US 20040267815A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2004/0267815 A1**

De Mes (43) **Pub. Date: Dec. 30, 2004**

(54) **SEARCHABLE PERSONAL BROWSING HISTORY**

(76) Inventor: **Arjan De Mes**, Leiden (NL)

Correspondence Address:
IBM CORPORATION
IPLAW IQ0A/40-3
1701 NORTH STREET
ENDICOTT, NY 13760 (US)

(21) Appl. No.: **10/809,575**

(22) Filed: **Mar. 25, 2004**

(30) **Foreign Application Priority Data**

Jun. 25, 2003 (GB) 0314782.4

Publication Classification

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

(52) **U.S. Cl.** **707/104.1**

(57) **ABSTRACT**

A system, method and program product for creating a searchable personal browsing history. In response to a user request to obtain a web page from the Internet, metadata and textual data are automatically extracted from the web page. Then, the extracted metadata and textual data are indexed and stored. Subsequently, the stored metadata and textual data are displayed in categories based on the indexing, to enable searching of the displayed categories of metadata and textual data.

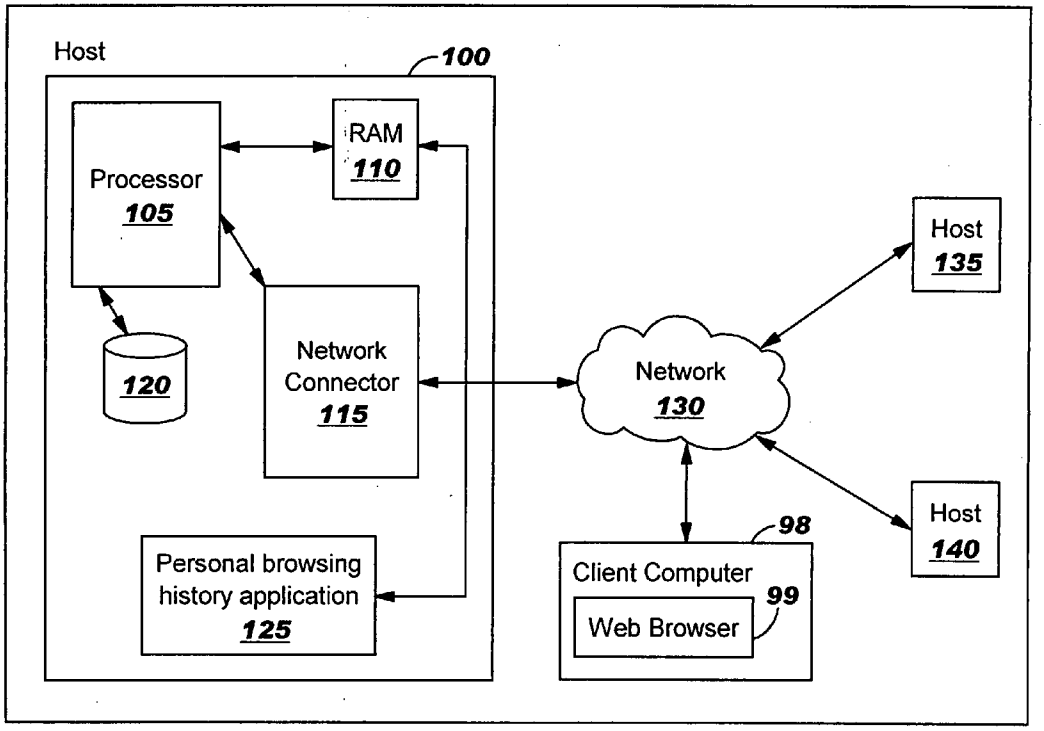


FIG. 1

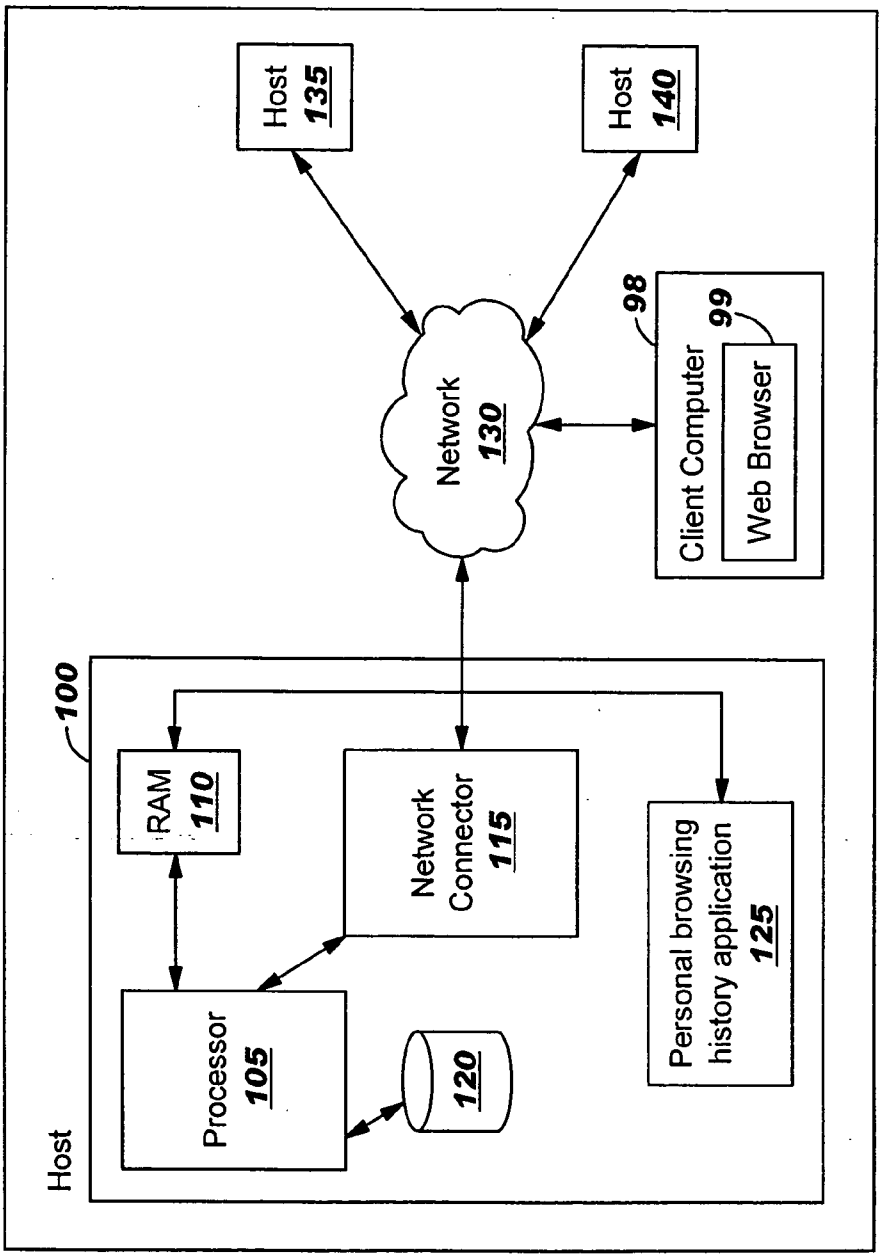

Host
135
Host
140
Network
130
98
Client Computer
99
Web Browser
100
RAM
110
Network
Connector
115
Personal browsing
history application
125
Processor
105
120
Host

