USPTO Patent Full-Text and Image Database

(**1** of **1**)

United States Patent	**7,716,225**
Dean , et al.	**May 11, 2010**

Ranking documents based on user behavior and/or feature data

Abstract

A system generates a model based on feature data relating to different features of a link from a linking document to a linked document and user behavior data relating to navigational actions associated with the link. The system also assigns a rank to a document based on the model.

Inventors:	**Dean; Jeffrey A.** (Palo Alto, CA)**, Anderson; Corin** (Mountain View, CA)**, Battle; Alexis** (Redwood City, CA)
Assignee:	**Google Inc.** (Mountain View, CA)
Family ID:	**42139440**
Appl. No.:	**10/869,057**
Filed:	**June 17, 2004**

Current U.S. Class:	**707/748**; 707/751
Current CPC Class:	G06F 40/134 (20200101); G06F 16/9535 (20190101); G06F 16/24578 (20190101); G06F 16/951 (20190101)
Current International Class:	G06F 7/00 (20060101); G06F 17/30 (20060101)
Field of Search:	;707/2,3,999.002,999.005

References Cited [Referenced By]

U.S. Patent Documents

5897627	April 1999	Leivian et al.
6006222	December 1999	Culliss

6014665	January 2000	Culliss
6078916	June 2000	Culliss
6088692	July 2000	Driscoll
6144944	November 2000	Kurtman et al.
6182068	January 2001	Culliss
6285999	September 2001	Page
6311175	October 2001	Adriaans et al.
6397211	May 2002	Cooper
6463430	October 2002	Brady et al.
6523020	February 2003	Weiss
6539377	March 2003	Culliss
6546388	April 2003	Edlund et al.
6546389	April 2003	Agrawal et al.
6714929	March 2004	Micaelian et al.
6738764	May 2004	Mao et al.
6782390	August 2004	Lee et al.
6799176	September 2004	Page
6804659	October 2004	Graham et al.
6836773	December 2004	Tamayo et al.
6947930	September 2005	Anick et al.
7007074	February 2006	Radwin
7058628	June 2006	Page
7065524	June 2006	Lee
7089194	August 2006	Berstis et al.
7100111	August 2006	McElfresh et al.
7231399	June 2007	Bem et al.
7356530	April 2008	Kim et al.

7398271	July 2008	Borkovsky et al.
7421432	September 2008	Hoelzle et al.
7523096	April 2009	Badros et al.
7584181	September 2009	Zeng et al.
2002/0083067	June 2002	Tamayo et al.
2002/0123988	September 2002	Dean et al.
2002/0184181	December 2002	Agarwal et al.
2003/0195877	October 2003	Ford et al.
2003/0197837	October 2003	Gyu Lee
2004/0088308	May 2004	Bailey et al.
2005/0071465	March 2005	Zeng et al.
2005/0071741	March 2005	Acharya et al.

Other References

Wang et al. "Ranking User's Relevance to a Topic through Link Analysis on Web Logs", WIDM'02, Nov. 8, 2002. cited by examiner .
Co-pending U.S. Appl. No. 10/706,991; Jeremy Bem et al.; "Ranking Documents Based on Large Data Sets"; filed Nov. 14, 2003, 38 pages. cited by other .
Co-pending U.S. Appl. No. 10/734,584; Jeremy Bem et al.; "Large Scale Machine Learning Systems and Methods"; filed Dec. 15, 2003, 35 pages. cited by other .
Co-pending U.S. Appl. No. 10/712,263; Jeremy Bem et al.; "Targeting Advertisements Based on Predicted Relevance of the Advertisements"; filed Nov. 14, 2003; 40 pages. cited by other .
Justin Boyan et al.; "A Machine Learning Architecture for Optimizing Web Search Engines"; Carnegie Mellon University; May 10, 1996; pp. 1-8. cited by other .
"Click Popularity-DirectHit Technology Overview"; http://www.searchengines.com/directhit.html; Nov. 10, 2003 (print date); 2 pages. cited by other .
http://www.httprevealer.com; "Creative Use of HttpRevealer--How does Google Toolbar Work?"; Apr. 19, 2004 (print date); pp. 1-6. cited by other .
J.H. Friedman, T. Hastie, and R. Tibshirani; "Additive Logistic Regression: a Statistical View of Boosting"; Dept. of Statistics, Stanford University Technical Report; Aug. 20, 1998. cited by other .
A.Y. Ng and M.I. Jordan; "On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes," in T. Dietterich, S. Becker and Z. Ghahramani (eds.), Advances in Neural Information Processing Systems 14, Cambridge, MA: MIT Press, 2002. cited by other .
F. Crestani, M. Lalmes, C. Van Rijsbergen and I. Campbell; "Is This Document Relevant? . . . Probably": A Survey of Probabilistic Models in Information Retrieval; ACM Computing Surveys, vol. 30, No. 4, Dec. 1998. cited by other .
Weis et al.: Rule-based Machine Learning Methods for Functional Prediction, Journal of AI Research, vol. 3, Dec. 1995, pp. 383-403. cited by other.

