

USPTO PATENT FULL-TEXT AND IMAGE DATABASE

(1 of 1)

United States Patent	7,603,350
Guha	October 13, 2009

****Please see images for: (Certificate of Correction) ****

Search result ranking based on trust

Abstract

A search engine system provides search results that are ranked according to a measure of the trust associated with entities that have provided labels for the documents in the search results. A search engine receives a query and selects documents relevant to the query. The search engine also determines labels associated with selected documents, and the trust ranks of the entities that provided the labels. The trust ranks are used to determine trust factors for the respective documents. The trust factors are used to adjust information retrieval scores of the documents. The search results are then ranked based on the adjusted information retrieval scores.

Inventors: Guha; Ramanathan (Los Altos, CA)

Assignee: Google Inc. (Mountain View, CA)

Family ID: 41138080

Appl. No.: 11/382,449

Filed: May 9, 2006

Current U.S. Class: 1/1; 709/245; 707/999.005

Current CPC Class: G06F 16/24578 (20190101); G06F 16/951 (20190101); Y10S 707/99935 (20130101)

Current International Class: G06F 7/00 (20060101)

References Cited [\[Referenced By\]](#)

U.S. Patent Documents

6336117	January 2002	Massarani
6360215	March 2002	Judd et al.
6421675	July 2002	Ryan et al.
6636854	October 2003	Dutta et al.
7031961	April 2006	Pitkow et al.
2006/0200487	September 2006	Adelman et al.
2006/0294086	December 2006	Rose et al.

Other References

Zoltan Gyongyi et al. Combating Web Spam with TrustRank, 2004, pp. 576-587, (<http://portal.acm.org/citation.cfm?id=1316689.1316740&coll;=&dl;=->). cited by examiner .

Google Directory Search, Feb. 1, 2004

(<http://web.archive.org/web/20040201210740/http://www.google.com/dirhp>). cited by examiner .

R. Guha, Open Rating Systems, WayBackMachine--Mar. 22, 2005,

(<http://web.archive.org/web/20050322033341/http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/open.sub.--rating.sub.--systems/wot.pdf>). cited by examiner .

Epinions.com, WayBackMachine--2004

(<http://web.archive.org/web/20041013122547/www.epinions.com/Movies>),

(<http://web.archive.org/web/20041211104641/www.epinions.com/mvie.sub.--mu-1131678/display.sub.--about.reviews>),

<http://web.archive.org/web/20041210125754/www.epinions.com/Movies-Foreign-sub.--Films>). cited by examiner.

Primary Examiner: Ali; Mohammad

Assistant Examiner: Shmatov; Alexey

Attorney, Agent or Firm: Fish & Richardson P.C.

Claims

What is claimed is:

1. A method performed by a data processing apparatus, the method comprising: receiving a search query, the search query comprising a query term and a query label term, the query

label term being a categorical identifier; identifying one or more resources in a search result set responsive to the search query, wherein each of the one or more resources has an associated matching label term that matches the query label term and an associated non-matching label term that does not match the query label term, each associated label term being the categorical identifier describing each resource and having been associated with the resource by a respective entity; determining, for each of the one or more resources: a trust rank of the entity that associated the associated matching label term with the resource, wherein the trust rank indicates whether a user trusts the associated matching label term of the resource, and increasing a relevance score of each resource that has the associated matching label term based on the respective trust rank, the relevance score indicating a degree of relevance between the respective resource and the query term; ranking each of the one or more resources in the search result set based on the respective relevance scores; annotating, with a name of the respective entity that associated the matching label term with each resource, indicia identifying each resource of the ranked search result set; and providing the indicia in a response to the search query.

2. The method of claim 1, wherein annotating, with the name of the respective entity further comprises: annotating, with the matching label term associated with each resource, the indicia identifying each resource of the ranked search result set.

3. The method of claim 1, further comprising: determining, for each of the one or more resources: a second trust rank of a second entity that associated a second matching label term with the resource; and aggregating the trust rank and the second trust rank to determine a trust factor.

4. A method executed by one or more processors causing the one or more processors to execute the steps of: receiving a search query, the search query comprising a query term and a query label term, the query label term being a categorical identifier; identifying one or more resources in a search result set responsive to the search query, wherein each of the one or more resources has two or more associated matching label terms that match the query label term and that each label term being the categorical identifier describing each resource and been associated with the resource by a different respective entity; determining, for each of the one or more resources: a trust rank of each entity that associated a matching label term with the resource, wherein the trust rank indicates whether a user trusts the associated matching label term of the resource, and increasing a relevance score of each resource that has the two or more associated matching label terms based on the respective trust factors, the relevance score indicating a degree of relevance between the respective resource and the query term; ranking each of the one or more resources in the search result set based on the respective relevance scores; annotating, with a name of each entity that associated the matching label terms with each resource, indicia identifying each resource of the ranked search result set; and providing the indicia in a response to the search query.

