



(19) **United States**

(12) **Patent Application Publication**
Virdy

(10) **Pub. No.: US 2013/0110817 A1**

(43) **Pub. Date:** **May 2, 2013**

(54) **SYSTEM AND METHOD FOR
GEOGRAPHICALLY ORGANIZING AND
CLASSIFYING BUSINESSES ON THE
WORLD-WIDE WEB**

(60) Provisional application No. 60/017,548, filed on May 10, 1996.

Publication Classification

(71) Applicant: **FACEBOOK, INC.**, Menlo Park, CA
(US)

(51) **Int. Cl.**
G06F 17/30 (2006.01)

(72) Inventor: **Ajaipal Singh Viridy**, Sterling, VA (US)

(52) **U.S. Cl.**
CPC **G06F 17/30864** (2013.01)
USPC **707/711; 707/740; 707/706**

(73) Assignee: **FACEBOOK, INC.**, Menlo Park, CA
(US)

(21) Appl. No.: 13/725,034

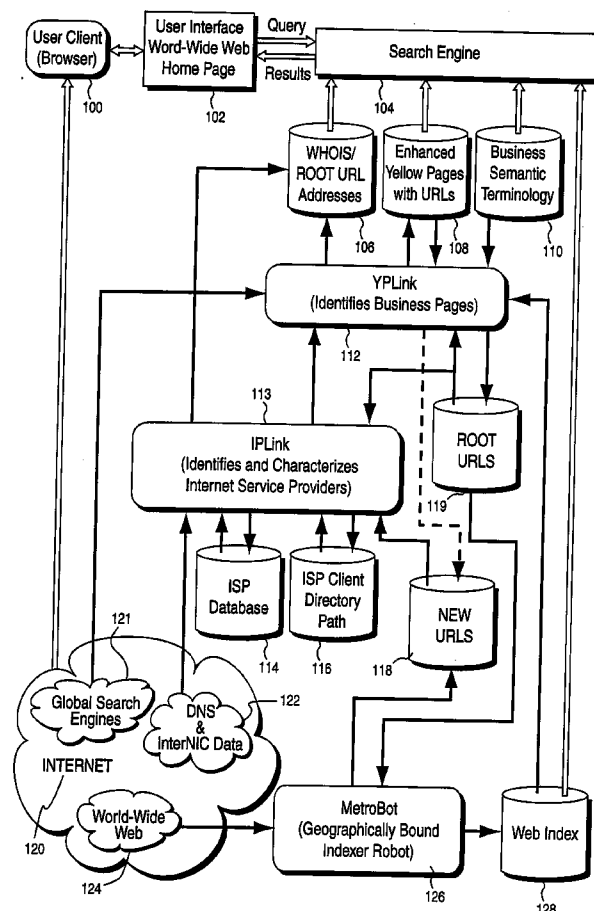
(22) Filed: **Dec. 21, 2012**

Related U.S. Application Data

(63) Continuation of application No. 12/685,799, filed on Jan. 12, 2010, which is a continuation of application No. 12/054,130, filed on Mar. 24, 2008, now Pat. No. 7,664,767, which is a continuation of application No. 10/705,822, filed on Nov. 13, 2003, now Pat. No. 7,349,892, which is a continuation of application No. 09/500,471, filed on Feb. 9, 2000, now Pat. No. 6,691,105, which is a continuation of application No. 08/844,522, filed on Apr. 18, 1997, now Pat. No. 6,148,289.

(57) **ABSTRACT**

A method and search engine for classifying a source publishing a document on a portion of a network, includes steps of electronically receiving a document, based on the document, determining a source which published the document, and assigning a code to the document based on whether data associated with the document published by the source matches with data contained in a database. An intelligent geographic- and business topic-specific resource discovery system facilitates local commerce on the World-Wide Web and also reduces search time by accurately isolating information for end-users. Distinguishing and classifying business pages on the Web by business categories using Standard Industrial Classification (SIC) codes is achieved through an automatic iterative process.



