US 20070266306A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2007/0266306 A1**

Koppel et al. (43) **Pub. Date: Nov. 15, 2007**

(54) **SITE FINDING**

(75) Inventors: **Moshe Koppel**, Efrat (IL); **Eyal Lanxner**, Jerusalem (IL)

Correspondence Address:
Martin D. Moynihan
PRTSI, Inc.
P.O. Box 16446
Arlington, VA 22215 (US)

(73) Assignee: **Egocentricity Ltd.**, Efrat (IL)

(21) Appl. No.: **11/878,254**

(22) Filed: **Jul. 23, 2007**

Related U.S. Application Data

(63) Continuation of application No. 09/605,987, filed on Jun. 29, 2000, now Pat. No. 7,257,766.

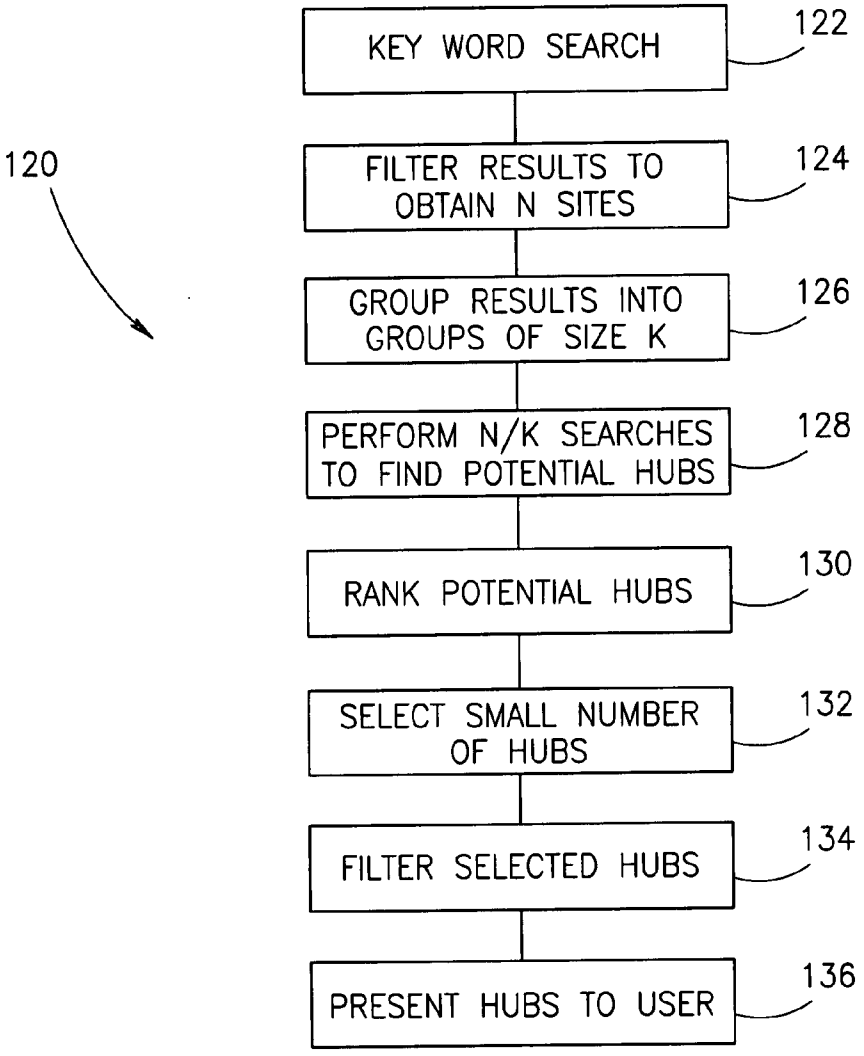
Publication Classification

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

(52) **U.S. Cl.** **715/501.1**

(57) **ABSTRACT**

A method of finding WWW pages, each of which includes at least one list of links to desired Internet resources, comprising: providing a list of URLs; automatically generating at least one query for an Internet search tool for WWW pages that include links to at least one URL of said list of URLs; executing said at least one generated query to provide search results that include at least one of said searched for WWW pages; and generating a response comprising at least one indication of one of said WWW pages, responsive to said search results.



