US 20020123988A1

(19) United States

(12) Patent Application Publication (10) Pub. No.: US 2002/0123988 A1

Dean et al. (43) Pub. Date: Sep. 5, 2002

(54) METHODS AND APPARATUS FOR EMPLOYING USAGE STATISTICS IN DOCUMENT RETRIEVAL

(75) Inventors: **Jeffrey A. Dean**, Menlo Park, CA (US); **Benedict Gomes**, Berkeley, CA (US); **Krishna Bharat**, Santa Clara, CA (US); **Georges Harik**, Mountain View, CA (US); **Monika H. Henzinger**, Menlo Park, CA (US)

Correspondence Address:
Straub & Pokotylo
Suite 83
1 Bethany Road
Hazlet, NJ 07730 (US)

(73) Assignee: **Google, Inc.**

(21) Appl. No.: **09/797,754**

(22) Filed: **Mar. 2, 2001**

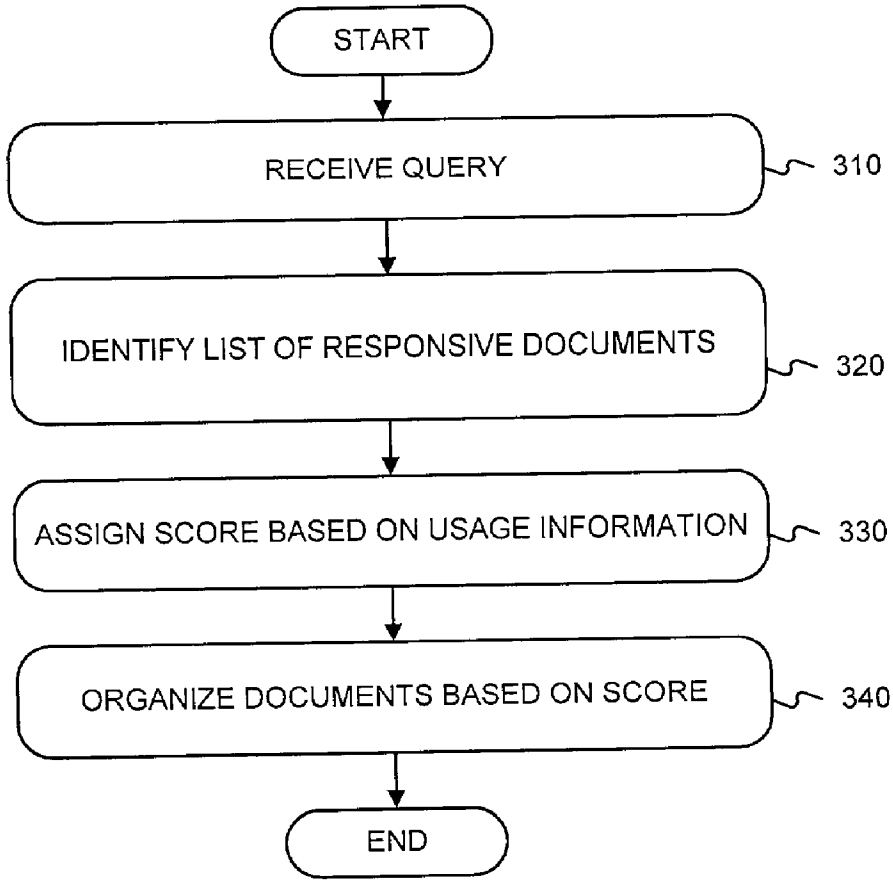
Publication Classification

(51) **Int. Cl.**[7] **G06F 7/00**

(52) **U.S. Cl.** **707/3**

(57) **ABSTRACT**

Methods and apparatus consistent with the invention provide improved organization of documents responsive to a search query. In one embodiment, a search query is received and a list of responsive documents is identified. The responsive documents are organized based in whole or in part on usage statistics.