FIG. 2

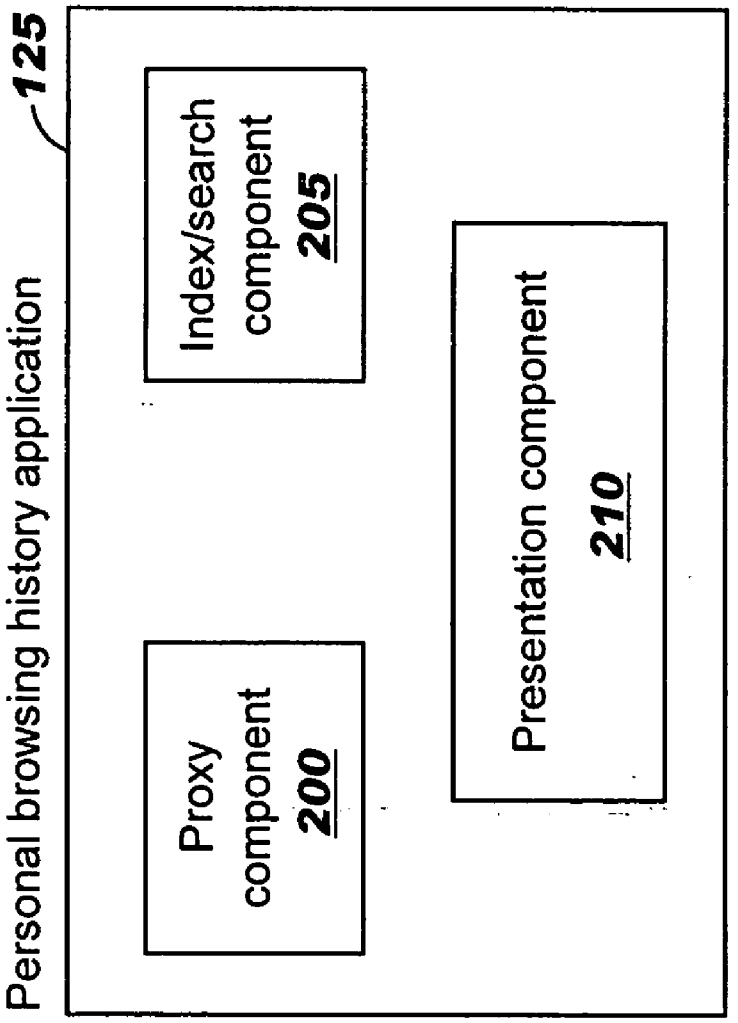

125
Personal browsing history application
Index/search component 205
