| (54) Title | Traffic prediction for web sites |
| (51) International Patent Classification(s) | H04L 12/26 (2006.01)  
G06F 15/173 (2006.01)  
G06F 15/16 (2006.01) |
| (21) Application No: | 2006292261  
(22) Date of Filing: | 2006.09.18 |
| (87) WIPO No: | WO07/035727 |
| (30) Priority Data |  
| (31) Number | 11/228,217  
| (32) Date | 2005.09.19  
| (33) Country | US |
| (43) Publication Date: | 2007.03.29 |
| (44) Accepted Journal Date: | 2010.05.13 |
| (71) Applicant(s) | Google Inc. |
| (72) Inventor(s) | Bennett, Victor |
| (74) Agent / Attorney | Shelston IP, Level 21 60 Margaret Street, Sydney, NSW, 2000 |
| (56) Related Art | US 6278966 B1  
US 6338066 B1 |
A traffic prediction component may automatically generate predicted traffic profiles for web sites based on tags that characterize the sites. An initial set of tags can be selected for a web site based on a set of predefined rules. An initial traffic profile may be selected based on the initial set of tags. The predicted profile of user traffic is then generated based on the initial set of tags and on the initial traffic profile.
TRAFFIC PREDICTION FOR WEB SITES

Field of the Invention

Implementations described herein generally relate to the prediction of user traffic for online sites.

Description of the Related Art

In an increasingly networked world, consumers frequently use online resources to access information. For example, search engines are a popular tool through which users enter a search query describing information of interest and receive back documents or links to documents that relate to the search query. Search engines may be tailored to a particular area of interest, such as a general web search engine, a news search engine, or a shopping search engine.

It may be desirable for an operator of an online resource, such as a search engine, to be able to intelligently predict information about the type of users that are likely to visit a particular web site. This type of “traffic prediction” can be used, for example, when the search engine displays advertisements that link to a particular web site. If the search engine provider knows a general profile that characterizes traffic that is likely to click on a particular advertisement or visit the web site referenced by the advertisement, the search engine provider may be able to spot invalid or non-genuine user activity that is not reflective of true user interest. Distinguishing invalid user activity (i.e., traffic) from genuine user activity can be particularly important when, for example, the search engine provider charges advertisers based on the traffic that is referred to the advertiser’s web site. In this situation, the search engine provider may like to be able to distinguish invalid user activity from genuine user activity and only charge the advertiser for the genuine user activity.

Predicting user activity can be useful in a number of online contexts in addition to the display of advertisements and outside of the context of a search engine. For example, a web site designer may wish to predict how a proposed change to a web site will affect user activity at or to the web site.

SUMMARY

According to a first aspect of the invention, there is provided a computer-implemented method of predicting a profile of user traffic at a web site, the method comprising:

characterizing the web site based on at least one feature of the web site by using a plurality of tags that describe the at least one feature of the web site;

selecting an initial traffic profile, that characterizes behaviors of users of another web site during a time that the users are accessing the other web site, based on the characterization of the web site; and

generating a predicted profile of user traffic at the web site based on the characterization of the web site and the initial traffic profile.

According to a second aspect of the invention, there is provided a computer-implemented method of modeling a web site, the method comprising:

assigning an initial set of tags to the model based on features of the web site, each of the tags describing at least one feature of the web site;

determining an initial traffic profile for the web site based on the assigned initial set of tags,

where the initial profile characterizes behaviors of users of another web site during a time that the users
are accessing the other web site; and

  recursively updating the set of tags and the traffic profile to obtain a final predicted traffic
profile for the web site.

According to a third aspect of the invention, there is provided a computer-readable memory device
containing instructions for execution by processing logic, the computer-readable memory device
comprising:

  instructions for characterizing a web site based on at least one feature of the web site by using
a plurality of tags that describe the at least one feature of the web site;

  instructions for selecting an initial traffic profile, that characterizes behaviors of users of
another web site during a time that the users are accessing the other web site, based on the
characterization of the web site; and

  instructions for generating a profile predicting user traffic at the web site based on the
characterization of the web site and the initial traffic profile.

According to a fourth aspect of the invention, there is provided a device comprising:

  means for assigning an initial set of tags to a model of the web site based on features of the
web site, each of the tags describing at least one feature of the web site;

  means for determining an initial traffic profile for the web site based on the assigned initial set
of tags, where the initial traffic profile characterizes behaviors of users of another web site during a
time that the users are accessing the other web site; and

  means for recursively updating the set of tags and the traffic profile to obtain a final predicted
traffic profile for the web site.

