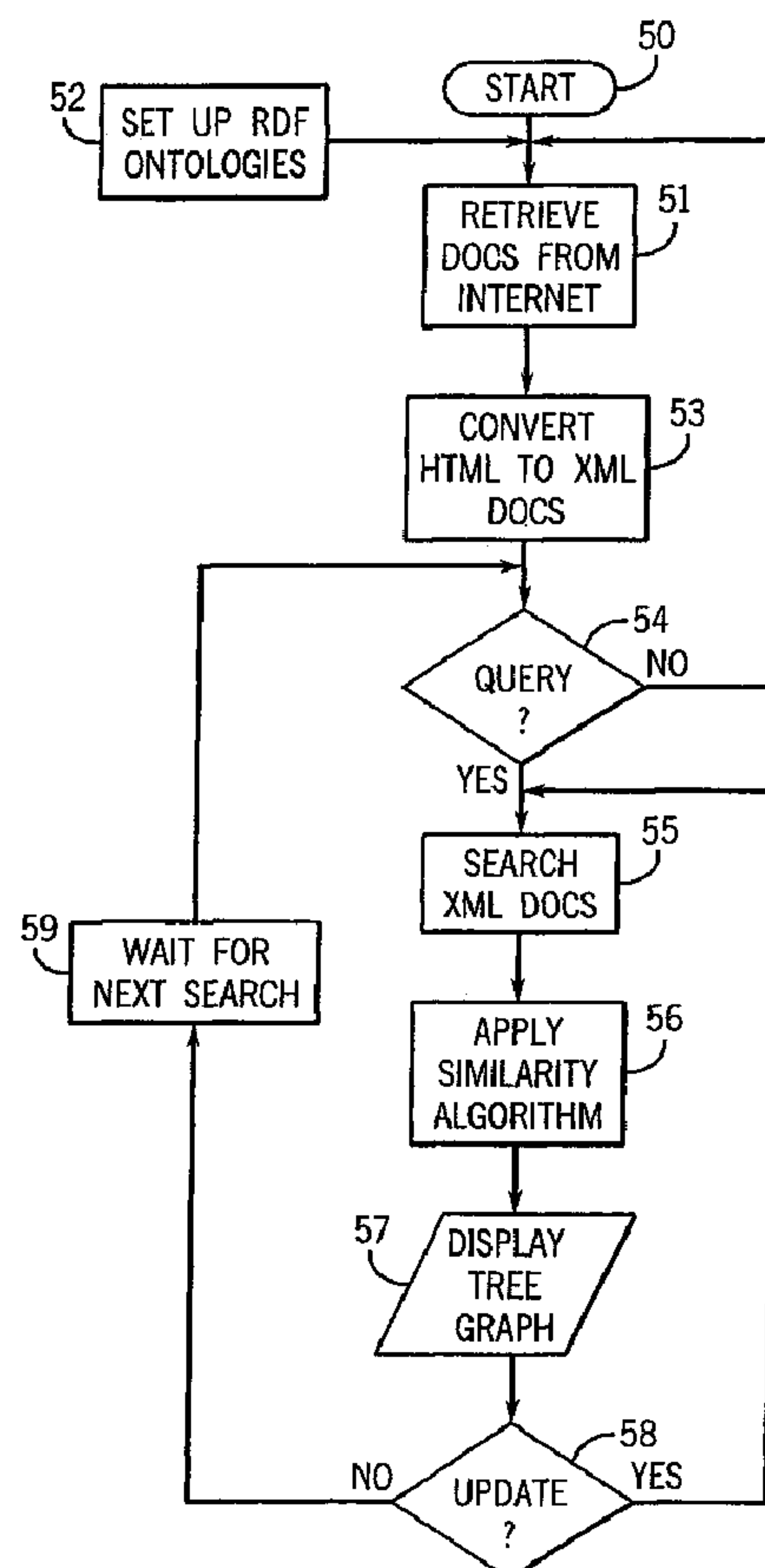




(86) Date de dépôt PCT/PCT Filing Date: 2002/12/12
(87) Date publication PCT/PCT Publication Date: 2003/07/10
(45) Date de délivrance/Issue Date: 2012/10/02
(85) Entrée phase nationale/National Entry: 2004/06/18
(86) N° demande PCT/PCT Application No.: US 2002/040089
(87) N° publication PCT/PCT Publication No.: 2003/056467
(30) Priorités/Priorities: 2001/12/21 (US60/341,755);
2002/05/29 (US10/157,704)

(51) Cl.Int./Int.Cl. *G06F 17/30* (2006.01)
(72) Inventeurs/Inventors:
POTOK, THOMAS E., US;
ELMORE, MARK THOMAS, US;
REED, JOEL WESLEY, US;
TREADWELL, JIM N., US;
SAMATOVA, NAGIZA FARIDOVNA, US
(73) Propriétaires/Owners:
UT-BATTELLE, LLC, US;
BWXT Y-12, L.L.C., US
(74) Agent: GOUDREAU GAGE DUBUC

(54) Titre : PROCEDE DE RASSEMBLEMENT ET DE RESUME D'INFORMATIONS INTERNET
(54) Title: METHOD FOR GATHERING AND SUMMARIZING INTERNET INFORMATION



(57) Abrégé/Abstract:

A computer method of gathering and summarizing information available through the Internet comprises collecting information from a plurality of Internet sites (14, 51) according to respective maps (52) of the Internet sites (14), converting the collected information

(57) **Abrégé(suite)/Abstract(continued):**

from HTML-language web pages to XML-language documents (26, 53) and storing the XML-language documents in a storage medium, searching for documents (55) according to a search query (13) having at least one term and identifying the documents (26) found in the search, and displaying the documents as nodes (33) of a tree structure (32) having links (34) and nodes (33) so as to indicate similarity of the documents to each other.

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

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
10 July 2003 (10.07.2003)

PCT

(10) International Publication Number
WO 2003/056467 A3

(51) International Patent Classification⁷: **G06F 17/30**

(21) International Application Number:
PCT/US2002/040089

(22) International Filing Date:
12 December 2002 (12.12.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/341,755 21 December 2001 (21.12.2001) US
10/157,704 29 May 2002 (29.05.2002) US

(71) Applicants: **UT-BATTELLE, LLC** [US/US]; 111 B Union Valley Road, P.O. Box 2008, MS 6498, Oak Ridge, TN 37831-6498 (US). **BWXT Y-12, L.L.C.** [US/US]; P.O. Box 2009, Oak Ridge, TN 37830 (US).

(72) Inventors: **POTOK, Thomas, E.**; 105 Windham Road, Oak Ridge, TN 37830 (US). **ELMORE, Mark, Thomas**;

102 East Melbourne Road, Oak Ridge, TN 37830 (US). **REED, Joel, Wesley**; 1015 Blinken St., Knoxville, TN 37932 (US). **TREADWELL, Jim, N.**; 104 River Place, Louisville, TN 37777 (US). **SAMATOVA, Nagiza, Fari-dovna**; 103 Harland Court, Oak Ridge, TN 37830 (US).

