the system is capable of storing and executing the methods according to the present invention.

The present invention permits document summaries and browsing aids to be rapidly produced by precomputing and caching query relevant information from a document at index creation time such that a summary or document browsing aid can be created at search time in a time-efficient manner from the cached information. By using the cached query relevant information at search time, the present invention also eliminates the need to have access to the entire document at search time in order to produce the summary or browsing aid.

It will be understood by those skilled in the art that while the foregoing description sets forth in detail preferred embodiments of the present invention, modifications, additions, and changes may be made thereto without departing from the spirit and scope of the invention. Having thus described our invention with the detail and particularity required by the Patent Laws, what is desired to be protected by Letters Patent is set forth in the following claims.

We claim:

1. A computer-assisted method for generating a summary of a document, comprising the steps of:

at an index creation time,

extracting from the document information that is relevant to at least one dummy query and is necessary to compile at least one temporary summary, and

caching at least part of the information; and

at a search time,

generating the summary from the information cached.

2. The computer-assisted method according to claim 1, wherein the dummy query includes terms most likely to be contained in a user query searching for the document.

3. The computer-assisted method according to claim 1, wherein the dummy query includes terms that appear frequently in the document after stop words are removed.

4. The computer-assisted method according to claim 1, wherein the dummy query includes terms in a passage term vector of the document.

5. The computer-assisted method according to claim 1, wherein the dummy query includes terms in a term vector of the document.

6. The computer-assisted method according to claim 5, wherein the dummy query also includes terms that are synonyms of the terms in the term vector of the document.

7. The computer-assisted method according to claim 5, wherein the dummy query also includes terms that appear frequently in documents in a top ranked documents

list resulting from a user query containing the terms in the term vector.

8. The computer-assisted method according to claim 1, wherein the dummy query includes terms that are in both the document and prior user queries.

9. The computer-assisted method according to claim 8, wherein the prior user queries result in the document appearing in a top ranked documents list.

10. The computer-assisted method according to claim 1, wherein the information is extracted using an information extraction algorithm that extracts specific information.

11. The computer-assisted method according to claim 10, wherein the information extraction algorithm extracts the summary or abstract coded in the document.

12. The computer-assisted method according to claim 1, wherein the information extracted is text spans.

13. The computer-assisted method according to claim 12, wherein the text spans are sentences.

14. The computer-assisted method according to claim 13, wherein:

the information is extracted by the step of assigning at least one score to each sentence in the document according to the relevance of the sentence to the at least one dummy query,

wherein at least the highest scoring sentence is extracted.

15. The computer-assisted method according to claim 14, wherein a pre-defined number of the highest scoring sentences are extracted.

16. The computer-assisted method according to claim 14, wherein the highest scoring sentences that have a score greater than a threshold are extracted.

17. The computer-assisted method according to claim 16, wherein up to a pre-defined number of the highest scoring sentences is extracted.

18. The computer-assisted method according to claim 14, wherein:

the information is extracted by the steps of:

assigning at least one score to each sentence in the document according to the relevance of the sentence to the at least one dummy query, and

extracting all of the sentences of a paragraph of the document which contains a number of the highest scoring sentences, and

the summary generated is the paragraph.

19. The computer-assisted method according to claim 14, wherein the score assigned to each sentence is based upon the similarity of the sentence to sentences of documents in a results document list created by execution of the at least one dummy query, and the dummy query contains terms most likely to be contained in a user query searching for the document.

20. The computer-assisted method according to claim 1, wherein the cached information is stored in an inverted document index with a document pointer and document attributes.

21. The computer-assisted method according to claim 1, wherein the cached information is stored in a query table.

22. The computer-assisted method according to claim 1, further including the steps of:

at index creation time,

compiling the at least one temporary summary from the extracted information, and

caching the at least one temporary summary as part of the information cached,

wherein the summary generated is one of the at least one temporary summaries.

23. The computer-assisted method according to claim 1, further including the steps of:

generating a link that associates the information with at least one position within the document to which the information relates; and

caching the link.

24. The computer-assisted method according to claim 1, wherein:

the information extracted is relevant to at least two dummy queries each made up of at least one term and is separately cached for each dummy query, and

the summary is generated from the information cached for the dummy query that substantially matches a user query having at least one term.

25. The computer-assisted method according to claim 24, wherein the information extracted is a summary or abstract coded in the document.

26. The computer-assisted method according to claim 24, wherein the information extracted is a set of results pages for each dummy query.

27. The computer-assisted method according to claim 24, wherein:

the information is extracted by steps of assigning at least one score to each sentence in the document according to the relevance of the sentence to the at least one dummy query, and

wherein at least the highest scoring sentence is extracted.

28. The computer-assisted method according to claim 24, wherein:

the information extracted includes a label consisting of each term of the corresponding dummy query; and

the summary generated consists of the sentences associated with the dummy query in which the terms of the label substantially match terms of the user query.

29. The computer-assisted method according to claim 24, wherein:

the information cached includes one document summary generated from the information extracted for each of the at least two dummy queries, and a label consisting of each term of the corresponding dummy query, and

the summary generated consists of the document summary associated with the dummy query in which the terms of the label substantially match the terms of the user query.