Primary Examiner: Alam; Hosain T
Assistant Examiner: Lin; Shew-Fen
Attorney, Agent or Firm: Harrity & Harrity, LLP

Claims

What is claimed is:

1. A method performed by one or more server devices, comprising: storing, in a memory associated with the one or more server devices, feature data associated with a plurality of first links, within a plurality of first source documents, that point to a plurality of first target documents, the feature data, for one of the plurality of first links, including one or more features of one of the plurality of first source documents that contains the one of the plurality of links, one or more features of one of the plurality of first target documents that is pointed to by the one of the plurality of links, and one or more features of the one of the plurality of first links; storing, in a memory associated with the one or more server devices, user

behavior data relating to user navigational activity with regard to the plurality of first source documents accessed by one or more users and the plurality of first links within the plurality of first source documents selected by the one or more users; training, using one or more processors of the one or more server devices and based on the feature data and the user behavior data, a model that identifies a probability that a particular link, with particular feature data, will be selected by a user, where training the model includes: analyzing the feature data associated with each of the plurality of first links that was selected by the one or more users and the feature data associated with each of the plurality of first links that was not selected by the one or more users to generate rules for the model; identifying, by one or more processors associated with the one or more server devices, a plurality of second links, within a plurality of second source documents, that point to a plurality of second target documents; determining, using one or more processors associated with the one or more server devices, feature data associated with each of the plurality of second links, the feature data, associated with one of the plurality of second links, including one or more features of the one of the plurality of second links, one or more features of one of the plurality of second source documents that contains the one of the plurality of second links, and one or more features of the one of the plurality of second target documents that is pointed to by the one of the plurality of second links; determining, using the model and based on the feature data, a probability that each of the plurality of second links will be selected by a user, where the determining includes: inputting, into the model, the feature data associated with the one of the plurality of second links, and outputting, by the model, the probability that the one of the plurality of second links will be selected by a user; calculating, using one or more processors associated with the one or more server devices, a rank for a particular target document of the plurality of second target documents based on the probability associated with one or more of the plurality of second links that point to the particular target document; and ordering the particular target document, with regard to at least one other document, based on the rank for the particular target document.

2. The method of claim 1, further comprising: obtaining data relating to the user navigational activity of the one or more users from client devices used by the one or more users.

3. The method of claim 1, where the user behavior data corresponds to a single user.

4. The method of claim 1, where the user behavior data corresponds to a class of users.

5. The method of claim 1, where the features associated with one of the plurality of first source documents include at least one of an entire address of the one of the plurality of first source documents, a portion of the address of the one of the plurality of first source documents, information regarding a web site associated with the one of the plurality of first source documents, a number of links in the one of the plurality of first source documents, presence of words in the one of the plurality of first source documents, presence of words in a heading of the one of the plurality of first source documents, a topical cluster with which the one of the plurality of first source documents is associated, or a degree to which a topical cluster associated with the one of the plurality of first source documents matches a topical cluster associated with a link.

6. The method of claim 1, where the features associated with one of the plurality of first links include at least one of a font size of anchor text associated with the one of the plurality of

first links, a position of the one of the plurality of first links within one of the plurality of first source documents, a position of the one of the plurality of first links in a list, a font color associated with the one of the plurality of first links, attributes of the one of the plurality of first links, a number of words in the anchor text associated with the one of the plurality of first links, actual words in the anchor text associated with the one of the plurality of first links, a determination of commerciality of the anchor text associated with the one of the plurality of first links, a type of the one of the plurality of first links, a context of words before or after the one of the plurality of first links, a topical cluster with which the anchor text of the one of the plurality of first links is associated, whether the one of the plurality of first links leads to a first target document on a same host or domain as one of the plurality of first source documents containing the one of the plurality of first links, or whether an address associated with the one of the plurality of first links embeds another address.

7. The method of claim 1, where the features associated with one of the plurality of first target documents include at least one of an entire address of the one of the plurality of first target documents, a portion of the address of the one of the plurality of first target documents, information regarding a web site associated with the one of the plurality of first target documents, whether the address of the one of the plurality of first target documents is on a same host as an address of a first source document that links to the one of the plurality of first target documents, whether the address of the one of the plurality of first target documents is associated with a same domain as the address of the first source document, words in the address of the one of the plurality of first target documents, or a length of the address of the one of the plurality of first target documents.

8. The method of claim 1, further comprising: generating a feature vector for each one of the plurality of first links based on the feature data associated with the one of the plurality of first links.

9. The method of claim 8, where analyzing the feature data associated with the plurality of first links and the instances where each of the plurality of the first links were selected by the one or more users and the instances where each of the plurality of first links were not selected by the one or more users includes: generating the rules for the model based on the instances where each of the plurality of the first links were selected by the one or more users and the instances where each of the plurality of first links were not selected by the one or more users and the feature vectors.

