PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/30

A1

(11) International Publication Number: WO 00/68837

(43) International Publication Date: 16 November 2000 (16.11.00)

(21) International Application Number: PCT/US00/12396

(22) International Filing Date: 5 May 2000 (05.05.00)

(30) Priority Data:
60/133,201 7 May 1999 (07.05.99) US

(71) Applicant: SEARCHLOGIC.COM CORPORATION [US/US]; Suite 310, 1800 30th Street, Boulder, CO 80301 (US).

(72) Inventors: STARZL, Timothy, W.; Suite 310, 1800 30th Street, Boulder, CO 80301 (US). STARZL, Ravi, S.; Suite 310, 1800 30th Street, Boulder, CO 80301 (US).

(74) Agent: HAMRE, Curtis, B.; Merchant & Gould P.C., P.O. Box 2903, Minneapolis, MN 55402-0903 (US).

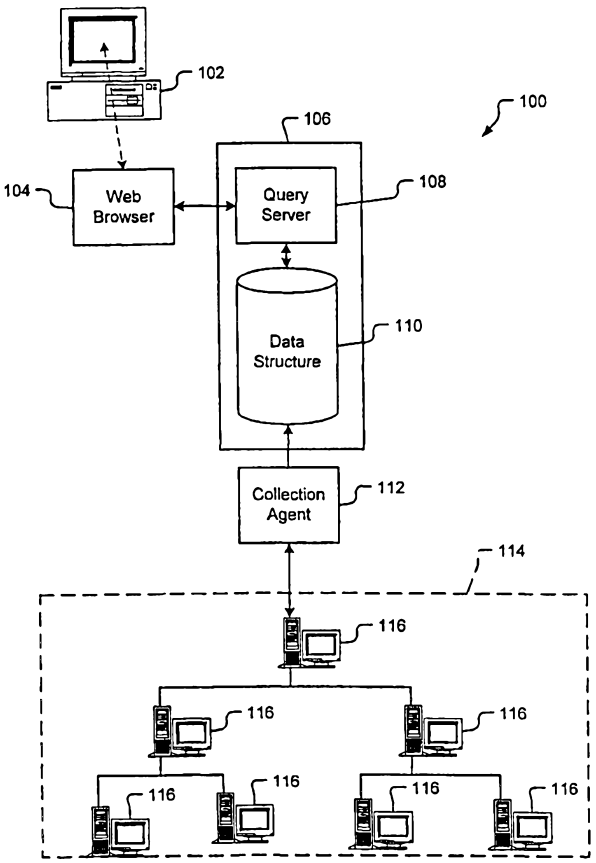
(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD AND SYSTEM FOR CREATING A TOPICAL DATA STRUCTURE

(57) Abstract

An automated system (200) and method for creating a topical data structure of documents or other items from an inter-linked system of documents, such as the WEB (114) and/or the Internet. The data structure (110) can then be searched using conventional means information to generate highly relevant results. The system (200) automatically discovers and collects topically relevant information by analyzing each document traversed and determining its topical relevancy prior to addition to the data structure (110). The topical relevancy information can further be used to confine traversal paths that are more likely to be topically relevant.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD AND SYSTEM FOR CREATING A TOPICAL DATA STRUCTURE

Technical Field of the Invention

The present invention relates to processes for discovering and collecting information located in an inter–linked environment such as the Internet and the World Wide Web ("Web") or in other archived, repository, database or stored information environment where the information is in a digital format, and is accessible electronically. More specifically, the present invention relates to improving both the topical or class relevancy of the information collected and the amount of relevant information collected from these environments.

Background of the Invention

The World Wide Web is an extremely large, inter–networked data system connecting hundreds of millions of informational sites and documents and is growing daily. The inter–linked relationships between these sites create a dynamic system of enormous complexity. Despite the information or "content" dependent utility of the Web, the existing Internet addressing system does not locate or identify sites based on their information content. Thus, one of the persistent problems associated with the Web is finding useful information. Indeed, while the rich, decentralized, dynamic and diverse nature of the Web can make casual Web surfing enjoyable, it has made serious navigation aimed at finding specific information extremely difficult.

In response to this problem, several types of Internet/Web navigation, location, finding or searching resources have evolved to facilitate the finding of sites based on content. One such resource relates to an automated information retrieval system, often referred to as an Internet or Web "search engine." Typical Web search engine systems use automated collection agents, software programs generally called "spiders", to automatically traverse the Web to discover and collect any accessible information sources. A spider automatically traverses the Web's hypertext link structure, recursively retrieving documents, pages, or resources that are discovered. These spiders return Web documents or document addresses (URLs) to a confined data structure or Information Retrieval System. Spiders may retrieve all or only a portion of a document such as the headers or metatags, or may only collect the page address. The term spider is understood here to include automated user agents, call

utilities, Web robots, bots, autonomous and mobile agents dedicated to this function, and other like utilities

The resources collected by the spider are typically stored in a database as part of an Information Retrieval System. The term "Information Retrieval System" or "IR system" refers to the data structure–based functions of storage, ordering, and presenting of previously discovered and collected information, as distinct from the processes of discovery and collection of data from the Web. Typically, in response to user supplied queries, the IR systems sort these previously collected documents, or representations of documents, and associate them with their Internet, archive, or other address for presentation to the user. For most Web Search Engine IR systems, all of the web pages that the spider discovers and collects are searched and sorted in an undifferentiated manner. Other such IR systems differentiate content within the IR data structure itself for more efficient ordering, storage and quicker access.

In response to a user supplied query, Web search engine IR systems typically analyze the collected Web documents using filters to perform a calculation and produce a relevance score. This score may be based on a number of factors such as the number of search or query terms that appear in the document, where and how often they appear. Some systems use other criteria such as number of links or frequency of use as scoring criteria. Usually a low score indicates the document is not relevant to the user query, and a high score indicates that it is likely to be relevant.

One drawback associated with typical spiders is that they are unconstrained, seeking to retrieve all accessible pages, and will inherently discover and collect large numbers of documents, resulting in very large data structures. Queries to such large data structures often return very large numbers of responses, including many non–relevant documents. Indeed, published and practical search engine experience has shown that much of what is returned to the user is not beneficial. One reason for the return of non–relevant documents relates to the frequent multiple meanings of words, the use of vernacular and idiomatic expressions, and the reliance of meaning on syntactic, semantic, and pragmatic structure. In a large population of documents there is an increased chance of words co–occurring without actually having a common meaning or relationship. Additionally, some documents merely summarize or reference other documents, and some documents, such as full text books or reference sources may, by their nature contain a significant amount of information not relevant to the intention of the initial query to the search engine.

Another problem associated with unconstrained spiders is that they typically cannot effectively traverse the entire Web. The longer a spider conducts an unconstrained traversal the larger the accumulation of found hyperlinks. In principle the entire Web can be discovered in this manner, however, in practice the process is intractable, and system resources are rapidly used up. Problems associated with practical spidering of the Web include the large and highly variable number of links on different pages, the high level of self-referential and recursive linking architectures, and cyclical link paths. Web search engines typically assign rules or "policies" to limit spider traversal, effectively causing significant portions of the Web to be left undiscovered.

