



(22) Date de dépôt/Filing Date: 2002/11/27

(41) Mise à la disp. pub./Open to Public Insp.: 2003/06/07

(45) Date de délivrance/Issue Date: 2010/11/16

(30) Priorité/Priority: 2001/12/07 (US10/017,750)

(51) Cl.Int./Int.Cl. *H04L 29/02* (2006.01),
G06F 17/30 (2006.01), *G06F 17/40* (2006.01),
H04L 12/66 (2006.01)

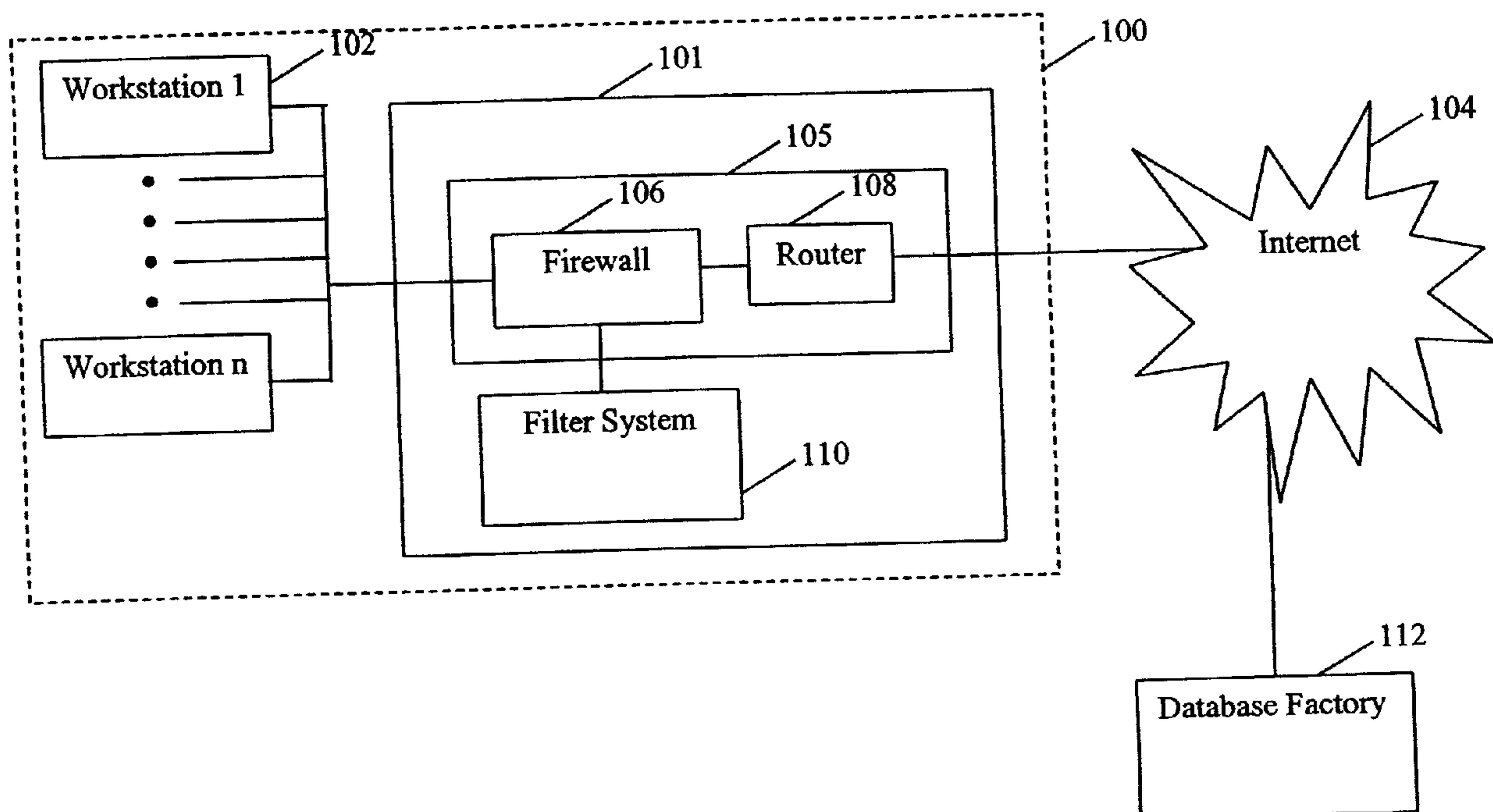
(72) Inventeurs/Inventors:
KESTER, HAROLD, US;
RUSKIN, DAN, US;
LEE, CHRIS, US;
ANDERSON, MARK, US

(73) Propriétaire/Owner:
WEBSense, INC., US

(74) Agent: FETHERSTONHAUGH & CO.

(54) Titre : SYSTÈME ET METHODE D'ADAPTATION D'UN FILTRE INTERNET

(54) Title: SYSTEM AND METHOD FOR ADAPTING AN INTERNET FILTER



(57) Abrégé/Abstract:

A system and method for updating a filtering system (110) which controls access to a website/page between a local area network (LAN) (100) and an Internet (104). The LAN (100) includes an Internet gateway system (105) coupled to a workstation (102) and configured to receive a URL request. The system controls access to the website/page associated with the URL based on one or more categories that are associated with the URL. The Internet gateway system (105) can determine the category that is associated with the URL by referencing a master database (204) or requesting the category from a database factory (112). The database factory (112) can receive URLs from multiple Internet gateway systems (105). The database factory (112) determines whether the identifier was previously categorized by the database factory (112) and provides the category to the Internet gateway system (105). Once the Internet gateway system (105) has the category, it applies rules associated with the category and user to filter access to the requested website/page.



SYSTEM AND METHOD FOR ADAPTING AN INTERNET FILTER**Abstract of the Disclosure**

A system and method for updating a filtering system (110) which controls access to a website/page between a local area network (LAN) (100) and an Internet (104). The LAN (100) includes an Internet gateway system (105) coupled to a workstation (102) and configured to receive a URL request. The system controls access to the website/page associated with the URL based on one or more categories that are associated with the URL. The Internet gateway system (105) can determine the category that is associated with the URL by referencing a master database (204) or requesting the category from a database factory (112). The database factory (112) can receive URLs from multiple Internet gateway systems (105). The database factory (112) determines whether the identifier was previously categorized by the database factory (112) and provides the category to the Internet gateway system (105). Once the Internet gateway system (105) has the category, it applies rules associated with the category and user to filter access to the requested website/page.

SYSTEM AND METHOD FOR ADAPTING AN INTERNET FILTER

Background of the Invention

Field of the Invention

5 A system and method for updating a filtering system which controls access to a website/page between a local area network (LAN) and an Internet.

Description of the Related Art

10 The Internet is a global system of computers that are linked together so that the various computers can communicate seamlessly with one another. Internet users access server computers to download and display informational pages. Once a server has been connected to the Internet, its informational pages can be displayed by virtually anyone having access to the Internet.

15 The easy access and inexpensive cost of retrieving Internet pages has led to several problems for controlling access to inappropriate information, such as pornography. Several solutions to this problem have been proposed, including rating systems similar to that used for rating movies so that a parent or employer could control access to Internet servers, or pages, that have a particular rating. Unfortunately, this mechanism requires each person running an Internet server to voluntarily rate their site. Because of the free-wheeling nature of the Internet, this
20 type of voluntary rating scheme is unlikely to be very efficient for preventing access to sites, such as those containing pornography, that most parents or businesses desire to block.

25 In addition to a rating scheme, others have developed databases that contain the uniform resource locator (URL) address of sites to be blocked. These databases are integrated into network computer systems and Internet firewalls so that a person wishing access to the Internet first has their URL request matched against the database of blocked sites. The user cannot access any URL found in the database. One such system is described in U.S. Patent No. 5,678,041 to Baker et al. Unfortunately, such systems rely on the completeness of the database of accessed sites to be
30 complete. Because new servers and URLs are being added to the Internet on a daily basis, as well as current servers being updated with new information, these databases do not provide a complete list of sites that should be blocked.

Summary of the Invention

The systems and methods have several features, no single one of which is solely responsible for its desirable attributes. Without limiting the scope as expressed by the claims which follow, its more prominent features will now be discussed briefly. After considering this discussion, and particularly after reading the section entitled "Detailed Description" one will understand how the features of the system and methods provide several advantages over traditional filter systems.

In accordance with one aspect of the invention, there is provided a system for collecting identifiers for updating a filtering system which controls access to Internet websites or pages between a local area network and an Internet. The system includes a workstation configured for a user to send an identifier to request an Internet website or page, an Internet gateway system coupled to the workstation and configured to receive the identifier and to allow or deny access to the Internet website or page associated with the identifier, a master database of identifiers along with one or more categories associated with each identifier, a filter system coupled to the Internet gateway system and configured to receive the identifier from the Internet gateway system, determine whether the identifier is in the master database, send the identifier to a database factory if the identifier is not in the master database, and apply one or more rules to one or more categories that are associated with the identifier, wherein the one or more categories are received from the database factory. The database factory is configured to receive the identifier from the filter system if the identifier was not in the master database, determine whether the identifier was previously categorized by the database factory, if the identifier was not previously categorized, determine the one or more categories to associate with the identifier and provide the one or more categories to the filter system, else provide the one or more categories that were previously associated with the identifier.

The identifier may be in the master database and may be associated with the one or more categories.

The filter system may be further configured to pre-filter the identifier or the Internet website or page associated with the identifier for a data characteristic that is indicative of the one or more categories, and may associate one or more indicators with the identifier.

