Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2396565 A1 2001/07/19

(21) 2 396 565

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2000/10/25
(87) Date publication PCT/PCT Publication Date: 2001/07/19
(85) Entrée phase nationale/National Entry: 2002/07/05
(86) N° demande PCT/PCT Application No.: US 2000/029352
(87) N° publication PCT/PCT Publication No.: 2001/052462
(30) Priorités/Priorities: 2000/01/12 (60/175,665) US;
2000/09/07 (60/231,195) US

(51) Cl.Int.^{7}/Int.Cl.^{7} G06F 15/16

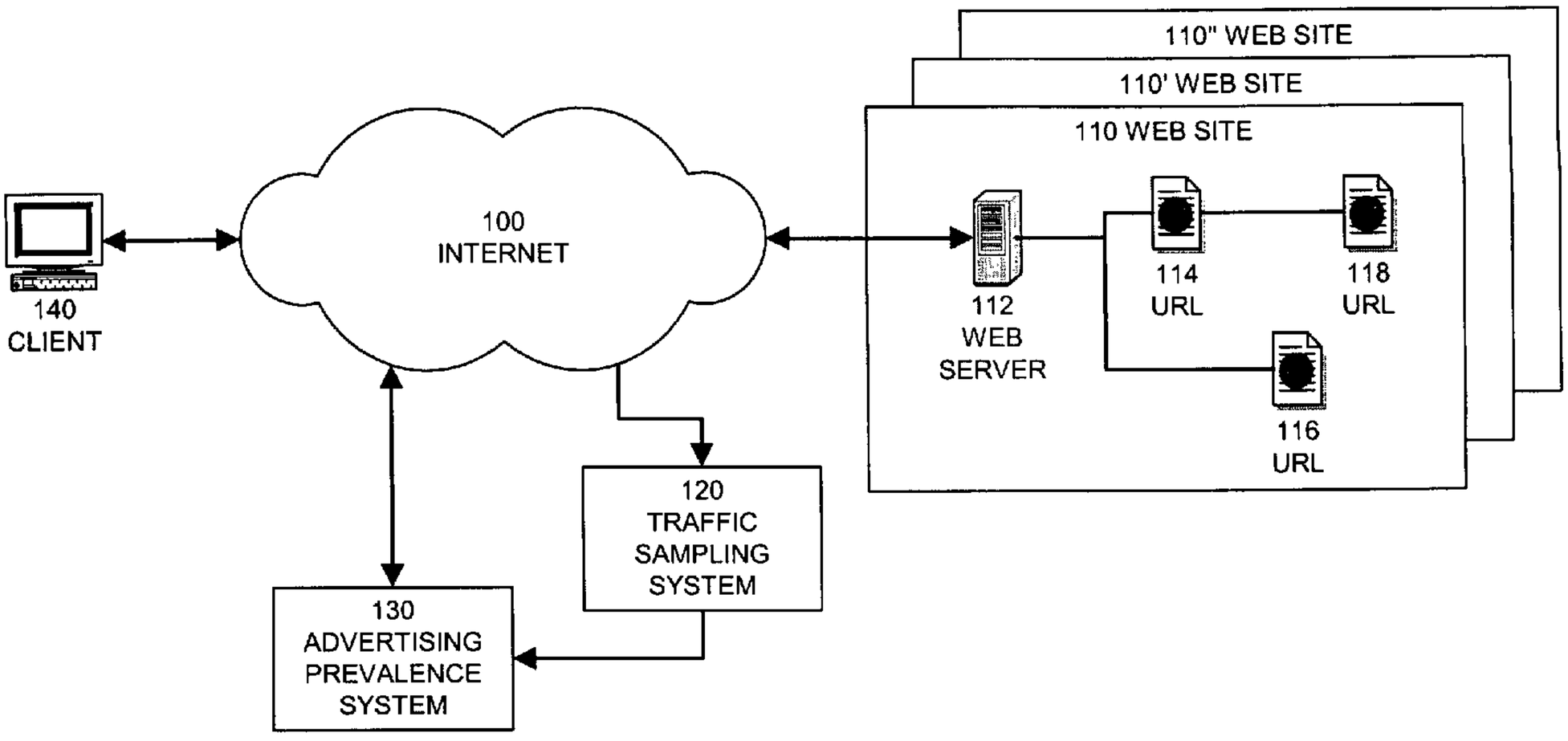
(71) Demandeur/Applicant:
JUPITER MEDIA METRIX, INC., US

(72) Inventeurs/Inventors:
LAUCKHART, GREGORY J., US;
HORMAN, CRAIG B., US;
KOROL, CHRISTA L., US;
BARTOT, JAMES T., US

(74) Agent: MARKS & CLERK

(54) Titre : SYSTEME ET PROCEDE PERMETTANT D'ESTIMER LA PREVALENCE DU CONTENU NUMERIQUE SUR LE WORLD-WIDE-WEB

(54) Title: SYSTEM AND METHOD FOR ESTIMATING PREVALENCE OF DIGITAL CONTENT ON THE WORLD-WIDE-WEB



(57) Abrégé/Abstract:
The present invention is a system, method, and computer program product for tracking and measuring digital content that is distributed on a computer network such as the Internet (100). The system collects online advertisement data, analyzes the data, and uses the data to calculate measurements of the prevalence of those advertisements. The system processes raw traffic data by cleansing and summarizing the traffic data prior to storing the processed data in a database (200). An advertisement sampling system (220) uses site selection and definition criteria and a probe map (524) to retrieve Web pages from the Internet (100), extract advertisements from those Web pages, classify each advertisement, and store the data in a database (200). A statistical summarization system accesses the processed raw traffic data and the advertisement data in the database (200) to calculate advertising prevalence statistics including the advertising frequency, impressions, and spending.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 July 2001 (19.07.2001)

PCT

(10) International Publication Number
WO 01/52462 A3

(51) International Patent Classification[7]: G06F 15/16

(21) International Application Number: PCT/US00/29352

(22) International Filing Date: 25 October 2000 (25.10.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/175,665	12 January 2000 (12.01.2000)	US
60/231,195	7 September 2000 (07.09.2000)	US

(71) Applicant: JUPITER MEDIA METRIX, INC. [US/US]; 4th Floor, 67 Broadway, New York, NY 10012 (US).

(72) Inventors: LAUCKHART, Gregory, J.; 13200 39th Avenue, NE, Seattle, WA 98125 (US). HORMAN, Craig, B.; Apartment #510, 201 Galer Street, Seattle, WA 98109 (US). KOROL, Christa, L.; Seattle, WA (US). BARTOT, James, T.; Apartment #228, 201 Galer Street, Seattle, WA 98109 (US).

(74) Agents: WASZKIEWICZ, Kenneth, P. et al.; Morgan & Finnegan, L.L.P., 345 Park Avenue, New York, NY 10154-0053 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

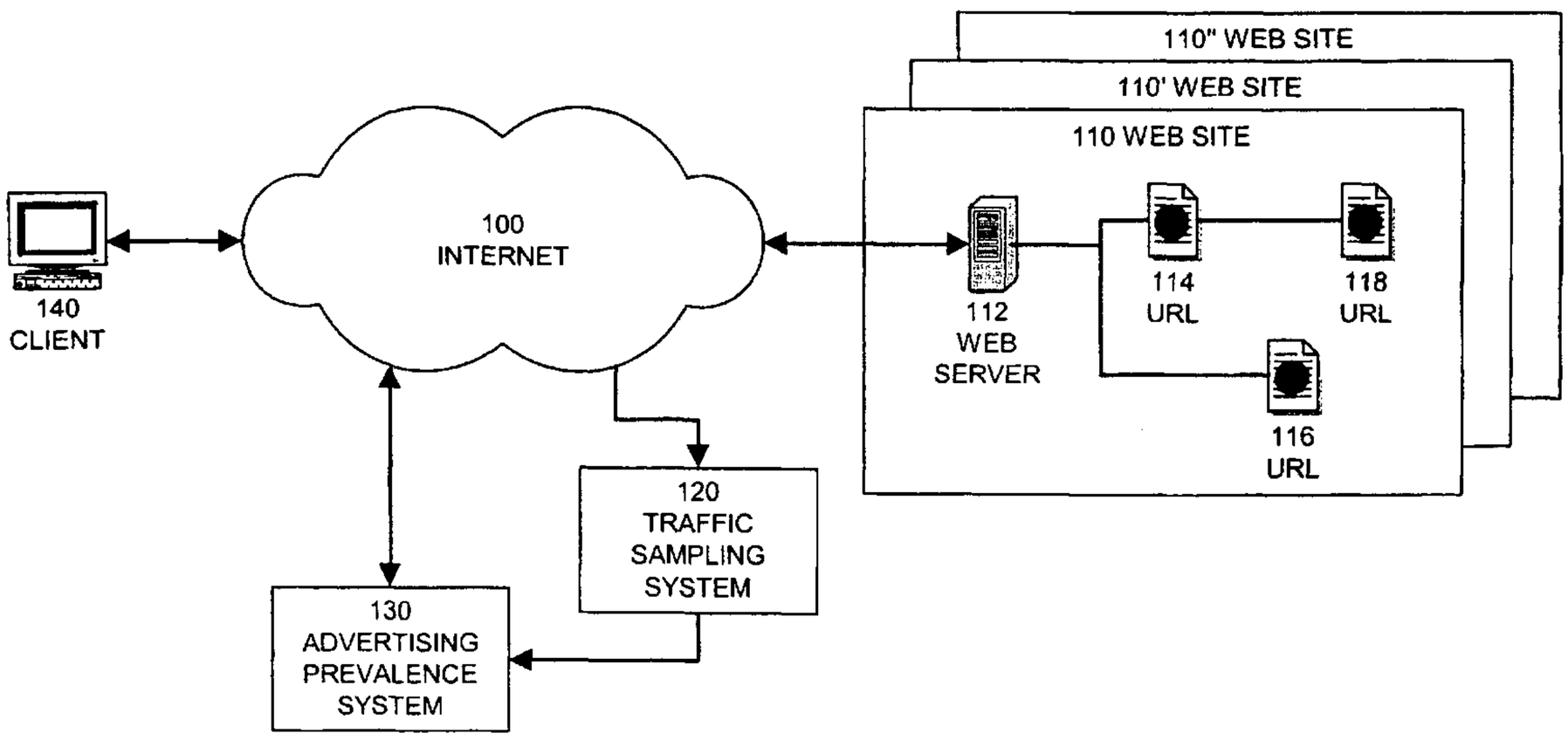
(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(88) Date of publication of the international search report:
31 January 2002

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR ESTIMATING PREVALENCE OF DIGITAL CONTENT ON THE WORLD-WIDE-WEB



(57) Abstract: The present invention is a system, method, and computer program product for tracking and measuring digital content that is distributed on a computer network such as the Internet (100). The system collects online advertisement data, analyzes the data, and uses the data to calculate measurements of the prevalence of those advertisements. The system processes raw traffic data by cleansing and summarizing the traffic data prior to storing the processed data in a database (200). An advertisement sampling system (220) uses site selection and definition criteria and a probe map (524) to retrieve Web pages from the Internet (100), extract advertisements from those Web pages, classify each advertisement, and store the data in a database (200). A statistical summarization system accesses the processed raw traffic data and the advertisement data in the database (200) to calculate advertising prevalence statistics including the advertising frequency, impressions, and spending.

WO 01/52462 A3

WO 01/52462 A3

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM AND METHOD FOR ESTIMATING PREVALENCE OF DIGITAL CONTENT ON THE WORLD-WIDE-WEB

CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims priority from, and incorporates by reference, the provisional application for letters patent, number 60/175,665, filed in the United States Patent and Trademark Office on January 12, 2000, and provisional application for letters patent, number 60/231,195, filed in the United States Patent and Trademark Office on September 7, 2000.

FIELD OF THE INVENTION

The present invention relates generally to a system, method, and computer program product for tracking and measuring digital content that is distributed on a computer network such as the Internet. More particularly, the present invention relates to a system, method, and computer program product that collects online advertisement data, analyzes the data, and uses the data to calculate measurements of the prevalence of those advertisements.

BACKGROUND OF THE INVENTION

The increase in the popularity of the Internet and the World-Wide-Web ("Web") is due, in part, to the interactive technologies that a Web page can employ. These interactive technologies directly affect the Web as an advertising medium because the technologies introduced new advertising formats such as fixed icon sponsorship advertisements, rotating banners and buttons, and interstitial advertisements (i.e., online advertisements that interrupts the user's work and takes over a significant percentage of the screen display). Even though the creation of the advertisement is different, the affect on the viewer is similar to traditional advertising. For example, a banner advertisement or logo icon on a Web page creates an

impression of the product for the viewer that is equivalent to a traditional billboard advertisement that promotes a product by presenting the brand name or slogan. Similarly, a sponsor's logo on a Web page creates an impression of the sponsor for the viewer that is equivalent to seeing a sponsor logo on the scoreboard at a college basketball game.

The rapid and volatile growth of the Internet over the last several years has created a high demand for quality statistics quantifying its magnitude and rate of expansion. Several traditional measurement methodologies produce useful statistics about the Internet and its users, but the complexity of the Internet has left some of these methodologies unable to answer many important questions.

Online advertising is one area where traditional methodologies do not lend well to measurement. Each day, thousands upon thousands of electronic advertisements appear and then disappear from millions of Web pages. The transitory nature of online advertising activity warrants a novel methodology to accurately measure advertising activity.

Existing advertisement tracking and measurement systems automate the collection of Web pages, but fail to automate the collection of the online advertisements. Since the content of an online advertisement changes or rotates over time, accurate reconstruction of the frequency of specific advertisements requires continuous sampling of relevant Web pages in the correct proportions. Furthermore, due to the sheer size of the Web, sampling algorithms must be finely tuned to optimize the allocation of resources (i.e., network bandwidth, database storage, processor time, etc.) and simultaneously enable maximum Internet coverage. The existing advertisement tracking and measurement systems fail to meet these needs because they are not optimized for resource allocation and do not continuously sample relevant Web pages in the correct proportion.

In view of the deficiencies of the existing systems described above, there is a need for an advertisement tracking and measurement system that uses resources more intelligently, is friendlier to the Web sites that it visits, is scalable, and produces accurate measurements. The invention disclosed herein addresses this need.

SUMMARY OF THE INVENTION

The present invention is a system, method, and computer program product for tracking and measuring digital content that is distributed on a computer network such as the Internet. The system collects online advertisement data, analyzes the data, and uses the data to calculate measurements of the prevalence of those advertisements.

In the preferred embodiment, traffic data from a variety of sources and complimentary methodologies fuels the traffic analysis system, an intelligent agent (i.e., software that interact with, learn from, and adapt to an environment). The traffic analysis system processes raw traffic data by cleansing and summarizing the traffic data prior to storing the processed data in a database. When the statistical summarization system calculates the advertising frequency, impressions, and spending, it relies upon the processed data from the traffic analysis system.

The advertisement sampling system, also known as the "prober" or "Cloudprober", use a robust methodology that continually seek out the most significant and influential Web sites to probe (i.e., monitor). Moreover, the site selection and definition performed by the present invention dictates the Web pages that comprise each Web site to ensure that complete, singularly branded entities are reported as such. The advertisement sampling system uses intelligent agent technology to retrieve Web pages at various frequencies to obtain a representative sample. This allows the Cloudprober to accurately assess how frequently each advertisement appears in the traffic data. After the Cloudprober fetches a Web page, the advertisement sampling system

extracts the advertisements from the Web page. In the preferred embodiment, the advertisement extractor, also known as the "extractor", invokes an automatic advertisement detection ("AAD") process, a heuristic extraction process, to automatically extract all of the advertisements from the Web page.

Following extraction of the advertisements from the Web page, the advertisement sampling system invokes a classification engine to analyze the advertisement fragments. The classifier processes each fragment to determine a classification for the fragment and then stores the fragment and classification data in a database. The result of the analyses and processing performed by the advertisement sampling system is a rich catalog of advertising activity that can be easily queried by a client.

The present invention uses a Web front end and user interface to access and update the data in the database. The Web front end provides a client, or user, of the present invention with a query interface to the database populated by the traffic analysis, advertisement sampling, and the statistical summarization systems. The user interface is a graphical user interface that includes a separate component for system account management, site administration, taxonomy administration, advertising content classification, and rate card collection. The user interface allows an account manager and operator to maintain and administer the present invention. The user interface also allows a media editor to review the data in the database to verify the accuracy and integrity of the vast amount of data collected by the present invention. This data integrity process routinely investigates unusual or outlying data points to calibrate the system and adapt it to an ever-changing environment.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying figures best illustrate the details of the present invention, both as to its structure and operation. Like reference numbers and designations in these figures refer to like elements.

Figure 1 is a network diagram depicting the environment for an advertising prevalence system according to the present invention.

Figure 2 depicts the network diagram of Figure 1, in greater detail, to show the relationships between the network environment and the elements that comprise the advertising prevalence system.

Figure 3 depicts the network diagram of Figure 2, in greater detail, to show the elements and sub-elements that comprise the advertising prevalence system and the connections to the network environment.

Figure 4A is an exemplary Web site that illustrates the expected values used in the calculation of the advertising prevalence statistics.

Figure 4B is an exemplary Web site that illustrates the observed values used in the calculation of the advertising prevalence statistics.

Figure 4C is an exemplary Web site that illustrates the weighted values used in the calculation of the advertising prevalence statistics.

Figure 4D is an exemplary Web site that illustrates an alternative method for the calculation of the advertising prevalence statistics.

Figure 5 illustrates an example of a database structure that the advertising prevalence system may use.