5. The method of claim 4, further comprising: determining, for each of the one or more resources: a second trust rank of a second entity that associated a second matching label term with the resource; and aggregating the trust rank and the second trust rank to determine a trust factor.

6. The method of claim 4, wherein annotating, with the name of each entity further comprises: annotating, with the two matching label terms associated with each resource, the indicia identifying each resource of the ranked search result set.

7. A system comprising: one or more computers; a computer-readable medium coupled to the one or more computers having instructions stored thereon which, when executed by the one or more computers, cause the one or more computers to perform operations comprising: receiving a search query, the search query comprising a query term and a query label term, the query label term being a categorical identifier; identifying one or more resources in a search result set responsive to the search query, wherein each of the one or more resources has an associated matching label term that matches the query label term and an associated non-matching label term that does not match the query label term, each associated label term being the categorical identifier describing each resource and having been associated with the resource by a respective entity; determining, for each of the one or more resources: a trust rank of the entity that associated the associated matching label term with the resource, wherein the trust rank indicates whether a user trusts the associated matching label term of the resource, and increasing a relevance score of each resource that has the associated matching label term based on the respective trust rank, the relevance score indicating a degree of relevance between the respective resource and the query term; ranking each of the one or more resources in the search result set based on the respective relevance scores; annotating, with a name of the respective entity that associated the matching label term with each resource, indicia identifying each resource of the ranked search result set; and providing the indicia in a response to the search query.

8. The system of claim 7, where in the operations further comprise: annotating, with the matching label term associated with each resource, the indicia identifying each resource of the ranked search result set.

9. The system of claim 7, wherein the operations further comprise: determining, for each of the one or more resources: a second trust rank of a second entity that associated a second matching label term with the resource; and aggregating the trust rank and the second trust rank to determine a trust factor.

10. A system comprising: one or more computers; a computer-readable medium coupled to the one or more computers having instructions stored thereon which, when executed by the one or more computers, cause the one or more computers to perform operations comprising: receiving a search query, the search query comprising a query term and a query label term, the query label term being a categorical identifier; identifying one or more resources in a search result set responsive to the search query, wherein each of the one or more resources has two or more associated matching label terms that match the query label term and that each label term being the categorical identifier describing each resource and been associated with the resource by a different respective entity; determining, for each of the one or more resources: a trust rank of each entity that associated a matching label term with the resource based on the trust information; and increasing a relevance score of each resource that has the two or more associated matching label terms based on the respective trust ranks, the relevance score indicating a degree of relevance between the respective resource and the query term; ranking each of the one or more resources in the search result set based on the respective relevance scores; annotating, with a name of each entity that

associated the matching label term with each resource, indicia identifying each resource of the ranked search result set; and providing the indicia in a response to the search query.

11. The system of claim 10, further comprising: determining, for each of the one or more resources: a second trust rank of a second entity that associated a second matching label term with the resource; and aggregating the trust rank and the second trust rank to determine a trust factor.

12. The system of claim 10, wherein the operations further comprise: annotating, with the two matching label terms associated with each resource, the indicia identifying each resource of the ranked search result set.

13. A computer storage medium encoded with a computer program, the program comprising instructions that when executed by data processing apparatus cause the data processing apparatus to perform operations comprising: receiving a search query, the search query comprising a query term and a query label term, the query label term being a categorical identifier; identifying one or more resources in a search result set responsive to the search query, wherein each of the one or more resources has an associated matching label term that matches the query label term and an associated non-matching label term that does not match the query label term, each associated label term being the categorical identifier describing each resource and having been associated with the resource by a respective entity; determining, for each of the one or more resources: a trust rank of the entity that associated the associated matching label term with the resource based on the trust information, wherein the trust rank indicates whether a user trusts the associated matching label term of the resource, and increasing a relevance score of each resource that has the associated matching label term based on the respective trust rank, the relevance score indicating a degree of relevance between the respective resource and the query term; ranking each of the one or more resources in the search result set based on the respective relevance scores; annotating, with a name of the respective entity that associated the matching label term with each resource, indicia identifying each resource of the ranked search result set; and providing the indicia in a response to the search query.

14. A computer storage medium encoded with a computer program, the program comprising instructions that when executed by data processing apparatus cause the data processing apparatus to perform operations comprising: receiving a search query, the search query comprising a query term and a query label term, the query label term being a categorical identifier; identifying one or more resources in a search result set responsive to the search query, wherein each of the one or more resources has two or more associated matching label terms that match the query label term and that each label term being the categorical identifier describing each resource and been associated with the resource by a different respective entity; determining, for each of the one or more resources: a trust rank of each entity that associated a matching label term with the resource, wherein the trust rank indicates whether a user trusts the associated matching label term of the resource, and increasing a relevance score of each resource that has the two or more associated matching label terms based on the respective trust ranks, the relevance score indicating a degree of relevance between the respective resource and the query term; ranking each of the one or more resources in the search result set based on the respective relevance scores; annotating, with a name of each entity that associated the matching label term with each

resource, indicia identifying each resource of the ranked search result set; and providing the indicia in a response to the search query.