Pre-filtering may be performed on text strings, graphics, and audio that are associated with the identifier or the Internet website or page.

The one or more indicators may include a category flag.

5 The filter system may use the one or more indicators to screen the identifier prior to sending the identifier to the database factory.

The Internet gateway system may include a firewall module configured to provide an electronic boundary between the workstation and the Internet and a router module configured to find a best path from the firewall module to the Internet website or page associated with the identifier.

10 The filter system may include a management module configured to provide an interface for a system administrator to select the one or more rules that are applied to the one or more categories associated with the identifier, an uncategorized database configured to store the identifier if the identifier is not in the master database and a manager module configured to send the stored identifier to the database factory and to
15 receive the one or more categories from the database factory.

The uncategorized database may include a request frequency that is associated with the identifier and may indicate the frequency of requests for the identifier in the uncategorized database.

20 The manager module may be configured to send the request frequency from the uncategorized database to the database factory.

The master database may include a request frequency that is associated with the identifier and may indicate the frequency of requests for the identifier in the master database.

25 The manager module may be configured to send the request frequency from the master database to the database factory.

The one or more rules may include blocking access to the Internet website or page based on the one or more categories associated with the identifier and the user.

The one or more rules may include allowing access to the Internet website or page based on the one or more categories associated with the identifier and the user.

30 The database factory may include a module configured to receive the identifier from the filter system and provide the one or more categories to the filter system, an identifier processing module configured to receive the identifier from the module and

determine whether the identifier has been previously categorized by the database factory, a categorization system module configured to categorize the identifier if not previously categorized by the database factory and a database of categorized identifiers configured to store the identifier and the one or more categories.

5 The module may be configured to receive a request frequency from the filter system to prioritize the identifier.

 The request frequency may be associated with the identifier and may indicate the frequency of requests for the identifier in the uncategorized database.

10 The request frequency may be associated with the identifier and may indicate the frequency of requests for the identifier in the master database.

 The identifier may be a uniform resource locator (URL).

 The identifier may be an Internet Protocol (I.P.) address.

 The identifier may be a domain name.

15 The system may further include a second filter system and a second Internet gateway system coupled to the second filter system and the database factory.

 The identifier processing module may be further configured to merge and sort the identifier and a second identifier received from the second Internet gateway system.

20 In accordance with another aspect of the invention, there is provided a method for adapting a filter system which controls access to Internet sites. The method involves receiving a request from a user in the form of an identifier to access a website or page, determining whether the identifier is in a master database of categorized identifiers and one or more categories associated with the identifier, if the identifier is not in the master database, determining whether the identifier is in an
25 uncategorized database, else applying one or more rules to the one or more categories associated with the identifier, if the identifier is not in the uncategorized database, posting the identifier to the uncategorized database, else updating an uncategorized database request frequency in the uncategorized database that is associated with the identifier, uploading the uncategorized database to a database factory, determining
30 whether each identifier has been previously categorized by the database factory, for each identifier that was not previously categorized, categorizing each identifier or website or page associated with the identifier to select one or more categories to

associate with each identifier and posting each identifier along with its selected one or more categories into a database of categorized sites and downloading the database of categorized sites to the filter system for incorporation into the master database.

5 The method may further involve updating a master database request frequency in the master database if the identifier is in the master database and uploading the master database request frequency and the associated identifier to the database factory.

 The one or more rules may include blocking access to the website or page based on the one or more categories associated with the identifier and the user.

10 The one or more rules may include allowing access to the website or page based on the one or more categories associated with the identifier and the user.

 The uncategorized database may further include additional data associated with the identifier.

 The additional data may include a request frequency.

15 The additional data may include an indicator.

 The additional data may include a trace ID.

 The additional data may include a primary language used by the filter system.

20 The method may further involve merging and sorting the uncategorized database based on each identifier and the additional data associated with each identifier.

 The method may further involve pre-filtering the identifier or website or page associated with the identifier for data characteristics that are indicative of the one or more categories and associating one or more indicators with the identifier.

25 Pre-filtering may be performed on text strings, graphics, and audio that are associated with the identifier and or website or page associated with the identifier.

 The one or more indicators may include a category flag.

 The method may further involve screening the identifier using the one or more indicators prior to sending the identifier to the database factory.

 The identifier may be a uniform resource locator (URL).

30 The identifier may be an Internet Protocol (I.P.) address.

 The identifier may be a domain name.

In accordance with another aspect of the invention, there is provided a method for collecting collection data for updating a filtering system which controls access to Internet sites. The method involves receiving a request from a user in the form of an identifier to access a website or page, determining whether the identifier is stored in a master database of categorized identifiers, if the identifier is stored, recalling a category that is associated with the identifier and applying a rule to the identifier that is associated with the category, if the identifier is not stored, determining whether the identifier is stored in an uncategorized database and if the identifier is not stored in the uncategorized database, posting the identifier to the uncategorized database, else updating an uncategorized database request frequency associated with the identifier.

The method may further involve updating a master database request frequency in the master database if the identifier is stored in the master database.

The method may further involve pre-filtering the identifier or website or page associated with the identifier for data characteristics that are indicative of one or more categories, determining whether data characteristics were found during the pre-filtering and if data characteristics were found, associating one or more indicators with the identifier in the uncategorized database.

The rule may block access to the website or page based on the category associated with the identifier and the user.

The rule may allow access to the website or page based on the category associated with the identifier and the user.

The identifier may be a uniform resource locator (URL).

The identifier may be an Internet Protocol (I.P.) address.

The identifier may be a domain name.

The uncategorized database may further include additional data associated with the identifier.

The additional data may include a trace ID.

The additional data may include a primary language used by the filter system.

The additional data may include the uncategorized database request frequency.

The master database may further include additional data associated with the identifier.

The master database may further include the master database request frequency.

In accordance with another aspect of the invention, there is provided a method for processing identifiers for updating a filtering system which controls access to Internet sites. The method involves receiving identifiers at a database factory from a filter system, wherein at least one of the identifiers corresponds to an Internet site requested to be accessed by a user, determining whether each of the identifiers has been previously categorized, for each identifier that was not previously categorized by the database factory, categorizing each of the identifiers, a website or page associated with the identifier, or additional data associated with the identifier, and associating output with each identifier and posting the identifier along with the associated output into a database of categorized sites.

The output may include one or more categories.

The output may include a statistical probability.

The output may include a multidimensional vector.

Categorizing each of the identifiers may be performed by an automated classification engine.

The identifiers may include an uncategorized identifier.

The identifiers may include a categorized identifier.

The method may further involve receiving at the database factory a master database request frequency which indicates the frequency that the categorized identifier was requested at the filter system.

The method may further involve merging and sorting the identifier and a second identifier received from a second filter system.

The method may further involve determining whether processing of the output associated with the uncategorized identifier is performed.

The processing of the output may include automatically posting the uncategorized identifier to the master database.

The processing of the output may include verifying that the output associated with the uncategorized identifier is correct.

The method may further involve processing the identifiers and the additional data.

Processing may include reassembling the identifiers and the additional data using a markup language.

The method may further involve decrypting the identifiers and the additional data and decompressing the identifiers and the additional data.

5 Decrypting may be performed using a data encryption standard (DES).

The identifier may be a uniform resource locator (URL).

The identifier may be an Internet Protocol (I.P.) address.

The identifier may be a domain name.

10 In accordance with another aspect of the invention, there is provided a system for collecting identifiers for updating a filtering system which controls access to a wide area network (WAN) of websites or pages. The system includes a master database including one or more identifiers received from a user to request access to an Internet website or page, and one or more categories associated with each of the one or more identifiers, an access system coupled to the WAN and configured to send an
15 identifier request if the identifier request is not in the master database and a database factory configured to receive the identifier request, select one or more categories to associate with the identifier request if the one or more categories were not previously associated with the identifier, and provide the selected one or more categories to the master database.

20 The access system may include an Internet gateway system coupled to a filter system.

The identifier may be a uniform resource locator (URL).

The identifier may be an Internet Protocol (I.P.) address.

The identifier may be a domain name.

25 The master database may include a request frequency that is associated with the identifier and indicative of the frequency of requests for the identifier at the filter system.

The database factory may be configured to receive the request frequency to prioritize the received identifiers.

30 Brief Description of the Drawings

FIGURE 1 is a block diagram of a site collection system for controlling access to Internet sites.

FIGURE 2 is a block diagram of a filter system.

FIGURE 3 is a flow diagram illustrating a process for collecting collection data.

FIGURE 4 is block diagram of a database factory.

[0011] FIGURE 5 is a flow diagram illustrating processing and uploading of collection data from the filter system to the database factory.

[0012] FIGURE 6 is a flow diagram illustrating processing of collection data by the database factory.

Detailed Description

[0013] The following detailed description is directed to certain specific embodiments of the invention. However, the invention can be embodied in a multitude of different systems and methods. In this description, reference is made to the drawings wherein like parts are designated with like numerals throughout.

[0014] In connection with the following description many of the components of the various systems, some of which are referred to as "module," can be implemented as software, firmware or a hardware component, such as a Field Programmable Gate Array (FPGA) or Application-Specific Integrated Circuit (ASIC), which performs certain tasks. Such components or modules may advantageously be configured to reside on the addressable storage medium and configured to execute on one or more processors. Thus, a module may include, by way of example, components, such as software components, object-oriented software components, class components and task components, processes, functions, attributes, procedures, subroutines, segments of program code, drivers, firmware, microcode, circuitry, data, databases, data structures, tables, arrays, and variables. The functionality provided for in the components and modules may be combined into fewer components and modules or further separated into additional components and modules. Additionally, the components and modules may advantageously be implemented to execute on one or more computers.