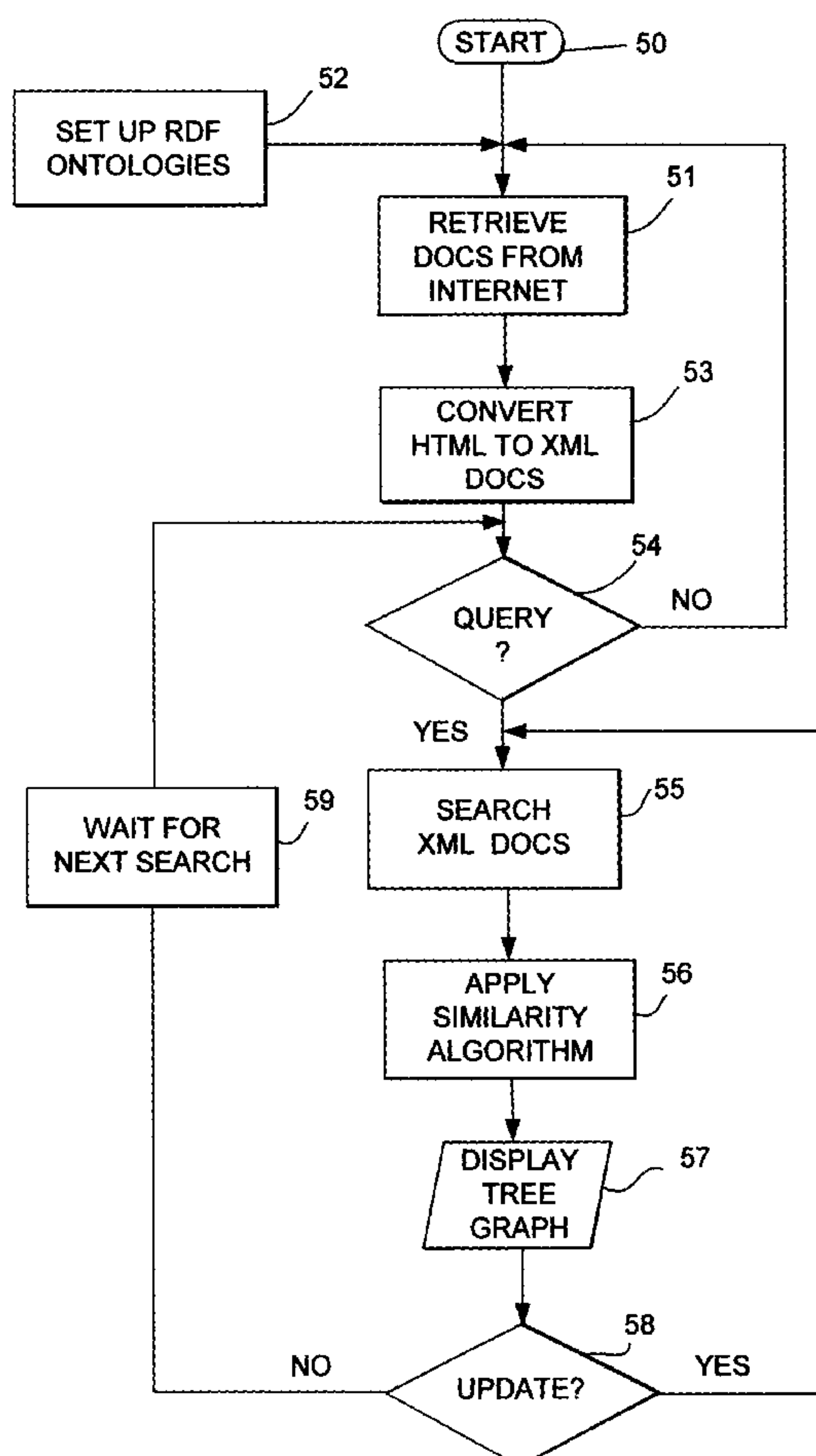
(74) Agent: **MCGOVERN, Michael, J.**; Quarles & Brady, 411 East Wisconsin Avenue, Milwaukee, WI 53202-4497 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, 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, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: METHOD FOR GATHERING AND SUMMARIZING INTERNET INFORMATION



(57) Abstract: A computer method of gathering and summarizing information available through the Internet comprises collecting information from a plurality of Internet sites (14, 51) according to respective maps (52) of the Internet sites (14), converting the collected information from HTML-language web pages to XML-language documents (26, 53) and storing the XML-language documents in a storage medium, searching for documents (55) according to a search query (13) having at least one term and identifying the documents (26) found in the search, and displaying the documents as nodes (33) of a tree structure (32) having links (34) and nodes (33) so as to indicate similarity of the documents to each other.

WO 2003/056467 A3

WO 2003/056467 A3



European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(88) Date of publication of the international search report:
22 April 2004

Published:

— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHOD FOR GATHERING AND SUMMARIZING INTERNET INFORMATION

TECHNICAL FIELD

[001] The field of the invention is software agents for gathering information available through the World Wide Web (WWW) of networks, also known as the Internet.

DESCRIPTION OF THE BACKGROUND ART

[002] There are two very generalized approaches to collecting and organizing information over the Internet. One approach is to use Internet search engines. These search engines typically have spidering programs that recursively traverse Internet links, capturing non-trivial terms on each page. These pages are then organized based on the terms encountered in each document. The strength of this approach is that a very wide number of documents can be spidered and made available for keyword searches. Some of the drawbacks are as follows: 1) Existing pages in the system are infrequently re-spidered, meaning that information can easily be out of date. 2) Internet pages have no consistent format, and therefore, the content of a page cannot be easily discerned. 3) The documents are organized based solely on the presence of a keyword in a document.

[003] The other broad approach is to gather and process Internet information using information agents to retrieve information. These agents provide a number of ways to retrieve and organize information. Information agents are capable of accessing information from multiple sources, and then filtering information by relevance to a user. The most basic systems use non-cooperating agents to perform an information retrieval task. Enhanced systems use cooperating agents, and finally, adaptive information agents that can deal with uncertain, incomplete, or vague information. Information agents can efficiently gather heterogeneous and frequently changing information from the

Internet. While the information agent concept is appealing, much of the literature in the area describes characteristics and attributes of agents, with little detail on specific advantages of the technology. Another technical problem is the lack of enough inherent structure in newspaper articles that would allow the information agents to transform the inherent structure to a common schema.

[004] Once the information has been retrieved, the next challenge is how to organize it. There are a number of methods available for doing this. The most basic approach is keyword searching within a document as a way of classifying the document. This simple approach yields mixed results because documents that contain the same words may have no semantic relationship to each other.

[005] A more sophisticated approach to organizing information uses a vector space model (VSM), where each unique word within a collection of documents represents a dimension in space, while each document represents a vector within that multidimensional space. Vectors that are close together in this multidimensional space form clusters, or groups of documents that are similar.

[006] Clustering techniques can be used for organizing documents into similar groups of documents. Through local and global weighing schemes this approach can be adapted to compare the similarity of one document to another. One of the limitations of clustering is that the entire document set must be available at the time of the analysis, and clustering algorithms require extensive computations, typically n^3 in complexity based on "n" documents.

[007] Another approach to organizing information is to use neural networks to determine patterns within documents. It is assumed that documents with similar word patterns are similar in content. These models are built on the premise that historic patterns will hold in the future. This is clearly not the case with newspaper articles where topics, people, and events change at frequent intervals.

[0008] There remains a need for more effective software agents for collecting and summarizing information available at Web sites on the Internet.

SUMMARY OF THE INVENTION

[0009] The invention is incorporated in a computer method for gathering and summarizing information available through a network, such as the Internet, an Intranet or other network, the method comprising: collecting information from a plurality of network sites according to respective maps of the Internet sites; converting the collected information from HTML-language web pages to XML-language documents and storing the XML-language documents in a storage medium; searching for documents according to a search query having at least one term; identifying the documents found in the search; and displaying the documents as nodes of a tree structure having links and nodes so as to indicate similarity of the documents to each other.