30. The computer-assisted method according to claim 24, wherein the information is sentences, and the computer-assisted method further includes the steps of:

generating links that associate the dummy query terms with the locations in the documents of the sentences that contain the terms; and

caching the links.

31. A computer-assisted method for generating a summary of a collection of documents, comprising the steps of:

at an index creation time,

for one or more documents in the collection of documents, extracting information that is relevant to at least one dummy query and is necessary to compile one or more temporary summaries for each of the one or more documents,

compiling the one or more temporary summaries from the extracted information,

caching the one or more temporary summaries; and

at a search time,

generating the summary for the collection of documents from the cached one or more temporary summaries.

32. A computer-assisted method for generating a query-biased document browsing aid, comprising the steps of:

at an index time,

extracting information that is relevant to at least one dummy query and is necessary to compile the browsing aid from the document;

caching at least part of the information; and

at a search time,

generating the browsing aid from the information cached.

33. The computer-assisted method according to claim 31, wherein the query-biased document browsing aid is a keyword gist.

34. The computer-assisted method according to claim 31, wherein the query-biased document browsing aid is a thumbnail image.

35. The computer-assisted method according to claim 31, wherein the query-biased document browsing aid is clustering.

36. The computer-assisted method according to claim 31, wherein the query-biased document browsing aid is categories.

37. An apparatus to enable a method for generating a summary of at least one document, comprising:

a means for extracting information that is relevant to at least one dummy query and is necessary to compile at least one temporary summary from the at least one document at an index creation time; and

a means for caching at least part of the information at an index creation time; and

a means for generating the summary from the information cached at a search time.

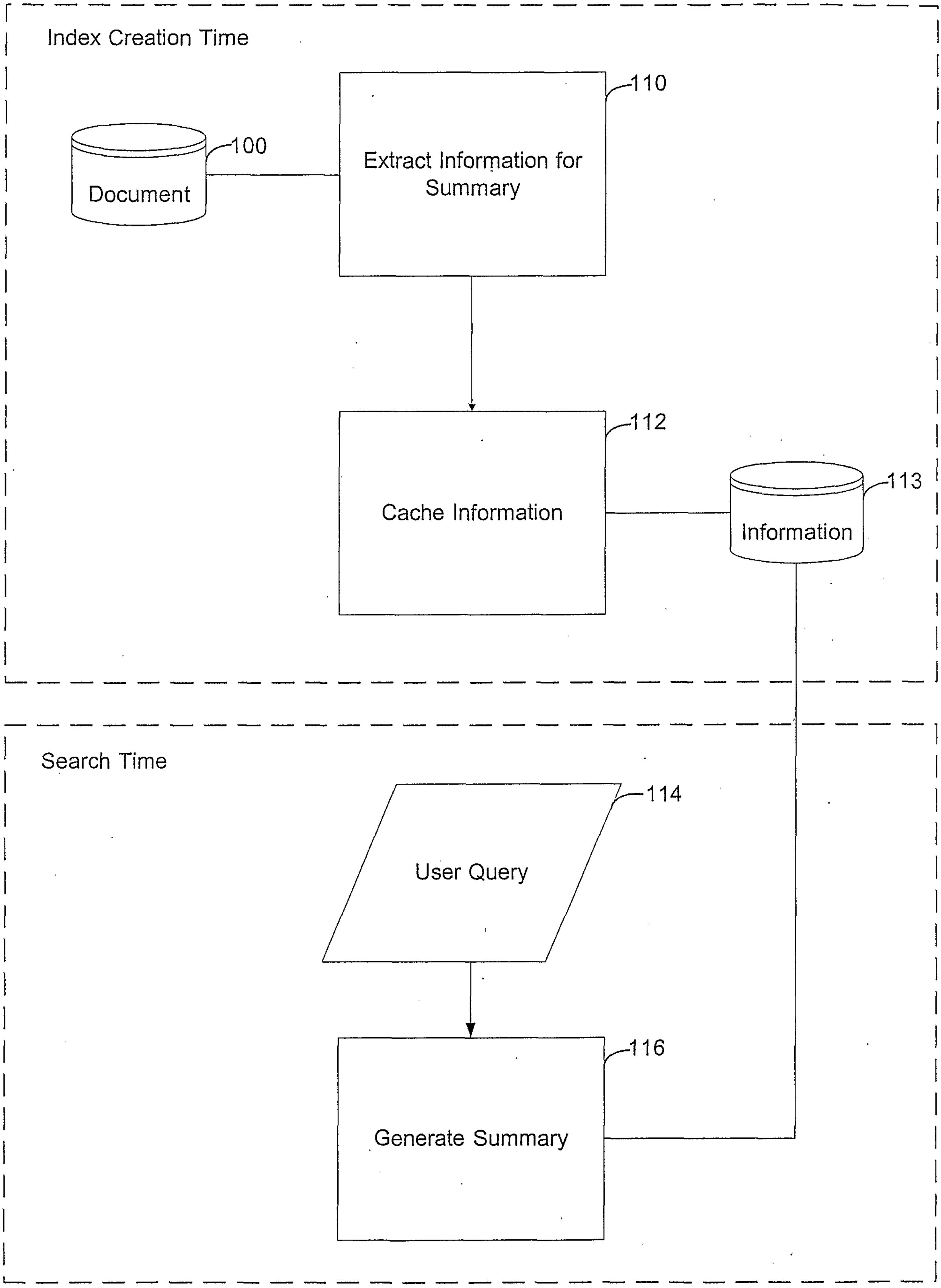

Index Creation Time
Document
100
Extract Information for Summary
110
Cache Information
112
Information
113
Search Time
User Query
114
Generate Summary
116