[0015] FIGURE 1 is a block diagram of a local area network (LAN) 100 coupled to an Internet 104 and a database factory 112 also coupled to the Internet 104. For ease of explanation only a single LAN is shown though two or numerous such networks would more typically be included. Similarly, two or more database factories could also be deployed.

[0016] The LAN 100 includes one or more workstations 102 coupled to an access system 101. The access system 101 includes an Internet gateway system 105 and a filter system 110. LANs may also include other devices such as servers (not shown). The LAN communicates via the Internet gateway system 105 in order to provide the workstation(s) 102 with communication to sites on the Internet 104. The LAN 100 can have an Ethernet 10baseT topology, or be based on any networking protocol, including wireless networks, token ring network, and the like.

[0017] The workstation 102 is coupled to the Internet gateway system 105. The workstation 102 can be a personal computer operating, for example, under the Microsoft Windows

Operating System. However, other computers, such as those manufactured by Apple, IBM, Compaq, Dell, Sun Microsystems or other system, can be used.

[0018] The Internet gateway system 105 couples the LAN 100 and the Internet 104. Internet gateway systems are well known in the art and normally communicate through connection devices, such as routers or other data packet switching technology, for translating Internet TCP/IP protocols into the proper protocols for communicating with the Internet 104. The Internet gateway system 105 used to implement a given system can vary as well as its location within the LAN 100. For example, Internet gateway system 105 could be located at the workstation(s) 102 or connected peripherally to the Internet 104. The Internet gateway system 105 illustrated in FIGURE 1 includes a firewall module 106 coupled to a router module 108.

[0019] The firewall module 106 provides an electronic boundary between devices on the LAN 100, such as the workstation(s) 102, and the Internet 104 to prevent unauthorized users from accessing computer resources on the LAN 100. More specifically, the firewall module 106 monitors data packets flowing to and from the Internet 104. Thus, all communications between the Internet 104 and the LAN first pass through the firewall module 106. The firewall module 106 can be one of the many firewall software programs commercially available, such as Firewall-1 (Check Point software, Redwood City, California). However, it should be realized that while the system described in FIGURE 1 has the firewall module 106 controlling access of data packets between the Internet 104 and the workstations 102, other similar access control systems are available and can be used. For example, the Microsoft proxy server (Microsoft Corp., Redmond, WA), Netscape proxy server (Netscape Corp) and the Cisco PIX Firewall (Cisco Corp.) are currently available and can also be used as the firewall module 106. Alternatively, a caching device can be utilized to provide access control. For example, the Inktomi Traffic Server (Inktomi Corp.) and the Network Appliance NetCache (Network Appliance Inc.) can be used.

[0020] The router module 108 is configured to find a best path for a data packet that is sent from the firewall 106 to the Internet 104. The router module 108 stores and forwards electronic messages between the firewall and the requested website/page, first determining all possible paths to the destination address and then picking the most expedient route, based on the traffic load and the number of hops.

[0021] Still referring to FIGURE 1, a filter system 110 is shown coupled to the firewall module 106. The filter system 110 receives user requests for accessing Internet websites/pages from the firewall module 106. Alternatively or additionally, the filter system can receive or monitor user requests for accessing the Internet from other points on the LAN. The filter system 110 determines whether the user will be allowed access to the requested website/page. Examples of techniques that can be used with the methods and systems disclosed herein are

disclosed in U.S. patent application no. 09/494,315, filed 1/28/2000, and entitled SYSTEM AND METHOD FOR CONTROLLING ACCESS TO INTERNET SITES, which is hereby incorporated by reference in its entirety.

[0022] The internet 104 in FIGURE 1 is a network or combination of networks spanning any geographical area, such as a local area network, wide area network, regional network, national network, and/or global network. Such networks may be hardwire, wireless, or a combination of hardwire and wireless.

[0023] The database factory 112 is shown connected to the filter system 110 via the Internet 104. Alternatively, the filter system 110 can communicate with the database factory 112 in other known ways such as a direct telephone link, a private network connection, or other suitable communication link.

[0024] FIGURE 2 is a block diagram of the filter system 110 from FIGURE 1 which communicates with the Internet gateway system 105. The filter system 110 can include a management module 200, a filter module 202, an upload/download manager module 208, a master database 204, and an uncategorized database 206.

[0025] A system administrator or the like interfaces with the filter system 110 via the management module 200 to select or create rules for users and/or groups of users. These rules can include, for example, allowing access to websites in selected categories and blocking access to websites in other categories. Rules can also include flexible filters. For example, rather than simply blocking or allowing access to the website/page, the system administrator selects or creates a flexible filter which is applied to the request. Example of flexible filters include, postponing the user's access, allowing the user to override denial of access, limiting the user's access based on a quota, and limiting the user's access based on a network load. Each requested website/page or category of website/pages can be associated with one or more rules.

[0026] The filter module 202 filters each request for Internet websites/pages using the master database 204 in conjunction with the rules. The filter module 202 analyzes the Internet website/page request from the workstation and then compares the Internet website/page request with the master database 204 of categorized Internet website/pages. It should be noted that the address could be a single page within an Internet website, or the default address of the website (e.g.: www.company.com). The master database 204 includes a list of websites/pages which can be in the form of URLs along with one or more categories associated with each URL. A URL (Uniform Resource Locator) is the address of a computer or a document on the Internet that consists of a communications protocol followed by a colon and two slashes (e.g.: http://), the identifier of a computer, and usually a path through a directory to a file. The identifier of the computer can be in the form of a domain name, for example www.m-w.com, or an Internet

Protocol (I.P.) address, for example 123.456.789.1. A unique domain name can correspond to multiple I.P. addresses. Though often addresses, components thereof (for example, I.P. address, domain name, and communication protocol), or other location identifiers can be used to identify computers or documents on the Internet, for ease of description the term URL is used hereafter. The master database 204 can also include additional data associated with the URL. For example, a request frequency for the categorized website/page can be included in the master database 204. If the URL of the categorized website/page is found in the master database 204, the request frequency in the master database 204 can be updated for the requested website/page. A reporter log (not shown) can be used to track requested websites/pages that are found in the master database 204.

[0027] The filter module 202 checks to see if the requested website/page address matches any addresses stored in the master database 204. If an address match between the requested address and the master database 204 is found, the filter system 110 applies the rule(s) associated with the one or more categories that match the requested address and the user. For example, if application of the rule by the filter module 202 indicates that the requested website/page is to be blocked, a pre-defined block page is sent to the user's browser explaining that the request is not allowed and why. Alternatively, the filter system 110 simply closes the connection that was requested by the Internet browser to the requested website/page.

[0028] If the filter module 202 does not find the URL in the master database 204 (i.e. the URL is uncategorized), the filter module 202 then determines how to proceed with the uncategorized Internet website/page. For example, user access to the requested website/page can be allowed when the filter module 202 determines that the website/page is uncategorized. Alternatively, the filter module 202 can block access to uncategorized sites.

[0029] Even when the requested website/page is not found in the master database 204, the filter module 202 can pre-filter, or scan, the requested website/page for specific characteristics. These specific characteristics can relate to one or more of the categories found in the master database 204. For example, the scan can identify whether the uncategorized website/page includes characteristics that are indicative of pornography. This scan can be accomplished by, for example, searching the requested website/page and URL for text strings, graphics, audio and the like which have a high correlation with pornography websites. The filter module 202 can then associate an indicator with the uncategorized website/page based on the results of the scan. The indicator can be, for example, a specific category flag that relates to characteristics found during the scan of the uncategorized website/page. Continuing with the example above, if a text string was found that was indicative of pornography, a pornography flag would be attached to the uncategorized website/page. Alternatively, the filter module 202 performs the categorization of the URL and adds the URL and associated categories to the master database 204.

[0030] For uncategorized websites/pages, the filter module 202 determines whether they are represented in the uncategorized database 206 of URLs. If they are not, the filter module 202 stores the URLs associated with the requested uncategorized websites/pages in the uncategorized database 206. The uncategorized database can include additional data associated with the URL. For example, the request frequency for the uncategorized website/page and/or one or more indicators identified during the filter module's scan of the uncategorized website/page can be included in the uncategorized database 206. If the URL of the uncategorized website/page is found in the uncategorized database 206, the request frequency can be updated for the requested website/page.

[0031] Still referring to FIGURE 2, the upload/download manager module 208 can transmit data from the uncategorized database 206 and the master database 204 to the database factory 112 (see FIGURE 1). The upload could be immediate or periodic depending on the level of service required. For example, a daily upload after normal business hours could be used. The upload/download module 208 can refer to the request frequency and/or one or more indicators to prioritize the URLs in the uncategorized database 206 for their transmission to the database factory 112. If data from the master database 204 is to be uploaded to the database factory 112, the upload/download module 208 can refer to a request frequency for websites/pages found in the master database 204. The request frequency can be used to prioritize the URLs in the master database 204 for their transmission to the database factory 112.