[0010] In a further aspect of the invention, information is collected from a plurality of Internet sites at a predefined time interval.

[0011] The method is carried out by at least one software agent computer program authored in the JAVA computer language.

[0012] The software agent computer program resides in a computer with an agent hosting program. The software agent computer program is a client program in relation to the agent hosting program. The hosting program allows a plurality of software agents to communicate, whether the software agents are all operating on one computer or are distributed over several computers. The communication can be peer-to-peer as well as host-client. It also possible to run the client as a JAVA applet running with a browser accessing the host through the Internet.

[0013] In another aspect of the invention the similarity of new documents is tested through dynamic clustering in which an additional document added to the plurality of documents is calculated by comparing the additional document

to a similarity matrix for the plurality of documents and without recalculating the matrix for all of the documents.

[0014] Other objects and advantages of the invention, besides those discussed above, will be apparent to those of ordinary skill in the art from the description of the preferred embodiments which follows. In the description reference is made to the accompanying drawings, which form a part hereof, and which illustrate examples of the invention. Such examples, however are not exhaustive of the various embodiments of the invention, and therefore reference is made to the claims which follow the description for determining the scope of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] Figs. 1-3 are screen displays in a user interface generated by the present invention;

[0016] Fig. 4 is block diagram of the present invention installed on a plurality of computers; and

[0017] Fig. 5 is a block diagram of a host computer for a plurality of agent programs of the present invention; and

[0018] Fig. 6 is a flow chart illustrating the method of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0019] Fig. 1 shows a window screen display 10 for summarizing information previously collected on the Internet. The window screen display 10 is subdivided into two frames 11 and 12. The collected information will search using a search query somewhat like an Internet search with a search engine such as Yahoo! In the left frame 11, a list of collections 14, in this case, newspapers, is displayed, each preceded by a check box 15 to select a collection for searching.

[0020] A text entry and display box 16 is provided for typing in search terms. A set of radio buttons 17 is provided for providing search logic, such as "phrase," "and"

and "or" functions for the search query. In this case, the term "oil" was searched, by clicking on the button 18 labeled "Search" to produce five articles 26 from the collections 14 shown. The user may check the articles for relevance in check boxes 18, "H" = high, "M" = medium, "L" = low and "I" = ignore. The button 20 labeled "All Articles" is provided for displaying all articles from selected collections 14.

[0021] Below the two frames 11, 12, are command buttons 21-25 labeled "Refresh" (the collections), "Remove Article," "Cluster," "Knowledge Discovery," and "Clear".

[0022] When the button 23 labeled "Cluster," is operated by selecting and executing it with the mouse or keyboard, the screen display 30 seen in Fig. 2 is shown on the screen of the computer. In the upper frame 31 is a tree graph 32 with nodes 33 and link 34. The five nodes 33 each represent one of the articles 26 seen in Fig. 1. Their proximity to one another and the length of the links from a common vertex 36, indicate their similarity to each other. In a lower frame 35 below the tree graph is a hierarchical folder organization containing the articles returned in the search.

[0023] Returning to Fig. 1, if the button 24 labeled "Knowledge Discovery," is operated by selecting and executing it with the mouse or keyboard, the screen display 40 seen in Fig. 3 is shown on the screen of the computer. Here the documents found in the search are organized in folders 41 with similar articles, which were not found in the search, but which are determined to be similar to the articles found in the search.

[0024] Fig. 6 shows the method used in producing the search results seen in Figs. 1-3. After the start of program operations represented by the start block 50 in Fig. 6, the articles or documents in the collections must be retrieved through the Internet and stored in a computer memory, as represented by process block 51. It should be noted that various storage media such as RAM memories, hard disks or CD-RW's may be used to store the documents. The

maps for searching web sites and a predefined time interval are set up in a preliminary action represented by process block 52. Next, the articles are converted from HTML web pages to XML files that can be better searched for key words, as represented by process block 53. The computer system then waits for a search query, as represented by decision block 54. If no search query is received, as represented by the "NO" branch from decision block 54, the program in the server will repeat the execution of blocks 51 and 53 at the predefined time interval. If a search query is entered, as represented by the "YES" branch from decision block 54, the documents or articles are retrieved in response to a search query, as represented by process block 55. Then, as represented by process block 56, the articles are analyzed for similarity by applying a similarity algorithm. Next, as represented by I/O block 57, the results of the search query and the similarity analysis are displayed or "clustered" through display of a tree graph. The results can also be displayed in other ways, such as a hierarchical folder presentation. Then, the search portion of the program will wait for the next search, as represented by process block 59, unless commands are entered to update the previous search as represented by the "YES" result from decision block 58. In that case, the program loops to re-execute process blocks 56 and 57. In executing block 56, a dynamic clustering algorithm is applied which results in only five percent of the matrix being recalculated.

[0025] As mentioned above, in order to search the Internet websites where the collections reside a time interval is defined for each website to be searched. This is included in a Resource Description Framework (RDF) ontology, allowing the computer software retrieval agent program to automatically address a site, retrieve relevant documents, and format the documents using the XML tag description language as described above. In this embodiment, each of the Internet websites is monitored by a respective search agent having a corresponding RDF ontology including the search time interval. When a new document is

found, the retrieval agent uploads the document, formats it in XML tag description language, and then sends the new document on for further processing, as described below.

[0026] The RDF ontological description for each website to be monitored includes the four key elements of information:

[0027] 1) Traversal directives - site-specific actions for traversing an Internet site. This includes the search depth limit from the root URL, and the time interval to wait between rechecking the site for new documents.

[0028] 2) Traversal maps - maps of an Internet newspaper site containing the pages of interest. The map starts with the root URL from which the agent is to begin a traversal of the site, and from which the agent can resolve relative URLs found at the site. A rule-based map of the pages of interest on the site is based on the URL structure of the site and is encoded via regular expressions.

[0029] 3) Document delimiters - markers to delimit the text of a document from other information on a given web page. The map of the Internet site includes information used by the retrieval agent to delimit the text of a document from the myriad of other information on the page (boilerplate, banners, advertisements, etc).

[0030] 4) Document structuring rules - rules for structuring the document text as XML. Again, regular expressions are used to reduce the various structural characteristics of an document, such as the title, author, and paragraphs.

[0031] Based on this RDF ontology, a retrieval agent checks each page link found at an Internet site against the traversal map to determine if the document page is of interest. If the document is of interest, and new to the system, then the agent retrieves the page, discerning the page text from the document delimiters, and cleaning it of extraneous information. The agent then marks up the clean text using XML, tagging the parts of the document (title, author, date, location, paragraphs, etc) depending on the site's document structuring rules. The agent continues to

monitor the site based on the traversal directives, and posting new information of interest as it becomes available.

[0032] As an example, a further description for claims for the layout of a site's ontology as represented in an RDF file. The ontology is defined for the Pacific Islands Report (PIR), a Hawaii-based newspaper focusing on news from the Pacific Islands.

[0033] Table 1 in Appendix A shows the overall layout of the PIR site. This site has two levels of interest, the root URL that forms a "table of contents" for the site, and the individual article pages. There are also a number of links that are not of interest, and are thus excluded from processing. For example, pages that do not conform the URL pattern of "http://pidp.ewc.Hawaii.edu/pireport/..." are excluded from processing, as will be described below.