Presentation component 210
Proxy component 200

FIG. 3

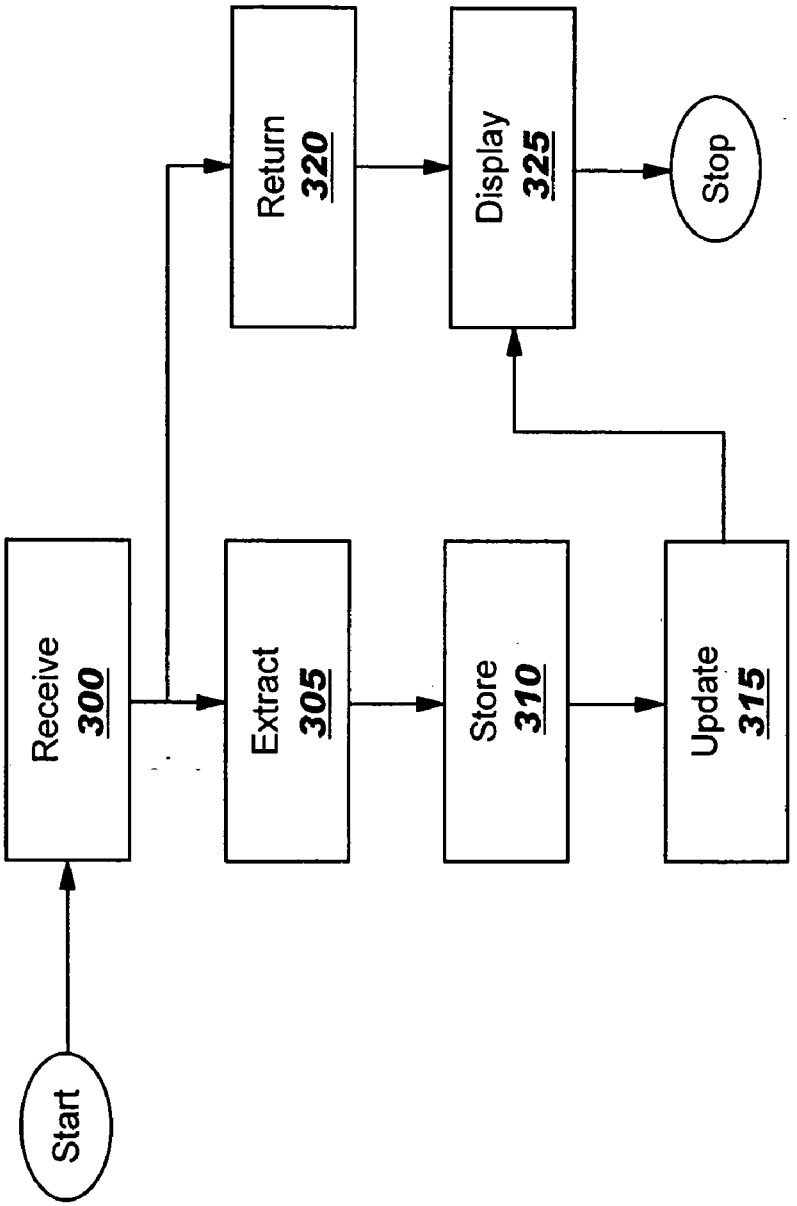

Start
Receive
300
Extract
305
Store
310