10. The method of claim 1, where the rules for the model comprise: a general rule applicable to a group of documents, and a specific rule applicable to a particular document.

11. The method of claim 1, further comprising: periodically updating the rules for the model based on changes in the user behavior data.

12. A method performed by one or more server devices, comprising: storing, in one or more memories associated with the one or more server devices, feature data associated with a plurality of first links within a plurality of first source documents that point to a plurality of first target documents, the feature data including features of the first source documents, features of the first target documents, and features of the first links; storing, in one or more memories associated with the one or more server devices, user behavior data relating to user

navigational activity with regard to the first links within the first source documents selected by one or more users; training, using one or more processors associated with the one or more server devices and based on the feature data associated with the feature data associated with the first links and the user behavior data relating to the first links, a model that identifies a probability that a particular link will be selected by a user, where training the model includes: analyzing the feature data associated with the first links that were selected by the one or more users and the feature data associated with the first links that were not selected by the one or more users to generate rules for the model; identifying a plurality of second links within a plurality of second source documents that point to a plurality of second target documents; determining feature data associated with the second links, the feature data associated with the second links including features of the second source documents, features of the second target documents, and features of the second links; determining, using the model, a probability that each of the second links will be selected using only the feature data associated with the second link as input to the model; assigning a weight to each of the second links based on the probability that the second link will be selected; assigning a rank to one of the second target documents based on the weights assigned to the second links that point to the one of the second target documents; and ordering the one of the second target documents, with regard to at least one other document, based on the rank assigned to the one of the second target documents.

13. The method of claim 12, further comprising: periodically updating the rules for the model based on changes to the user behavior data.

14. The method of claim 12, where the user behavior data corresponds to a single user.

15. The method of claim 12, where the user behavior data corresponds to a plurality of users.

16. One or more server devices, comprising: means for storing, in a memory, feature data associated with a plurality of links within source documents that point to target documents, the feature data including data associated with features of the source documents, data associated with features of the links, and data associated with features of the target documents, the data associated with the features of one of the source documents including at least one of an entire address of the source document, a portion of the address of the source document, information regarding a web site associated with the source document, a number of links in the source document, presence of words in the source document, presence of words in a heading of the source document, a topical cluster with which the source document is associated, or a degree to which a topical cluster associated with the source document matches a topical cluster associated with a link, the data associated with the features of one of the links including at least one of a font size of anchor text associated with the link, a position of the link within a source document, a position of the link in a list, a font color associated with the link, attributes of the link, a number of words in the anchor text associated with the link, actual words in the anchor text associated with the link, a determination of commerciality of the anchor text associated with the link, a type of the link, a context of words before or after the link, a topical cluster with which the anchor text of the link is associated, whether the link leads to a target document on a same host or domain, or whether an address associated with the link embeds another address, and the data associated with the features of one of the target documents including at least one of an

entire address of the target document, a portion of the address of the target document, information regarding a web site associated with the target document, whether the address of the target document is on a same host as an address of a source document that links to the target document, whether the address of the target document is associated with a same domain as the address of the source document, words in the address of the target document, or a length of the address of the target document; means for storing, in a memory, user behavior data relating to user navigational activity with regard to the source documents accessed by one or more users and the links within the source documents selected by the one or more users and the links within the source documents that were not selected by the one or more users; means for training, based on the feature data and instances where the links were selected by the one or more users and instances where the links were not selected by the one or more users, a model that identifies a probability that a link, with particular feature data, will be selected by a user, where the means for training includes: means for analyzing the feature data associated with the links that were selected by the one or more users and the feature data associated with the links that were not selected by the one or more users to generate rules for the model; means for identifying a particular link within a first document that points to a second document; means for determining the feature data associated with the particular link; means for determining, based on inputting the feature data into the model, a probability that the particular link will be selected by a user; means for assigning a weight to the particular link based on the probability that the particular link will be selected; means for assigning a rank to the second document based on the weight assigned to the particular link; and means for ordering the second document, with respect to at least one other document, based on the assigned rank.

17. The one or more server devices of claim 16, where the data associated with the features of the one of the source documents includes at least two of: the entire address of the source document, the portion of the address of the source document, the information regarding a web site associated with the source document, the number of links in the source document, the presence of words in the source document, the presence of words in a heading of the source document, the topical cluster with which the source document is associated, or the degree to which a topical cluster associated with the source document matches a topical cluster associated with a link.

18. The one or more server devices of claim 16, where the data associated with the features of the one of the links includes at least two of: the font size of anchor text associated with the link, the position of the link within a source document, the position of the link in a list, the font color associated with the link, the attributes of the link, the number of words in the anchor text associated with the link, the actual words in the anchor text associated with the link, the determination of commerciality of the anchor text associated with the link, the type of the link, the context of words before or after the link, the topical cluster with which the anchor text of the link is associated, whether the link leads to a target document on a same host or domain, or whether an address associated with the link embeds another address.