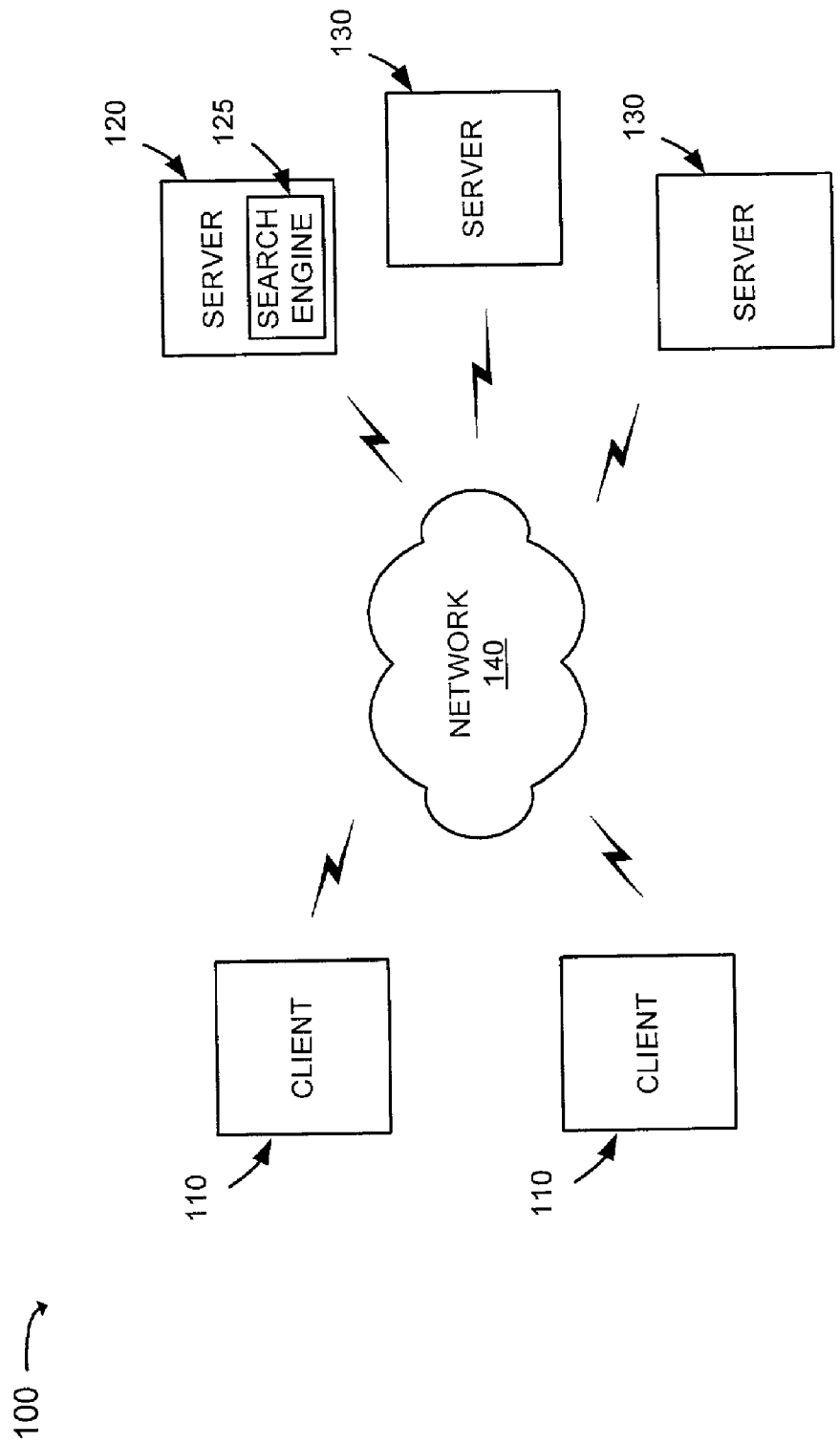

100
120
125
SERVER
SEARCH ENGINE
130
SERVER
130
SERVER
NETWORK
140
110
CLIENT
110
CLIENT

FIG. 1

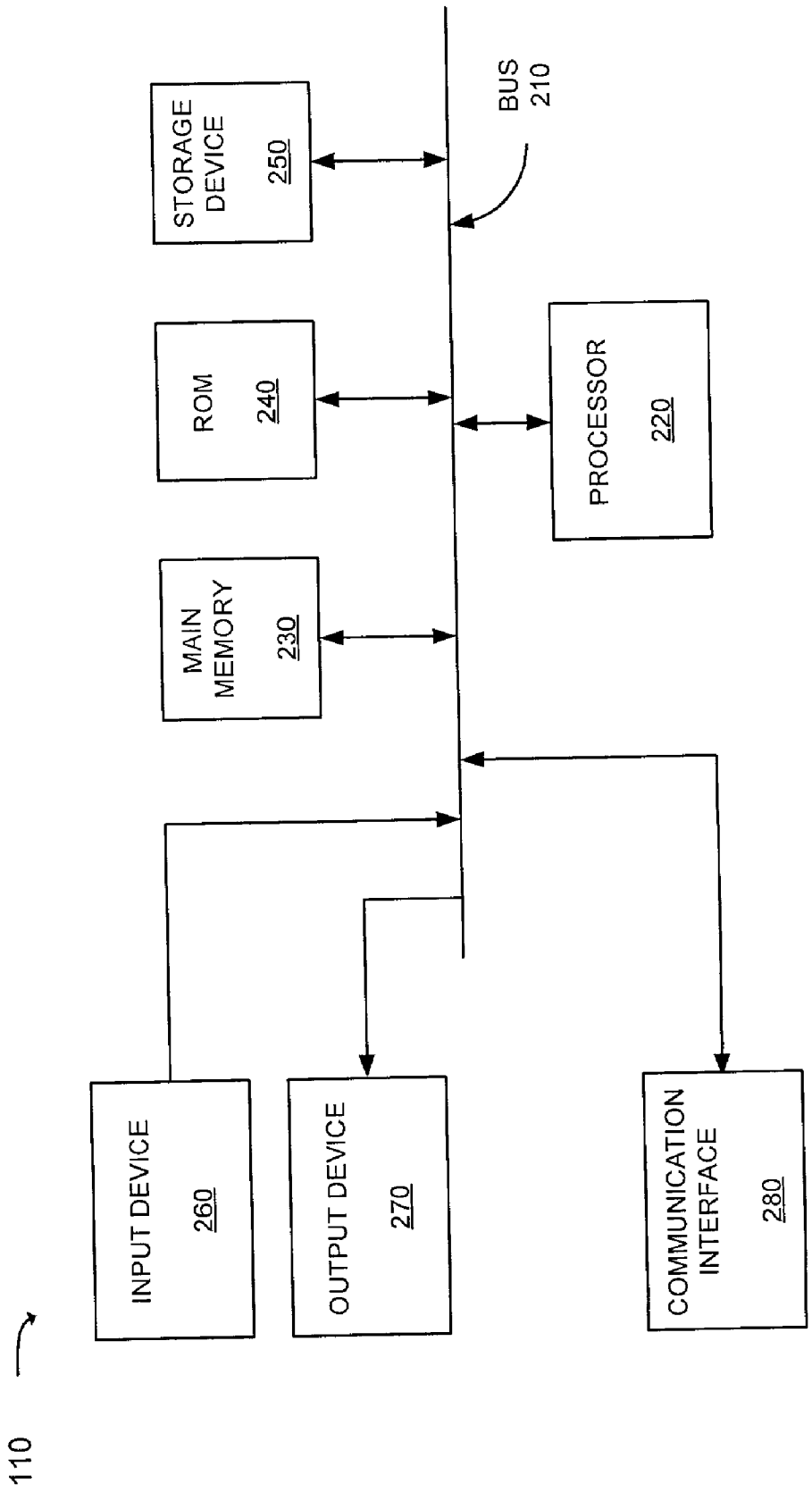

110
BUS
210
STORAGE DEVICE
250
ROM
240
MAIN MEMORY
230
PROCESSOR
220
INPUT DEVICE