One response to the lack of relevancy in Web search engine returns has been the development of "Web directories." These directories are manually created by people who examine each page or resource and determine whether the resource should be included in the directory. Web directories are distinguished from search engines in that they only collect or accept content that is relevant to a topic or category within the directory. Although each directory typically has highly relevant resources, the throughput of manual processing creates directories that are unsatisfactorily small, on the scale both of the total Web and when compared to the size of Web search engines. Moreover, since people must manually perform the task of accepting or rejecting each and every resource, the cost of maintaining and updating the directories is significantly high.

It is with respect to these considerations and others that the current invention has been made.

Summary of the Invention

The present invention relates to an automated system and method for creating a topical data structure, which can then be searched using conventional IR means. The term "topical" relates to the concepts of human-derived topic, class, category, grouping, natural grouping, taxonomic grouping, taxon, theme, cluster, or subject, and which may be identified through measures of relatedness, similarity, likeness, clustering, nearness, or other like measures. Since the data structure is topical, i.e., primarily restricted to topically related information, the results from the search show substantially improved query relevancy. Additionally, since the discovery and collection system is automated many more documents can be incorporated into the data structure, and the cost of generating and updating the data structure is relatively low.

In accordance with preferred aspects, the present invention relates to a system or method for discovering and collecting information from an inter–linked system of documents, such as the Web and/or the Internet. The system or method recursively traverses the system of inter–linked documents, analyzes each document traversed to extract a signature for each document, wherein the signature is related to the content of the document, and then compares the signature for each document to predetermined signature criteria related to that topic to determine the relevancy of each document to that topic. Once the relevancy of the document is determined, the method adds or combines relevant documents to create the topical data structure. The analysis and comparison is done by a filter system that may be external to an information retrieval system where the topical data structure resides.

In accordance with other preferred aspects, the system or method utilizes a spider to traverse the Web, wherein the spider feeds document information to the filter system. The spider may further be combined with a filter system to deliver topically relevant documents to the data structure and to confine the traversal paths taken by the spider. Thus, the spider may receive relevancy information about document signatures so the spider may determine whether paths are relevant or conforming. The spider may further elect to traverse only relevant paths based on this determination. The spider may further be configured to jump a predetermined number of irrelevant documents in determining whether paths are relevant.

In accordance with other aspects, at least one filter may determine relevancy based on a predetermined scale and provides relevancy information according to the predetermined scale. Additionally, more than one filter may be used to determine the relevancy of each document. This information can then be further evaluated to determine whether additional analysis is necessary in determining whether to include or reject a document from the topical data.

In an embodiment of the invention the predetermined criteria is derived from a collection of sample documents to determine topical signatures and preferably using some form of analysis, such as lexical, relational, statistical, linguistic, or inferential content analysis.

The constrained results produced may subsequently be used in any IR system, such as a document search engine, a hierarchical directory, a vector space construct, any clustering algorithm driven data structure, array or construct, or any data storage and query format.

The invention may be implemented as a computer process, a computing system or as an article of manufacture such as a computer program product. The computer program product may be a computer storage medium readable by a computer system and encoding a computer program of instructions for executing a computer process. The computer program product may also be a propagated signal on a carrier readable by a computing system and encoding a computer program of instructions for executing a computer process.

A more complete appreciation of the present invention and its improvements can be obtained by reference to the accompanying drawings, which are briefly summarized below, to the following detail description of presently preferred embodiments of the invention, and to the appended claims.

Brief Description of the Figures

Fig. 1 is a block diagram of the computer system shown in Fig. 1 connected to server computers through a computer network.

Fig. 2 is a block diagram of a computer system that may be used to implement a method and apparatus embodying the improved spider of the present invention.

Fig. 3 illustrates the functional components of a prior art Web discovery and collection system.

Fig. 4 illustrates the functional components of a Web discovery and collection system of the present invention.

Fig. 5 illustrates the functional components of a Web discovery and collection system of an alternative embodiment of the present invention.

Fig. 6 is a flowchart illustrating the operational characteristics of an embodiment of the invention.

Detailed Description of the Invention

The logical operations of the various embodiments of the present invention are implemented (1) as a sequence of computer implemented steps or program modules running on a computing system and/or (2) as interconnected hardware or logic modules within the computing system. The implementation is a matter of choice dependent on the performance requirements of the computing system implementing the invention. Accordingly, the logical operations making up the embodiments of the present invention described herein are referred to alternatively as operations, steps or modules.

An interconnected computer system 100 that may incorporate the present invention is shown in Fig. 1. The client computer system 102 operates a traditional browser application 104. The browser application 104 communicates with an information retrieval system 106, which is located on either computer system 102 or on another server computer system (not shown). The retrieval system 106 comprises a suitable query server 108 and a data structure 110, preferably a database or text base. The information retrieval system 106 communicates with a collection agent 112 for collecting information from the Web 114 and storing those collected resources in the data structure 110. The data structure stores the various resources, and may be configured to index or otherwise sort the information for future reference. The query server 108 receives a query from browser 104 and uses the query to search the data structure for relevant information. Once the relevant information is retrieved, that information is then presented to a user of computer 102 through the interface that is displayed through the browser 104.

The collection agent 112 traverses the Web 114, which generally has a network of informational sites that are linked via the hypertext transfer protocol (HTTP). Each of the sites resides on a server computer system, such as server computer systems 116 as shown in Fig. 1. As discussed in more detail below, the collection agent 112 in various embodiments of the present invention is capable of differentiating the various web resources during traversal so that the resulting data structure comprises mostly resources that are relevant to a particular topic. Thus, the query server 108 generally returns only relevant information. Moreover, since the data structure is topic specific, the query server is better able to perform advanced algorithms on the data structure to retrieve highly relevant information and can present that information accordingly. Further still, since the data structure is constrained to containing only topical information, the collection agent must accept and collect far fewer documents than prior art collection agents, and thus is able to traverse a significantly larger portion of the Web before the IR system reaches capacity, further improving the results provided by the query server.

In one embodiment of the invention, the computer 102 is a desktop computer system. In alternative embodiments, the invention is used in combination with any number of other computer systems or environments, such as in handheld computer environments, laptop or notebook computer systems, multiprocessor systems, micro–processor based or programmable consumer electronics, network PCs, mini computers, main frame computers and the like. The invention may also be practiced

in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network in a distributed computing environment, programs may be located in both local and remote memory storage devices.

The computer 102, as well as most computer systems 116, incorporates a system of resources for implementing an embodiment of the invention, such as the system 200 shown in Fig. 2. The system 200 incorporates a computer 202 having at least one central processing unit (CPU) 204, a memory system 206, an input device 208, and an output device 210. These elements are coupled by at least one system bus 212.

The CPU 204 is of familiar design and includes an Arithmetic Logic Unit (ALU) 214 for performing computations, a collection of registers 216 for temporary storage of data and instructions, and a control unit 218 for controlling operation of the system 200. The CPU 204 may be a microprocessor having any of a variety of architectures including, but not limited to those architectures currently produced by Intel, Cyrix, AMD, IBM and Motorola.