[0034] The root is at the URL <http://pidp.ewc.hawaii.edu/pireport/>. From this, a number of articles are linked, using the date in the path names of the articles, for example, the URL for the first article is <http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-01.htm>, where the last number, 01, represents the article number for that day. On this day, there were twenty-six articles. On other sites, it is quite likely to have several tables of contents of articles. For example, one may contain local news, while another contains state news, and yet another contains national news.

[0035] Next, Table 2 in Appendix A shows the HTML for a typical news article from this newspaper. The HTML in Table 2 shows HTML tags for formatting the page, then the text of the article itself, followed by more formatting tags. The HTML tags do not provide any structuring of the article text; it merely changes the display of the text. Without understanding the content of the page, there is no way to automatically determine what the title of the article is or who wrote it.

[0036] The converted XML document is shown in Table 3 of Appendix A. The file contains a significant amount of information beyond that merely stored within the article

text, for example, the time stamp of when the article was retrieved, the ontology metadata information, the raw HTML, the clean text, as well as the actual text of the article marked up in XML.

[0037] Software programs and agents can then readily process this information. The XML representation in Table 3 can be used to display the article contents within a Web browser using style sheets. Likewise, the article is structured, so that queries and searches can be performed over the XML tags. The RDF ontology will now be described in more detail.

[0038] As an example, the RDF ontology for Pacific Island Reporter is presented across Tables 4 and 5 in Appendix A. Of the five key elements of this ontological information, 1) article metadata, 2) traversal directives, 3) traversal maps, 4) article delimiters, and 5) article structuring rules, Table 4 captures the first two elements.

[0039] The article metadata includes the `<ORNL:newspaperName>` tag that contains the name of the newspaper. In this example, it is the "Pacific Islands Report." The `<ORNL:rootURLStr>` tag contains the root URL of the newspaper site. This is the page from which the agent will begin its traversal of site's contents and is also the base URL used to resolve relative links found within the site. `<ORNL:collection>` is the tag that describes the collection (based on region of the world) to which the articles will be added.

[0040] The traversal directives are contained within the `<rdf:Description ID="agentDirective">` tag set. These directives include the `<ORNL:searchDepthLimit>` tag that defines how many nesting levels deep the search is to go. Although this can be used in filtering articles, its main function is as a failsafe measure in the event a search goes awry. For example, it prevents the agent from traversing into an archive, where thousands of old articles may be stored. How often an agent will revisit a given site to

check for new articles is controlled by the `<ORNL:minutesWaitBetweenDownloadSessions>` tag.

[0041] The portion of the RDF in Table 5 captures the third and fourth key elements of information, the traversal map and the article delimiters.

[0042] The traversal map represents pages on the site that are of interest. For example, current news articles of interest are represented in the site map, while classified ads are explicitly blocked. The map is represented by a series of regular expressions that are used to classify the links found on the site into one of three categories. In the first category, a link is to a page that contains links of interest. Such a page may be thought of as a table of contents page. In the second category, a link is to an article of interest, while in third category, a link is to a page of no interest. The key aspect here is that only the pages of relevance are considered.

[0043] Continuing in Table 5, the `<rdf:Description="tocMetaData">` tag contains one or more table of contents (toc) regular expressions. These are an unordered list, and thus wrapped in the `<rdf:Bag>` container tags. The `<ORNL:urlRegEx>` tag contains a regular expression to categorize the link. Those links that match the regular expression are considered to be table of contents pages, and are recursively scoured for links to pages of interest. For PIR, there was only one type of table of contents to describe, thus there is only one description within the `<rdf:Bag>` container tags.

[0044] The `<rdf:Description="articleMetaData">` tag contains one or more unordered article descriptions. The `<rdf:Description ID="article">` tag contains information for one type of article of interest found at a site; this tag set contains an association of three sub-tags, `<ORNL:urlRegEx>`, `<ORNL:startOfTextStr>`, and `<ORNL:endOfTextStr>`. The `<ORNL:urlRegEx>` tag contains a regular expression with which the retrieval agent tests links found on the site. Those links that pass this regular

expression test are considered to be article pages. In this example, the regular expression:

[0045] [http://pidp\ .ewc\ .hawaii\ .edu/pireport/\[0-9\]{4}/\(January|February|March|April|May|June|July|August|September|October|November|December\)/\[0-9\]{2}-\[0-9\]{2}-\[0-9\]{2}\.htm](http://pidp\ .ewc\ .hawaii\ .edu/pireport/[0-9]{4}/(January|February|March|April|May|June|July|August|September|October|November|December)/[0-9]{2}-[0-9]{2}-[0-9]{2}\.htm)

is used to test the links for articles of interest.

[0046] The fourth key element of information, article delimiters, is also contained within the <rdf:Description ID="article"> tag. Article delimiters are only needed for pages that contain articles. Note, however, that a page may be both an article and a table of contents, that is, the page contains both article text and links of other pages of interest. In such a case, a regular expression for such a page would appear in both the <rdf:Description ID="article"> tag and in the <rdf:Description="tocMetaData"> tag.

[0047] The <ORNL:startOfTextStr> tag contains a character string that delimits the beginning of the article text, and the <ORNL:endOfTextStr> tag contains a character string that delimits the end of the article text. The goal is to be able to find a consistent combination of characters that delimit the article text for all articles matching the regular expression contained in the associated <ORNL:urlRegEx> tag. Note that these delimiting character strings must match the HTML found at the newspaper's web site, whether or not the HTML is well-formed. So far, we have not found a site where this cannot be done. Note that in this PIR example, these characters are HTML tags, but that is not the case with all sites.

[0048] The fifth key element of information, article structuring rules, have been added to the text processing software stored in the host computer, and works very well for converting the raw article text to XML. The implementation would be very similar to the article delimiters, where the consistent structure of an article would be identified throughout the pages of a site.

[0049] To summarize, the key point is that an XML document has been converted from an unstructured HTML document using an RDF ontology.

[0050] The document information is stored as a vector space model (VSM). Using this method, each unique word in a collection of documents represents a dimension in space and each document in this space is represented by a vector.

[0051] When a document is added to the existing set of documents, the first action is to remove the stop words. These are words that are common in speech, but carry little meaning, such as the words "the," or "and." The remaining words are then counted to determine the frequency of a given word within a given document (its local frequency) and compute the frequency of each term over the entire set of documents (its global frequency). These frequency counts are recorded in the local and global document frequency tables. The local document frequency table contains an entry for each document that records the frequency of each term in that document. The global frequency count table contains frequency counts for how often each unique term appears in the entire document set. From these local and global frequencies a document-term weighting is calculated by the following function:

$$Weight_{dt} = LF_{dt} * \left(1 + \sum \frac{LF_{dt} / GF_t * \log_2(LF_{dt} / GF_t)}{\log_2 n} \right) \quad 1)$$

[0052] Where LF is the local frequency for term t in document d, GF is the global frequency for term t, and n is the total number documents in the set. To avoid recalculating all of the vectors every time a new document is added or deleted, only a portion of them are recalculated.

[0053] The approach is to create a list of the matrix cells which is ordered by when they were last updated. Using this list, each time a new document is added to the document set, the oldest five percent of the matrix is updated. In other words, each time a document is added to the document set, the pairs of document vectors corresponding to least recently updated five percent of the

matrix cells are recalculated, and then those matrix cells are updated using the new vectors. Documents being removed from the system are handled in a very similar manner. This allows documents to be quickly added to the system as they stream in and removed from the system as they are no longer needed.