Description

FIELD OF THE INVENTION

The present invention relates to search engines, and more specifically to search engines that use information indicative of trust relationship between users to rank search results.

BACKGROUND OF THE INVENTION

The development of information retrieval systems has predominantly focused on improving the overall quality of the search results presented to the user. The quality of the results has typically been measured in terms of precision, recall, or other quantifiable measures of performance. Information retrieval systems, or `search engines` in the context of the Internet and World Wide Web, use a wide variety of techniques to improve the quality and usefulness of the search results. These techniques address every possible aspect of search engine design, from the basic indexing algorithms and document representation, through query analysis and modification, to relevance ranking and result presentation, methodologies too numerous to fully catalog here.

An inherent problem in the design of search engines is that the relevance of search results to a particular user depends on factors that are highly dependent on the user's intent in conducting the search--that is why they are conducting the search--as well as the user's circumstances, the facts pertaining to the user's information need. Thus, given the same query by two different users, a given set of search results can be relevant to one user and irrelevant to another, entirely because of the different intent and information needs. Most attempts at solving the problem of inferring a user's intent typically depend on relatively weak indicators, such as static user preferences, or predefined methods of query reformulation that are nothing more than educated guesses about what the user is interested in based on the query terms. Approaches such as these cannot fully capture user intent because such intent is itself highly variable and dependent on numerous situational facts that cannot be extrapolated from typical query terms.

In part because of the inability of contemporary search engines to consistently find information that satisfies the user's information need, and not merely the user's query terms, users frequently turn to websites that offer additional analysis or understanding of content available on the Internet. For the purposes of discussion these sites are called vertical knowledge sites. Some vertical knowledge websites, typically community sites for users of shared interests, allow users to link to content on the Internet and provide labels or tags describing the content. For example, a site may enable a user to link to the website of an automobile manufacturer, and post comment or description about a particular car being offered by the manufacturer; similarly, such a site could enable a user to link to a news report on the website of a news organization and post comment about the report. These and other vertical knowledge sites may also host the analysis and comments of experts or others with knowledge, expertise, or a point of view in particular fields, who again can comment on

content found on the Internet. For example, a website operated by a digital camera expert and devoted to digital cameras typically includes product reviews, guidance on how to purchase a digital camera, as well as links to camera manufacturer's sites, new products announcements, technical articles, additional reviews, or other sources of content. To assist the user, the expert may include comments on the linked content, such as labeling a particular technical article as "expert level," or a particular review as "negative professional review," or a new product announcement as "new 10 MP digital SLR". A user interested in a particular point of view, type of information, or the like then search within the domain of such a site for articles or links that have certain associated labels or comments. For example, a user could search the aforementioned digital camera site for all camera reviews labeled "digital SLR".

One of the underlying aspects of vertical knowledge sites that makes them appealing to users is that some of the user who participate on them come to be perceived as being trustworthy in their comments, analysis, opinions and recommendations. This degree of trust is valuable to users as a way of evaluating the often bewildering array of information that is available on the Internet. Indeed, many popular vertical knowledge sites, blogs, news outlets, so forth, are primarily devoted to facilitating dissemination of the opinions of individual experts or commentators, while other vertical knowledge sites such as forums, rating sites, and community sites operate to share and disseminate the opinions of many users in a community. In either case, many users access these sites because of an underlying sense of trust in at least some of the others users who are providing their opinions. Of course, in some cases a particular user's (or author's) views may not be trusted by others. For example, on most forums there will be one or more users who are viewed by others as being uninformed, biased, hostile or otherwise not trustworthy. Particularly for new users, identifying which other members of a community are trustworthy and which are not can be a difficult and time consuming process.

Some vertical knowledge sites now provide various types of indicators or proxies for the trustworthiness of particular individuals who participate at the site. Auction sites use rating systems to identify trusted buyers and sellers. Forum sites use membership criteria and other factors to distinguish between posters. But many sites simply rely on general reputation of their experts to instill a sense of trust in users who visit the site. Thus, at best a user can currently search within the context or domain of a particular website for comments, opinions or the like made by individuals who are trusted by the user or by others.

The problem remains that when the user returns to a general search engine, outside of the vertical knowledge site, the user is unable to obtain search results that reflect the trustworthiness of the documents themselves or the trustworthiness of any commentary or opinions that may be associated with the search result documents. Thus, none of the additional reputation based information that is associated with users in the vertical knowledge site is available to the general search engine in order to provide more meaningful search results to other users.