The system memory 206 comprises a main memory 220, in the form of media such as random access memory (RAM) and read only memory (ROM), and may incorporate or be adapted to connect to secondary storage 222 in the form of long term storage mediums such as hard disks, floppy disks, tape, compact disks (CDs), flash memory, etc. and other devices that store data using electrical, magnetic, optical or other recording media. The main memory 220 may also comprise video display memory for displaying images through the output device 208. The memory can comprise a variety of alternative components having a variety of storage capacities such as magnetic cassettes memory cards, video digital disks, Bernoulli cartridges, random access memories, read only memories and the like may also be used in the exemplary operating environment. Memory devices within the memory system and their associated computer readable media provide non–volatile storage of computer readable instructions, data structures, programs and other data for the computer system.

The system bus 212 may be any of several types of bus structures such as a memory bus, a peripheral bus or a local bus using any of a variety of bus architectures.

The input and output devices are also familiar. The input device can comprise a small keyboard, a mouse, a microphone, a touch pad, a touch screen, etc.

The output device can comprise a display, a printer, a speaker, a touch screen, etc. Some devices, such as a network interface or a modem can be used as input and/or output devices. The input and output devices are connected to the computer through system buses 212.

The computer system 200 further comprises an operating system and usually one or more application programs. The operating system comprises a set of programs that control the operation of the system 200, control the allocation of resources, provide a graphical user interface to the user, facilitate access to local or remote information, and may also include certain utility programs such as the email system. An application program is software that runs on top of the operating system software and uses computer resources made available through the operating system to perform application specific tasks desired by the user. In general, applications are responsible for generating displays in accordance with the present invention, but the invention may be integrated into the operating system.

In order to better understand the present invention, a brief discussion of a prior art discovery and collection system, as shown in Fig. 3, is provided. The information retrieval system 302, which is similar to informational retrieval system 106 (Fig. 1), communicates with a spider 304. As discussed in the Background Section, the term, "spider," is understood here to include prior art automated user agents, call utilities, Web robots, bots, autonomous and mobile agents dedicated to the function, and other like utilities. Spider 304 automatically traverses the Web's hypertext structure, recursively retrieving all web documents 306 that are found. The documents 306 may be actual textual documents, images, pages, or other resources found on the Web, as well as their addresses, and are referred to hereinafter as either documents, pages or resources.

The web documents 306 are stored in data structure 312, hereinafter referred to as the database or data structure, of the information retrieval system 302. The database 312 may have a filter so that specific predetermined types, structures or formats of pages are not accepted in the database (e.g. duplicates, spam pages, by character set, by domain). Alternatively, the database may not have a filter, but in any case may create an index of the information stored in the database 312. The information is available to the user through user interface 314, which may comprise a browser and query server, such as the browser 104 and query server 108 shown in Fig. 1.

Importantly, the prior art spider 304 does not differentiate documents 306 based on topical content. Instead, each document that is traversed is returned to the database 312, creating a large, undifferentiated collection of items.

An embodiment of the invention is shown in Fig. 4. The embodiment uses a discovery and collection system 400 having a spider 402 and a topical content filter 404. The spider 402 is any software program or system capable of traversing the Web, and which, in this case, must also be capable of interfacing with and using a linguistic, lexical or other text filter such as filter 404. The content filter 404 analyzes the documents 306 returned by the spider 402 and accepts or rejects each document based on predetermined topical content criteria. The criteria may be a lexical or linguistic signature or some other basis. Based on the acceptance or rejection of these documents by the content filter 404, the database 312 comprises topical information. In an alternative embodiment (not shown) the filter 404 is integrated with the information retrieval system 302 in a manner that pre-filters the content accepted by the information retrieval system 302. In each case, the information returned to the database 312 has been differentiated, based on topical content, from other information on the Web through the use of the system 400.

Another embodiment of the present invention is shown in Fig. 5. This embodiment may operate in conjunction with information retrieval system 302, having the same type of database 312 and user interface 314 as described above in conjunction with Figs. 3 and 4. Moreover, the information retrieved and stored in the database 312 is topical, i.e., content based as described above in conjunction with Fig. 4. However, the embodiment shown in Fig. 5 comprises a unique filtering spider 500.

The filtering spider 500 traverses the Web and performs its own filtering analysis on each document or site that the filtering spider 500 encounters. By performing the analysis during traversal, an embodiment of the filtering spider 500 can be configured to elect different paths of traversal, thereby only traversing selected web documents 306. Thus, the spider 500 may avoid paths that are less likely to produce relevant information and concentrate on paths that are more likely to produce relevant information. This process is referred to herein as "link-tunneling." One purpose of the link-tunneling approach is to limit the content presented for incorporation into a topical data structure to only material that displays pre-specified characteristics associated with that topic. Additionally, targeted "link-tunneling" methods of traversal can capture the topic knowledge of site authors, as

expressed in their linking decisions. The effect of this system is the selection of a constrained population of resources for inclusion in a data structure, around which a topical or subject oriented information retrieval system can be defined.

The embodiments shown in Figs. 4 and 5 retrieve relevant information using differentiating spider systems that incorporate "linguistic signatures". A linguistic or lexical signature relates to any extractable attribute or representation of content, i.e., subject matter, that provides a basis for document or subject recognition or differentiation and usually beyond that provided by the simple presence or absence of a keyword, a group of keywords, or Boolean expression. Designed constructs of keywords representing a subject or topic may be extracted or generated that reflect this equivalent function. Additionally, differentiation of discovered material by comparison to a linguistic signature or template, may be topically or categorically related by a predefined linguistic, lexical, textual, semantic, syntactic, mythographic, semiotic, pictographic, hieroglyphic, graphic, structural, hybrid or other content related attributes.

The ability to differentiate, select or reject a document on the basis of its content requires the use of topical signature data for differentiation. The discovery or development of this signature refers to any of a class of processes for the mathematical, logical, or linguistic extraction and characterization of document, atomic, molecular or elemental components (words, lexies, associative patterns, frequencies, word clusters, word class relationships, etc.) to produce a set of differentiating representations or characteristics. These representations are referred to as "linguistic signatures" in this disclosure. The methods referenced here include: lexical analysis, semantic analysis, syntactical analysis, textual analysis, clustering analysis, auto–categorization, vector analysis, statistical analysis, heuristics, pragmatic methods and/or any models, algorithms or relationships using these methods. Also included within a definition of the system is the application of a linguistic signature, derived or extracted by any means, by the filter 402 or filtering spider 500 as a conformity test for unknown, heterogeneous documents.

Differentiation by "linguistic signature" according to subject matter of a web document 306 is to be understood as the automated assignment of document membership or the identification of non–membership within a pre–defined subject, category, class, or topic area. Acceptance, differentiation or rejection may be into, or in reference to, any topical, subject, categorical, hierarchical, relational or other

organizational system, scheme, ontology, taxonomy, or concept hierarchy, using any relatedness–based classification measure or method.

A class, category, subject or topic may be identified by human judgment or agency, or may be identified as a measure of relatedness, similarity, or clustering of a group of documents. A class, category, subject or topic "linguistic signature" may be determined in substantially the same manner as described above for the determination of document "linguistic signature" as applied over a sufficiently large group of documents judged to be members of the class, category, subject or topic so as to allow for the creation of a representative signature. The method includes any method for the development or identification of lists, strings, arrays, files, algorithms, expressions, collections or groupings of such elements that are characteristic of the subject class, category, subject or topic.