According to a fifth aspect of the invention, there is provided a computer-implemented method of
predicting a profile of user traffic at a web site, the method comprising:

  determining an initial set of tags that characterize at least one feature of the web site based on
a set of predefined rules, the initial set of tags including a tag that relates to an intended use of the web
site or a tag that relates to an industry classification of the web site;

  comparing the initial set of tags to tags determined for other web sites to locate a second web
site that has tags that are similar to the initial set of tags;

  using a traffic profile of the second web site as an initial traffic profile for the web site;

  recursively updating the initial traffic profile and the initial set of tags to generate a model of
the web site;

  using the model to predict traffic at the web site;

  comparing the predicted traffic to observed traffic; and

  outputting a signal when the observed traffic does not match the predicted traffic.

Unless the context clearly requires otherwise, throughout the description and the claims, the
words "comprise", "comprising", and the like are to be construed in an inclusive sense as opposed to an
exclusive or exhaustive sense; that is to say, in the sense of "including, but not limited to".

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 a diagram conceptually illustrating, at high level, an exemplary implementation
as will be described in more detail below, generates user traffic profiles for web sites. Instead of being implemented on a server connected to network 240, traffic prediction component 225 may more generally be implemented on any computing device, whether or not connected to a network. Network 240 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, or a combination of networks. Three clients 210 and two servers 220 and 230 are illustrated as connected to network 240 for simplicity. In practice, there may be more clients and/or servers. Also, in some instances, a client may perform one or more functions of a server and a server may perform one or more functions of a client.

A client 210 may include 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. Users of clients 210 may access or receive information from servers 220 and 230.

EXEMPLARY COMPUTING DEVICE ARCHITECTURE

Fig. 3 is an exemplary diagram of computing device 300, such as one of clients 210, server 220, or server 230. Computing device 300 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 components of computing device 300.

Processor 320 may include any type of processor, microprocessor, or processing logic that may interpret and execute instructions. Main memory 330 may include a random access memory (RAM) or another type of dynamic storage device that may store information and instructions for execution by processor 320. ROM 340 may include a ROM device or another type of static storage device that may store 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 mechanism that permits a user to input information to computing device 300, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms, etc. Output device 370 may include a mechanism that outputs information to the user, including a display, a printer, a speaker, etc. Communication interface 380 may include any transceiver-like mechanism that enables computing device 300 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 240.

Traffic prediction component 235 may be implemented in software and stored in a computer-readable medium, such as memory 330. A computer-readable medium may be defined as one or more physical or logical memory devices and/or carrier waves.

The software instructions defining traffic prediction component 235 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 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 present invention. Thus, implementations consistent with the principles of the invention are not limited to any specific combination of hardware circuitry and software.

**TRAFFIC PREDICTION COMPONENT 235**

Traffic prediction component 235 may generally operate to categorize web sites in a way consistent with differences in the web sites that lead to or are correlated with different traffic profiles for the sites. A traffic profile for a site, as this phrase is used herein, refers to any set of data that tends to characterize users of the web site or how the users tend to behave at the site. Examples of factors that can be included in a traffic profile include: how long a typical user will spend at the site, how "deep" or how many links a typical user will click on within the site, the screen size of a typical user, the type of browser of a typical user, or user demographic information.

Fig. 4 is a flow chart illustrating exemplary operations of traffic prediction component 235 in generating a traffic prediction model for a web site. A web site may be described (i.e., categorized) by a plurality of labels or tags that describe various aspects of the site. An initial set of tags that describes the site may be heuristically determined for the web site (act 401). For example, a predetermined set of rules may be used to generate the initial set of tags.

Based on the initial tags that describe the web site, an initial traffic profile for the site may be determined (act 402). The initial traffic profile may be based on a traffic profile observed at a site that is similar to the site being modeled. The initial tags may be used to determine which sites are similar to the site being modeled.

The web site tags and the traffic profile may then be recursively modified until the traffic profile and the site tags stabilize (act 403). The final set of site tags and the final traffic profile can be thought of as a model of the site, as generated by traffic prediction component 235. The traffic profile described by the model represents the predicted traffic profile of the web site.