Update
315
Return
320
Display
325
Stop

FIG. 4

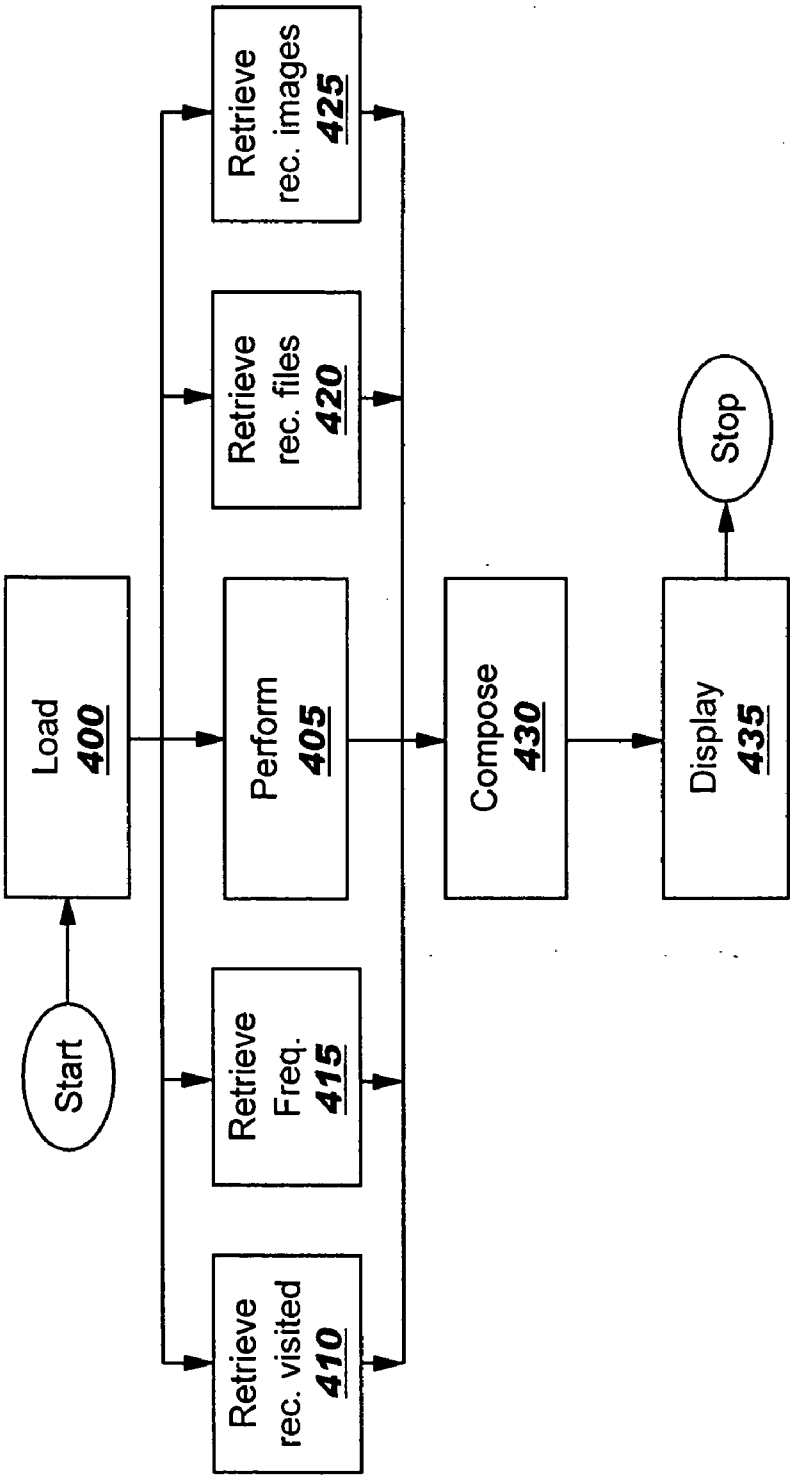
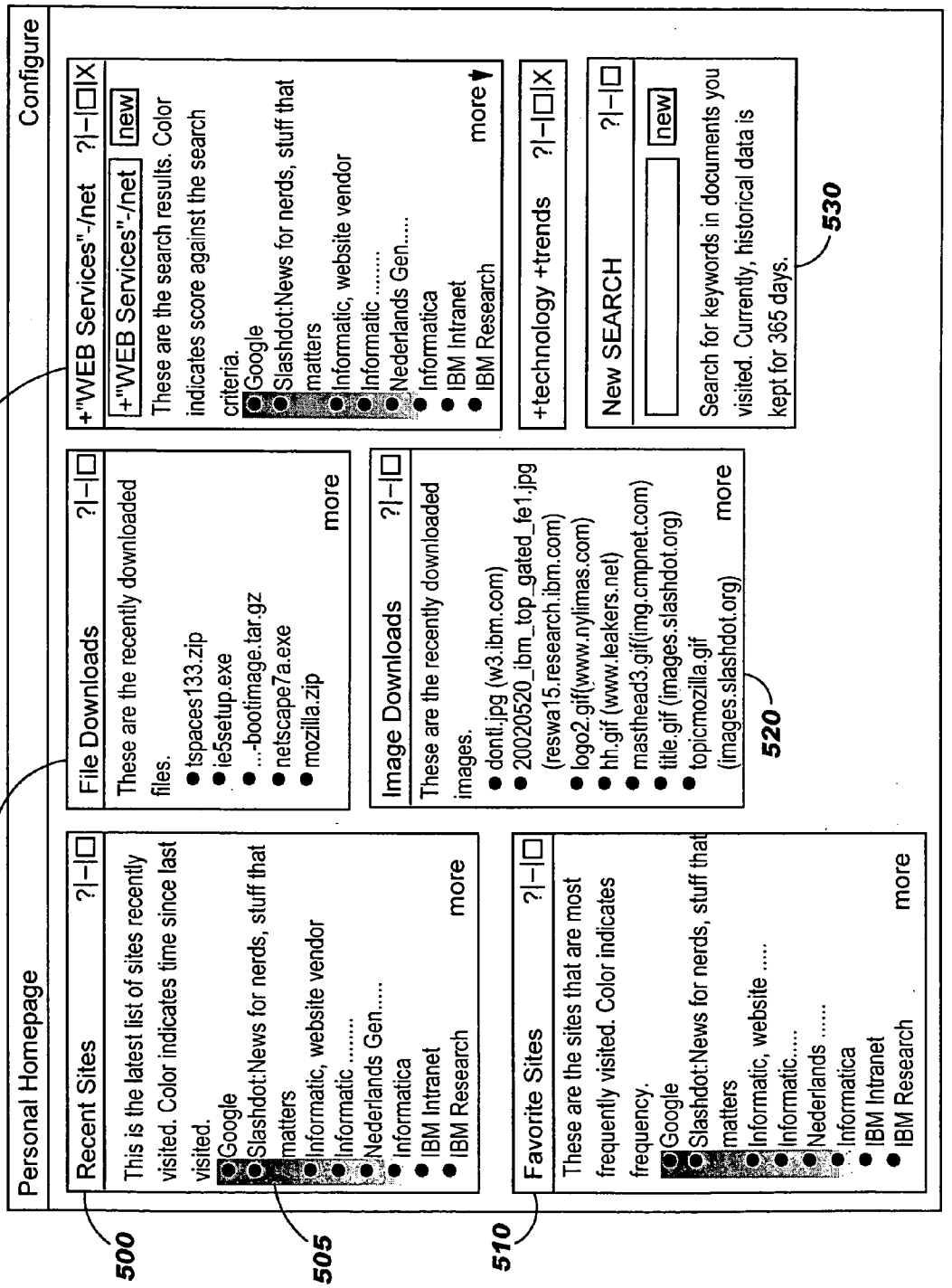

