CLAR 1006-3

METHOD AND APPARATUS FOR RESPONDING TO END-USER REQUEST FOR INFORMATION

Inventors: Remigiusz K. Paczkowski,
Matthew E. Westover,
Dominic Bennett,
Benjamin Lindsey,
Scott Vandevelde,
Jason Fairchild, and
Adam William Pingel

RELATED APPLICATIONS

[0001] This application claims the benefit of US provisional Patent Application Nos. 60/603,140 entitled "Method and Apparatus for Responding to End-User Request for Information" by Westover et al. filed on August 19, 2004; 60/637,684 entitled "Method and Device Publishing Cross-Network User Behavioral Data" by Wohlers et al. filed on December 20, 2004; 60/662,680 entitled "Method and Device for Publishing Behavioral Observations to Customers" by Eagle et al. filed on March 17, 2005; and 60/660,798 entitled "Method and Apparatus for Responding to End-User Requests for Information" by Westover et al. filed on March 11, 2005. This application continues-in-part the disclosure of US Patent Application No. 11/015,583 entitled "Search Engine for a Computer Network" by Anthony G. Martin filed on December 17, 2004. These identified applications are incorporated by reference for all purposes.

FIELD OF THE INVENTION

[0002] The present invention relates to an advanced search engine. The advanced search engine may include a client component for monitoring an end-user's browsing activity, a remote server (may comprise one or more computers) for storing and processing data received from the client component, and a module that process web pages and serves search results to end-users. The advanced search engine may collect web pages for keywords of proven interest, fetch web pages requested by end-users, generate snippets or abstracts of the web pages, eliminate duplicate web pages, rank the importance of the web pages, and provide

relevant web pages or links to web pages in response to an end-user search request for information regarding one or more keywords, for example. Technical problems solved, measures used and results obtained are discussed below.

BACKGROUND

[0003] One approach to search engines, taken by Google, is to organize the world's information and make it universally accessible and useful. Another approach, once taken by Dogpile, is to have a meta-search engine aggregate the results of other search engines. These approaches create a great haystack of results. For instance, the keyword "cheap travel" returns about 18,000,000 results from Google, about 85,800,000 from Yahoo and 68,377,619 from MSN, as of summer 2005!

[0004] Much work has been done to float the "needles" to the top of the results haystack, to devise methods of ranking links returned in response to a query. Google's published patent applications propose, in their titles, to use local inter-connectivity, article information, location awareness and other factors to decide on the position of results. Yahoo, Overture and Microsoft also have worked to refine their presentation of results.

[0005] In any set of information, a search term is sometimes not enough to determine what results are sought. In one sense, the search term may be ambiguous, as extensively discussed in Bharat et al., "Generating User Information for Use in Targeted Advertising", US 205/0131762 A1 published June 16, 2005 and in Carrasco et al., "Disambiguation of Search Phrases Using Interpreation Clusters", US 2005/0015366 A1 published January 20, 2005. The term "jaguar" might refer to cars, animals, a football team, or an operating system. Even if a term were unambiguous, different users might prefer to access different information. For instance, teenage travelers, business travelers and luxury travelers look for different travel arrangements and accommodations, potentially using similar search terms.

[0006] It is desirable to return the most relevant results, whether in response to a search or, more generally, on an information feed. The growing number of documents published on web sites (and of documents accessible on private servers) invites development of alternative or improved technology to quickly return relevant results responsive to users' queries. In effect, to find the 50 or 100 most relevant web sites for a particular user whose keyword is "cheap travel" and effectively summarize them for the user. This further invites

CLAR 1006-3

development of technologies that personalize the information returned, whether content, sponsored content or advertising, based on the interests of the user.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] **FIG. 1** is a schematic diagram of an example computer that may be used in various embodiments.

[0008] **FIG. 2** shows a schematic diagram of a computing environment in which this technology may operate.

[0009] **FIG. 3** schematically illustrates the tracking of a conversion process.

[0010] **FIG. 4** depicts a high level block diagram of components that can function together to produce a search engine that personalizes sponsored and non-sponsored information delivery.

[0011] **FIG. 5** is an example of a cookie or data record that reports a URL clicked.

[0012] **FIG. 6** is an example of a cookie that stores the most recent commercial keyword of interest.

[0013] **FIG. 7** is an example of summary data maintained by the behavior watcher that may be written to persistent memory.

[0014] **FIG. 8** illustrates a bit string further summarized by category.

[0015] **FIG. 9** illustrates buckets that may be applied to recency of using search words, clicking through, entering URL destinations or other indications of user interest.

[0016] **FIG. 10** illustrates buckets that can be applied to frequency of visits.

[0017] **FIG. 11** illustrates how a category ID can be associated with both the category name and a revenue priority.

[0018] **FIG. 12** compares the insight that a behavior watching module on a user's machine gives regarding web surfing with the limited view that a server can achieve.

[0019] **FIG. 13** depicts a web page highly relevant to the keyword "cars".

[0020] **FIG. 14** is a counter example of a page weakly relevant to "cars".

[0021] **FIG. 15** juxtaposes the results for "cars" from Yahoo!, msn Search and Google with an embodiment of the search engine described herein.

[0022] **FIG. 16** illustrates user differentiation using results from a behavior watching module.

CLAR 1006-3

[0023] **FIG. 17** depicts traffic to a particular domain, from a particular search engine, by keyword.

[0024] **FIG. 18** depicts an architecture for following links and retrieving content.

[0025] **FIG. 19** illustrates the importance of recent behavior in response to banner ads, measured by click through rates.

[0026] **FIG. 20** depicts how the window scoring routine begins with the first word of text within the document.

DETAILED DESCRIPTION

[0027] The following detailed description is made with reference to the figures. Preferred embodiments are described to illustrate, not to limit the scope of the claims. Those of ordinary skill in the art will recognize a variety of equivalent variations on the description that follows. Persons of ordinary skill in the art will recognize, however, that the embodiments described can be practiced without one or more of the specific details. In other instances, well-known details are not shown or described to avoid obscuring aspects of the embodiment.

[0028] Being computer-related, it can be appreciated that the components disclosed herein may be implemented in hardware, software, or a combination of hardware and software (e.g., firmware). Software components may be in the form of computer readable program code stored in a computer-readable storage medium, such as memory, mass storage device, or removable storage device. For example, a computer readable medium may comprise computer-readable program code for performing the function of a particular component. Likewise, computer memory may be configured to include one or more components, which may then be executed by a processor. Components may be implemented separately in multiple modules or together in a single module.

[0029] Embodiments and aspects of embodiments described below can be applied to solve various technical problems. One problem applies to a large network: how to monitor and usefully aggregate patterns of communication among users, search engines and documents accessed. In one scenario, the users are surfing the Internet at widely disbursed personal computers; the search engines include Baidu, Google, Yahoo! and MSN; the documents are pages posted on web sites around the world. In another, employees use an

enterprise intranet with an enterprise search engine to locate reference documents exposed on workgroup servers. Addressing these problems may involve positioning a behavior observing module where it can monitor the communication channels in use and report observations to a server, preferably without disrupting the users' routines. In one embodiment, the behavior observing module may run on the user's personal computer (such as a desktop, laptop or handheld computer or media center device). The module can observe both communications and the status of the user's machine when the communications take place. For example, what search terms and results did a user follow to a particular web site? The module can achieve distributed processing and substantially reduce the resources required to aggregate communication behavior by filtering reports and categorically encoding activity. It may report observations to the server as resources are available or when a user browses to an affiliated domain. As part of the ordinary uploading of cookies to domains accessed, observation cookies can be transmitted to the server. The server can aggregate reported patterns of communication. One technical result is receiving reports from communication monitors positioned across a large network and aggregating patterns, including the status or state of individual computers when communications took place. In turn, the aggregated patterns of communications can be used to pre-organize information for retrieval or publication, in anticipation of a query or contact.

[0030] A related technical problem is how to organize over-abundant electronic records based on the current and recent status of a particular terminal connected to a network, to enhance the relevance of the first records presented to a user. The over-abundant electronic records may come from web sites world wide, such as the web sites for "jaguar." Or, they may be documents stored on workgroup servers. They are over-abundant in the sense that they are too numerous to display on a user's screen without repeatedly pressing "page down" or the like. The current and recent status of the terminal, preferably associated with a particular user, may include web sites visited within the last 24 hours (or some other period) before a search query was submitted. Or, it may include a history of documents retrieved from workgroup servers. Either type of activity may be preprocessed and categorically classified. The period for reporting terminal status may precede a query or request for a personalized electronic journal that selects and filters the information based on the current and recent status of the terminal. The technical method again involves a behavior watching

module running on the user's computer. In this embodiment, the module summarizes terminal status, publishes the summary to an electronic record (e.g., a cookie) and communicates the record to a search engine or other site that the user contacts. Reporting categorical summaries of status or activity distributes processing and reduces the need for server-based resources. The search engine uses the information, for instance, to determine what sense of "jaguar" is of interest? A highly involved auto category user who searches for "jaguar" would get Jaguar auto related links, while a person with no relevant category involvement would get a mix of auto, animal, etc. links. Category profiles may be developed to further categorize a user's interests. A new parent in the suburbs might be recognized from content accessed and be assigned to a different category for home accessories or cars than a single gen-X 20-something person. Life change events also might be recognized, such as marriage, home buying or parenthood. Like the "jaguar" example, for a "travel Italy" inquiry, the search engine might determine the style of travel that interests the user and organize the over-abundant electronic records accordingly. The technical result is respond to a query or contact based on an electronic report of the status or recent activity of a terminal, selecting from over-abundant electronic records a particular set of records that are most likely relevant to the current and recent status of the terminal.

[0031] Web-wide behavioral targeting differs substantially from site-side behavioral targeting. Practicing site-side behavioral targeting, a group of affiliated sites attempt to identify commercial behaviors. The sites typically serve ads, such as portals and news sites. Messages are displayed on the affiliated sites, responsive to behavior that is recognized from visits to the affiliated sites. The results of site-side behavioral targeting are better than non-behaviorally targeted campaigns, but depend on insight that can be gained from a narrow portion of user's behavior, as illustrated by **FIG. 12**, which is described below. Web-wide behavioral targeting, in contrast to site-side, can identify commercial surfing behavior and, optionally, non-commercial surfing behavior of interest, across all sites, even sites to do not accept advertising. This is significant because 90% of commercial Web activity takes place at sites or on pages that do not display advertising. Therefore, portals see less than 10% of user's total commercial activity, as measured by page views. Non-portal publishers see less than 3% of their users' total commercial activity. A behavior watching module may simplify the relationship with web site operators and allow messages to be displayed across a larger

network of publishers than participate in site-side behavioral targeting. Web-wide behavioral targeting produces better results than site-side behavioral targeting.