260
OUTPUT DEVICE
270
COMMUNICATION INTERFACE
280

FIG. 2

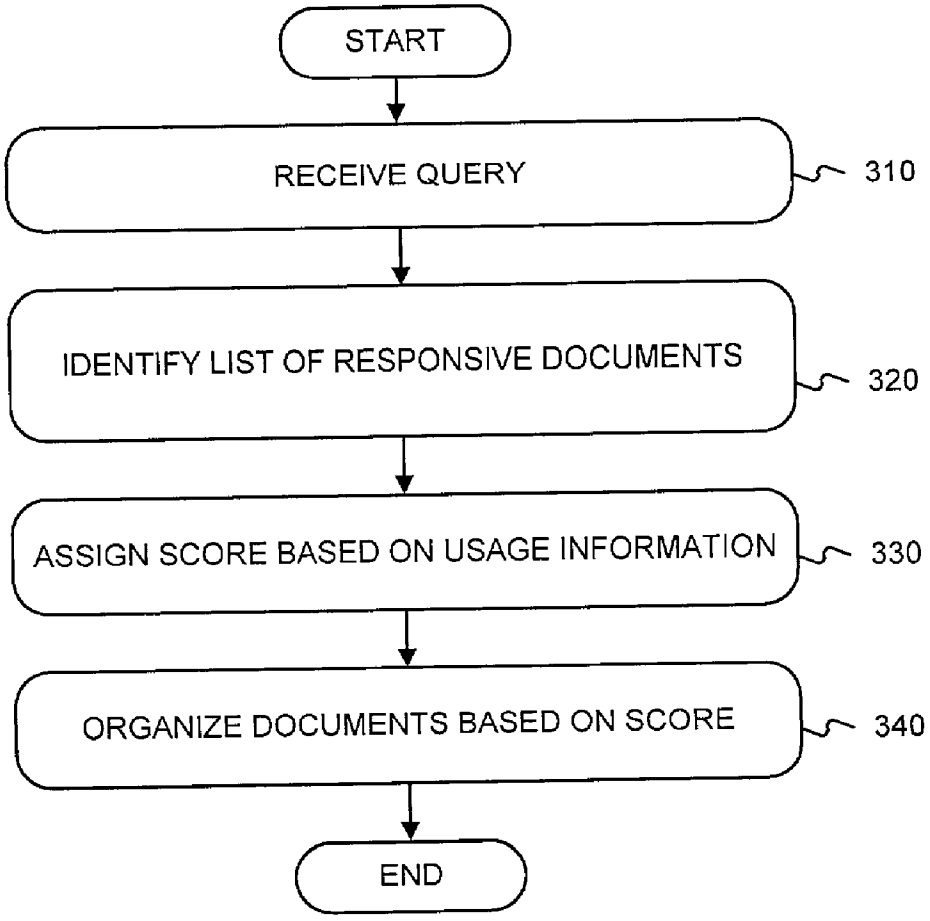

START
RECEIVE QUERY
310
IDENTIFY LIST OF RESPONSIVE DOCUMENTS
320
ASSIGN SCORE BASED ON USAGE INFORMATION
330
ORGANIZE DOCUMENTS BASED ON SCORE
340
END