[0032] FIGURE 3 is a flow diagram illustrating a process performed by the filter system 110 to collect uncategorized websites/pages. The collection process begins at a state 502 where the filter module 202 receives a user request in the form of a URL to access a website/page. As was noted above, the requested URL and the identification of the user can be received from the Internet gateway system 105 or from a direct monitoring of traffic on the LAN by the filter system 110. Next, the process moves to a decision block 504 where the filter module 202 determines whether the URL is in the master database 204. If the URL is in the master database 204, the process proceeds to a state 506 where the request for the website is posted in a reporter log. The reporter log is available to the system administrator for tracking requests for websites/pages. Alternatively, the request for a website that is found in the master database 204 is posted to the request frequency in the master database 204. The process moves to a state 508 where the filter module 202 recalls the one or more categories that are associated with the requested website/page. The filter module 202 can then apply one or more rules associated with the requesting user and the one or more categories.

[0033] Returning to the decision block 504, if the URL is not in the master database 204, the process continues to a decision state 510 where the filter module 202 determining whether

to pre-filter the uncategorized URL. The system administrator can select whether pre-filtering is to be performed by the filter module 202. If the filter module 202 does not perform pre-filtering, the process proceeds to a state 512 where the filter module 202 posts the URL to the uncategorized database 206 as uncategorized. Next, at a state 514, if the URL was already posted in the uncategorized database 206, the filter module 202 updates the request frequency associated with the URL. The process then returns to state 502 where the filter module waits to receive the next request for a website/page. Additionally, the filter module can allow or deny access to the user based upon a rule for uncategorized URLs.

[0034] Returning to decision state 510, if the filter module 202 is to perform pre-filtering, the process moves to a state 516 where the filter module 202 scans or analyzes the URL and/or website/page associated with the requested URL for specific characteristics that are indicative of one or more categories. The process continues to a decision state 518 where the filter module 202 determines whether any data characteristics were found during the scan. If data characteristics were found, the process moves to a state 520 where an indicator, for example, a flag, is associated with the requested URL. The process then continues to state 512 as described above where the URL is stored in the uncategorized database with the indicator.

[0035] FIGURE 4 is block diagram of the database factory 112 connected to the Internet 104. The database factory 112 provides the master databases to filter system(s) and processes websites/pages that are associated with uncategorized URLs and other information, for example, frequency usage in the master database. For example, the database factory 112 receives uncategorized URLs and any additional data associated with the URL from the filter system 110 and downloads categorized URLs to the filter system. The database factory 112 can also upload the request frequency for website/pages found in the master database 204. Additional techniques available to the database factory 112 for collecting URLs include, for example, using a domain name system (DNS), using an Internet search engine, mining high trafficked website/page directories, and receiving suggested sites from the public.

[0036] The database factory 112 can include an upload/download module 701, a URL processing module 700, a categorization system module 702, and a database 704 of categorized URLs.

[0037] One function of the upload/download module 701 is to receive URLs and any additional data associated with the URLs from the filter system 110. In one embodiment, the URLs include URLs from the uncategorized database 206 and URLs from the master database 204. The additional data can include a request frequency for a website\page found in the master database 204, a request frequency for a website/page found in the uncategorized database 206, an indicator associated with a URL, a trace ID, and a primary language used by a filter system. For

ease of explanation, the term collection data will be used to include URLs and any additional data associated with the URL. Additionally, the upload/download module 701 downloads the master database to the filter system(s), as will be described more fully below.

[0038] The URL processing module 700 receives the collection data from the upload/download module 701. The URL processing module 700 processes the collection data. Processing can include merging, sorting, and determining a language for the collection data from multiple filter systems. The URL processing module 700 determines whether each URL in the collection data requires categorization. If the URL has not been previously categorized, the categorization system module 702 receives the URL and any additional data associated with the URL from the URL processing module 700.

[0039] The categorization system module 702 categorizes URLs which are then added to the database 704 of categorized sites. The categorization system module 702 can analyze each URL, the website/page associated with the URL, and any additional data associated with the URL to determine its appropriate category or categories.

[0040] The categorization system module 702 can include an automated categorization or classification engine to determine the appropriate category or categories of the URL. The automated categorization engine can determine statistical probabilities and multidimensional vectors during the categorization process. Categorization can be based upon word analysis, adaptive learning systems, and image analysis. The categorization system module 702 can interface with a human checker to determine the appropriate category or categories of the URL. The categorization system module 702 can include the automated categorization engine and the human checker to determine the appropriate category or categories of the URL. For example, the automated categorization engine can initially determine the appropriate category. The human checker can verify that the URL is correctly categorized. The categorization system 702 determines whether the human checker is required to review the categorization results for each URL. If a human checker is involved, his results can also be utilized to refine the automated categorization engine. Once categorized, the categorization system module 702 posts the URL along with its associated one or more categories into the database 704 of categorized sites.

[0041] The categorization system module 702 can include a language analyzer. The language analyzer determines the language of the website for each URL. Determining the language can facilitate the categorization process by allowing each human checker to be language dependent.

[0042] The database 704 of categorized sites can include URLs and their associated categories. The database 704 can be stored in a relational database management system, such as Oracle, Sybase, Informix, Microsoft Server, and Access.

[0043] Once the categorization system module 702 has posted the URL and its associated category or categories into the database 704, the upload/download module 701 thereafter routinely copies the database 704 to the filter system(s) 110. As can be imagined, the system can include thousands of filter systems, each of which is updated regularly by the upload/download module 701 to provide an updated database of categorized URLs. Moreover, the upload/download module 701 can transfer portions of the database 704, such as updates, to the filter system 110 so that the entire database does not need to be transmitted.

[0044] FIGURE 5 is a flow diagram illustrating processing and uploading of collection data from the filter system 110 to the database factory 112. The process begins at a start state 600. Next, at a state 602, the upload/download manager module 208 (see FIGURE 2) requests a download of URLs and their associated categories from the database factory 112. This request can be periodic, random, or at a set time. For example, the request can be made when the number of URLs or the number of stored bits in the uncategorized database reaches a selected level. The request can be made when a selected maximum request frequency for any of the URLs in the uncategorized database 206 is reached. The request can be made as a result of an uncategorized URL being associated with an indicator. For example, when a URL is associated with a pornography flag, the upload/download manager module 208 instigates an upload to the database factory 112. Additionally, the request can be in response to polling by the database factory 112. Alternatively, the upload/download module 701 can initiate the download. For example, the upload/download module 701 can initiate the process based on the current processing capacity of the categorization system module 702. If the categorization system module 702 is currently being underutilized, the upload/download module 701 can seek a filter system(s) and initiate an upload of the filter system's uncategorized database 206.

[0045] The process continues to a decision state 604 where the upload/download manager module 208 determines whether pre-filtering of the uncategorized URLs was performed by the filter module 202. This pre-filtering, or scanning, is performed to determine whether the requested website/page includes specific characteristics. These specific characteristics can relate to one or more of the categories found in the master database 204. For example, the scan can identify whether the uncategorized website/page includes characteristics that are indicative of pornography. The filter module 202 can then associate an indicator with the uncategorized website/page based on the results of the scan. The indicator can be, for example, a specific category flag that relates to characteristics found during the scan of the uncategorized website/page. The upload/download module 208 can refer to the one or more indicators to prioritize the URLs in the uncategorized database 206 for uploading to the database factory 112.

[0046] If pre-filtering was not performed, the process moves to a decision state 606 where the upload/download manager module 208 determines whether all of the uncategorized URLs are to be uploaded to the database factory 112. This provides the option to not upload the uncategorized database 206 but still receive a download of categorized sites from the database factory 112. The system administrator can select whether all of the uncategorized URLs are to be uploaded. If all of the uncategorized URLs are not to be uploaded, the process proceeds to a decision state 607 where the upload/download manager module 208 determines whether all of the URLs are to be uploaded. If all of the URLs are not to be uploaded, the process moves to a state 608 where the upload/download manager module 208 receives categorized URLs from the database factory 112. The upload/download module 701 can copy the database 704 to the filter system(s) 110. Thus each filter system 110 can be updated regularly by the upload/download module 701 to provide an updated database of categorized URLs. Moreover, the upload/download module 701 can transfer portions of the database 704, such as updates, to the filter system 110 so that the entire database does not need to be transmitted. The upload/download manager module 208 posts the categorized URLs into the master database 204.

[0047] Returning to decision state 606, if all of the uncategorized URLs are to be categorized by the database factory 112, the process moves to a state 610 where the upload/download manager 208 retrieves all URLs from the uncategorized database 206. The process moves to a state 612 where the uncategorized URLs and any additional data associated with the URLs, i.e. collection data, can be formatted. The additional data can include request frequencies and/or indicators associated with the URLs. For ease of explanation, the term collection data is being used to include URLs and any additional data associated with the URL. The collection data is not required to be formatted and thus may be directly uploaded to the database factory 112. Moreover, the selection of a format for the collection data can depend on the type of data connection that the database factory 112 has with the filter system 110. For a data connection via the Internet 104, the upload/download module 208 can use a markup language, for example, Extensible Markup Language (XML), Standard Generalized Markup Language (SGML), and HyperText Markup Language (HTML), to format the collection data.

[0048] The collection data can be further processed prior to its upload to the database factory 112. For example, limit block 614, compression block 616, and encryption block 618 can be performed to process the collection data for upload to the database factory 112. While these blocks may facilitate the upload of the collection data, they are not required to be performed. The collection data can be uploaded without applying blocks 614, 616, 618. Thus, the collection data can be directly uploaded to the database factory 112 without applying blocks 612 through blocks 618.