SUMMARY

A search engine system provides search results that are ranked according to a measure of the trust associated with entities that have provided labels for the documents in the search

results. The system maintains information describing annotations provided by any of the entities. An entity creates an annotation for a particular item (or collection) of web content as a way of associating a label with a particular item of web content, such as a web page. For example, an entity such as a digital camera expert operating a website devoted to digital cameras, may create an annotation associating the label "professional review" with a particular review of a digital camera on some third party site (e.g., on the site of a news publication). In addition, the system maintains information about trust relationships between entities, such as individual users, indicating whether (or the degree to which) one entity trusts another entity. Thus, users can indicate whether or the extent to which they trust the digital camera expert. The system can determine a trust rank for any particular entity, based on the trust relationships between the various entities.

A user provides a query to the system; the query contains at least one query term and optionally includes one or more labels of interest to the user. The system retrieves a set of search results comprising documents that are relevant to the query term(s). The system determines which query labels are applicable to which of the search result documents. The system determines for each document an overall trust factor to apply to the document based on the trust ranks of those entities that provided the labels that match the query labels. Applying the trust factor to the document adjusts the document's information retrieval score, to provide a trust adjusted information retrieval score. The system reranks the search result documents based at on the trust adjusted information retrieval scores.

The present invention has various embodiments, including as a computer implemented process, as computer apparatuses, as integrated circuits, and as computer program products that execute on general or special purpose processors.

The features and advantages described in this summary and the following detailed description are not all-inclusive. Many additional features and advantages will be apparent to one of ordinary skill in the art in view of the drawings, specification, and claims hereof.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1 illustrates a generalized system architecture for a search engine system in accordance with one embodiment.

FIG. 2 is a flowchart of a method for collecting trust information.

FIG. 3 is a flowchart of a method for collecting annotation information.

FIG. 4 is a flowchart of a method for processing search results with trust ranks.

FIG. 5 is an example of a search query including a label and corresponding search results.

The figures depict various embodiments of the present invention for purposes of illustration only. One skilled in the art will readily recognize from the following discussion that alternative embodiments of the illustrated and described structures, methods, and functions may be employed without departing from the principles of the invention.

DETAILED DESCRIPTION

System Overview

FIG. 1 presents a diagram illustrating a system that ranks search results using trust in accordance with an embodiment of the present invention. Referring to FIG. 1, a user operates a browser 104 located on a client 103. Client 103 can generally include any node on a network including computational capability and including a mechanism for communicating across the network. The client 103 can be any type of client, including any type of computer (e.g., desktop computer, workstation, notebook, mainframe, terminal, etc.), handheld device (personal digital assistant, cellular phone, etc.), or the like. The client device 103 need only have the capability to communicate over a network (e.g. Internet, telephony, LAN, WAN, or combination thereof) with the search engine system 100. Typically, a client device 103 will also support appropriate networking applications and components, all of which are known to those of skill in the art.

Browser 104 can generally include any type of web browser capable of viewing a web site, such as the INTERNET EXPLORER.TM. browser distributed by the Microsoft Corporation of Redmond, Wash. Search engine system 100 can generally include any computational node including a mechanism for servicing queries from a client for computational and/or data storage resources. Note that the code and data involved in processing queries typically resides within a memory within search engine system 100. Copies of the code and data can also reside within non-volatile storage which is included in the system 100. The figure does not show a number of conventional components (e.g. network, firewalls, routers, domain name servers, load balancers, etc.) in order to not obscure the relevant details of the embodiment.

During operation, a user's client 103 can communicate with search engine system 100 to search for documents relevant to a query. A query general comprises one or more query terms and optionally includes one or more labels. Query terms are words that the user believes are relevant to the user's information need; labels are words, phrases, markers or other indicia that have been associated with certain web content (pages, sites, documents, media, etc.) by others as descriptive or categorical identifiers. For example, a query "cancer label:symptoms" includes the query term cancer and a query label "symptoms", and thus is a request for documents relevant to cancer, and that has been labeled as relating to "symptoms." Labels such as these can be associated with documents from any entity, whether the entity created the document, or is a third party. The entity that has labeled a document has some degree of trust, as further described below.

The browser 104 can also access any web site available on the network (not shown) belonging to a person, or any other type of entity such as a company, enterprise, community group, and so forth. For example, in FIG. 1 browser 104 accesses Expert's web site 105, which is hosted by a third-party server, or alternatively by the server associated with the search engine system 100. Expert's web site 105 can generally be any type of web content provided by Expert. For example, Expert's web site 105 can include Expert's personal home page, along with pages directed towards Expert's interests, as well as Expert's blog, an e-commerce storefront for goods that Expert sells, and so forth. In addition, Expert's site 105 can include pages containing links to other sites (including entire sites, portions of a site, or

individual pages) along with annotations 106 associated with such linked content. An annotation 106 includes a label 107 and a URL pattern associated with the label; the URL pattern can be specific to an individual web page or to any portion of a web site or pages therein. The term "document" is used herein to refer to any and all types of content that can be accessed by a client over a network.