Fig. 5 is a diagram conceptually illustrating the generation of a model 500 for a web site, such as a model generated by the operations shown in Fig. 4. Model 500 may include a set of descriptive tags 510 that describe or categorize the site. Descriptive tags 510 may generally be generated with the goal of categorizing the web site in such a way that different web sites are correlated with different traffic profiles. Numerous different tags are possible. Example tags that may be used include tags that indicate the primary language used by the site (e.g., English, French, etc.); tags that categorize the site by type or intent, such as a shopping site or news site; tags that categorize the site by industry classification, such as a site targeted to computer enthusiasts or a site targeted to home contractors; tags that relate to the geographical location of the site's host; tags that indicate the average price of a product for sale (for shopping sites); or tags that characterize the hyper-text markup language (HTML) layout of the site, such as tags that relate to the portion of the visual layout of the site that is text or graphics. One of ordinary skill in the art will recognize that additional tags could be used.

The initial descriptive tags for a site may be set based on heuristic signals 515. Heuristic signals 515 may generally include predetermined rules that are used to set one or more of the tags in the set of descriptive tags 510. Heuristic signals 515 may include, for example, directory information signals 516, which may set the tags to reflect a classification of the web site based on a known directory of web sites. One such known directory of web sites that may be used is The Open Directory.
Project (http://dmoz.org/), which is a human-compiled directory of the web that classifies web sites using a hierarchical directory structure. Heuristic signals 515 may additionally include signals 517 based on a lexical analysis of the web page. Signals 517 may, for example, be used to set tags that characterize the predominant language used by the site or the type of words used by the site (e.g., an indication of whether the site uses many technical terms). Heuristic signals 515 may additionally include signals 518 based on geographical features of the web site, such as the location of the web site host, as determined by the IP address of the web site. One or more tags may be set that define the geographical location of the web site. The above examples of heuristic signals 515 are exemplary, one of ordinary skill in the art will recognize that other heuristic signals could be used to generate tags. For example, heuristic signals based on text from sites that link to the site being modeled can be used to generate tags. For example, a tag may be set if a site is linked to by numerous known shopping sites (or other types of sites). As another example of heuristic signals 515, the design of a web page may be matched with whether the web page is exemplary of a particular type of site. For example, a web site with many web pages that include a picture, a description, and a price is likely to be an online store.

Model 500 may additionally include a traffic profile 520 that describes or categorizes the site. Traffic profile 520 may include values or tags used to characterize expected traffic, including the behavior of the traffic, at a site. Traffic profile 520 may include values that describe, for example, how long a typical user will spend at the site, how "deep" or how many links a typical user will click on within the site, the screen size of a typical user, the type of browser of a typical user, and demographic information of a typical user.

Traffic profile 520 may be initially set to reflect an estimated traffic profile 521. Initial traffic profile 521 may be set based on a known traffic profile of traffic at another site that has a similar set of descriptive tags as the initial set of descriptive tags 510. For example, if the site being modeled is a shopping site, the initial estimated traffic profile 521 may be a known (i.e., observed) traffic profile from another shopping site.

The recursive analysis of act 403 may be performed to update the set of descriptive tags 510 and traffic profile 520. In one implementation, descriptive tags 510 may be updated stochastically. For instance, occurrences of tags (based on traffic profile 520) that may potentially apply to the web site are tracked, and if the tags occur often enough, then the tags are considered to be statistically likely to apply to the web site. Conversely, tags that occur infrequently may be considered to be spurious tags that do not apply to the web site.

As an example of training of model 500, consider a web site comparing the virtues of different vendors of gasoline. This site is a product comparison site, although user behavior on this site is likely to be significantly different than user behavior on a product comparison site that compares products that the user is likely to directly buy online. That is, the user is not browsing the gasoline vendor comparison site intending to find products to buy, but is instead browsing to view the information available on the site. In this example, the traffic profile for the site is likely to be similar to sites that compare or provide information on different companies rather than other product comparison sites. Model 500, when trained on this site, will tend to generate a traffic profile consistent with a site that compares or provides information on different companies.
As another example of the training of model 500, consider a news web site. Visitors to news web sites tend to focus their attention on specific articles and read through them until they have finished reading the text of an article. This behavior can be contrasted with traffic at a product comparison site, at which users are more likely to stop reading a particular product page once they find a product that meets their needs. Model 500, when trained on a news site, will tend to generate a traffic profile consistent with behavioral differences such as this.

**EXEMPLARY USE OF MODELS GENERATED BY TRAFFIC PREDICTION COMPONENT 235**

Specific examples of the use of model 500 will now be described. As previously mentioned, traffic prediction component 235 may be used to automatically train a model 500 to obtain an expected traffic profile for a site, without requiring physical measurement or observation of traffic at the site. The predicted traffic profile for the site can be used in a number of ways.