Figure 6 is a functional block diagram of the advertising prevalence system that shows the configuration of the hardware and software components.

Figure 7A is a flow diagram of a process in the advertising prevalence system that measures the quality of online advertising and the activity generated by an online advertisement.

Figure 7B is a flow diagram that describes, in greater detail, the process of sampling traffic data from Figure 7A.

Figure 7C is a flow diagram that describes, in greater detail, the process of generating a probe map based on sampled traffic data from Figure 7A.

Figure 7D is a flow diagram that describes, in greater detail, the process of probing the Internet 100 to gather sample data from Figure 7A.

Figure 7E is a flow diagram that describes, in greater detail, the process of classifying the advertising data from Figure 7A.

Figure 7F is a flow diagram that describes, in greater detail, the process of calculating advertising statistics from Figure 7A.

DETAILED DESCRIPTION OF THE INVENTION

Figure 1 depicts the environment for the preferred embodiment of the present invention that includes the Internet 100, and a Web site 110, traffic sampling system 120, advertising prevalence system 130, and client 140. The present invention uses intelligent agent technology to gather data related to the attributes, placement, and prevalence of online advertisements. This data provides a user with up-to-date estimates of advertisement statistics and helps the user to gain a competitive advantage.

As shown in Figure 1, the Internet 100 is a public communication network that allows the traffic sampling system 120 and advertising prevalence system 130 to communicate with a client

140 and a Web site 110. Even though the preferred embodiment uses the Internet 100, the present invention contemplates the use of other public or private network architectures such as an intranet or extranet. An intranet is a private communication network that functions similar to the Internet 100. An organization, such as a corporation, creates an intranet to provide a secure means for members of the organization to access the resources on the organization's network. An extranet is also a private communication network that functions similar to the Internet 100. In contrast to an intranet, an extranet provides a secure means for the organization to authorize non-members of the organization to access certain resources on the organization's network. The present invention also contemplates using a network protocol such as Ethernet or Token Ring, as well as, proprietary network protocols.

The traffic sampling system 120 is a program that monitors and records Web activity on the Internet 100. The traffic sampling system 120 is an intermediary repository of traffic data between a Web surfer (not shown) on the Internet 100 and a Web server 112. The Web server 112 shown in Figure 1 is a conventional personal computer or computer workstation that includes the proper operating system, hardware, communications protocol (e.g., Transmission Control Protocol/Internet Protocol), and Web server software to host a collection of Web pages. The Web surfer (not shown) communicates with the Web server 112 by requesting a Uniform Resource Locator ("URL") 114, 116, 118 associated with the Web site 110, typically using a Web browser. Any program or device that can record a request for a URL made by a Web surfer (not shown) to a Web server 112 can perform the functions that the present invention requires of the traffic sampling system 120. The traffic sampling system 120 then aggregates the traffic data for each Web site 110 for use by the advertising prevalence system 130.

The present invention can use any commercially available traffic sampling system that provides functionality similar to the Media Metrix audience measurement product. Other possible mechanisms to obtain a traffic data sample include:

1. "Proxy Cache Sampling" gathers data such as user clickstream data, and Web page requests from a global distributed hierarchy of proxy cache servers. This data passes through an intermediate mechanism that provides pre-fetch and caching services for Web objects. As of May 1999, traffic statistics calculated by the present invention represent the distillation of raw data from nine first-tier and approximately 400 second-tier caches in the United States, as well as an additional 1100 worldwide.
2. "Client-Side Panel Collection" retrieves sample data from each panelist via a client-side mechanism and transfers that data to a collection repository. The client-side mechanism may monitor the browser location bar, use browser, a client-side proxy, or TCP/IP stack hooks.
3. A "Transcoder" is a proxy that rewrites HTML, usually for the purpose of adding elements for generation of advertisement revenue or page headers/footers. Free Internet service providers ("ISPs") typically use this technique.
4. Any content distribution mechanism that replicates Web page or site content in a manner meant to ease network congestion or improve user experience.
5. Any content filtering mechanism that evaluates requests for URLs and takes actions to allow or disallow such requests.
6. From server logs maintained by Internet service providers ("ISPs") or individual Web sites.

Figure 2 expands the detail of the advertising prevalence system 130 in Figure 1 to show the relationships between the network environment and the elements that comprise the advertising prevalence system 130. The advertising prevalence system 130 includes a traffic analysis system 210, advertisement sampling system 220, and statistical summarization system 230 that communicate data to the database 200 for storage. The account manager 260, operator 262, and media editor 264 can access the database 200 through the user interface 240 to perform administrative functions. The client 140 can access the database 200 through the Web front end 250.

The traffic analysis system 210 receives raw traffic data from the traffic sampling system 120. The traffic analysis system 210 cleanses the raw traffic data by removing information from the traffic data that may identify a particular user on the Internet 100 and then stores the anonymous data in the database 200. The traffic analysis system 210 estimates the global traffic to every significant Web site on the Internet 100. This present invention uses this data not only for computing the number of advertising impressions given an estimate of the frequency of rotation on that page, but also in the probe mapping system 320. In one embodiment, the traffic analysis system 210 receives traffic data from a cache site on the Internet 100. The goal is to accurately measure the number of page views by individual users, and therefore the number of advertising impressions.

The advertisement sampling system 220 uses the anonymous traffic data to determine which URLs to include in the sample retrieved from the Web server 112. The advertisement sampling system 220 contacts the Web server 112 through the Internet 100 to retrieve a URL 114, 116, 118 and extract the advertisements therein along with the accompanying characteristics that describe the advertisements. The success rate for retrieval of creatives is high. Analysis

indicates that the present invention captures over 95% of creatives served. The advertisement sampling system 220 stores these advertisement characteristics in the database 200. The advertisement sampling system 220, for example, the Cloudprober, Online Media Network Intelligent Agent Collection ("OMNIAC"), or the Cloudprober, repeatedly probes prominent Web sites, extracts advertisements from each Web page returned by the probe, and classifies the advertisements in each Web page by type, technology and advertiser.

The traffic analysis system 210 and the advertisement sampling system also present the data retrieved from the Internet 100 to the statistical summarization system 230 for periodic processing. The statistical summarization system 230 calculates the advertising frequency, impressions, and spending on per site per week basis.

The graphical user interface for the present invention includes the user interface 240 and Web front end 250. The account manager 260, operator 262, and media editor 264 access the user interface 240 to administer access by the client 140 to the Web front end 250 (e.g., account and password management), define sites and probe instructions, and manage the advertising taxonomy, content classification, and rate card collection for the advertising prevalence system 130. The Web front end 250 is the Web browser interface that a client 140 uses to retrieve the advertisement measurement results from the database 200 as generated by the traffic analysis system 210, advertisement sampling system 220, and the statistical summarization system 230.

Figure 3 further expands the detail of the advertising prevalence system 130 to depict the logical components comprising the elements of the advertising prevalence system 130 shown in Figure 2. Figure 3 also depicts the relationships between the network environment and those logical components.

The traffic analysis system 210 includes an anonymity system 310 and traffic summarization process 312.

The anonymity system 310 cleanses the data received from the traffic sampling system 120 by removing information that identifies a particular user on the Internet. The data is rendered anonymous by passing all user information (e.g., originating internet protocol ("IP") number or cookies) through a cryptographically secure one-way hash function; this assures the utmost privacy for Web users without devaluing the resulting data. The anonymity system 310 presents the cleansed data to the traffic summarization system 312 which in turn stores the aggregated URL count information in database 200.

The traffic summarization process 312 receives cleansed data from the anonymity system 310. The anonymous traffic data is summarized to yield traffic totals by week or month for individual URLs, domains, and Web sites. The traffic summarization process 312 scales the data by weighting factors to extrapolate total global traffic from the sample.

The advertisement sampling system 220 in Figure 3 includes a probe mapping system 320, Web page retrieval system 322, Web browser emulation environment 324, advertisement extractor 326, and a structural classifier 328.

The probe mapping system 320 generates a probe map, i.e., the URLs 114, 116, 118 that the advertisement sampling system 220 will visit. This probe map assists the advertisement sampling system 220 with the measurement of the rotation of advertisements on individual Web sites. The preferred embodiment of the present invention continuously fetches various Web pages in the probe map. In an alternative embodiment, the present invention visits each URL in the probe map approximately every 6 minutes. Another embodiment can vary the fetching rate by considering several factors including the amount of traffic that visits the Web site as a whole

and the individual Web page in question, the number of advertisements historically seen on the Web page, and the similarlity of the historically observed ad rotation to other sampled pages.

The Web page retrieval system 322 uses this probe map generated by the probe mapping system 320 to determine which Web pages it needs to sample and the frequency of the sampling.

For each URL in the probe map generated by the probe mapping system 320, the Web page retrieval system 322 fetches a Web page, extracts each advertisement from the Web page, and stores the advertisement's attributes in the database 200. The data retrieved from each URL in the probe map is used to calculate the frequency with which each advertisement is shown on a particular Web site

For each Web page, the Web browser emulation environment 324 simulates the display of the Web page in a browser. This simulation guarantees that the present invention will detect not only static advertisements, but also dynamic advertisements generated by software programs written in a language such as JavaScript, Perl, Java, C, C++, or HTML that can be embedded in a Web page.

The advertisement extractor 326 extracts the online advertisements from the result of the simulation performed by the Web browser emulation environment 324. The advertisement extractor 326 identifies features of the advertising content (i.e., "fragments") extracted from the Web pages returned by the probe mapping system 320 that are of particular interest. Advertisements are the most interesting dynamic feature to extract, however, an alternative embodiment of the present invention may use the extraction technology to collect any type of digital content including promotions, surveys, and news stories. The advertisement extractor 326 can use various advertisement extraction methods, including rule-based extraction, heuristic extraction, and comparison extraction.

Rule-based extraction relies upon a media editor 264 to use the user interface 240 to create rules. The user interface 240 stores the rules in the database 200 and the advertisement extractor 326 applies the rules to each Web page that the Web page retrieval system 322 retrieves. The effect of running a rule is to identify and extract an HTML fragment from the Web page (i.e., the part of the page containing the advertisement). The advertisement extractor 326 first converts the HTML representation of the fetched Web page into a well-formed XML representation. Following this conversion, the rules are applied to the parse tree of the XML representation of the Web page.

Heuristic extraction relies upon the similarity of advertisements at the HTML or XML source code level because the advertisements are typically inserted by an advertisement server when the Web page is generated in response to the Web browser emulation environment 324 request to display the Web page. Heuristic extraction analyzes the source code for clues (e.g., references to the names of known advertisement servers) and extracts fragments that surround those clues. The advantage of this method is that the extraction is automatic and the media editor need not create the rules.

Comparison extraction repeatedly fetches the same Web page. This extraction method compares the different versions of the Web page to determine whether the content varies from version to version. The portion of the Web page that varies with some degree of frequency is usually an advertisement and is extracted.

The structural classifier 328 parses each advertisement and stores the structural components in the database 200 and passes those components to the statistical summarization system 230. Each advertisement fragment extracted by the advertisement extractor 326 is analyzed by the structural classifier 328. The process performed by the structural classifier 328

comprises duplicate fragment elimination, structural fragment analysis, duplicate advertisement detection.

The structural classifier 328 performs duplicate fragment elimination by comparing the current advertisement fragment to other fragments in the database 200. Two advertisement fragments are duplicates if the fragments are identical (e.g., each fragment has the exact same HTML content). If the structural classifier 328 determines that the current fragment is a duplicate of a fragment in the database, the advertisement sampling system 220 logs another observation of the fragment and continues processing fragments.

The structural classifier 328 performs structural fragment analysis on the XML representation of the Web page by determining the "physical type" of the fragment (i.e., the HTML source code used to construct the advertisement). Physical types that present invention recognizes include banner, form, single link, and embedded content. Banner advertisement fragments include a single HTML link having one or two enclosed images and no FORM or IFRAME tag. Form advertisement fragments include a single HTML form having no IFRAME tag. Single link advertisement fragments include a link with textual, but no IMG, FORM, or IFRAME tags. Embedded content advertisement fragments reference an external entity using an IFRAME tag. After performing this analysis, the structural classifier 328 updates the advertisement fragment in the database. For a banner advertisement fragment, the structural classifier 328 stores the link and image URL's in the database 200. A form advertisement fragment requires the creation of a URL by simulating a user submission that sets each HTML control to its default value. The structural classifier 328 stores this URL and the "form signature" (i.e., a string that uniquely describes the content of all controls in the form) in the database 200. For a single text advertisement fragment, the structural classifier 328 stores the

URL for the link and all text contained within the link in the database 200. For embedded content advertisement fragments, the structural classifier 328 stores the URL associated with the external reference in the database 200. This URL is loaded by the system, and the referenced document is loaded. Once the loaded document has been structurally analyzed, the original fragment inherits any attributes that result from analysis of the new fragment.

The structural classifier 328 performs duplicate advertisement detection on each advertisement fragment that has a known physical type because these fragments represent advertisements. Each unique advertisement has information, including which site definitions are associated with the fragment, stored in the database 200. The structural classifier 328 determination of uniqueness depends on different criteria for each type of fragment. The first step for every type of definition is to resolve all URLs associated with the record. URLs that refer to images are loaded, and duplicate images are noted. HTML link URLs, also known as "click URLs", are followed each time a new ad is created. The final destination for a click URL, after following all HTTP redirects, is noted. This is also done for simulated link submission URLs associated with form definitions. Once all URLs have been resolved, the structural classifier 328 determines whether the advertisement is unique. Banner advertisement fragments are considered unique if they have the same number of images, if the images are identical, and if the destination URL is identical. Form advertisement fragments are considered unique if they have the same signature, and the same destination URL. Single link advertisement fragments are considered unique if they have the same textual content and the same destination URL.

The statistical summarization system 230 calculates the advertisement statistics for each unique advertisement in the database 200. The present invention calculates, for each Web site, the advertising impressions (i.e., the number of times a human being views an advertisement).

The present invention calculates the advertising impressions, I, using the formula $I = T \times R$, where T is the traffic going to the site, and R is the rotation of advertisements on that site. The present invention also calculates the spending, S, using the formula $S = I \times RC$, where I is the advertising impressions for a Web site, and RC is the rate code for the Web site. Most advertising buys are complicated deals with volume purchasing discounts so our numbers do not necessarily represent the actual cost of the total buy.

The Web front end 250 is a graphical user interface that provides a client 140 with a query interface to the database 200 populated by the traffic analysis system 210, advertisement sampling system 220, and the statistical summarization system 230. The client 140 can use the Web front end 250 to create, store, edit and download graphical and tabular reports for one or more industry categories depending on the level of service the client 140 selects.

The user interface 240 in Figure 3 includes a separate component for system account management 340, site administration 342, taxonomy administration 344, advertising content classification 346, and rate card collection 348.

The account manager 260 uses the system account management 340 module of the user interface 240 to simplify the administration of the Web front end 250. The account manager 260 uses the system account management 340 module to create and delete user accounts, manage user account passwords, and check on the overall health of the Web front end 250.

The operator 262 uses the site administration 342 module of the user interface 240 to simplify the administration of the site definitions. Analysts from the Internet Advertising Bureau estimate that over 90% of all Web advertising dollars are spent on the top fifty Web sites. Site selection begins by choosing the top 100 advertising by considering data from Media Metrix,

Neilsen/Net Ratings, and the proxy traffic data in the database 200. These lists are periodically updated to demote Web sites with low traffic levels and promote new sites with high traffic levels. The present invention also includes Web sites that provide significant content in key industries. A site chosen for inclusion in the site definitions must have the structure of the site analyzed to remove sections that do not serve advertisements, originate from foreign countries, or are part of a frame set. Sites that originate from a foreign country, such as yahoo.co.jp, sell advertising in the host country, and therefore are not applicable to the measurements calculated by the present invention. Web sites that use an HTML frameset are treated very carefully to only apply rotation rates to the traffic from the sections of the frameset that contain the advertisement. These combined exclusions are key to making accurate estimates of advertising impressions. The present invention also tags sections that cannot be measured directly, due to registration requirements (e.g., mail pages). Since Web sites change frequency, this structural analysis is repeated periodically. Eventually the analysis stage will automatically flag altered sites to allow even more timely updates.

The media editor 264 uses the taxonomy administration 344, advertising content classification 346, and rate card collection 348 modules of the user interface 240. The taxonomy administration 344 module simplifies the creation and maintenance of the attributes assigned to advertisements during content classification including the advertisements industry, company, and products. The taxonomy names each attribute and specifies its type, ancestry and segment membership. For example, a company Honda, might be parented by the Automotive industry and belong to the industry segment Automotive Manufactures. The advertising content classification 346 component assists the media editor 264 with performing the content classification.

The structural classifier 328 performs automated advertisable assignment to determine what the advertisement is advertising. This process include assigning "advertiseables" (i.e., attributes describing each "thing" that the advertisement is advertising) to each advertisement fragment. In another embodiment of the present invention, the advertisement sampling system 220 uses an extensible set of heuristics to assign advertisables to each advertisement. In the preferred embodiment, however, the only automatic method employed is location classification. Location classification relies on the destination URL in order to assign a set of advertisables to an advertisement. A media editor 264 uses the user interface 240 to maintain the set of classified locations. For example, the first time a media editor observes an advertisement in which the click-thru URL is www.honda.com, he can enter this URL as pertaining to the advertiser "Honda Motors". Any subsequent advertisement that includes the same click-thru URL will also be recognized as a Honda advertisement. A classified location comprises a host, URL path prefix, and set of advertisables. Location classification assigns a classified location advertisables to an advertisement if the host in the destination URL matches the host of the classified location and the path prefix in the classified location matches the beginning of the path in the destination URL.