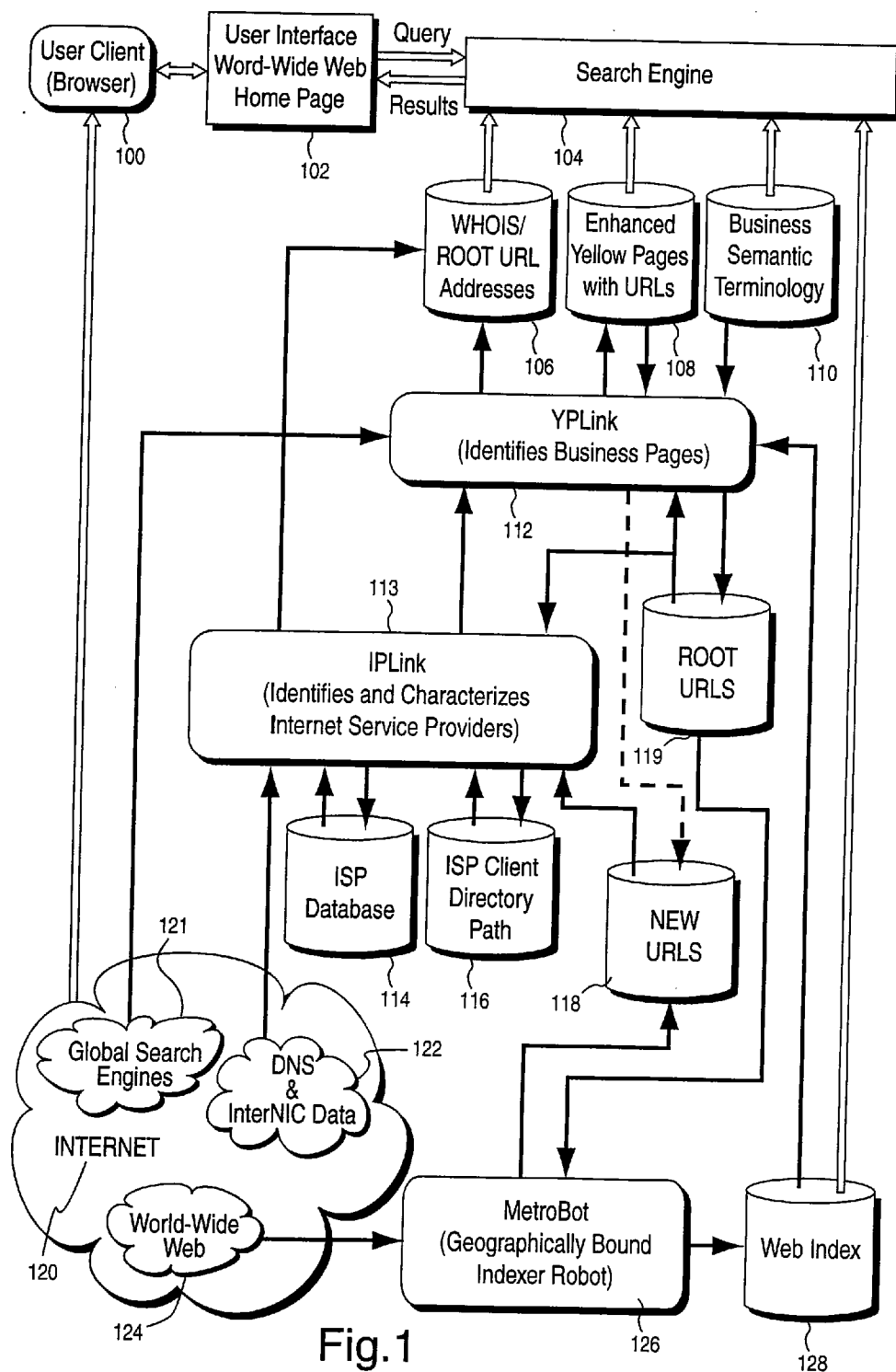