FIG. 3

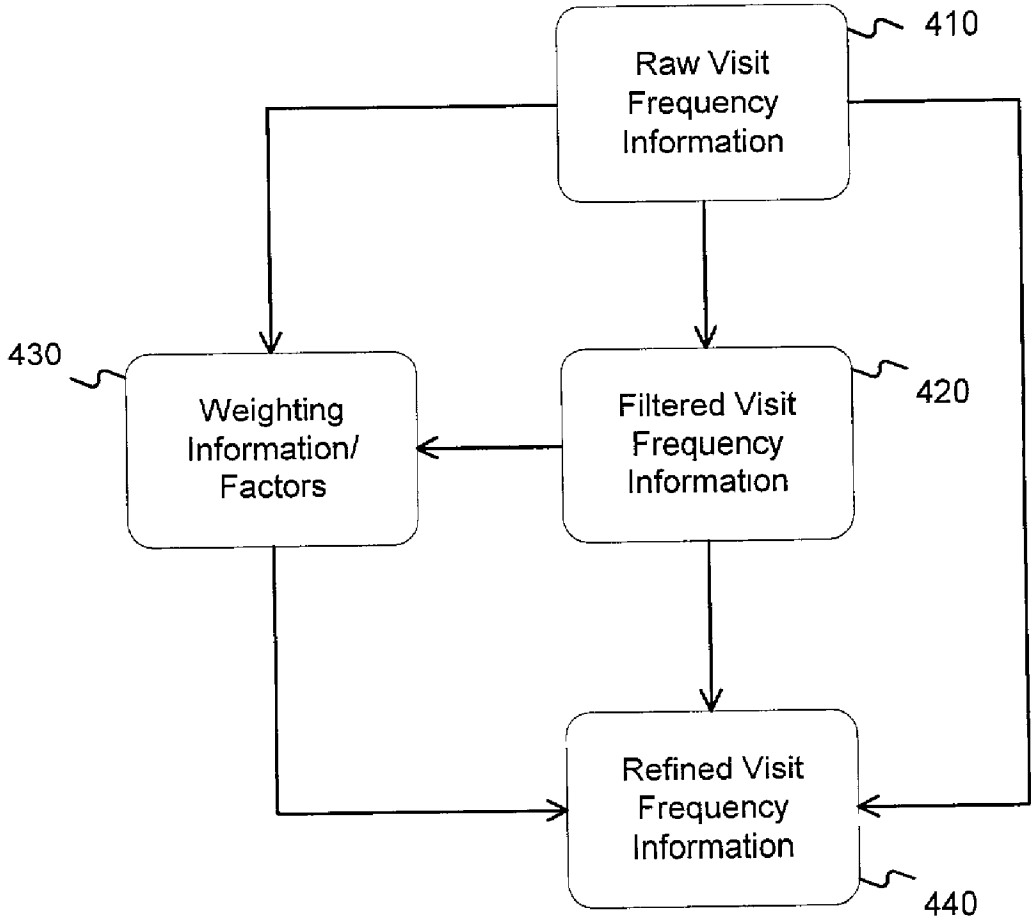

Raw Visit Frequency Information
410
Filtered Visit Frequency Information
420
Weighting Information/ Factors
430
Refined Visit Frequency Information
440

FIG. 4

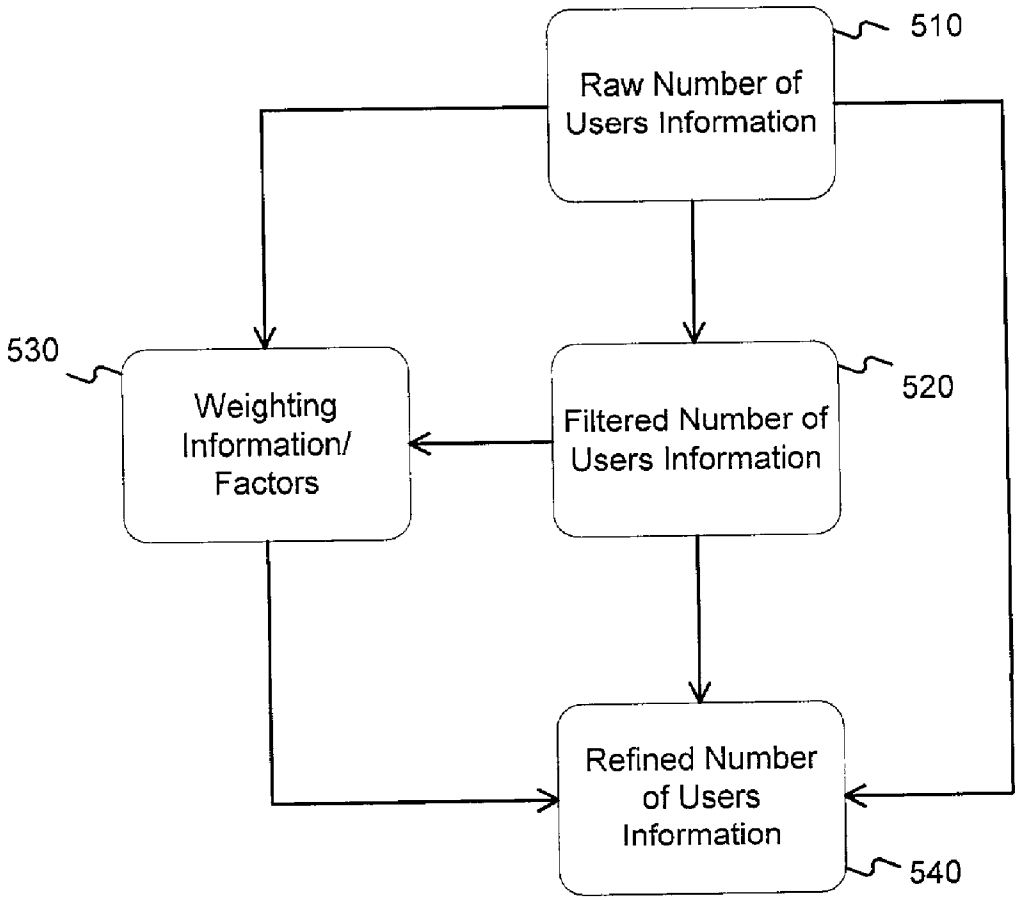

510
Raw Number of Users Information
530
Weighting Information/ Factors
520
Filtered Number of Users Information
Refined Number of Users Information
540