Fig. 6 illustrates the operational flow process 600 which relates to an embodiment of the invention. Process 600 begins with traverse operation 602 which traverses the Web, or another inter–linked data structure, using a provided link, Uniform Resource Locator (URL) or other address information (hereinafter "links"). Preferably a first link is provided to operation 602, indicating the first site to visit. The spider 304 or the filtering spider 500 carries out the traversal. Operation 602 may mark the link in some manner so that the process can recognize, at a later time, that the link has been analyzed at some earlier time. Similarly, this step may analyze the link to determine if the link has been marked in the past. If the link has been analyzed, the process may elect to either re–analyze the document or recursively determine the next link to analyze. Other embodiments utilize tables of links. A first table stores potential links. That is, once a page is visited all links on the page are put in the first table, where they remain until removed by the process. A second table is used to store traversed links, and another to store topically rejected links. By comparing a link in the first table to those in the second table, the process can determine if it has been traversed.

Once at the given site, page capture and decomposition operation 604 retrieves the document located at the site and parses the information. This operation may involve an in–depth lexical analysis, or other analysis of the document to extract a "signature" for the document. The signature is reflective of the subject matter or content of the document.

Next, operation 606 performs a comparison on the signature that has been generated by operation 604. The filtering operation 606 may be any method suitable

for the comparison of the document "linguistic signature" to a pre–determined class, category, subject or topic "linguistic signature", so as to determine within some specified level of precision, the membership of the subject document within the subject class. The method references any means suitable to allow a determination of whether a document falls within, or out of, a particular pre–specified class, topic, subject or category. In particular, in an embodiment of the present invention, the filtering operation 606 utilizes a linguistic signature to determine conformity of collected data sets to preexisting human–derived topic, category, class or subject cognitive criteria. For example, one use for this system is the automated production of an information resource similar to a content–based Web Directory.

The filtering step 606 may compare the document signature with a predefined signature to produce a weighted score related to the probable degree of relevance for the document.

In order to determine a predefined signature, personnel responsible for the data structure may decide what topic(s) the data structure should include and what untargeted topic(s) may use language similar to that of the target topic(s). Using information related to the language of the targeted topic and not related to untargeted topics, a definition of the goals for the inclusion filters and exclusion filters for the topical data structure is generated. As an example, a topical database for the topic of golf, i.e., the game, may require the inclusion of documents having the word golf in them, unless they refer to cars named GOLF which are made by Volkswagen.

This process may involve the selection by the database collection personnel of one or more electronic texts as representative of the topic selected. These documents may be manually selected or automatically selected from a web directory or other search resource that can provide topically representative documents. A class, category, subject or topic may be identified by human judgment or agency, or may be identified as a measure of relatedness, similarity, or clustering of a group of documents. In addition, for some topics it may be important to select documents representative of the exclusions that are identified by the database personnel and to place these into separate corpora for analysis. Such topics and documents may use overlapping terminology but are not targeted by the topical database. Generally, more than one document will be required to form a corpus of documents for analysis. However, one document of sufficient length and topical specificity may also be used for the purpose of further analysis.

The topical document collections are then analyzed for a lexical signature. The ability to differentiate, select or reject a document based on its content requires the use of such signature data for differentiation. As described above, the discovery or development of this signature refers to any of a class of processes for the mathematical, logical, or linguistic extraction and characterization of document, atomic, molecular or elemental components (words, lexes, associative patterns, frequencies, word clusters, word class relationships, etc) to produce a set of differentiating representations or characteristics. Preferably, the sample documents are analyzed using some form of quantitative or semi-quantitative analysis beyond that provided by the simple presence or absence of a keyword, a group of keywords, or Boolean expressions that are derived by qualitative analysis of the topic by the database collection personnel. In addition, the relationships between words and non-lexical features of the document (graphics, encoding, hyperlinks) may also be analyzed for features of a signature.

A simple signature may be expressed as a simple list of keywords extracted from the representative document(s). In this case, it is preferable that a minimum of three keywords be used to provide the most basic data for a Boolean-logic-based filter for the presence or absence of keywords in any given document. Even under this simplest case, the previously mentioned quantitative and semi-quantitative methods should be employed to extract or assist in the extraction of meaningful lexical features of the signature.

The signature extraction process produces a series of features of the document. These features can then be applied within the topical filter. The filter process may involve application of the feature extraction process in reverse. However, the process for filter process does not have to be the same analysis as that used to extract the signature. For example, a keyword frequency analysis could be employed to extract the lexical signature and then those keywords could be employed in a Boolean filter, a co-association matrix, or may be extended using a semantic nearness function.

Not every type of extracted feature in a signature will be able to be employed in every type of possible topical filter. Therefore, if a particular type of topical filter is to be used, it is important to make sure the feature extraction method used will produce features that are compatible with the filter and vice versa. Moreover, more than one filter may be employed in this step of the process. An array of topical filters may be employed for document analysis for both the inclusion and exclusion

of pages into the topical database. Additional topical filters may also generate lexical metrics about the pages at this step in the process to be associated with the document into the topical database. These additional topical filters need not necessarily be part of the acceptance/rejection of the document into the topical database.

Following the filtering operation 606, the process determines, at step 608, whether the document meets the requisite criteria to be accepted (included) or rejected (excluded). In one embodiment, the filtering step produces a topical relevancy score and operation 608 compares the topical relevancy score against a minimum threshold value. If the score for the document is above the minimum threshold value, the document is determined to meet the criteria. In such a case, flow branches YES and the document is added to the conforming list at add operation 610.

Once a document is added to the conforming list at 610, identify link act 612 identifies the next link, typically a link on the conforming page. This link is provided to operation 602 and the process begins again. If there are no links on the conforming page that have not been analyzed, then the identify link act 612 recursively determines the next link to analyze.

If the document is determined to not conform to the predetermined criteria at operation 608, such as when the score is below the minimum threshold, the process flow branches NO to determination step 614. Determination step 614 determines whether pages on the non-conforming page should be analyzed. This determination involves a comparison of the depth level for non-conforming documents to a predetermined number of levels to be searched. For example, the process may be configured to not analyze any sub-links on a non-conforming page and therefore the predetermined number of levels would equal one. In such a case, determination step 614 would always branch YES since the current document is, by definition at level one. However, if the predetermined level was set to two, then the sub-pages of a first non-conforming page are analyzed. But, if any of those pages are non-conforming, then their sub-pages are not analyzed, since they are at level two.

That is, to a spider with a specific lexical filter, and an instruction set to reject all links that are not immediately followed by a relevant page, the immediately non-conformal links become essentially invisible, leaving a simpler architecture to map and record. Additionally, the database and information retrieval systems are substantially unburdened by this method. The risk with this election is that some

conformal sites are missed because they are not directly linked to another conformal site. Moreover, two-level traversal matching relates to allowing two levels of conformity testing prior to rejection of the link thread. This setting allows the system to "jump" over sites that otherwise would stand as impermeable barriers to discovery.