[0054] Next the information is analyzed and clustered for presentation. From the VSM, a similarity matrix is calculated that provides a pairwise comparison of each document in the system. The dot product (which is the cosine of the angle between the vector pair) as used as the measure of similarity between two document vectors. This generates a global similarity matrix of size "n x n," where "n" is the number of documents contained in the document collection. Only the upper triangular portion of this matrix is needed to be stored since it is a symmetric matrix.

[0055] To further analyze the documents, a clustering algorithm is applied to them. Many approaches are available, such as Ward's Method. This method initially treats each document as a cluster. Among all cluster pairs, the method then locates the most similar pair of clusters using the dissimilarity matrix, and agglomerates this pair of clusters into a single cluster. The dissimilarity matrix is then updated to reflect the merged clusters using the following function:

$$D_{MC} = \left[\frac{((A_n + C_n) * D_{AD} + (B_n + C_n) * D_{BD} - C_n * D_{AB})}{A_n + B_n + C_n} \right] \forall C \quad 2)$$

where D represents the dissimilarity measure between two document, M is the new cluster built when clusters A and B are merged and where C represents the cluster whose dissimilarity is being updated. Also, A_n and B_n are the number of documents that make up the clusters being merged to make cluster M, and C_n is the number of documents that make up the cluster being updated. This merging process is repeated until all of the documents are in a single cluster.

[0056] The information is presented in a cluster diagramming graph called a Phylips Tree (Fig. 2). The nodes

33 of the tree 32 represent each document while the links 34 between the nodes 33 represent relationships. In general, the closer two nodes 33 are, the more similarity there is between two documents. If links from two nodes 33 share a vertex 36, then these documents are the closest in the set of documents. The longer the links 34 are between documents, the greater the dissimilarity is between the documents.

[0057] To organize and classify Internet newspaper information, cooperative and adaptive information agents are used. These agents work together to gather and organize information. A number of different agent types, and implemented a communication protocol enabling them to interact. For example, one type of agents gathers information from individual newspapers, another type of agents analyzes the articles and organizes the information. The agents operate with an agent-community hosting program which provides transparent agent communication and mobility across any Internet connected host.

[0058] Within an agent-community hosting program, an agent community can be quickly created using a set of computers with each machine executing the agent host program. The agent host program allows agents to be transmitted and received among machines. This allows agents to be truly mobile, moving from machine to machine as needed. This capability helps facilitate communication among agents within a community. Agents can also interact with systems and agents that are not part of the community. Agent mobility through the Internet is very limited based on the necessary security limitations enforced over the Internet. The agent hosting program uses the Foundation for Intelligent Physical Agent (FIPA) compliant agent communication language (ACL) messages. This allows any FIPA compliant agent to be able to interact with the agent host program.

[0059] Within the agent host community (see Fig. 4), each agent host 73, 74 has a name server responsible for knowing what agents are currently being hosted. In addition, the name server is responsible for answering queries from agents trying to locate other agents in the community. For

example, an agent may want to broadcast information to all of the agents within the community. The name server in each agent host 73, 74 is used to locate all of the agents so that the message can be delivered.

[0060] Figure 4 illustrates a system with multiple agent hosting computers 71, 72, identified as "Machine 1" and "Machine 2.". Agents A-F and G-Z can move from one machine to another by changing agent hosts 73, 74. The RDF ontologies 75 move with the agent A-F and G-Z. The agent contexts provide machine specific environments for the agent to work.

[0061] When an agent is received at a machine 71, 72, the agent host 73, 74 provides it with an agent context. This agent context is the agent's only point of contact with the machine it is running on. The agent is not allowed to directly communicate with the agent host or other agents. This provides an architectural layer for security in the agent host system. The agent host program is written in JAVA and uses JAVA Remote Method Invocation (RMI) for communicating information from one agent to another. The agent host program as well as the agent programs run as applications under an operating system such as Windows, Unix, Linux, or other known operating systems.

[0062] In a further specific embodiment of the invention seen in Fig. 5, a first group of agents 80-82 performs most of the information retrieval and processing, and a second group of agents 83-84 performs most of the user interface functions. Although these have certain conceptual parallels to a typical client-server system, in this system, there are peer processes where any peer may initiate communication. The host 85 is implemented using a set of information retrieval agents 80-82, whose task is to gather news related, non-redundant information from Internet newspapers, and to format the information using XML. A whiteboard agent 86 acts as an information clearinghouse. The information agents 80-82 submit their articles to the whiteboard agent 86, who then manages the information by ensuring that there are no duplicate articles, archiving stale articles that

beyond a given number of days old, and providing articles to agents that have "subscribed" to the whiteboard 86. There is a group of cluster agents 84 that organizes articles into a vector space model (VSM), then into a cluster of articles.

[0063] The initial challenge of the information agents 80-82 is to gather and organize heterogeneous Internet information. This is accomplished through the transformation of HTML-formatted information into XML-formatted information. The conversion of HTML information to XML is a two-step process:

[0064] An RDF ontology is defined to enable a common semantic representation and structuring of heterogeneous information. A site can be viewed as a directed graph, from which, RDF provides a solid way of modeling the linked pages. Furthermore, these RDF instructions can be understood and followed by a software agent.

[0065] Once an agent can understand an RDF file that describes the layout of an Internet newspaper site and its semantics, then this agent can periodically access the site, retrieve articles of interest, and convert the unstructured heterogeneous information into an XML-formatted document. Each converted article will then contain a rich set of XML tags ranging from the time and date the article was downloaded, to the URL location of the information, to XML tags that format the article itself.

[0066] Each of the information agents 80 monitors the Internet newspapers site looking for new articles. Any time a new article is found, the information agent retrieves the article, formats it, and then posts it to the whiteboard agent 86.

[0067] The ontological description of the site includes the root URL from which the agent is to begin a traversal of the site and from which the agent is to resolve relative URLs found at the site. It also includes a series of one or more regular expressions that will describe table-of-contents pages on the newspaper site. Finally, the site description includes a series of one or more regular expressions that describe article pages of interest on the

site along with information used by the agent to discern the text of an article from the myriad of other information on the page (boilerplate, banners, advertisements, etc). The meta-information includes the newspaper's name and the name of the collection under which the newspaper is classified, as well as site-specific actions taken by the agents and includes the search depth limit (how many hops) from the root URL and the number of minutes to wait between rechecking the site for new articles.

[0068] Based on the RDF ontology, the information agents 80 monitor and manages information at an Internet newspaper site. The agents 80 check each link found at a site against the ontological criteria to determine table-of-contents pages and article pages. If an article page of interest is found, the agent checks with the whiteboard agent 86 to verify that the article was not already incorporated into the system. If the article is indeed new, the agent 80 reads the page, discerns clean article text, i.e., just the raw text from the news article from the other information on the page. The agent 80 then marks up the clean text using XML, tagging the parts of the article (title, author, date, location, paragraphs, etc) depending on the site, and then posts the information to the whiteboard agent 86. The agent 80 continues to monitor the site, posting new information of interest as it becomes available.

[0069] A client agent 87 that contains a graphical user interface is also used. The client agent 87 communicates with both the whiteboard agent 86 and cluster agent 84 to perform searches and clustering.

[0070] The whiteboard agent 86 maintains all of the current articles, making sure there are no duplicates, and removes any articles that are beyond a given time period. The cluster agent 84 subscribes to the whiteboard agent 86 and thus is notified any time an article is added or removed from the whiteboard. When the cluster agent 84 is notified of a new article (as discussed below), it examines the contents of the article and adjusts its search and clustering tables appropriately. Likewise, the tables are adjusted when the whiteboard agent 86 removes an article.