[0049] If further processing is desired, the process moves to a state 614 where the upload/download manager 208 checks the limits of the collection data. The upload/download manager 208 can limit the collection data to a maximum size for uploading to the database factory 112. For example, the collection data from a single filter system could be limited to a maximum of 20 Mbytes. The process continues to a state 616 where the collection data is compressed so that the collection data takes up less space. Next, at a state 618 the collection data is encrypted so that it is unreadable except by authorized users, for example, the database factory 112.

[0050] Flow continues to a state 620 where the collection data is uploaded to the database factory 112. As explained above, the collection data can include any additional data associated with the URL, for example, request frequencies and/or indicators. The process then moves to state 608 as described above to receive categorized URLs from the database factory 112.

[0051] Returning to decision state 604, if pre-filtering was performed by the filter module 202, the process moves to a state 622 where the upload/download manager 208 retrieves URLs that were associated with an indicator. The indicator can be, for example, a specific category flag that relates to characteristics found during the scan of the uncategorized website/page. Multiple indicators can be associated with a single URL. The upload/download module 208 can refer to the one or more indicators, and/or the request frequency to prioritize the URLs in the uncategorized database 206 for uploading to the database factory 112. The process then continues to state 612 where formatting can be performed on the URL and on any associated data as described above.

[0052] Returning to decision state 607, if all URLs are to be uploaded to the database factory 112, the process moves to a state 611 where the upload/download manager 208 retrieves all URLs that have been requested by users of the filter system 110. For example, the URLs from the uncategorized database 206 along with the URLs from the master database 204 are retrieved along with additional data, for example, request frequency. Alternatively, the URLs from the uncategorized database 206 along with data from the reporter log (not shown) is retrieved. The process then continues to state 612 where formatting can be performed on the URLs retrieved in state 611 and on any associated data as described above.

[0053] FIGURE 6 is a flow diagram illustrating the processing of collection data by the database factory 112. The process begins at a state 800 where the upload/download module 701 receives collection data from the filter system 110 (see FIGURE 1). The time for receiving the collection data can be periodic, random, at a set time, or in response to polling. The upload/download module 701 and/or the upload/download manager module 208 can initiate the upload to the database factory 112. As explained above, the collection data can include any additional data associated with the URL, for example, request frequencies associated with URLs

from the master database 204 and/or request frequencies associate with URLs from the uncategorized database 206, and/or indicators.

[0054] Next, at a state 802, the URL processing module 700 receives the collection data from the upload/download module 701. The collection data can be formatted or unformatted. Additionally, the collection data can be encrypted and/or compressed or not.

[0055] The process continues to a state 804 where the URL processing module 700 decrypts and uncompresses the collection data if decryption and/or uncompression is required. The process moves to a state 806 where the collection data is reassembled into a list of URLs and any additional data associated with the URL.

[0056] The process moves to a state 808 where the URL processing module 700 merges and sorts the collection data. The system can include thousands of filter systems, each of which is regularly uploading collection data from its filter system 110. As explained above, the collection data can include any additional data associated with the URL, for example, request frequencies and/or indicators. The URL processing module 700 can merge and sort the uploaded data for a filter system(s) based on the URL or any additional data associated with the URL. For example, the URL processing module 700 can refer to one or more indicators, and/or request frequencies to sort and merge the URLs from one or more filter systems.

[0057] The URL processing module 700 determines whether each URL in the collection data requires categorization. If the URL has not been previously categorized, the categorization system module 702 receives the URL and any additional data associated with the URL from the URL processing module 700.

[0058] Next, at a state 810 the categorization system module 702 categorizes URLs which are then added to the database 704 of categorized sites. As explained above, the categorization system module 702 can analyze each URL, the website/page associated with the URL, and any additional data associated with the URL to determine its appropriate category or categories. The categorization system module 702 can include an automated categorization/classification engine to determine the appropriate category or categories of the URL. Categorization can be based upon word analysis, adaptive learning systems, and image analysis. The categorization system module 702 can interface with a human checker to determine the appropriate category or categories of the URL. The categorization system module 702 can include the automated categorization engine and the human checker to determine the appropriate category or categories of the URL. The categorization system 702 determines whether the human checker is required to review the categorization results for each URL. If a human checker is involved, his results can also be utilized to refine the automated categorization engine.

[0059] The process continues to a state 812 where the categorization system module 702 posts the URL along with its associated one or more categories into the database 704 of categorized sites. The database 704 of categorized sites can include URLs and their associated categories.

[0060] While the above detailed description has shown, described, and pointed out novel features of the invention as applied to various embodiments, it will be understood that various omissions, substitutions, and changes in the form and details of the device or process illustrated may be made by those skilled in the art without departing from the spirit of the invention. The scope of the invention is indicated by the appended claims rather than by the foregoing description. All changes which come within the meaning and range of equivalency of the claims are to be embraced within their scope.

**THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE
PROPERTY OR PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS:**

1. A system for collecting identifiers for updating a filtering system which
controls access to Internet websites or pages between a local area network and
an Internet comprising:

a workstation configured for a user to send an identifier to request an
Internet website or page;

an Internet gateway system coupled to the workstation and configured
to receive the identifier and to allow or deny access to the Internet
website or page associated with the identifier;

a master database of identifiers along with one or more categories
associated with each identifier;

a filter system coupled to the Internet gateway system and configured
to receive the identifier from the Internet gateway system, determine
whether the identifier is in the master database, send the identifier to a
database factory if the identifier is not in the master database, and
apply one or more rules to one or more categories that are associated
with the identifier, wherein the one or more categories are received
from the database factory; and

the database factory configured to receive the identifier from the filter
system if the identifier was not in the master database, determine
whether the identifier was previously categorized by the database
factory, if the identifier was not previously categorized, determine the
one or more categories to associate with the identifier and provide the
one or more categories to the filter system, else provide the one or
more categories that were previously associated with the identifier.

2. The system of Claim 1, wherein the identifier is in the master database and is associated with the one or more categories.
- 5 3. The system of Claim 1, wherein the filter system is further configured to pre-filter the identifier or the Internet website or page associated with the identifier for a data characteristic that is indicative of the one or more categories, and associating one or more indicators with the identifier.
- 10 4. The system of Claim 3, wherein pre-filtering can be performed on text strings, graphics, and audio that are associated with the identifier or the Internet website or page.
- 15 5. The system of Claim 3, wherein the one or more indicators can include a category flag.
- 20 6. The system of Claim 5, wherein the filter system uses the one or more indicators to screen the identifier prior to sending the identifier to the database factory.
- 25 7. The system of Claim 1, wherein the Internet gateway system comprises:
 - a firewall module configured to provide an electronic boundary between the workstation and the Internet; and
 - a router module configured to find a best path from the firewall module to the Internet website or page associated with the identifier.
- 30 8. The system of Claim 7, wherein the filter system comprises:
 - a management module configured to provide an interface for a system administrator to select the one or more rules that are applied to the one or more categories associated with the identifier;

an uncategorized database configured to store the identifier if the identifier is not in the master database; and

a manager module configured to send the stored identifier to the database factory and to receive the one or more categories from the database factory.

9. The system of Claim 8, wherein the uncategorized database includes a request frequency that is associated with the identifier and indicates the frequency of requests for the identifier in the uncategorized database.

10. The system of Claim 9, wherein the manager module is configured to send the request frequency from the uncategorized database to the database factory.

11. The system of Claim 10, wherein the master database includes a request frequency that is associated with the identifier and indicates the frequency of requests for the identifier in the master database.

12. The system of Claim 11, wherein the manager module is configured to send the request frequency from the master database to the database factory.

13. The system of Claim 8, wherein the one or more rules include blocking access to the Internet website or page based on the one or more categories associated with the identifier and the user.

14. The system of Claim 8, wherein the one or more rules include allowing access to the Internet website or page based on the one or more categories associated with the identifier and the user.

15. The system of Claim 1, wherein the database factory comprises:

a module configured to receive the identifier from the filter system and provide the one or more categories to the filter system;

an identifier processing module configured to receive the identifier from the module and determine whether the identifier has been previously categorized by the database factory;

5 a categorization system module configured to categorize the identifier if not previously categorized by the database factory; and

a database of categorized identifiers configured to store the identifier and the one or more categories.

10

16. The system of Claim **15**, wherein the module is configured to receive a request frequency from the filter system to prioritize the identifier.

15

17. The system of Claim **16**, wherein the request frequency is associated with the identifier and indicates the frequency of requests for the identifier in the uncategorized database.

20

18. The system of Claim **16**, wherein the request frequency is associated with the identifier and indicates the frequency of requests for the identifier in the master database.

19. The system of Claim **15**, wherein the identifier is a uniform resource locator (URL).

25

20. The system of Claim **15**, wherein the identifier is an Internet Protocol (I.P.) address.

21. The system of Claim **15**, wherein the identifier is a domain name.

30

22. The system of Claim **15**, further comprising:

a second filter system; and

a second Internet gateway system coupled to the second filter system and the database factory.

5 23. The system of Claim 22, wherein the identifier processing module is further configured to merge and sort the identifier and a second identifier received from the second Internet gateway system.

10 24. A method for adapting a filter system which controls access to Internet sites, the method comprising:

receiving a request from a user in the form of an identifier to access a website or page;

15 determining whether the identifier is in a master database of categorized identifiers and one or more categories associated with the identifier;

20 if the identifier is not in the master database, determining whether the identifier is in an uncategorized database, else applying one or more rules to the one or more categories associated with the identifier;

25 if the identifier is not in the uncategorized database, posting the identifier to the uncategorized database, else updating an uncategorized database request frequency in the uncategorized database that is associated with the identifier;

uploading the uncategorized database to a database factory;

30 determining whether each identifier has been previously categorized by the database factory;

for each identifier that was not previously categorized, categorizing each identifier or website or page associated with the identifier to select one or more categories to associate with each identifier; and

posting each identifier along with its selected one or more categories into a database of categorized sites; and

5 downloading the database of categorized sites to the filter system for incorporation into the master database.