Fig. 6 is a flow chart illustrating an exemplary use of a predicted traffic profile for a site. As shown, a site of interest may be identified, such as by receiving or identifying a URL of the site (act 601). The URL may be used to look-up the predicted traffic profile for the site (act 602), such as a traffic profile previously generated by traffic prediction component 235. The actual traffic profile to the site may then be observed over a predetermined time frame (act 603). The actual traffic profile may be compared to the observed traffic profile and a determination made as to whether the actual traffic profile differs from the predicted traffic profile by more than a threshold level (act 604). If so, this may be an indication that the actual traffic to the site is not genuine (act 605).

One example in which it may be desirable to detect non-genuine traffic is in the context of online advertisements. Advertisers may be charged to display their advertisements based on the number of users that “click-through” an advertisement to the web site of the advertiser. It is desirable that advertisers are only charged for genuine click-throughs (i.e., click-throughs by human users that are genuinely interested in the advertisement). Advertisement click-throughs by automated programs or by malicious users that are not actually interested in the advertisement will tend to generate traffic profiles that do not match the predicted traffic profile.

Another way in which the predicted traffic profile 520, as determined by traffic prediction component 235, may be used is by using the prediction information relating to the time a user is likely to spend at the site to optimize advertisements displayed to the user. For example, different advertisements may be cycled to the user based on the amount of time that the user is expected to browse the site.

Web designers may use traffic prediction component 235 to predict how users will interact with various web layouts. Predicting how a user is likely to interact with a page may help the designer arrange elements within the page.

Yet another use of traffic prediction component 235 may be in the area of web server load balancing. Predicting how users will interact with a web site could be used to intelligently manage load-balancing for the site, such as by intelligently pre-fetching or delaying graphic loads until the user is expected to request them.

**CONCLUSION**

Traffic prediction component 235, as described above, can be used to automatically generate
predicted traffic profiles for web sites based on tags that characterize the sites. The tags can include, for example, tags that indicate the industry with which the site is related and the intended use of the site. The tags can be automatically sent based on predetermined heuristics.

The foregoing description of exemplary embodiments of the invention provides illustration and description, but are 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 a series of acts have been described with regard to Figs. 4 and 6, the order of the acts may be varied in other implementations consistent with the invention. Moreover, non-dependent acts may be implemented in parallel.

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 invention. Thus, the operation and behavior of the aspects of the invention 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.

Further, certain portions of the invention may be implemented as "logic" or as a "component" that performs one or more functions. This logic or component may include hardware, such as an application specific integrated circuit or a field programmable gate array, software, or a combination of hardware and software.

No element, act, or instruction used in the description of the invention 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.
THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:-

1. A computer-implemented method of predicting a profile of user traffic at a web site, the method comprising:
   characterizing the web site based on at least one feature of the web site by using a plurality of tags that describe the at least one feature of the web site;
   selecting an initial traffic profile, that characterizes behaviors of users of another web site during a time that the users are accessing the other web site, based on the characterization of the web site; and
   generating a predicted profile of user traffic at the web site based on the characterization of the web site and the initial traffic profile.

2. The method of claim 1, wherein selecting the initial traffic profile includes selecting the initial traffic profile based on an observed traffic profile associated with another web site that has similar tags as the web site.

3. The method of claim 2, wherein selecting the initial traffic profile based on an observed traffic profile associated with another web site includes selecting the initial traffic profile associated with another web site that is associated with a second plurality of tags that are similar to the plurality of tags.

4. The method of claim 1, wherein generating the predicted profile of traffic includes recursively updating the plurality of tags and the predicted profile of user traffic.

5. The method of claim 1, wherein the plurality of tags include tags that relate to an intended use of the web site.

6. The method of claim 1, wherein the plurality of tags include tags that relate to an industry classification of the web site.

7. The method of claim 1, wherein the plurality of tags include at least one of:
   tags that indicate a primary language used by the web site;
   tags that relate to a geographical location of a host of the web site; or
   tags that relate to a visual layout of the web site.

8. The method of claim 1, wherein the traffic profile includes values that estimate how long a typical user will spend at the web site.

9. The method of claim 1, wherein the predicted profile includes values that estimate how many links a typical user will select at the site.
10. The method of claim 1, wherein the predicted profile includes values that estimate a screen
size of a typical user of the web site or a type of browser of the typical user of the web site.

11. The method of claim 1, further comprising:
   heuristically determining initial values for the plurality of tags.

12. A computer-implemented method of modeling a web site, the method comprising:
   assigning an initial set of tags to the model based on features of the web site, each of the tags
describing at least one feature of the web site;
   determining an initial traffic profile for the web site based on the assigned initial set of tags,
   where the initial profile characterizes behaviors of users of another web site during a time that the users
are accessing the other web site; and
   recursively updating the set of tags and the traffic profile to obtain a final predicted traffic
profile for the web site.