[0071] This has been a description of the preferred embodiments of the invention. The present invention is intended to encompass additional embodiments including modifications to the details described above which would nevertheless come within the scope of the following claims.

Appendix A

Table 1. Site Map of the Pacific Islands reporter.

Root	http://pidp.ewc.hawaii.edu/pireport/
Article	http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-01.htm
Article	http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-02.htm
...	
Article	http://pidp.ewc.hawaii.edu/pireport/2001/June/06-05-26.htm
Link	http://www.eneushawaii.com/
Link	http://www.eastwestcenter.org/events-en.asp
Feature	http://pidp.ewc.hawaii.edu/pireport/2001/May/eww%2005-17.htm
Archive	http://166.122.164.43/archive/
Link	http://pidp.ewc.hawaii.edu/pireport/2001/previous.htm
...	

Table 2. HTML Code for an Article from
the Pacific Islands Reporter

```

<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<title>CORAL REEF EXCAVATION WORRIES FIJI TOURISM INDUSTRY - June 4, 2001</title>
</head>
<body
topmargin="10" leftmargin="10" style="src='../template for stories.htm'
background='../images/backgrnd.gif' bgcolor="FFFFFF" text="000000" link="0000FF"
vlink="000080" alink="FF0000">
<p><strong><font face="Times New Roman" size="5">P</font><font face="Times New Roman"
size="4">ACIFIC</font><big><font face="Times New Roman"> </font></big><font
face="Times New Roman" size="5">I</font><font face="Times New Roman"
size="4">SLANDS</font><big><font
face="Times New Roman"> </font></big><font face="Times New Roman" size="5">R</font><font
face="Times New Roman" size="4">EPORT</font></strong></p>
<p><strong><em><i><font face="Times New Roman" size="4" color="FF0000">Pacific Islands
Development Program/East-West Center<br>
</font><font face="Times New Roman" color="FF0000" size="2">With Support From Center for Pacific
Islands Studies/University of Hawai'i</font></i></em></strong></p>

<hr>
<b><font SIZE="4">
<p>CORAL REEF EXCAVATION WORRIES FIJI TOURISM INDUSTRY</p>
</font></b><font SIZE="4">
<p>SUVA, Fiji Islands -June 3, 2001 - PINA Nius Online----Fiji hotel owners have expressed concern over the
large amount of live coral being excavated and exported to the United States, Ministry of Tourism Director Eroni
Luveniyali
said.</p>
<p>The concern was among issues raised at last week's Fiji National Tourism Council annual meeting, a Ministry
of Information news release said.</p>
<p>Thirty representatives -- both from government and the tourism industry -- attended the meeting in Nadi.</p>
<p>Mr. Luveniyali said many hotel and resort owners have requested that live corals must not be touched or
removed illegally as it endangers the lives of other marine resources.</p>
<p>Tourists who mostly go diving for recreational purposes will be severely affected if the practice continues, he
said.</p>
<p>Mr. Luveniyali said the problem is Fiji's alone, but also one prevalent in other Pacific Island countries.</p>
<p>A recommendation was made at the meeting for a subcommittee to be formed -- comprised of Ministry of
Tourism, Agriculture and Fisheries and Immigration Department officials -- to find ways and means of addressing
the issue.</p>
</font><i><font SIZE="2">
<p>Pacific Islands News Association -PINA-<br>
Website: <font><a href="http://www.pinanius.org">http://www.pinanius.org</a>&nbsp;</font></p>
</i>

<hr>
<table border="0" cellpadding="2" width="100%">
<tr>
<td valign="bottom" align="left"><font face="Times New Roman" size="3">Go back to</font><font
size="3"> </font><font

```

Table 3. XML Code for the Article from
the Pacific Islands Reporter

```

<article>
  <fileBuildTimeMilliSec>
    991680761171
  </fileBuildTimeMilliSec>
  <downloadDate>
    <year> 2001 </year>
    <month> Jun </month>
    <day> 4 </day>
  </downloadDate>
  <articleURL> http://pidp.ewc.hawaii.edu/pireport/2001/June/06-04-05.htm </articleURL>
  <collection> Pacific </collection>
  <newspaperName> Pacific Islands Report </newspaperName>
  <articleParentURL> http://pidp.ewc.hawaii.edu/pireport/graphics.htm </articleParentURL>
  <articleRootURL> http://pidp.ewc.hawaii.edu/pireport/ </articleRootURL>
  <articleDepthFromRoot> 2 </articleDepthFromRoot>
  <articleContentEncoding> null </articleContentEncoding>
  <articleContentType> text/html </articleContentType>
  <articleDate> 991680957000 </articleDate>
  <articleExpiration> 0 </articleExpiration>
  <articleLastMod> 991628284000 </articleLastMod>
  <articleRawHTML>

    ... (omitted for the table)
  </articleRawHTML>
</rdfFileName>
  C:\Program Files\Server V3.0\Server\DownloadAgent\Rdf\pireport.rdf
</rdfFileName>
<articleCleanText>

    ... (omitted for the table)

</articleCleanText>
<xmlMarkedUpText>
  <newspaperName> Pacific Islands Report </newspaperName>
  <url> http://pidp.ewc.hawaii.edu/pireport/2001/June/06-04-05.htm </url>
  <title> CORAL REEF EXCAVATION WORRIES FIJI TOURISM INDUSTRY </title>
  <city> SUVA, Fiji Islands </city>
  <date> June 3, 2001 </date>
  <newsService> - PINA Nius Online </newsService>
  <paragraph number="1">
    Fiji hotel owners have expressed concern over the large amount of live coral being excavated and
    exported to the United States, Ministry of Tourism Director Eroni Luveniyali said.
  </paragraph>
  <paragraph number="2">
    The concern was among issues raised at last week s Fiji National Tourism Council annual meeting, a
    Ministry of Information news release said.
  </paragraph>
  ...
  <paragraph number="7">
    A recommendation was made at the meeting for a subcommittee to be formed -- comprised of
    Ministry of Tourism, Agriculture and Fisheries and Immigration Department officials -- to find ways and
    means of addressing the issue.
  </paragraph>
  <paragraph number="8">
    Pacific Islands News Association -PINA-
    Website: http://www.pinanius.org
  </paragraph>
</xmlMarkedUpText>
</article>

```

Table 4. RDF for the Pacific Islands Report (Part A)

```

<? xml version="1.0" ?>
<rdf:RDF xmlns:ORNL = "http://csm.ornl.gov/VIPAR">

  <rdf:Description about = "http://pidp.ewc.hawaii.edu/pireport/">

    <ORNL:newspaperName>
      Pacific Islands Report
    </ORNL:newspaperName>

    <ORNL:rootURLStr>
      http://pidp.ewc.hawaii.edu/pireport/
    </ORNL:rootURLStr>

    <ORNL:collection>
      Pacific
    </ORNL:collection>

    <rdf:Description ID="agentDirective">
      <ORNL:searchDepthLimit>
        2
      </ORNL:searchDepthLimit>

      <ORNL:minutesWaitBetweenDownloadSessions>
        60
      </ORNL:minutesWaitBetweenDownloadSessions>
    </rdf:Description>
  </rdf:Description>

```

Table 5. RDF for the Pacific Islands Report (Part B)

Continued from Table 4

```

<rdf:Description ID = "tocMetaData">
  <rdf:Bag>
    <ORNL:urlRegEx>
      http://pidp.ewc.hawaii.edu/pireport/graphics.h
      tm
    </ORNL:urlRegEx>
  </rdf:Bag>
</rdf:Description>

<rdf:Description ID="articleMetaData">
  <rdf:Bag>
    <rdf:Description ID="article">
      <ORNL:urlRegEx>
        http://pidp\.ewc\.hawaii\.edu/pireport/[0
        -9]{4}/
        (January|February|March|April|May|June|July|Au
        gust|September|October|November|December)/[0-
        9]{2}-[0-9]{2}-[0-9]{2}\.htm
      </ORNL:urlRegEx>
      <ORNL:startOfTextStr>
        <b><font SIZE="4">
      </ORNL:startOfTextStr>
      <ORNL:endOfTextStr>

```

CLAIMS

1. A computer system for gathering and summarizing information, the computer system being operable on at least one computer having a software operating system, the computer system comprising:

an agent hosting program for running under said software operating system;

a plurality of agent programs operating with said agent hosting program, said plurality of agent programs including programs for collecting documents from respective information sources;

wherein the documents comprise metadata and text that are retrieved; and

wherein said agent program operates according to a stored search ontology providing a map of each respective information source, including documents having both metadata and text to be retrieved, and said search ontology also providing a time interval between search updates for the information source.