25. The method of Claim 24, further comprising:

10 updating a master database request frequency in the master database if the identifier is in the master database; and

 uploading the master database request frequency and the associated identifier to the database factory.

15

26. The method of Claim 24, wherein the one or more rules include blocking access to the website or page based on the one or more categories associated with the identifier and the user.

20 27. The method of Claim 24, wherein the one or more rules include allowing access to the website or page based on the one or more categories associated with the identifier and the user.

25 28. The method of Claim 24, wherein the uncategorized database further includes additional data associated with the identifier.

29. The method of Claim 28, wherein the additional data includes a request frequency.

30 30. The method of Claim 28, wherein the additional data includes an indicator.

31. The method of Claim 28, wherein the additional data includes a trace ID.

32. The method of Claim 28, wherein the additional data includes a primary language used by the filter system.
- 5 33. The method of Claim 28, further comprising merging and sorting the uncategorized database based on each identifier and the additional data associated with each identifier.
34. The method of Claim 28, further comprising:
- 10 pre-filtering the identifier or website or page associated with the identifier for data characteristics that are indicative of the one or more categories; and
- associating one or more indicators with the identifier.
- 15 35. The method of Claim 34, wherein the pre-filtering is performed on text strings, graphics, and audio that are associated with the identifier and or website or page associated with the identifier.
- 20 36. The method of Claim 34, wherein the one or more indicators can include a category flag.
37. The method of Claim 36, further comprising screening the identifier using the one or more indicators prior to sending the identifier to the database factory.
- 25 38. The method of Claim 37, wherein the identifier is a uniform resource locator (URL).
39. The method of Claim 37, wherein the identifier is an Internet Protocol (I.P.) address.
- 30 40. The method of Claim 37, wherein the identifier is a domain name.

- 41.** A method for collecting collection data for updating a filtering system which controls access to Internet sites, the method comprising:

receiving a request from a user in the form of an identifier to access a website or page;

determining whether the identifier is stored in a master database of categorized identifiers;

if the identifier is stored, recalling a category that is associated with the identifier and applying a rule to the identifier that is associated with the category;

if the identifier is not stored, determining whether the identifier is stored in an uncategorized database; and

if the identifier is not stored in the uncategorized database, posting the identifier to the uncategorized database, else updating an uncategorized database request frequency associated with the identifier.

- 42.** The method of Claim **41**, further comprising:

updating a master database request frequency in the master database if the identifier is stored in the master database.

- 43.** The method of Claim **41**, further comprising:

pre-filtering the identifier or website or page associated with the identifier for data characteristics that are indicative of one or more categories;

determining whether data characteristics were found during the pre-filtering;
and

if data characteristics were found, associating one or more indicators with the
identifier in the uncategorized database.

44. The method of Claim 43, wherein the rule blocks access to the website or page
based on the category associated with the identifier and the user.

45. The method of Claim 43, wherein the rule allows access to the website or page
based on the category associated with the identifier and the user.

46. The method of Claim 43, wherein the identifier is a uniform resource locator
(URL).

47. The method of Claim 43, wherein the identifier is an Internet Protocol (I.P.)
address.

48. The method of Claim 43, wherein the identifier is a domain name.

49. The method of Claim 43, wherein the uncategorized database further includes
additional data associated with the identifier.

50. The method of Claim 49, wherein the additional data includes a trace ID.

51. The method of Claim 49, wherein the additional data includes a primary
language used by the filter system.

52. The method of Claim 49, wherein the additional data includes the
uncategorized database request frequency.

53. The method of Claim 43, wherein the master database further includes
additional data associated with the identifier.

54. The method of Claim **53**, wherein the master database further includes the master database request frequency.

5 **55.** A system for collecting identifiers for updating a filtering system which controls access to a wide area network (WAN) of websites or pages, comprising:

10 a master database including one or more identifiers received from a user to request access to an Internet website or page, and one or more categories associated with each of the one or more identifiers;

15 an access system coupled to the WAN and configured to send an identifier request if the identifier request is not in the master database; and

20 a database factory configured to receive the identifier request, select one or more categories to associate with the identifier request if the one or more categories were not previously associated with the identifier, and provide the selected one or more categories to the master database.

56. The system of Claim **55**, wherein the access system comprises an Internet gateway system coupled to a filter system.

25 **57.** The system of Claim **56**, wherein the identifier is a uniform resource locator (URL).

58. The system of Claim **56**, wherein the identifier is an Internet Protocol (I.P.) address.

30

59. The system of Claim **56**, wherein the identifier is a domain name.

60. The system of Claim **56**, wherein the master database includes a request frequency that is associated with the identifier and indicative of the frequency of requests for the identifier at the filter system.

5 **61.** The system of Claim **60**, wherein the database factory is configured to receive the request frequency to prioritize the received identifiers.