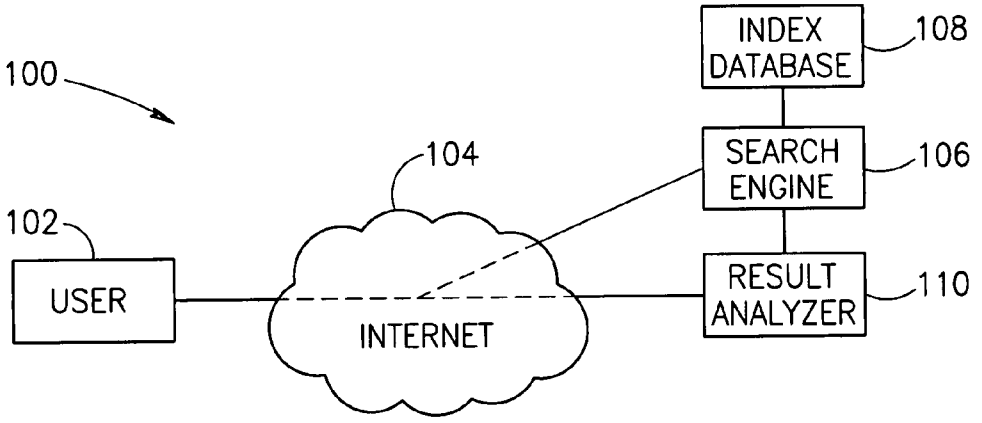

100
INDEX DATABASE
108
SEARCH ENGINE
106
104
102
USER
INTERNET
RESULT ANALYZER
110

FIG.1

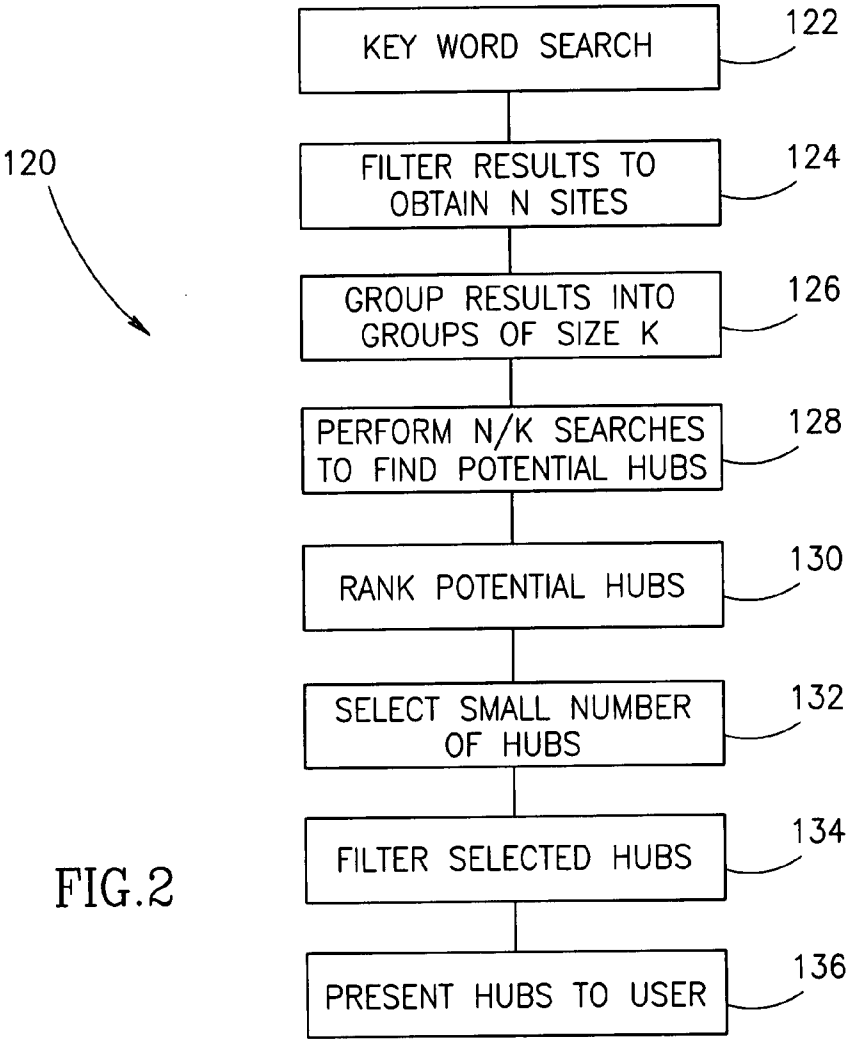

120
KEY WORD SEARCH
122
FILTER RESULTS TO OBTAIN N SITES
124
GROUP RESULTS INTO GROUPS OF SIZE K
126
PERFORM N/K SEARCHES TO FIND POTENTIAL HUBS
128
RANK POTENTIAL HUBS
130
SELECT SMALL NUMBER OF HUBS
132
FILTER SELECTED HUBS
134
PRESENT HUBS TO USER
136

FIG.2

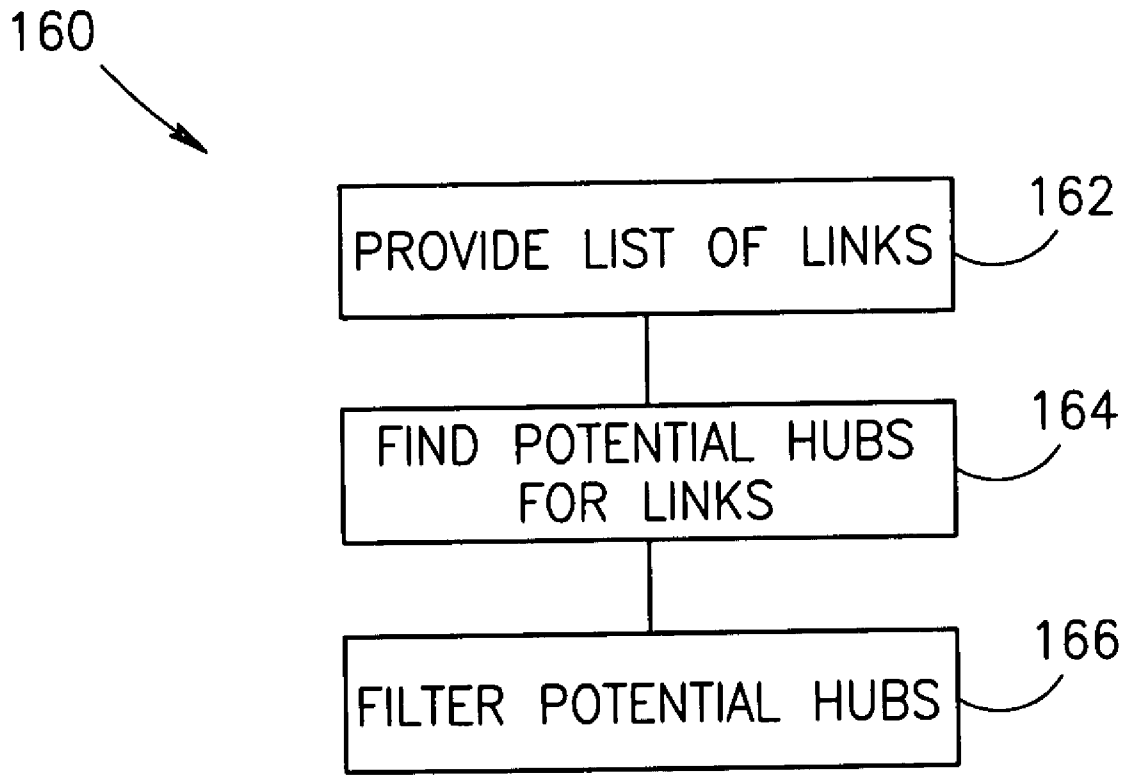

160
PROVIDE LIST OF LINKS
162
FIND POTENTIAL HUBS FOR LINKS
164
FILTER POTENTIAL HUBS
166

FIG.3

SITE FINDING

RELATED APPLICATIONS

[0001] This application is a continuation of U.S. patent application Ser. No. 09/605,987 filed on Jun. 29, 2000, the disclosure of which is incorporated herein by reference.

FIELD OF THE INVENTION

[0002] The present invention relates to searching for information on a data network, and especially to searching utilizing an analysis of the results of search engines.

BACKGROUND OF THE INVENTION

[0003] It is known in the art to analyze data networks, such as journals and journal citations, to determine meta knowledge about the field.

[0004] IBM Inc., described a method of determining hubs and authorities on the Internet, in U.S. Pat. No. 5,884,305, in a U.S. patent application Ser. No. 08/813,749 filed Mar. 7, 1997, mentioned in the patent and in "Authoritative Sources in a Hyperlinked Environment", by Jon M. Kleinberg, in IBM research report RJ10076(91892), topic area "Computer Science", May 29, 1997, the disclosures of which are incorporated herein by reference. Hubs are Internet sites that contain links to many other sites in a same field and authorities are sites that are pointed to by a significant number of relevant sites in a field. An iterative process was suggested to determine, from among a predetermined set of sites, a kernel of sites that match a hub or authority definition. In the Kleinberg paper, it is noted that the Internet is to be considered a different type of data network than journal articles.