2. The computer system of claim 1, further comprising a second host computer having a software operating system and further comprising:

an agent hosting program for running under said software operating system;

a plurality of agent programs operating with said agent hosting program, said plurality of agent programs including programs for collecting documents from information sources;

wherein said plurality of agent programs operate according to a stored search ontology providing a map of each respective information source and a time interval between search updates for the information source.

3. The computer system of claim 1, wherein said at least one of said agent programs is relocatable from one of said host computers to the other of said host computers and is operable on said other one of said host computers.

4. The computer system of claim 1, wherein information sources are Internet web sites, and wherein the documents are web pages including both metadata and text to be retrieved.

5. A computer system for gathering and summarizing information, the computer system being operable on at least one computer having a software operating system, the computer system comprising:

- an agent hosting program for running under said software operating system;

- a plurality of agent programs operating with said agent hosting program, said plurality of agent programs including programs for collecting documents from respective information sources;

- wherein said plurality of agent programs operate according to a stored search ontology providing a map of each respective information source and a time interval between search updates for the information source; and

- an agent for applying a clustering algorithm to documents found in the search of the information sources; and

- a user interface agent for providing a display of results of a search of an information source and results of applying the clustering algorithm; and

- an agent program for interfacing said user interface agent, said agent for applying a clustering algorithm and said plurality of agent programs for collecting documents from respective information sources.

6. The computer system of claim 5, wherein the information sources are Internet web sites, and wherein the documents are web pages including both metadata and text to be retrieved.

