METHOD AND SYSTEM FOR CHARACTERIZATION OF ONLINE BEHAVIOR

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application is related to and claims priority to U.S. Provisional Application Serial No. 60/323,798, filed September 20, 2001 and entitled "Method and System for Characterization of Online Behavior," the disclosure thereof being expressly incorporated herein by reference.

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0002] This invention relates to a method and system for collecting and classifying computer network traffic, particularly Internet traffic. Specifically, the method and system of the invention characterize the interaction between users and web sites to facilitate construction of more accurate online behavior profiles for marketing purposes.

2. Discussion of the Background

[0003] The Internet has rapidly grown into a center for conducting commerce with unprecedented efficiency and commercial advantage; however, the Internet also presents numerous new challenges to the development and execution of appropriate business models and processes. To design and implement effective marketing and business plans, companies need to gain a better understanding of consumer behavior and preferences while they are conducting Internet commerce.

[0004] Behavior profiles are created using network usage data collected through various methods. Once the data is collected, it is analyzed to determine the behavior of a particular user. In order to create an accurate behavior profile, it is useful to generalize Internet usage by identifying the types of Web sites a particular type of user accesses and the way that type of user accesses a particular type of Web site.

[0005] For example, it would be valuable to a merchant to know that users from a geographical area regularly purchase books from a merchant with a

website, for example, Amazon.com™. However, there is a need for more generalized data, than that which is specific to a particular merchant. It is desirable to have a system that can create what are known as "generalized behavior profiles". By generalized behavior profile is meant a collection of data that describes the attributes and usage patterns of online shoppers, but does not contain personally identifiable information of such online shoppers. More specifically, it is valuable to know that users in a particular geographical area regularly conduct electronic commerce by accessing online catalog and shopping sites by following links on a Web portal site.

[0006] To build accurate generalized behavior profiles, it is desirable to know typical behaviors and actions that lead up to a purchase. For example, it would be desirable to know that many users searched one online merchant site for books to purchase and then went to a different online merchant site to make the actual purchase. For the specific merchant whose site was used for browsing, this information indicates that the site is successful in attracting potential shoppers, but perhaps prices need to be more competitive to retain those initially coming to the site. More generally, this information may indicate that price is a factor which overrides factors such as convenience and functionality of a website in the decisions of similar consumers.

SUMMARY OF THE INVENTION

[0007] In accordance with the present invention there is provided a method of collecting and classifying data about computer network traffic as user website visit sessions, comprising:

determining whether a data request to at least one website comprises a user ongoing interaction with a website or whether it is a new site visit by a user;

if the request is a new visit, creating a record of a new site visit;

if the request is a user ongoing interaction which occurred more than a predetermined amount of time since the last visit, creating a record of the data request as a new site visit;

if the request is a user ongoing interaction occurring at less than a predetermined amount of time since the last visit, create a record of the visit as a continuation of the prior visit; and

characterising the network traffic based on the records created;

characterised in that the method further comprises the steps of:

assembling a data table made up of all known hosts that direct a user to a specific website; and

classifying the computer network traffic corresponding to said sessions by tracking each user request to all known hosts for each specific website;

and wherein the network traffic is characterised for each website independent of which host is used by a user to request data from each website.

BRIEF SUMMARY OF THE DRAWING

[0008] Figure 1 is a basic flow chart of the method to characterize computer network traffic data.

DETAILED DESCRIPTION OF THE INVENTION

[0009] In a step 101, a data table that contains all known hosts (the part of the URL path name that precedes the backslash) that direct a user to a particular location (site) is established. This is referred to as the Site Members Table. By way of example, let us consider the merchant known as Barnes and Noble. In the case of Barnes and Noble, the following hosts are all members of the Barnes and Noble site: *www.bn.com*, *www.barnesandnoble.com*, *music.barnesandnoble.com*, *shop.barnesandnoble.com*, and *207.36.92.145*. By compiling a table containing all hosts which are members of the Barnes and Noble site, this allows the capture of unique users, visits, and the calculation of visit durations for the Barnes and Noble site regardless of what host is entered or how the user navigates on the site.

[0010] To calculate durations accurately, the logic of the method and system must also have a capability for handling components (ads, images, etc.) of a Web site coming from servers (hosts) that are not members of a given site. An additional data table is compiled in step 103, and contains hosts that are classified as "forced content" (e.g. ads, or banners, or counters). Further, URL content types that are classified as images (e.g. .gif, .jpg, etc.) are also compiled into this data table in step 103. This data table is referred to as the Host/URL Types Table.