[0005] A paper entitled "Mining the Web's Link Structure", by S. Chakrabarti et al, in IEEE Computer, Aug. 1999, the disclosure of which is incorporated herein by reference, describes analyzing link structures of WWW pages to determine hubs and authorities. At a site "http://www.google-.com", available on Feb. 1, 2000 and for some time before, a tool "googlescout" is suggested for detecting WWW sites that are similar to a shown site, for example for finding competition.

[0006] A WWW page "www.cgl.uwaterloo.ca/Project/Vanish/webquery_1.html", apparently available at least from Dec. 11, 1996, the disclosure of which is incorporated herein by reference, describes the "webquery" project, in which a quality of a site that turns up in a search is evaluated based on the number of sites linked to the site and the number of sites links in the site.

SUMMARY OF THE INVENTION

[0007] An object of some embodiments of the invention is finding one or more hub sites or lists of WWW pages that cover a topic presented by a set of input sites. In an embodiment of the invention, the hubs or page lists are selected by virtue of their including links to a significant number of the sites in the set of the input sites. An expected advantage of using hubs is that each hub may concentrate in it a large number of links to relevant sites, beyond those provided in the input set, and also include additional information which can help a human user select certain sites for browsing.

[0008] An aspect of some embodiments of the invention relates to selecting a potential hub based on a statistical analysis of an Internet link structure, for example, using an approximation of a number of links from the potential hub to a set of input sites, rather than determining which sites from the input set are actually pointed to. In one embodiment of the invention, this determination is made by searching for potential hubs that include links to groups of input sites and then ranking the resulting potential hubs, based on the number of groups pointed to by each potential hub. As a potential hub might include links to more than one site in an input group, the approximation may be significantly different from the actual number of links between a potential hubs and individual member sites of input groups. It is noted that-in some embodiments, there is no final determination of which particular site is pointed to by the potential hub.

[0009] An aspect of some embodiments of the invention relates to a method of automatically determining a hub-potential of a site, for example for ranking hubs in a set of potential hubs or for finding potential hubs in a search. In one embodiment of the invention, a hub potential of a site is determined based on structural properties of the site, for example, the existence of a list of links and/or the existence of a text paragraph (e.g., a review or description) of many of the links. Optionally, the number of links is determined by counting the occurrence of the phrases which indicate the presence of links, such as "http:" or "href". Alternatively or additionally, a hub-potential may be determined based on the usage of key terms of the topic in the site in general and/or in anchor portions of the site in particular, such as a main title or a section heading. Alternatively or additionally, a hub-potential may be determined based on a usage of hub-typical words or phrases, such as "list of links", "links", "index", "list", "compilation"and/or "resources". Optionally, these words or phrases receive a higher scoring based on their location in the site, for example in a title or before a long list of links.

[0010] In one embodiment of the invention, the potential hubs are ranked and/or filtered before being analyzed in greater depth. Alternatively or additionally, the hub generation process may create a small set of potential hubs to begin with, for example using a threshold setting. Such ranking may include, for example, selecting only a subset of those sites that point to the input set of sites, for example based on the existence of a topic word in those sites, prior to analyzing the sites for hub-potential. In another example, potential hubs that are found using a search engine are required to both include a topic word and at least one link to one group of sites from the input site.

[0011] In an embodiment of the invention, hub potential is characterized by rules, which may be phrased in a search engine command language, so a search for the hubs using the search engine returns sites with a higher potential of being desired hubs. In an embodiment of the invention, the particular features of a search engine, for example, searching for URLs or links, disjunctive search and/or pipes, are used to perform one or more of the above activities, for example, group comparison, rule application and/or thresholding of potential hubs, more efficiently.

[0012] In one embodiment of the invention, an input set of sites is generated by a user providing a topic or topic words and generating, for example by one or more search engine(s)

and/or Internet indexes, a list of sites relevant to that topic. Optionally, the list of sites is filtered prior to being used as a basis for finding hubs, for example by removing redundant and/or mirroring sites.

[0013] Alternatively or additionally, an input set of sites is generated from a user provided site. The user provided site can be analyzed to find a second set of sites that is similar to the provided site. One exemplary method of determining similarity is by finding hubs as defined above which point to the site and selecting links from those hubs as similar sites. Another exemplary method is to receive a short list of examples for such similar sites. Another exemplary method is finding sites that contain similar text to the provided site. Optionally, the user provides a set of sites, rather than a single site.

[0014] Optionally, hubs that point to the similar sites and not to the provided sites are determined. In some embodiments, these hubs are treated as hubs to which a link to the provided site should be added, for example by suggestion to the hub operators.

[0015] Alternatively or additionally, an input set of sites is generated by analyzing a user provided hub or a hub obtained from previous use of hub-finder or a hub constructed by combining search results/analysis of existing hubs or other user provided information.

[0016] Alternatively or additionally to providing a hub as an input, a list of a user's favorite bookmarks or recently or frequently traveled sites may be used as an input instead. Such lists may be considered to comprise a profile of a user, for example for advertisement targeting or for finding friends or partners. Such a user profiling tool can be used, in some embodiments of the invention, to extrapolate from an existing, studied group of users to a large group which is not studied in detail but whose browsing habits are known.

[0017] A set of sites may be filtered, manually or automatically, prior to being used as an input set, for example, a user manually selecting a subset of links or a topic word for use in analyzing the suitability of the links.

[0018] Optionally, the resulting hubs are considered a set of hubs which are similar to the input hub or at least an aspect of the input hub, and may thus be presented to a user

[0019] In one embodiment of the invention, a set of similar hubs is analyzed, to harvest information which may be useful, for example to the owner of the provided hub. In one example, the links of the similar hubs are collated, filtered and/or ranked, to detect links or textual descriptive material of links that are missing from the input hub and might be desirable. In another example, links that exist in the provided hub are ranked based on the particulars of the appearance of such links in the similar hubs. In another example, a new hub is created, possibly ad-hoc, based on the analyzed similar hubs.

[0020] The similar hubs that are found may be real hubs searched for in the Internet. Alternatively to finding Internet hubs, interest hubs of users may be determined. A database of user's browsing habits or favorite links may be considered as hubs, one for each user. The search for hubs then comprises searching in this database for users, whose interest hubs are relevant to a provided set of input hubs. The expansion of sites into hubs may be performed on the Internet, in which case the found hubs reflect the common association of links. These hubs may be used to find links that exist in the database of user habits. Alternatively also the expansion of sites into hubs is performed in the user habits database, in which case the found hubs reflect the preferences of the particular users in the database. A similarity between user browsing habits (or favorite links) and hub sites, which may be noted, is that both are lists of links that are organized by a thinking being to reflect a particular thought, topic or personality.