FIG. 1

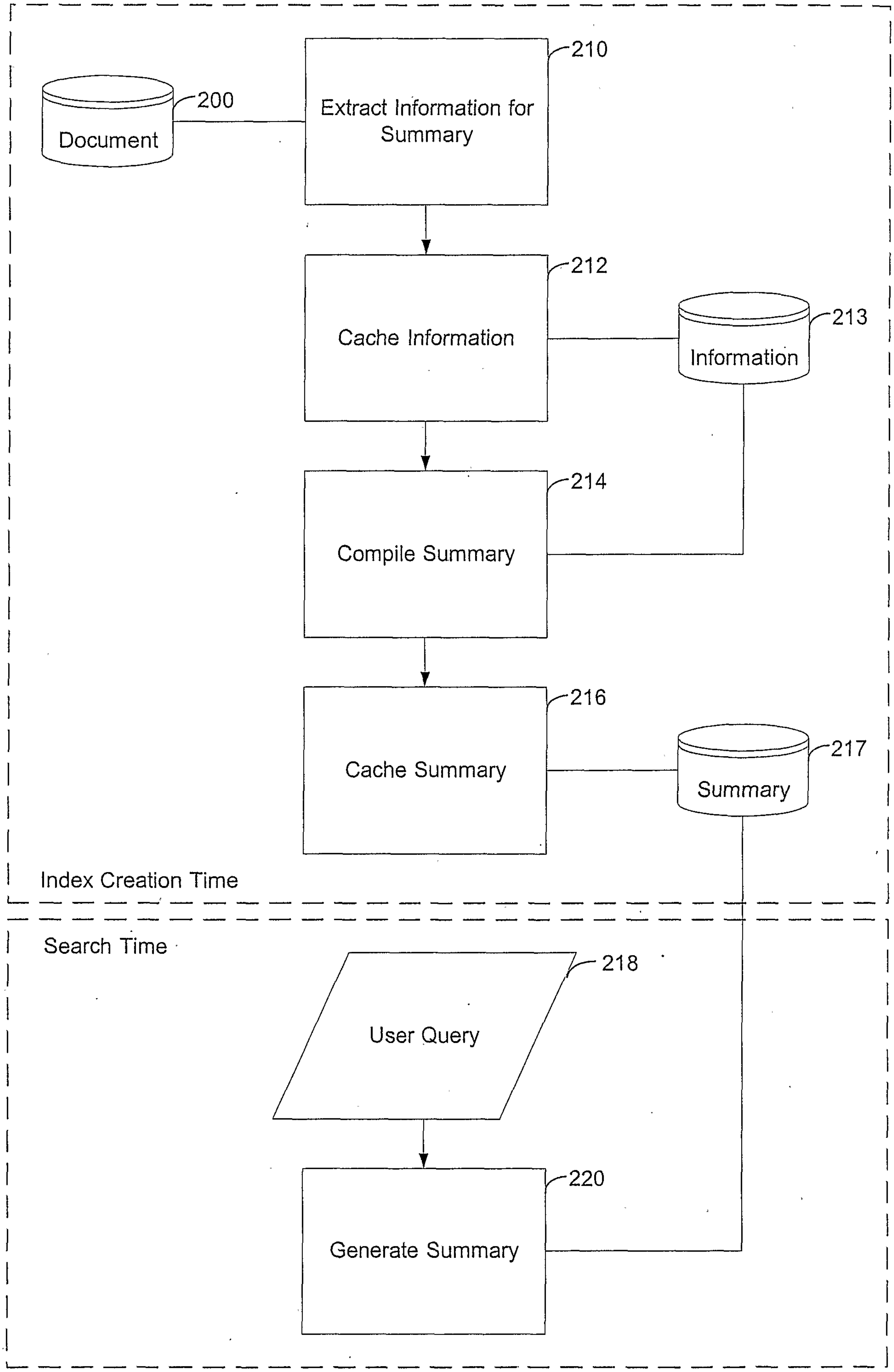

Document
200
Extract Information for Summary
210
Cache Information
212
Information
213
Compile Summary
214
Cache Summary
216
Summary
217
Index Creation Time
Search Time
User Query
218
Generate Summary
220