The structural classifier 328 performs human advertisable assignment and verification as a quality check of the advertisable data. This phase is the most human intensive. A media editor 264 uses a graphical user interface module in the user interface 240 to display each advertisement, verifies automatic advertisable assignments, and assigns any other appropriate advertisables that appear appropriate after inspection of the advertisement and the destination of the advertisement. The location classification database is also typically maintained at this time.

The media editor 264 uses the rate card collection 348 module to enter the contact and rate card information for a Web site identified by the traffic analysis system 210, as well as, designated advertisers. Rate card entry includes the applicable quarter (e.g., Q4 2000), advertisement dimensions in pixels, fee structure (e.g., CPM, flat fee, or per click), cost schedule for buys of various quantities and duration. The media editor also records the URL address of the online media kit and whether rates are published therein. Contact information for a Web site or advertiser includes the homepage, name, phone and facsimile numbers, email address, and street address.

Figures 4A through 4C illustrate the preferred method for calculating the advertising prevalence statistics. The calculation of the advertising prevalence statistics is an iterative process that uses expected values derived by the traffic analysis system 210 and observed values derived by the advertising prevalence system 220 to calculate the weighted values and the advertising prevalence statistics. Figures 4A through 4C each depict a network on the Internet 100 that includes two Web sites served by Web server P 410 and Web server Q 420. Figure 4A illustrates exemplary expected traffic values for the network. Figure 4B illustrates exemplary observed traffic values for the network. Figure 4C illustrates exemplary weighted traffic values for the network.

The first step in the process is to normalize the results from the traffic analysis system 210. The traffic analysis system 210 provides the traffic received by each Web page in the traffic data sample. Figure 4A depicts the exemplary traffic received at each Web page 411-416, 421-424 in the Internet 100 with the label "Traffic =". The probe map generated by the probe mapping system 320 includes an entry for each Web page 411-416, 421-424. The probe map also includes an "area" that each Web page 411-416, 421-424 consumes in the probe map.

Figure 4A depicts the exemplary area that each Web page 411-416, 421-424 consumes in the probe map with the label "Area =". The normalized results are calculated by dividing the area that a Web page consumes in the probe map by the sum of the area for each Web page in the traffic sample. In Figure 4A, the normalized value, or chance, for Web page P1 411 is the area for Web page P1 (i.e., 15) divided by the sum of the area for Web page P1, P2, P3, P4, P5, P6, Q1, Q2, Q3, and Q4 (i.e., 120). The normalized value is, therefore, 0.125, or 12.5%. In addition to the normalized, the system also determines the scale by dividing the traffic for a Web page by the area for the Web page. In Figure 4A, the scale for Web page P1 411 is the traffic for Web page P1 (i.e., 150) divided by the area for Web page P1 (i.e., 15), therefore, the scale for Web page P1 is 10. Table 1 summarizes the scale and chance values for the remaining Web page in Figure 4A.

Web Page	Area	Scale	Chance
P1	15	10	12.5%
P2	10	1	8.3%
P3	14	1	12%
P4	12	0.25	10%
P5	8	0.5	6.7%
P6	4	1	3.3%
Q1	30	0.5	25%
Q2	4	0.5	3.3%
Q3	15	2	12.5%
Q4	8	0.5	6.7%

Table 1.

Figure 4B depicts the exemplary Web page fetches at each Web page 411-416, 421-424 in the Internet 100 with the label "Fetches =". Figure 4B also depicts the exemplary number of views of advertisement that appear on each Web page 411-416, 421-424 with the label "A1 Views =" to indicate the number of views of advertisement A1, "A2 Views =" to indicate the number of views of advertisement A2, etc.

Figure 4C depicts the exemplary Web page weighted fetches at each Web page 411-416, 421-424 in the Internet 100 with the label "Fetches =". Figure 4C also depicts the exemplary number of views of advertisement that appear on each Web page 411-416, 421-424 with the label "A1 Views =" to indicate the number of views of advertisement A1, "A2 Views =" to indicate the number of views of advertisement A2, etc. The next step in the calculation process is to calculate the Scaled Fetches for each Web site 410, 420 by summing the product of the observed fetches from Figure 4B and the scale from Figure 4A, for each Web page 411-416, 421-424 in the Web site. Next, the calculation computes the Traffic for each Web site 410, 420 by summing the traffic from Figure 4A for each Web page 411-416, 421-424 in the Web site. The rate card, or CPM, is a value assigned by the media editor 264 for each Web site 410, 420. Table 2 summarizes the Scaled Fetches, Traffic, and CPM for Figures 4A through 4C.

Site	Scaled Fetches	Traffic	CPM
P	193.5	185	$35.00
Q	43	51	$50.00

Table 2.

The next in the calculation process is to compute the Scaled Observations for each advertisement on each Web site 410, 420 by summing the product of the advertisement views from Figure 4B and the scale from Figure 4A, for each Web page 411-416, 421-424 in the Web site 410, 420. The final step in the calculation is to compute the advertising prevalence statistics (i.e., Frequency, Impressions, and Spending) for each advertisement in each Web site 410, 420. Frequency is computed by dividing the scaled observations by the scaled fetches for each advertisement in each Web site 410, 420. Impressions is computed by multiplying the Frequency by the Traffic from Table 2 above for each advertisement in each Web site 410, 420. Spending is computed by multiplying the Impressions by the CPM from Table 2 above for each advertisement in each Web site 410, 420. Table 3 summarizes the Scaled Observations, Frequency, Impressions, and Spending for Web site P 410 using the data in Figures 4A through 4C. Table 4 summarizes the Scaled Observations, Frequency, Impressions, and Spending for Web site Q 410 using the data in Figures 4A through 4C.

	Scaled Observations	Frequency	Impressions	Spending
A1	55.0	0.28	52.58	$1.84
A2	85.0	0.44	81.27	$2.84
A3	6.0	0.03	5.74	$0.20
A4	3.5	0.02	3.35	$0.12
A5				

Table 3.

	Scaled Observations	Frequency	Impression	Spending

			s	
A1	29.5	0.69	34.99	1.75
A2	12.0	0.28	14.23	0.71
A3	12.0	0.28	14.23	0.71
A4	12.0	0.28	14.23	0.71
A5	1.5	0.03	1.78	0.09

Table 4.

Figure 4D illustrates an alternative embodiment for calculating the advertising prevalence statistics. In this embodiment, the prober is tuned to optimize rotation measurement accuracy. Statistical estimates of accuracy in the field are difficult to perform, due to the non-stationary nature of advertising servers. When probing every 6 minutes, it has a 0.06% resolution in rotational frequency over a one-week measurement period.

Also in the alternative embodiment of Figure 4D, the probes are distributed among the sites to accurately measure ad rotation on each site. The number of probing URLs assigned to a site is determined from three variables. The first is a constant across all sites; a certain number of probing URLs are required to accurately measure rotation on even the smallest site. Half of the probes are assigned with this variable. The second variable, weighted at 40%, is the amount of traffic going to a site, as each probing URL represents a proportion of total Internet traffic. The twenty largest sites receive over 75% of these probes. Finally the complexity of site, as measured by the total number of unique URLs found in our proxy traffic data, is taken into account, with more complicated sites receiving extra probing URLs. This accounts for the remaining 10% of the probe distribution. Probing URLs can be chosen using a Site Shredder

algorithm to break the site into regions (i.e., sets of pages whose advertisement rotation characteristics are likely to be similar) for probing. The distribution of regions is mathematically designed to maximize site coverage and, therefore, advertisement rotation accuracy. A single URL is chosen to represent the advertising rotation from each region. This URL is chosen as the most heavily trafficked page containing advertisements in that region. The algorithm avoids date specific pages or pages referring to a time-limited event such as the August 1999 total lunar eclipse.

The alternative embodiment of Figure 4D calculates advertisement impressions by combining the estimates of rotation and traffic for each Web site 430. To do this the system breaks the site down into its constituent stems using the Site Shredder algorithm. The rotation of advertisements in each advertisement slot is calculated and applied to estimate advertising impressions on its associated stem. The advertisement rotation on stems without probes is estimated from an average, weighted by traffic, of advertisement rotation of probes on a similar level.

For instance, in Figure 4D, the sample site tree has five probes URLs 431-435, $P_{1\text{-}5}$, placed on five main branches off a main page and 14 secondary branches. The number on each page is the sample traffic going to that page. Probe P_1 on the home page, "www.testsite.com", measures the rotation, R, to be applied to the traffic going to that main page, with traffic of 88 page views. Branch A has a single probe, P_2, placed on the top-level page of that branch with a probing URL "www.testsite.com/A/". The rotation of this single probing URL is estimated as R_A and is applied to the traffic for that entire stem, a total of 21 page views. Branch C has a probe, P_3, on a heavily trafficked secondary branch page, with a probing URL "www.testsite.com/C/third.html". The rotation, R_C, of this page is applied to all the secondary

branch pages on that stem and also up one level in the tree, across a total of 25 page views. Branch E receives a large portion of the traffic for the site, a total of 61 page views, and therefore is assigned two probes, P_4 and P_5. These are on two secondary branch pages, "www.testsite.com/E/first.html" and "www.testsite.com/E/third.html". The rotation of each is applied the traffic to those individual pages. For the remaining 18 page views on that branch (ten page views from two secondary pages and eight from the top level page of that branch) a weighted rotation is calculated, $R_E = ((13 \times R_{E1}) + (30 \times R_{E3}))/(13+30)$. The analysis of stem rotation results in advertising impressions for over 96% of the site. The impressions for the final two branches, B and D, are calculated with an average rotation from adjacent branches, weighted by traffic,

$$R_B = R_D = ((21 \times R_A) + (25 \times R_C) + (61 \times R_E)) \div (21 + 25 + 61).$$

This analysis results in total impressions across the site for each unique advertisement. The final calculation performed by the alternative embodiment of Figure 4D is spending, the product of the Impressions and the Rate Card.

Figure 5 illustrates a database structure that the advertising prevalence system 130 may use to store information retrieved by the traffic sampling system 120 and the Web page retrieval system 320. The preferred embodiment segments the database 200 into partitions. Each partition can perform functions similar to an independent database such as the database 200. In addition, a partitioned database simplifies the administration of the data in the partition. Even though the preferred embodiment uses database partitions, the present invention contemplates consolidation of these partitions into a single database, as well as making each partition an independent database and distributing each database to a separate general purpose computer

workstation or server. The partitions for the database 200 of the present invention include sampling records 510, probing definitions 520, advertising support data 530, and advertising summary 540. The preferred embodiment of the present invention uses a relational database management system, such as the Oracle8*i* product by Oracle Corporation, to create and manage the database and partitions. Even though the preferred embodiment uses a relational database, the present invention contemplates the use of other database architectures such as an object-oriented database management system.

The sampling records 510 partition of database 200 comprises database tables that are logically segmented into traffic data 512, advertisement view logging 514, and advertising structure 516 areas.

The traffic data 512 area contains data processed by the traffic sampling system 120, anonymity system 310, and statistical summarization system 230. The data stored in this schema includes a "munged" URL, and the count of traffic each URL receives per traffic source over a period of time. A "munged" URL is an ordinary URL with the protocol field removed and the order of the dotted components in the hostname reversed. For example, the present invention transforms an ordinary URL, such as http://www.somesite.com/food, into a munged URL by removing the protocol field (i.e., "http://") and reversing the order of the dotted components in the hostname (i.e., "www.somesite.com"). The resulting munged URL in this example is "com.somesite.www/food". The present invention uses this proprietary URL format to greatly enhance the traffic data analysis process. The traffic sampling system 120 populates the traffic data 512 area in database 200. The probe mapping system 320 accesses the data in the traffic data 512 area to assist the Web page retrieval system 322 and the statistical summarization system 230 with the calculation of the advertising impression and spending statistics.

The advertisement view logging 514 area logs the time, URL, and advertisement identifier for each advertisement encountered on the Internet 100. This area also logs each time the system does not detect an advertisement in a Web page that previously included the advertisement. In addition, the system logs each time the system detects a potential advertisement, but fails to recognize the advertisement during structural classification. The structural classifier 328 and the Web page retrieval system 322 of the advertisement sampling system 220 populate the advertisement view logging 514 area in database 200. The statistical summarization system 230 accesses the data in the advertisement view logging 514 area to determine the frequency that each advertisement appears on each site.

The advertisement structure 516 area contains data that characterizes each unique advertisement located by the system. This data includes the content of the advertisement, advertisement type (e.g., image, HTML form, Flash, etc.), the destination URL linked to the advertisement, and several items used during content classification and diagnostics, including where the advertisement was first seen, and which advertisement definition originally produced the advertisement. The structural classifier 328 component of the advertisement sampling system 220 populates the advertisement structure 516 area in database 200. The user interface 240 accesses the data in the advertisement structure 516 area to display each advertisement to the media editor 264 during classification editing. The Web front end 250 also accesses the data in the advertisement structure 516 area to display the advertisements to the client 140.

The probing definitions 520 partition of database 200 comprises database tables that are logically segmented into site definition 522, probe map 524, and advertisement extraction rule definition 526 areas.

The site definition 522 area carves the portion of the Internet 100 that the system probes into regions. The primary region definition is a "site", a cohesive entity the system needs to analyze, sample, and summarize. The system defines each site in terms of both inclusive and exclusive munged URL prefixes. A "munged URL prefix" is a munged URL that represents the region of all munged URLs for which it is a prefix. An "inclusive munged URL prefix" specifies that a URL is part of some entity. An "exclusive munged URL prefix" specifies that a URL is not part of some entity, overriding portions of the entity included by an inclusive prefix. To illustrate, the following is list of munged URLs that may result from the processing of a set of URLs in a traffic sample.

1. com.somesite/
2. com.somesite/foo
3. com.somesite/foo/bar
4. com.somesite/foo/blah
5. com.someothersite/

If the site definition for "somesite" includes the inclusive URL prefix "com.somesite/" and the exclusive URL prefix "com.somesite/foo/bar", the application of this site definition to above sample URLs yields a system that includes URL 1, 2, and 4. URL 3 is not part of the site definition due to the explicit exclusion of "com.somesite/foo/bar". URL 5 is not part of the site definition because it was never included in the inclusive URL prefix "com.somesite/". The user interface 240 populates the site definition 522 area in database 200. The probe mapping system 320 accesses the data in the site definition 522 area to determine which URLs to probe. The statistical summarization system 230 accesses the data in the site definition 522 area to determine traffic levels to sites by summing traffic to URLs included in a site.

The probe map 524 area contains a weight for each URL in each site that the system is measuring. This weight determines the likelihood that the system will choose a URL for each probe. The system generates the weights by running complex iterative algorithms against the traffic data and the probing records in the database 200. An analysis of the traffic data can discern which URLs have been visited, how often users have visited those URLs. The result of the analysis guarantees that the system performs advertisement sampling of these URLs in similar proportions, given certain constraints such as a maximum number of probes to allocate to any single URL. The data in the sampling records 510 partition of the database 200 is useful for determining which URLs are in need of special handling due to past behavior (e.g., a URL is sampled less infrequently if the system has never detected an advertisement in the URL). The probe mapping system 320 populates the probe map 524 areas in the database 200. The probe mapping system 320 accesses the data in the probe map 524 area to allocate the probes. The statistical summarization system 230 accesses the data in the probe map 524 area to determine which URLs should have their rotations scaled to counter the effect of probe map constraint enforcement.

The advertisement extraction rule definition 526 area describes Extensible Markup Language ("XML") tags, typically representing a normalized HTML document, that indicate those portions of the content that the system considers to be advertisements. The system defines an extraction rule in terms of "XML structure" and "XML features". "XML structure" refers to the positioning of various XML nodes relative to others XML nodes. For example, an anchor ("A") node containing an image ("IMG") node is likely an advertisement. After using this structural detection process to match the advertisement content, the system examines the features of the content to determine if the content is an advertisement. To continue the previous example,

if the image node contains a link ("href") feature that contains the sub-string "adserver", it is very likely an advertisement. Features may match based on a simple sub-string, as in the example, or a more complicated regular expression. Another form of extraction rule may point to a specific node in an XML structure using some form of XML path specification, such as a "Xpointer". The media editor 264 populates the advertisement extraction rule definition 526 area in the database 200. The advertisement extractor 326 of the advertisement sampling system 220 accesses the data in the advertisement extraction rule definition 326 area to determine which portions of each probed page represent an advertisement.

The advertising support data 530 partition of database 200 comprises database tables that are logically segmented into advertisable taxonomy 532, advertising information 534, rate card 536, and extended advertisable information 538 areas.

The advertisable taxonomy 532 area contains a hierarchical taxonomy of advertisables, attributes that describe what the advertisement is advertising. This taxonomy includes industries, companies, products, Web sites, Web sub-sites, messages, etc. Each node in the hierarchy has a type that specifies what kind of entity it represents and a parent node. For example, the hierarchy may specify that products live within companies, which in turn live within industries. The media editor 264 populates the advertisable taxonomy 532 area in the database 200. The user interface 240 accesses the data in the advertisable taxonomy 532 area to generate statistical data and record where companies, industries, etc. tend to advertise. The Web front end 250 also accesses the data in the advertisable taxonomy 532 area to display this information to the client 140.