FIG. 5

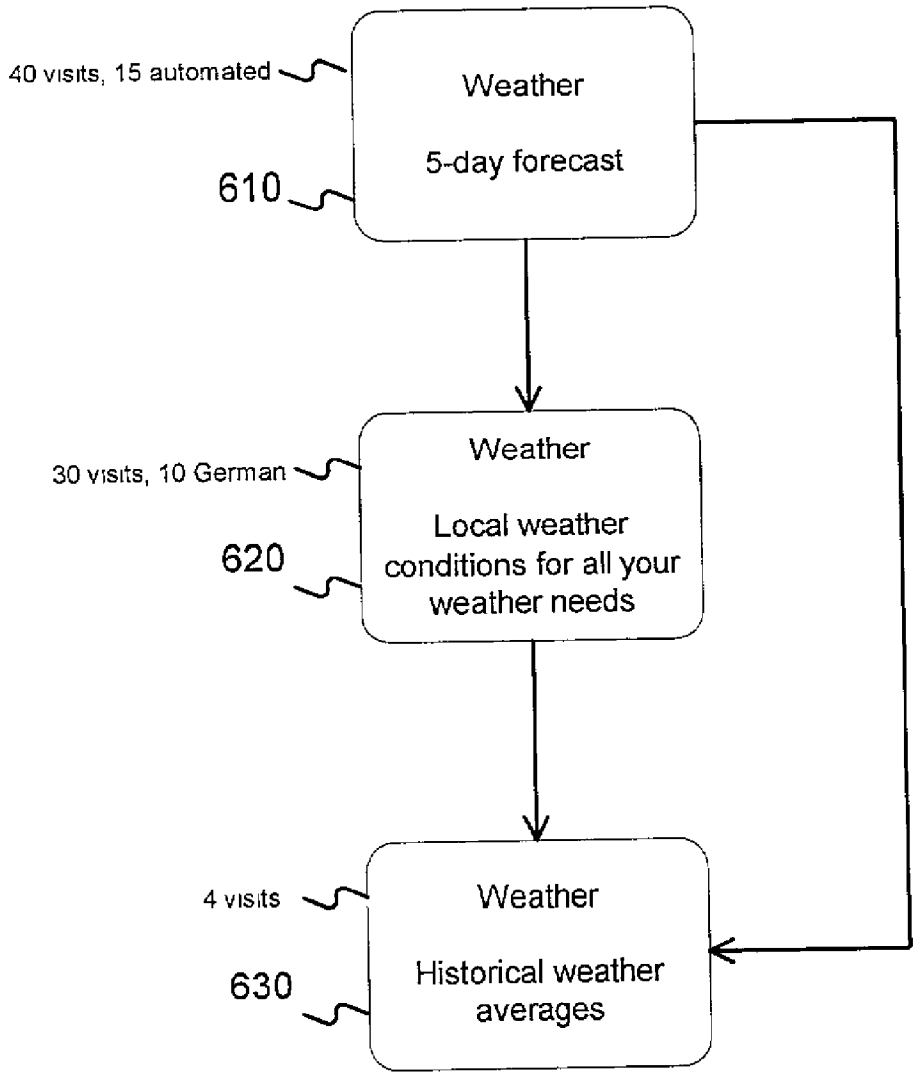

40 visits, 15 automated
Weather
5-day forecast
610
30 visits, 10 German
Weather
Local weather conditions for all your weather needs
620
4 visits
Weather
Historical weather averages
630

FIG. 6

METHODS AND APPARATUS FOR EMPLOYING USAGE STATISTICS IN DOCUMENT RETRIEVAL

BACKGROUND OF THE INVENTION

[0001] A. Field of the Invention

[0002] The present invention relates generally to information search and retrieval and, more particularly, to employing usage data to improve information search and retrieval.

[0003] B. Description of Related Art

[0004] The World Wide Web ("web") contains a vast amount of information. Locating a desired portion of the information, however, can be challenging. This problem is compounded because the amount of information on the web and the number of new users inexperienced at web research are growing rapidly.

[0005] People generally surf the web based on its link graph structure, often starting with high quality human-maintained indices or search engines. Human-maintained lists cover popular topics effectively but are subjective, expensive to build and maintain, slow to improve, and do not cover all esoteric topics.

[0006] Automated search engines, in contrast, locate web sites by matching search terms entered by the user to an indexed corpus of web pages. Generally, the search engine returns a list of web sites sorted based on relevance to the user's search terms. Determining the correct relevance, or importance, of a web page to a user, however, can be a difficult task. For one thing, the importance of a web page to the user is inherently subjective and depends on the user's interests, knowledge, and attitudes. There is, however, much that can be determined objectively about the relative importance of a web page.

[0007] Conventional methods of determining relevance are based on matching a user's search terms to terms indexed from web pages. More advanced techniques determine the importance of a web page based on more than the content of the web page. For example, one known method, described in the article entitled "The Anatomy of a Large-Scale Hypertextual Search Engine," by Sergey Brin and Lawrence Page, assigns a degree of importance to a web page based on the link structure of the web page.

[0008] Each of these conventional methods has shortcomings, however. Term-based methods are biased towards pages whose content or display is carefully chosen towards the given term-based method. Thus, they can be easily manipulated by the designers of the web page. Link-based methods have the problem that relatively new pages have usually fewer hyperlinks pointing to them than older pages, which tends to give a lower score to newer pages.

[0009] There exists, therefore, a need to develop other techniques for determining the importance of documents.