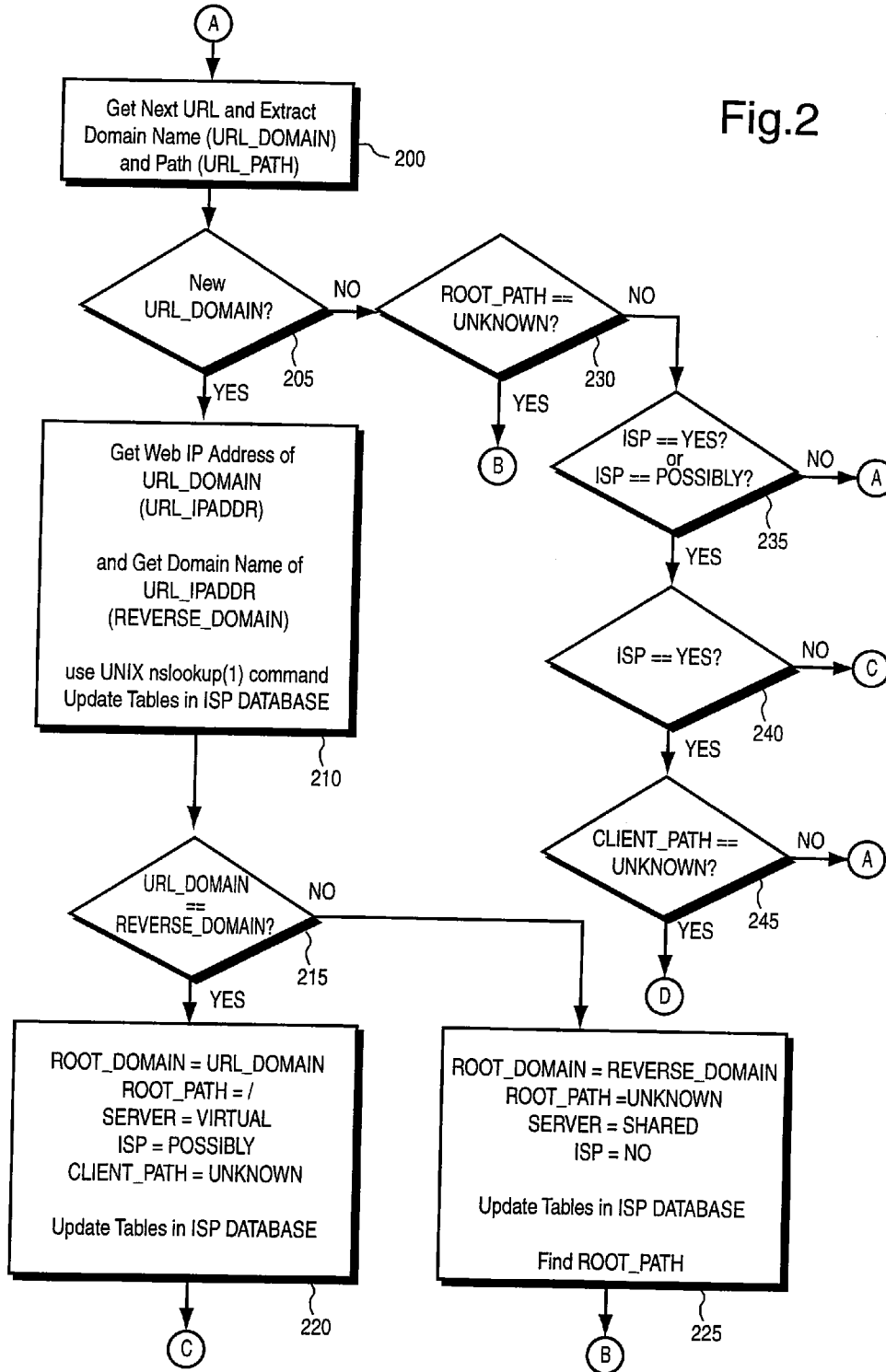