19. The one or more server devices of claim 6, where the data associated with the features of the one of the target documents includes at least two of: the entire address of the target document, the portion of the address of the target document, the information regarding a web site associated with the target document, whether the address of the target document is on a same host as an address of a source document that links to the target document,

whether the address of the target document is associated with a same domain as the address of the source document, the words in the address of the target document, or the length of the address of the target document.

Description

BACKGROUND

1. Field of the Invention

Systems and methods consistent with the principles of the invention relate generally to information retrieval and, more particularly, to creating a ranking function based on user behavior and/or feature data and using the ranking function to assign ranks to documents.

2. Description of Related Art

The World Wide Web ("web") contains a vast amount of information. Locating a desired portion of the information, however, can be challenging. This problem is compounded because the amount of information on the web and the number of new users inexperienced at web searching are growing rapidly.

Search engines attempt to return hyperlinks to web documents in which a user is interested. The goal of the search engine is to provide links to high quality documents to the user. Identifying high quality documents can be a tricky problem and is made more difficult by spamming techniques.

SUMMARY

According to one aspect, a method may include generating a model based on user behavior data associated with a group of documents. The method may also include assigning weights to links based on the model, where the links may include references from first documents to second documents in a set of documents, and assigning ranks to the second documents based on ranks of the first documents and the weights assigned to the links.

According to another aspect, a system may include means for generating a model based on user behavior data relating to links in a group of documents and feature data associated with the links. The system may also include means for assigning weights to references in a set of documents based on the model and means for assigning ranks to documents in the set of documents based on the weights assigned to the references.

According to yet another aspect, a system may include a memory and a processor. The memory may store user behavior data relating to a group of documents and feature data associated with the documents. The processor may generate a model based on the user behavior data and the feature data, assign a weight to a link from a first document to a second document in a set of documents based on the model, and assign a rank to the second document based on a rank of the first document and the weight assigned to the link.

According to a further aspect, a method may include generating a model based on different types of feature data associated with a group of documents. The method may also include assigning a weight to a link from a first document in a set of documents to a second document in the set of documents based on the model and assigning a rank to the second document based on a rank of the first document and the weight assigned to the link.

According to another aspect, a method may include generating a model based on feature data relating to features of a linking document and a link associated with the linking document, where the linking document references a linked document via the link. The method may also include assigning a rank to a document based on the model.

According to yet another aspect, a method may include generating a model based on feature data relating to different features of a link from a linking document to a linked document and user behavior data relating to navigational actions associated with the link and assigning a rank to a document based on the model.

According to a further aspect, a method may include generating a model based on user behavior data and document feature data and assigning ranks to documents based on the model.

According to another aspect, a method may include determining a weight for a link from a linking document to a linked document based on feature data associated with at least one of the link, the linking document, or the linked document, and assigning a rank to at least one of the linking document or the linked document based on the determined weight for the link.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, explain the invention. In the drawings,

FIG. 1 is an exemplary diagram of a simple linked database;

FIG. 2 is a diagram of an exemplary information retrieval network in which systems and methods consistent with the principles of the invention may be implemented;

FIG. 3 is an exemplary diagram of a client or server according to an implementation consistent with the principles of the invention;

FIG. 4 is a functional block diagram of an exemplary modeling system according to an implementation consistent with the principles of the invention;

FIG. 5 is a flowchart of exemplary processing for determining document ranks according to an implementation consistent with the principles of the invention;

FIG. 6 is a flowchart of exemplary processing for presenting search results according to an implementation consistent with the principles of the invention; and

FIG. 7 is a diagram of an exemplary linked database.

DETAILED DESCRIPTION

The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention.

Overview

A linked database may include documents with links among them. A "document," as the term is used herein, is to be broadly interpreted to include any machine-readable and machine-storable work product. A document may include, for example, an e-mail, a file, a combination of files, one or more files with embedded links to other files, a news group posting, a blog, a web advertisement, etc. In the context of the Internet, a common document is a web page. Web pages often include textual information and may include embedded information (such as meta information, images, hyperlinks, etc.) and/or embedded instructions (such as Javascript, etc.).

A "link," as the term is used herein, is to be broadly interpreted to include any reference to/from a document from/to another document or another part of the same document. A "forward link" (sometimes referred to as an "outbound link") is used to refer to a link from a source document to a target document from the standpoint of the source document. A "backward link" (sometimes referred to as an "inbound link") is used to refer to a link from a target document to a source document from the standpoint of the source document.

FIG. 1 is an exemplary diagram of a simple linked database that includes three documents: document A, document B, and document C. As shown, document A includes a single forward link to document C and two backward links from documents B and C. Document B includes a single forward link to document A. Document C includes a single forward link to document A and a single backward link from document A.

Systems and methods consistent with the principles of the invention may provide a reasonable surfer model that indicates that when a surfer accesses a document with a set of links, the surfer will follow some of the links with higher probability than others. This reasonable surfer model reflects the fact that not all of the links associated with a document are equally likely to be followed. Examples of unlikely followed links may include "Terms of Service" links, banner advertisements, and links unrelated to the document.