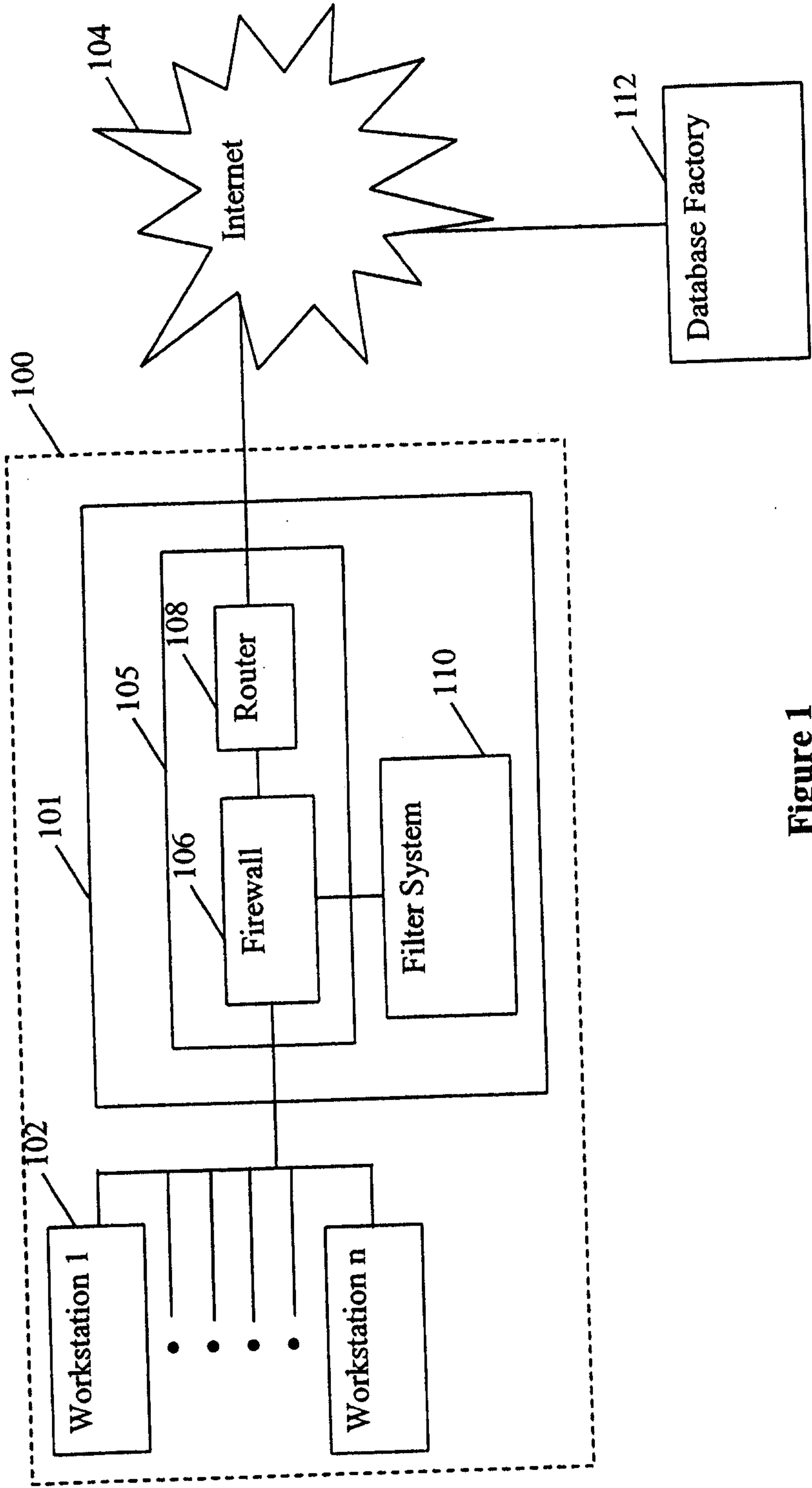


Figure 1

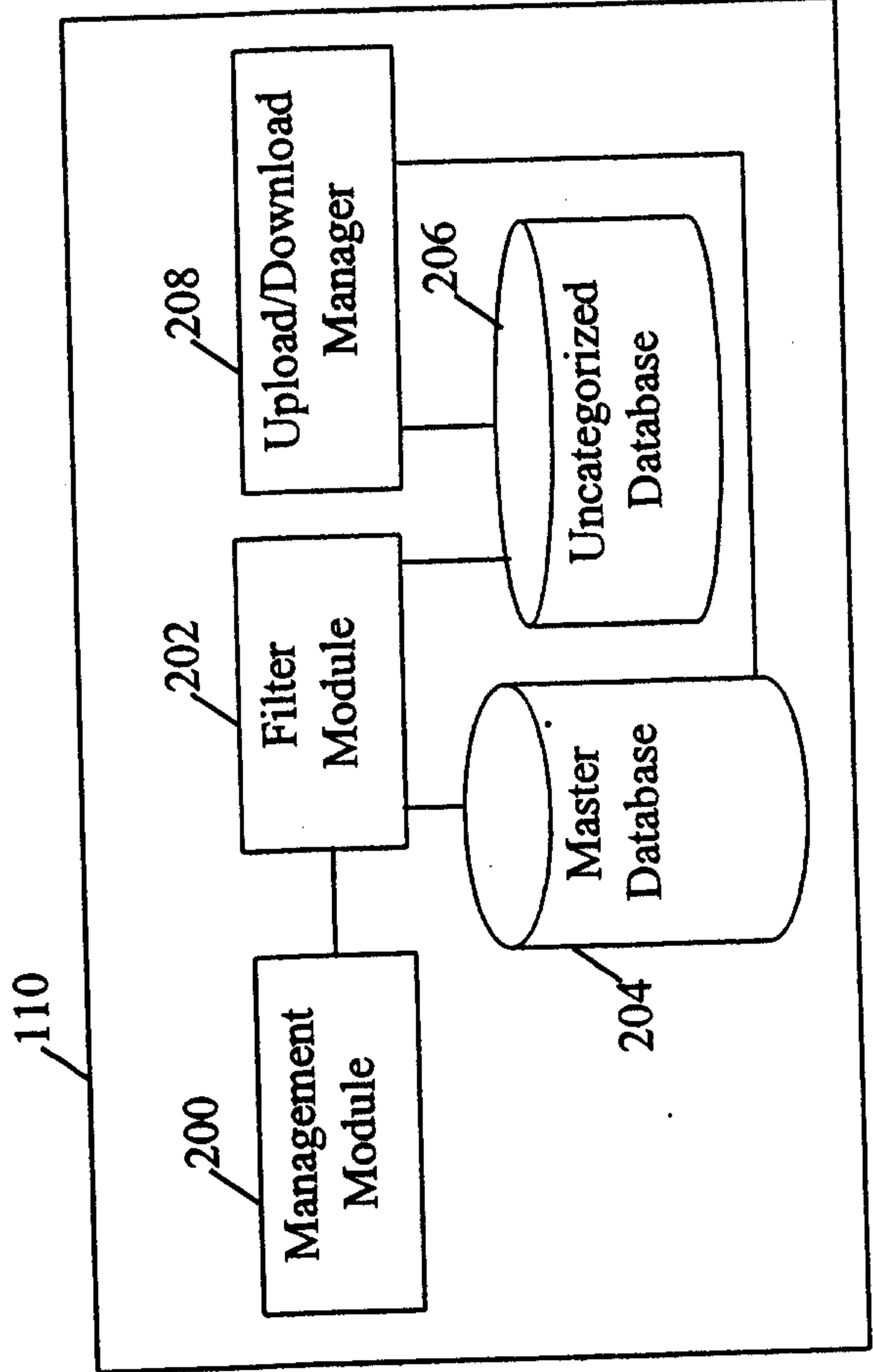


Figure 2

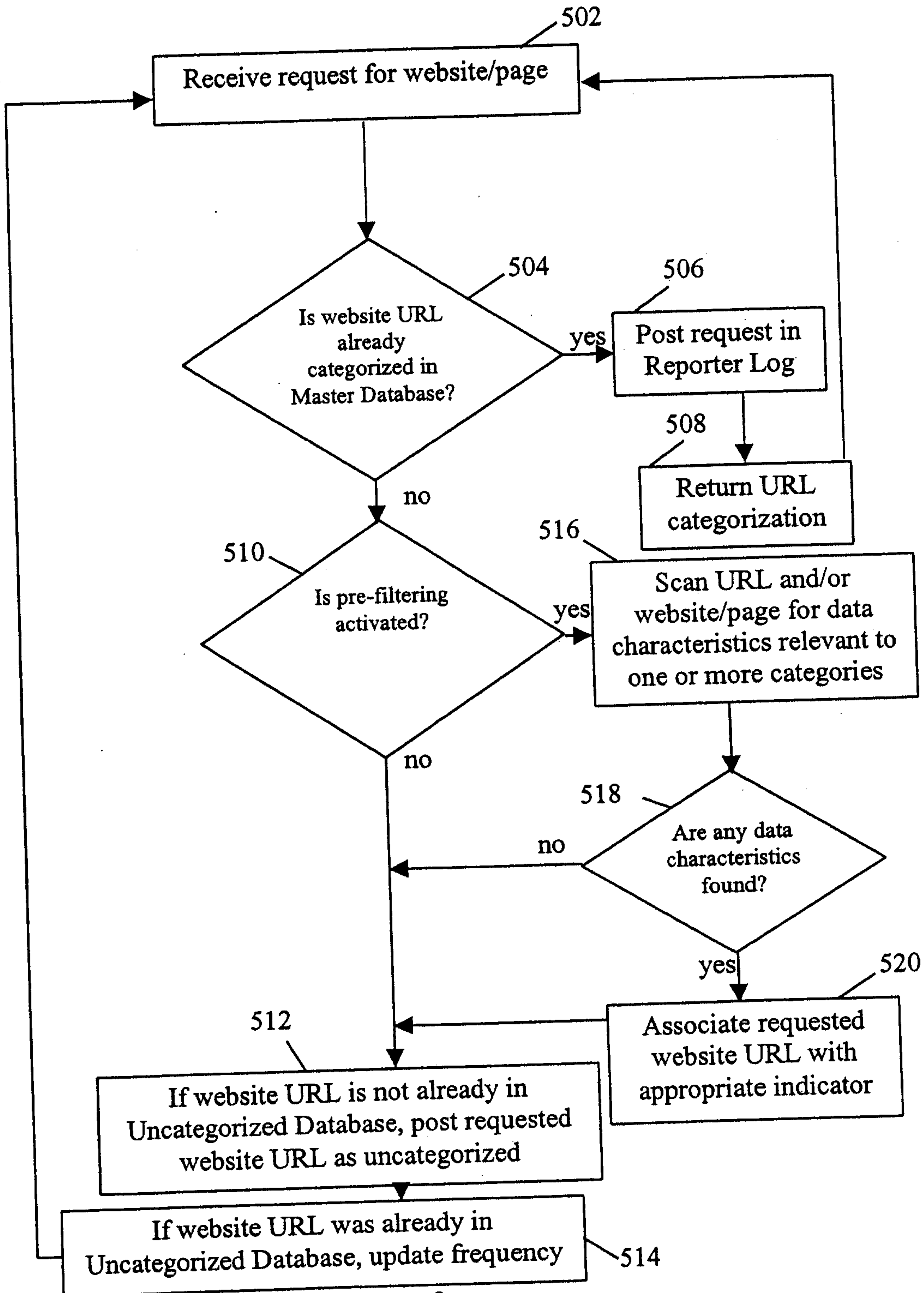


Figure 3