FIG. 2

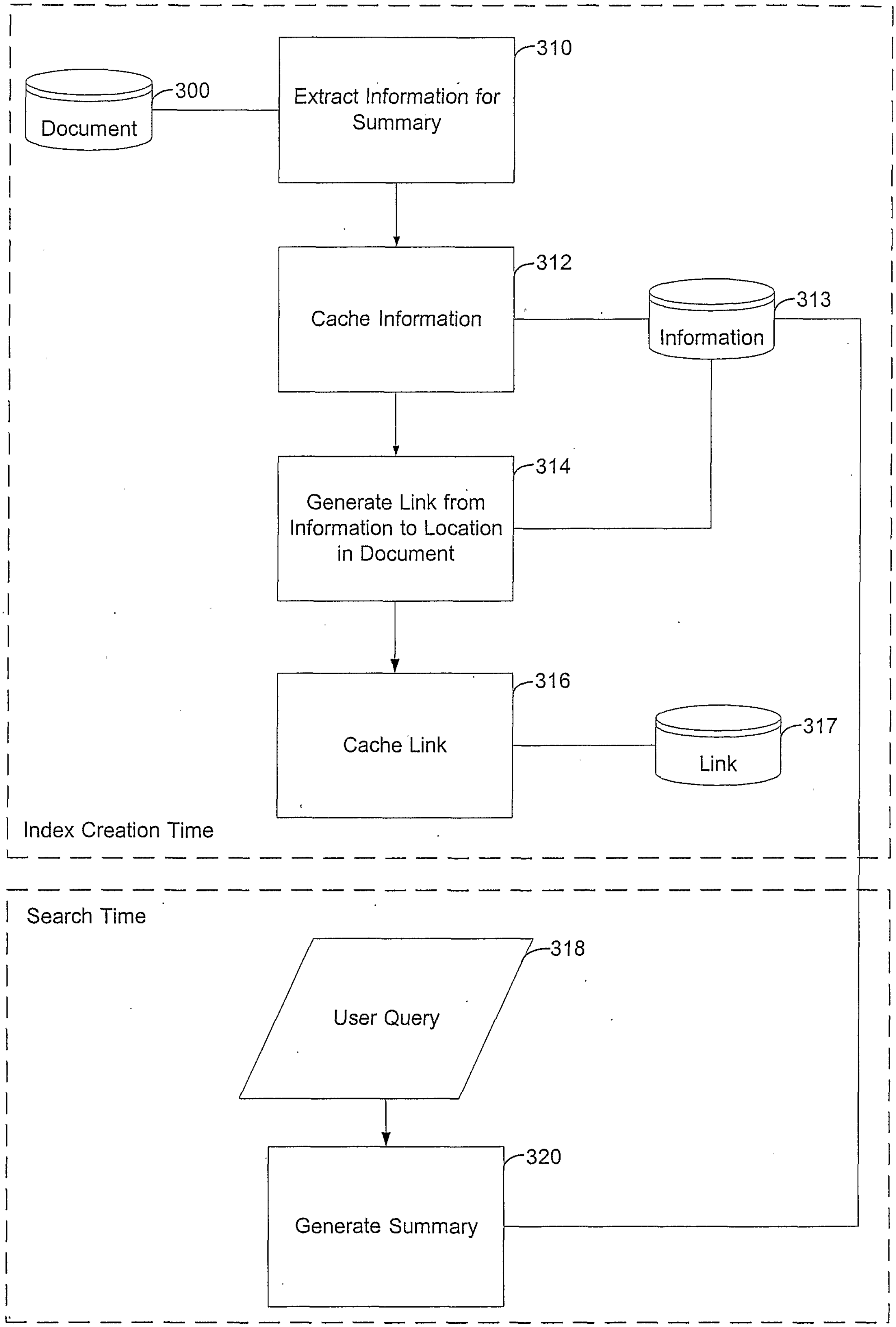

Document
300
Extract Information for Summary
310
Cache Information
312
Information
313
Generate Link from Information to Location in Document
314
Cache Link
316
Link
317
Index Creation Time
Search Time
User Query
318
Generate Summary
320