Start
Load
400
Retrieve
rec. visited
410
Retrieve
Freq.
415
Retrieve
rec. files
420
Retrieve
rec. images
425
Perform
405
Compose
430
Display
435
Stop

FIG. 5


Personal Homepage
Configure
Recent Sites
This is the latest list of sites recently visited. Color indicates time since last visited.
Google
Slashdot:News for nerds, stuff that matters
Informatic, website vendor
Informatic .....
Nederlands Gen.....
Informatica
IBM Intranet
IBM Research
more
File Downloads
These are the recently downloaded files.
tspaces133.zip
ie5setup.exe
...-bootimage.tar.gz
netscape7a.exe
mozilla.zip
more
Image Downloads
These are the recently downloaded images.
dontl.jpg (w3.ibm.com)
20020520_ibm_top_gated_fe1.jpg (reswa15.research.ibm.com)
logo2.gif(www.nytimas.com)
hh.gif (www.leakers.net)
masthead3.gif(img.cmpnet.com)
title.gif (images.slashdot.org)
topicmozilla.gif (images.slashdot.org)
more
+"WEB Services"-/net
+"WEB Services"-/net
new
These are the search results. Color indicates score against the search criteria.
Google
Slashdot:News for nerds, stuff that matters
Informatic, website vendor
Informatic .....
Nederlands Gen.....
Informatica
IBM Intranet
IBM Research
more
+technology +trends
New SEARCH
new
Search for keywords in documents you visited. Currently, historical data is kept for 365 days.
Favorite Sites
These are the sites that are most frequently visited. Color indicates frequency.
Google
Slashdot:News for nerds, stuff that matters
Informatic, website .....
Informatic.....
Nederlands .....
Informatica
IBM Intranet
IBM Research
more
500
505
510
515
520
525
530

SEARCHABLE PERSONAL BROWSING HISTORY

FIELD OF THE INVENTION

[0001] The invention relates generally to computer systems and deals more particularly with a tool for tracking web browsing.

BACKGROUND OF THE INVENTION

[0002] The World Wide Web (WWW) has evolved into a very useful tool for banking, shopping, booking hotels, rental cars and airline tickets, checking stock prices and searching for other types of information. The WWW comprises a vast multitude of individual webpages and files, and it is difficult to remember which web pages have been previously visited. Consider an example of searching the WWW using Google (Google is a registered trademark of Google Technology Inc) or Yahoo (Yahoo is a registered trademark of Yahoo! Inc.) search engine for a topic such as knowledge management. The search engine displays the results as a list of titles and hyperlinks to knowledge management websites. If the user selects a particular hyperlink from the search results a corresponding web page is displayed. Embedded within this web page may be other hyperlinks which direct a user to other knowledge management web pages which may or may not be of interest to the user. Once the user has found the web page with the information that he or she needs, the user can either print, download or bookmark the web page for future reference. However, a problem may occur later when the user tries to locate a web page which the user did not save, print or download this web page. In such a case, the users may resort to another search to attempt to find the same comparable web page.

[0003] It is known to cache web pages for later use. Most web browsers maintain in the client computer's local file system a cache of recently visited web pages and other web resources. Before displaying them in the web browser, an HTTP request is used to check with the original server that the cached web pages are the most current pages available. However, a web browser cache suffers the disadvantage that it is not well controlled and temporary in nature. It also requires periodic scanning/indexing in order for the information stored in the cache to be of any use to a user. Further, some web pages are never placed in the cache. Therefore the cache does not give a full indication of the web pages or web resources that a user has accessed over a particular period of time.

[0004] Another method of storing recently visited web pages is to save the web pages for off-line viewing. This facility is offered in current versions of Microsoft Internet Explorer. To save a visited web page for off line viewing, a user can bookmark the web page currently being accessed. Microsoft Internet Explorer provides a "wizard" which presents the user with a number of options to customise the content for off line viewing. A disadvantage with the foregoing approach is that a user has to actively select the web pages to be bookmarked.