[0032] Referring now to **FIG. 1**, there is shown a schematic diagram of an example computer that may be used in various embodiments. Depending on its configuration, the computer shown in the example of **FIG. 1** may be employed as a client computer, a server computer, or other data processing device. The computer of **FIG. 1** may have less or more components to meet the needs of a particular application. As shown in **FIG. 1**, the computer may include a processor 101, such as those from the Intel Corporation, Advanced Micro Devices or IBM, for example. The computer may have one or more buses 103 coupling its various components. The computer may include one ore more input devices 102 (e-g., keyboard, mouse), a computer-readable storage medium (CRSM) 105 (e.g., floppy disk, CD-ROM), a CRSM reader 104 (e.g., floppy drive, CD-ROM drive), a display monitor 109 (e.g., cathode ray tube, flat panel display), a communications interface 106 (e.g., network adapter, modem) for coupling to a network, one or more data storage devices 107 (e.g., hard disk drive, optical drive, FLASH memory), and a main memory 108 (e.g., RAM). Software embodiments may be stored in a computer-readable storage medium 105 forwarding into a data storage device 107 or main memory 108. Software embodiments in main memory 108 may be executed by processor 101. In the example of **FIG. 1**, main memory 108 is shown as having a behavior watching and message delivery program 120, which is further described below. Behavior watching and message delivery program 120 and other programs in main memory 108 may be loaded from a computer-readable storage medium 105, a data storage device 107, or over the Internet by way of communications interface 106, for example. Behavior watching and message delivery program 120 and other programs in main memory 108 may be executed by processor 101.

[0033] **FIG. 2** shows a schematic diagram of a computing environment in which this technology may operate. In the example of **FIG. 2**, the computing environment includes one or more web server computers 160 (i.e., 160-1, 160-2, ...) , one or more client computers 110, one or more message server computers 140, and other computers not specifically shown. In the example of **FIG. 2**, a client computer 110 communicates with server computers (e.g., a web server computer or a message server computer) over the Internet. As such, arrows 201 denote Internet connections in this example. Intermediate nodes such as gateways, routers.

bridges, Internet service provider networks, public-switched telephone networks, proxy servers, firewalls, and other network components are not shown for clarity.

[0034] A client computer 110 is typically, but not necessarily, a personal computer such as those running the Microsoft Windows TM operating system, far example. A consumer may employ a suitably equipped client computer 110 to get on the Internet and access computers coupled thereto. For example, a client computer 110 may be used to access web pages from a web sever computer 160.

[0035] A web server computer 160 may be a server computer hosting a website, which comprises web pages designed to attract consumers surfing on the Internet. A web server computer 160 may include web pages supporting advertisements, downloadable computer programs, products available for online purchase, and so on. As can be appreciated, a website may be on one or more server computers.

[0036] A message server computer 140 may include the functionalities of a web server computer 160. In one embodiment, a message server computer 140 further includes a database 171. Database 171 may be a commercially available database, such as those available from the Oracle Corporation. Database 171 may store client data received from behavior watching and message delivery programs 120 running in client computers 110. The client data may be transmitted from a client computer 110 to message server computer 140 in a data packet 121. The client data may include navigation and behavioral data obtained by a behavior watching and message delivery program 120 by monitoring a consumer's on-line activities. In the example of **FIG. 2**, message server computer 140 is shown as communicating with one client computer 110 for clarity of illustration. In practice, message server computer 140 receives data packets 121 containing client data from a plurality of client computers 110, each of which has a behavior watching and message delivery program 120. A message server computer 140 may also include downloadable computer programs and files for supporting, updating, and maintaining software components on a client computer 110. The message server computer 140 may include or be connected with a search engine 650.

[0037] Web server computers 160 and message server computers 140 are typically, but not necessarily, server computers such as those available from Sun Microsystems, Hewlett-Packard, or International Business Machines. A client computer 110 may communicate with a web server computer 160 or a message server computer 140 using client-

server protocol. It is to be noted that client-server computing is well known in the art and will not be further described here.

[0038] As shown in **FIG. 2**, a client computer 110 may include a web browser 112 and a behavior watching and message delivery program 120. Web browser 112 may be a commercially available web browser or web client. In one embodiment, web browser 112 comprises the Microsoft Internet Explorer TM web browser. A web browser allows a consumer on a client computer to access a web page. In the example of **FIG. 2**, web browser 112 is depicted as displaying a web page 313 from a web server computer 160, A web page, such as web page 113, has a corresponding address referred to as a "URL" (Uniform Resource Locator), Web browser 112 is pointed to the URL of a web page to receive that web page in client computer 110. Web browser 112 may be pointed to a URL by entering the URL at an address window of web browser 112, or by clicking a link pointed to that URL, for example.

[0039] In one embodiment, behavior watching and message delivery program 120 is downloadable from a message server computer 140 or a web server computer 160. Behavior watching and message delivery program 120 may be downloaded to a client computer 110 in conjunction with the downloading of another computer program. For example, behavior watching and message delivery program 120 may be downloaded to client computer 110 along with a utility program 181 that is provided free of charge or at a reduced cost. Utility program 181 may be a wallet or calendar program, for example. Utility program 181 may be provided to a consumer in exchange for the right to deliver advertisements to that consumer's client computer 110 via behavior watching and message delivery program 120. In essence, revenue from advertisements delivered to the consumer helps defray the cost of creating and maintaining the utility program. Behavior watching and message delivery program 120 may also be provided to the consumer along with free or reduced cost access to an online service, for example.

[0040] Behavior watching and message delivery program 120 is a client-side program in that it is stored and run in a client computer 110. Behavior watching and message delivery program 120 may comprise computer readable program code for displaying advertisements in a client computer 110 and for monitoring the online activity of a consumer on the client computer 110. It is to be noted that the mechanics of monitoring a consumer's online activity, such as determining where a consumer is navigating to the URL of web pages received in

client computer 110, the domain names of websites visited by the consumer, what the consumer is typing on a web page, what keyword the consumer is providing to a search engine, whether the consumer clicked on a link or an advertisement, when the consumer activates a mouse or keyboard, and the like, is, in general, known in the art and not a further described here. For example, behavior watching and message delivery program 120 may learn of consumer online activities by receiving event notifications from web browser 112.

[0041] Behavior watching and message delivery program 120 may record the consumer's online activity for reporting to message server computer 140. The recorded consumer online activity is also referred to as "client data," and provided to message server computer 140 using data packets 121. Message server computer 140 may use the client data to provide targeted advertisements to the consumer. Message server computer 140 may include the advertisement or data for displaying the advertisement in a message unit 141. In the example of **FIG. 2**, the targeted advertisement is labeled as advertisement 116 and displayed in a presentation vehicle 115. Presentation vehicle 115 may be a pop-under, pop-up, separate browser window, custom browser window, or other means for displaying an advertisement on a computer screen. Techniques for delivering advertisements to client computers using a client-side program are also disclosed in commonly-owned U.S. Application No. 10/152,204, entitled "Method and Apparatus for Displaying Messages in Computer Systems," filed on May 21, 2002 by Scott G. Eagle, David L. Goulden, Anthony G. Martin, and Eugene A. Veteska, which is incorporated herein by reference in its entirety.

[0042] As will be more apparent below, behavior watching and message delivery programs are primarily used to obtain client data far building a search engine index, not necessarily to display presentation vehicles in a client computer 110, That is, a behavior watching and message delivery program does not necessarily have to display advertisements in a client computer 110. This is advantageous in that consumers may be allowed to obtain a free or reduced cost utility program 181 (or other benefits) without having to see advertisements from the provider or sponsor of the utility program.

[0043] **FIG. 3** schematically illustrates the tracking of a conversion process. A "conversion" occurs when a consumer makes an online purchase or registration in response to an advertisement presented to her. As will be more apparent below, the embodiments may use conversion as a factor in ranking a keyword-link combination. In the example of **FIG. 3**, web

pages 202 (i.e., 202-1, 202-2,...) may be sequentially displayed in the same or separate windows of web browser 112. Each web page 202 includes a page identifier 21 13 (i.e., 210-1,210-2, ...), which may be a URL. Behavior watching and message delivery program 120 records the 20 URLs of web pages 202 viewed by the consumer as well the amount of time the consumer spent with each web page as navigation data. For purposes of the present disclosure, the web pages "viewed" by consumers are those web pages received in client computers. In the example of **FIG. 3**, navigation data 627 comprises log entries 117 (i.e., 117-1, 117-2, ...). Each log entry 117 includes a machine ID anonymously identifying the client computer 110 (or the consumer), a page identifier, and a time stamp indicating when the log entry 417 was made. The time stamps between log entries 117 provide an estimate of the amount of time the consumer spent viewing the indicated web page. A log entry 117 may be created for each web page 202 viewed by the consumer. For example, a log entry 117-1 may be created when the consumer clicks on a link 501 to receive web page 202-2. In client computer 110, a log entry 117-2 may be created when the consumer receives web page 202-3 in client computer 110, and so on.

[0044] In the example of **FIG. 3**, web page 202-2 is also referred to as a "landing page" because it is the web page directly pointed to by an advertisement having a link in web page 202-1. A website that sells products on line may also have a "confirmation page" 202-5. A confirmation page is a web page provided to the consumer to confirm a just completed online purchase. Alternatively, a confirmation page may involve confirmed submission of registration information. A website may have "intermediate pages" 202-3, 202-4, and so on between a landing page 202-2 and a confirmation page 202-5. An intermediate page may be an online product catalog, shopping cart, and other types of web pages. The page identifiers of landing and confirmation pages of popular or partner websites may be stored in database 171 of message server computer 140, the page identifiers of the landing and confirmation pages may be compared to those in navigation data 627 of client computers 110 to determine the number of consumers making online purchases from a particular web page. As can be appreciated, web pages from which consumers make purchases may be highly relevant, and may thus be deemed a consumer preferred page. Techniques for monitoring consumer purchase behavior are also disclosed in commonly-owned U. S. Application No. 10/464,419, entitled "Generation of Statistical Information in a Computer Network," filed by David L.