FIG. 3

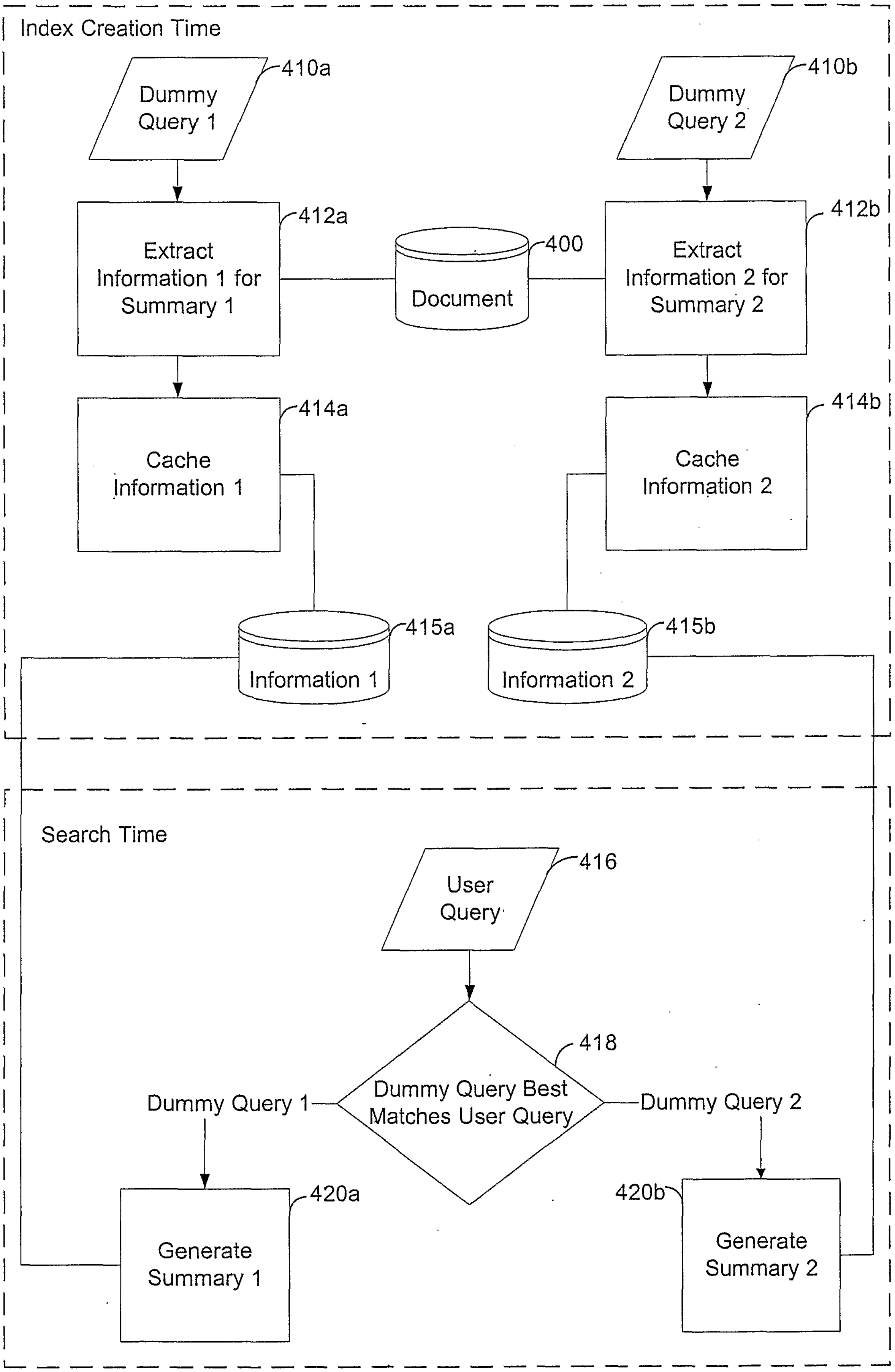

Index Creation Time
Dummy Query 1
410a
Dummy Query 2
410b
Extract Information 1 for Summary 1
412a
Document
400
Extract Information 2 for Summary 2
412b
Cache Information 1
414a
Cache Information 2
414b
Information 1
415a
Information 2
415b
Search Time
User Query
416
Dummy Query Best Matches User Query
418
Dummy Query 1
Dummy Query 2
Generate Summary 1
420a
420b
Generate Summary 2