Similarly, three-level (and up) traversal matching further allows for conditional or transient acceptance of non-conformal links can be specified for the spider. In such a case, more links may be discovered. However each level retained requires additional processing and memory capacity, and contributes to the growth of the link-validation burden. Decisions as to the number of levels to be traversed will depend upon the density of information sources, and the degree of completeness desired for the topical information space being developed. Such a system allows for the retention of link trails or threads through three or more non-conformal layers. It is important to note that continued traversal of the link thread does not imply the retention or recording of the non-conformal pages to the information retrieval data structure. These pages are rejected, but records of their linking relationships are retained through the prescribed number of layers. The system may retain connection to an indefinitely large link thread despite the absence of conformal pages. However this approach requires progressively larger computing resources.

If the document is determined, at determination step 614 to not be in the last level for analysis, then flow branches NO to mark level operation 616. Mark operation 616 provides for the identification of the link as a sub-link of a non-conforming page, thereby allowing later analysis of link levels relative to a non-conforming parent page. Following operation 616, identify next link 612 identifies the next link to be analyzed and passes it to operation 602 and the process 600 is repeated. If the document is determined, at determination step 614 to be in the last level for analysis, then flow branches YES to determination module 618. Determination module 618 determines whether all the links have been analyzed. This module may be configured to stop after a predetermined number of documents have been analyzed or collected. Otherwise, this module may be configured to only quit once each and every document in the system has been analyzed. If there are more documents to be analyzed, which is typically the case in large-scale information systems such as the Web, then flow branches NO to operation 620. Operation 620 recursively determines the next link and passes it to operation 602 and the process 602 is repeated. If determination module has determined that all the

potential links have been analyzed, then flow branches YES to end step 622 and process 600 ends.

In an embodiment of the invention, the conforming list created at operation 610 comprises the link or URL, for all the items that are added to the topical database 312, Figs. 4 and 5. In an alternative embodiment, each time a page is determined to be conforming at step 608, the page is added to the list at 610, and is then forwarded to an additional processing module, (not shown). This module performs a more intensive analysis on the document, as opposed to merely comparing a signature for the document to a template. The full analysis may comprise lexie identification, grouping, correlation, pattern recognition, pattern matching, fitting and other analysis techniques. Following this analysis, the page is either determined to be in or out of topic. If it is out of topic, the page is rejected as described above at step 608 and flow branches to operation 614. If it is determined to be in topic, then the page is forwarded to the topical database. Additionally, the page may be forwarded to a topical hierarchy directory interface and potentially a learning engine of strategy level modeling or a neural network for pattern recognition.

Once the database has been populated with topically related information, the information retrieval system may operate in the conventional manner. However, since only topically related information exists in the database, the system is more likely to produce relevant information. Also, since the database is not filled with a significantly large amount of irrelevant data, the results of query searches are more complete as well. That is, since the invention allows for the discovery and inclusion of defined subsets of resources, differentiated from other unrelated resources, in an automated or semi-automated manner, a high relevancy resource is generated. Because the system is automated, the depth or completeness achieved by this system can be as great or greater than provided by a typical, prior-art Web directory approach. The sources discovered and collected by this process may be incorporated into any conventional information retrieval system, may be subject to further processing, ordering, characterization, or organization, and may be presented as either a directory hierarchy or as a searchable data structure.

A significant benefit derived from the present invention relates to the fact that the constrained content approach removes a very large portion of the processing burden from the information retrieval internal system, placing it instead on an exogenous filter system. Additionally the reduced number of entries, and the tighter

linguistic and topical focus of the entries, allows for specialized and more efficient processing functions.

In addition to advantages already discussed for discovery, collection and storage topical differentiation also has important advantages in the areas of information organization, refinement, and presentation. The system may take advantage of "natural" or common usage methods for organizing collected information derived from the topic area itself. Further, the specialized uses of language often associated with specific topics can be used by this system as guides and markers to refine and differentiate topical groupings. In comparison, for global systems that must integrate many or all subjects or topics, this specialized usage is a significant contributor to the noise and imprecision within the process. In addition, the use of a topical format lends itself readily to thematic graphical and design expression for display and presentation within the context of the specific topic.

The invention disclosed here is distinct from prior teaching within this field in that it parses or segments the processing of information into separate pre-acceptance and information retrieval system stages, resulting in a substantial and useful change in the processing profile and capabilities for large scale Web or Internet search resources.

Another aspect of this system is the ability to control the degree of precision used to select or reject links or documents. This is accomplished by selecting the degree of precision of the linguistic signature applied, and by the stringency of conformity required for acceptance. Additionally this system allows for the ability to specify immediate rejection of a link thread on the basis of page non-conformity or to allow the link thread to be explored despite page non-conformity. Links may be followed despite non-conformal page status for any specified number of steps or layers, or indefinitely, without the collection of non-conformal pages, so as to discover discontinuous regions (non-topically inter-linked) of a topical information space. This method allows the system to "jump" over intervening or blocking pages to any prescribed depth.

Significant advantages are gained from a system using a data set that has been filtered or constrained during the discovery and collection process. The purpose of this approach is to insulate and protect the system from the burden of undifferentiated data sets. This method reduces the number of instances that the information retrieval system must process, prior to its being exposed to them. This approach also narrows and focuses the range of operations required of the

information retrieval system through the imposition of a topic, class, category or subject limitation. These modifications from standard search practice serve to substantially reduce the processing overhead and burden, allowing for substantial improvement in performance.

The present invention is the method, apparatus, computer storage medium or propagated signal containing a computer program for providing a discovery and collection system for creating a topical database as recited within the claimed attached hereto. Thus the present invention is presently embodied as a method, apparatus, computer storage medium or propagated signal containing a computer program for traversing the Web, analyzing sites and/or documents and delivering only relevant documents to a database. Additionally, the system may restrict or confine the paths that are traversed in the Web using relevancy information. While the invention has been particularly shown and described with reference to preferred embodiments thereof, it will be understood by those skilled in the art that various other changes in the form and details may be made therein without departing form the spirit and scope of the invention.

Claims

What is claimed is:

1. A method of creating a topical data structure of information located on an inter-linked system of informational documents, the method comprising:

recursively traversing the system of inter-linked documents;

analyzing each document traversed to generate a signature for each document, wherein the signature is related to the content of the document;

comparing the signature for each document to predetermined criteria to determine the topical relevancy of each document; and

adding topically relevant documents to the topical data structure.

2. A method as defined in claim 1 wherein the analyzing and comparing acts are performed by an external filter, the external filter being external to the topical data structure system.

3. A method as defined in claim 2 wherein the traversing act is performed by a spider and wherein the spider is separate from the filter, the spider supplying documents to the filter.

4. A method as defined in claim 2 wherein the inter-linked system of informational documents comprises non-relevant paths, and wherein:

the traversing act is performed by a spider and wherein the spider is combined with the filter;

the filter providing topical relevancy information to the spider, and the spider using the topical relevancy information to determine whether paths are non-conforming paths; and

the spider eliminating non-conforming paths from the recursive traversal of the system.

5. A method as defined in claim 4 wherein the inter-linked system comprises non-conforming documents and wherein the spider is configurable to jump non-conforming documents in determining whether paths are non-conforming.