Fig.2



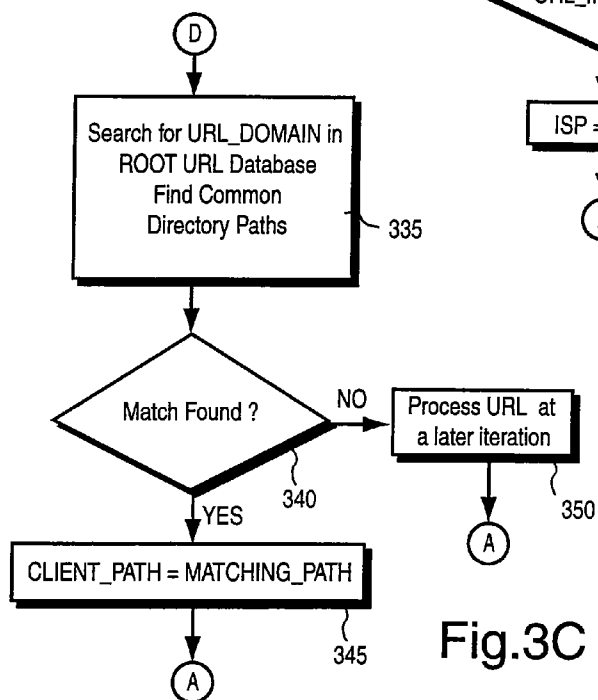
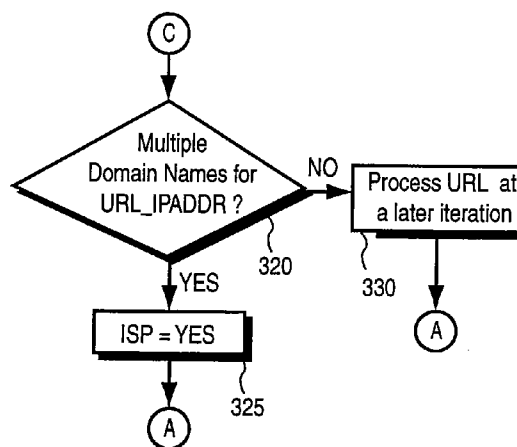
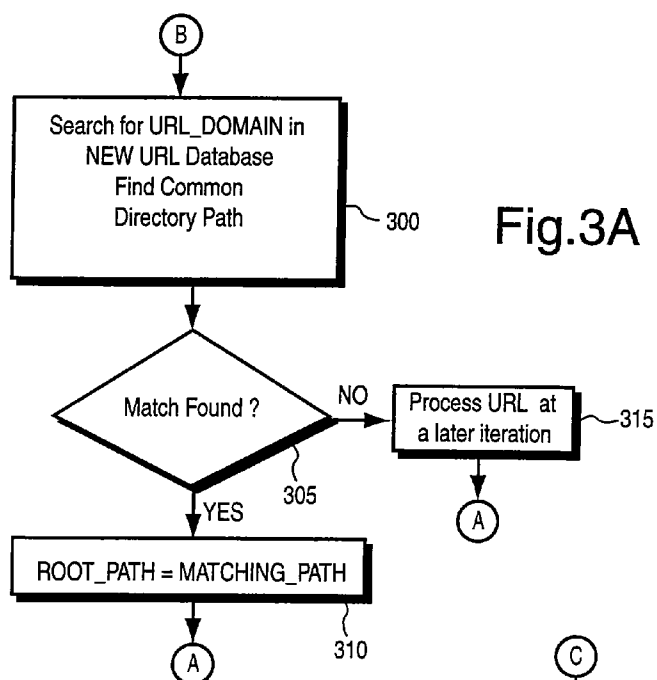
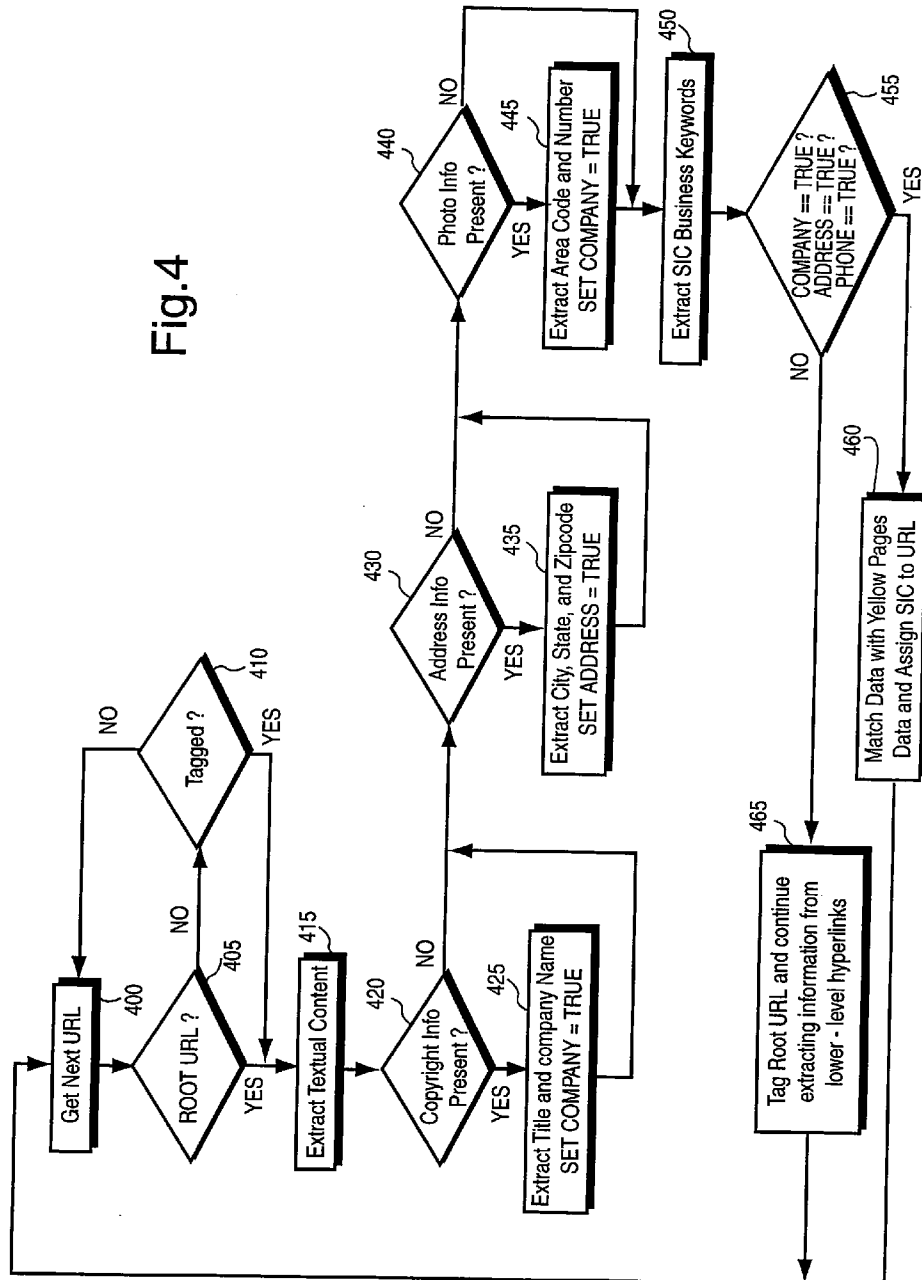