The advertising information 534 area contains the data that describe what each unique advertisement recorded by the system advertises. This tables in this area associate advertisables

with advertisements. For example, the system may associate a company type of advertisable with a specific advertisement to indicate that the advertisement is advertising the company. The system uses the following methods to associate an advertisable with an advertisement:

1. A "direct classification" assigns an advertisable directly to the advertisement. For example, a media editor 264 creates a direct classification by specifying that a particular advertisement advertises the "Honda" advertisable.
2. A "location classification" assigns an advertisable to a location prefix that the system uses to match the destination of the advertisement. For example, a media editor 264 creates a location classification by specifying that the location "com.honda" indicates an advertisement for Honda. An advertisement that points to "com.honda.www/cars", therefore, associates the advertisement with Honda.
3. An "ancestral classification" assigns an ancestor of the advertisable to an advertisement. For example, if a direct classification assigns Honda to an advertisement, the "automotive" industry advertisable is a predecessor of Honda. Ancestral classification uses this relationship to associate automotive to the advertisement.

The media editor 264 populates the advertising information 534 area in the database 200. The user interface 240 accesses the data in the advertising information 534 area to generate statistical data.

The rate card 536 area contains data describing the cost of advertisements on a Web site. These costs include monetary values for each specific shape, size, or length of run that advertisers on the Internet 100 use to determine the cost of advertisement purchases. The system stores rate card data for each Web site that the system probes. The media editor 264 populates

the rate card 536 area in the database 200. The user interface 240 accesses the data in the rate card 536 area to generate statistical data.

The extended advertisable information 538 area contains additional information about specific types of advertisables not readily captured in the taxonomy hierarchy. Specifically, this includes additional information related to Web sites and companies, such as company contact information, Web site, and media kit URLs. This information extends the usefulness of the system by providing additional information to the client 140 about probed entities. For example, a client 140 may follow a hyperlink to company contact information directly from a system report. The media editor 264 populates the extended advertisable information 538 area in the database 200. The Web front end 250 accesses the data in the extended advertisable information 538 area to deliver value-added information to a client 140.

The advertising summary 540 partition of database 200 comprises database tables that are logically segmented into advertising statistics 542, data integrity 544, and advertising information summary 546 areas.

The advertising statistics 542 area describes how often an advertisement appears on each Web site. The system calculates and stores the following statistics in this area.

1. The proportion of page views that display an advertisement to the total number of page view. The system determines this statistic by analyzing the probing records.
2. The number of impressions that an advertisement received. The system determines this statistic by measuring traffic levels for the Web site using the site definition and traffic data, and multiplying that measurement by the proportion of page view calculated above.

3. The amount of spending that an advertisement received. The system determines this statistic by applying the rate card information to the number of impressions that the advertisement receives calculated above.

The statistical summarization system 230 populates the advertising statistics area 542 in the database 200. The Web front end 250 accesses the data in the advertising statistics 542 area to report spending, impressions, and advertising rotation to the client 140.

The data integrity 544 area contains in-depth information about statistical outliers and other potential anomalies resulting from trend and time slice analyses. This automated monitoring and analysis guarantees that the system will contain accurate analysis data. In addition, the system uses real world advertising information, as an input to the system, to verify the accuracy of the analysis data. The data integrity analysis system, performed by the statistical summarization system 230, populates the data integrity 544 area in the database 200. The operator 262 accesses the data integrity 544 area to detect potential errors and monitor general system health.

The advertising information summary 546 area summarizes advertising information in a format that is compact and easy to distribute. The system extracts the data in this area from the advertising support data 530 partition. While the data is not as descriptive as the data in the advertising support data 530 partition, it provides the ability to quickly perform a precise query. The advertising support data 530 partition associates each advertisement with a company, product, or industry. If the system associates multiple advertisables of the same type with an advertisement, a single advertisable is chosen to summary those associates using an assignment priority system, as follows:

1. Advertisables associated with an advertisement using direct classification receive the highest possible priority, "M".
2. Advertisables associated with an advertisement using location classification receive priority equal to the string length of the location prefix to which they are assigned, therefore, a long location prefix string will receive a higher priority than a short location prefix string.
3. Advertisables associated with an advertisement using ancestral classification receive the priority of the assigned ancestor.
4. The advertisement receives the highest priority advertisable in each type.
5. When two ancestors having the same type and priority are assigned to an advertisement, a conflict occurs and must be corrected by the media editor 264.

The statistical summarization system 230 populates the advertising information summary 546 area in the database 200. The Web front end 250 accesses the advertising information summary 546 area to generate reports for the client 140.

The following description discusses one embodiment of the database structure illustrated in Figure 5. This data model is encoded in an Oracle database. The table structure comprises three environments, the core schema, analysis schema, and front end. The core schema describes the back-end environment which allow the Cloudprober to direct live autonomous processes that continuously scour the Web noting advertising activity and operators and media editors for the present invention to direct, monitor and augment information provided by the Cloudprober. The analysis schema is the back-end environment that allows the advertisement sampling system, also known as OMNIAC, to apply rigorous data analysis procedures to information gathered

from the Web. The front end schema assists a client of the present invention with accessing data, building database query strings, and generating reports.

The database objects comprising the "core schema" are most frequently used by various components of the OMNIAC system. Code bases that rely on this schema include implementation of the back end processes that pull advertisements from the Web. Additionally, database schemas utilized by other components associated with OMNIAC are composed of some or all of the tables in the core schema. The core schema is conceptually composed of four sub-schemas including advertising, advertisements, probing, and sites. The advertising sub-schema holds information about "advertiseable" entities along with which entities each advertisement is advertising. The advertisements sub-schema describes the advertisements that the system has located and analyzed. The probing sub-schema defines "when", "where", and "how" for the probing process. The sites sub-schema describes Web sites, including structural site definitions and rate card information.

Of the four sub-schemas, Advertising serves the most general purpose and is therefore the most frequently referenced. The primary table in this sub-schema is ADVERTISABLE, which defines *advertisables*. Many of the conceptual entities in OMNIAC's universe are advertisables: industries, companies, products, services and Web sites are all defined here. The *type* field, referencing the ADVERTISABLE_TYPE table, differentiates between different types of advertisables, and the *parent* field organizes records hierarchically, establishing such relationships as industry-contains-company and company-produces-product.

In addition to the inherent grouping implied by the parent-child relationship defined in ADVERTISABLE, ADVERTISABLE_GROUP_MEMBER is used to further group

advertisables. Examples of groups defined in this way include automotive classes, travel industry segments, and types of computer hardware.

Other tables in the Advertising sub-schema serve to define what is advertised by each advertisement. ADVERTISES is used to associate advertisables directly with advertisements. LOCATION_ADVERTISES, CLASSIFIED_LOCATION and LOCATION_MATCHES also indirectly associate advertisables with advertisements via the advertisement's destination location.

"Advertisements" referred to above are references to records in AD, the primary table in the Advertisements sub-schema. The Advertisements sub-schema serves to define each advertisement in OMNIAC's universe. Every unique advertisement has a record in AD, along with one or more records in AD_DEFINITION. *Advertisement definitions* are unique XML fragments OMNIAC has retrieved from the Web. *Ads* are unique advertisements defined by sets of advertisement definitions determined to be equivalent during automated classification.

Other tables in Advertisements contain advertisement attributes, referenced by AD and AD_DEFINITION. AD_TECHNOLOGY describes known Web technologies used to render advertisements, while TEXT describes textual content for certain advertisements. FUZZY_WEB_LOCATION contains fuzzy locations found in advertisements. A *fuzzy location* is a URL that needs to be processed by the system, such as an anchor or image. Once OMNIAC has loaded a fuzzy location, a reference is made to MIME_CONTENT if the URL references an image, or DEST_WEB_LOCATION if the URL references another HTML page.

Moving on, the Probing sub-schema controls the behavior of OMNIAC's probing and advertisement extraction components. The primary purpose of this schema is to define target sets. A *target set* is a conceptual construct that instructs OMNIAC to fetch a set of pages at

certain intervals, extracting advertisements using a set of rules called *extraction rules*. Each target set is defined by a row in TARGET_SET.

The frequencies, locations, and extraction rules that make up each target set are defined in STROBE, AD_WEB_LOCATION, and EXTRACTION_RULE, respectively. The many-to-many relationships between rows in these tables are defined in TS_RUNS_AT, TS_PROBES, and TS_APPLIES.

The fourth and final sub-schema is Sites, which simply records information about Web sites. Each site or subsite defined in the advertisable hierarchy has a corresponding record in SITE_INFO, along with a number of rows in SITE_DOMAIN and SITE_MONTHLY_DATA. SITE_DOMAIN describes the physical structure of a site in terms of inclusive and exclusive URL stems. SITE_MONTHLY_DATA records advertising rate cards, third party traffic estimates, and cache statistics for each site on a monthly basis.

The analysis schema is an extension to the core schema that includes a number of additional tables populated by OMNIAC's analysis module. The *analysis module* is the unit in charge of processing information held in the core schema, producing a trim dataset that accurately describes advertising activity.

Like the core schema, the analysis schema is composed of four conceptual sub-schemas composed of tables implementing common functionality. These sub-schemas include advertising decomposition, advertisement view summarization, slot statistics, and site statisitics. The advertising decomposition sub-schema holds information about each advertisement in the system, including attributes and what the advertisement is advertising. The advertisement view summarization sub-schema summarizes advertisement views, recording how many times each advertisement was seen in each slot over the course of a day. The slot statistics sub-schema

describes advertisement rotation for each slot during each time period. The site statistics sub-schema describes site information, including advertisement rotation for each time period.

The primary table in the Advertising Decomposition sub-schema is AD_INFO, which contains de-normalized records describing advertisement attributes. AD_INFO records are keyed off of ID's in the AD table; an AD_INFO record exists for each AD record that has been completely classified and represents a valid advertisement. AD_INFO is populated by the analysis module from the advertising relationships described in the core schema tables ADVERTISES and LOCATION_ADVERTISES.

Fields in AD_INFO that specify what is advertised by an advertisement are: CATEGORY (industry), ORGANIZATION (company), ORGANIZATION_GROUP (industry segment), ORGANIZATION_OVERGROUP, COMMODITY (product/service), COMMODITY_GROUP (product/service segment), COMMODITY_OVERGROUP, and MESSAGE.

AD_INFO also includes fields describing a number of non-advertising attributes. FORMAT, referencing AD_SLOT_TYPE.ID, specifies the form factor of an advertisement. TECHNOLOGY, referencing AD_TECHNOLOGY2.ID, specifies the technology used to implement the advertisement. DEFINITION, IMAGE, and DESTINATION specify the AD_DEFINITION, IMAGE, and DEST_WEB_LOCATION records associated with the advertisement. These three fields mirror fields in the AD table.

The Advertising Decomposition schema contains a few tables in addition to AD_INFO. ADV_IMPLICATION is a cache of advertisable implications derived from the hierarchy in ADVERTISABLE. This is used to speed operation of the analysis module. AD_INFO_FLATTENED is a more readily queried version of AD_INFO containing

advertisement/advertisable pairs for each of the fields in AD_INFO that reference ADVERTISABLE. Finally, AD_TECHNOLOGY2 describes advertisement technologies understood by the analysis module that are presentable to the user in the front end.

The Advertisement View Summarization sub-schema covers the single table PLACEMENT_SUMMARY. PLACEMENT_SUMMARY is keyed off of day, advertisement, and slot, and contains, in the CNT field, the number of times an advertisement was seen in a slot on a particular day.

The analysis module populates PLACEMENT_SUMMARY by aggregating hits recorded in the APD_*n* tables, one of which exists for each day, *n* being the ID of the day in question. These tables are created and populated by the back-end as advertisement hits flow into the system.

The third sub-schema in the Analysis schema is Slot Statistics. This sub-schema describes advertisement behavior in the context of slots in addition to information about the slots themselves. A *slot* is a location on the Web in which advertisements rotate, currently defined in terms of the location ID (a reference to AD_WEB_LOCATION.ID) and extraction rule ID (a reference to EXTRACTION_RULE.ID).

The primary table in the Slot Statistics is SLOT_AD_VIEWS, which records the total views and relative frequency for each advertisement in each slot during each time period. The primary key of this table is composed of the fields PERIOD_TYPE, PERIOD, LOCATION_ID, RULE_ID and AD_ID. Two fields exist outside of the primary key: CNT holds the total number of advertisement views, and FREQUENCY holds the relative frequency.

Also in this sub-schema is SLOT_SUMMARY, which records general slot information outside the context of individual advertisements. Accordingly, this table is keyed off the

PERIOD_TYPE, PERIOD, LOCATION_ID and RULE_ID fields. The CNT field records total advertisement views in the slot; this field is divided into the SLOT_AD_VIEWS.CNT to determine relative frequency. Also in SLOT_SUMMARY is a SLOT_TYPE field that specifies the type of advertisement seen most frequently in the slot, and SITE_ID, which specifies which site the slot resides within.

The final table in the Slot Statistics sub-schema is SLOT_TYPE_COUNT. This table is used to determine which value to use in SLOT_SUMMARY.SLOT_TYPE. The number of times each advertisement format was seen is recorded, and the slot type that receives the most views is stuck into SLOT_SUMMARY.SLOT_TYPE.

Figure 6 is a functional block diagram of the advertising prevalence system 130. Memory 610 of the advertising prevalence system 130 stores the software components, in accordance with the present invention, that analyze traffic data on the Internet 100, sample the advertising data from that traffic data, and generate summarization data that characterizes the advertising data. The system bus 612 connects the memory 610 of the advertising prevalence system 130 to the transmission control protocol/internet protocol ("TCP/IP") network adapter 614, database 200, and central processor 616. The TCP/IP network adapter 614 is the mechanism that facilitates the passage of network traffic between the advertising prevalence system 130 and the Internet 100. The central processor 616 executes the programmed instructions stored in the memory 610.

Figure 6 shows the functional modules of the advertising prevalence system 130 arranged as an object model. The object model groups the object-oriented software programs into components that perform the major functions and applications in the advertising prevalence system 130. A suitable implementation of the object-oriented software program components of

Figure 6 may use the Enterprise JavaBeans specification. The book by Paul J. Perrone et al., entitled "Building Java Enterprise Systems with J2EE" (Sams Publishing, June 2000) provides a description of a Java enterprise application developed using the Enterprise JavaBeans specification. The book by Matthew Reynolds, entitled "Beginning E-Commerce" (Wrox Press Inc., 2000) provides a description of the use of an object model in the design of a Web server for an Electronic Commerce application.

The object model for the memory 610 of the advertising prevalence system 130 employs a three-tier architecture that includes the presentation tier 620, infrastructure objects partition 630, and business logic tier 640. The object model further divides the business logic tier 640 into two partitions, the application service objects partition 650 and data objects partition 660.

The presentation tier 620 retains the programs that manage the graphical user interface to the advertising prevalence system 130 for the client 140, account manager 260, operator 262, and media editor 264. In Figure 6, the presentation tier 620 includes the TCP/IP interface 622, the Web front end 624, and the user interface 626. A suitable implementation of the presentation tier 620 may use Java servlets to interact with the client 140, account manager 260, operator 262, and media editor 264 of the present invention via the hypertext transfer protocol ("HTTP"). The Java servlets run within a request/response server that handles request messages from the client 140, account manager 260, operator 262, and media editor 264 and returns response messages to the client 140, account manager 260, operator 262, and media editor 264. A Java servlet is a Java program that runs within a Web server environment. A Java servlet takes a request as input, parses the data, performs logic operations, and issues a response back to the client 140, account manager 260, operator 262, and media editor 264. The Java runtime platform pools the Java servlets to simultaneously service many requests. A TCP/IP interface 622 that uses Java servlets

functions as a Web server that communicates with the client 140, account manager 260, operator 262, and media editor 264 using the HTTP protocol. The TCP/IP interface 622 accepts HTTP requests from the client 140, account manager 260, operator 262, and media editor 264 and passes the information in the request to the visit object 642 in the business logic tier 640. Visit object 642 passes result information returned from the business logic tier 640 to the TCP/IP interface 622. The TCP/IP interface 622 sends these results back to the client 140, account manager 260, operator 262, and media editor 264 in an HTTP response. The TCP/IP interface 622 uses the TCP/IP network adapter 614 to exchange data via the Internet 100.

The infrastructure objects partition 630 retains the programs that perform administrative and system functions on behalf of the business logic tier 640. The infrastructure objects partition 630 includes the operating system 636, and an object oriented software program component for the database management system ("DBMS") interface 632, system administrator interface 634, and Java runtime platform 638.

The business logic tier 640 retains the programs that perform the substance of the present invention. The business logic tier 640 in Figure 6 includes multiple instances of the visit object 642. A separate instance of the visit object 642 exists for each client session initiated by either the Web front end 624 or user interface 626 via the TCP/IP interface 622. Each visit object 642 is a stateful session bean that includes a persistent storage area from initiation through termination of the client session, not just during a single interaction or method call. The persistent storage area retains information associated with either the URL 114, 116, 118 or the client 140, account manager 260, operator 262, and media editor 264. In addition, the persistent storage area retains data exchanged between the advertising prevalence system 130 and the

traffic sampling system 120 via the TCP/IP interface 622 such as the query result sets from a database 200 query.

When the traffic sampling system 120 finishes collecting information about a URL 114, 116, 118, it sends a message to the TCP/IP interface 622 that invokes a method to create a visit object 642 and stores information about the connection in the visit object 642 state. Visit object 642, in turn, invokes a method in the traffic analysis application 652 to process the information retrieved by the traffic sampling system 120. The traffic analysis application 652 stores the processed data from the anonymity system 310 and probe mapping system 320 in the traffic analysis data 662 state and the database 200. Figures 7A and 7B describe, in greater detail, the process that the traffic analysis application 652 follows for each URL 114, 116, 118 obtained from the traffic sampling system 120. Even though Figure 6 depicts the central processor 616 as controlling the traffic analysis application 652, it is to be understood that the function performed by the traffic analysis application 652 can be distributed to a separate system configured similarly to the advertising prevalence system 130.