6. A method as defined in claim 2 wherein more than one external filter is used to analyze and compare the documents to determine topical relevancy.

7. A method as defined in claim 2 wherein the filter determines topical relevancy based on a predetermined scale and provides topical relevancy information according to the predetermined scale; and wherein the act of adding topically relevant documents can be configured to add documents having different levels of topical relevancy.

8. A method as defined in claim 1 wherein the predetermined criteria is derived from a collection of sample documents to determine topical signature.

9. A method as defined in claim 8 wherein the predetermined criteria is derived from the sample documents using quantitative analysis.

10. A method as defined in claim 1 wherein the analyzing of each document for generating a signature is performed using quantitative analysis.

11. A method as defined in claim 10 wherein the analysis is lexical analysis.

12. A method as defined in claim 10 wherein the analysis is relational analysis.

13. A method as defined in claim 10 wherein the analysis is statistical analysis.

14. A method as defined in claim 10 wherein the analysis is linguistic analysis.

15. A method as defined in claim 10 wherein the analysis is inferential content analysis.

16. A method as defined in claim 1 wherein the inter–linked system is the Internet.

17. A method as defined in claim 1 wherein the act of comparing the signature of each document to predetermined criteria determines a minimum level of topical relevancy and wherein the method further comprises the followings acts prior to adding relevant documents to the data structure:

for each document satisfying the minimum level of topical relevancy, performing a second level of analysis on the document; and

based on the second level of analysis, determining whether the document is topically relevant and should be added to the data structure.

18. A discovery and collection system for analyzing documents found on an inter-linked system of documents, the discovery and collection system providing topically related documents to an information retrieval system having a searchable data structure, the searchable data structure providing users document information in response to user supplied queries, said discovery and collection system comprising:

a spider for discovering documents on the inter-linked system; and

a filter that receives document information from the spider, the filter determines the topical relevancy of each document received from the spider and the filter delivers the topically relevant documents to the information retrieval system.

19. A discovery and collection system as defined in claim 18 wherein the spider receives information from the filter related to the topical relevancy of documents and wherein the spider uses the topical relevancy information in determining paths of discovery in the inter-linked system of documents.

20. A discovery and collection system as defined in claim 18 wherein the inter-linked system of documents comprises conforming and non-conforming documents and wherein the documents have links to other documents, and wherein the spider evaluates the topical relevancy of documents and based on this evaluation determines the topical relevancy of paths within the inter-linked system and wherein the spider is adapted to eliminate topically irrelevant paths from the traversal process.

21. A discovery and collection system as defined in claim 20 wherein the spider is configurable to jump a predetermined number of non-conforming documents in the evaluation the topical relevancy of paths within the inter-linked system of documents.

22. A discovery and collection system as defined in claim 18 wherein the filter analyzes each document and generates a lexical signature, the filter compares the

lexical signature to predetermined criteria and determines the topical relevancy of each document based on the comparison.

23. A computer program product readable by a computer and encoding instructions for executing a computer process for creating a topical data structure, said process comprising:

traversing an inter–linked system of documents, wherein the inter–linked system comprises topically relevant and irrelevant document paths and wherein the system spans a plurality of server computer systems;

analyzing traversed documents for topical relevancy; and

combining topically relevant documents to create the topical data structure.

24. A computer program product as defined in claim 23 wherein the analysis of the traversed documents is performed on the server computer system having the document being analyzed.

25. A computer program product as defined in claim 23 wherein the traversing act is confined to topically relevant document paths.

26. A computer program product as defined in claim 25 wherein the topically irrelevant document paths comprise at least a predetermined minimum number of contiguously linked topically non–relevant documents.

27. A computer program product as defined in claim 23 wherein the predetermined criteria is a lexical signature generated by performing quantitative analysis on at least one sample document.

28. A computer program product as defined in claim 23 wherein the predetermined criteria is a lexical signature generated by performing semi–quantitative analysis on at least one sample document.

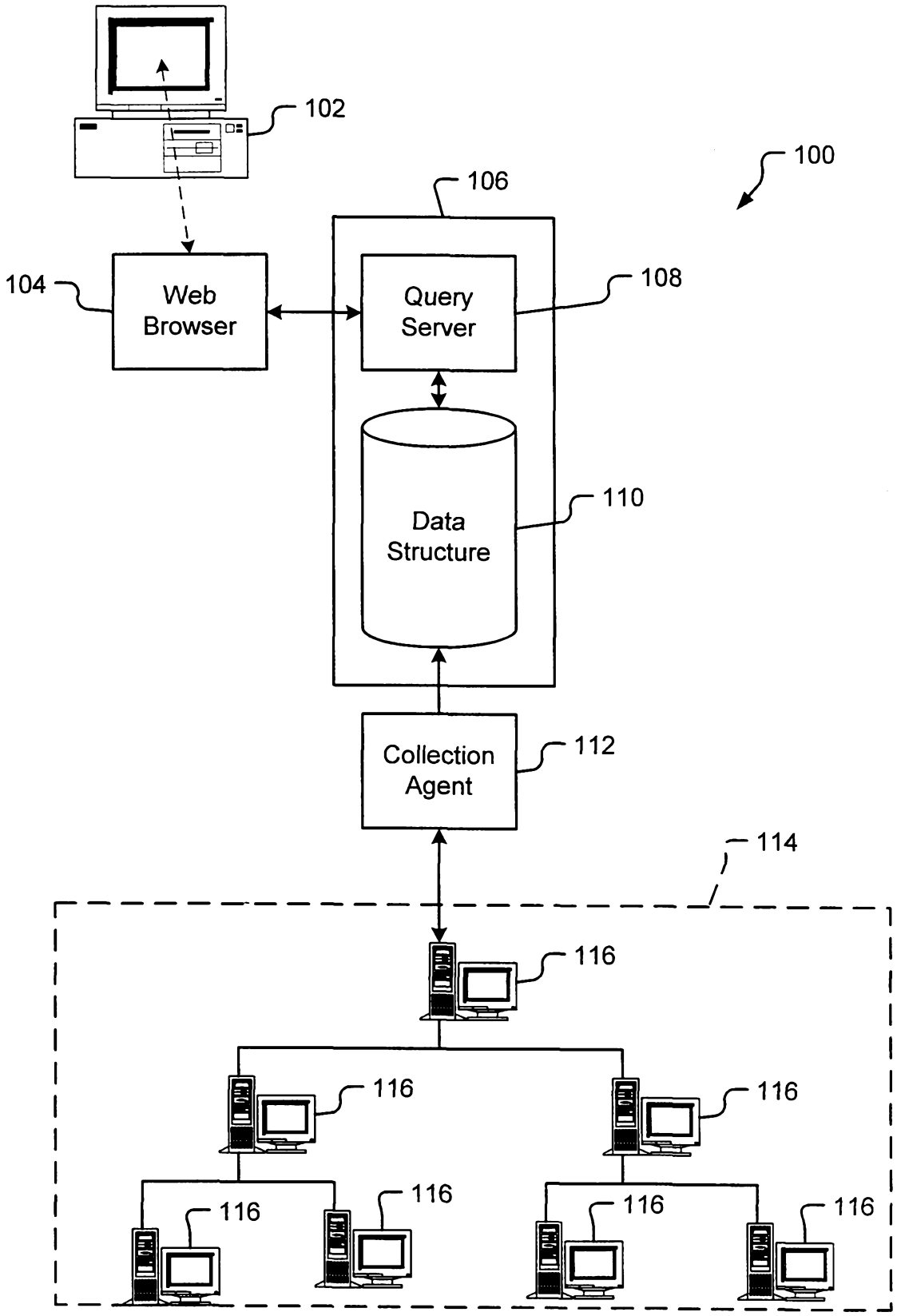

102
100
106
104
Web Browser
Query Server
108
Data Structure
110
Collection Agent
112
114
116
116
116
116
116
116
116