[0021] An aspect of some embodiments of the invention relates to a method of presenting a list of hub sites. Alternatively or additionally, to providing as a list of sites, the sites may be provided along with auxiliary information, for example, information about link structure, such as number of links, number of unique links (not in other pages), number of popular links (on at least k pages), amount of explanation for each link, method of ordering of links in the page (alphabetic, topical, regional, ranked, etc.), information copied from the target pages, such as he links themselves and/or explanations about the links. Copied information may be collated, for example, by target link (or equivalent links), or grouped according to other criteria, such as length, alphabetic, topic, rank, region and/or repetition.

[0022] There is thus provided in accordance with an exemplary embodiment of the invention, a method of finding WWW pages, each of which includes at least one list of links to desired Internet resources, comprising:

[0023] providing a list of URLs;

[0024] automatically generating at least one query for an Internet search tool for WWW pages that include links to at least one URL of said list of URLs;

[0025] executing said at least one generated query to provide search results that include at least one of said searched for WWW pages; and

[0026] generating a response comprising at least one indication of one of said WWW pages, responsive to said search results. Optionally, the method comprises displaying said response to a user. Alternatively or additionally, said at least one URL comprises a plurality of URLs. Alternatively or additionally, said response is generated using a single search step and no iterations. Alternatively or additionally, said method comprises ranking said search results. Optionally, ranking of a WWW page is responsive to a number of groups of URLs pointed to by said WWW page.

[0027] In an exemplary embodiment of the invention, said generating at least one search query, comprises:

[0028] dividing said list of URLs into a plurality of groups and generating at least a single query for each group, wherein said at least a single query does not differentiate which URL in said group is pointed to by the results of the search,

[0029] wherein said executing comprises executing said generated at least one query for a plurality of said groups, generating a plurality of result lists. Optionally, all of said groups have a same number of members. Alternatively, at least three of said groups have a different number of members from each other.

[0030] In an exemplary embodiment of the invention, the method comprises collating said result lists into a single list

of search results. Optionally, the method comprises ranking the contents of at least one of said result lists. Optionally, said collating is responsive to said ranking of said at least one of said result lists. Alternatively or additionally, said ranking is applied to said result list after it is generated. Optionally, the method comprises filtering said at least one result list responsive to said ranking.

[0031] In an exemplary embodiment of the invention, said ranking is applied to said result list during said execution. Optionally, said ranking is applied by adding at least one limitation to said at least one generated search query.

[0032] In an exemplary embodiment of the invention, said ranking comprises ranking responsive to a number of said URLs pointed to by said result list. Alternatively or additionally, said ranking comprises ranking responsive to a morphological property of pages of said at least one result list. Optionally, said morphological property comprises the existence of a link list.

[0033] In an exemplary embodiment of the invention, said ranking indicates a probability of a ranked page being a hub. Alternatively or additionally, said ranking comprises ranking responsive to the presence of at least one key word in pages of said at least one result list. Optionally, said key word comprises a word that is related to a content of said list of URLs. Alternatively or additionally, said key word comprises a word that serves as a statistical indicator that the page is a hub. Optionally, said key word is selected from the group "links", "index" and "resource".

[0034] In an exemplary embodiment of the invention, said providing comprises a user providing a list of URLs. Optionally, said user provided list of URLs comprises at least a part of a URL bookmark file.

[0035] In an exemplary embodiment of the invention, a method according to claim **1**, wherein said providing comprises a user providing a WWW page including a list of URLs. Alternatively or additionally, said providing comprises:

[0036] a user providing one or more topic words; and

[0037] executing a preliminary search to find a list of URLs related to said one or more topic words. Alternatively or additionally, said providing comprises:

[0038] a user providing a WWW page; and

[0039] executing a preliminary search to find a list of URLs that point to pages similar to the provided WWW page. Optionally, said executing said at least one generated query comprises executing said at least one query to ignore WWW pages that include links to said user provided URL.

[0040] In an exemplary embodiment of the invention, the method comprises filtering said search results before said generating. Alternatively or additionally, said search tool comprises a search engine. Optionally, said executing said at least one query comprises executing using a pipe feature of said search engine to limit a second search step to a list of sites found in a first search step using said search engine.

[0041] In an exemplary embodiment of the invention, said response comprises a list of said WWW pages. Optionally, said response includes link statistics for said WWW pages. Optionally, said link statistics include a number of links in each WWW page. Alternatively or additionally, said link statistics include an indicator of a uniqueness of links in each WWW page. Alternatively or additionally, said link statistics include an indicator of an amount of information associated with links in each WWW page.

[0042] In an exemplary embodiment of the invention, said response comprises a list of links listed in at least one of said WWW pages. Optionally, said response comprises a list of links listed in at least a given number of said WWW pages. Optionally, said given number is greater than 1. Alternatively, said given number is greater than 2.

[0043] In an exemplary embodiment of the invention, said list is arranged by WWW pages. Alternatively or additionally, said list comprises information associated with a link in its corresponding WWW page. Alternatively or additionally, said list indicates pages not including a link to any URL in a predetermined list of URLs. Alternatively or additionally, said list indicates pages not including a link from the contents of any URL in a predetermined list of URLs. Optionally, said predetermined list is provided by a user.

[0044] There is also provided in accordance with an exemplary embodiment of the invention, a method of finding WWW pages, each of which includes at least one list of links to desired Internet resources, comprising:

[0045] providing at least one URL;

[0046] generating a list of URLs related to said at least one URL;

[0047] determining at least one WWW page that includes links to at least one URL of said list of URLs but not to said provided at least one URL; and

[0048] generating a response comprising at least one indication of one of said at least one WWW page. Optionally, the method comprises displaying said response to a user. Alternatively or additionally, said at least one WWW page comprises a plurality of WWW pages. Optionally, said providing comprises providing a WWW page including having a link to said at least one URL.

[0049] In an exemplary embodiment of the invention, said providing comprises providing a list of a plurality of URLs. Alternatively or additionally, generating a list of related URLs, comprises generating a list of competition URLs. Alternatively or additionally, generating a list of related URLs, comprises generating a list of similar URLs. Alternatively or additionally, generating a list of related URLs, comprises finding WWW pages characterized in that a common WWW page includes links to at least one of said WWW pages and at least one of said at least one URL. Alternatively or additionally, said determining comprises executing a query on a search engine.