After the traffic analysis application 652 processes a URL 114, 116, 118 identified by the traffic sampling system 120, the visit object 642 invokes a method in the advertising sampling application 654 to retrieve the URL 114, 116, 118 from the Web site 110. The advertising sampling application 654 processes the retrieved Web page by extracting embedded advertisements and classifying those advertisements. The advertising sampling application 654 stores the data retrieved by the Web page retrieval system 322 and processed by the Web browser emulation environment 324, advertisement extractor 326, and the structural classifier 328 in the advertising sampling data 664 state and the database 200. Figures 7A, 7C, and 7D describe, in greater detail, the process that the advertising sampling application follows for each

URL 114, 116, 118 identified by the traffic sampling system 120. Even though Figure 6 depicts the central processor 616 as controlling the advertising sampling application 654, a person skilled in the art will realize that the processing performed by the advertising sampling application 654 can be distributed to a separate system configured similarly to the advertising prevalence system 130.

After the traffic analysis application 652 and the advertisement sampling system 654 process the URL 114, 116, 118 identified by the traffic sampling system 120, the visit object 642 invokes a method in the statistical summarization application 656 to compute summary statistics for the data. The statistical summarization application 656 computes the advertising impression, spending, and valuation statistics for each advertisement embedded in URL 114, 116, 118. The statistical summarization application 656 stores the statistical data in the statistical summarization data 666 state and the database 200. Even though Figure 6 depicts the central processor 616 as controlling the statistical summarization application 656, a person skilled in the art realizes that the function performed by the statistical summarization application 656 can be distributed to a separate system configured similarly to the advertising prevalence system 130.

Figure 7A is a flow diagram of a process in the advertising prevalence system 130 that measures the value online advertisements by tracking and comparing online advertising activity across all major industries, channels, advertising formats, and types. Process 700 begins, at step 710, by sampling traffic data from the Internet 100. Figure 7B describes step 710 in greater detail. Step 720 uses the sampled traffic data from step 710 to perform site selection, and define and refine site definitions for the advertising prevalence system 130. Step 730 uses the result of the site selection and definition process to generate a probe map based on the sampled traffic data. Figure 7C describes step 730 in greater detail. Step 740 uses the probe map from step 730

to visit the Internet 100 to gather sample data from the probe sites identified in step 730. Figure 7D describes step 740 in greater detail. For each URL retrieved in step 740, step 750 extracts the advertisements from the URL, step 760 classifies each advertisement, and step 770 calculates the statistics for each advertisement. Figures 7E and 7F describe, respectively, steps 760 and 770 in greater detail. Finally, process 700 performs data integrity checks in step 780 to verify the integrity of the data and analysis results in the system.

Figure 7B is a flow diagram that describes, in greater detail, the process of sampling traffic data from Figure 7A, step 710. Process 710 begins in step 711 by gathering data from a Web traffic monitor such as the traffic sampling system 120. Process 710 strips the user information from the data retrieved by the Web traffic monitor in step 712 to cleanse the data and guarantee the anonymity of the sample. For each URL in the cleansed sample, step 713 measures the number of Web page views observed in the traffic data. Step 714 completes process 710 by statistically extrapolating the measured number of Web page view in the sample to whole universe of the Internet 100.

Figure 7C is a flow diagram that describes, in greater detail, the process of generating a probe map based on sampled traffic data from Figure 7A, step 730. Process 730 begins in step 731 by analyzing a subset of the sample traffic data that falls within eligible site definitions. Following the analysis in step 731, step 732 builds an initial probe map based on the sample traffic data. Step 733 analyzes the historic advertisement measurement results in the database 200 for the URLs in the initial probe map. Step 734 uses these historic results, as well as, system parameters to optimize the sampling plan. Step 735 completes process 730 by monitoring the sample results and adjusting the system as necessary.

Figure 7D is a flow diagram that describes, in greater detail, the process of probing the Internet 100 to gather sample data from Figure 7A, step 740. Process 740 begins in step 741 by fetching a Web page from the Internet 100. The Web page from step 741 is passed to a Web browser emulation environment in step 742 to simulate the display of that Web page in a browser. This simulation allows the advertising prevalence system 130 to detect advertisements embedded in the Web page. These advertisements may be embedded in JavaScript code, Java applet or servlet code, or common gateway interface code such as a Perl script. In addition, the simulation in step 742 allows the advertising prevalence system 130 to detect dynamic and interactive advertisements in the Web page. After the simulation in step 742, step 743 extracts the advertisement data from the Web page and step 744 stores the advertisement data in the database 200. Step 745 determines whether process 740 needs to fetch another Web page to gather more sample data. In the preferred embodiment, process 740 continuously samples Web pages from the Internet 100. A person skilled in the art realizes that the functionality performed by step 745 can be associated with a scheduling system that will schedule the probing of the Internet 100 to gather the sample advertising data.

Figure 7E is a flow diagram that describes, in greater detail, the process of classifying the advertising data from Figure 7A, step 760. Process 760 begins the analysis of advertisement fragments in step 761 by determining whether the fragment is a duplicate. When the advertising prevalence system 130 encounters an advertisement fragment for the first time, step 762 analyzes the internal structure of the fragment. Following step 762, or when step 761 determines that the advertisement fragment is a duplicate, step 763 retrieves the external content of the advertisement from the Web page. Step 764 then compares the external content to previously observed advertisements. Step 765 analyzes the result of the comparison in step 764 to

determine whether the advertisement is a duplicate. When the advertising prevalence system 130 encounters an advertisement for the first time, step 766 begins processing the new advertisement by recording the structure of the new advertisement in the database 200. Step 767 then performs automated advertisement classification and stores the classification types in the database 200. Step 768 completes processing of a new advertisement by performing human verification of the advertisement classifications. Following step 768, or when step 765 determines that the advertisement is a duplicate, step 769 updates the advertisement viewing log in the database 200 to indicate the observation of the advertisement.

Figure 7F is a flow diagram that describes, in greater detail, the process of calculating advertising statistics from Figure 7A, step 770. Process 770 begins the calculation of the advertising statistics in step 771 by summarizing the advertising measurement results. In step 772, process 770 uses the probe map generated in step 730 to weight the advertising measurement results. The advertising frequency is calculated in step 773 for each Web page request. Step 774 uses the sample traffic data from step 710 and the advertising frequency from step 773 to calculate the advertising impressions for each advertisement. Step 775 completes process 770 by calculating the advertisement spending by combining the advertising impressions from step 774 and the rate card data input by the media editor 264 with the rate card collection 348 module of the user interface 240.

Although embodiments disclosed in the present invention describe a fully functioning system, it is to be understood that other embodiments exist which are equivalent to the embodiments disclosed herein. Since numerous modifications and variations will occur to those who review the instant application, the present invention is not limited to the exact construction

and operation illustrated and described herein. Accordingly, all suitable modifications and equivalents which may be resorted to are intended to fall within the scope of the claims.

We claim:

1. A system for estimating prevalence of digital content on the World-Wide-Web, comprising:

an estimating device for estimating the global traffic to a plurality of Web sites to provide traffic data;

a sampling device for statistically sampling the contents of said plurality of Web sites to provide sampling data;

a storage device for storing said traffic data and said sampling data; and

an accessing device for accessing said traffic data and said sampling data stored in said storage device.

2. The system of claim 1, wherein said estimating device being a globally distributed set of proxy cache servers.

3. The system of claim 1, wherein said estimating device computes for each Web site, the number of impressions of an advertisement on a Web page on said each Web site.

4. The system of claim 1, wherein said sampling device includes:

a prober for periodically fetching pages from each Web site;

an extractor for extracting fragments from said pages; and

a classifier for classifying said fragments.

5. The system of claim 1, wherein said accessing device generates reports in accordance with a predetermined criteria.

6. A method of estimating prevalence of digital content on the World-Wide-Web, comprising the steps of:

estimating the global traffic to a plurality of Web sites to provide traffic data;

statistically sampling the contents of said plurality of Web sites to provide sampling data;

storing said traffic data and said sampling data;

accessing said traffic data and said sampling data stored in said storage device to generate reports.

7. A system for estimating the prevalence of digital content on a network, wherein the network connects to at least one network site having at least one network server to access at least one uniform resource locator, the system comprising:

a database;

a traffic analysis system that receives a traffic data sample from a traffic sampling system and stores the traffic data sample in the database, wherein the traffic sampling system is connected to the network, and wherein the traffic data sample includes said at least one uniform resource locator;

an digital content sampling system connected to the network, wherein the digital content sampling system retrieves at least one digital content resource from said at least one uniform resource locator and stores said at least one digital content resource in the database; and

a statistical summarization system that creates summarization data that describes said at least one digital content resource and stores the summarization data in the database.

8. The system of claim 7, further comprising:

a Web front end connected to the network, wherein a client can use the Web front end to access the database, and wherein the client uses a browser to connect to the Web front end; and

9. The system of claim 7, further comprising:

a user interface that an account manager, operator, or media editor can use to administer the system.

10. The system of claim 7, wherein the network is the Internet, and wherein the network site is a Web site.

11. The system of claim 7, wherein the traffic analysis system further comprises:

an anonymity system that receives the traffic data sample from the traffic sampling system and produces a clean traffic data sample; and

a traffic summarization system that produces a summarization of the clean traffic data sample and stores the traffic data sample in the database.

12. The system of claim 11, wherein the anonymity system produces a clean traffic data sample by removing network address or cookie data from the traffic data sample.

13. The system of claim 11, wherein the summarization of the clean traffic data sample includes a reference to said at least one uniform resource locator and a tally of the number of times said at least one uniform resource locator was requested.

14. The system of claim 7, wherein the digital content sampling system further comprises:

a probe mapping system that uses the summarization data to create a probe map for the network, wherein the probe map includes a mapping for said at least one uniform resource locator;

a uniform resource locator retrieval system that retrieves said at least one uniform resource locator from the network server;

a browser emulation environment that conducts a simulation of the display of said at least one uniform resource locator in a browser;

a digital content extractor that retrieves said at least one digital content resource from said at least one uniform resource locator and stores said at least one digital content resource in the database;

a structural classifier that determines at least one classification type for said at least one digital content resource and stores said at least one classification type in the database; and

a statistical summarization of the prevalence of the digital content.

15. The system of claim 14, wherein the probe map comprises:

a probability of the likelihood that said at least one uniform resource location will be sampled; and

a scale that determines the contribution of said at least one uniform resource location to .

16. The system of claim 14, wherein the simulation includes executing a program embedded in said at least one uniform resource locator.

17. The system of claim 16, wherein the program is a JavaScript script, Java applet, Perl script, or common gateway interface program.

18. The system of claim 14, wherein the simulation includes executing dynamic digital content in said at least one uniform resource locator.

19. The system of claim 18, wherein the dynamic content is an interlaced GIF image, MPEG movie, or MP3 audio file.

20. The system of claim 14, wherein the digital content extractor retrieves said at least one digital content resource from said at least one uniform resource locator by applying a rule set defined by a media editor.

21. The system of claim 14, wherein the digital content extractor retrieves said at least one digital content resource from said at least one uniform resource locator by using an automated digital content detection system.

22. The system of claim 21, wherein the automatic digital detection system comprises:

a structural detector that locates particular XML structures; and

a feature detector that locates particular XML features within said structures.

23. The system of claim 14, wherein the structural classifier determines said at least one classification type for said at least one advertisement.

24. The system of claim 7, wherein the user interface comprises:

a system account management interface, wherein the account manager uses the system account management interface to create and modify an account for the client on the system;

a site administration interface, wherein the operator uses the site administration interface;

a taxonomy administration interface, wherein the media editor uses the taxonomy administration interface;

an advertising content classification interface, wherein the media editor uses the advertising content classification interface; and

a rate card collection interface, wherein the media editor uses the rate card collection interface.

25. A system for estimating prevalence of dynamic content on a network, comprising:

a memory device; and

a processor disposed in communication with said memory device, said processor configured to:

collect a sample of traffic data to a plurality of Web sites;

compute a number of impressions of a Web advertisement from each of a plurality of Web sites to generate traffic data,

retrieve sample contents of each of said Web sites to generate sampling data, and

generate prevalence estimates of said dynamic content from said traffic data and said sampling data.

26. The system of claim 25 wherein said processor is further configured to sample said contents by retrieving Web pages from each of said Web sites, extract fragments from said Web pages and classify said fragments.

27. The system of claim 25 wherein said processor is further configured to generate said traffic data by retrieving anonymous traffic data samples.

28. The system of claim 27 wherein said processor is configured to retrieve anonymous data samples by removing data from traffic data samples which identify users on said network.

29. The system of claim 25 wherein said processor is further configured to classify fragments within said sampling data.

30. The system of claim 29 wherein said processor is further configured to classify fragments by analyzing each fragment for uniqueness, and adding information to a database regarding the uniqueness of said fragment.

31. The system of claim 30 wherein said processor is configured to classify said fragments by detecting duplicate fragments.

32. The system of claim 25 wherein said processor is further configured to interact with a user interface for use in administering said system.

33. The system of claim 25 wherein said processor is further configured to generate said traffic data to include uniform resource locator information regarding said plurality of Web sites.

34. The system of claim 25 wherein said processor is further configured to perform data integrity monitoring of said sample data.

35. The system of claim 25 wherein said processor is configured to serve as an automatic advertisement detection system.

36. The system of claim 35 wherein said processor is configured to serve as an automatic advertisement detection system by using heuristics to detect advertising within HTML or XML documents, and normalizing detected HTML or XML content into a hierarchical form.

37. A method for using a computer to estimate prevalence of dynamic content on a network, comprising:

computing a number of impressions of a Web advertisement from each of a plurality of Web sites to generate traffic data;

retrieving sample contents of each of said Web sites, using said computer, to generate sampling data; and

generating prevalence estimates of said dynamic content from said traffic data and said sampling data.

38. The method of claim 37 wherein said retrieving comprises retrieving Web pages from each of said Web sites, extracting fragments from said Web pages and classifying said fragments.

39. The method of claim 37 wherein said traffic data is generated by retrieving anonymous traffic data samples.

40. The method of claim 39 wherein said retrieving comprises retrieving anonymous data samples by removing data from traffic data samples which identify users on said network.

41. The method of claim 37 further comprising classifying fragments within said sampling data.

42. The method of claim 41 wherein said classifying fragments comprises analyzing each fragment for uniqueness, and adding information to a database regarding the uniqueness of each said fragment.

43. The method of claim 42 further comprising classifying said fragments by detecting duplicate fragments.

44. The method of claim 37 further comprising interacting with a user interface to administer said system.

45. The method of claim 37 further comprising generating said traffic data to include uniform resource locator information regarding said plurality of Web sites.

46. The method of claim 37 further comprising performing data integrity monitoring of said sample data.

47. The method of claim 37 further comprising performing automatic advertisement detection by using heuristics to detect advertising within HTML or XML documents, and normalizing detected HTML or XML content into a hierarchical form.

48. A computer readable medium comprising:

code for computing a number of impressions of a Web advertisement from each of a plurality of Web sites to generate traffic data;

code for retrieving sample contents of each of said Web sites to generate sampling data; and

code for generating prevalence estimates of dynamic content from said traffic data and said sampling data.

49. The computer readable medium of claim 48 further comprising code to extract fragments from said Web pages and classify said fragments.

50. A system for estimating prevalence of dynamic content on a network, comprising:

means for computing a number of impressions of a Web advertisement from each of a plurality of Web sites to generate traffic data;

means for retrieving sample contents of each of said Web sites, using said computer, to generate sampling data; and

means for generating prevalence estimates of said dynamic content from said traffic data and said sampling data.

51. The system of claim 50 further comprising:

means for classifying fragments extracted from said Web pages.

52. The system of claim 50 further comprising:

means for anonymizing said traffic data.

53. A system of estimating prevalence of dynamic content on the World-Wide-Web, comprising:

means for estimating global traffic to a plurality of Web sites to provide traffic data;

means for statistically sampling the contents of said plurality of Web sites to provide sampling data;

means for storing said traffic data and said sampling data; and

means for accessing said traffic data and said sampling data stored in said storage device to generate prevalance estimates and reports therefrom.