FIG. 1

READY



40

ARTICLES

+

IRAN RETURNS TO GUNBOAT DIPLOMACY IN DEALING WITH THE CASPIAN SEA MOSCOW-27 F

+

U.S. not to certify N. Koreas nuclear record By Nicholas Kraiev THE WASHINGT

+

N Korea 'may end' nuclear pact North Korea has repeated its threat to ditch a nu

+

Kazakh capital based on Washington By Christopher Pala SPECIAL TO THE WASHIN

+

Chavez to use troops in case of oil strike CARACAS, Venezuela President Hugo C

41

KNOWLEDGE DISCOVERY

SEARCH

READY

FIG. 3

4 / 5

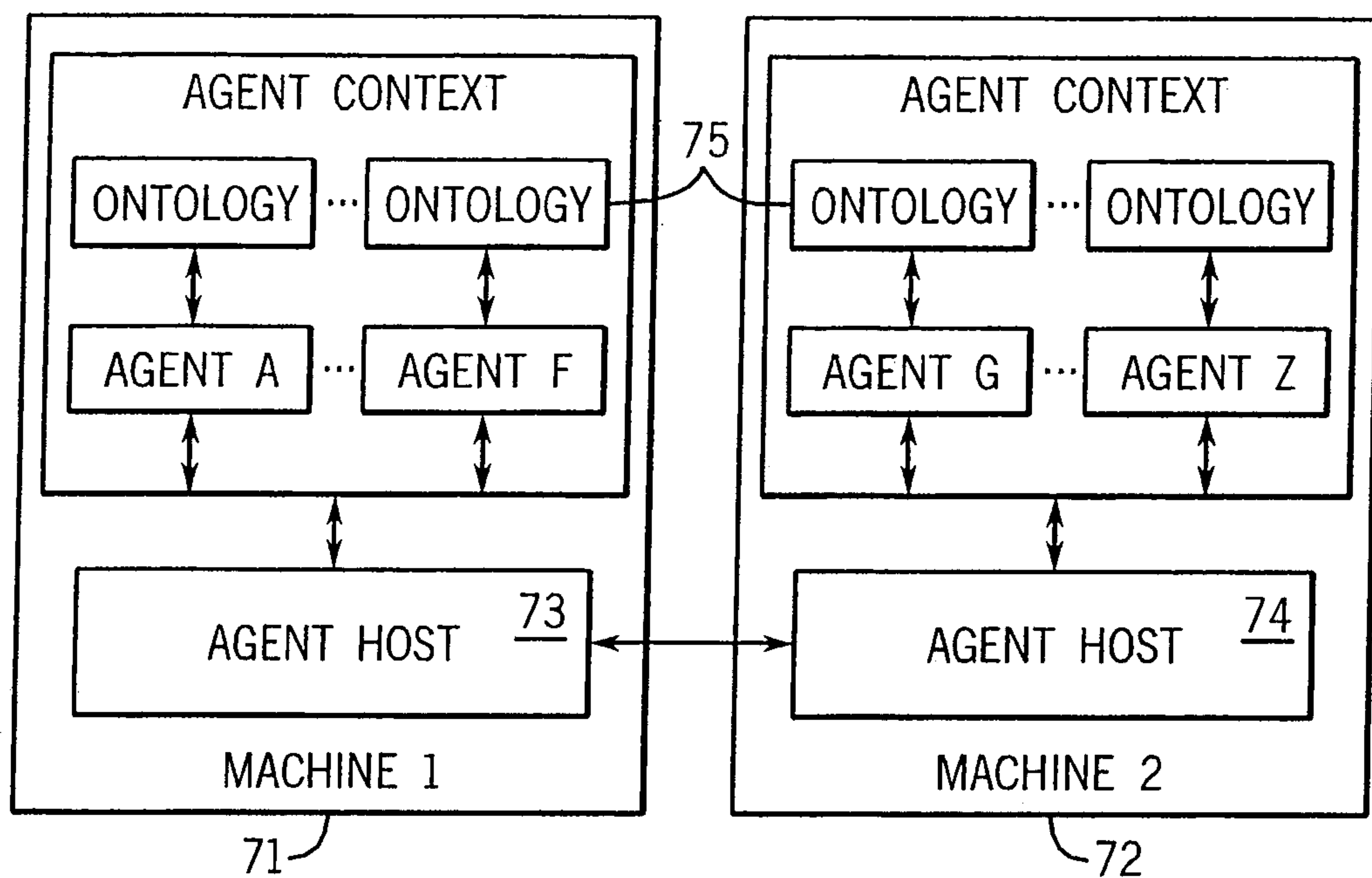


FIG. 4

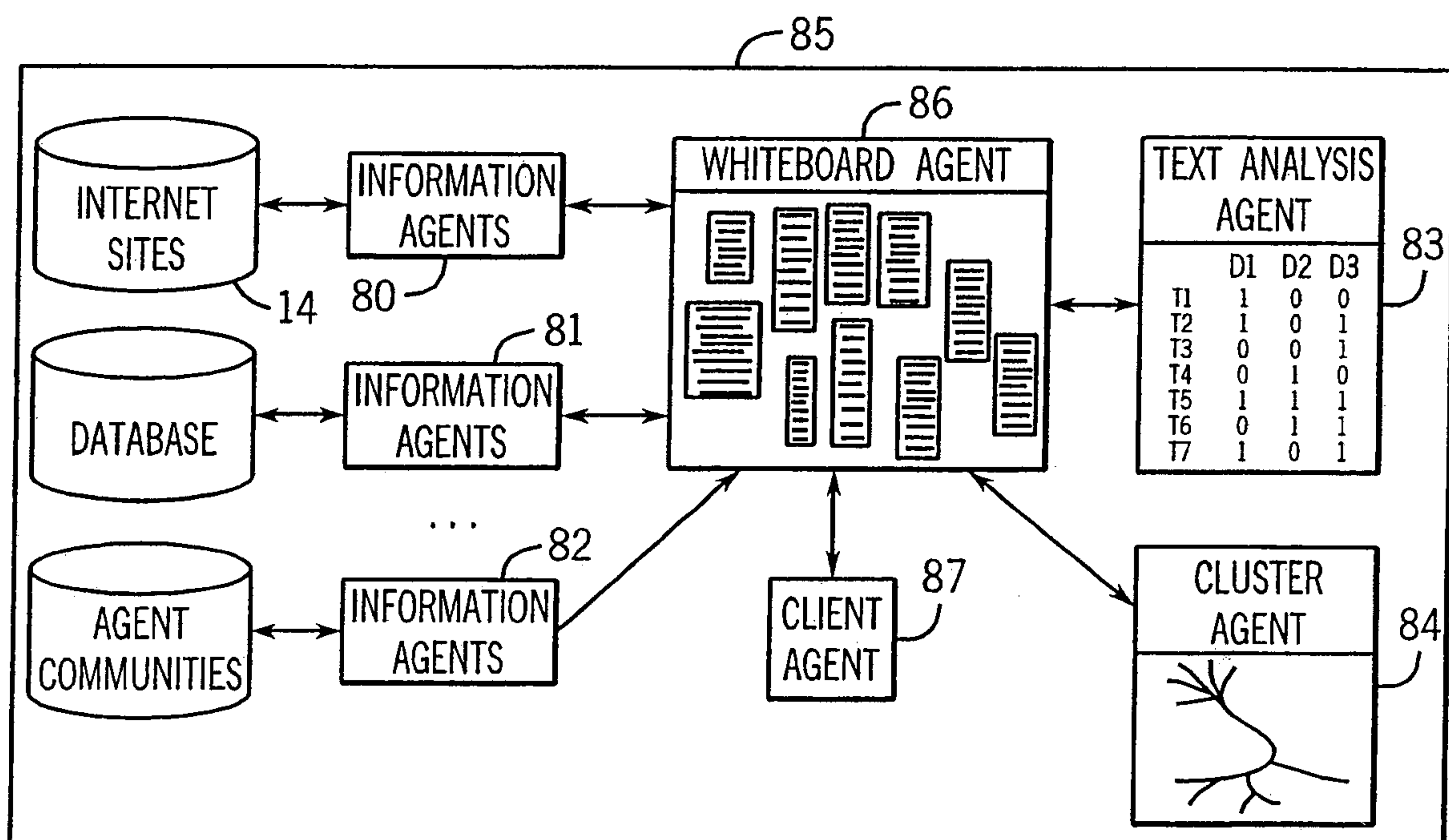


FIG. 5

5 / 5

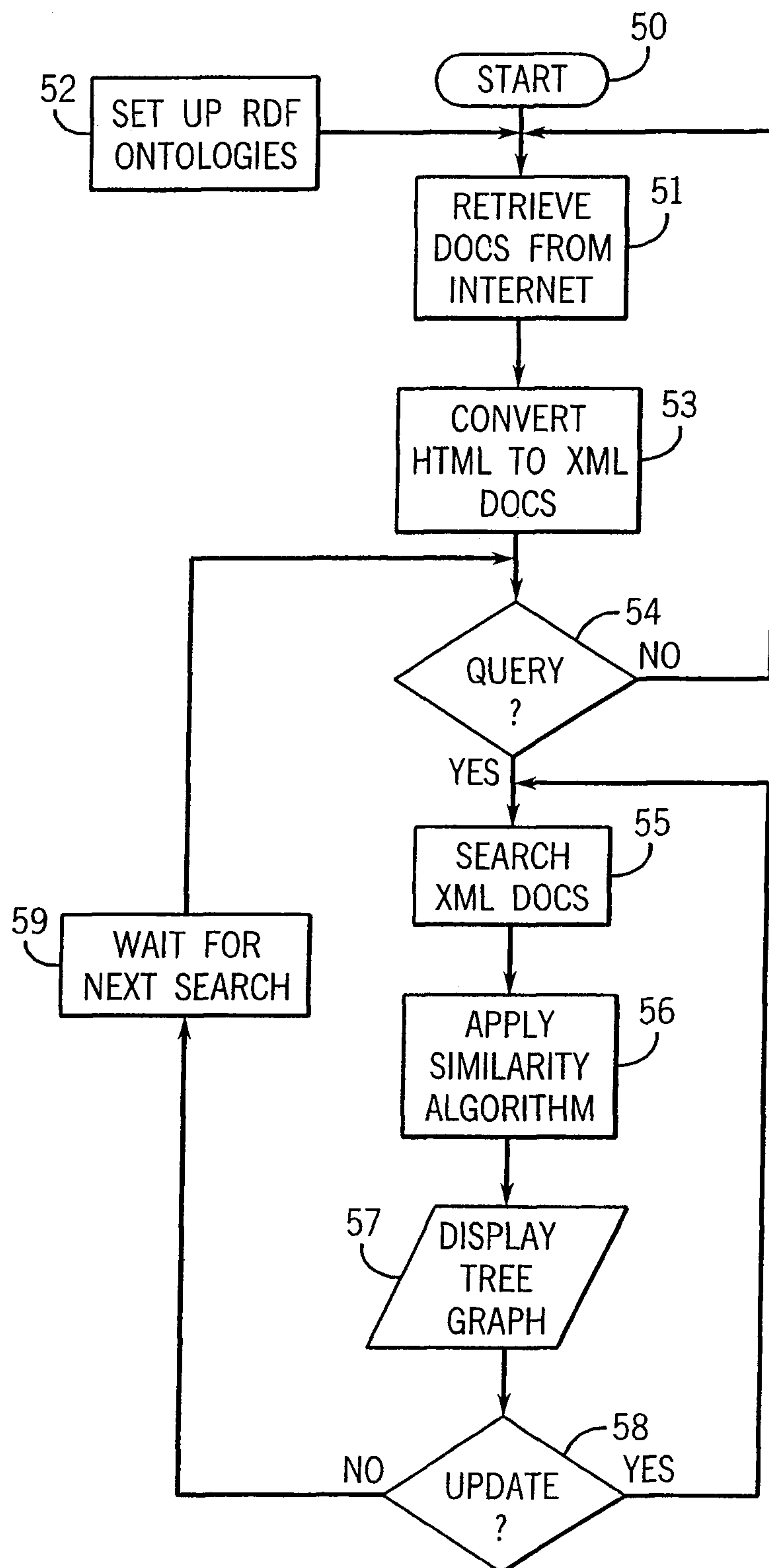


FIG. 6