BRIEF DESCRIPTION OF THE DRAWINGS

[0050] Particular embodiments of the invention will be described with reference to the following description of some embodiments of the invention in conjunction with the figures, wherein identical structures, elements or parts which appear in more than one figure are optionally labeled with a same or similar number in all the figures in which they appear, in which:

[0051] FIG. **1** is a schematic illustration of a configuration of a search engine in accordance with an exemplary embodiment of the invention;

[0052] FIG. **2** is a flowchart of a method for finding hubs, in accordance with an exemplary embodiment of the invention; and

[0053] FIG. **3** is a flowchart of a method of finding sites similar to a provided list of sites, in accordance with an exemplary embodiment of the invention.

DETAILED DESCRIPTION OF SOME EMBODIMENTS GENERAL

[0054] FIG. **1** is a schematic illustration of a configuration **100** of a search engine **106** in accordance with an exemplary embodiment of the invention. A user **102** uses search engine **106** for finding sites of interest on an Internet **104**. The connection to search engine **106** is typically also through Internet **104**, but is not required. Typically, search engine **106** utilizes a database **108** that contains indexes and other information relating to WWW pages known to search engine **106**. In a typical search engine, a user provides terms and the engine responds with a list of sites that include some of the terms. Some, more advanced search engines also provide sites that appear to be related for various reasons. In some embodiments of the invention, a directory including an index of which sites link to which other sites is used as a search tool.

[0055] A search engine result analyzer **110** is optionally provided, to analyze the results of the search of index **108** by search engine **106** and to provide analyzed results to user **102**. Optionally, as will be described below, analyzer **110** also executes particular searches on search engine **106**. Although result analyzer **110** is optionally configured to work best with a particular search engine **106**, a same analyzer can work with a plurality of search engines.

[0056] An analysis of search results is generally desired as search engines do not typically provide a single or small number or exactly matching sites, rather, based on keywords or subject fields, a large number of sites that might be suitable are provided. Wading through a long list of sites is extremely time consuming. One reason for this required wading is the lack of suitable software for determining if a particular site is really relevant to user **102**. Also, valuable sites are often missed. Even indexing sites, such as Yahoo!, which use human indexers, often do not supply a suitable site, for several reasons, including, (a) not being up to date; (b) lack of coverage over much of the Internet; (c) lack of suitable manpower and/or time for such manpower to cover all the myriad subjects on the Internet; and (d) lack of a suitable index structure.

[0057] Typical reasons that a user browses the Internet for information include:

[0058] (a) searching for an answer to a particular question, optionally answered by an authority on that question;

[0059] (b) looking for an overview of a particular field; and

[0060] (c) searching for a set of sites, from which the user can derive his or her own conclusions.

[0061] The inventors of the present invention have realized that in many fields there are interested people who have compiled their own listing of relevant sites and an analysis of each relevant site; such listing sites are known as hubs. Thus, it is generally useful to provide a user with a short list of such hubs. The inventors have also realized that reviewing such hubs by a user may be a better way for the user to find a dependable and knowledgeable authority in a field, than by merely relying on an automated program that analyses links between sites. Neglecting a search for potential authorities, in accordance with some embodiments of the invention can allow a faster method to be used for finding hubs. Once these hubs are determined, there are other, further types of analysis that can be usefully presented to a user and answer other information gathering questions the user might have.

[0062] Following are several methods of analyzing search results to assist an interested user **102** in finding one or a small number of relevant sites or hub sites. Although not explicitly described in each of the below-described methods, additional filtering steps for rejecting certain sites as being unsuitable may be provided. Also it is noted that a block portion of one method may be suitable for inclusion, as is, in another method, as is also described in the exemplary implementation method, described herein.

[0063] A user may have a particular question to which he desires an answer. However, the words used in the question often do not match the words used in the field, or in the particular site that holds the answer to the user's question. In some cases, there is no common way of describing the subject of the question. Each hub can be considered, among other things, to be a dictionary of synonyms. Once a user finds one hub, the common usage of names to describe the subject of the question, generally becomes clear.

Finding Hubs

[0064] FIG. **2** is a flowchart **120** of a method for finding hubs, in accordance with an exemplary embodiment of the invention.

[0065] First, a keyword search (**122**) is performed. Alternatively, other ways of locating a plurality of sites related to the subject matter, may be used, for example, the listing of links in an existing hub may be used. Other method of providing link lists are described below. Optionally, the number of sites returned is limited, for example to 50.

[0066] An optional filtering step **124** may be performed to remove sites that are clearly unsuitable, for example multiple results in a same domain. Other possible filtering rules include: removing internal sites of other search engines, removing sites with low ratings or based on the site size or creation/revision date. After filtering, there are N sites, where N may be a result of the filtering or the filtering may be adapted to achieve a desired value for N.

[0067] The filtered search results are then grouped into groups (**126**), of size K, for example K=4. In some embodiments of the invention, K is a function of N. Alternatively or additionally, K is a function of the results in the group, for example, one group may have a small number of high ranking results while another group has a large number of low-ranking results. It is not, however, required that all groups be the same size. Various grouping methods may be used, for example, randomly selecting sites, based on order in search results (either selecting blocks of sites, or selecting evenly or non-evenly spaced sites from the search results), based on a ranking method, to create groups with balanced ranks or search order (e.g., two high and two low ranks)

and/or grouping similar sites together. Optionally, the size of the group may be inversely related to the ranking of sites in the groups.

[0068] A plurality of potential hubs are determined in step **128**, by searching for sites that include links to any site in one of the groups. Only N/K searches are required. In an exemplary search engine **106**, for each group the search is for sites that include reference to or link to the http address of at least one of the sites in the group, e.g. the search term being: "www.site1.com OR www.site2.com OR www.site3.com/stuff OR www.site4.com". Optionally, this search includes a request for ranking of search results by the search engine.

[0069] The results of all the searches are collated and then optionally ranked (**130**). In an exemplary ranking scheme, a two digit number is used, the tens being the number of searches the site came up on and the ones being the existence and number of special keywords that appear in the potential hub (to be described below). A four digit scheme may also be used. Also, different ranking methods or different weights for the different factors may be used. Exemplary special keywords are words that indicate that a site is more likely to be a hub (described below) or words from the subject topic or from the original search. In some cases, such topic words can be gleaned from the original search results (**122**), for example from the page topics or provided by a user.

[0070] In a step **132**, a small number of hubs are selected for further consideration, for example based on the ranking.