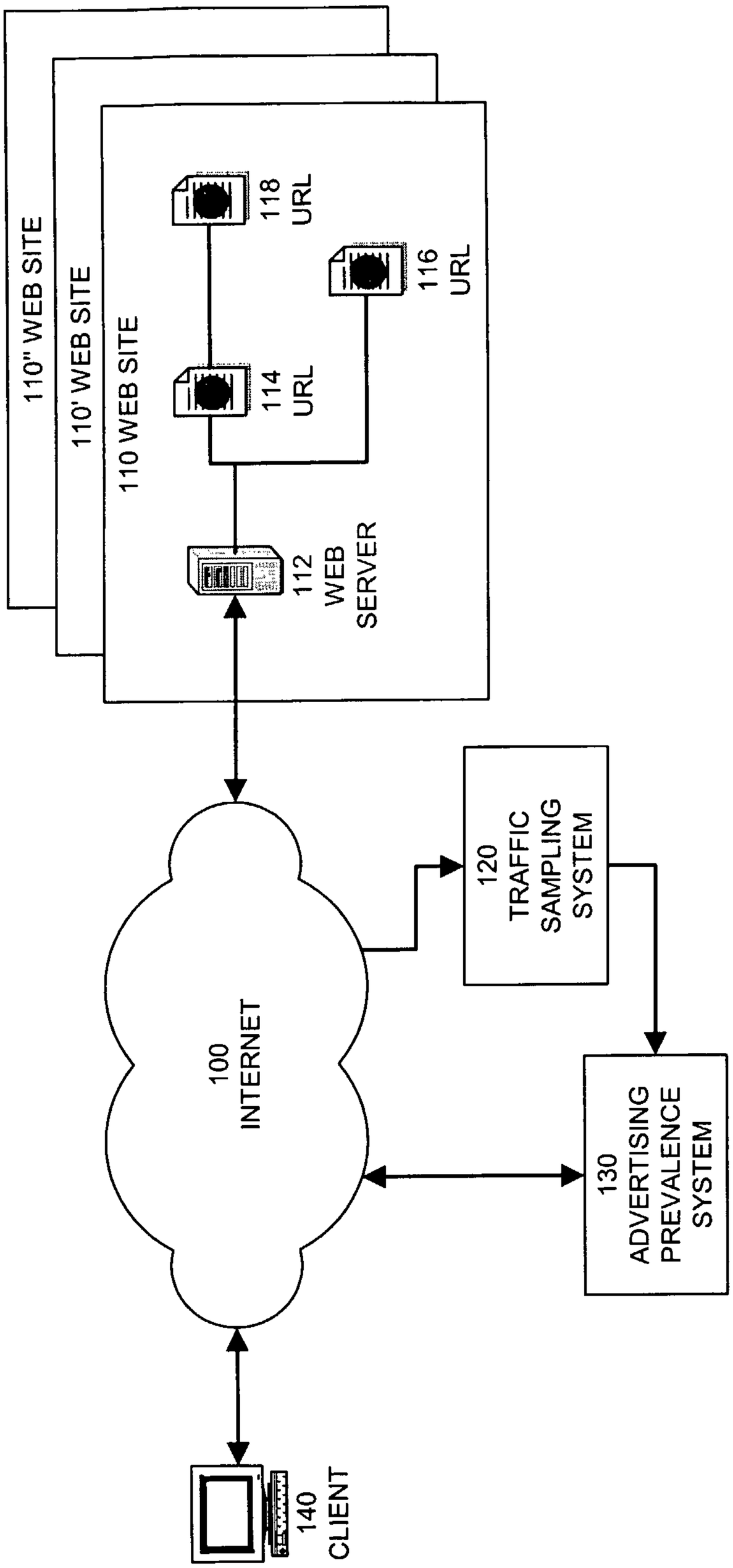

110" WEB SITE
110' WEB SITE
110 WEB SITE
118 URL
116 URL
114 URL
112 WEB SERVER
120 TRAFFIC SAMPLING SYSTEM
100 INTERNET
130 ADVERTISING PREVALENCE SYSTEM
140 CLIENT

FIG. 1

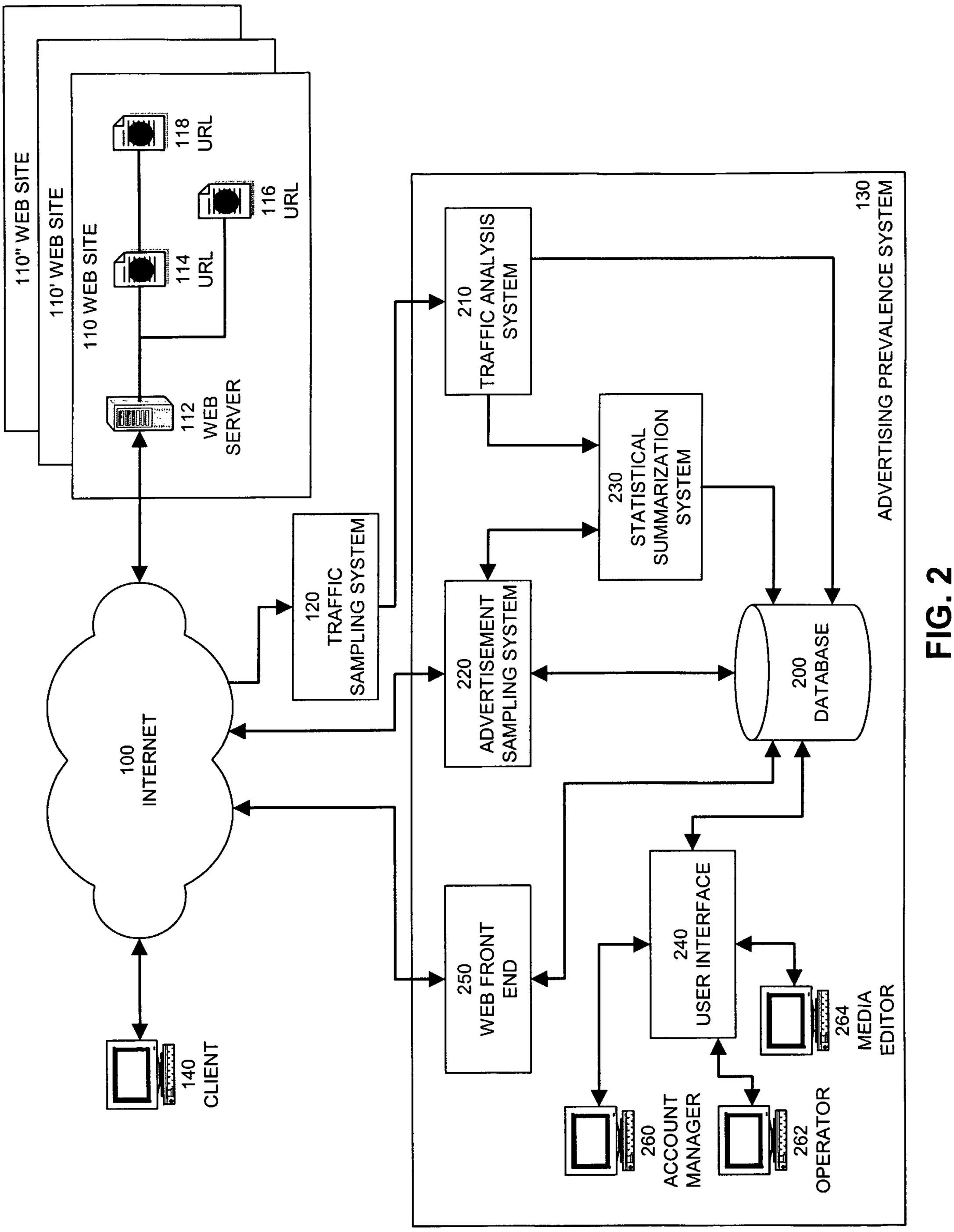

110" WEB SITE
110' WEB SITE
110 WEB SITE
118 URL
114 URL
116 URL
112 WEB SERVER
130 ADVERTISING PREVALENCE SYSTEM
210 TRAFFIC ANALYSIS SYSTEM
230 STATISTICAL SUMMARIZATION SYSTEM
120 TRAFFIC SAMPLING SYSTEM
220 ADVERTISEMENT SAMPLING SYSTEM
200 DATABASE
100 INTERNET
240 USER INTERFACE
250 WEB FRONT END
260 ACCOUNT MANAGER
262 OPERATOR
264 MEDIA EDITOR
140 CLIENT

FIG. 2

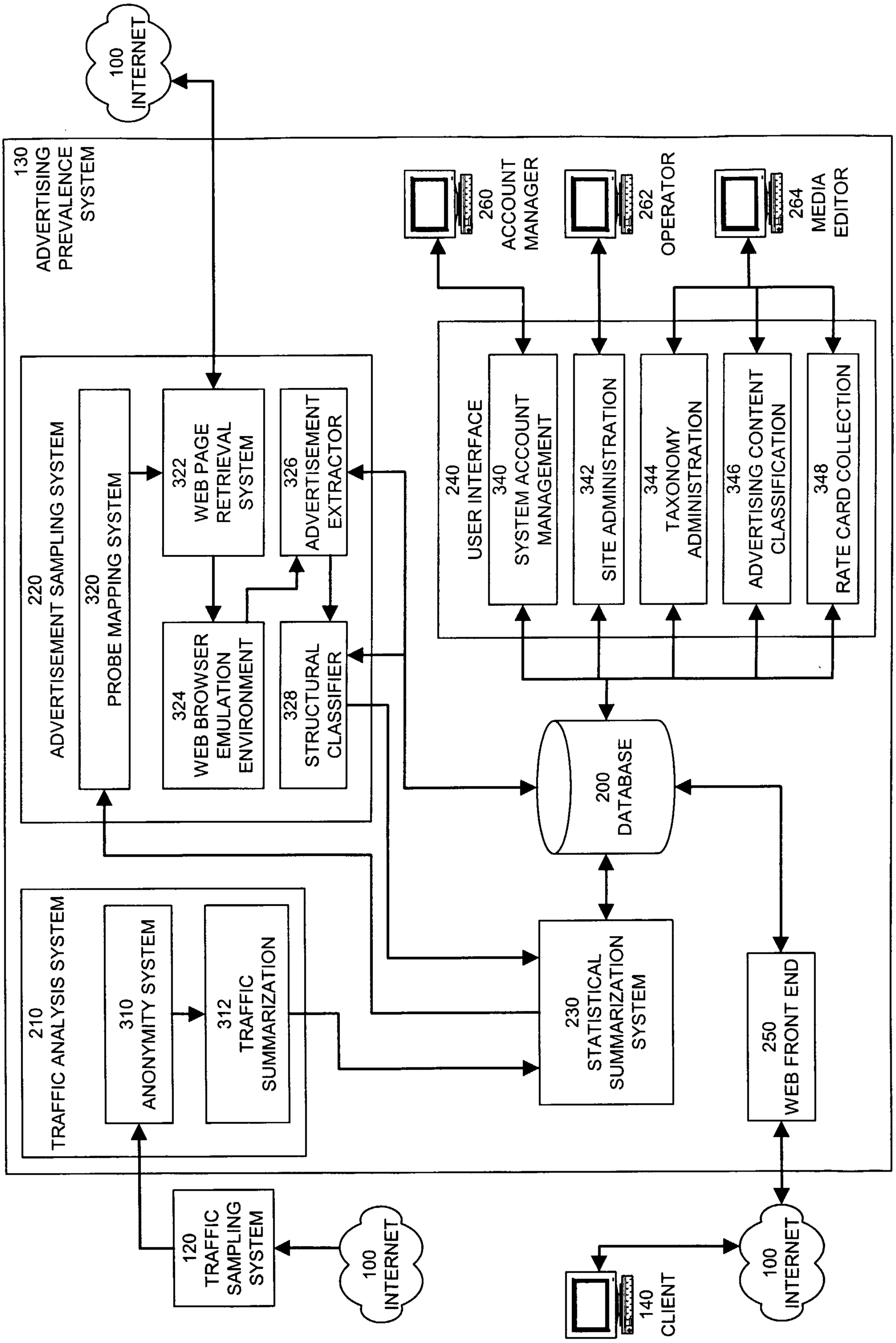

100 INTERNET
130 ADVERTISING PREVALENCE SYSTEM
260 ACCOUNT MANAGER
262 OPERATOR
264 MEDIA EDITOR
220 ADVERTISEMENT SAMPLING SYSTEM
320 PROBE MAPPING SYSTEM
322 WEB PAGE RETRIEVAL SYSTEM
326 ADVERTISEMENT EXTRACTOR
324 WEB BROWSER EMULATION ENVIRONMENT
328 STRUCTURAL CLASSIFIER
240 USER INTERFACE
340 SYSTEM ACCOUNT MANAGEMENT
342 SITE ADMINISTRATION
344 TAXONOMY ADMINISTRATION
346 ADVERTISING CONTENT CLASSIFICATION
348 RATE CARD COLLECTION
200 DATABASE
210 TRAFFIC ANALYSIS SYSTEM
310 ANONYMITY SYSTEM
312 TRAFFIC SUMMARIZATION
230 STATISTICAL SUMMARIZATION SYSTEM
250 WEB FRONT END
120 TRAFFIC SAMPLING SYSTEM
100 INTERNET
140 CLIENT
100 INTERNET

FIG. 3

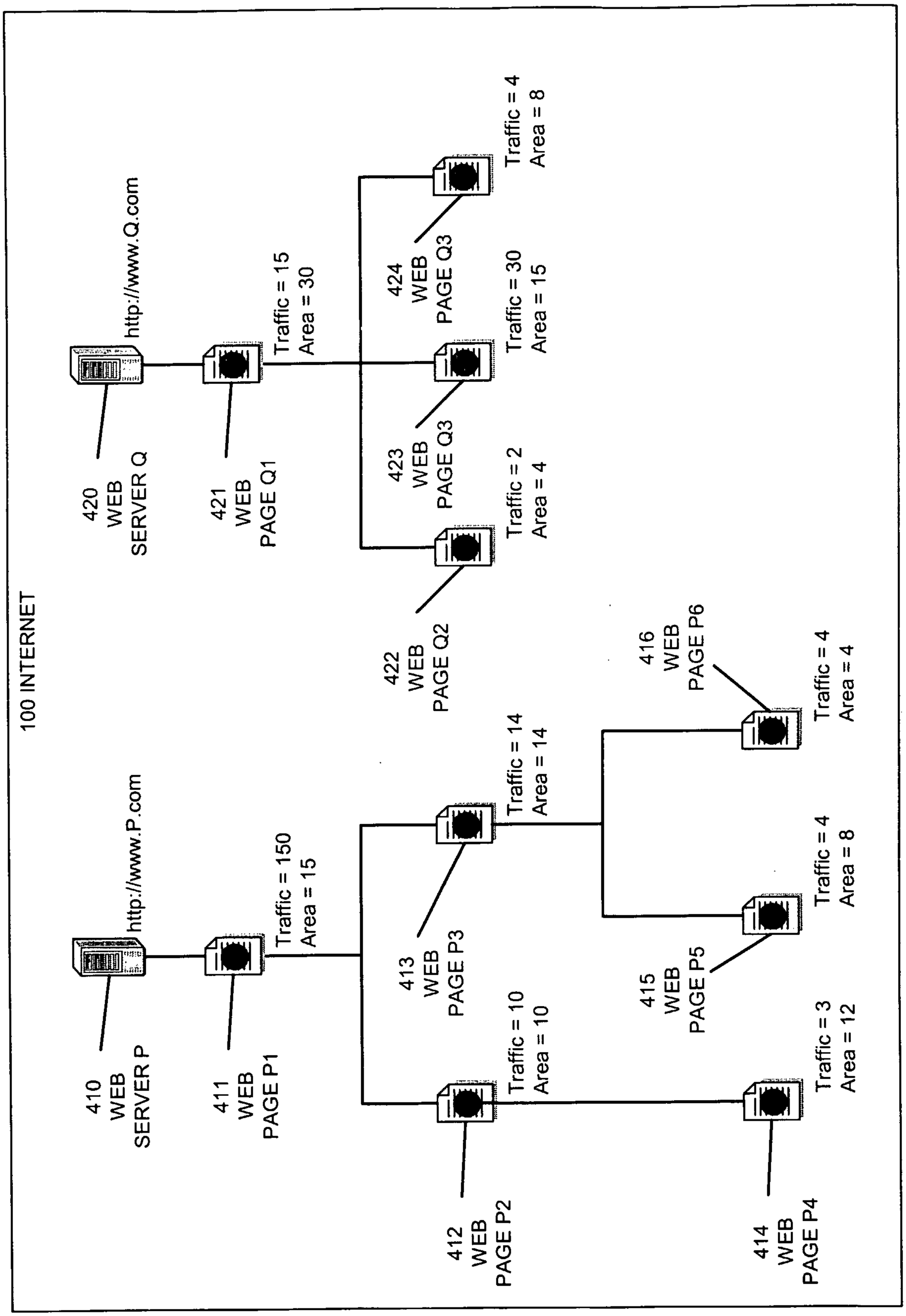

100 INTERNET
410 WEB SERVER P
http://www.P.com
411 WEB PAGE P1
Traffic = 150
Area = 15
412 WEB PAGE P2
Traffic = 10
Area = 10
413 WEB PAGE P3
Traffic = 14
Area = 14
414 WEB PAGE P4
Traffic = 3
Area = 12
415 WEB PAGE P5
Traffic = 4
Area = 8
416 WEB PAGE P6
Traffic = 4
Area = 4
420 WEB SERVER Q
http://www.Q.com
421 WEB PAGE Q1
Traffic = 15
Area = 30
422 WEB PAGE Q2
Traffic = 2
Area = 4
423 WEB PAGE Q3
Traffic = 30
Area = 15
424 WEB PAGE Q3
Traffic = 4
Area = 8