Exemplary Information Retrieval Network

FIG. 2 is an exemplary diagram of a network 200 in which systems and methods consistent with the principles of the invention may be implemented. Network 200 may include multiple clients 210 connected to multiple servers 220-240 via a network 250. Network 250 may include a local area network (LAN), a wide area network (WAN), a telephone network, such as the Public Switched Telephone Network (PSTN), an intranet, the Internet, a memory device, or a combination of networks. Two clients 210 and three servers 220-240 have been illustrated as connected to network 250 for simplicity. In practice, there may be more or fewer clients and servers. Also, in some instances, a client may perform the functions of a server and/or a server may perform the functions of a client.

Clients 210 may include client entities. An entity may be defined as a device, such as a personal computer, a wireless telephone, a personal digital assistant (PDA), a lap top, or another type of computation or communication device, a thread or process running on one of these devices, and/or an object executable by one of these devices. Servers 220-240 may include server entities that gather, process, search, and/or maintain documents in a manner consistent with the principles of the invention. Clients 210 and servers 220-240 may connect to network 250 via wired, wireless, and/or optical connections.

In an implementation consistent with the principles of the invention, server 220 may include a search engine 225 usable by clients 210. Server 220 may crawl a corpus of documents (e.g., web pages), index the documents, and store information associated with the documents in a repository of crawled documents. Servers 230 and 240 may store or maintain documents that may be crawled by server 220. While servers 220-240 are shown as separate entities, it may be possible for one or more of servers 220-240 to perform one or more of the functions of another one or more of servers 220-240. For example, it may be possible that two or more of servers 220-240 are implemented as a single server. It may also be possible for a single one of servers 220-240 to be implemented as two or more separate (and possibly distributed) devices.

Exemplary Client/Server Architecture

FIG. 3 is an exemplary diagram of a client or server entity (hereinafter called "client/server entity"), which may correspond to one or more of clients 210 and servers 220-240, according to an implementation consistent with the principles of the invention. The client/server entity may include a bus 310, a processor 320, a main memory 330, a read only memory (ROM) 340, a storage device 350, an input device 360, an output device 370, and a communication interface 380. Bus 310 may include a path that permits communication among the elements of the client/server entity.

Processor 320 may include a conventional processor, microprocessor, or processing logic that interprets and executes instructions. Main memory 330 may include a random access memory (RAM) or another type of dynamic storage device that stores information and instructions for execution by processor 320. ROM 340 may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by processor 320. Storage device 350 may include a magnetic and/or optical recording medium and its corresponding drive.

Input device 360 may include a conventional mechanism that permits an operator to input information to the client/server entity, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms, etc. Output device 370 may include a conventional mechanism that outputs information to the operator, including a display, a printer, a speaker, etc. Communication interface 380 may include any transceiver-like mechanism that enables the client/server entity to communicate with other devices and/or systems. For example, communication interface 380 may include mechanisms for communicating with another device or system via a network, such as network 250.

As will be described in detail below, the client/server entity, consistent with the principles of the invention, performs certain searching-related operations. The client/server entity may perform these operations in response to processor 320 executing software instructions contained in a computer-readable medium, such as memory 330. A computer-readable medium may be defined as a physical or logical memory device and/or carrier wave.

The software instructions may be read into memory 330 from another computer-readable medium, such as data storage device 350, or from another device via communication interface 380. The software instructions contained in memory 330 may cause processor 320 to perform processes that will be described later. Alternatively, hardwired circuitry may be used in place of or in combination with software instructions to implement processes consistent with the principles of the invention. Thus, implementations consistent with the principles of the invention are not limited to any specific combination of hardware circuitry and software.

Exemplary Modeling System

FIG. 4 is an exemplary functional block diagram of a modeling system 400 according to an implementation consistent with the principles of the invention. According to one implementation, one or more of the functions described below may be performed by server 220. According to another implementation, one or more of these functions may be performed by an entity separate from server 220, such as a computer associated with server 220 or one of servers 230 or 240.

Modeling system 400 may include model generating unit 410 and model applying unit 420 connected to a repository 430. Repository 430 may include a physical or logical memory device that stores information associated with documents that were crawled and indexed by, for example, server 220 (FIG. 2) or an entity separate from server 220. For example, repository 430 may store documents that form a linked database.