SUMMARY OF THE INVENTION

[0010] Systems and methods consistent with the present invention address this and other needs by identifying compounds based on the overall context of a user query. One aspect of the present invention is directed to a method of organizing a set of documents by receiving a search query and identifying a plurality of documents responsive to the search query. Each identified document is assigned a score based on usage information, and the documents are organized based on the assigned scores.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, explain the invention. In the drawings,

[0012] **FIG. 1** is a diagram illustrating an exemplary network in which concepts consistent with the present invention may be implemented;

[0013] **FIG. 2** illustrates a flow diagram, consistent with the invention, for organizing documents based on usage information;

[0014] **FIG. 3** illustrates a flow chart describing the computation of usage data;

[0015] **FIG. 4** illustrates a few techniques for computing the frequency of visits, consistent with the invention.

[0016] **FIG. 5** illustrates a few techniques for computing the number of users, consistent with the invention; and

[0017] **FIG. 6** depicts an exemplary method, consistent with the invention.

DETAILED DESCRIPTION

[0018] The following detailed description of the invention refers to the accompanying drawings. The detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents.

[0019] A. Overview

[0020] In one embodiment, a search query is received and a list of responsive documents is identified. The list of responsive documents may be based on a comparison between the search query and the contents of the documents, or by other conventional methods. Usage statistics are determined for each document, and the documents are organized based in whole or in part on the usage statistics. These usage statistics may include, for example, the number of visitors to the document (perhaps over a period of time), the frequency with which the document was visited (perhaps over a period of time), or other measures.

[0021] A. Architecture

[0022] **FIG. 1** illustrates a system **100** in which methods and apparatus, consistent with the present invention, may be implemented. The system **100** may include multiple client devices **110** connected to multiple servers **120** and **130** via a network **140**. The network **140** may include a local area network (LAN), a wide area network (WAN), a telephone network, such as the Public Switched Telephone Network (PSTN), an intranet, the Internet, or a combination of networks. Two client devices **110** and three servers **120** and **130** have been illustrated as connected to network **140** for simplicity. In practice, there may be more or less client devices and servers. Also, in some instances, a client device may perform the functions of a server and a server may perform the functions of a client device.

[0023] The client devices **110** may include devices, such mainframes, minicomputers, personal computers, laptops,

personal digital assistants, or the like, capable of connecting to the network **140**. The client devices **110** may transmit data over the network **140** or receive data from the network **140** via a wired, wireless, or optical connection.

[**0024**] **FIG. 2** illustrates an exemplary client device **110** consistent with the present invention. The client device **110** may include a bus **210**, a processor **220**, a main memory **230**, a read only memory (ROM) **240**, a storage device **250**, an input device **260**, an output device **270**, and a communication interface **280**.

[**0025**] The bus **210** may include one or more conventional buses that permit communication among the components of the client device **110**. The processor **220** may include any type of conventional processor or microprocessor that interprets and executes instructions. The main memory **230** may include a random access memory (RAM) or another type of dynamic storage device that stores information and instructions for execution by the processor **220**. The ROM **240** may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by the processor **220**. The storage device **250** may include a magnetic and/or optical recording medium and its corresponding drive.

[**0026**] The input device **260** may include one or more conventional mechanisms that permit a user to input information to the client device **110**, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms, etc. The output device **270** may include one or more conventional mechanisms that output information to the user, including a display, a printer, a speaker, etc. The communication interface **280** may include any transceiver-like mechanism that enables the client device **110** to communicate with other devices and/or systems. For example, the communication interface **280** may include mechanisms for communicating with another device or system via a network, such as network **140**.

[**0027**] As will be described in detail below, the client devices **110**, consistent with the present invention, may perform certain document retrieval operations. The client devices **110** may perform these operations in response to processor **220** executing software instructions contained in a computer-readable medium, such as memory **230**. A computer-readable medium may be defined as one or more memory devices and/or carrier waves. The software instructions may be read into memory **230** from another computer-readable medium, such as the data storage device **250**, or from another device via the communication interface **280**. The software instructions contained in memory **230** causes processor **220** to perform search-related activities described below. Alternatively, hardwired circuitry may be used in place of or in combination with software instructions to implement processes consistent with the present invention. Thus, the present invention is not limited to any specific combination of hardware circuitry and software.

[**0028**] The servers **120** and **130** may include one or more types of computer systems, such as a mainframe, minicomputer, or personal computer, capable of connecting to the network **140** to enable servers **120** and **130** to communicate with the client devices **110**. In alternative implementations, the servers **120** and **130** may include mechanisms for directly connecting to one or more client devices **110**. The servers **120** and **130** may transmit data over network **140** or receive data from the network **140** via a wired, wireless, or optical connection.

[**0029**] The servers may be configured in a manner similar to that described above in reference to **FIG. 2** for client device **110**. In an implementation consistent with the present invention, the server **120** may include a search engine **125** usable by the client devices **110**. The servers **130** may store documents (or web pages) accessible by the client devices **110** and may perform document retrieval and organization operations, as described below.

[**0030**] B. Architectural Operation

[**0031**] **FIG. 3** illustrates a flow diagram, consistent with the invention, for organizing documents based on usage information. At stage **310**, a search query is received by search engine **125**. The query may contain text, audio, video, or graphical information. At stage **320**, search engine **125** identifies a list of documents that are responsive (or relevant) to the search query. This identification of responsive documents may be performed in a variety of ways, consistent with the invention, including conventional ways such as comparing the search query to the content of the document.

[**0032**] Once this set of responsive documents has been determined, it is necessary to organize the documents in some manner. Consistent with the invention, this may be achieved by employing usage statistics, in whole or in part.