For example, Expert may create an annotation 106 including the label 107 "Professional review" for a review 114 of Canon digital SLR camera on a web site "www.digitalcameraworld.com", a label 107 of "Jazz music" for a CD 115 on the site "www.jazzworld.com", a label 107 of "Classic Drama" for the movie 116 "North by Northwest" listed on website "www.movierental.com", and a label 107 of "Symptoms" for a group of pages describing the symptoms of colon cancer on a website 117 "www.yourhealth.com". Note that labels 107 can also include numerical values (not shown), indicating a rating or degree of significance that the entity attaches to the labeled document.

Expert's web site 105 can also include trust information. More specifically, Expert's web site 105 can include a trust list 109 of entities whom Expert trusts. This list may be in the form of a list of entity names, the URLs of such entities' web pages, or by other identifying information. Expert's web site 105 may also include a vanity list 111 listing entities who trust Expert; again this may be in the form of a list of entity names, URLs, or other identifying information. The trust list 109 and vanity list 111 may be encoded in any computer readable form supported by Expert's website. An "entity" may be a specific person, group, organization, website, business, institution, government agency or the like.

In one embodiment of the present invention, a web crawler (not shown) obtains labels and trust information and sends it to search engine system 100 to facilitate subsequent usage in search result ranking. This process is described in more detail below with reference to FIGS. 2-4.

Expert's web site 105 can also include a "trust button" 112. A user visiting Expert's web site can click on trust button 112 to cause a corresponding trust relationship to be recorded at the search engine system 100. This trust relationship indicates that the user trusts Expert, and may be understood as well to indicate that the user trusts the labels that Expert has provided for various documents and site. Note that if Expert's web site 105 is provided by a third-party server, the trust button 112 can be implemented as an inline frame (iframe) provided by such server, which examines a cookie to identify user and then provides an identifier for the user and an identifier for Expert to the search engine system 100. When accessed, the search engine system 100 receives information indicative of the trust relationship between the user and Expert. This information may take the form of a tuple of data <entity1, entity2, trust_value> identifying the trusting entity, the trusted entity, and a trust value. For example, the tuple <3365, 1230, 1> indicates that the entity (e.g., a user) with the identifier 3365 trusts the entity with the identifier 1230 (e.g., Expert) with a trust value of 1. The trust information is stored in a trust database 190.

In one embodiment, a user can have different levels of trust in an entity for different topics. For example, a user may trust an entity with respect to politics and economics, but not with respect to sports and entertainment. The trust information tuple can thus be extended as <entity1, entity2, topic, trust_value>, where topic identifies the particular topic for which

entity1 trusts entity2 (the topic can be identified by a text string, a topic identifier, or other means). The user manifestation of topic based trust can vary. For example, the trust button 112 can be linked to a further selection or list of topics. For a trust list, the user can include a list of the trusted topics for each entity he or she trusts. The trust information provided by the users with respect to others is used to determine a trust rank for each user, which is a measure of the overall degree of trust that users have in the particular entity.

In order to obtain search results, user's client 103 sends a search query to the search engine system 100, which is received by a front end server 110. The front end server 110 is adapted to receive a query from the client 102 and pass it to the query processor 120, and then subsequently receive the search results back from the query processor 120 and provide them to the client 102. The front end server 110 is further adapted to handle the queries from multiple different clients 102 concurrently.

The front end server 110 passes the query to the search engine 180, which processes a user's query (using a content server and index 170) to obtain a set of search results that are relevant to the query term(s). Where the query includes labels (as these are optional), the search engine 180 accesses an annotation database 140 to determine which of the documents in the search results have one or more of the query labels associated with them. Using the annotation information and trust information from the trust database 190, the search engine 180 determines a trust factor for each document. A document's trust factor is a function of the trust ranks associated with the entities that have labeled the document with labels that match the query labels. The search engine 180 adjusts each document's underlying information retrieval score using the document's trust factor, and then reranks the search results using the adjusted scores. The trust adjusted search results are returned to the front end server 110, which passes them to the client 103.

FIG. 5 illustrates an example of search query including a label and corresponding search results. Search query field 506 is shown including a search query 500 that includes the query term "cancer" and the indicated label "symptoms." This query has been provided to the search engine system 100, and some of the search results 502 (a set of documents) as shown in trust-adjusted ranked order. Of interest in each search result 502 (document) is the indicia of the matching label 504 "symptom" for each shown search result, as well as other labels 504 such as "Tests/diagnosis," and "Treatment." Also shown next to each result is the name 508 of the entity that provided the matching label, here the user named Guha; the name 508 of another user Roni, is also shown on one of the results. The results are listed in an order based on the trust ranks of the entities 508 which have provided the labels 504 that match the query label of "symptoms." Thus, the user can determine the quality and significance of each search result in part by seeing which entities have provided the labels that match the query labels.