FIG. 4

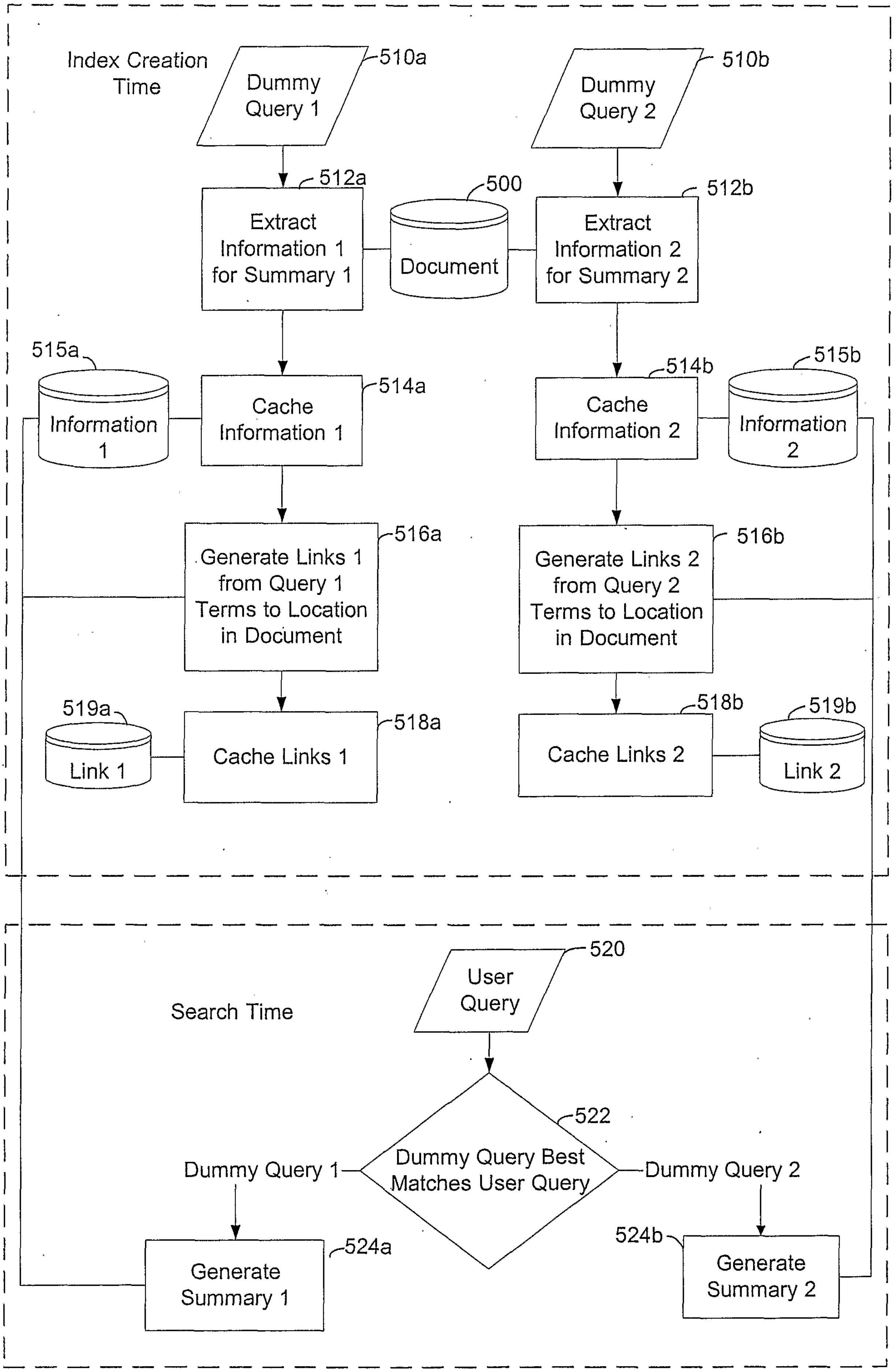

Index Creation Time
Dummy Query 1
510a
Dummy Query 2
510b
Extract Information 1 for Summary 1
512a
500
Document
Extract Information 2 for Summary 2
512b
515a
Information 1
Cache Information 1
514a
514b
Cache Information 2
515b
Information 2
Generate Links 1 from Query 1 Terms to Location in Document
516a
Generate Links 2 from Query 2 Terms to Location in Document
516b
519a
Link 1
Cache Links 1
518a
518b
Cache Links 2
519b
Link 2
520
Search Time
User Query
522
Dummy Query 1
Dummy Query Best Matches User Query
Dummy Query 2
Generate Summary 1
524a
524b
Generate Summary 2