[**0033**] As shown at stage **340**, scores are assigned to each document based on the usage information. The scores may be absolute in value or relative to the scores for other documents. This process of assigning scores, which may occur before or after the set of responsive documents is identified, can be based on a variety of usage information. In a preferred implementation, the usage information comprises both unique visitor information and frequency of visit information, as described below in reference to **FIGS. 4 and 5**. The usage information may be maintained at client **110** and transmitted to search engine **125**. The location of the usage information is not critical, however, and it could also be maintained in other ways. For example, the usage information may be maintained at servers **130**, which forward the information to search engine **125**; or the usage information may be maintained at server **120** if it provides access to the documents (e.g., as a web proxy).

[**0034**] At stage **350**, the responsive documents are organized based on the assigned scores. The documents may be organized based entirely on the scores derived from usage statistics. Alternatively, they may be organized based on the assigned scores in combination with other factors. For example, the documents may be organized based on the assigned scores combined with link information and/or query information. Link information involves the relationships between linked documents, and an example of the use of such link information is described in the Brin & Page publication referenced above. Query information involves the information provided as part of the search query, which may be used in a variety of ways to determine the relevance of a document. Other information, such as the length of the path of a document, could also be used.

[**0035**] In one implementation, documents are organized based on a total score that represents the product of a usage score and a standard query-term-based score ("IR score"). In

particular, the total score equals the square root of the IR score multiplied by the usage score. The usage score, in turn, equals a frequency of visit score multiplied by a unique user score multiplied by a path length score.

[0036] The frequency of visit score equals log2(1+log(VF)/log(MAXVF). VF is the number of times that the document was visited (or accessed) in one month, and MAXVF is set to 2000. A small value is used when VF is unknown. If the unique user is less than 10, it equals 0.5*UU/10; otherwise, it equals 0.5*(1+UU/MAXUU). UU is the number of unique hosts/IPs that access the document in one month, and MAXUU is set to 400. A small value is used when UU is unknown. The path length score equals log(K-PL)/log(K). PL is the number of '/' characters in the document's path, and K is set to 20.

[0037] **FIG. 4** illustrates a few techniques for computing the frequency of visits, consistent with the invention. The computation begins with a raw count **410**, which could be an absolute or relative number corresponding to the visit frequency for the document. For example, the raw count may represent the total number of times that a document has been visited. Alternatively, the raw count may represent the number of times that a document has been visited in a given period of time (e.g., 100 visits over the past week), the change in the number of times that a documents has been visited in a given period of time (e.g., 20% increase during this week compared to the last week), or any number of different ways to measure how frequently a document has been visited. In one implementation, this raw count is used as the refined visit frequency **440**, as shown by the path from **410** to **440**.

[0038] In other implementations, the raw count may be processed using any of a variety of techniques to develop a refined visit frequency, with a few such techniques being illustrated in **FIG. 4**. As shown by **420**, the raw count may be filtered to remove certain visits. For example, one may wish to remove visits by automated agents or by those affiliated with the document at issue, since such visits may be deemed to not represent objective usage. This filtered count **420** may then be used to calculate the refined visit frequency **440**.

[0039] Instead of, or in addition to, filtering the raw count, the raw count may be weighted based on the nature of the visit (**430**). For example, one may wish to assign a weighting factor to a visit based on the geographic source for the visit (e.g., counting a visit from Germany as twice as important as a visit from Antarctica). Any other type of information that can be derived about the nature of the visit (e.g., the browser being used, information concerning the user, etc.) could also be used to weight the visit. This weighted visit frequency **430** may then be used as the refined visit frequency **440**.

[0040] Although only a few techniques for computing the visit frequency are illustrated in **FIG. 4**, those skilled in the art will recognize that there exist other ways for computing the visit frequency, consistent with the invention.

[0041] **FIG. 5** illustrates a few techniques for computing the number of users, consistent with the invention. As with the techniques for computing visit frequency illustrated in **FIG. 4**, the computation begins with a raw count **510**, which could be an absolute or relative number corresponding to the number of users who have visited the document. Alternatively, the raw count may represent the number of users that have visited a document in a given period of time (e.g., 30 users over the past week), the change in the number of users that have visited the document in a given period of time (e.g., 20% increase during this week compared to the last week), or any number of different ways to measure how many users have visited a document. The identification of the users may be achieved based on the user's Internet Protocol (IP) address, their hostname, cookie information, or other user or machine identification information. In one implementation, this raw count is used as the refined number of users **540**, as shown by the path from **510** to **540**.

[0042] In other implementations, the raw count may be processed using any of a variety of techniques to develop a refined user count, with a few such techniques being illustrated in **FIG. 5**. As shown by **520**, the raw count may be filtered to remove certain users. For example, one may wish to remove users identified as automated agents or as users affiliated with the document at issue, since such users may be deemed to not provide objective information about the value of the document. This filtered count **520** may then be used to calculate the refined user count **540**.

[0043] Instead of, or in addition to, filtering the raw count, the raw count may be weighted based on the nature of the user (**530**). For example, one may wish to assign a weighting factor to a visit based on the geographic source for the visit (e.g., counting a user from Germany as twice as important as a user from Antarctica). Any other type of information that can be derived about the nature of the user (e.g., browsing history, bookmarked items, etc.) could also be used to weight the user. This weighted user information **530** may then be used as the refined user count **540**.

[0044] Although only a few techniques for computing the number of users are illustrated in **FIG. 5**, those skilled in the art will recognize that there exist other ways for computing the number of users, consistent with the invention. Similarly, although **FIGS. 4 and 5** illustrate two types of usage information that may be used to organize documents, those skilled in the art will recognize that there exist other such type of information and techniques consistent with the invention.