CLAR 1006-3

Goulden and Dominic Bennett on June 17, 2003, which is incorporated herein by reference in its entirety.

Process Flow

[0045] **FIG. 4** is a high level block diagram of components that can function together to produce a search engine that personalizes sponsored and non-sponsored information delivery. As background, behavior watching modules currently capture behavioral web-navigation information on over 40 million machines in the U.S. and abroad. The information collected will expand and be adapted as needed. This navigational behavior information is currently used to power web-wide GMT ad vehicles (e.g., pop-ups and slide-unders) as well as power many data-rich applications that give unique insights into consumer behavior across the web. The behavior watching module has the ability to see the entire search life-cycle, from observing the initial keyword entered, to the resultant links displayed, to post-click metrics including dwell times, pages viewed, repeat visits, and post-click purchases or registrations. By aggregating a user's behavior, category profiles and life change events can be determined. User sophistication with search engine usage and degree of category involvement can be determined over time. Using this information, one can select and re-rank the search results, filtered by humans who requested the searches, not by bots.

Data Collection

[0046] Web usage statistics are collected (401) using behavior watching modules (120) for users searching on selected search engines across the Internet or an enterprise intranet. The modules send back information related to their searches and how effective their searches were on each particular search engine for each particular keyword. The information is captured in a database, which is loaded daily or at some other concurrent frequency. Information available includes what the user saw and how they responded. URLs seen by the user may be displayed in algorithmic or natural sections the results. Pertinent information available for a single search term includes machine id, keyword, search engine where search was performed, resultant URLs, algorithmic URLs, bidded URLs, paid inclusion URLs, whether or not a URL was clicked, number of pages viewed, dwell time, repeat visits and user metrics such as category involvement and search engine sophistication.

CLAR 1006-3

[0047] Some statistics from US users can bring to life the analytical reach of data collection and ranking. The one million most frequently searched keywords presently account for just more than half of the searches conducted on the major search engines, 53% of the searches. The 45,000 most frequently searched keywords account for 38%. The distribution of keywords entered as searches can be represented by a Zipf distribution, which plots as a straight line on a graph with logarithmic scale on both axes. See, e.g., Jacob Nielson, "Diversity is Power for Specialized Sites", Alertbox (June 16, 2003) accessed August 13, 2005 at http://www.useit.com/alertbox/20030616.html. It is estimated that 2,000 behavior watching modules will generate enough data to rank the 10,000 most frequent searches. A larger group of 20,000 behavior watching modules will cover the 45,000 most frequently searched keywords. A base of 40 million behavior watching modules may capture 30 billion rows of data per month, filtered from 150 billion page views observed. Among the page views observed, on the order of 8 billion commercial events per month are noted, including more than 2.5 million purchases. These numbers and even a passing familiarity with statistics should excite the reader to aggregating the intelligence and behavior of a multiplicity of searchers, and presenting by popular acclaim the most significant web sites on the first page of results from a search engine. A search engine that uses aggregated consumer behavior is well-positioned to rate the authority and popularity of pages/documents as responsive to a search request.

[0048] **FIGS. 5-11** are examples of data records or cookies that can communicate information from the behavior watching module to a ranking and link-following server and to a search engine that produces personalized results. **FIG. 5** is an example of a cookie or data record that reports a URL clicked. A keyword 501, time stamp 502 and list of one or more URLs selected by the user 503 are conveyed, for instance in a cookie when a user reaches a web site affiliated with one of the servers or in a data record when the behavior watching module reports activity, independently of Internet surfing behavior. Additional data regarding conversion status, return conversion status, rank in a list that a user saw, etc. can be included in such a cookie or data record.

[0049] **FIG. 6** is an example of a cookie from that stores the most recent commercial keyword of interest. This may be the most recent commercial keyword of interest from a search or a keyword associated with a URL or click-through. More than one recent

commercial keyword of interest may be stored in one or more cookies, following the example of **FIG. 6.** One attribute or field of this cookie is the version of the crumb 602. Another is the time written 604, which may be expressed in seconds from 1970 to the time the cookie is written. A commercial flag 606 may be used to categorize the keyword type, for instance between key words that are of current commercial interest and key words that are for sale. The keyword may be included 608, either in text or as a shorthand code. A character set 610 such as UFT-8 may be included to assist with decoding of the keyword 608. This may be important in international settings.

[0050] **FIG. 7** is an example of summary data maintained by the behavior watcher 120 that may be written to persistent memory. In this example, a category ID 802 is a shorthand for a particular category. Bins are allocated for a current period 702, such as a day. The periods may be of equal length for different lengths. In one embodiment, a bit string implements the bins. A history of 60 days may be implemented by 60 bits plus storage for a category ID. Each bin corresponds to a particular time segment. A bit flag indicates whether there was activity related to the category during the particular time segment. At a regular time corresponding to the shortest period represented by any bin, the bit string is updated. Data is shifted into older bins, making room in the most recent bin. Data regarding recent activity is summarized into the most recent bin, labeled zero in the figure. If bins represent unequal amounts of time, special provisions are made at the boundaries between shorter and longer time bins, either algorithmically or by allocating additional storage. For instance, activity during any one-hour segment during a preceding day may result in setting the bit flag in a day-long time segment. The day-long time segment is flag "0" only if the hour segment flag at midnight is "0" when the midnight update takes place. This is an example of an algorithmic provision at the boundary between a one-hour segment and a day-long segment. Alternatively, storage may be allocated for bins to record number of visits in a particular time segment. This may significantly increase the amount of storage required, without necessarily changing how a reported-behavior aware server 132 responds. Compact storage is preferred when there are hundreds of categories and the practical limit for the size of a cookie is 2K to 4K bytes. Cookies can be sent to a server that archives user activity data. The user activity data can be used for analysis or targeting of ads. A bit string can be further summarized by category, shown in FIGS. 8-10.

CLAR 1006-3

[0051] **FIG. 9** illustrates buckets that may be applied to recency of using search words, clicking through, entering URL destinations or other indications of user interest. The most recent visit is recorded by membership in a particular bucket. One of 15 or 16 buckets can be identified with four bits. In the figure, the buckets are of unequal length. A single, most recent visit record such as illustrated by **FIG. 6** or an entry kept in memory can be used to assign a category to a particular bucket.

[0052] **FIG. 19** illustrates the importance of recent behavior in response to banner ads, measured by click through rates. A user who had behavior within the last hour in an advertising category was 30 times as likely to respond to a banner ad as one with no behavior over the last 7 days in the category. Behavior in the last hour, two hours or three hours has been observed to generate a huge increase in responsiveness to banner ads. Even behavior within the last seven days correlates strongly with a user's interest. Preferably, the selection of recency bucket sizes reflects the observed power of recent behavior illustrated by the figure.

[0053] **FIG. 10** illustrates buckets that can be applied to frequency of visits. When the bit string of **FIG. 7** is organized by days, activity in a category can be assigned to buckets of different visit frequencies, such as 0, 1, 2, 3, 4 or 5, 6 to 10, 11 to 30, or 31 to 60 days out of the last 60 days in which the user indicated interest. The number of bits set in the bit string is counted and summarized using this categorization.

[0054] With the categorical coding of recency and frequency in mind, we turn to **FIG. 8**. A single cookie may record data for many category ID's 802. For each category ID, three or more metrics of user interest may be recorded. One or more recency categories 804 can be recorded. The recording of recency can be applied to key words submitted to search engines or to other kinds of user behavior, described above. One or more frequency categories 806 can be recorded, applied to key words submitted or other kinds of user behavior. Banner clicks 808 can be separately categorized by recency, frequency or both. The comments in data column 810 are editorial comments to explain the data in the table and are not part of the data. Category ID 98409 experienced heavy and recent user interest. Category ID 65625 experienced heavy user interest last month. Category ID 68530 experienced light, recent user interest. Category 147374 experienced light activity last month, which probably also should be reflected by a nonzero value under frequency 806.

CLAR 1006-3

[0055] The behavior watcher module 120 preferably sorts the category history information in **FIG. 8** so that the most important categories are listed first. Alternatively, the system could leave the category sorting to the reported-behavior responsive server. As illustrated in **FIG. 11**, a category ID 1102 can be associated with both the category name 1104 and a revenue priority 1106. It may be useful to sort category history information by revenue priority 1106, especially if only a portion of the client-sided user behavior data is transmitted to a reported-behavior responsive server.

[0056] The statistics returned regarding web usage may or may not distinguish between commercial and non-commercial keyword searches. This distinction is made at the time of loading into the database where keywords are checked against the ever-changing list of commercial terms, for instance, terms bidded by Overture. If a keyword is determined to be commercial, then it is assigned a keyword id, which may be compatible with the Overture keyword id list. If the keyword is not deemed commercial, then another id is assigned in the data loading process.

[0057] These two types of data (commercial and non-commercial) are loaded into separate sections of a data warehouse. At the time of a new search engine loading run, keyword data is extracted from both the commercial and non-commercial tables. The results are joined and unique keyword ids are assigned from a master table. A keyword can sometimes be found in both the commercial and non-commercial contexts. For instance, if the commercial nature of a keyword is tied to whether it's being bidded upon, a keyword which is not commercial today may become a commercial tomorrow, if it receives a bid. This duality of keywords creates non-unique keyword ids as the same keyword may have both a keyword id created by a bidding process and a second one created through the load process. To eliminate this, it is preferred to create and maintain a single unique keyword id for every keyword. This list is updated when new keywords are discovered, and assigns existing ids to keywords already in the system.

[0058] Some preprocessing may be performed by the behavior watching module to simplify the URLs reported. URLs are unwrapped and cleaned in a separate process. URLs are often wrapped by search engines to enable the serving search engine to track clicks on served URLs. There are many different forms of URL wrapping. For example, a wrapped URL from yahoo.com might be: http://rds.yahoo.com/S=2766679/K=bmw/v=2/SID=w/l=

CLAR 1006-3

WS1/R=2/SS=100040736/H=1/SHE=0/*-http://www.bmwmotorcycles. com/. Unwrapping the URL produces http://www.bmwmotorcycles.com

[0059] From the server's perspective, the process begins with receipt (402) of behavioral information. The server uses whatever information it receives. From the user's perspective, the behavior watching module will report its observations and the user will receive search results ordered aggregating the user's information with others and/or will receive personalized to the user's recent behavior.