Fig.4



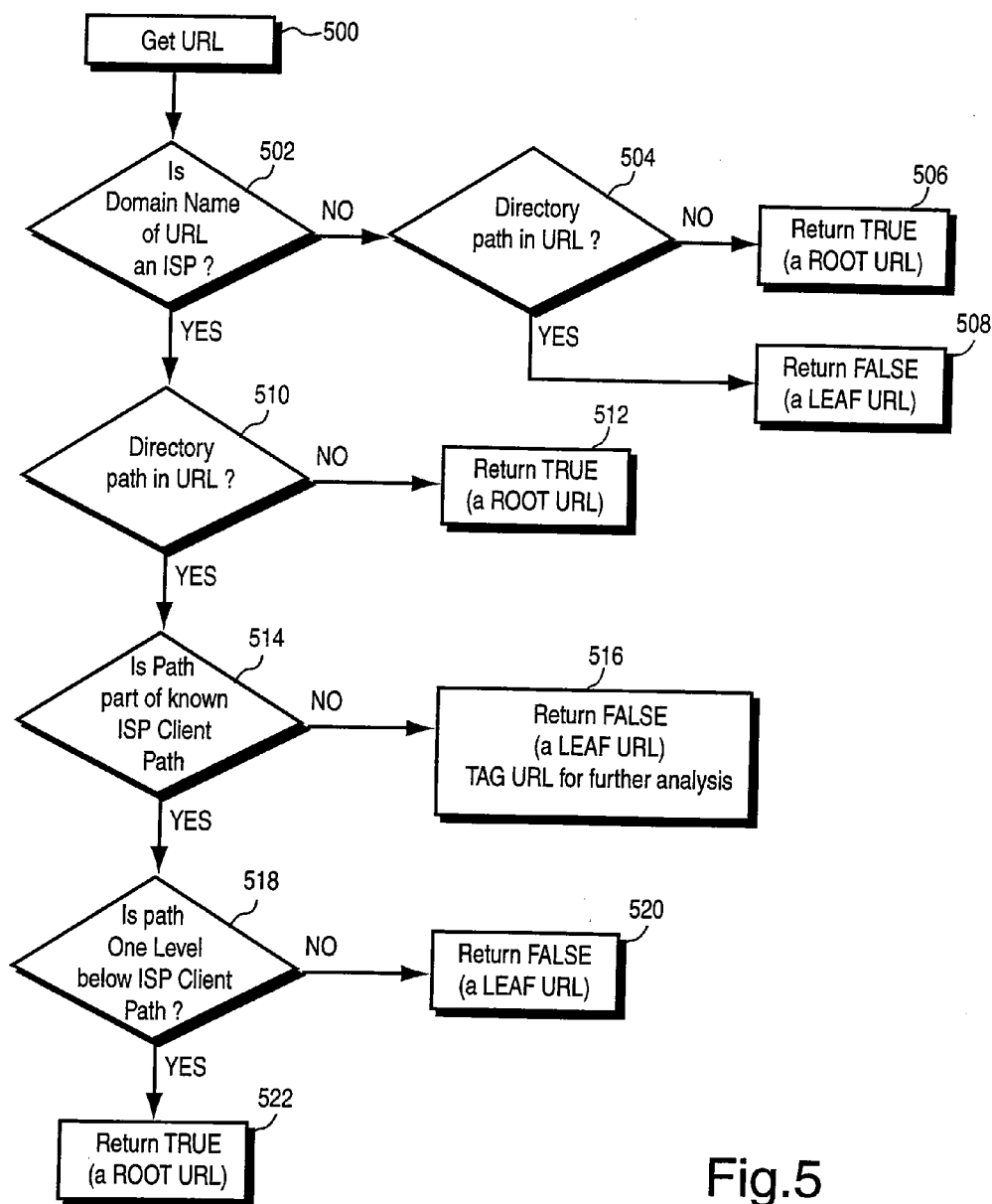


Fig.5

The diagram shows an HTML document structure enclosed in a rectangular box. It includes a header section with meta tags, a body section with contact information, and a footer with a copyright notice. Hand-drawn squiggly lines with numerical labels point to specific parts of the document: label 600 points to the opening <HTML> tag, label 602 points to the 'Company Name' text in the body, and label 604 points to the 'Copyright' text in the footer.

```
<HTML>  
<HEAD>  
<TITLE>Company Name's Home Page</TITLE>  
  
Company Name  
Street Address  
City, State Zip  
Phone: ###-###-###  
Fax: ###-###-###  
  
Copyright (C) 1996 Company Name. All rights reserved.
```

Fig.6

SYSTEM AND METHOD FOR GEOGRAPHICALLY ORGANIZING AND CLASSIFYING BUSINESSES ON THE WORLD-WIDE WEB

BACKGROUND OF THE INVENTION

[0001] The present invention generally relates to a resource discovery system and method for facilitating local commerce on the World-Wide Web and for reducing search time by accurately isolating information for end-users. For example, distinguishing and classifying business pages on the Web by business categories using the Standard Industrial Classification (SIC) codes is achieved through an automatic iterative process which effectively localizes the Web.

DESCRIPTION OF THE RELATED ART

[0002] Resource discovery systems have been widely studied and deployed to collect and index textual content contained on the World-Wide Web. However, as the volume of accessible information continues to grow, it becomes increasingly difficult to index and locate relevant information. Moreover, global flat file indexes become less useful as the information space grows causing user queries to match too much information.

[0003] Leading organizations are attempting to classify and organize all of Web space in some manner. The most notable example is Yahoo, Inc. which manually categorizes Web sites under fourteen broad headings and 20,000 different sub-headings. Still others are using advanced information retrieval and mathematical techniques to automatically bring order out of chaos on the Web.

[0004] Solutions to solve this information overload problem have been addressed by C. Mic Bowman et al. using Harvest: A Scalable, Customizable Resource Discovery and Access System. Harvest supports resource discovery through topic-specific content indexing made possible by a very efficient distributed information gathering architecture. However, these topic specific brokers require manual construction and they are geared more for academic and scientific research than commercial applications.

[0005] Cornell's SMART engine developed by Gerard Salton uses a thesaurus to automatically expand a user's search and capture more documents. Individual, Inc. uses this system to sift through vast amounts of textual data from news sources by filtering, capturing, and ranking articles and documents based on news industry classification.

[0006] The latest attempts for automated topic-specific indexing include the Excite, Inc. search engine which uses statistical techniques to build a self-organizing classification scheme. Excite Inc.'s implementation is based on a modification of the popular inverted word indexing technique which takes into account concepts (i.e., synonymy and homonymy) and analyzes words that frequently occur together. Oracle has developed a system called ConText to automatically classify documents under a nine-level hierarchy that identifies a quarter-million different concepts by understanding the written English language. ConText analyzes a document and then decides which of the concepts best describe the document's topic.

[0007] The systems described above all attempt to organize the vast amounts of data residing on the Web. However, these mathematical information retrieval techniques for classifying documents only work when the message of a document is

directly correlated to the words it contains. Attempts to isolate documents by regions or to separate business content from personal content in an automated fashion is not addressed by any conventional system or structure.

SUMMARY OF THE INVENTION

[0008] It is therefore an object of the present invention to provide a method and system for overcoming the above-mentioned problems of the conventional methods and techniques.

[0009] The invention is based on a heuristic algorithm which exploits common Web page design principles. The key challenge is to ascertain the owner of a Web page through an iterative process. Knowing the owner of a Web page helps identify the nature of the content business or personal which, in turn, helps identify the geographic location.

[0010] In a first aspect of the invention, a method of classifying a source publishing a document on a portion of a network, includes steps of electronically receiving a document, based on the document, determining a source which published the document, and assigning a code to the document based on whether data associated with the document published by the source matches with data contained in a database.