The document information may also, or alternatively, include feature data associated with features of documents ("source documents"), links in the source documents, and possibly documents pointed to by these links ("target documents"). Examples of features associated with a link might include the font size of the anchor text associated with the link; the position of the link (measured, for example, in a HTML list, in running text, above or below the first screenful viewed on an 800.times.600 browser display, side (top, bottom, left, right) of document, in a footer, in a sidebar, etc.); if the link is in a list, the position of the link in the list; font color and/or attributes of the link (e.g., italics, gray, same color as background, etc.); number of words in anchor text associated with the link; actual words in the anchor text associated with the link; commerciality of the anchor text associated with the link; type of the link (e.g., image link); if the link is associated with an image (i.e., image link), the aspect ratio of the image; the context of a few words before and/or after the link; a topical cluster with which the anchor text of the link is associated; whether the link leads somewhere on the same host or domain; if the link leads to somewhere on the same domain, whether the link URL is shorter than the referring URL; and/or whether the link URL embeds another URL (e.g., for server-side redirection). This list is not exhaustive and may include more, less, or different features associated with a link.

Examples of features associated with a source document might include the URL of the source document (or a portion of the URL of the source document); a web site associated with the source document; a number of links in the source document; the presence of other words in the source document; the presence of other words in a heading of the source document; a topical cluster with which the source document is associated; and/or a degree to which a topical cluster associated with the source document matches a topical cluster associated with anchor text of a link. This list is not exhaustive and may include more, less, or different features associated with a source document.

Examples of features associated with a target document might include the URL of the target document (or a portion of the URL of the target document); a web site associated with the target document; whether the URL of the target document is on the same host as the URL of the source document; whether the URL of the target document is associated with the same domain as the URL of the source document; words in the URL of the target document; and/or the length of the URL of the target document. This list is not exhaustive and may include more, less, or different features associated with a target document.

Repository 430 may also store user behavior data associated with documents. The user behavior data may include, for example, information concerning users who accessed the documents, such as navigational actions (e.g., what links the users selected, addresses entered by the users, forms completed by the users, etc.), the language of the users, interests of the users, query terms entered by the users, etc. In an alternate implementation, the user behavior data may be stored external to repository 430 and provided as an input to model generating unit 410.

The user behavior data might be obtained from a web browser or a browser assistant associated with clients 210. A browser assistant may include executable code, such as a plug-in, an applet, a dynamic link library (DLL), or a similar type of executable object or process that operates in conjunction with (or separately from) a web browser. The web browser or browser assistant might send information to server 220 concerning a user of a client 210.

For example, the web browser or browser assistant may record data concerning the documents accessed by the user and the links within the documents (if any) the user selected. Additionally, or alternatively, the web browser or browser assistant may record data concerning the language of the user, which may be determined in a number of ways that are known in the art, such as by analyzing documents accessed by the user. Additionally, or alternatively, the web browser or browser assistant may record data concerning interests of the user, which may be determined, for example, from the favorites or bookmark list of the user, topics associated with documents accessed by the user, or in other ways that are known in the art. Additionally, or alternatively, the web browser or browser assistant may record data concerning query terms entered by the user. The web browser or browser assistant may send this data for storage in repository 430.

In one implementation, repository 430 may also store data specific to certain classes of users. For example, repository 430 might store data corresponding to the language of a user class compared to the language of the source document and the language of the link and/or target document. Repository 430 might also/alternatively store data corresponding to a topical cluster associated with the interests of the user class compared to a topical cluster associated with the source and/or target documents. Repository 430 might also/alternatively store data corresponding to a set of query words associated with the user class compared to the content of the source and/or target documents.

Model generating unit 410 may determine link data for the various document links based on information stored in repository 430. The link data associated with a particular link might include how often the link was selected and the feature data associated with the link (including features associated with the link, the source document containing the link, and the target document referenced by the link). In one implementation, model generating unit 410 may analyze the user behavior data in repository 430 to generate positive and negative instances for the links. For example, model generating unit 410 may consider selection of a particular link in a document as a positive instance for that link and non-selection of the other links in the document as negative instances for those links. In the case where no links in a document are selected, model generating unit 410 may consider non-selection of the links as negative instances for the links.

To illustrate this, assume that a document W includes forward links to documents X, Y, and Z and the user behavior data indicates that the following selections occurred (W, X), (W, X), and (W, Z), where (W, X) means that the link from document W to document X was selected. In this case, three positive instances occurred: two for the link from W to X and one for the link from W to Z; and six negative instances occurred: one for the link from W to X, three for the link from W to Y, and two for the link from W to Z.

Model generating unit 410 may generate a feature vector for each of the links based on the link data. The feature vector associated with a link may be a function of the feature data associated with the link (including features associated with the link, the source document containing the link, and the target document referenced by the link). For example, the feature vector might indicate the font size of the anchor text associated with the link, the web site associated with the source document, the URL of the target document, and/or other feature data, as described above. The feature vector may take different forms.

Model generating unit 410 may then build a model of whether a link is likely to be selected based on the link's positive and negative instances, the link's associated feature vector, and possibly other information in repository 430. The model may be considered a dynamic model in that it is built from data that changes over time. Model generating unit 410 may use a conventional technique, such as a naive bayes, a decision tree, logistic regression, or a hand-tailored approach, to form the model.

The model may include general rules and document-specific rules. Model generating unit 410 may learn the general rules based on the user behavior data and the feature vector associated with the various links. For example, model generating unit 410 may determine how users behaved when presented with links with different associated feature data. From this information, model generating unit 410 may generate general rules of link selection.