[0060] De-duping may also occur at the URL level. Information for two URLs which are identical is aggregated into one single URL. If two URLs differ even slightly however, (e.g., by a slash) then the two distinct versions are kept and another attempt at de-duplication is made as described below, for instance, using a combination of title and generated snippet.

[0061] Conversion data can be associated to a specific search by an algorithm that ties a search URL click to a specific conversion event, which occurred within a predetermined window. Usage and conversion data are matched for advertiser domains (URLs) that have clicks at the machine id, query time, advertiser domain level. For a particular machine with a click on a particular advertiser domain, if a conversion stat is observed within a predetermined window, then the conversion is attributed to that search click. If the conversion falls outside of the predetermined window, then the search click is not attributed.

[0062] For machine id-advertiser domain pairs that have a conversion stat attributed, subsequent future conversions are attributed as repeat conversions. These attributes also may be carried along and are available for use by a ranking algorithm (404). Metrics included with search data include number of visits, time spent (dwell time) and pages viewed.

[0063] Domain event data is joined to user data (with conversion metrics) at the machine id-advertiser domain level across sources, for combining search behavior for U.S. machines at google.com, msn.com, and yahoo.com. These results are put into time series order within machine id and advertiser domain. Domain events which occur within a predetermined time period following a search click are assigned as post-click metrics for that search click on that particular advertiser domain. If post-click metrics cannot be assigned to a particular search-click, the record is thrown out.

[0064] Several of the domain event data elements are subject to inaccuracies manifested in the client-sent stats. Both time spent and pages viewed are occasionally

misreported, and at other times, accurately reported, but in need of logically driven limits to be imposed.

[0065] Time spent can be misrepresented by machines having bad or inaccurate clocks. It can be accurately represented but in need of caps in such a case when a machine is left on a particular domain for an extended period of time. In order to cap outliers and to maintain a reasonable threshold for time spent on a site post click (403), a time limit of 30 minutes has been employed. A cap for pages viewed has also been implemented and set at 5. Other time and pages viewed caps may be substituted. These caps can be implemented on the behavior watching client side or after data is received at a host. In addition, user activity can be monitored by the behavior watching module so that extended periods of inactivity are not counted as dwell time.

[0066] When data is joined and aggregated at the machine-id, keyword, and URL level, the resulting data structure may include: machine-id, keyword, keyword_id, URL, URL_ID, domainid (corresponds to the domain of the URL), clicks, dwell_per_click, pages_per_click, conversions_per_click, rank_position (from search results list viewed by the user). Optionally, only keywords of predetermined interest (402) may be processed. For instance, keywords having commercial interest, such as bidded keywords, may be processed.

[0067] Outliers optionally may be removed (403), to avoid scoring anomalies. One example of an outlier is a link that is returned only once by a search engine and followed with enthusiasm by the user when was returned. A single strong sample point can give a link an unbeatable average score. A link that appeared just before a keyword was rescored could potentially be ranked in the top position for that keyword on the basis of a single sample! Accordingly, one example of optionally removing outliers is to not rank links unless they have been followed a predetermined number of times. For instance, if a link has not been selected by users and followed at least 10 or 100 times, it might remain unranked until its activity level reached the predetermined level or threshold.

Ranking Algorithm

[0068] A URL ranking algorithm (404) has been developed to identify and rank links for any given keyword. Many variations on combining the aggregated observations have

merit, as discussed below. One combination uses URL click rate and dwell time metrics (time spent at the domain and/or number of pages viewed), to select and rank URLs.

[0069] Optionally, user responses can be normalized for the position of a URL in a search result set. Position produces an inherent bias in URL click data for search results, which may be desirable or not. URLs occupying higher ranks garner higher clicks. In order to account for this bias, a normalization algorithm was developed to put clicks on links in disparate positions on equal footing.

[0070] Data is aggregated for each position and average click through rates, average time spent, and average pages viewed are calculated. For each rank position (1 – n), there are at least three average aggregate measures of that position's importance: click through rate, time spent and pages viewed. Normalization of these measures can be expressed as:

$$\text{Pages Viewed Inflation Factor for Position } n = \frac{\text{Avg Pages Viewed (Position 1)}}{\text{Avg Pages Viewed (Position } n\text{)}}$$

$$\text{Dwell Time Inflation Factor for Position } n = \frac{\text{Avg Dwell Time (Position 1)}}{\text{Avg Dwell Time (Position } n\text{)}}$$

$$\text{Pages Viewed Inflation Factor for Position } n = \frac{\text{Avg Pages Viewed (Position 1)}}{\text{Avg Pages Viewed (Position } n\text{)}}$$

This embodiment calculates and ranks top URLs for any keyword (404) based upon observed user metrics.

[0071] This embodiment may re-rank results based upon a time share metric, which corresponds to an individual machine's percentage vote. The algorithm takes into account user web surfing patterns and effectively places users on equal voting for relevant links. The premise is that a user has a certain amount of time which is spent on a site post a search click. These times are totaled to form the individual user's total time value which was spent viewing sites post search clicks. Percentages of the total time are then calculated for each URL click made by that particular user, resulting in a time fraction vote.

[0072] In using this methodology, users who in general spend less time surfing the Web have the same voting power as users who tend to spend longer amounts of time. This

evens the playing field across all categories. Alternatively, other embodiments may take into account other factors which make up an individual user's profile. Users identified as category experts may have a higher vote. For example, a user highly involved in the electronics category may have his vote count more for links clicked than a user who is new to the category. The voting blocks may take place within a particular category, and not across all categories as a whole.

[0073] The following detailed computational example applies to a specific keyword-URL pair through the aggregation and cleansing process.

[0074] Links in higher positions garner higher clicks and hence possess higher click through rates. It follows that users also tend to spend longer amounts of time and view more pages at URLs occupying these higher ranks. In order to account for this bias, a normalization process is applied.

[0075] Average clicks, average dwell time, and average pages viewed are calculated for each position regardless of keyword or URL combinations. These numbers are shown below as Position Averages. Average fraction clicks, dwell time and page views are also shown for this keyword, URL combination. These average fractions correspond to the percentage of each metric devoted by all machines to each keyword, URL, position grouping.

Kwd	URL	POS	Position Average Share of Clicks	Position Average Share of Dwell Time	Position Average Share of Page Views	Average Fraction Clicks	Average Fraction Dwell Time	Average Fraction Page Views
01	101	1	0.15	0.14	0.18	.25	.14	.08
01	101	3	0.13	0.12	0.15	.21	.17	.09
01	101	7	0.09	0.11	0.1	.32	.18	.11

[0076] Inflation factors are calculated for each position and applied to the appropriate observed metrics, normalizing them for position. The normalizing equation:

$$\text{Inflation Factor For Position } n = \frac{\text{Average (Position 1)}}{\text{Average (Position } n\text{)}}$$

Clicks, dwell time, and page views in lower positions are factored up by the appropriate inflation factor observed for that particular position. In the example below, the average fraction dwell time for position 7 pre-normalization was .32. After the application of the

CLAR 1006-3

inflation factor, the average fraction dwell time was .53. This number is now a normalized dwell time.

Kwd	URL	POS	Clicks Inflation Factor For Position *n*	Dwell Time Inflation Factor For Position *n*	Page Views Inflation Factor For Position *n*	Normalized Clicks	Normalized Dwell Time	Normal Page V
01	101	1	1.00	1.00	1.00	0.25	0.14	0.08
01	101	3	1.15	1.17	1.20	0.24	0.20	0.11
01	101	7	1.67	1.27	1.80	0.53	0.23	0.20

[0077] Weighted averages are calculated for each metric at each position, and totals are calculated across the positions for clicks, and all of the weighted average fraction measurements; clicks, time and page views.

[0078] Over time, with new reports of web usage from behavior watching modules, URLs for a particular keyword will adjust their positions. A URL in position 1 this week may be in position 7 the next. Adjusting for position is therefore a cleanup and adjustment process.

Kwd	URL	POS	Absolute Clicks	Weighted Average Fraction Clicks	Weighted Average Fraction Time	Weighted Average Fraction Page Views
01	101	1	155	38.75	21.70	12.40
01	101	3	124	30.05	24.59	13.39
01	101	7	98	52.27	22.45	19.40
01	101	ALL	377	121.06	68.74	45.20

[0079] The final result is the Weighted Average of Normalized Totals for Keyword 01, URL 101, at any position. Computed for every keyword, URL combination a single score is calculated for each metric needed for the v17.1 algorithm. These metrics are now normalized for position, and for multiple rank occurrences.

KWD	URL	POS	Absolute Clicks	Weighted Average Fraction Clicks	Weighted Average Fraction Time	Weighted Average Fraction Page Views
01	101	ANY	377	0.32	0.18	0.12

[0080] Ranks for scoring can be based on time that a user spends viewing pages on the domain. The information received from the behavior watching module may limit the maximum amount of time that will be assigned for any viewing session or it may track the

user's behavior, such as window navigation between programs, mouse clicks or mouse movement, and disregard periods of inactivity when calculating dwell time.

[0081] Ranking may ignore links that were selected by users less than a predetermined number of times, which may be predetermined as a fixed number or a function of traffic for the keyword or category. Ignoring outlier links (403) may avoid giving a high ranking to a link that was rarely presented by the search engines and followed only once or twice.

[0082] Ranks for scoring also can be based on a combination of click through rate, dwell time and the number of pages or documents viewed after following the link. Combining these factors, in some instances one of the factors will dominate: all or more than two thirds of the ranking weight may be assigned to just one of click through rate, dwell time or number of pages or documents viewed. Alternatively, they may be equally weighted, plus or minus 10%, or the factors may be assigned weighting ratios of approximately 2-1-1, plus or -10%, so that one factor is given approximately half of the combined weighting.

[0083] Another factor that can be used in ranking is return visits. If the user returns to the domain within a predetermined time after leaving it or within a predetermined number of navigation events, the user's return to the site can be assigned significance. Return to the site may reflect a favorable impression after considering other sites.