Obtaining and Storing Trust Information

FIG. 2 presents a flowchart illustrating the process of obtaining and storing trust information in accordance with an embodiment of the present invention; other variations of the steps or procedures of this process can be readily developed to achieve the same results.

During this process, the system 100 receives 202 information indicating that the user trusts labels provided by the entity. This can involve receiving this information directly from the user. For example, the user can click on a "trust button" on a web page belonging to the entity, which causes a corresponding record for a trust relationship to be recorded in the trust database 190. In general any type of input from the user indicating that such as trust relationship exists can be used.

As indicated above, the system can also use a web crawler to examine web pages to locate information indicating that which user trusts a particular entity. While examining web pages, the web crawler can look for a number of relationships, including: (1) links from the user's web page to web pages belonging to trusted entities; (2) a trust list that identifies entities that the user trusts; or (3) a vanity list which identifies users who trust the owner of the vanity page.

The system can also examine web visitation patterns of the user and can infer from the web visitation patterns which entities the user trusts. For example, the system can infer that a particular user trust a particular entity when the user visits the entity's web page with a certain frequency.

The search engine system 100 can also examine computer-based records belonging to the user, such as an email contact list or an instant-messaging chat list, and can infer from these computer-based records that the user trusts entities included in such contact or chat lists. More particular, for each entity included in the user's contact/chat list, a trust record indicating a trust relationship between the user and entity can be entered in the trust database 190. Alternatively, a user may specifically mark or identify entities in the user's contact or instant messaging list as being trusted, or with a particular degree of trust, or trusted topics.

The search engine system 100 can also infer trust transitively, by assigning a trust value between a user and first entity, where the user directly trusts a second entity, and the second entity in turn trusts the first entity.

Next, the search engine system 100 stores 204 a corresponding record for the trust relationship between the user and the entity in the trust database 190 to facilitate subsequent data retrieval operations. The trust database 190 can include any type of data structure or database system that can facilitate data retrieval of trust information.

In one embodiment, the trust information in the trust database 190 is stored in a square trust matrix M , where each matrix value $M_{sub.i,j}$ stores a value indicative of entity i 's trust of entity j . As noted above, in one embodiment, the trust value is 0 or 1, though other values or ranges of values can be used. The matrix information can be stored in any suitable storage structure, including tables, arrays, bit maps, trees, or the like.

For each entity i , a trust rank $TR_{sub.i}$ can be computed 206 from the trust matrix M . The $TR_{sub.i}$ of entity i is the $i_{sup.th}$ component of the eigenvector of TR that is associated with the eigenvalue of 1. Where trust information includes topic information, a separate trust matrix M is constructed for each topic t and a topic-specific trust rank $TR_{sub.it}$ is computed. The topic matrices can aggregated at various levels according to topical hierarchies, and

aggregated topic-specific trust ranks may also be computed therefrom. While FIG. 2 shows the computation 206 of trust rank following storing 204, these steps need not occur in this order, and in one embodiment, the computation 206 of trust rank is executed in an independent process.

Note that trust relationships can change. For example, the system can increase (or decrease) the strength of a trust relationship for a trusted entity. The search engine system 100 can also cause the strength of a trust relationship to decay over time if the trust relationship is not affirmed by the user, for example by visiting the entity's web site and activating the trust button 112. The search engine system 100 may also expose a user interface to the trust database 190 by which the user can edit the user trust relationships, including adding or removing trust relationships with selected entities.

The trust information in the trust database 190 is also periodically updated by crawling of web sites, including sites of entities with trust information (e.g., trust lists, vanity lists); trust ranks are recomputed based on the updated trust information.

Obtaining and Storing Label Information

FIG. 3 presents a flowchart illustrating the process of obtaining and storing annotations (including labels) on web content in accordance with an embodiment of the present invention; other variations of the steps or procedures of this process can be readily developed to achieve the same results.

During this process, the system gathers 302 annotations 106 associated with an entity. This can involve using a web crawler to examine web pages associated with the entity to identify annotations 106 created by the entity. One system and methodology for analysis of annotations provided by entities is described in U.S. application Ser. No. 11/202,423, filed on Aug. 10, 2005 and entitled "Programmable Search Engine", which is incorporated by reference herein.

The system can also receive annotations 106 from an entity via an annotation interface 160 that is configured receive annotation files. An annotation file contains annotations. As indicated previously, an annotation includes a pattern for a uniform resource locator (URL) for the URLs of documents, and a label to be applied to a document whose URL matches the URL pattern. Schematically, an annotation may take the form:

<label, URL_pattern>

where label is a term or phrase, and URL_pattern is a specification of a pattern for a URL.