[0045] Furthermore, although **FIGS. 4 and 5** illustrate determining usage information on a document-by-document basis, other techniques consistent with the information may be used to associate usage information with a document. For example, rather than maintaining usage information for each document, one could maintain usage information on a site-by-site basis. This site usage information could then be associated with some or all of the documents within that site.

[0046] **FIG. 6** depicts an exemplary method employing visit frequency information, consistent with the invention. **FIG. 6** depicts three documents, **610**, **620**, and **630**, which are responsive to a search query for the term "weather." Document **610** is shown to have been visited 40 times over the past month, with 15 of those 40 visits being by automated agents. Document **620**, which is linked to from document **610**, is shown to have been visited 30 times over the past month, with 10 of those 30 visits coming from Germany. Document **630**, which is linked to from documents **610** and **620**, is shown to have been visited 4 times over the past month.

[0047] Under a conventional term frequency based search method, the documents may be organized based on the frequency with which the search query term ("weather") appears in the document. Accordingly, the documents may be organized into the following order: **620** (three occurrences of "weather"), **630** (two occurrences of "weather'), and **610** (one occurrence of "weather').

[0048] Under a conventional link-based search method, the documents may be organized based on the number of other documents that link to those documents. Accordingly, the documents may be organized into the following order: **630** (linked to by two other documents), **620** (linked to by one other document), and **610** (linked to by no other documents).

[0049] Methods and apparatus consistent with the invention employ usage information to aid in organizing documents. Based purely on raw visit frequency, the documents may be organized into the following order: **610** (40 visits), **620** (30 visits), and **630** (4 visits). If these raw visit frequency number are refined to filter automated agents and to assign double weight to visits from Germany, the documents may be organized in the following order: **620** (effectively 40 visits, since the 10 from Germany count double), **610** (effectively 25 visits after filtering the 15 visits from automated agents), and **630** (effectively 4 visits).

[0050] Instead of using the usage information alone, the usage information may be used in combination with the query information and/or the link information to develop the ultimate organization of the documents.

[0051] C. Conclusion

[0052] The foregoing description of preferred embodiments of the present invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. For example, although the preceding description generally discussed the operation of search engine **125** in the context of a search of documents on the world wide web, search engine **125** could be implemented on any corpus.

[0053] The scope of the invention is defined by the claims and their equivalents.

What is claimed:

1. A computer implemented method of organizing a collection of documents by employing usage information, comprising:

receiving a search query;

identifying a plurality of documents responsive to the search query;

assigning a score to each document based on at least the usage information; and

organizing the documents based on the assigned scores.

2. The method of claim 1, wherein the documents are hyperlinked pages from the world wide web.

3. The method of claim 1, wherein the usage information for a document comprises the number of users who have visited the document.

4. The method of claim 3, wherein the usage information for a document comprises the change, over a period of time, in the number of users who have visited the document.

5. The method of claim 3, wherein the usage information for a document excludes certain predefined users.

6. The method of claim 3, wherein the usage information for a document is weighted based on the nature of user.

7. The method of claim 1, wherein the usage information for a document comprises the frequency with which the document has been visited.

8. The method of claim 7, wherein the usage information for a document comprises the change, over a period of time, in the frequency with which the document has been visited.

9. The method of claim 7, wherein the usage information for a document excludes certain predefined visits.

10. The method of claim 7, wherein the usage information for a document is weighted based on the nature of the visit.

11. The method of claim 1, wherein the usage information for a document comprises a combination of unique visitors to the document and a frequency with which the document has been visited.

12. The method of claim 1, wherein the usage information is stored at a server that provides access to the documents.

13. The method of claim 1, wherein the usage information is stored at a client that accesses the documents.

14. The method of claim 1, wherein the score assigned to a document is relative to the score assigned to other documents.

15. The method of claim 1, wherein the score assigned to a document is an absolute score.

16. A computer implemented method of organizing a collection of documents by employing usage information, comprising:

receiving a search query;

identifying a plurality of documents responsive to the search query; and

organizing the documents based on at least the usage information.

17. The method of claim 16, wherein the usage information for a document comprises the number of unique visitors to the document.

18. The method of claim 16, wherein the usage information for a document comprises the frequency with which the document has been visited.

19. The method of claim 16, wherein the documents are organized based on the usage information alone.

20. The method of claim 16, further comprising organizing the documents based on the usage information and the search query.

21. The method of claim 16, wherein the documents contain link information.

22. The method of claim 21, further comprising organizing the documents based on the usage information and the link information.

23. The method of claim 16, further comprising organizing the documents based on the usage statistics, the search query, and the link information.

24. The method of claim 16, wherein the usage information for a document is based on the usage information for the site to which the document belongs.

25. A computer-readable medium containing one or more instructions for organizing a collection of documents, the instructions comprising:

receiving a search query;

identifying a plurality of documents responsive to the search query;

assigning a score to each document based on the usage information; and

organizing the documents based on the assigned scores.

26. An apparatus for organizing a collection of documents, comprising:

at least one memory having program instructions, and

at least one processor configured to execute the program instructions to perform the operations of:

receiving a search query;

identifying a plurality of documents responsive to the search query;

assigning a score to each document based on the usage information; and

organizing the documents based on the assigned scores.

27. A machine-readable medium having stored thereon a plurality of records, each of the records comprising:

a) a first field containing a document identifier; and

b) a second field containing a value corresponding to the importance of the document, the value being a function of both usage data for the document and link information for the document.

* * * * *