FIG. 4A

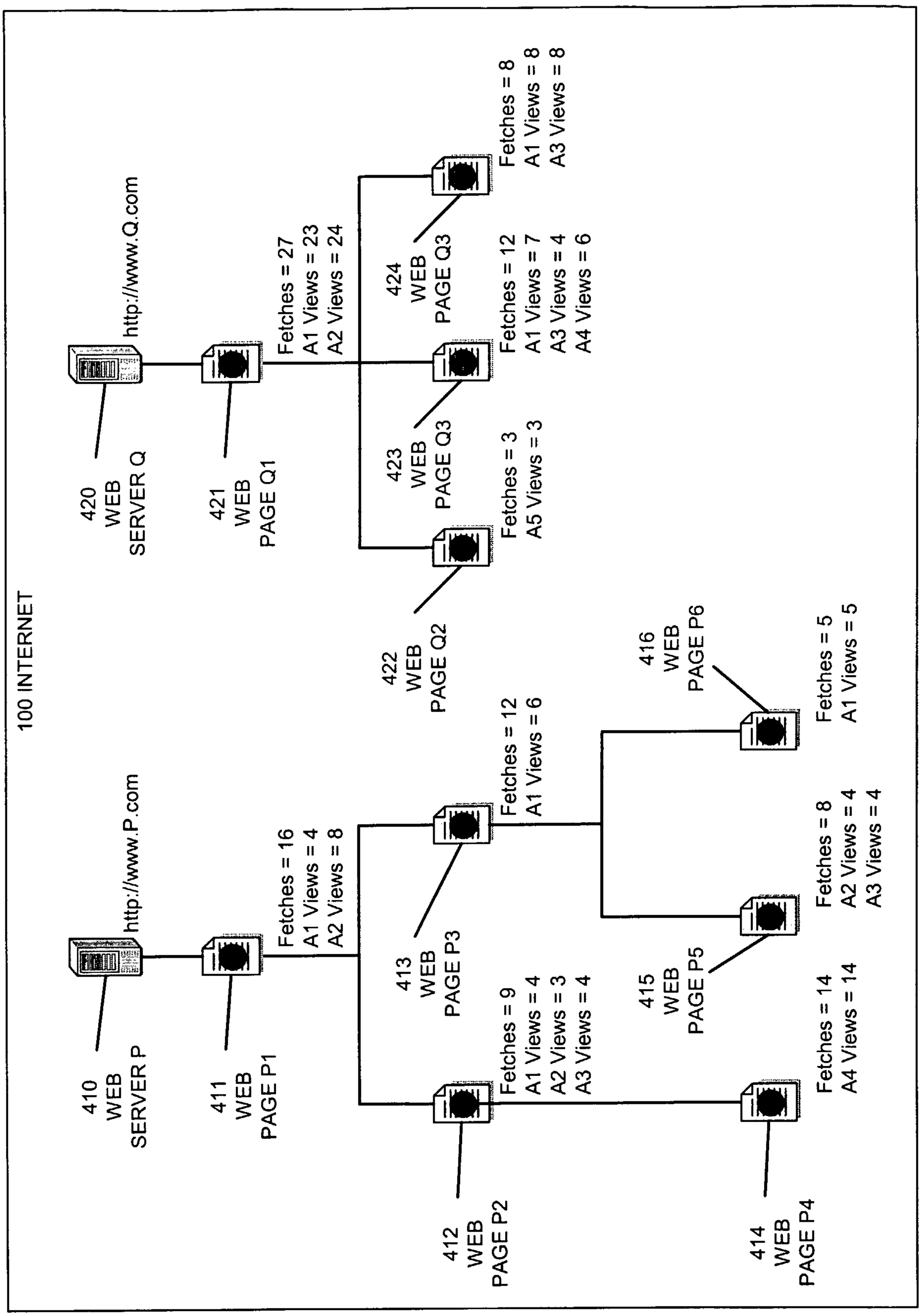

100 INTERNET
410 WEB SERVER P
http://www.P.com
411 WEB PAGE P1
Fetches = 16
A1 Views = 4
A2 Views = 8
412 WEB PAGE P2
Fetches = 14
A4 Views = 14
413 WEB PAGE P3
Fetches = 9
A1 Views = 4
A2 Views = 3
A3 Views = 4
414 WEB PAGE P4
Fetches = 8
A2 Views = 4
A3 Views = 4
415 WEB PAGE P5
Fetches = 12
A1 Views = 6
416 WEB PAGE P6
Fetches = 5
A1 Views = 5
420 WEB SERVER Q
http://www.Q.com
421 WEB PAGE Q1
Fetches = 27
A1 Views = 23
A2 Views = 24
422 WEB PAGE Q2
Fetches = 3
A5 Views = 3
423 WEB PAGE Q3
Fetches = 12
A1 Views = 7
A3 Views = 4
A4 Views = 6
424 WEB PAGE Q3
Fetches = 8
A1 Views = 8
A3 Views = 8

FIG. 4B

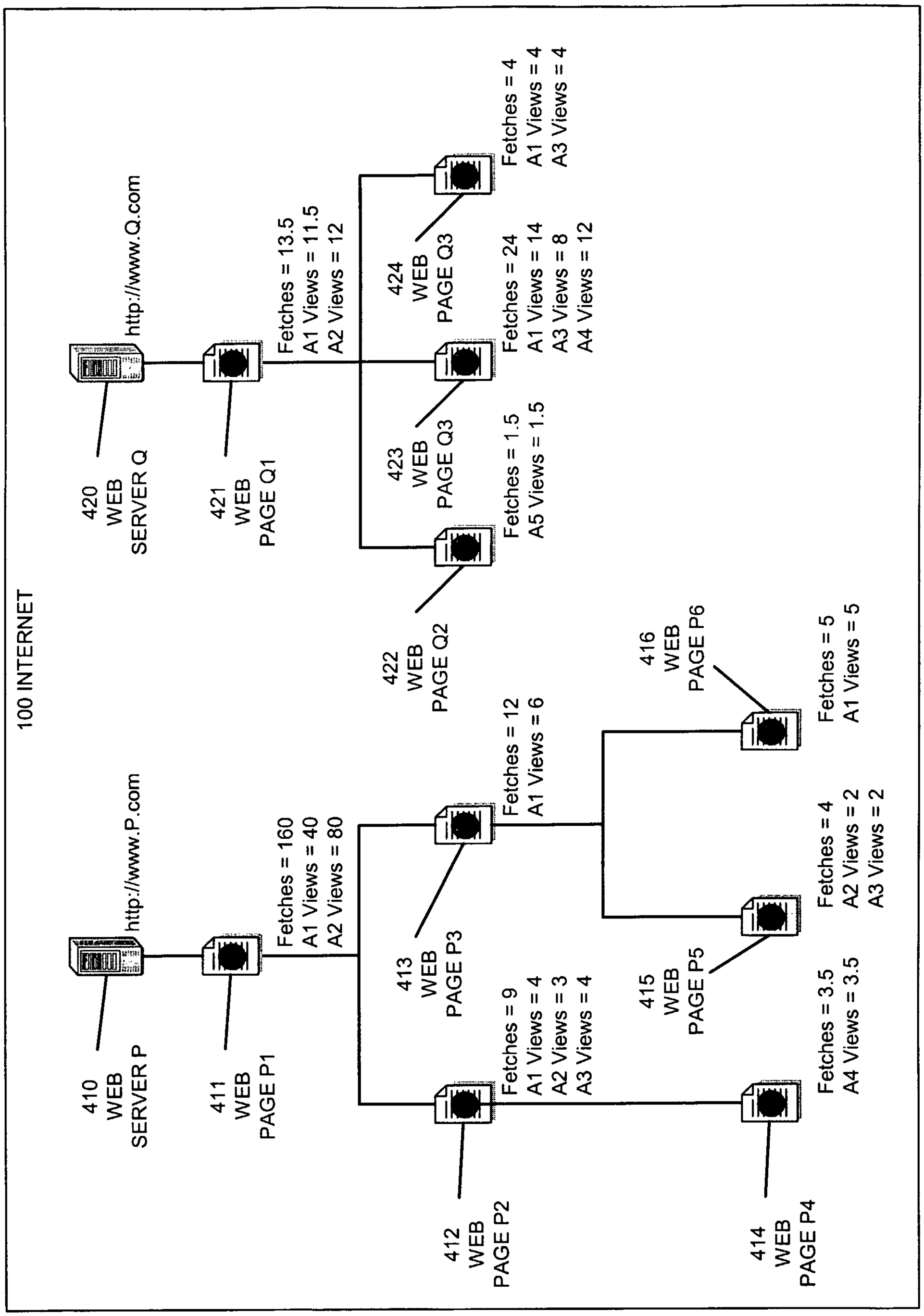

100 INTERNET
http://www.Q.com
420 WEB SERVER Q
421 WEB PAGE Q1
Fetches = 13.5
A1 Views = 11.5
A2 Views = 12
424 WEB PAGE Q3
Fetches = 24
A1 Views = 14
A3 Views = 8
A4 Views = 12
423 WEB PAGE Q3
Fetches = 1.5
A5 Views = 1.5
422 WEB PAGE Q2
http://www.P.com
410 WEB SERVER P
411 WEB PAGE P1
Fetches = 160
A1 Views = 40
A2 Views = 80
413 WEB PAGE P3
Fetches = 12
A1 Views = 6
416 WEB PAGE P6
Fetches = 5
A1 Views = 5
Fetches = 9
A1 Views = 4
A2 Views = 3
A3 Views = 4
415 WEB PAGE P5
Fetches = 4
A2 Views = 2
A3 Views = 2
Fetches = 3.5
A4 Views = 3.5
412 WEB PAGE P2
414 WEB PAGE P4

FIG. 4C

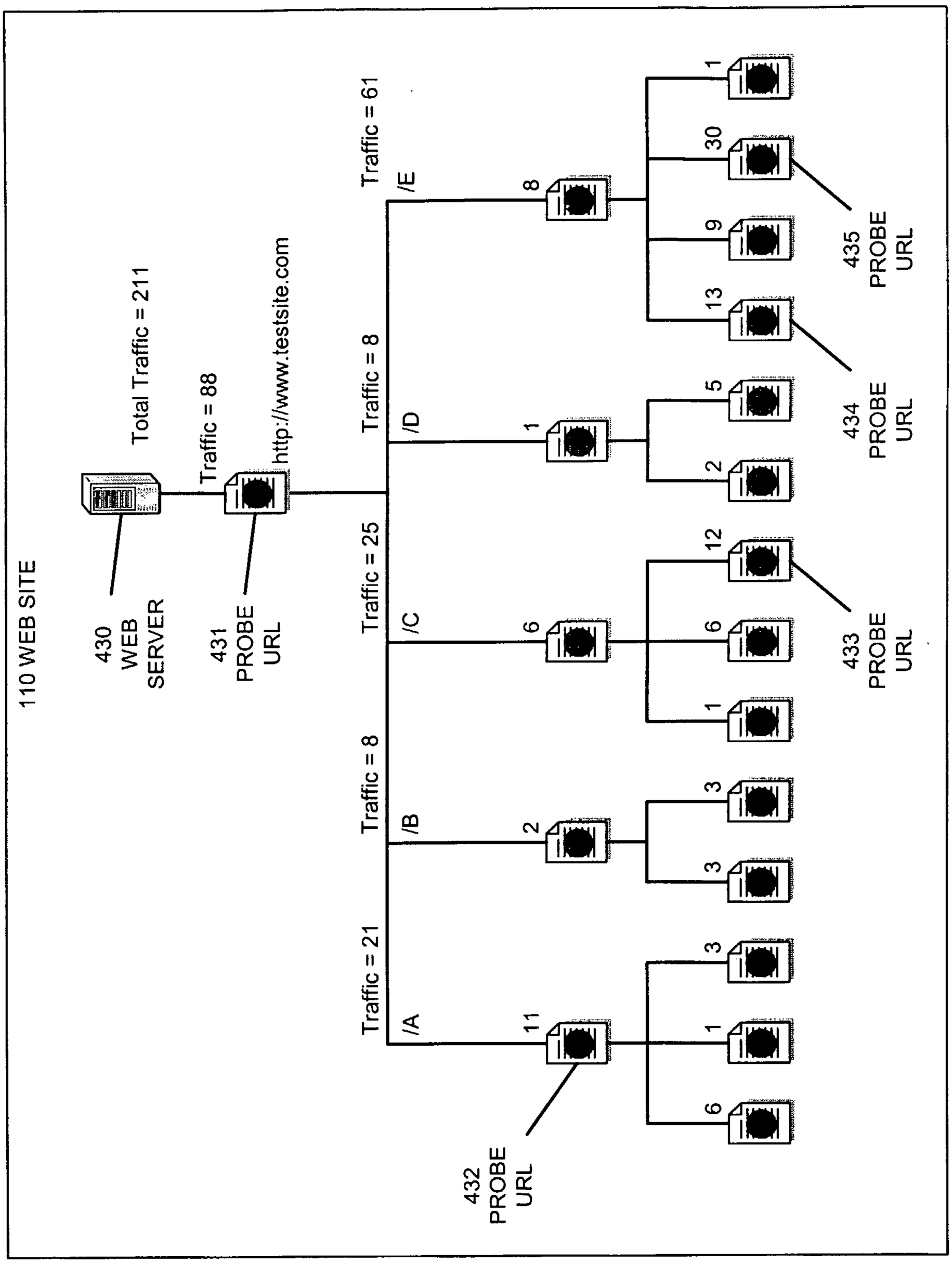

110 WEB SITE
430 WEB SERVER
Total Traffic = 211
431 PROBE URL
Traffic = 88
http://www.testsite.com
Traffic = 8
Traffic = 25
Traffic = 8
Traffic = 21
/A
/B
/C
/D
/E
11
2
6
1
8
432 PROBE URL
433 PROBE URL
434 PROBE URL
435 PROBE URL
1
3
6
12
2
5
13
9
30

FIG. 4D

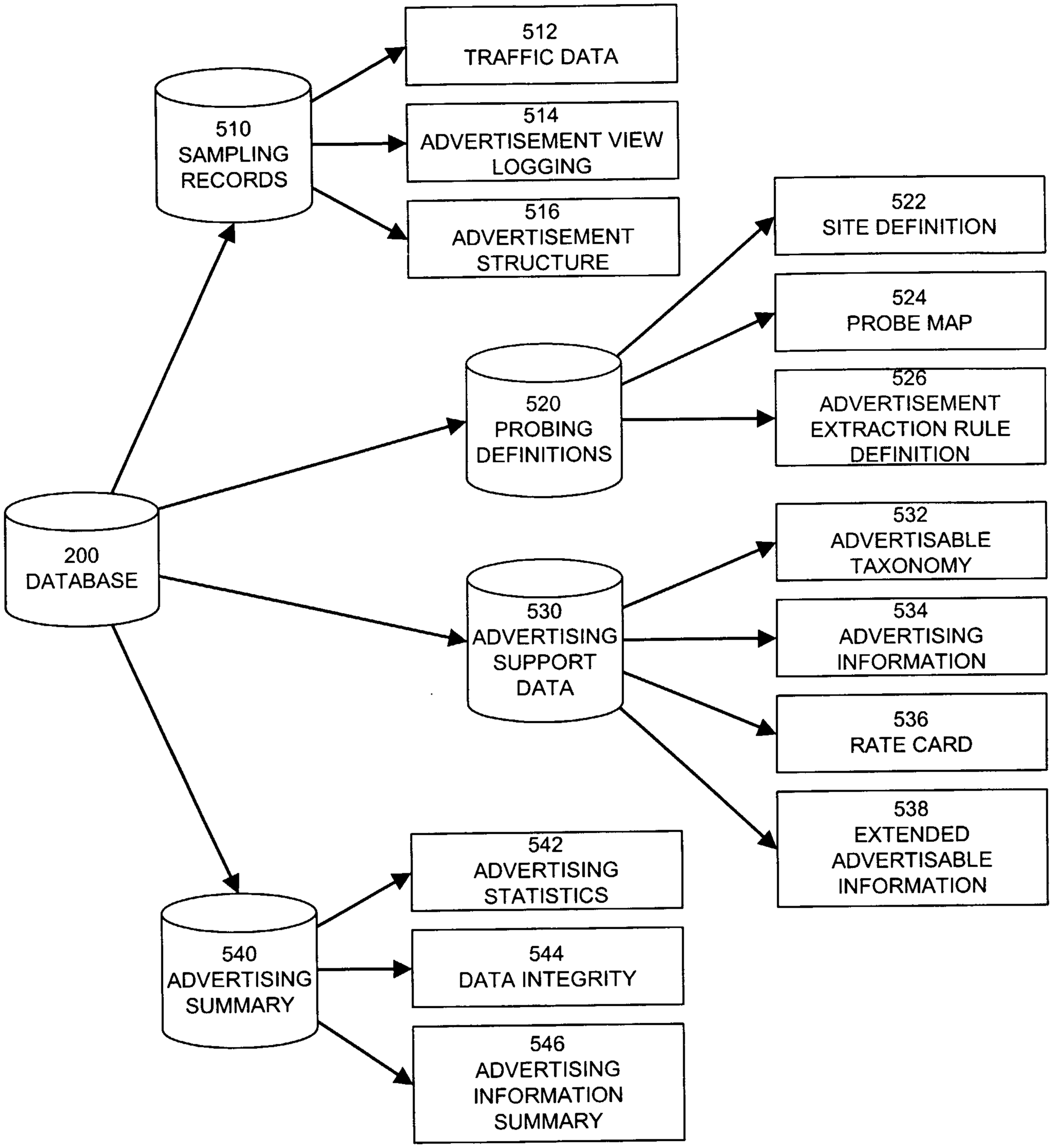

512
TRAFFIC DATA
510
SAMPLING
RECORDS
514
ADVERTISEMENT VIEW
LOGGING
516
ADVERTISEMENT
STRUCTURE
522
SITE DEFINITION
524
PROBE MAP
520
PROBING
DEFINITIONS
526
ADVERTISEMENT
EXTRACTION RULE
DEFINITION
200
DATABASE
532
ADVERTISABLE
TAXONOMY
530
ADVERTISING
SUPPORT
DATA
534
ADVERTISING
INFORMATION
536
RATE CARD
538
EXTENDED
ADVERTISABLE
INFORMATION
542
ADVERTISING
STATISTICS
540
ADVERTISING
SUMMARY
544
DATA INTEGRITY
546
ADVERTISING
INFORMATION
SUMMARY