For example, model generating unit 410 may generate a rule that indicates that links with anchor text greater than a particular font size have a higher probability of being selected than links with anchor text less than the particular font size. Additionally, or alternatively, model generating unit 410 may generate a rule that indicates that links positioned closer to the top of a document have a higher probability of being selected than links positioned toward the bottom of the document. Additionally, or alternatively, model generating unit 410 may generate a rule that indicates that when a topical cluster associated with the source document is related to a topical cluster associated with the target document, the link has a higher probability of being selected than when the topical cluster associated with the source document is unrelated to the topical cluster associated with the target document. These rules are provided merely as examples. Model generating unit 410 may generate other rules based on other types of feature data or combinations of feature data.

Model generating unit 410 may learn the document-specific rules based on the user behavior data and the feature vector associated with the various links. For example, model generating unit 410 may determine how users behaved when presented with links of a particular source document. From this information, model generating unit 410 may generate document-specific rules of link selection.

For example, model generating unit 410 may generate a rule that indicates that a link positioned under the "More Top Stories" heading on the cnn.com web site has a high probability of being selected. Additionally, or alternatively, model generating unit 410 may generate a rule that indicates that a link associated with a target URL that contains the word "domainpark" has a low probability of being selected. Additionally, or alternatively, model generating unit 410 may generate a rule that indicates that a link associated with a source document that contains a popup has a low probability of being selected. Additionally, or alternatively, model generating unit 410 may generate a rule that indicates that a link associated with a target domain that ends in ".tv" has a low probability of being selected. Additionally, or alternatively, model generating unit 410 may generate a rule that indicates that a link associated with a target URL that includes multiple hyphens has a low probability of being selected. These rules are provided merely as examples. Model generating unit 410 may generate other document-specific rules.

Model applying unit 420 may assign weights to links based on the dynamic model generated by model generating unit 410. The weight of a link may be a function of the rules applicable to the feature data associated with the link. A link's weight may reflect the probability that the link will be selected.

Model applying unit 420 may then assign ranks to documents based on the ranks of their linking documents (i.e., those documents with forward links to the documents), such as described in U.S. Pat. No. 6,285,999, entitled "METHOD FOR NODE RANKING IN A LINKED DATABASE," the contents of which are incorporated herein by reference. In implementations consistent with the principles of the invention, however, the document ranks are modified based on the dynamic weighting model described herein. For example, ranks for the documents may be generated according to the function:

.function..times..times..function..times..times..times..function..times. ##EQU00001## where A is a document for which a rank is being generated, B.sub.1, . . . , B.sub.n are documents connected by backward links to document A, r(B.sub.1), . . . , r(B.sub.n) are the ranks of the B documents, w.sub.1, . . . , w.sub.n are the weights assigned to the backward links,

|B.sub.1|, . . . , |B.sub.n| are the number of forward links associated with the B documents, .alpha. is a constant in the interval [0, 1], and N is the total number of documents in the linked database. The rank of a document may be interpreted as the probability that a reasonable surfer will access the document after following a large number of forward links.

Model applying unit 420 may store the document ranks so that when the documents are later identified as relevant to a search query by a search engine, such as search engine 225, the ranks of the documents may be quickly determined. Since links periodically appear and disappear and user behavior data is constantly changing, model applying unit 420 may periodically update the weights assigned to the links and, thus, the ranks of the documents.

Exemplary Processing

FIG. 5 is a flowchart of exemplary processing for determining document ranks according to an implementation consistent with the principles of the invention. Processing may begin with the storing of information, such as user behavior data and feature data, in a repository. As described above, documents maintained by servers connected to a network or a combination of networks, such as the Internet, may be crawled and indexed. User behavior data and feature data may also be determined for all or a subset of the documents. The user behavior data may include, for example, information concerning users who accessed the documents, such as navigational actions of the users, the language of the users, interests of the users, query terms entered by the users, etc. The feature data may include features associated with source documents, links in the source documents, and possibly target documents pointed to by the links. Examples of feature data have been provided above.

Positive and negative instances for the links may be determined based on information in the repository (act 510). For example, the user behavior data may be analyzed to determine which links were selected and which links were not selected. Selection of a link in a document may be identified as a positive instance for that link and non-selection of the other links in the document may be identified as negative instances for those links. Also, selection of no links in a document may be identified as negative instances for all of the links in the document.

Feature vectors may be generated for the links (act 520). The feature vector associated with a link may be a function of the feature data for the link. For example, the feature vector might include fields that provide feature data, such as the feature data described above, associated with the link.

A model may then be generated (act 530). The model may be generated based on the links' positive and negative instances, the links' associated feature vectors, and possibly other information in the repository. As described above, the model may include general rules and document-specific rules. The general rules are applicable across documents and the document-specific rules are applicable to specific documents.