Fig. 1

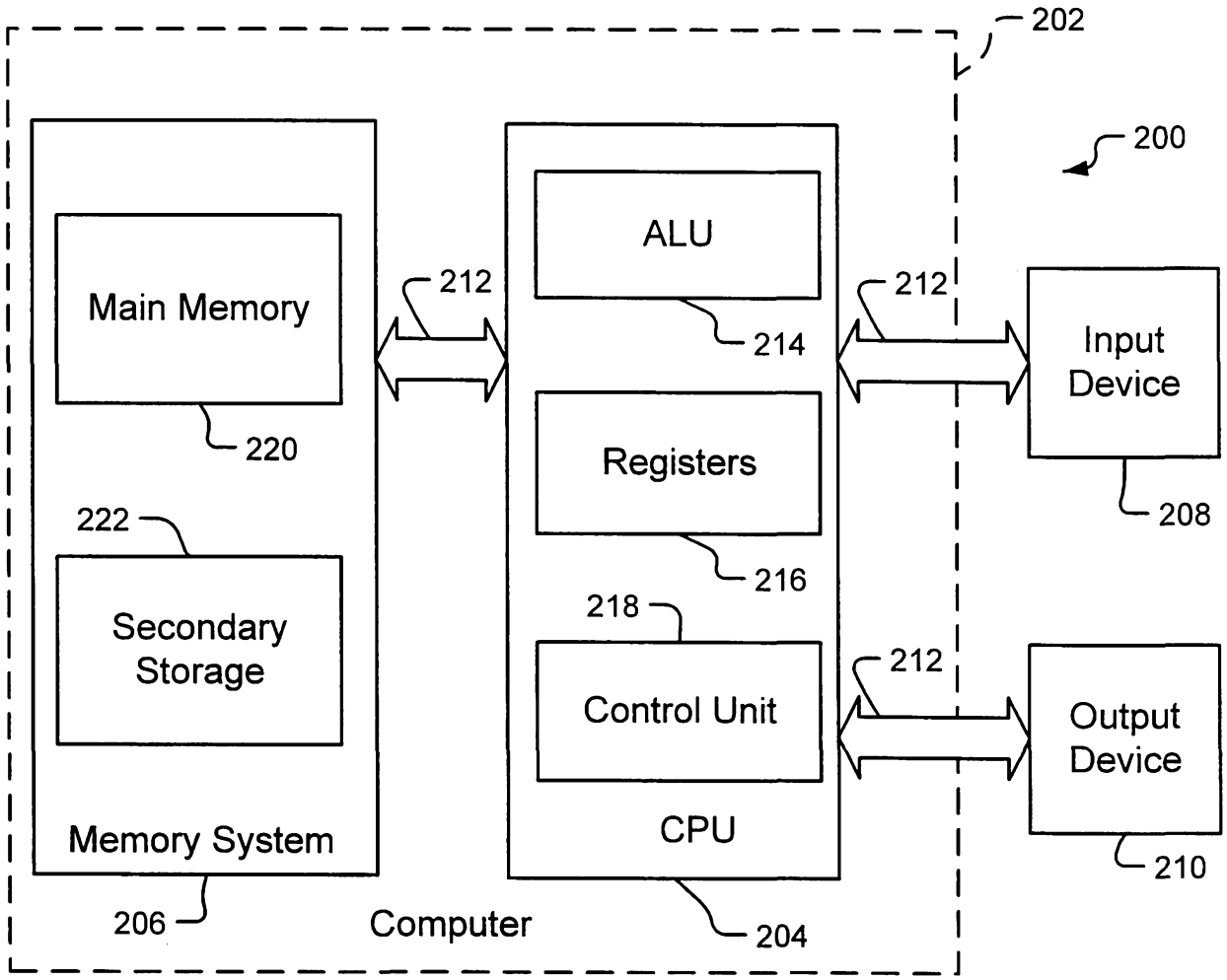

202
200
Main Memory
220
212
ALU
214
212
Input Device
208
Registers
216
222
Secondary Storage
218
Control Unit
212
Output Device
210
Memory System
206
CPU
204
Computer