For example, the annotation

<"professional review", www.digitalcameraworld.com/review/>

would be used to apply the label "professional review" to any document whose URL includes a URL prefix matching the network location "www.digitalcameraworld.com/review/"; all documents in this particular host's directory are considered by the entity providing the

annotation to be "professional reviews" of digital cameras. In one embodiment, the URL pattern can include wildcards as well as regular expressions. In a typical embodiment, the annotation database 104 includes thousands, even millions of such annotations. There is a many-to-many relationship between labels and URL patterns, in that a given label may be applied to any document matching multiple different URL patterns, and a given URL pattern may be associated with many different labels.

From the received annotation information, the system stores the annotations in the annotation database 140. The annotation database 140 includes an inverted index (or equivalent) that indexes each annotation to whatever URL patterns have been defined for it (whether by the same or different entities). In some cases, an entity may simply create a link from its site to a particular item of web content (e.g., a document) and provide a label 107 as the anchor text of the link. In this case, a crawler extracts the label from the link and generates the appropriate annotation by using the label and the associated URL of the linked content. Each annotation is associated with the entity that provided the annotation. The trust rank TR for the associated entity can thereby be accessed as follows. First, given a document identifier (e.g., a URL), the annotations 106 associated with the document can be determined from the annotation database 140. For each identified annotation, the entity that provided the annotation can be determined, and that entity's trust rank TR retrieved from the trust database 190.

The search engine system 100 periodically updates the annotation database 140 by crawling web sites and extracting annotations therein, as well as receiving annotation files via the annotation interface 160.

Generated Trust Ranked Search Results

FIG. 4 presents a flowchart illustrating one process of generating trust ranked search results in accordance with an embodiment of the present invention; other variations of the steps or procedures of this process can be readily developed to achieve the same results.

The front end server 110 receives 402 a query from a user. The query includes at least one query term, and optionally includes one or more labels, as described above. The query is provided to the search engine 180.

For example, a query may be

digital camera more: professional review

where the terms "digital camera" are the query terms, and the token "label:" indicates that the following term is a label of interest. Thus, in this example, the user is searching for documents that have been labeled (by one or more individuals, such as other users, experts, etc.) as being professional reviews of digital cameras.

Another example would be the query

abortion label:statistics label:pro-choice.

In this example, the user would be searching for documents that relevant to abortion and had been labeled as either statistics or pro-choice related. The particular syntax used to indicate labels (as differentiated from query terms) is not limited to the use of "label:" and other tokens and formats can be used as well. For example, the query can take the form such as abortion label:(statistics pro-choice) where the terms in the parenthetical label are the labels of interest (as disjuncts).

The front end server 110 (or other module) includes a parser that performs the initial parsing of the query into query term and labels.

The search engine 180 retrieves 404 a set of documents that are relevant to the query term(s). The search engine 180 can use any type of information retrieval model for selecting and scoring relevant documents. One suitable information retrieval model is a link-based model, such as PageRank which is described in U.S. Pat. No. 6,285,999, and incorporated by reference herein.

The retrieved documents are ranked ordered in terms of an underlying (base) information retrieval score used by the search engine 180. For each document, the labels associated with the document are determined 406 from the annotation database 140 by matching the URL of the document to the URL patterns in the annotations. Thus, if the URL of a particular search result document matches the three different annotations for three different labels, each of these labels will be determined as associated with the document.

The query labels are then matched against the retrieved annotation labels, to identify which query labels apply to the document; this can be done by matching the query labels to the document's annotated labels. For each such annotation label matching a query label, the trust rank TR of entity that provided the annotation label is retrieved 408. A document may have been labeled with the same label by several different entities. For example, three different entities may have labeled a particular review of digital camera with the label "professional review" while a fourth labeled the same review with the label "negative review." If the query label included "professional review" then only the first three annotation labels would be deemed matched. It should be noted that steps 406 and 408 can performed in batch against all the search results, or iteratively for each search result, depending the desired programming model.

The trust ranks of the entities associated with the matching labels are aggregated 410 to create an aggregated trust rank. Thus, in this example, the trust ranks of the three different experts would be aggregated into a single trust rank associated with the particular label "professional review" for the review document. This aggregation is performed for each label associated with the document.

For example, assume that a particular document, such as a review of a Casio digital camera, has the following annotations associated therewith:

TABLE-US-00001 Label Entity Trust Rank "Professional Review" Phil Photo 8 "Professional Review" Earl Expert 6 "Professional Review" Chris Click 7 "Digital SLR" Phil Photo 8 "Digital SLR" Eddy Shooter 2 "Best buy" Betsy Buyer 3

The trust ranks of the entities could be aggregated with respect to these labels. The trust ranks can be aggregated using a weighting function. Aggregation using a linear weighting function would be:

TABLE-US-00002 Linear Aggregated Label Trust Rank "Professional Review" 21 "Digital SLR" 10 "Best buy" 3

This approach basically applies a fixed trust weight (e.g., 1) to the trust ranks prior to summing them.