[0011] For example, the host *ads.ebay.com* is an advertisement, and may appear in the datastream, and thereby suggest a visit to the merchant known as eBay, when a user is actually visiting the merchant using the URL *amazon.com*. The method and system must recognize that although that host is not a member of the current site, it is a component of that site visit to the merchant known as

Amazon, and may also appear on many other sites. The logic underlying the method and system therefore treats ads, images, counters, etc., as part of a running visit, and does not allow the hosts classified as ads, images, counters, etc., to "break" a running visit or spawn the creation of a new site visit. Lines that are generated from secure socket layer (SSL) traffic will also be included as part of the most recent site visit.

[0012] The method and system can create concurrently running site-visit sessions in the event a user has multiple browsers open. Web sites with frames can also be perceived as multiple visits with overlapping times if the frames are served from hosts belonging to different sites.

[0013] The method and system employ the following rules:

1. A new site visit will begin if the line includes a host that is a valid "site member" and:
   - An open site visit is not already running for that site;
   - The URL does not have a content type of "image" (e.g., .gif, .jpg, etc.); and
   - The host is not classified as "forced content" (e.g., ads, banners, counters, etc.)
2. If the host is a site member of an "open" site visit and there are no intervening lines and less than five (5) minutes' time has passed, it will be included in the open site visit.
3. If the line contains a host that is a site member of an open site visit (Site "A"), but there have been one or more intervening lines, it will roll into the site visit from Site A if the following conditions are met:
   - The intervening lines contain known (or classified) hosts and the elapsed time is less than/equal to five (5) seconds or
   - The intervening lines have null host fields or contain only unknown hosts and the elapsed time is less than/equal to thirty (30) seconds.

   Note that if the intervening lines contain a host that is a member of Site B, an overlapping site visit for Site B would start.
4. If the host field is not populated and the port number indicates a secure socket layer (SSL) transaction, then the line is rolled into the most-

recent open site visit. Lines that contain any other port number with a null host field are ignored.

5. If the line contains a host that is not a member of any open site visit and the host is classified as "forced content" (e.g., ads, banners, counters, etc.), the line will roll into the most-recently created site visit.
6. If the line that contains a host that is not a member of an open site visit and the URL content type is defined as "image" (e.g., .gif, .jpg, etc.), it will roll into the most-recent site visit.

[0014] The level of greatest detail at which data is collected is referred to as a Collector Line, and contains all captured elements (user, times, bytes, port number, content types, etc.) from a single request from a server (e.g., a single .gif file on a Web site.) The Site Visits Table is an aggregate table that captures start time, end time, duration, and bytes for each intentional visit to a particular site.

[0015] In step 107, each Collector Line from step 105 is analyzed, and it is determined whether or not the host field of the Collector Line is populated. If not, then at step 113, the port number is analyzed to determine if an SSL transaction is indicated. If an SSL transaction is indicated, then at step 125, the current site visit is continued, and times are updated to include the new Collector Line. If an SSL transaction is not indicated, then, at step 127, the Collector Line is ignored, and the next Collector Line is analyzed.

[0016] If, in step 107, the host field is populated, then the host is checked against the Site Members Table in step 109 to determine if the host is a member of an open site visit. If so, a check is performed in step 117 to determine if there have been any intervening Collector Lines since the last line was added to this site visit. If not, the time interval between the time of the current Collector Line and the time that the last Collector Line was added to the site visit is analyzed in step 115. If the interval is less than or equal to an arbitrary but suitably large time interval, then, in step 125, the current site visit is continued, and times are updated to include the new Collector Line. If the interval is greater than the arbitrary but suitably large time interval, then a new site visit is created in step 129. In a preferred embodiment, the arbitrary but suitably large time interval is within the range of approximately zero to fifteen minutes. In a more preferred embodiment,

the arbitrary but suitably large time interval is within the range of approximately zero to ten minutes. In the most preferred embodiment, the arbitrary but suitably large time interval is approximately five minutes.