[0011] In a second aspect, a search engine is provided for use on a network for distinguishing between business web pages and personal web pages. The search engine includes a mechanism for parsing the content of a hyper-text markup language (HTML) at a web address and searching for criteria contained therein, a mechanism for analyzing a uniform resources locator (URL) of the web address to determine characteristics thereof of a web page at the web address, a mechanism for determining whether the criteria match with data contained in a database, and a mechanism for cross-referencing a match, determined by the determining mechanism, to a second database, to classify a source which published the web page.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The foregoing and other objects, aspects and advantages will be better understood from the following detailed description of a preferred embodiment of the invention with reference to the drawings, in which:

[0013] FIG. 1 shows the process flow diagram of a geographically bound resource discovery system including three main components of the invention (sometime referred to below as "MetroSearch") identified as MetroBot, IPLink, and YPLink;

[0014] FIG. 2 depicts the IPLink flow chart, the process for identifying ISPs and Client Directory Paths;

[0015] FIGS. 3A-3C are sub-processes of the IPLink flow chart shown in FIG. 2;

[0016] FIG. 4 depicts the flow chart of YPLink for identifying business pages;

[0017] FIG. 5 is a flow diagram for determining if a given uniform resources locator (URL) is a Root URL or a Leaf URL; and

[0018] FIG. 6 is a template of a typical business home page.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT OF THE INVENTION

[0019] Referring now to the drawings, and more particularly to FIG. 1, there is shown the general arrangement of a preferred embodiment according to the present invention.

[0020] The underlying insight behind the invention is that individuals and organizations responsible for the design, creation, and maintenance of their home page generally follow some basic unwritten rules. These rules can be exploited to automatically identify the owner of the home page with a high probability of success. Once the owner of the home page is determined, an SIC code is assigned to it by looking up the owner in a Yellow Pages database. If a matching entry exists, then the owner is a business, otherwise the owner is deemed to be an individual with a personal home page.

[0021] FIG. 1 shows a preferred architecture for implementing a geographically bound resource discovery system. The main components of interest are MetroBot 126, IPLink 113, and YPLink 112.

[0022] The World-Wide Web ("the Web") 124 is based on a client-server architecture. The Web is the graphical, multimedia portion of the Internet 120. The client side program is a Web browser 100 and the server side is a computer running the HTTPD program 102. The Web server is accessed through the Internet by specifying a Uniform Resource Locator (URL). User-entered queries are sent to a back-end processor or search engine 104 which gathers results from various databases 106, 108, 110, and 128, and formats the request and presents them back to the user.

[0023] MetroBot 126 is an indexer robot which traverses hyperlinks in HTML documents and indexes the content into a searchable Web index database 128. These hyperlinks or URLs point to other Web pages making it possible to recursively traverse large portions of the Web from a single, well-chosen URL (seed URL). MetroBot begins its traversal from known Root URL 119 such as the home page of a local service provider (SP), such as an internet service provider (ISP). New links that are discovered are stored in New URLs database 118. These links are processed by IPLink 113 and YPLink 112 to extract new Root URLs at which point the whole process repeats itself. Furthermore, YPLink periodically supplements its New URL list by querying global search engines 121 using strategic keywords (e.g., regional city, county, state names, zip codes, and industry specific terms).

[0024] The first level of localization is achieved by limiting URLs to registered domain names 106. IPLink extracts domain names from the New URL database and then queries the InterNIC database 122 where records of registered domain names containing company name, contact, street address, and Internet Protocol (IP) addresses are kept. This InterNIC database can be accessed through the Unix whois(1) command. YPLink merges the InterNIC address database 108 with the Yellow Pages data 110. This process is described in detail below.

[0025] The next level of localization is more complex since most businesses do not have their own registered domain name. Instead, they have their home page hosted on local SPs (or ISPs) or Online Service Providers (OSPs) Web Servers.

[0026] The first step in solving this problem is for IPLink 113 to characterize URLs by their IP addresses. FIGS. 2 and 3A-3C shows the IPLink flow logic. IPLink identifies the following attributes based on the IP addresses of New URLs:

[0027] True/Virtual Web Servers vs. Shared Web Servers.

[0028] ISP vs. Non-ISP hosts.

[0029] Root Domain of URLs.

[0030] Root Path of URLs.

[0031] Client Directory Paths if host is an ISP.

[0032] A new URL is retrieved from the New URL database 200 and is parsed into the domain name and directory path portions. If it is a new domain 205, then its Web IP address (i.e., www.domain.name) is retrieved using the Internet Domain Name Service 122. The Unix nslookup(1) utility 210 returns an IP address given a domain name. The corresponding IP address is stored in the ISP database 114. A reverse lookup 210 of the Web IP address is also performed to determine 215 if the given URL is hosted on a true (or virtual) Web server 220 or a shared Web server 225. A domain name with its own unique Web IP address indicates a true or virtual Web server (non-ISP host). Multiple domain names for a single Web IP address indicates a shared Web server (ISP host).

[0033] The official domain name (Root Domain) 220 and 225 for the IP address is the domain name of the ISP (master/slave name server information returned by whois(1) can also be used to accurately identify the ISP if the Root Domain does not correspond to the ISP). Root Domain is only used for displaying URL information on search results not for further processing.

[0034] Turning to FIG. 3A, for shared servers 225, the Root Path is determined by searching 300 for the given domain name in the New URL database 118 and finding common directory paths 305. If no match is found 315, the URL will automatically be processed at a later iteration 230, otherwise the Root Path is set to the matching path 310.

[0035] Turning to FIG. 3B, for virtual servers 220, the Root Path is simply the root directory ('/'). These servers may or may not be ISPs. If multiple domain names exist for the given IP address 320, then it is classified as an ISP 325, otherwise it is processed at a later iteration 330, 235 and 240. It is possible for organizations to become ISPs in the future by simply adding/hosting new domain names on their existing Web servers.