[0084] Conversion from browser to buyer or registered lead can considered to be particularly worthwhile as a factor. Again, conversion may include both a purchase in the domain and a registration. In some instances, such as car or home purchases, registration may be more realistic measure, because the purchase may be impractical or infrequently completed at a web site. Return conversion also may be taken into account.

[0085] Results may be segregated for analysis by search engine and ranks scored. Then, the separate rank scores may be combined into an overall ranking.

[0086] Statistical or other analysis can be applied within categories or keywords to determine which combination of ranking factors best attracts users to follow a link responsive to a search. It is anticipated that ranking information will be used differently among categories of keywords. Time spent will be important in the auto category. Conversions will be much more important music downloads category.

CLAR 1006-3

[0087] Optionally, click segmentation bands may be applied. These bands give precedence to URLs with high numbers of clicks. Employing these bands may improve the resultant links on selected algorithms.

[0088] Segmentation bands are identified based upon total clicks received by a particular URL. For instance:

Tier 1: URLs with 100+ clicks
Tier2: URLs with between 50 and 99 clicks
Tier3: URLs with between 10 and 49 clicks
Tier4: URLs with less than 10 clicks

[0089] URLs for a particular keyword are first put into the appropriate segmentation band. Once the band is identified, these URLs are set in descending order by rank score.

[0090] A predetermined number of links, such as the top 15 links (405), may be selected for data collection, to be followed by a spider engine (406).

[0091] Three tables are generated as output from the rank process:
Keyword table (KEYWORDID, KEYWORD)
URL table (DOMAINID, URL_ID, URL)
Rank Table (keyword, URL, rank)

Following Links

[0092] Traditional crawling programs at other search engines (ex: Slurp at Yahoo!, Googlebot at Google, MSNBot at MSN) crawl the entire web in search of relevant pages to index to be used in determining the rank order of links to display for a given keyword. The embodiment disclosed here, in contrast, is given a succinct number of URLs to crawl, which may optionally be selected (405) from links reported by the behavior watching module (401). These links are pre-ranked, hence this information retrieval process needs not determine the relative importance of a given URL from its connections to others, but rather to obtain the best possible descriptive information from the URL.

[0093] This embodiment takes a specific set of URLs and performs several specific tasks: It strips out all HTML tags and returns first 1,000k or another predetermined chunk of the text on the page to a file. It takes and stores a mapping from the text object's value into a

uniform scalar space to be used as a text signature or text fingerprint. It calculates an MD5 or other fingerprint of the document (with or without html tags). It calculates a summary count of the characters within the text extracted from the document.

[0094] This method may be implemented by a Java application which operates in a Linux environment as illustrated by **FIG. 18**. The application (1811) is a brokered one which allows for a large and scalable number of threads to run simultaneously across multiple Linux servers. In the current configuration, each Linux server (1821, 1822) can handle up to 10 spider instances (1831-33, 1834-36). Each spider in turn runs 200 worker threads (1841-43, 1844-46). At capacity, this configuration is able to spider 800 links a second.

[0095] Total threads working for a single broker can be arrived at by the following equation: With i number of spiders each having j number of workers (threads):

$$\text{Total threads for i \# of Spiders: } \sum_{i=1}^{n}\left(\sum_{j=1}^{n} wor\,\text{ker}_j\right)$$

[0096] The dual-broker model (1821, 1822) can segregate keywords by keyword velocity. General keywords are funneled through a robust, heavy duty version of the ranking algorithm. Fast moving keywords (e.g., news, current events) can be processed through a nimble, express version of the ranking algorithm, which uses less history. Keyword velocity is a measure of how quickly the popularity of a keyword changes. The highest velocity keywords can be selected by comparing the number of keyword searches in the last 24 hours (day 0) against the 24 hours before that (day-1). A different time span, such as four or eight hours, can be used, of course. How far the ration day 0/day-1 varies from "1" is the keyword velocity. If the ration is less than 1, the keyword is becoming less popular, "old news." If the ratio is much more than 1, the keyword may relate to a new story. Generally, a predetermined number of relatively high velocity keywords are re-indexed at a predetermined interval or as resources permit. In one embodiment, the top 10,000 keywords are re-indexed each day. While one metric of keyword velocity or volatility has been described, variations are anticipated.

[0097] The heavy duty version handles the ranked keyword URL pairs. These ranked keywords URL pairs are made available through an Oracle table on a database. The

CLAR 1006-3

URL_TABLE includes: DOMAINID, URLID, URL, LENGTH, SIGNATURE_H, SIGNATURE_T, SPIDER_DATE and HOST

[0098] The DOMAINID, URLID, and URL fields are populated from a reference database prior to following the links. After the link-following process for a specific URL, the LENGTH, SIGNATURE_H, SIGNATURE_T, SPIDERDATE, and HOST fields are written back to the database.

[0099] Brokers use Java Database Connectivity (JDBC) to connect in to the Oracle database. The broker accesses the URL_TABLE from the ranking process. The broker makes a request for 1/100th of the total number of domains which are available in the URL table for which SPIDER_DATE is null. All URLs associated with these domains are extracted by the broker where they are grouped by domain. Individual spider boxes talk to the Broker via Remote Method Invocation (RMI) requesting URLs for domains 1,000 domains at a time. Domains are then passed from the spider to a worker who takes all of the URLs associated with its domain and operates upon those URLs.

[00100] URLs are passed to the workers grouped by domain in order to accommodate generally accepted crawling or link following practices so as not to swamp domains with thousands of requests simultaneously. It is a generally accepted practice to not access a single domain with more than one request at a time. The link following process (406) respects this generally accepted principle by assigning each worker all URLs associated with a given domain.

[00101] The link following process (406) is a robust, scalable, application which fetches content and calculates statistics from a specific URL. Once a worker receives a domain and its associated URLs, it accesses that URL using HTTP protocols. If a good response code is received, a link following worker goes to work on that page. The worker receives a 200 response code (status OK) more than 98% of the time. If the page returns an HTTP code indicating a redirect (codes 301, 302, 303 and 307), further action must be taken by the worker or system in order to obtain information about that URL. A worker will follow up to 5 redirects from an initial URL before abandoning. Once the worker reaches an end point, the following tasks take place: Acquire HTTP return code from the URL. If a good response code is acquired: Identify title meta tag if available; calculate an MD5 fingerprint of

the entire document (both HTML and text); parse HTML from the page; and write back first 1,000k of text to disk.

[00102] Once the content is parsed and written back to the disk, a subsequent operator takes over. This operator makes several calculations used for the document fingerprint and writes those and other statistics back to the Oracle database. The system writes back the following fields to Oracle: URL_TABLE, DOMAINID, URLID, URL, LENGTH, SIGNATURE_H, SIGNATURE_T, SPIDER_DATE and HOST.

[00103] LENGTH is a count of characters in the text of the document (first 1,000k). This feature can be used for de-duping URLs later in the process (408). SIGNATURE_H is the MD5 hash code signature. SIGNATURE_T is a CRC32 checksum code of the text (first 1000k). SPIDER_DATE indicates the date and time that the particular URL was accessed. HOST pertains to which spider machine stored the text of the URL.

[00104] The following system may create three different measures designed to aid in document de-duplication (409). This de-duplication process aims at identifying documents that are identical or very similar within a given keyword result set. In a prior step not separately shown, URLs are de-duped at the URL level. Easily identified duplicates such as two occurrences of the exact same URL are eliminated. The system attempts to eliminate URLs that do not appear to point to the same page, but in fact do. In one embodiment, mathematical signatures (fingerprints) are taken for each URL and compared to other URLs within a given keyword result set. Three exemplary signatures are a length signature, an MD5 signature and a CRC32 checksum. Other signatures may be substituted.

[00105] For the length signature, the character length of the text document is calculated. This measure aids in the de-duping process to aid in giving context to a page which has been identified as a duplicate. For instance, if two sites show identical MD5 and CRC32 signatures, but have very disparate URLs, the signature is analyzed. If the length signature is low, meaning the page is small, it is likely that these two URLs share, for instance, a standard warning screen as would be found prior to entering an adult content site.

[00106] An MD5 signature typically is a way to verify data integrity. The MD5 algorithm takes as input a message of arbitrary length and produces as output a 128-bit "fingerprint" or "message digest" of the input. The MD5 algorithm is intended for digital signature applications, where a large file must be "compressed" in a secure manner. The

system computes an MD5 signature for the entire document, reducing the identity comparison process to a 128-bit comparison, for instance.

[00107] A CRC32 checksum generates the cyclic redundancy checksum polynomial of 32-bit lengths. This is usually used to validate the integrity of data being transmitted. The CRC is a "digital fingerprint" of a file, With CRC32 you can "melt down" a huge 20 MB (or even much bigger) file to have a small, handy reference to it, a single 32-bit number like 7d9c42fb (hexadecimal notation) which reflects the entire contents of this huge file. The system computes a CRC32 signature of the text of the document, giving insights into the text content of the page.

[00108] Another signature that can be calculated and used is Rabin's fingerprinting algorithm, for instance Broeder's implementation, which produces a compact checksum.

[00109] Any of the checksums or fingerprints can be applied to the whole document, the whole document less HTML tags stripped away, the selected chunk of the document that is cached, the title and snippets or some other predetermined excerpt from the document. More or less than 1,000k of the document can be used.

[00110] The process completes a run for a particular URL with data being written to an Oracle database and a spider box. The Oracle database receives fingerprint information (length, MD5, crc32), spider date/time, and host location information written to URL_TABLE and a spider date/time stamp written to KEYWORD table. The spider box receives files for data links that it followed: URL, title (if it was obtained during the initial fetch from the URL) and text of the document (first 1,000k) to be used for snippet generation. The text contains elements of the meta description and the body of the document

Snippet Generation

[00111] A snippet generation process generates titles and snippets for display (407). The snippet process takes a keyword phrase and URL combination, comes up with the best title describing that URL, and creates the best snippet (i.e., abstract, description) for that URL outlining in a 200 character space the information contained in the URL that pertains to the keyword. Snippet generation follows the link following process. Snippets are created from the text of the document retrieved from the chosen URL.

[00112] The keyword "somec bicycles" produces the following sample text for display:

CLAR 1006-3

Upland Sports Group: **Somec Bicycle** Frames
... is a mystique to Italian **bicycles**; an aura born of precision, innovative materials and artistry. Societa Mecchanica, better known as **Somec**, has contributed to ...
www.uplandsg.com/**somec**/ - 12k - Aug 17, 2004 - Cached - Similar pages

In this example, the title is "Upland Sports ... Frames". The snippet is the two lines following the title. The URL is on the bottom line.