[0017] If it is determined, in step 117, that there were intervening Collector Lines since the last line was added to this site visit, the time interval between the time of the current Collector Line and the time that the last Collector Line was added to the site visit is analyzed in step 121. If the time interval is less than or equal to an arbitrary but suitably small time interval, then, in step 125, the current site visit is continued, and times are updated to include the new Collector Line. If the time interval is greater than the arbitrary but suitably small time interval, then in step 123, the host is checked against the Host/URL Types Table to determine if the host is known or classified. In a preferred embodiment, the arbitrary but suitably small time interval is within the range of approximately zero to twenty seconds. In a more preferred embodiment, the arbitrary but suitably small time interval is within the range of approximately zero to ten seconds. In the most preferred embodiment, the arbitrary but suitably small time interval is approximately five seconds.

[0018] If, in step 123, the host is not known or classified, and the time interval is less than or equal to an arbitrary but suitably moderate time interval, then in step 125, the current site visit is continued, and times are updated to include the new Collector Line. If the host is known or classified or if the time interval is greater than the arbitrary but suitably moderate time interval, then in step 129, a new site visit is created. In a preferred embodiment, the arbitrary but suitably moderate time interval is within the range of approximately zero to sixty seconds. In a more preferred embodiment, the arbitrary but suitably moderate time interval is within the range of approximately zero to forty-five seconds. In the most preferred embodiment, the arbitrary but suitably moderate time interval is approximately thirty seconds.

[0019] In step 109, if the host is not a member of an open site visit, then the host and URL are checked in step 111 against the Host/URL Types Table to determine if they are classified as forced content, or if the URL is classified as an image. If so, then in step 125, the current site visit is continued, and times are

updated to include the new Collector Line 125. If not, the host is checked against the Site Members Table in step 119, to determine if the host is a member of another site. If not, the Collector Line is ignored, and the next Collector Line is analyzed in step 127. If the host is a member of another site, then a new site visit is created in step 129.

[0020] The duration of a site visit is determined by subtracting the start time from the latest end time. Because the data lines are analyzed in order by start time, the latest end time may not be the last time received. For example, a request to a web server for a web page may involve multiple requests for different parts of the page. Specifically, one request line may have a start time of 10:02 a.m. and an end time of 10:05 a.m., and a second request line may have a start time of 10:03 a.m. and an end time of 10:04 a.m. Thus, the latest end times are used in the determination of the site visit duration.

[0021] There are two indicators used to determine when a site visit ends: when the end user closes the session, or there is an arbitrary but suitably large period of inactivity. The most preferred arbitrary but suitably large period of inactivity is fifteen minutes. When either of these indicators is present, all open site visits for the end user are closed.

CLAIMS

1. A method of collecting and classifying data about computer network traffic as user website visit sessions, comprising:

determining whether a data request to at least one website comprises a user ongoing interaction with a website or whether it is a new site visit by a user;

if the request is a new visit, creating a record of a new site visit;

if the request is a user ongoing interaction which occurred more than a predetermined amount of time since the last visit, creating a record of the data request as a new site visit;

if the request is a user ongoing interaction occurring at less than a predetermined amount of time since the last visit, create a record of the visit as a continuation of the prior visit; and

characterising the network traffic based on the records created;

characterised in that the method further comprises the steps of:

assembling a data table made up of all known hosts that direct a user to a specific website; and

classifying the computer network traffic corresponding to said sessions by tracking each user request to all known hosts for each specific website;

and wherein the network traffic is characterised for each website independent of which host is used by a user to request data from each website.

2. The method of claim 1, further comprising creating records of a website visit which includes components of a website visit originating from hosts that are not members of a specific website.

3. The method of claim 2, wherein said components of a website visit originating from hosts that are not members of a specific website comprise forced content and images, and are recorded as data requests to the specific website.

4. The method of claim 1, further comprising collecting and classifying computer network traffic for said sessions by concurrently running said sessions for each browser opened by users having multiple browsers open.

5. The method of claim 1, further comprising detecting and recording the port numbers of each data request and determining whether the data requests are secure socket layer transactions.

6. The method of claim 1, further comprising determining the duration of a website visit by subtracting the start time of a session for a user from the latest end time.

7. The method of claim 6, wherein the end of a website visit by a user is determined by detecting when a user closes a session.

8. The method of claim 6, wherein the end of a website visit is determined by detecting when a predetermined amount of time of inactivity occurs.

9. A method performed by a computer system substantially as hereinbefore described with reference to the accompanying drawings.

10. A system for performing a method as claimed in any one of the preceding claims.