Weights may be generated for links based on the model (act 540). The weight of a link may be a function of the rules applicable to the feature data associated with the link. A link's weight may reflect the probability that the link will be selected.

Document ranks may then be determined based on the link weights (act 550). One possible function for determining the rank of a document has been described above in Eqn. 1. The rank of a document may be interpreted as the probability that a reasonable surfer will access the document after following a number of forward links.

FIG. 6 is a flowchart of exemplary processing for presenting search results according to an implementation consistent with the principles of the invention. Processing may begin with a user providing search terms as a search query for searching a document corpus. In one implementation, the document corpus includes documents available from the Internet and the vehicle for searching this corpus is a search engine, such as search engine 225 (FIG. 2). The user may provide the search query via web browser software on a client, such as client 210 (FIG. 2).

The search query may be received by the search engine and used to identify documents (e.g., web pages) related to the search query (acts 610 and 620). A number of techniques exist for identifying documents related to a search query. One such technique might include identifying documents that contain the search terms as a phrase. Another technique might include identifying documents that contain the search terms, but not necessarily together. Other techniques might include identifying documents that contain less than all of the search terms, or synonyms of the search terms. Yet other techniques are known to those skilled in the art.

Ranks may be determined for the identified documents based on the model (act 630). In one implementation, document ranks are pre-calculated and determining the ranks of the documents may include simply looking up the document ranks. In another implementation, the document ranks are not pre-calculated. In this case, the ranks of the documents can be determined based on the model rules, as described above with regard to the processing of FIG. 5.

The documents may then be sorted based on their ranks (act 640). In practice, however, a document's rank may be one of several factors used to determine an overall rank for the document. The documents may then be sorted based on their overall ranks.

Search results may be formed based on the sorted documents (act 650). In an implementation consistent with the principles of the invention, the search results may include links to the documents, possibly including a textual description of the links. In another implementation, the search results may include the documents themselves. In yet other implementations, the search results may take other forms.

The search results may be provided as a HTML document, similar to search results provided by conventional search engines. Alternatively, the search results may be provided according to a protocol agreed upon by the search engine and the client (e.g., Extensible Markup Language (XML)).

Example

FIG. 7 is a diagram of an exemplary linked database that includes three documents: document A, document B, and document C. As shown, document A includes two forward links to documents B and C and a single backward link from document C. Document B includes a single forward link to document C and a single backward link from document A. Document C includes a single forward link to document A and two backward links from documents A and B.

Assume that the backward link from document C to document A has an associated weight of 0.5, the backward link from document A to document B has an associated weight of 0.6, the backward link from document B to document C has an associated weight of 0.9, and the backward link from document A to document C has an associated weight of 0.4. The weights for these backward links may be determined based on the user behavior data and feature data associated with the links, as described above. While a typical value for a is 0.1, assume that a is 0.5 for this example.

Based on Eqn. 1, the ranks of documents A-C may be determined as follows:

.function..times..times..times..times..times..times..function..times..times..times..times..function..times..function. ##EQU00002## The solution in this case is r(A).apprxeq.0.237, r(B).apprxeq.0.202, and r(C).apprxeq.0.281.

CONCLUSION

Systems and methods consistent with the principles of the invention may determine ranks for documents based on the ranks of linking documents (i.e., those documents with forward links to the documents) and a dynamic link weighting model. The model may be used to adjust the contribution that various links make in the ranking process.

The foregoing description of preferred embodiments of the present invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention.

For example, while series of acts have been described with regard to FIGS. 5 and 6, the order of the acts may be modified in other implementations consistent with the principles of the invention. Further, non-dependent acts may be performed in parallel.

In one implementation, server 120 may perform most, if not all, of the acts described with regard to the processing of FIGS. 5 and 6. In another implementation consistent with the principles of the invention, one or more, or all, of the acts may be performed by another entity, such as another server 130 and/or 140 or client 110.

It has been described that ranks are determined for documents based on user behavior data. According to one implementation, the user behavior data is associated with a set of users. According to another implementation, the user behavior data is associated with a subset, or class, of users. In this case, the weights assigned to the links may be tailored to the user class. According to yet another implementation, the user behavior data is associated with a single user. In this case, the weights assigned to the links may be tailored to the user.

It will also be apparent to one of ordinary skill in the art that aspects of the invention, as described above, may be implemented in many different forms of software, firmware, and hardware in the implementations illustrated in the figures. The actual software code or specialized control hardware used to implement aspects consistent with the principles of the invention is not limiting of the present invention. Thus, the operation and behavior of the aspects were described without reference to the specific software code--it being understood that one of ordinary skill in the art would be able to design software and control hardware to implement the aspects based on the description herein.

No element, act, or instruction used in the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article "a" is intended to include one or more items. Where only one item is intended, the term "one" or similar language is used. Further, the phrase "based on" is intended to mean "based, at least in part, on" unless explicitly stated otherwise.

Doc. Garuda Website - Ranking documents based on user behavior and/or feature data
Source: archive.org