[00113] Titles are usually generated from the title of the page retrieved when a link is followed. Most sites annotate the title of the page for search engines through the use of HTML meta tags. A tag identifying the title is present on over 97% of all URLs.

[00114] In the 3% of URLs for which the HTML tags do not supply a title, the process composes a title. If there is text available for the URL, the process takes the first approximately 70 characters of text (respecting word boundaries) and creates a title. If there was no text generated from the URL, the domain name is stripped from the URL (all information between www and .com) and displayed as the title.

[00115] Snippet generation is a mix of art and science. The process creates snippets leveraging mathematical equations and linguistic science. In one embodiment, snippets can be comprised of 1 single sub-snippet, or up to 3 sub-snippets separated by ellipses (...). A scoring algorithm decides which sub-snippets when combined (or not in the case of a single sub-snippet) produce the best score.

[00116] The snippet scoring algorithm is a multi-step process which scores various portions of the document's text. In four parts, it includes keyword tokenization, window scoring, window trading and final determination.

[00117] Keyword tokenization is applied because keywords are not always single words. Keywords are often multi-word phrases. The process tokenizes or single outs individual words within a phrase. Identifying individual word tokens typically includes searching for word separators such as spaces, periods, commas, or colon/semicolons. Once the tokenization of the keyword phrase is complete, the window scoring routine can commence.

[00118] In one version of window scoring, windows of three different sizes are calculated within the text of the document, for instance, for sub-snippet of lengths 200

CLAR 1006-3

characters, 100 characters and 66 characters. For various display formats and user preferences, a total snippet length of 50, 100, 200, 500 or 1,000 characters may work best. The example here was 200 characters. **FIG. 20** depicts how the window scoring routine begins with the first word of text (2003) within the document (2004). In the first case, the window spans 200 characters. This window receives a score based upon several metrics. Once the score is calculated, the window moves 1 character to the right and calculates another score for the next window. Alternatively, the window may move one word to the right, aligning either the window start or window end with word boundaries. This process iterates (2004, 2005) until a score for the final window is obtained. Intelligence may be built into the algorithm to eliminate the need to compute the score for each entire window, but rather add the additional information to the previous window's score and subtract the information which was lost in the move. Titles (2002) also may be captured, as described above.

[00119] When the process is complete, there may exist:

i windows of length 200 (where i = document length - 200)

j windows of length 100 (where j = document length - 100)

k windows of length 66 (where k = document length - 66)

[00120] Window scoring may be based on one or more metrics, such as the number of unique tokens found within the window, the total number of tokens found within the window, and/or the longest string of tokens found within the window. A formula for each window is computed from a combination of these metrics and assigned to that window.

[00121] In the case where there is one 200 character snippet, the window with the highest score is chosen. The two highest scoring windows of length 100 are chosen for the two sub-snippet model. The three highest scoring windows of length 66 are chosen for the three sub-snippet model.

[00122] The best scores are calculated for each model (1, 2, or 3 sub-snippets). A final algorithm may be applied when 2 or 3 windows are eligible for a snippet. If the global window score can be increased by one window giving up characters to another, then that action is seen as a gain and it is taken. If the global window score cannot be raised in this manner, the snippets are used without trading.

[00123] The output from snippet generation may include 5 different scores: Score of single sub-snippet model; score of non-traded two sub-snippet model; score of traded two

CLAR 1006-3

sub-snippet model; score of the non-traded three sub-snippet model; and/or score of the traded three sub-snippet model. Of these, the single highest score is chosen and that sub-snippet model is applied to that keyword, URL combination.

[00124] For a sample keyword = "red dog run", the following steps may be followed:

[00125] Step1: Tokenize keyword into three tokens:

Token1 = red
Token2 = dog
Token3 = run

[00126] Step2a: Locate instances of the tokens within the text document

-------red----------dog--------dog---------
-----run-------red----------red-----------
-------red-dog-run---------dog----------
---run----red-------run----run-----------

[00127] Step2b: Score the windows and identify the top ones. In this example, the three sub-snippet model, the best 3 windows were calculated.

-------red----------dog--------dog---------
[------run------red--]--------red----------- (1)
---[----red-dog-run--]-------dog----------- (2)
--run-----[-red-----run---]-run------------ (3)

[00128] Step3: Allow for trading to occur. In this case, if window 1 can give up some of the non-token containing characters within it's left edge to window 3. This allows window 3 to expand and include the final token 'run', increasing the overall global score of the snippet.

-------red----------dog--------dog---------
------[run------red--]--------red----------- (1)
---[----red-dog-run--]-------dog----------- (2)

CLAR 1006-3

---run-----[-red-----run----run]---------- (3)

[00129] In an alternative embodiment, the snippet generation process may involve the creation of an approximately 200 byte field used as a descriptor for the associated link. Snippet generation takes place post spidering and is created from the complete text of the document associated with the chosen URL or at least the portion of the document stored.

Personalization and Ranking

[00130] Within the ranking algorithm, there is the ability to select anonymous users who, based upon their behavioral profile, would have their votes for particular categories of links count more than other users.

[00131] Users who are heavy searchers (based upon their observed search behavior) would have their votes count more on links that they click more than the votes of novice searchers on that same link. In this way, the search experts would help produce more relevant ranking results.

[00132] Similarly, users who are highly involved in a particular category would have their votes count higher in that category than users who have no involvement in that category. Using behavior watching modules, one can identify users who are highly involved in various categories such as digital cameras, fantasy sports, or automobiles. For example, a user identified as being highly involved in the digital camera category would have his vote count more for links he clicked after a search for 'Cannon G3' than a user who is new to the category searching on that same keyword.

[00133] Identification of a user's category involvement status also drives personalization. A user with a high degree of involvement in a particular category would get different results from a user identified as less involved. This personalized results serving would require the presence of a cookie like object available on a particular machine. This lifestyle cookie would provide the search engine with a behavioral profile of the user, obtained from the users category navigational patterns. These category specific navigational patterns would be obtained from information contained in a categorization structure that also can be used for targeted advertising. For commercial purposes, a budget category or likely budget can be inferred from sites visited. Visitors to IKEA and Target are likely to have a

different budget for apparel than visitors to Sachs Fifth Avenue or Bloomingdale's. Similarly, Hyatt Hotels are in a different budget category than youth hostels.

[00134] Personalization based on observed communications is much more powerful than user-entered customization, because research shows that only 8-14 percent of users manually personalized their content. Personalization highly correlates with pages viewed at a domain: users who personalize have been reported to view 130 percent more pages at the domain than users who do not personalize.

[00135] Sometimes different behavioral profiles can be leveraged to make a difference in search results. Other times, differences between two users' behavioral profiles does not help in the context of a particular search keyword.

[00136] Some examples are helpful. First, an ambiguous search terms example: A highly involved auto category user who searches for "jaguar" would get more Jaguar auto related links than jaguar animal related links as compared to a normal mix of auto and animal related links for someone with no identifiable category involvement. Identification and usage of these behaviorally profiles could slant results, without completely replacing results. In the example above, the auto category involved user could get 100% auto results, or just a larger percentage of auto results than found among popular websites.

[00137] Next, a sub-category identification example: Three users search for the keyword "rental car". Three separate sets of results come up, each personalized for the users. Each user has a particular behavioral profile obtained from their past navigational patterns observed within the travel category. These behaviors are readily identifiable from the observed communications.

[00138] User 1: Frequent business traveler – his rental car results would be slanted toward the business traveler car rental results, possibly more about frequent rental points, etc.

[00139] User 2: Budget traveler – his rental car results would be slanted toward the budget traveler; rent-a-wreck type results, specials on sub-compact cars etc.

[00140] User 3: Luxury Traveler – his rental car results would be slanted toward the high-end luxury traveler; sports car rentals, classic car rentals, etc.

[00141] **FIG. 12** compares the insight that a behavior watching module on a user's machine gives regarding web surfing with the limited view that a server can achieve. These statistics were generated to compare what a Yahoo! server knows about user behavior,

compared to what the behavior watching module sees. The x-axis 1202 identifies categories of web sites. The y-axis is a relative scale, with Yahoo! server observation scaled to "1". In each stack, Yahoo! 1203 has a visibility of 1. The behavior watching module has a relative score 1204 that is higher and varies strongly between categories. This surprising degree of difference strongly distinguishes the on-machine behavior watching module from visibility that even a widely used search engine with many affiliated sites can achieve.

[00142] **FIG. 13** depicts a web page highly relevant to the keyword "cars". Applying a ranking algorithm, the CarsDirect home page scores 97 of 100. Factors used to calculate this score may include initial visit metrics, subsequent visit metrics and user level metrics. Observing users' responses to a particular site across search engines responsive to the keyword "cars", the behavior watching module observed a click through rate of 12 percent. An average user viewed five pages at this web site, following the initial link, before they left the site. The average user spent 55 seconds looking at the pages and the conversion rate, here a registration more likely than a purchase, was 80 percent. On subsequent occasions, within a predetermined time after the search or the first visit, an average returnee viewed 15 pages and spent 300 seconds on the site. Returnees registered with a conversion rate of 15 percent. Among users following this link for this keyword, category (e.g., car category) involvement was high, ranked "Expert". Search engine usage was "Maven", indicating that common users were more frequent than computer experts. For this category, among U.S. users, the geographic location of the web site was important. Web sites in the U.S. were more interesting to users (e.g., car buyers) than overseas web sites.

[00143] **FIG. 14** is a counter example of a page weakly relevant to "cars". The title of this article matches the keyword, but the click through rate and other metrics indicate that users considered this page much less relevant to searching for "cars" than the CarsDirect home page. The calculated relevancy score for this page is only 11.

[00144] **FIG. 15** juxtaposes the results for "cars" from Yahoo!, MSN Search and Google with an embodiment of the search engine described herein. Using a relevancy index, this embodiment produces more relevant search results. The relevancy index is the ranking score, which reflects users' responses to links in search result lists. Sample scores are: MSN 598, Google 524 and Yahoo! 487. This significantly improved score for this embodiment

strongly distinguishes the use of an on-machine behavior watching module and the aggregate behaviors of a multitude of users from the ranking algorithms of the leading search engines.