[0005] Another approach can be found in a paper written by Manber U et al (to appear in 1997 Usenix Technical Conference . . . , Jan. 6-10, 1997), (web reference http://webglimpse.org/pubs/webglimpse/pdf) from the Department of Computer Science, University of Arizona, Tucson. The paper discusses a tool called WebGlimpse which analyses collections of webpages. WebGlimpse analyses a given WWW archive for example a website, a collection of specific documents or a private history cache and computes neighborhoods i.e. the most relevant documents according to a user's specification. Once this has been completed, search boxes are added to selected pages, remote pages are collected if relevant and the pages are cached locally. Users are able to browse the website using any of the added search boxes. A disadvantage of this approach is that a user has to actively indicate to WebGlimpse that the user wishes to archive a particular website or a particular web page. Also, if a user later wants to locate a web page seen earlier, and the web page has not been archived, the user still must try to retrace his or her steps using their preferred search engine.

[0006] Yet another approach is discussed in a paper entitled 'Lifestreams: organising your electronic life' written by Freeman, E et al, from the department of Computer Science, Yale University, New Haven, United States. This paper describes a system which provides a time ordered stream of documents which functions as a diary of a persons electronic life. The paper describes creating a time ordered stream of documents starting with a person's electronic birth certificate. The time-ordered document stream moves toward the present day with more current documents that the user has added to the time-ordered document stream. A disadvantage of this approach is that a user must actively create a document which is subsequently added to the time-ordered document stream. Also, this approach is not suitable for saving web pages for off-line viewing because the user is required to actively indicate which web pages are to be saved.

[0007] An object of the present invention is to provide an improved method and system for storing web pages and other web resources accessed by a user.

[0008] Another object of the present invention is to provide a method and system of the foregoing type which also presents the accessed web resources to the user in a meaningful way.

SUMMARY

[0009] The invention resides in a system, method and program product for creating a searchable personal browsing history. In response to a user request to obtain a web page from the Internet, metadata and textual data are automatically extracted from the web page. Then, the extracted metadata and textual data are indexed and stored. Subsequently, the stored metadata and textual data are displayed in categories based on the indexing, to enable searching of the displayed categories of metadata and textual data.

[0010] In accordance with a feature of the present invention, the user does not have to actively select that a data resource should be saved. Thus, the present invention provides an accurate account of the data resources accessed over a communications network by the user. The user may define the types of categories to be displayed in the searchable personal browsing history thereby personalising the data displayed. Further, a user may search the searchable personal browsing history and thereby create a view within the searchable personal browsing history defined by the search results and one or more user defined categories.

[0011] In accordance with another feature of the present invention, the extracted metadata and textual data are stored with a reference to the data resource's original location. This avoids need for a complete copy of the data resource to be stored in a data store.

[0012] In accordance with another feature of the present invention, a calculation is performed on the extracted metadata to create statistical information relating to a user's browsing activity. An advantage of this approach is that a user is able to view his or her browsing activity in categorised views which provides efficient access to the required information. Preferably the calculated statistical information provides a user with categories of recently visited web pages, most frequently visited web pages, recently visited downloads and/or recently visited images.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] **FIG. 1** illustrates a computer system in which is executed the personal browsing history application program according to the present invention.

[0014] **FIG. 2** illustrates program components of the personal browsing history application program of **FIG. 1**.

[0015] **FIG. 3** is a flowchart illustrating entry of historical web browsing data into the personal browsing history application program of **FIG. 2**.

[0016] **FIG. 4** is a flowchart illustrating operation of the personal browsing history application program of **FIG. 2** when generating a display of a personal browsing history.

[0017] **FIG. 5** is an example of a display screen showing a user's personal browsing history generated according to the steps of **FIG. 4**.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0018] **FIG. 1** is a block diagram of a computer system in which the present invention may be incorporated. A client/server data processing host computer **100** is connected to other client/server data processing host computers **135** and **140** via a network **130** such as the Internet. Client/server data processing host **100** includes a processor **105** for executing programs that control the operation of the client/server data processing host **100**, a RAM volatile memory **110**, a non-volatile memory **120**, and a network connector **115** for use in interfacing with the network **130** for communication with the other client/server hosts **135** and **140**. **FIG. 1** also illustrates a client computer **98** with a web browser **99** for accessing hosts **100**, **135** and **140**. In an alternate embodiment of the present invention, client computer **98** resides on an intranet (not shown) to enable connection to host **100**. Host computer **100** also includes a personal browsing history application program **125** according to the present invention.

[0019] Program **125** may be deployed as a standalone client application interfacing with a user's web browser **99** of a user's client computer **98**. Program **125** accesses, over network **130**, data resources requested from client/server data processing hosts **135** and **140**. Alternatively, the personal history application program **125** may be deployed as a server application on client/server data processing hosts **135** or **140** where the client/server data processing host **100** can access the personal history application **125** via the communication network **130**. For the remainder of this patent application, the personal browsing history application program **125** will be described as being deployed as a client application on the client/server data processing host **100** and accessing over communication network **130**, a plurality of data resources requested from client/server data processing hosts (herein referred to as a web server) **135** and **140**.