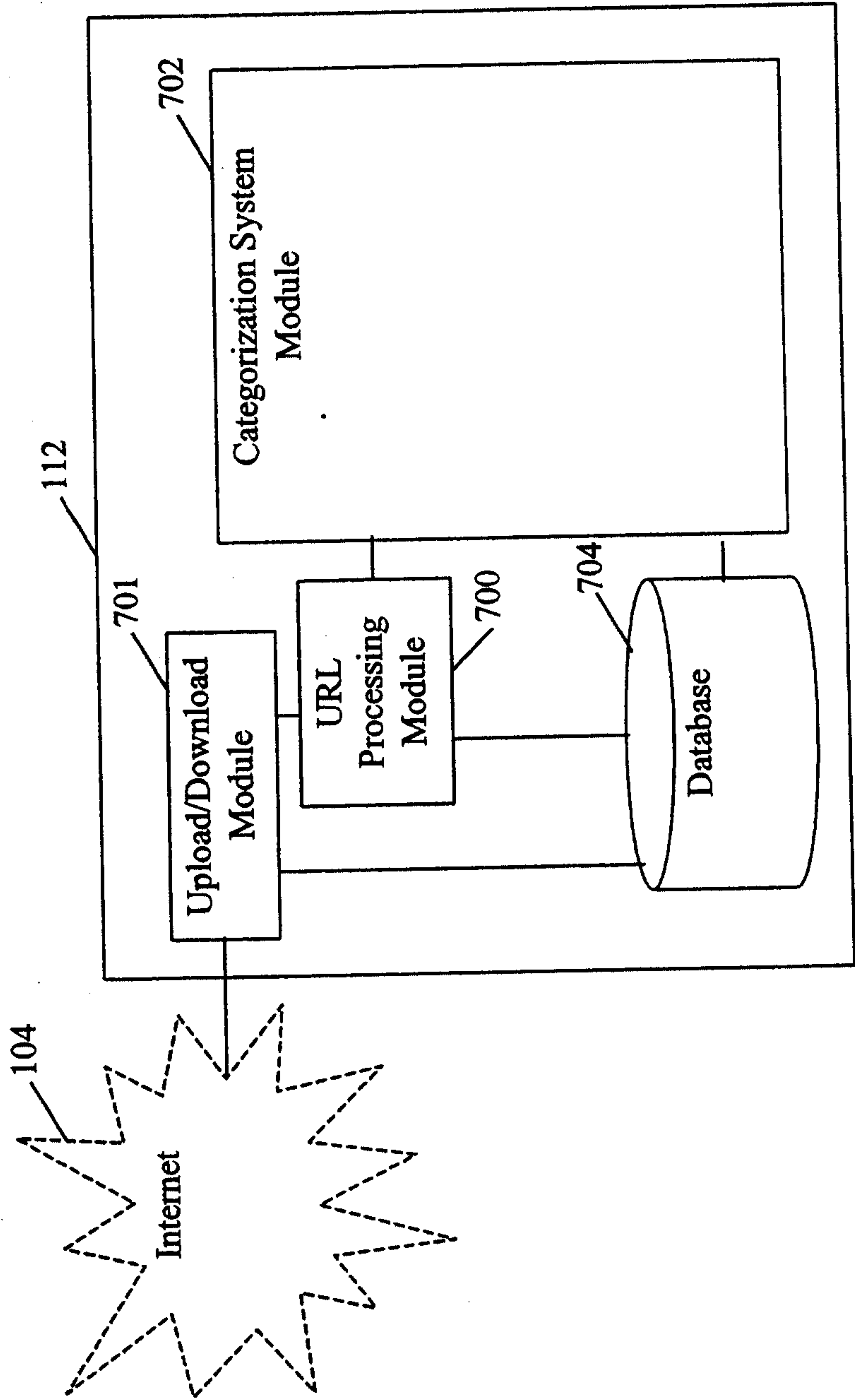


Figure 4

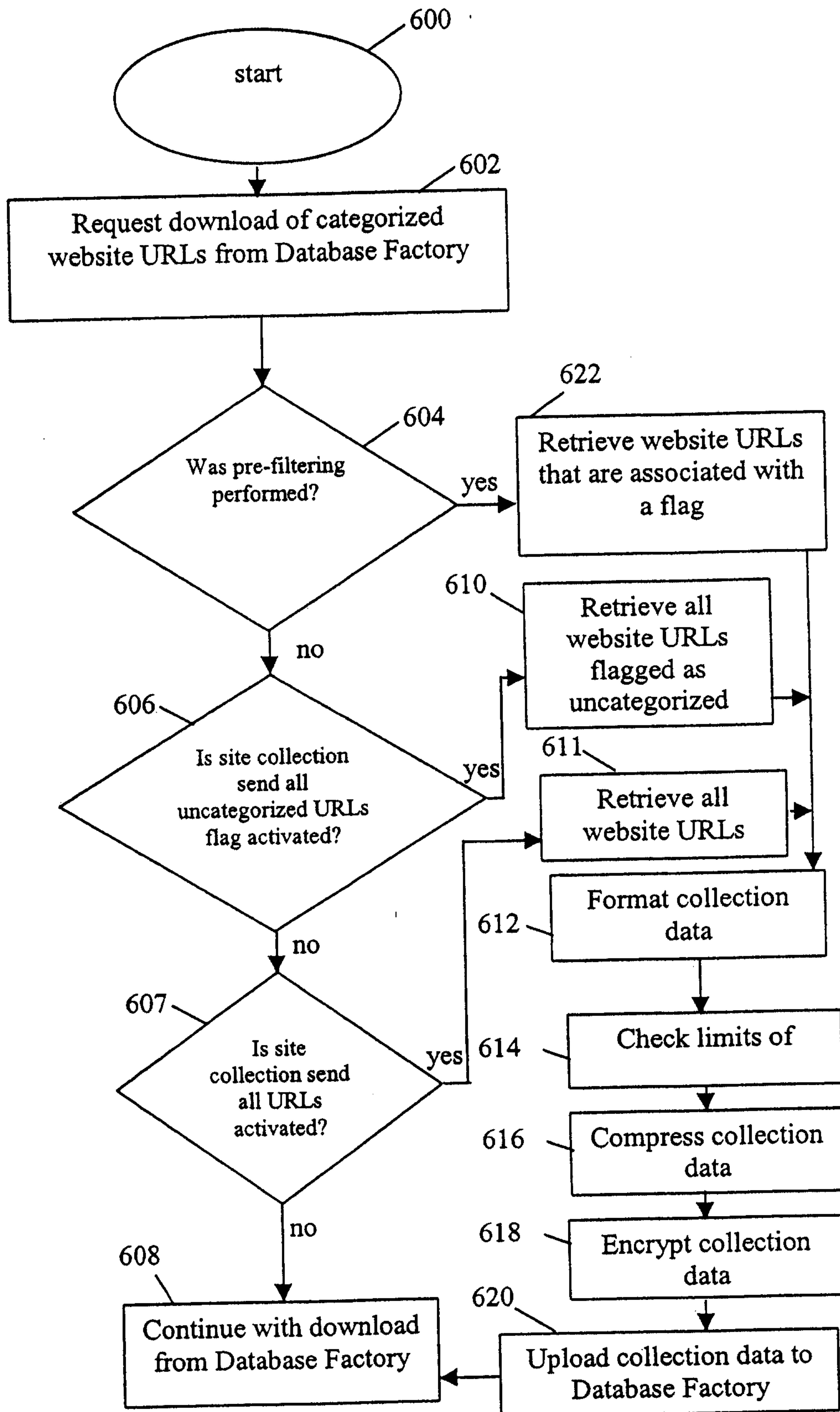
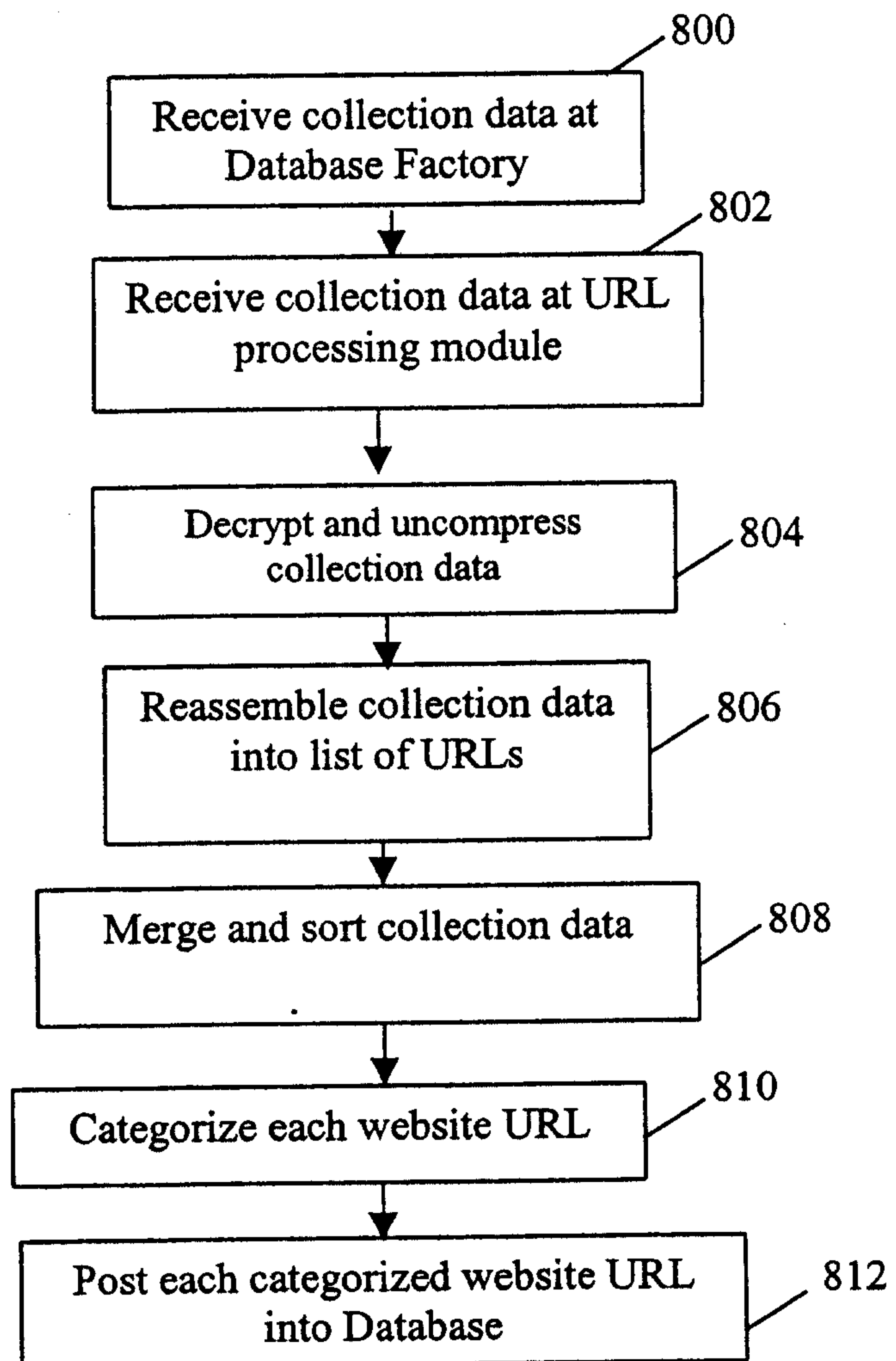


Figure 5

**Figure 6**