[0071] In an optional step **134**, the selected hubs are filtered to remove sites that are not desirable, for example based on an analysis of their content. Exemplary analysis rules that can be applied are: counting the number of links from the site; counting the number of links which also appear on other potential hubs; eliminating potential hubs which are almost identical to other potential hubs. Typically, but not necessarily, a larger number of links indicates a more desirable hub. If however, the number of links is too high, this may indicate an omnibus hub that may be too difficult to use, if it is not organized. A later optional step of analyzing the amount of content associated with each link and/or the organization of the links may be used to determine if such an omnibus hub is suitable for the user.

[0072] The filtered hubs are then presented to user **102** (**136**). In an exemplary embodiment of the invention, the filtered hubs are presented as a list of links. Alternatively or additionally, the sites may be provided along with auxiliary information, for example, information about link structure, such as number of links, number of unique links (not in other pages), number of popular links (on at least k pages or top q pages), amount of explanation for each link, method of ordering of links in the page (alphabetic, topical, regional, ranked , etc.), information copied from the target pages, such as the links themselves and/or explanations about the links. Alternatively or additionally, the list may indicate or be separated into pages that include many (or any) links to the user provided URL(s) and those pages that do not.

[0073] It is noted that in some embodiments of the invention hubs are found without finding authorities at the same time. A potential advantage is obviating a need for an iterative process to identify the hubs. In an exemplary embodiment of the invention, the step of determining an authority is performed manually by a user browsing through a short list of found hubs, to see which link in the hubs is suggested, by the hub, or by its contents, as a suitable authority. It is expected that in many cases, this method of finding an authority will yield better results than automatic determination of authorities, as many hubs are composed by experts and contain many hints that an automated method might not grasp, while a human user will. In this context it should be noted that many sites include links to other sites, for many reasons, which may have nothing to do with searching for information. However, when a hub includes a list of links, the list itself is often put together with some thought or logic.

[0074] In an exemplary embodiment of the invention, after relevant hubs for a field of interest are found, these hubs may be analyzed to determine whether or not a particular target site is pointed to, by the found hubs. This determination may be used, for example, for updating the hubs by the hub owner or a party offering a service of hub updating and supplementing.

[0075] Alternatively or additionally, the hubs are analyzed to detect and/or utilize inconsistencies between a search engine index and the actual state of WWW sites in the world. In one example, a new WWW site can be detected by finding it on a hub. In another example, a search result listing can be analyzed to detect relevant or virtual hubs and then the links in the hubs compared to the search results so as to provide to a user with a list limited to new sites.

[0076] Alternatively or additionally to using a keyword search as an input for a hub finding method, the input list of sites may be a listing of possible competitors. Such a list may be generated, for example, using a "find similar sites" feature common in many search engines, as applied to a site of interest, or provided by a user or found automatically by analyzing collections of sites. Optionally, the site of interest is indicated, in the search query, as desirable not to be found in the resulting hubs. This allows the finding of hubs that should contain a link to the site of interest but do not.

[0077] In a competition scenario, the most relevant hubs may be characterized, at least in part, by the number of competition sites pointed to by the potential hubs.

[0078] One method of finding such hubs is to search for hubs that include links to any one of the first N (e.g., 10) competitor sites. The search can be performed for example as shown in FIG. **2** at **128**, in which groups of sites are search together or potential hubs may be found for each individual. The results can filtered, for example as described above. The links in the found potential hubs can be collated and ranked. Optionally, if the provided or determined competitor list is too short (e.g., below a threshold length or not providing reasonable results, for example if the number of total links or hubs to the competitors as a group and/or as individuals is greater than a threshold value), it may be augmented, for example with other relevant sites, such as sites that include topic words found in the competition sites or provided by the user.

Find Matching Set

[0079] FIG. **3** is a flowchart **160** of a method of finding hubs similar to a provided list of sites and/or an existing hub, in accordance with an exemplary embodiment of the invention. The similarity is embodied in a similarity of the

content, type and/or other characteristics of links from and/or to the sites. In some cases, sites that are similar to individual ones of the provided sites are sought. In other cases, a found site is similar to a combination of the provided example sites. This method is useful, for example, for finding a group of sites that may be of interest to a particular user.

[0080] In a step **162**, a list of links is provided. This list may be provided in many ways, for example being gleaned from a provided set of sites to which similar sites are sought. In another example, this list may comprise a list of "favorites" or bookmarks or a user, of the list of links in such a list of favorites. In another example, the links are copied from the link list of a particular hub.

[0081] In a step **164**, the method of FIG. **2** is desirably applied to find potential hubs for the links. Other methods may be used as well.

[0082] In an optional filtering step **166**, some of the potential hubs are filtered out.

[0083] The resulting potential hubs may indicate other users whose interests are similar to those of a user who provided a "favorites" list.

Variations

[0084] In the above description, several filtration methods are described. An additional filtration method of hubs or sites can be based on the presence or lack of presence of topic words. Even if a user does not provide such topic words to begin with, these words may be automatically gleaned, for example from title or summary sections of relevant pages or by analyzing the text of URLs. It is noted that some search engines can be controlled to search for common words only in "summary" parts of the page or in anchor portions of the page (near links).

[0085] Another filtration method takes into account the presence of links that are essentially garbage links, such as promotions or advertising. These links may be repetitive or only a part of the link, for example the domain name is repeated. Optionally, a database of such links and their fields is maintained, so that if these links are actually of interest, this fact can be determined by the field of search matching that in the database.

[0086] Another filtration method analyzes a site based on its hub-likeness. Such an analysis may be based on the number and organization of links, existence of special sections titled, for example, "additional links" and/or the use of words common in hubs, such as "links", "index", and "resource".

[0087] A typical reason for a user searching for a hub (usually a plurality of hubs) is as a starting point for searching on a subject X. Not all hubs are equally suitable. Desirably, a "best" hub will optionally meet as many as possible of the following criteria:

[0088] (a) There are many links to relevant sites.

[0089] (b) The links are divided into meaningful categories.

[0090] (c) Each link is followed by some explanation concerning the site it points to.

[0091] These criteria can be checked using methods described herein, to rank hubs. In particular, the division of the links into categories can be determined by clustering methods. For example, taking each group of links and finding whether there are many or few hubs that include many of the links in the group.

[0092] As noted above, a search may be limited to a search in links of potential hubs or selected (by a user or automatically) ones of the hubs. Some search engines, such as "infoseek" include such a tool. Alternatively, a search can be limited by requiring the presence of selected ones of the links. Typically, the length of the search clause is limited, so that it may need to be repeated several times, each time with other of the relevant links. Alternatively or additionally, searching may be performed using a smart agent or through a web site (or software tool) dedicated to the application of the present invention.