[0020] **FIG. 2** illustrates the program components of the personal browsing history application program **125**—a proxy program component **200**, a search/index program component **205** and a presentation program component **210**. The proxy component **200** causes the personal browsing history application **125** to keep a local representation of recently accessed data resources. These data resources may be web pages, graphics, downloads or any other resource that are accessed over the network **130**. The proxy component **200** also determines, on receipt of a request for a data resource, whether server **100** can handle the request itself or if another proxy server must be contacted to handle or assist in handling the request for the data resource. The latter situation can occur in a corporate environment where requests for data resources outside of the corporate Intranet are configured to be sent to a proxy server before allowing access to the Internet. If the proxy component **200** determines that it can handle the request for a data resource directly, the proxy component **200** accesses the network **130** and contacts the web server **135** or **140** to provide the data resource. The web server **135** or **140** sends the request back to the proxy component **200** residing on the host **100**. Once the request is received by the proxy component **200**, the request is sent to the user's browser and the index/search component **205** automatically begins to process the data resource. The storing of a representation of an accessed data resource requires no active input from the user, it is carried out automatically by the index/search component **205** when the proxy component **200** inspects each accessed data resource.

[0021] The index/search component **205** extracts metadata and textual data from a data resource and indexes the extracted data to form a textual index for searching. In the preferred embodiment of the present invention, this extraction is based on a known mark up language such as HTML. HTML is used to specify the formatting, the presentation and the text and images that comprise the contents of a web page. A typical piece of HTML tagging is as follows:

[0022] <html>

[0023] <head>

[0024] <meta name="keywords" content="corporate home page"/>

[0025] <title>My Company</title>

[0026] </head>

[0027] <body TEXT="000000" BGCOLOR="FFFFFF" leftmargin=0 topmargin=0 marginwidth=0 marginheight=0> The body tag specifies how to display the text and graphics to a user.

[0028] <h1>This is a heading tag </h1>

[0029] The start of a new paragraph

[0030] </body>

[0031] </html>

[0032] When the index/search component **205** receives a data resource such as a web page from the proxy component **200**, the index/search component traverses each of the html tags and extracts metadata and textual data from the data resource. Examples of the metadata are the URL of the web page, the last modified date, fields specified as metadata in the HTML, the title of the web page, and the amount of text on the web page specfied in a word count. The textual data, i.e. the natural language information embedded in the web page between a body tag (<body></body>) is also extracted. Both metadata and textual data are stored with a reference to the original location of the data resource. The reference to the original location of the data resource may comprise an HTTP request or other appropriate protocol.

[0033] The presentation program component **210** displays a searchable personal browsing history created by the personal history application **125**, as described in more detail below with reference to **FIG. 4**.

[0034] **FIG. 3** illustrates how the personal browsing history application **125** operates when accessing a network **130** such as the Internet. At step **300** the user accesses the network (for example, requests a web page) using the personal browsing history application **125** configured to work with the user's browser. A web page or other web resource such as a downloadable file or graphic image may be accessed in the normal manner by entering in a Uniform Resource Locator (URL) into the URL address input box in the user's browser. The browser sends a request message for the web page or other web resource to the proxy component **200**, and the proxy component **200** determines whether it can handle the request itself or whether another proxy server must handle the request. If the proxy component **200** can handle the request itself, a request for a data resource is sent through the network **130** to the web server **135** or **140** depending on which web server can provide the requested data resource specified by the URL. In response to the request, the web server **135** or **140** looks up the path name of the requested data resource and sends back the data resource in a reply message through the network **130** to the personal browsing history application **125**. At step **320** the proxy component **200** forwards the requested resource to the web browser, where it is loaded into the browser window and displayed to the user at step **325**. At step **305** the index/search component **210** extracts metadata and textual data from the contents of the data resource as described previously. As described below, the metadata and the textual data extracted by the index/search component **210** are used to dynamically create a searchable personal browsing history which represents the user's browsing activity when accessing data resources over network **130**. The metadata and the textual data extracted in step **305** are stored in a data store at step **310**. At step **315** the stored metadata and textual data are indexed (as described below with reference to **FIG. 5**) to reflect any recently stored metadata and textual data in step **310**. A reference to the data resource's original location is also stored at step **310** such that the extracted metadata and the textual data create a textual index along with a reference to the data resource's original web location. Each time the proxy component **200** receives a requested resource, the textual index is updated to reflect the addition of a new data resource. The stored metadata and textual data are indexed each time a data resource is accessed over the network **130** thereby allowing the user to constantly view and search the data resources that they have accessed.

[0035] Step **320** is carried out in parallel with steps **305**, **310**, and **315**. In step **320**, the requested data resource is supplied to the browser and displayed to the user at step **325**. The above steps allow the personal history browsing application **125** to work in the background, constantly extracting, storing and re-indexing the extracted metadata and textual data, while the user is browsing the WWW.

[0036] Consider now how the personal browsing history may be used. A user may vaguely remember a web page or other web resource that he or she read some time ago, but not remember where the web page or other web resource is located. As illustrated in **FIG. 4**, a user can locate a data resource that the user had previously accessed by first loading the presentation component **210** from a menu option within the user's web browser. Then, the user's browser sends a request to the proxy component **200** to initiate the searchable personal browsing history. In response, the proxy component **200** loads the presentation component into the user's browser to display the searchable personal browsing history. At step **400** the proxy component **200** loads the custom user settings for the searchable personal browsing history. The user settings define information about how the user would prefer the searchable personal browsing history to be personalised. The user settings are defined in a user profile and may be modified at any time by the user. The user settings consist of information such as which sections may be displayed in the presentation component **210**, access rights of others to the personal history application **125** and password settings. Usability settings may include the color of the text to be displayed in the presentation component within the user's browser when viewing the searchable personal browsing history.