[00145] **FIG. 16** illustrates user differentiation using results from a behavior watching module. A large sample of user behavior, particularly shopping in the apparel category, is reported in this figure. Users were classified in nine boxes along attributes of amount of browsing in the category (1601) and degree of loyalty (1602) to a particular web domain (alternatively, to a vendor, brand or other category). Amount of browsing was classified as light, medium or heavy. One quarter of users were in each of the heavy and medium usage categories and the remaining half in the light usage category. Loyalty is classified as loyal, a switcher among sources or loyal to competitors. The 3-by-3 arrangement of attributes produces nine boxes. Among loyal shoppers, about one million unique observations were made. The heavy users had a higher click through rate when a particular result was displayed in response to a search, the rate being 7.90%, compared to 5.40% for medium users and just 3.00% for light users. After a click-through event, heavy users converted to purchase or registration at a rate of 8.00%, compared to 7.20% for medium users and 6.50% for light users. Not surprisingly, there are more switchers and competitive loyal unique observations than for the particular domain observed. Personalization of content, such as ads, may depend on which of the nine boxes the user fits. For instance, a loyal, heavy user will get a different message (a la frequent flier programs) than a medium to light switcher. A competitor-loyal heavy user might get a trial offer that would be completely misdirected to a loyal user. The strategy for advertising a heavy, loyal user is to insulate them from competitors. For medium to light loyal users, the strategy may target up-selling and cross-selling. For heavy to medium switchers, the objective is to increase share. Heavy to medium competitive loyal users are targeted for acquisition. Light switchers and light competitive loyal users may be ignored, to focus advertising resources.

[00146] Return visit data from the behavior watching module can assist an advertiser in measuring the effectiveness of a particular ad. User differentiation by box can further be associated with selection of ads and evaluation of ad effectiveness.

[00147] Cross-browsing of users also can be reported. Users can be selected by follow-through, for instance all click-throughs or all users with conversions. The users with a

conversion at a particular domain (or vendor or brand, for instance) can be rated by the frequency of their visits to competitors' domains (or vendors or brands).

[00148] **FIG. 17** depicts traffic to a particular domain, from a particular search engine, by keyword. This information may be useful to a company devising ads and also when bidding on keywords for sponsored content positioning on search engine result pages. Data from the behavior watching module can yield impressions, click, click-through rate, conversion rate and position on the search engine's results page.

Some Particular Embodiments

[00149] The present invention may be practiced as a method or device adapted to practice the method. The same method can be viewed from the perspective of a user at their terminal or personal computer or on the server side, collecting information from users. The invention may be an article of manufacture such as media impressed with logic to carry out computer-assisted method.

[00150] A device embodiment, from the user perspective, may be embodied in a module running on the user's computer and collecting behavioral observations, coupled to a server that responds to the behavioral observations with information personalized to the user.

[00151] While the present invention is disclosed by reference to the preferred embodiments and examples detailed above, it is understood that these examples are intended in an illustrative rather than in a limiting sense. Computer-assisted processing is implicated in the described embodiments. Accordingly, the present invention may be embodied in methods aggregating of communication patterns, pre-processing links responsive to keyword searches, responding to keyword searches using aggregated communication patterns to rank the responsive links, and responding to keyword searches using recent and current navigation information systems to resolve ambiguities and/or personalize responses based on user characteristics. Other embodiments, as devices, include logic and resources to carry out thes methods. As systems, still other embodiments include behavior watching modules on terminals, servers that process or respond to the behavioral data, or both. Other embodiments include media impressed with logic to carry out the methods, data streams impressed with logic to carry out the methods, or computer-accessible services that carry out the methods. It is contemplated that modifications and combinations will readily occur to those skilled in the

art, which modifications and combinations will be within the spirit of the invention and the scope of the following claims.

[00152] The embodiments disclosed may be practiced as a method or device adapted to practice the method. They may be practiced as a system. The same method can be viewed from the perspective of a server collecting data or a behavior watching module generating data. The invention may include an article of manufacture such as machine readable memory including instructions to carry out the method or a data stream including instructions to carry out the method.

[00153] One embodiment is a method of selectively collecting web pages that may be returned to users in response to search requests. This method includes receiving from behavior watching modules operating on a multiplicity of users' computers, information regarding the users' search engine usage across one or more search engines. This information includes at least keywords (including phrases) submitted by particular users to the search engines, links selected by the particular users from results returned by the search engines, and at least one of dwell time or documents viewed by the particular users when following the selected links. The method further includes using the search engine usage information to choose keywords of interest and to choose links selected by the users corresponding to the keywords of interest.

[00154] The multiplicity of users in this embodiment may exceed 2000 or 100,000 users, depending on the desired coverage of frequently submitted search terms. The information received from behavior watching modules may span a plurality of unaffiliated search engines, providing data that no single search engine could collect.

[00155] This embodiment further may include ranking, for the chosen keywords, the corresponding links based on user access rates calculated from the search engine usage information and at least one of the dwell time or the documents viewed. The method further may include limiting the collection of content based on a predetermined application of the ranking, such as following only the top 10 or top 15 links for a particular keyword.

[00156] In some embodiments, the search engine usage information is known to have been qualified before receipt to discount dwell time during periods when a particular user is likely to have been inattentive.

CLAR 1006-3

[00157] This embodiment further may include selecting snippets from the content of a document reached by following one of the chosen links. This snippet selecting may include applying a window of predetermined length to the document, repeatedly shifting the window through the document. The keyword of interest, if it is a phrase, is divided into words. The method further includes repeatedly calculating one or more window scores for the words in the window, including a count of instances of the words in the window and a measure of adjacency of the words in the window. With the window scores, a plurality of non-overlapping window positions are chosen as snippets. The snippets or references to the snippets are stored in a machine readable memory. This method may be refined in many ways. The repeated shifts of the window may align the window so that its start or end matches a word boundary. As the window is repeatedly shifted, the score within the window may be calculated by reducing the score for characters removed from the window and increasing the score for characters added to the window.

[00158] Snippet generation may be refined by trading off lengths of the snippets, approximately maintaining a combined length of all the snippets. As the lengths of the snippets are traded off, the one or more window scores are recalculated. The method further includes choosing a particular trade-off of snippet links using the recalculated window scores, optionally merging the snippets, and storing the chosen snippets or references to the chosen snippets in a machine readable memory.

[00159] Viewed from the perspective of a network of behavior watching modules, the method includes using a multiplicity of behavior watching modules operating on a multiplicity of users' computers, the behavior watching modules collecting information regarding the users' search engine usage across one or more search engines. The information collected includes at least keywords (including phrases) submitted by particular users to the search engines, links selected by the particular users from results returned by the search engines, and at least one of dwell time or documents viewed by the particular users when following the selected links. The method further includes the behavior watching modules electronically reporting the search engine usage information to one or more affiliated servers. As above, the multiplicity of users in this embodiment may exceed 2000 or 100,000 users, depending on the desired coverage of frequently submitted search terms. The information

received from behavior watching modules may span a plurality of unaffiliated search engines, providing data that no single search engine could collect.

[00160] The behavior watching modules may qualify the search engine usage information before reporting, to discount dwell time during periods when a particular user is likely to have been inattentive.

[00161] The behavior watching module embodiment may further include a particular user invoking a search engine with a keyword and receiving results from the search engine organized to reflect the reported search engine information.

[00162] A system embodiment includes at least one listener module and at least one computer or cluster of computers in communication with the listener module. The listener module includes a network interface that receives from behavior watching modules operating on a multiplicity of users' computers connected by network, information regarding the users' search engine usage across one or more search engines. The information includes at least keywords (including phrases) submitted by particular users to the search engines, links selected by the particular users from results returned by the search engines, and at least one of dwell time or documents viewed by the particular users when following the selected links. The computer or cluster of computers in communication with the listener module operate one or more modules that include logic and resources adapted to process the search engine usage information and choose keywords of interest and the links selected by the users corresponding to the keywords of interest, to follow the chosen links to collect at least part of the content of documents addressed by the chosen links, and to associate the collected content with the corresponding keyword of interest.

[00163] The multiplicity of users in this embodiment may exceed 2000 or 100,000 users, depending on the desired coverage of frequently submitted search terms. The information received from behavior watching modules may span a plurality of unaffiliated search engines, providing data that no single search engine could collect.

[00164] Another embodiment is a method of selecting snippets from content of a document that may be returned to users in response to search requests. The snippets selecting includes applying a window of predetermined length to the document, repeatedly shifting the window through the document. If the keyword of interest is a phrase, is divided into words. The method further includes repeatedly calculating one or more window scores for the words

CLAR 1006-3

in the window, including a count of instances of the words in the window and a measure of adjacency of the words in the window. With the window scores, a plurality of non-overlapping window positions are chosen as snippets. The snippets or references to the snippets are stored in a machine readable memory. This method may be refined in many ways. The repeated shifts of the window may align the window so that its start or end matches a word boundary. As the window is repeatedly shifted, the score within the window may be calculated by reducing the score for characters removed from the window and increasing the score for characters added to the window.

[00165] Snippet generation may be refined by trading off lengths of the snippets, approximately maintaining a combined length of all the snippets. As the lengths of the snippets are traded off, the one or more window scores are recalculated. The method further includes choosing a particular trade-off of snippet links using the recalculated window scores, optionally merging the snippets, and storing the chosen snippets or references to the chosen snippets in a machine readable memory.

[00166] Snippet generation may be combined with duplicate elimination. The method may further include evaluating to more documents based on their titles and the calculated snippets. If the titles and snippets or fingerprints or checksums calculated from the titles and snippets match or are very similar, the documents may be declared and handled his duplicates.

[00167] In snippet generation, the count of instances of the words in the window may include both counting how many of the words in the keyword appear in the window and how many times the words appear in the window.

[00168] A further embodiment is a method of ranking including receiving, from behavior watching modules operating on a multiplicity of users computers, information regarding the users search engine usage across the plurality search engines. Information received includes at least keywords (including phrases) submitted by particular users to search engines, links selected by the particular users from results returned by the search engines, and at least one of dwell time or documents viewed by the particular users when following the selected links. The method further includes ranking links corresponding to a particular carry keyword, using, from the search engine usage information, both (user access rates) and at least one of ((the 12th time) or (the documents viewed)). Then, the ranks are stored in machine readable memory, available in response to search in vote with a particular keyword.