[0093] In some embodiments, it is desirable to rank the results based on the number of links to a site in the results. Some search engines provide such a result. In other search engines, the number of displayed results can be limited to zero or one, so only the total count needs to be looked at. Alternatively, a connection to the search engine is broken as soon as the number of results is provided.

[0094] As noted above, the searching for potential hubs is optionally statistical, in that groups of sites are treated as single units. However, in some cases it is useful to treat at least one hub on an individual basis, for example if the hub is deemed to have been updated lately or based on its rank. Such a hub may be retrieved and/or the hub may be compared to each of the links, to see which links the hub actually includes.

Application of the Above Methods

[0095] The above methods may be applied in many ways, only exemplary ones of which are described below.

[0096] In one exemplary application, a service is provided to find hubs to which a WWW site should belong. Money can be charged based on clicks to the site and/or purchases at the site following travel through the links. The above analysis methods can be used, for example, to suggest to the site and/or to the hubs the suitability of listing the site. A service provider may sign contracts with the hubs to list sites at the service provider's request. The service provider can also provide an indexing service of pointing to the hubs of interest in a particular field. Unlike current Internet indexes that are all centralized, the service provider provides a distributed index, of which the service provider may not own any part, but optionally controls the existence of at least a limited number of relevant sites for the field that the index covers, in the particular slant (role) of that distributed index site. Optionally, but not necessarily, at least part of each distributed index part is arranged in a standardized format.

[0097] In an exemplary embodiment of the invention, the service provider can contact the hubs, to determine that additional site links are desired, and/or the sites, to determine that additional listings in hubs are desired. Once either an interested site owner or hub owner are determined, a partner (hub or site) for completing the listing transaction can usually be found. Alternatively, competing hubs may be set up by the service provider.

[0098] In one exemplary implementation, the service provider provides the list of sites as a file including also promotional material, so that the site owner will copy the link list with the promotional material. Alternatively or additionally, the links provided are not links directly to the targets, instead, the link passes through a server controlled by the service provider, for example to track access for charging purposes, which server can, for example, add promotional material and/or assure that the site links are up-to-date.

[0099] It is noted that one or more of the tasks of mapping sites, personalization of search engines, ranking of relevant sites and/or alerting to new sites may be provided in a significantly more efficient manner than known in the art using the methods described above, in some embodiments of the invention.

[0100] As part of competitive hubs or as a service at participating hubs, the service provider can update the site listing periodically, to reflect the changes in the Internet. Alternatively, a virtual hub (based on the method of FIG. **2**, for example), may be generated ad hoc, at a user's request. The listing may be generated in real-time or it may be updated periodically, for example once a week.

Specific Algorithms

[0101] In a particular implementation of the invention, which includes some of the above described methods, the following algorithms are implemented. Comments provided for a particular method step are not repeated for all the methods. First described are component algorithms. Then, composite algorithms that build on the component algorithms are described.

Finding Potential Hubs

[0102] This algorithm corresponds generally to the method of FIG. **2**.

Name : Centers(T)
Input : a set, T, of one or more target urls.
Output : a set of hubs that link to many urls in T.
Algorithm:

1. T' = filter(T) < delete blacklist, long url >. The source set T is filtered, for example removing long URLs and URLs on a black list.
2. Partition T' into sets of size K: $t_{1,...,}$ t_n using partitionmethod. The set is divided into small groups of URLs.
3. Define Link-to(t_i) = all non internal links to t_i.
4. For each i, compute Link-to'(t_i)
   i. Choose searchengine/settings
   ii. Submit query: link:t_{i1} ... link:t_{iK} | topic "links" "index" "resources"
   iii. Terminate query after timeout
5. RawScore(c) = |{ t_i | c∈Link-to'(t_i) , i=1,...,n}| This is a measure of whether the site acts like a hub (number of links)
6. TitleScore(c): This is a measure of whether the site looks like a hub.
   i. {"links" "index" "resources"} = +3
   ii. one topic word = +3; two = +5; three = +6
7. Score(c) = 10*RawScore + TitleScore(c)
8. Centers'(T) = {c | Score(c) > scorethreshold, Score(c) in top rankthreshold}
9. Centers(T) = Filter(Centers'(T)) < for each family keep highest score, delete dated >. Various hubs are removed, for example old hubs and mirror sites of other hubs.

Note:
for T = single url: skip 1, 2, 4iii; RawScore = 1
Parameters: K, partitionmethod, searchengine/settings, timeout, scorethreshold, rankthreshold.

Finding Potential Hubs Missing a Link

[0103] This algorithm is one of the variations described with reference to FIG. **2**, for finding hubs that do not point to a particular site that belongs to a topic of the hubs.

Name : CentersAbsent(T, url)
Input : a set, T, of one or more target urls; url
Output : a set of hubs which link to many urls in T but don't link to url.
Algorithm:

1. T' = filter(T) < delete blacklist, long url >
2. Partition T' into sets of size K: $t_{1,...,}$ t_n using partitionmethod
3. Define Link-to(t_i) = all non internal links to t_i .
4. For each i, compute Link-to-absent(t_i)
   i. Choose searchengine/settings
   ii. Submit query: -link:url link:t_{i1} ... link:t_{iK} | topic "links" "index" "resources" (Topic can be a parameter or it may be determined from the search results)
   iii. Terminate query after timeout
5. RawScore(c) = |{ ti | c∈Link-to-absent(ti) , i=1,...,n}|
6. TitleScore(c):
   i. {"links" "index" "resources"} = +3
   ii. one topic word = +3; two = +5; three = +6
7. Score(c) = 10*RawScore + TitleScore(c)
8. Centers'(T) = {c | Score(c) > scorethreshold, Score(c) in top rankthreshold}
9. Centers(T) = Filter(Centers'(T)) < delete dated >

Parameters: K, partitionmethod, searchengine/settings, timeout, scorethreshold, rankthreshold

Finding Sites Relevant to a Topic

[0104] This algorithm allows a user to selectively provide a list of sites or a topic, as an input into the other methods.

Name : Relevant(topic)
Input : topic

-continued

Output : sites about topic
1. Get top k_S sites on searchengine
2. Filter <remove duplicates, junk>
Parameters: searchengine, k_S

Finding Hubs About a Topic

[0105] This algorithm is one implementation of the method of FIG. **2**.

Input: topic
Output: Hubs on topic
Algorithm:
1. Compute Relevant(topic)
2. Compute Centers(Relevant(topic, searchengine))

Find Similar Hubs

[0106] This algorithm identifies hubs that are similar to a known hub, relate to a same subject field and thus are useful as a starting point for searching.