[0037] The metadata and textual data that was extracted from the accessed data resource at step **305** of **FIG. 3** are retrieved from the data store. The metadata is used to calculate statistical information on the activity of the user accessing over network **130** a plurality of data resources. The type of calculations that may be performed enable the determination of the most recently visited web pages at step **410**, the most frequently visited web pages at step **415**, the most recently downloaded files by the user at step **420**, and the most recently downloaded images by the user at step **425**. Thus, the statistical information allows a user to see his or her past browsing activity categorised by the type of calculation performed. At step **405** the user is able to perform a key word search in the index of the stored metadata and textual data. The keyword search is performed by typing search criteria into a search input box. The index/search component **205** uses the search criteria to locate and retrieve the information requested by the user. At step **430** the personal browsing history application **125** creates a searchable personal browsing history which is tailored to the search results, the statistical information and the configuration settings as defined by the user and displayed at step **435**. The searchable browsing history may contain the results of multiple searches (iterations of step **405**) and their results.

[0038] **FIG. 5** illustrates a searchable personal browsing history as generated by the personal browsing history application **125** and displayed in step **435** of **FIG. 4**. The

searchable personal browsing history is a dynamic view changing each time the user performs a new search on the index in step **405** of **FIG. 4** or accesses over a network **130** one or more data resources. The searchable personal browsing history comprises several different sections, recently visited sites **500**, favorite sites **510**, downloaded files **515**, image downloads **520** and search sections **525** and **530** for inputing search criteria. In the search section **525**, the example search criteria shown are '+"web services" -.net'. The searchable personal browsing history locates within the indexed data, all references to "web services" and scores the results according to the most relevant. The scoring is displayed to the user by a color gradient bar **505**, the higher the score the more intense the colour. The scoring is defined by the metadata extracted from the web resource at step **305** of **FIG. 3**. The search results in each section depend on the information contained within the metadata and in the textual data thereby, displaying information that is only relevant to the user's browsing activity. The user is therefore able to dynamically see which web resources he or she has visited at a particular point in time and quickly locate the information he or she had seen before. The searchable personal browsing history dynamically updates the view every time the user visits another web page or downloads a file or image.

1. A method for creating a searchable personal browsing history, the method comprising the steps of:

in response to a user request to obtain a web page from the Internet, automatially extracting metadata and textual data from the web page obtained from the Internet;

indexing the extracted metadata and textual data and storing the indexed metadata and textual data; and

subsequently displaying the stored metadata and textual data in categories based on the indexing, and enabling searching of the displayed categories of metadata and textual data.

2. A method as claimed in claim 1 wherein the extracted metadata and textual data are stored with a reference to a location on the Internet from which the data resource was originally obtained.

3. A method as claimed in claim 1 wherein the indexing of stored metadata and textual data is updated each time new metadata and textual data is extracted from a new web page received from the Internet.

4. A method as claimed in claim 1 further comprising the step of a user searching the displayed categories of metadata and textual data.

5. A method as claimed in claim 1 further comprising the step of calculating statistical information on the extracted metadata relating to a user's browsing activity.

6. A method as claimed in claim 5 wherein the statistical information comprises recently visited web pages, most frequently visited web pages, recently visited downloads and recently visited images.

7. A computer program product for creating a searchable personal browsing history, said computer program product comprising:

a computer readable medium;

first program instructions to respond to a user request to obtain a web page from the Internet, by automatially extracting metadata and textual data from the web page obtained from the Internet;

second program instructions to index the extracted metadata and textual data and store the indexed metadata and textual data; and

third program instructions to subsequently display the stored metadata and textual data in categories based on the indexing, and enable searching of the displayed categories of metadata and textual data; and wherein

said first, second and third program instructions are recorded on said medium.

8. A program product as claimed in claim 7 wherein the extracted metadata and textual data are stored with a reference to a location on the Internet from which the data resource was originally obtained.

9. A program product as claimed in claim 7 wherein the indexing of stored metadata and textual data is updated each time new metadata and textual data is extracted from a new web page received from the Internet.

10. A program product as claimed in claim 7 further comprising fourth program instructions to calculate statistical information on the extracted metadata relating to a user's browsing activity; and wherein said fourth program instructions are recorded on said medium.

11. A program product as claimed in claim 10 wherein the statistical information comprises recently visited web pages, most frequently visited web pages, recently visited downloads and recently visited images.

12. A system for creating a searchable personal browsing history, said system comprising:

means for responding to a user request to obtain a web page from the Internet, by automatially extracting metadata and textual data from the web page obtained from the Internet;

means for indexing the extracted metadata and textual data and storing the indexed metadata and textual data; and

means for subsequently displaying the stored metadata and textual data in categories based on the indexing, and enabling searching of the displayed categories of metadata and textual data.

* * * * *