CLAR 1006-3

[00169] Various weightings of user access rates, dwell time and number of pages viewed can be applied. One aspect of the data to which weightings are applied may be that the dwell time is known to have been qualified before receipt to discount periods when a particular user is likely to have been inattentive. Another aspect is that the data may be filtered to remove the outlying observations of search engine usage, eliminating links selected by particular users less than a predetermined number of times. Similarly, dwell times that exceed a predetermined length of time may be discounted after data is received.

[00170] Taking into account user access rates, dwell time and number of pages viewed, one variation has one of the factors dominating the other two. The dominant factor may be the click-through right, the dwell time or the number of pages viewed at the domain. Another variation gives balanced waiting to the factors, approximately equal plus or -10%. Yet another variation assigns waiting ratios of approximately 2-1-1, plus or -10%, so that one of the three factors is given approximately half the combined waiting.

[00171] Two additional factors that may be considered in ranking our return visits within a predetermined time the link was selected and conversion events. As described above conversion events may be purchases or registrations. Conversions and repeat conversions may separately be given weight.

[00172] For ranking, the combined data from a single user may be given a weight of one or all weight reflecting the amount of time spent searching or the number of links followed. But a single user is given a normalized weight, result biasing by behavior is mitigated by reducing the bots behavior to a single vote.

[00173] The rankings in step of this embodiment may further include groupings like the links into bands and ranking the select links within the bands, as described above. The ranking step also might include separately will ranking links selected by users of particular search engines and then combining rankings from the particular search engines into an overall ranking.

[00174] A further embodiment is a method resolving ambiguity in a keyword search terms submitted by user to a search engine. This method includes receiving, from a behavior watching module operating on the user's computer, information regarding the user's recent navigation behavior. The navigation behavior includes accessing documents that are not all associated with a particular server-side behavior data collection network and further includes

at least one of a mouse click-through, enter key stroke or other selection action and at least one keyword derived from context of the selection action. The server determines a plurality of interest categories to which the search term plausibly belongs and selects among the plurality of interest categories using the recent navigation information.

[00175] One aspect of the behavior watching modules operation is that it may categorize behaviors by interest categories and recency. The documents accessed may be web pages and the server-side behavior data collection network may be an advertising network.

[00176] From this perspective of the behavior watching module, this embodiment may be practiced as a method of supplementing a search request with information that reflects a user's recent navigation behavior among documents accessible via a network. This method includes operating a behavior watching module on the users computer, collecting information regarding the users recent navigation behavior. The recent navigation behavior information includes at least identifying documents electronically accessed by the user; categorizing the documents accessed; and indicating when documents in the categories were accessed. The behavior watching module submits at least a summary of the recent navigation behavior information along with a query to one or more servers including a search engine. The user receives from the search engine a set of results by asked to an interest category corresponding to the submitted recent navigation behavior information. Alternatively, the user may receive from the search engine a set of results by us to a budget category corresponding to the submitted recent navigation behavior information. More generally, any category that can be discerned from the recent navigation behavior can be reflected in a bias of search engine results received in response to the query.

[00177] Another embodiment is a method of personalizing content in response to a keyword search term submitted by a user to a search engine. As with the disambiguation method above, this method includes receiving, from a behavior watching module operating on the user's computer, information regarding the user's recent navigation behavior. The navigation behavior includes accessing documents that are not all associated with a particular server-side behavior data collection network and further includes at least one of a mouse click-through, enter key stroke or other selection action and at least one keyword derived from context of the selection action. The server determines a plurality of likely budget categories

to which results belong and selects among the likely budget categories using the recent navigation information.

[00178] For embodiments using recent navigation behavior, the navigation behavior may reflect recent navigations, including navigations with in the last hour and the last day. Alternatively, it may include recent navigations within periods of time determined to correlate to user responsiveness to information. The user navigation information may be summarized into interest categories were likely budget categories when it is received or after it is received. More generally, any category that can be discerned from the recent navigation behavior can be reflected in a bias of search engine results received in response to the query.

[00179] The embodiments and various aspects of the embodiments described above may be practiced as a machine readable memory including instructions to carry out the methods and aspects of methods described or a data stream including the machine-readable instructions. Further, a device may include one or more servers, personal computers or other computer devices having logic and resources adapted to practice the methods and aspects of methods described.

[00180] We claim as follows:

CLAIMS

1. A method of selectively collecting web pages that may be returned to users in response to search requests, the method including:

receiving, from behavior watching modules operating on a multiplicity of users' computers, information regarding the users' search engine usage across one or more search engines, including at least

keywords submitted by particular users to the search engines, wherein the keywords include phrases,

links selected by the particular users from results returned by the search engines, and

at least one of dwell time or documents viewed by the particular users when following the selected links;

using the search engine usage information, choosing keywords of interest and the links selected by the users corresponding to the keywords of interest;

following the chosen links to collect at least part of the content of documents addressed by the chosen links; and

associating the collected content with the corresponding keyword of interest.

2. The method of claim 1, wherein the multiplicity of users exceeds 2,000 users and the received search engine usage information spans a plurality of unaffiliated search engines.

3. The method of claim 2, further including: ranking, for the chosen keywords, the corresponding links based on user access rates calculated from the search engine usage information and at least one of the dwell time or the documents viewed.

4. The method of claim 3, further including limiting the collection of content based on a predetermined application of the ranking.

5. The method of claim 1, wherein the multiplicity of users exceeds 20,000 users.

6. The method of claim 1, wherein the search engine usage information is known to have been qualified before receipt to discount dwell time during periods when a particular user is likely to have been inattentive.

7. The method of claim 1, further including selecting snippets from the content of a document reached by following one of the chosen links, the snippet selecting method including:

applying a window of predetermined length to the document, repeatedly shifting the window through the document;

dividing of the keyword of interest (if it is a phrase) into words;

repeatedly calculating one or more window scores for the words in the window, including

a count of instances of the words in the window, and

a measure of adjacency of the words in the window;

choosing as snippets a plurality of non-overlapping window positions using the window scores; and

storing the snippets or references to the snippets in a machine readable memory.

8. The method of claim 7, wherein the repeated shifts of the window align the window start or the window end with a word boundary.

9. The method of claim 7, further including:

trading off lengths of the snippets, approximately maintaining a combined length of all snippets;

recalculating the one or more window scores to the snippets of changed length;

choosing a particular trade off of snippet lengths using the recalculated window scores, optionally merging the snippets; and

storing the chosen snippet(s) or references to the chosen snippet(s) in a machine readable memory.

10. A machine readable memory including program instructions adapted to:

receiving, from behavior watching modules operating on a multiplicity of users' computers, information regarding the users' computers, information regarding the user' search engine usage across one or more search engines, including at least

keywords submitted by particular users to the search engines, wherein the keywords include phrases,

links selected by the particular users from results returned by the search engines, and

at least one of dwell time or documents viewed by the particular users when following the selected links;

using the search engine usage information, choosing keywords of interest and the links selected by the users corresponding to the keywords of interest;

following the chosen links to collect at least part of the content of documents addressed by the chosen links; and

associating the collected content with the corresponding keyword of interest.

11. A device including one or more servers having logic and resources adapted to:

receiving, from behavior watching modules operating on a multiplicity of users' computers, information regarding the users' computers, information regarding the user' search engine usage across one or more search engines, including at least

keywords submitted by particular users to the search engines, wherein the keywords include phrases,

links selected by the particular users from results returned by the search engines, and

at least one of dwell time or documents viewed by the particular users when following the selected links;

using the search engine usage information, choosing keywords of interest and the links selected by the users corresponding to the keywords of interest;

following the chosen links to collect at least part of the content of documents addressed by the chosen links; and

associating the collected content with the corresponding keyword of interest.

12. A method of generating data using a network of behavior watching modules that transmit data to an affiliated system of servers, the method including:

using a multiplicity of behavior watching modules operating on a multiplicity of users' computers, the behavior watching modules collecting information regarding the users' search engine usage across one or more search engines, including at least

keywords submitted by particular users to the search engines, wherein the keywords includes phrases,

links selected by the particular users from results returned by the search engines, and

at least one of dwell time or documents viewed by the particular users when following the selected links; and

the behavior watching modules electronically reporting the search engine usage information to one or more affiliated servers.

13. The method of claim 12, wherein the multiplicity of behavior watching modules exceeds 2,000 users and the behavior watching modules collect search engine usage information across a plurality of unaffiliated search engines.

14. The method of claim 12, wherein the multiplicity of behavior watching modules exceeds 20,000 users.

15. The method of claim 12, wherein the behavior watching modules qualify the search engine usage information before reporting to discount dwell time during periods when a particular user is likely to have been inattentive.

16. The method of claim 12, further including a particular user invoking a search engine with a keyword and receiving results from the search engine organized to reflect the reported search engine information.

17. A device including a personal computer having logic and resources adapted to

receiving, from behavior watching modules operating on a multiplicity of users' computers, information regarding the users' computers, information regarding the user' search engine usage across one or more search engines, including at least

keywords submitted by particular users to the search engines, wherein the keywords includes phrases,

links selected by the particular users from results returned by the search engines, and

at least one of dwell time or documents viewed by the particular users when following the selected links;

using the search engine usage information, choosing keywords of interest and the links selected by the users corresponding to the keywords of interest;

following the chosen links to collect at least part of the content of documents addressed by the chosen links; and

associating the collected content with the corresponding keyword of interest.

18. A computer-implemented system that selectively collects web pages that may be returned to users in response to search requests, the system including:

at least one listener module with a network interface that receives from behavior watching modules operating on a multiplicity of users' computers connected by a

network, information regarding the users' search engine usage across one or more search engines, including at least

keywords submitted by particular users to the search engines, wherein the keywords include phrases,

links selected by the particular users from results returned by the search engines, and

at least one of dwell time or documents viewed by the particular users when following the selected links; and

at least one computer or cluster of computers operating one or more modules in communication with the listener module that includes logic and resources adapted to process the search engine usage information and choose keywords of interest and the links selected by the users corresponding to the keywords of interest,

follow the chosen links to collect at least part of the content of documents addressed by the chosen links, and

associate the collected content with the corresponding keyword of interest.

19. The system of claim 18, further including at least 2,000 behavior watching modules operating on the users' computers, in communication with the listener module.