[0036] The directory path where the ISP stores its customers Web pages is called the ISP Client Directory Path 116. This data is initially created manually for a few local ISPs (seed ISPs). This path is identified automatically 335 by searching for the given domain name in the Root URL database 119 and finding common directory paths 340, as shown in FIG. 3C. If no match is found 350, then it is processed at a later iteration 245. Matching paths 345 point to the ISPs Client Directory Path. This process improves over subsequent iterations when enough data is gathered and patterns can be recognized from a large set of ISP Web Servers.

[0037] IPLink encompasses the first phase of identifying and characterizing IP addresses. The next phase is to automatically identify businesses hosted on ISP Web servers.

[0038] FIG. 4 shows the YPLink flow chart. YPLink determines if a Web page belongs to a business or an individual. YPLink takes its input, a URL, from IPLink. FIG. 4 shows the flow diagram for the YPLink process. The first step after retrieving a URL 400 is determining if it is a "Root URL" or a "Leaf URL" 405.

[0039] A Root URL is the entry point for an organization's or individual's home page on the World-Wide Web. A Root URL may or may not be the same as the Home page. Leaf

URLs, on the other hand, are links below an organization's Root URL. Four factors are considered in determining a Root URL:

- [0040] 1. Is the URL hosted on a Service Provider's Web Server?
- [0041] 2. Is the URL on a virtual Web Server?
- [0042] 3. Does the URL contain a directory path?
- [0043] 4. Is the directory path a known Service Provider's Client Directory?

[0044] IPLink determines the SP Client Directory Path as described above. The ISP database 114 contains information about Client Directories for various ISPs.

[0045] FIG. 5 shows the Root URL flow logic. A given URL is retrieved 500 and parsed into two components: domain name and directory path. The domain name is analyzed to see if it is an ISP 502. If multiple IP addresses are associated with the domain name, then the domain name is an ISP. If the domain name is not an ISP, then the directory path component is checked 504. A missing directory path signifies a Root URL 506, otherwise it is a Leaf URL 508.

[0046] If the domain name is an ISP 510, then it is also a Root URL if no directory path exists 512. If a directory path exists 514, then the path is compared to a list of known ISP Client Directory paths. No match 516 indicates a Leaf URL, otherwise the directory path level is analyzed 518 for final Root URL determination. If the path is one directory level below the Client Directory path then it is a Root URL 522, otherwise it is a Leaf URL 520.

[0047] After a URL is determined to be Root URL, then the home page it points to is analyzed 415 to see if it follows some basic guidelines. A typical home page layout is illustrated in FIG. 6. Other than following HTML requirements, there is no rule or standards for the layout of textual content. The key pieces of information required to ascertain the owner of a Web page are 1) company name, 2) zip code, and 3) telephone number. These three pieces of information do not have to exist in the Root URL. They can reside anywhere among various Leaf URLs beneath a Root URL. In many cases, this information is stored in a file called about.html. However, the same information could be stored in other, similarly named files, as would be known to those skilled in the art taking the present specification as a whole. The process described below extracts this information automatically and assigns it to the Root URL being analyzed.

[0048] The company's name is usually included in the HTML TITLE tag 600. However, the company's name could be included in other locations, as would be known to those ordinarily skilled in the art within the purview of the present specification. The layout of the address, if present, usually is in a standard recognizable format 602. Most businesses also tend to include copyright notices near the bottom of their documents. A string search for "copyright", "©", and "©" is performed near the bottom 604 of the home page. The company name usually appears near the copyright notice. A match of the organization or individual's name in the copyright field 420 and the TITLE field 425 provides the first indication of the owner of the home page. If no match is found, then the URL is tagged for further analysis during the next iteration.

[0049] The next step is to analyze the URL for address 430 information. Addresses have an easily identifiable format. In the U.S., the format is the city name followed by a comma and then followed by the full state name or abbreviation and finally a five or nine digit zip code. However, other common

formats/methods also are possible and would be known to those ordinarily skilled in this art field to locate the zip code. This string is parsed in the HTML file after stripping all tags 435. The only information required is the 5-digit zip code since the city and state can be determined by this field alone. YPLink stores addresses associated with Root URLs and domain names in an address database 106.

[0050] If a phone format field is present then it is also extracted and stored 440. U.S. phone field is a 10-digit field where the first three digits representing the area code are optionally enclosed in parentheses or separated by a dash, space, or a period, and then followed by a 7-digit number which is separated by a dash, space, or a period after the third digit 445. Other similar methods of identifying a phone number are known to those ordinarily skilled in the art.

[0051] The pair consisting of the company name and zip code are usually enough to identify a business 455. A query is constructed using this pair and sent to a Yellow Pages database server. This database is indexed by business names and zip codes. If a single match is found, then the resulting SIC code is assigned to the corresponding Root URL 460. If multiple entries are matched, then the phone field is also included in the query to assure that only a single entry is retrieved. If no match is found, then the URL is tagged 465 for further analysis of lower-level hyperlinks during the next iteration. The matching data is stored in an enhanced Yellow Pages database 108.

[0052] If no match is found at any level, then the page is tagged 450 as a personal page with an SIC code assigned according to the closest match based on the Business Semantic Terminology database 110. This database is a proprietary thesauri of keywords relating business categories in the Yellow Pages and other emerging industries such as Internet technology to extended SIC codes.

[0053] While the invention has been described in terms of a single preferred embodiment, those skilled in the art will recognize that the invention can be practiced with modification within the spirit and scope of the appended claims.

[0054] For example, while the invention above has been described primarily in terms of (e.g., implemented in) a software process and a system employing software and hardware, the invention could also be implemented with hardware as would be known by one of ordinary skill in the art taking the present specification as a whole.