Fig. 2

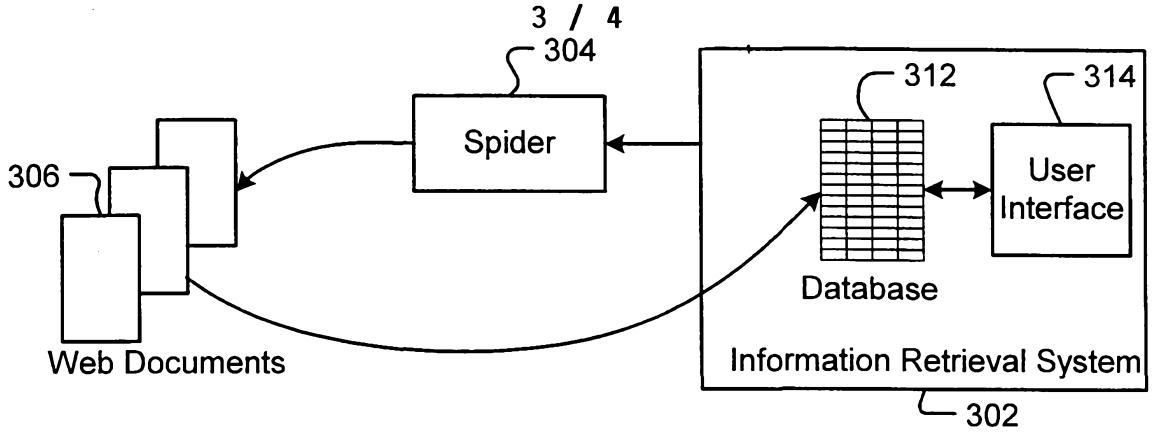


Fig. 3 Prior Art

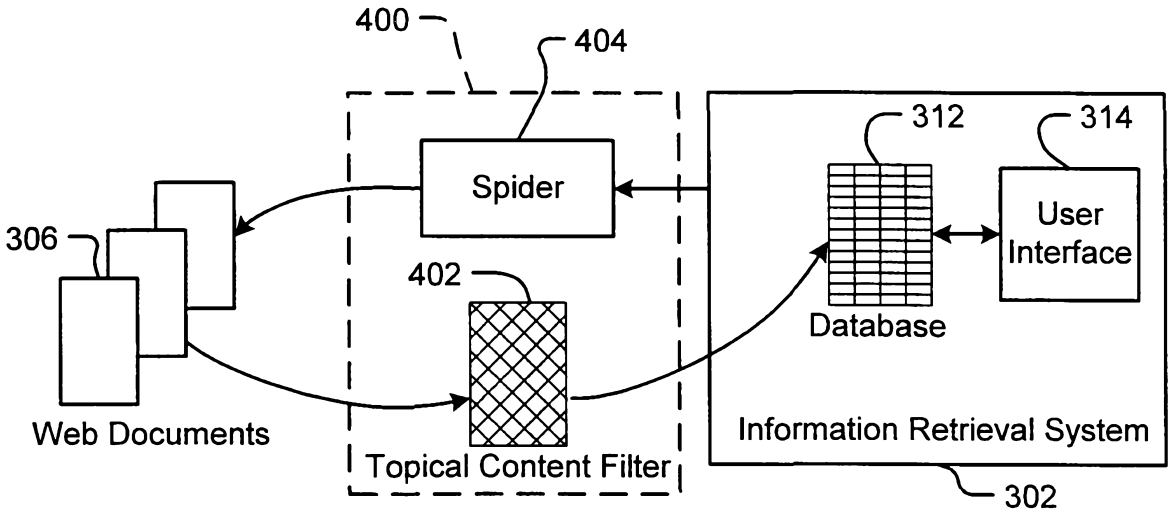


Fig. 4

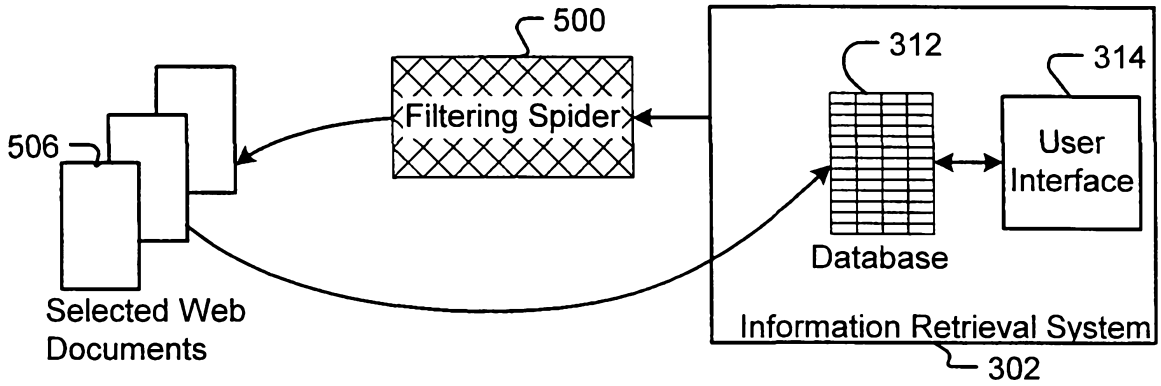


Fig. 5

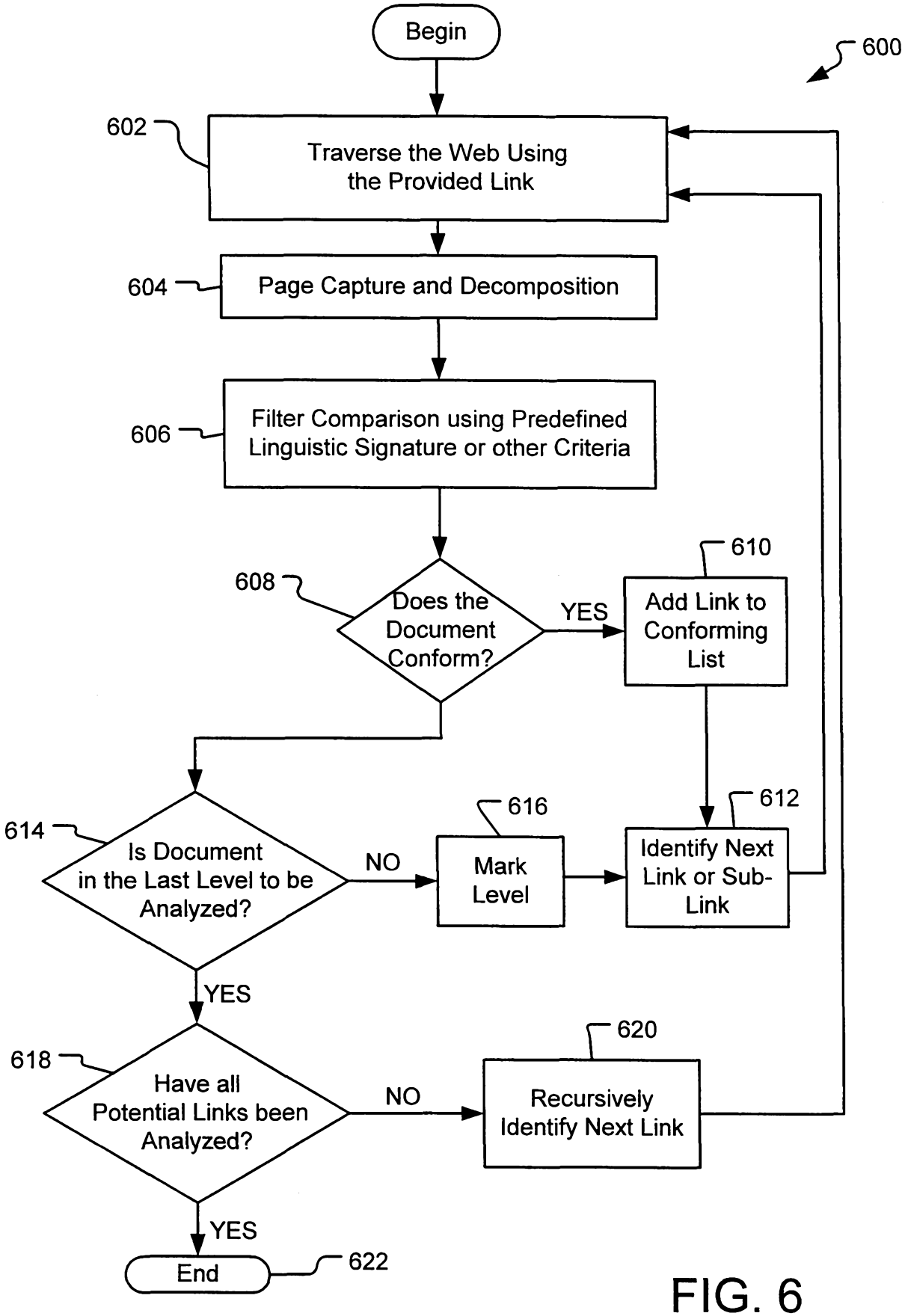

Begin
600
602
Traverse the Web Using the Provided Link
604
Page Capture and Decomposition
606
Filter Comparison using Predefined Linguistic Signature or other Criteria
608
Does the Document Conform?
YES
610
Add Link to Conforming List
612
Identify Next Link or Sub-Link
614
Is Document in the Last Level to be Analyzed?
NO
616
Mark Level
YES
618
Have all Potential Links been Analyzed?
NO
620
Recursively Identify Next Link
YES
End
622

FIG. 6