The trust ranks TR can be aggregated in a number of different ways other than linear weighting. One variation is an asymptotic weighting function. For example, one approach here would be to sum the log of the trust ranks to form the aggregated trust rank. Another variation of the weighting function is to apply a weight that decays with the increasing number of instances of a particular label. For example, the trust ranks can be ordered by age of the associated annotation, and weighted to decrease (alternatively increase) the weight applied to the trust rank for the oldest annotation. Another way in which the trust ranks can be aggregated is by use of a sigmoid weighting function.

The foregoing are various examples of different ways that trust ranks can be aggregated; other way of combining the trust ranks of individual entities can be readily implemented, and these too are deemed to come within the concept of aggregation.

Once the trust ranks have been aggregated for each label that matches a query label, the aggregated trust ranks are applied 412 to the document's base information retrieval score. The aggregated trust ranks can be combined into a trust factor, using any type of aggregation function. For example, trust factor can be the sum total of the aggregated trust ranks. Second, the aggregated trust ranks can be aggregated asymptotically or via a sigmoid function to form the trust factor.

However the trust factor is determined from the aggregated trust ranks, it is then applied to the base information retrieval score, for example by multiplying with the trust factor. Other ways of applying the trust factor to the base information retrieval score as also possible. The resulting score is said to be trust-adjusted information retrieval score.

After (or as) the trust-adjusted information retrieval scores have been determined, the search engine 180 reranks 414 the search results by these adjusted scores. The result set is returned to the front end server 110, which in turn passes it to the client 103. The user can then review the search results.

The search engine system 100 is adapted to receive queries that do not include labels, and still provide trust based ranking. In this case, if certain annotations are applicable to a search result document, then the trust rank for the entities providing these annotations is retrieved, aggregated, and applied to the base information retrieval score of the document as well, and the documents are reranked accordingly.

In discussion of the various embodiments, examples have been discussed using search queries including a single query term and a single label, as well as examples referencing

individual documents. It is understood that the embodiments and invention are not so limited, and should be generally understood as equally operable with queries including a plurality of query terms and a plurality of labels. Thus any reference herein, including in the claims below, to various words in the singular noun form, such as query term, label, annotation, document, and so forth, are not intended to be limited to the singular, but should be read as including at least one or a plurality thereof, unless such a construction is expressly indicated as not intended or appropriate for the circumstances.

The present invention has been described in particular detail with respect to one possible embodiment. Those of skill in the art will appreciate that the invention may be practiced in other embodiments. First, the particular naming of the components, capitalization of terms, the attributes, data structures, or any other programming or structural aspect is not mandatory or significant, and the mechanisms that implement the invention or its features may have different names, formats, or protocols. Further, the system may be implemented via a combination of hardware and software, as described, or entirely in hardware elements. Also, the particular division of functionality between the various system components described herein is merely exemplary, and not mandatory; functions performed by a single system component may instead be performed by multiple components, and functions performed by multiple components may instead be performed by a single component.

Some portions of above description present the features of the present invention in terms of algorithms and symbolic representations of operations on information. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. These operations, while described functionally or logically, are understood to be implemented by computer programs. Furthermore, it has also proven convenient at times, to refer to these arrangements of operations as modules or by functional names, without loss of generality.

Unless specifically stated otherwise as apparent from the above discussion, it is appreciated that throughout the description, discussions utilizing terms such as "calculating" or "determining" or "identifying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system memories or registers or other such information storage, transmission or display devices.

Certain aspects of the present invention have been described using commands, mnemonics, tokens, formats, syntax, and other programming conventions. The particular selection of the names, formats, syntax, and the like are merely illustrative, and not limiting. Those of skill in the art can readily construct alternative names, formats, syntax rules, and so forth for defining context files and programming the operations a programmable search engine via context processing.

Certain aspects of the present invention include process steps and instructions described herein in the form of an algorithm. It should be noted that the process steps and instructions of the present invention could be embodied in software, firmware or hardware, and when embodied in software, could be downloaded to reside on and be operated from different platforms used by real time network operating systems.

The present invention also relates to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general-purpose computer selectively activated or reconfigured by a computer program stored on a computer readable medium that can be accessed by the computer. Such a computer program may be stored in a computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, CD-ROMs, magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus.

The algorithms and operations presented herein are not inherently related to any particular computer or other apparatus. Various general-purpose systems may also be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatus to perform the required method steps. The required structure for a variety of these systems will be apparent to those of skill in the art, along with equivalent variations. In addition, the present invention is not described with reference to any particular programming language. It is appreciated that a variety of programming languages may be used to implement the teachings of the present invention as described herein, and any references to specific languages are provided for disclosure of enablement and best mode of the present invention.

Finally, it should be noted that the language used in the specification has been principally selected for readability and instructional purposes, and may not have been selected to delineate or circumscribe the inventive subject matter. Accordingly, the disclosure of the present invention is intended to be illustrative, but not limiting, of the scope of the invention, which is set forth in the following claims.

* * * * *

source: archive.org

doc. Garuda Website