FIG. 5

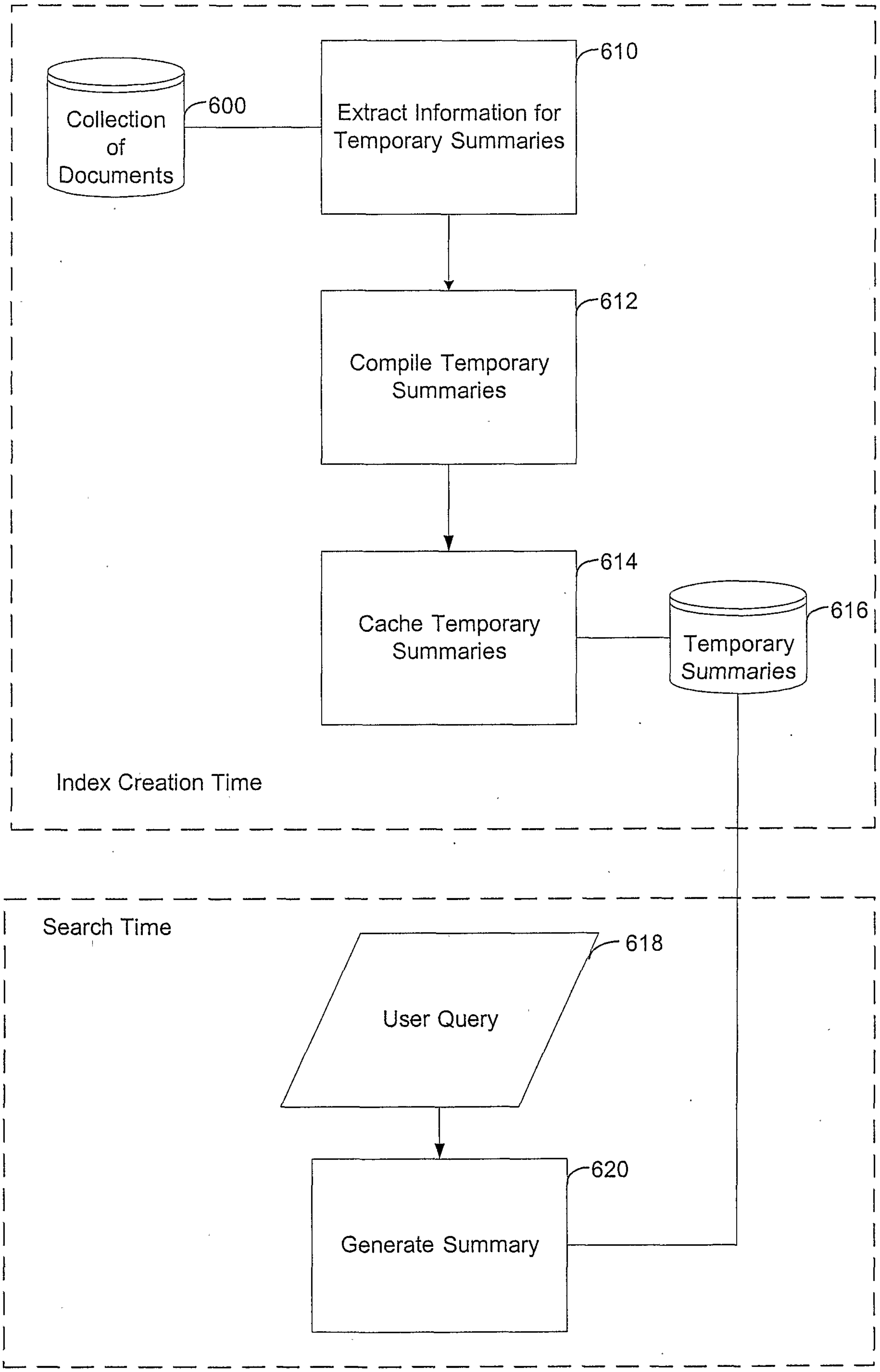

Collection of Documents
600
Extract Information for Temporary Summaries
610
Compile Temporary Summaries
612
Cache Temporary Summaries
614
Temporary Summaries
616
Index Creation Time
Search Time
User Query
618
Generate Summary
620

FIG. 6

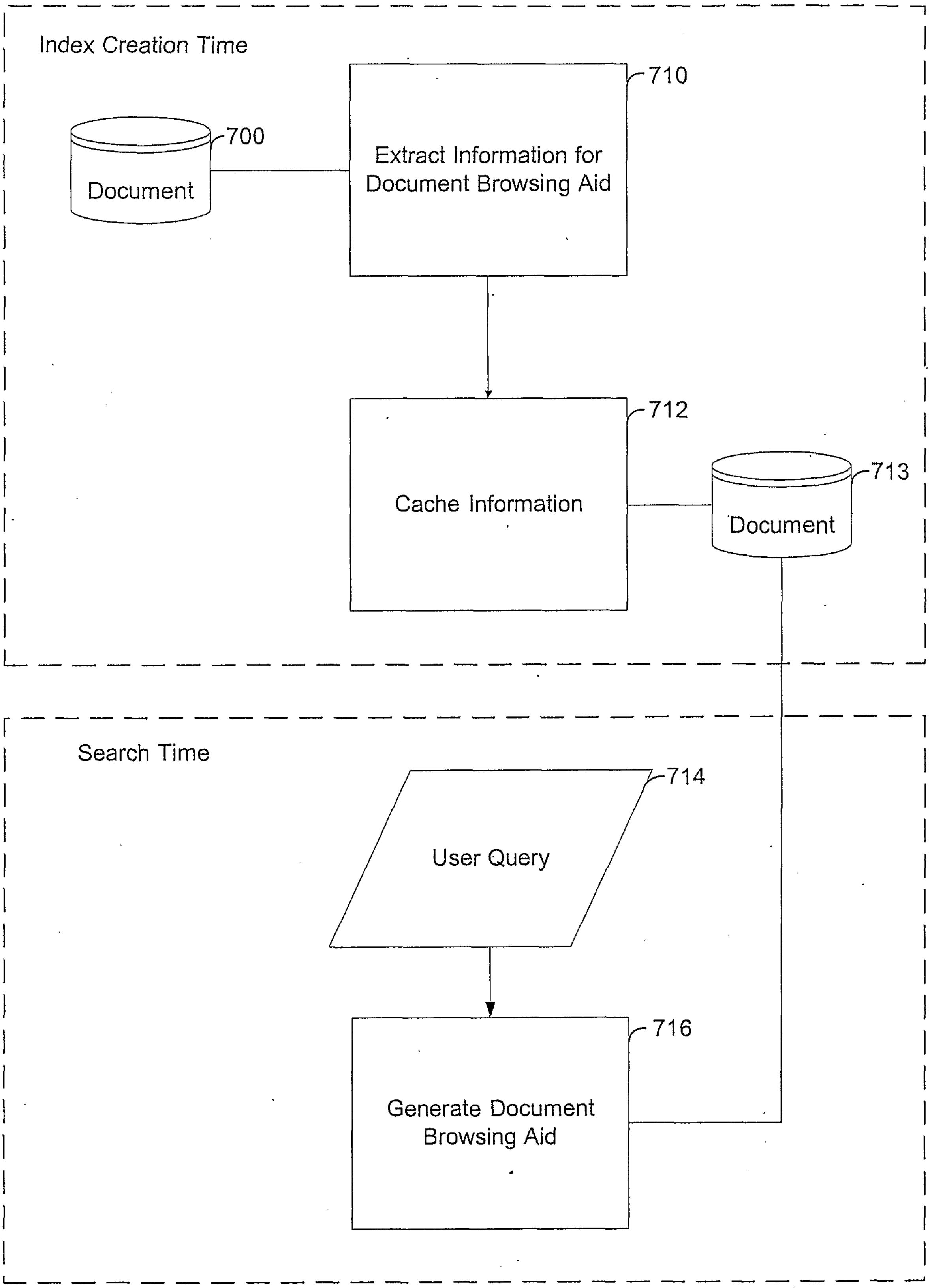

Index Creation Time
700
Document
710
Extract Information for Document Browsing Aid
712
Cache Information
713
Document
Search Time
714
User Query
716
Generate Document Browsing Aid