Name : SimilarHubs (for single hub c)
Input : hub c, topic
Output : Similar hubs
Algorithm:
1. Compute Targets(c) = external links in c
2. If |Targets(c)| = 20 − j
   i. Get Relevant(topic) <top j>
   ii. Targets'(c) = Targets(c) ∪ Relevant(topic)
3. If |Targets(c)| = k ≧ 200. A parameter of the method.
   i. Targets'(c) = |Targets(c)|/Floor(|Targets(c)|/100) of Targets(c)
   <Select periodically>
4. Else Targets'(c) = Targets(c)
5. Compute Centers (Targets'(c))

Find Potential Hubs That Should List a Site

[0107] This algorithm identifies hubs that do not point to a particular site (FIG. **2**).

Name : Place Target
Input : url, topic
Output : hub sites not linked to url
Algorithm:
1. Get Competitors(url)
2. For each t ∈ Competitors(url) compute CentersAbsent(t, url)
3.
4. If |CentersAbsent(t, url)| < 20
   Compute Relevant(topic)
5. Compute CentersAbsent(Relevant(topic) , url)
6. PlaceLink(url) =
   [CentersAbsent(t, url) ∪ CentersAbsent(Relevant(topic) , url)] −
   Similar(Link-to(url))
7. If c ∈ PlaceLink(url) but c ∈ −CentersAbsent(t, url)
Score(c) = 10 + RawScore(c) (in CentersAbsent(Relevant(topic) , url))

Physical Implementation

[0108] Search result analyzer **110** may be implemented in various ways, optionally without limiting its ability to provide the services described above.

[0109] In one example, search analyzer **110** is integrated with search engine **10**, possibly in a same computer or in a LAN thereof.

[0110] In another example, search analyzer **110** is a separate WWW server that contacts search engine **106** via the Internet or directly.

[0111] In another example, search analyzer **110** is, at least in part, a client software executing on user **102**. This client software may be permanent or it may be a network programmed, e.g. Java, applet that is downloaded by user **102** at need.

[0112] It will be appreciated that the above described methods of hub and site finding may be varied in many ways, including, changing the order of steps, which steps are performed on-line or off-line, such as table or index preparation, and the exact implementation used, which can include various hardware and software combinations. In addition, a multiplicity of various features has been described. It should be appreciated that different features may be combined in different ways. In particular, not all the features are necessary in every preferred embodiment of the invention. Software as described herein is preferably provided on a computer readable media, such as a diskette or an optical disk. Alternatively or additionally, it may be stored on a computer, for example in a main memory or on a hard disk, both of which are also computer readable media. Where methods have been described, also computer hardware programmed to perform the methods is within the scope of the description. When used in the following claims, the terms "comprises", "includes", "have" and their conjugates mean "including but not limited to".

[0113] It will be appreciated by a person skilled in the art that the present invention is not limited by what has thus far been described. Rather, the scope of the present invention is limited only by the following claims.

1. A method of finding WWW pages, each of which includes at least one list of links to desired Internet resources, comprising:

providing a list of URLs;

automatically generating at least one query for an Internet search tool for WWW pages that include links to at least one URL of said list of URLs;

executing said at least one generated query to provide search results that include at least one of said searched for WWW pages;

generating a response comprising at least one indication of one of said WWW pages, responsive to said search results;

wherein said response comprises a list of links listed in at least one of said WWW pages; and

wherein said list indicates pages not including a link to any URL in a predetermined list of URLs.

2. A method according to claim 1, comprising displaying said response to a user.

3. A method according to claim 1, wherein said at least one URL comprises a plurality of URLs.

4. A method according to claim 1, wherein said response is generated using a single search step and no iterations.

5. A method according to claim 1, comprising ranking said search results.

6. A method according to claim 5, wherein ranking of a WWW page is responsive to a number of groups of URLs pointed to by said WWW page.

7. A method according to claim 1, wherein said generating at least one search query, comprises:

dividing said list of URLs into a plurality of groups and generating at least a single query for each group, wherein said at least a single query does not differentiate which URL in said group is pointed to by the results of the search,

wherein said executing comprises executing said generated at least one query for a plurality of said groups, generating a plurality of result lists.

8. A method of finding WWW pages, each of which includes at least one list of links to desired Internet resources, comprising:

providing a list of URLs;

automatically generating at least one search query for an Internet search tool for WWW pages that include links to at least one URL of said list of URLs;

executing said at least one generated query to provide search results that include at least one of said searched for WWW pages;

generating a response comprising at least one indication of one of said WWW pages, responsive to said search results;

wherein said generating at least one search query, comprises:

dividing said list of URLs into a plurality of groups and generating at least a single query for each group, wherein said at least a single query does not differentiate which URL in said group is pointed to by the results of the search,

wherein said executing comprises executing said generated at least one query for a plurality of said groups, generating a plurality of result lists; and

wherein all of said groups have a same number of members.

9. A method of finding WWW pages, each of which includes at least one list of links to desired Internet resources, comprising:

providing a list of URLs;

automatically generating at least one search query for an Internet search tool for WWW pages that include links to at least one URL of said list of URLs;

executing said at least one generated query to provide search results that include at least one of said searched for WWW pages;

generating a response comprising at least one indication of one of said WWW pages, responsive to said search results;

wherein said generating at least one search query, comprises:

dividing said list of URLs into a plurality of groups and generating at least a single query for each group, wherein said at least a single query does not differentiate which URL in said group is pointed to by the results of the search,

wherein said executing comprises executing said generated at least one query for a plurality of said groups, generating a plurality of result lists; and

wherein at least three of said groups have a different number of members from each other.

10. A method according to claim 7, comprising:

collating said result lists into a single list of search results.

11. A method according to claim 10, comprising ranking the contents of at least one of said result lists.

12. A method according to claim 11, wherein said collating is responsive to said ranking of said at least one of said result lists.

13. A method according to claim 11, wherein said ranking is applied to said result list after it is generated.

14. A method according to claim 13, comprising filtering said at least one result list responsive to said ranking.

15. A method according to claim 11, wherein said ranking is applied to said result list during said execution.

16. A method according to claim 15, wherein said ranking is applied by adding at least one limitation to said at least one generated search query.

17. A method according to claim 11, wherein said ranking comprises ranking responsive to a number of said URLs pointed to by said result list.

18. A method according to claim 11, wherein said ranking comprises ranking responsive to a morphological property of pages of said at least one result list.

19. A method according to claim 18, wherein said morphological property comprises the existence of a link list.

20. A method according to claim 11, wherein said ranking indicates a probability of a ranked page being a hub.

* * * * *