13. The method of claim 12, wherein determining the initial traffic profile for the web site
includes determining the initial traffic profile based on an observed traffic profile from another web
site.

14. The method of claim 12, wherein determining the initial traffic profile includes selecting
the initial traffic profile from another web site that is associated with a second set of tags similar to the
initial set of tags.

15. The method of claim 12, where the initial set of tags includes tags that relate to an intended
use of the web site.

16. The method of claim 12, wherein the initial set of tags includes tags that relate to an
industry classification of the web site.

17. The method of claim 12, wherein the initial set of tags includes at least one of:
tags that indicate a primary language used by the web site;
tags that relate to a geographical location of a host of the web site; or
tags that relate to a visual layout of the web site.

18. The method of claim 12, wherein the initial traffic profile includes values that estimate how
long a typical user will spend at the web site.

19. The method of claim 12, wherein the initial traffic profile includes values that estimate how
many links a typical user will select at the web site.
20. The method of claim 12, wherein the initial traffic profile includes values that estimate a screen size of a typical user of the web site or a type of browser of the typical user of the web site.

21. A computer-readable memory device containing instructions for execution by processing logic, the computer-readable memory device comprising:

- instructions for characterizing a web site based on at least one feature of the web site by using a plurality of tags that describe the at least one feature of the web site;
- instructions for selecting an initial traffic profile, that characterizes behaviors of users of another web site during a time that the users are accessing the other web site, based on the characterization of the web site; and
- instructions for generating a profile predicting user traffic at the web site based on the characterization of the web site and the initial traffic profile.

22. The computer-readable memory device of claim 21, wherein the computer-readable memory device stores:

- tags that relate to an intended use of the web site; or
- tags that relate to an industry classification of the web site.

23. A device comprising:

- means for assigning an initial set of tags to a model of the web site based on features of the web site, each of the tags describing at least one feature of the web site;
- means for determining an initial traffic profile for the web site based on the assigned initial set of tags, where the initial traffic profile characterizes behaviors of users of another web site during a time that the users are accessing the other web site; and
- means for recursively updating the set of tags and the traffic profile to obtain a final predicted traffic profile for the web site.

24. A computer-implemented method of predicting a profile of user traffic at a web site, the method comprising:

- determining an initial set of tags that characterize at least one feature of the web site based on a set of predefined rules, the initial set of tags including a tag that relates to an intended use of the web site or a tag that relates to an industry classification of the web site;
- comparing the initial set of tags to tags determined for other web sites to locate a second web site that has tags that are similar to the initial set of tags;
- using a traffic profile of the second web site as an initial traffic profile for the web site;
- recursively updating the initial traffic profile and the initial set of tags to generate a model of the web site;
- using the model to predict traffic at the web site;
- comparing the predicted traffic to observed traffic; and
- outputting a signal when the observed traffic does not match the predicted traffic.
25. A computer-implemented method of predicting a profile of user traffic at a web site substantially as herein described with reference to any one of the embodiments of the invention illustrated in the accompanying drawings and/or examples.

26. A computer-implemented method of modeling a web site substantially as herein described with reference to any one of the embodiments of the invention illustrated in the accompanying drawings and/or examples.

27. A computer-readable memory device containing instructions for execution by processing logic, the computer-readable memory device substantially as herein described with reference to any one of the embodiments of the invention illustrated in the accompanying drawings and/or examples.

28. A device substantially as herein described with reference to any one of the embodiments of the invention illustrated in the accompanying drawings and/or examples.
BEGIN. HEURISTICALLY ESTIMATE INITIAL TAGS FOR SITE.

401

ESTIMATE INITIAL TRAFFIC PROFILE BASED ON INITIAL TAGS FOR SITE.

402

RECURSIVELY MODIFY TAGS AND TRAFFIC PROFILE UNTIL TAGS STABILIZE.

403

END.

FIG. 4
FIG. 6

BEGIN

RECEIVE IDENTIFICATION OF SITE OF INTEREST

LOOK-UP PREDICTED TRAFFIC PROFILE OF SITE

OBSERVER ACTUAL TRAFFIC PROFILE OF SITE

DOES PREDICTED TRAFFIC MATCH ACTUAL?

YES

ACTUAL TRAFFIC MAY NOT BE GENUINE

NO

END