FIG. 7

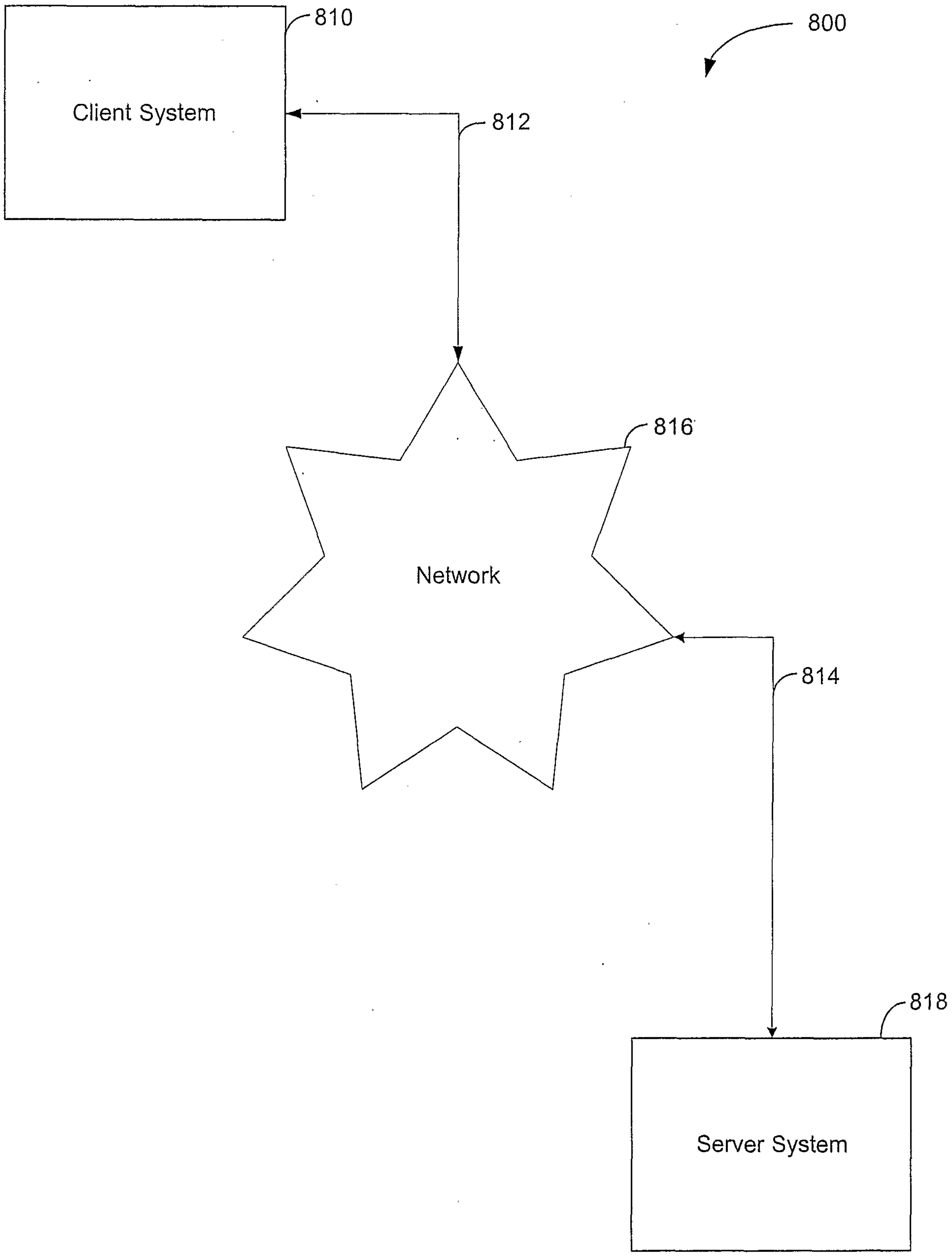


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/14635

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 17/21, 17/28, 17/30
US CL : 707/5, 500, 531
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 707/1-5, 500, 530 ,531

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 6,009,442 A (CHEN et al) 28 December 1999 (28.12.1999) column 1, line 64 through column 2, line 43.	1-7, 10-18, 20-23, 31-37
Y	US 5,838,323 A (ROSE et al) 17 November 1998 (17.11.1998) column 6, line 24 through column 7 line 49.	1-7, 10-18, 20-23, 31-37
Y	US 5,689,716 A (CHEN) 18 November 1997 (18.11.1997), column 3, line 9 through column 5 line 32.	1-7, 10-18, 20-23, 31-37
Y	US 5,050,071 A (HARIS et al.) 17 September 1991 (17.09.1991) Column 20, lines 11-44.	1-7, 10-18, 20-23, 31-37
A	Intelligent Document Retrieval, IBM Technical Disclosure Bulletin, 1 June 1991, Vol. 34, No. 5, pages 403-405, whole document relevant	1-37

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

* Special categories of cited documents:

"A"	document defining the general state of the art which is not considered to be of particular relevance	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E"	earlier application or patent published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search 31 August 2001 (31.08.2001)	Date of mailing of the international search report 18 OCT 2001
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Jack M Choules James R. Matthews Telephone No. (703) 305-3900

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US01/14635

Continuation of B. FIELDS SEARCHED Item 3:
EAST, USPAT, EPO, JPO, IBM TDB, DERWENT, IEEE
index, summary, document, abstract, query, dummy, pseudo, example, topical,