Having thus described my invention, what I claim as new and desire to secure by Letters Patent is as follows:

1. A method of classifying a source publishing a document on a portion of a network, comprising steps of:

- electronically receiving a document;
- based on the document, determining a source which published the document; and
- assigning a code to said document based on whether data associated with the document published by the source matches with data contained in a database.

2. The method according to claim 1, wherein said portion of said network comprises a graphical multimedia portion of said network, said source comprises a Web site publishing a home page, and said network comprises the Internet.

3. The method according to claim 2, wherein said graphical multimedia portion of said network comprises the World-Wide Web (WWW) and said document comprises a Web document,

wherein said step of assigning a code includes determining that the Web site comprises a first entity when there is a

- match of the Web site with said data, and determining that the Web site comprises a second entity when there is no match of the Web site with said data.
4. The method according to claim 1, wherein said step of determining a source includes:
- extracting a domain name from a predetermined uniform resources locator (URL) database;
 - querying a database for storing registered domain names; and
 - merging an address database with predetermined data.
5. The method according to claim 4, wherein said predetermined data comprises Yellow Pages data, wherein said step of determining further comprises:
- characterizing uniform resource locators (URLs) by their Internet Protocol (IP) addresses including identifying a plurality of attributes based on the IP addresses of new URLs, a new URL being retrieved and parsed into a domain name and directory path portions, and
 - determining, based on said domain name, whether a selected URL is hosted on one of a true server and a shared server.
6. The method according to claim 5, said step of determining further comprising:
- for a shared server, determining a root path by searching for the given domain name in a new URL database and identifying common directory paths,
 - wherein, when no match is present, the URL is processed subsequently at a later iteration, and, when a match is present, the root path is set to a matching path.
7. The method according to claim 6, wherein said step of assigning a code comprises:
- automatically identifying a business associated with the source publishing said document, said business being hosted on a Service Provider (SP) Web server.
8. The method according to claim 7, wherein said step of assigning a code further comprises:
- receiving a URL based on said determining step; and
 - a URL determining step for determining whether said URL comprises one of a root URL and a leaf URL.
9. The method according to claim 8, wherein said root URL comprises an entry point for a home page on the World-Wide Web, and a leaf URL comprises a link below a root URL, wherein said URL determining step comprises:
- parsing said URL into a domain name component and a directory path component;
 - analyzing the domain name in said domain name component to determine whether it is associated with an SP;
 - when the domain name is not associated with an SP, checking the directory path component to judge whether a directory path is missing, a missing directory path indicating a root URL;
 - when the domain name is associated with an SP, checking whether a directory path does not exist to thereby determine that said domain name comprises a root URL, and when a directory path exists, then comparing the path to known SP Client Directory paths.
10. The method according to claim 9, further comprising:
- when said URL is determined to be a root URL, analyzing a home page associated with said root URL automatically to extract home page data contained therein and assigning the home page data to the Root URL being analyzed.
11. The method according to claim 10, further comprising:
- comparing said home page data with data in a predetermined business organizations database,
 - wherein, when there is a match, said code is assigned to the corresponding root URL, and, when no match is found, said URL is identified for subsequent analysis of lower-level hyperlinks during a next iteration of said method.
12. The method according to claim 11, wherein when no match is found at any level, said home page is identified as a personal page.
13. A method of automatically assigning a document a code for distinguishing a first-type page from a second-type page, comprising steps of:
- electronically receiving a document;
 - based on the document, determining a source which published the document; and
 - assigning a code to said document based on whether the source matches with data contained in a database.
14. A search engine for use on a network for distinguishing between business web pages and personal web pages, comprising:
- means for parsing the content of a hyper-text markup language (HTML) at a web address and searching for criteria contained therein;
 - means for analyzing a uniform resources locator (URL) of the web address to determine characteristics thereof of a web page at the web address;
 - means for determining whether said criteria match with data contained in a database; and
 - means for cross-referencing a match, determined by said determining means, to a second database, to classify a source which published the web page.
15. A search engine according to claim 14, wherein said criteria include at least one of an address, a telephone numbers, a facsimile number, a contact and a key-word contained in said HTML, and
- wherein the characteristics of said web page include a geographical location and a web host computer.
16. A search engine according to claim 14, wherein said database includes a Business Semantic Terminology database having information related to business categories in a Yellow Pages directory.
17. A search engine according to claim 14, wherein said second database includes a Yellow Pages database.
18. A search engine according to claim 14, wherein said web page comprises hyperlinks, and said means for parsing comprises an indexer robot for traversing said hyperlinks in said web page and a web index database,
- said indexer robot for indexing a content of said web page into said web index database.
19. A search engine according to claim 14, wherein said means for analyzing comprises:
- means for determining whether said URL comprises one of a root URL and a leaf URL.
20. A search engine according to claim 19, wherein said root URL comprises an entry point for the web page on the World-Wide Web, and a leaf URL comprises a link below a root URL, said search engine further comprising:
- means for parsing said URL into a domain name component and a directory path component;
 - means for analyzing the domain name in said domain name component to determine whether it is associated with an SP;

means for checking the directory path component to judge whether a directory path is missing, when the domain name is not associated with a service provider (SP), a missing directory path indicating a root URL, and for checking whether a directory path does not exist to thereby determine that said domain name comprises a root URL, when the domain name is associated with an SP;

means for comparing the path to known SP Client Directory paths, when a directory path exists;

means for analyzing a home page associated with said root URL, when said URL is determined to be a root URL, thereby automatically to extract home page data contained therein; and

means for assigning the home page data to the Root URL being analyzed.

* * * * *