FIG. 5

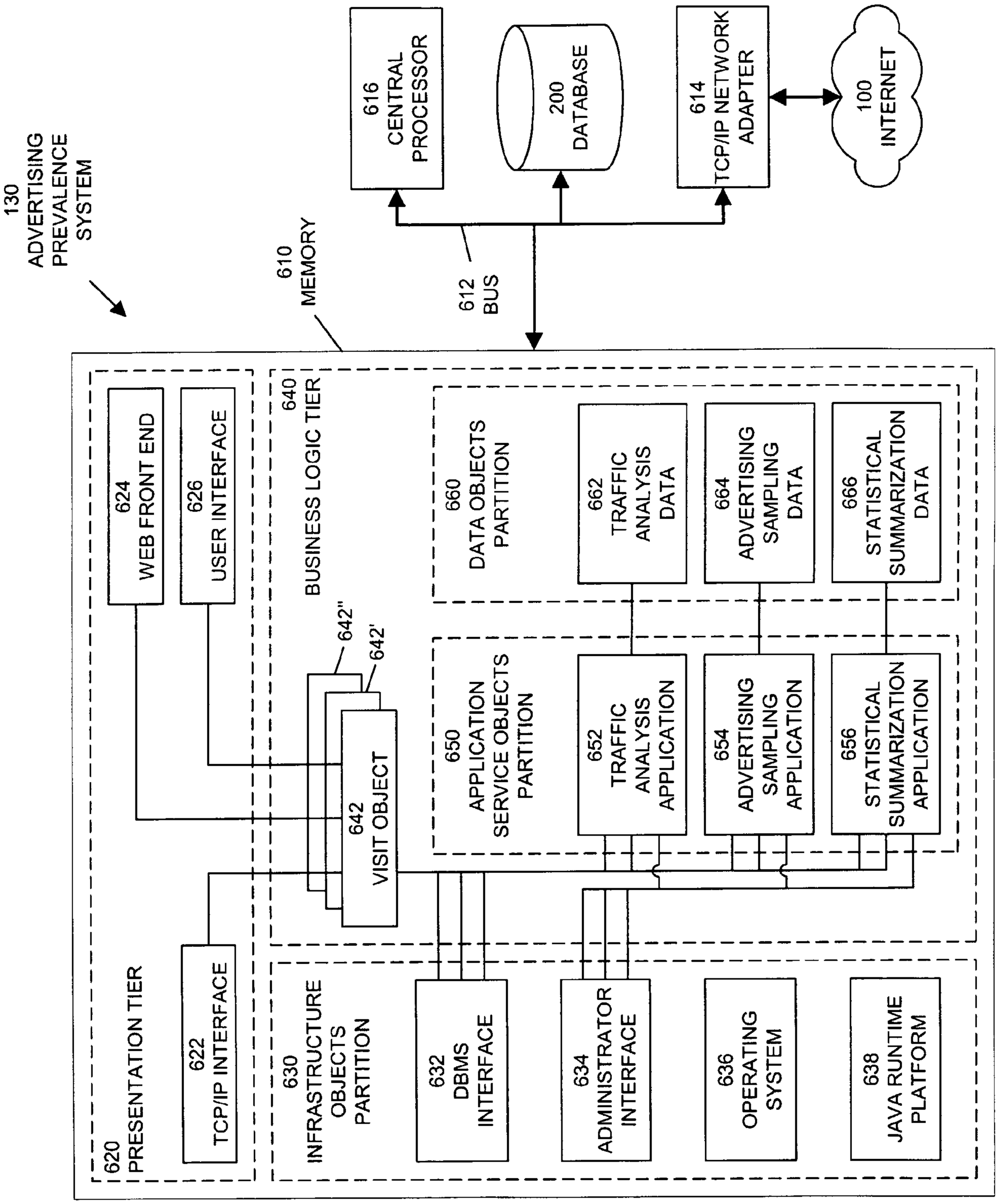

130
ADVERTISING
PREVALENCE
SYSTEM
610
MEMORY
612
BUS
616
CENTRAL
PROCESSOR
200
DATABASE
614
TCP/IP NETWORK
ADAPTER
100
INTERNET
620
PRESENTATION TIER
624
WEB FRONT END
626
USER INTERFACE
622
TCP/IP INTERFACE
640
BUSINESS LOGIC TIER
642"
642'
642
VISIT OBJECT
650
APPLICATION
SERVICE OBJECTS
PARTITION
652
TRAFFIC
ANALYSIS
APPLICATION
654
ADVERTISING
SAMPLING
APPLICATION
656
STATISTICAL
SUMMARIZATION
APPLICATION
660
DATA OBJECTS
PARTITION
662
TRAFFIC
ANALYSIS
DATA
664
ADVERTISING
SAMPLING
DATA
666
STATISTICAL
SUMMARIZATION
DATA
630
INFRASTRUCTURE
OBJECTS
PARTITION
632
DBMS
INTERFACE
634
ADMINISTRATOR
INTERFACE
636
OPERATING
SYSTEM
638
JAVA RUNTIME
PLATFORM

FIG. 6

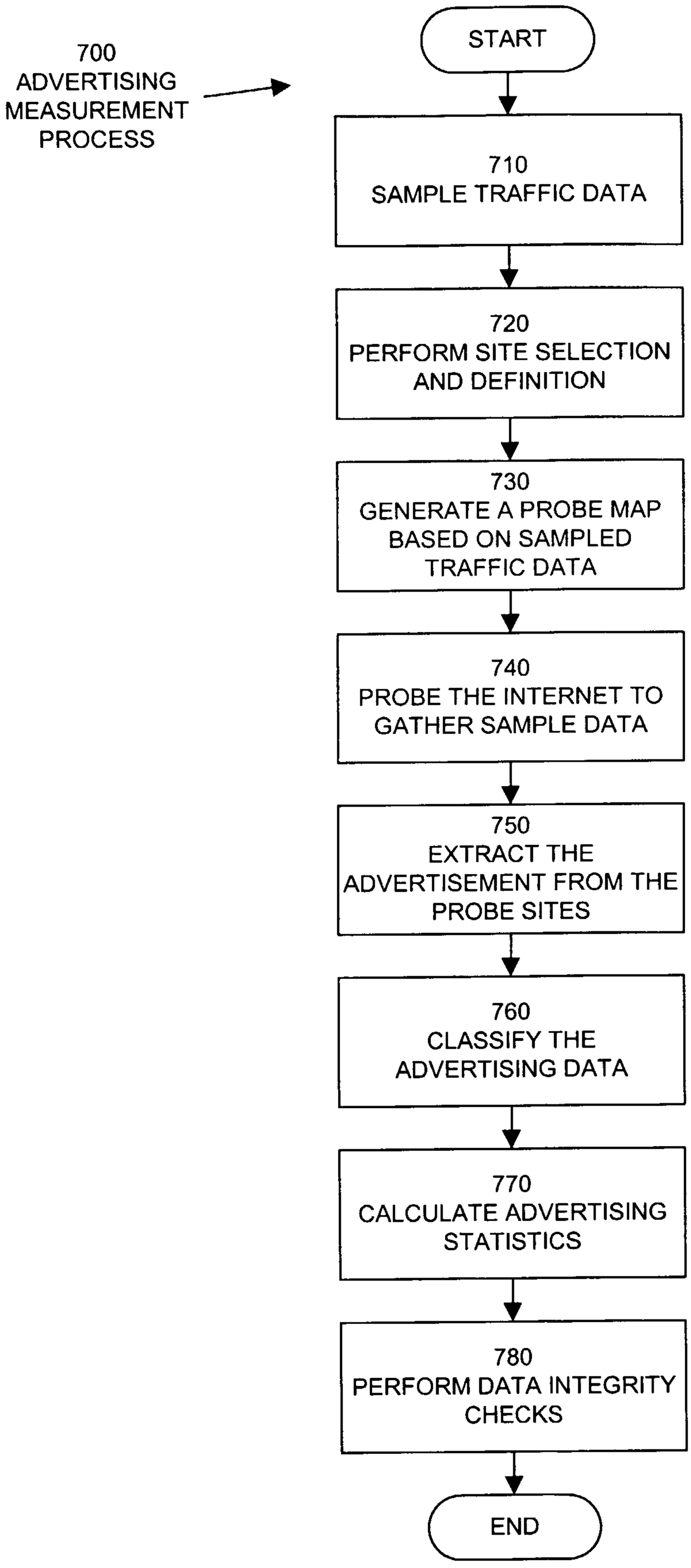

700
ADVERTISING
MEASUREMENT
PROCESS
START
710
SAMPLE TRAFFIC DATA
720
PERFORM SITE SELECTION
AND DEFINITION
730
GENERATE A PROBE MAP
BASED ON SAMPLED
TRAFFIC DATA
740
PROBE THE INTERNET TO
GATHER SAMPLE DATA
750
EXTRACT THE
ADVERTISEMENT FROM THE
PROBE SITES
760
CLASSIFY THE
ADVERTISING DATA
770
CALCULATE ADVERTISING
STATISTICS
780
PERFORM DATA INTEGRITY
CHECKS
END

FIG. 7A

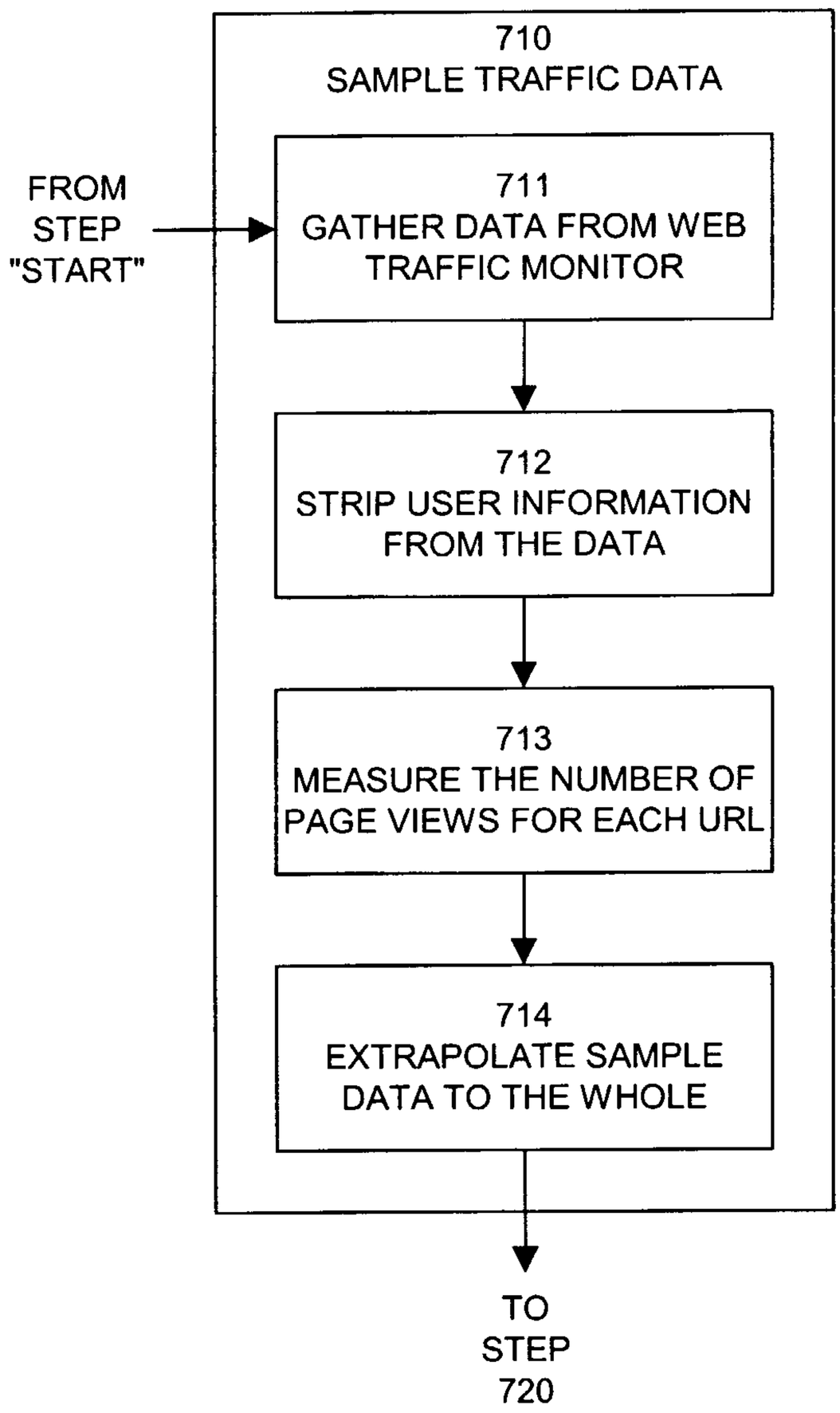

710
SAMPLE TRAFFIC DATA
FROM STEP "START"
711
GATHER DATA FROM WEB TRAFFIC MONITOR
712
STRIP USER INFORMATION FROM THE DATA
713
MEASURE THE NUMBER OF PAGE VIEWS FOR EACH URL
714
EXTRAPOLATE SAMPLE DATA TO THE WHOLE
TO STEP
720

FIG. 7B

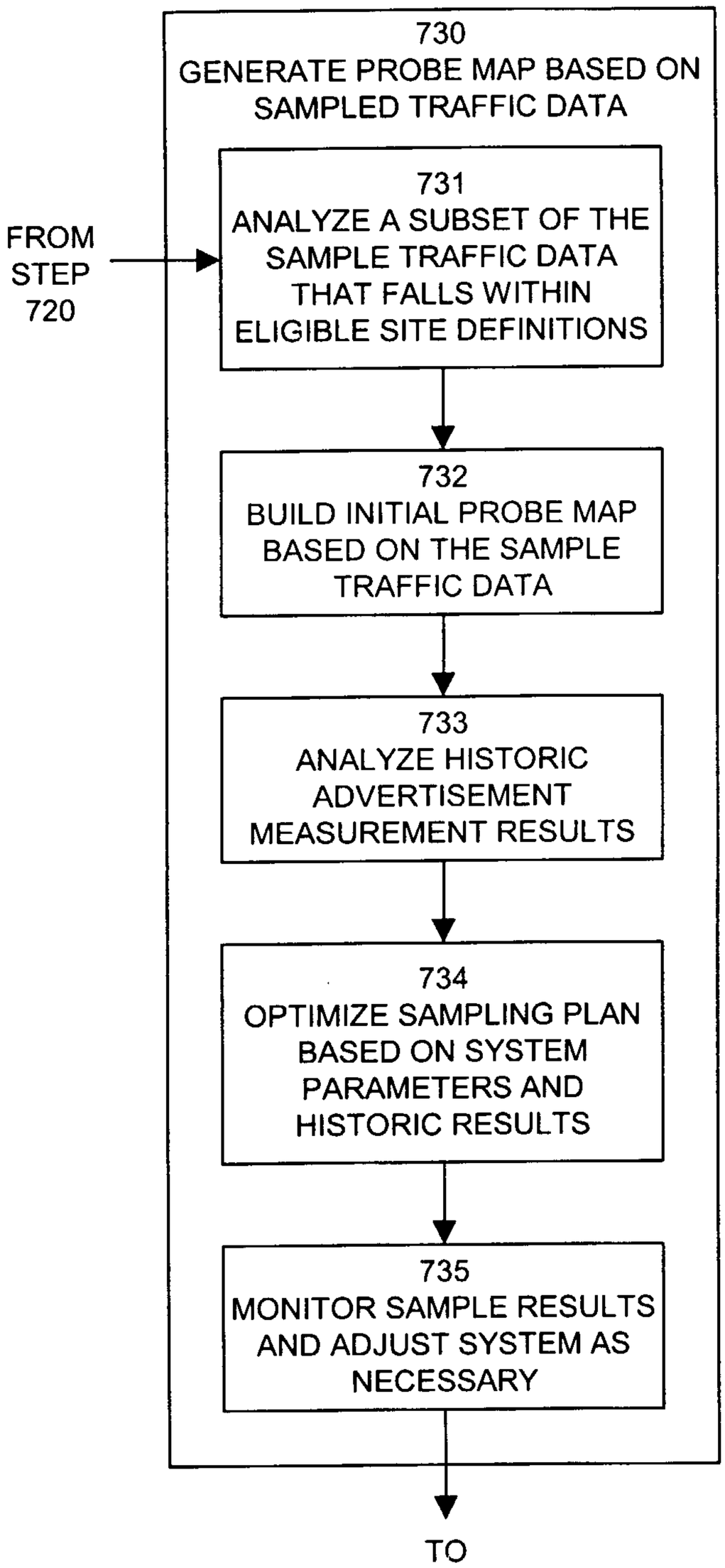

730
GENERATE PROBE MAP BASED ON SAMPLED TRAFFIC DATA
FROM STEP 720
731
ANALYZE A SUBSET OF THE SAMPLE TRAFFIC DATA THAT FALLS WITHIN ELIGIBLE SITE DEFINITIONS
732
BUILD INITIAL PROBE MAP BASED ON THE SAMPLE TRAFFIC DATA
733
ANALYZE HISTORIC ADVERTISEMENT MEASUREMENT RESULTS
734
OPTIMIZE SAMPLING PLAN BASED ON SYSTEM PARAMETERS AND HISTORIC RESULTS
735
MONITOR SAMPLE RESULTS AND ADJUST SYSTEM AS NECESSARY
TO STEP 740

FIG. 7C

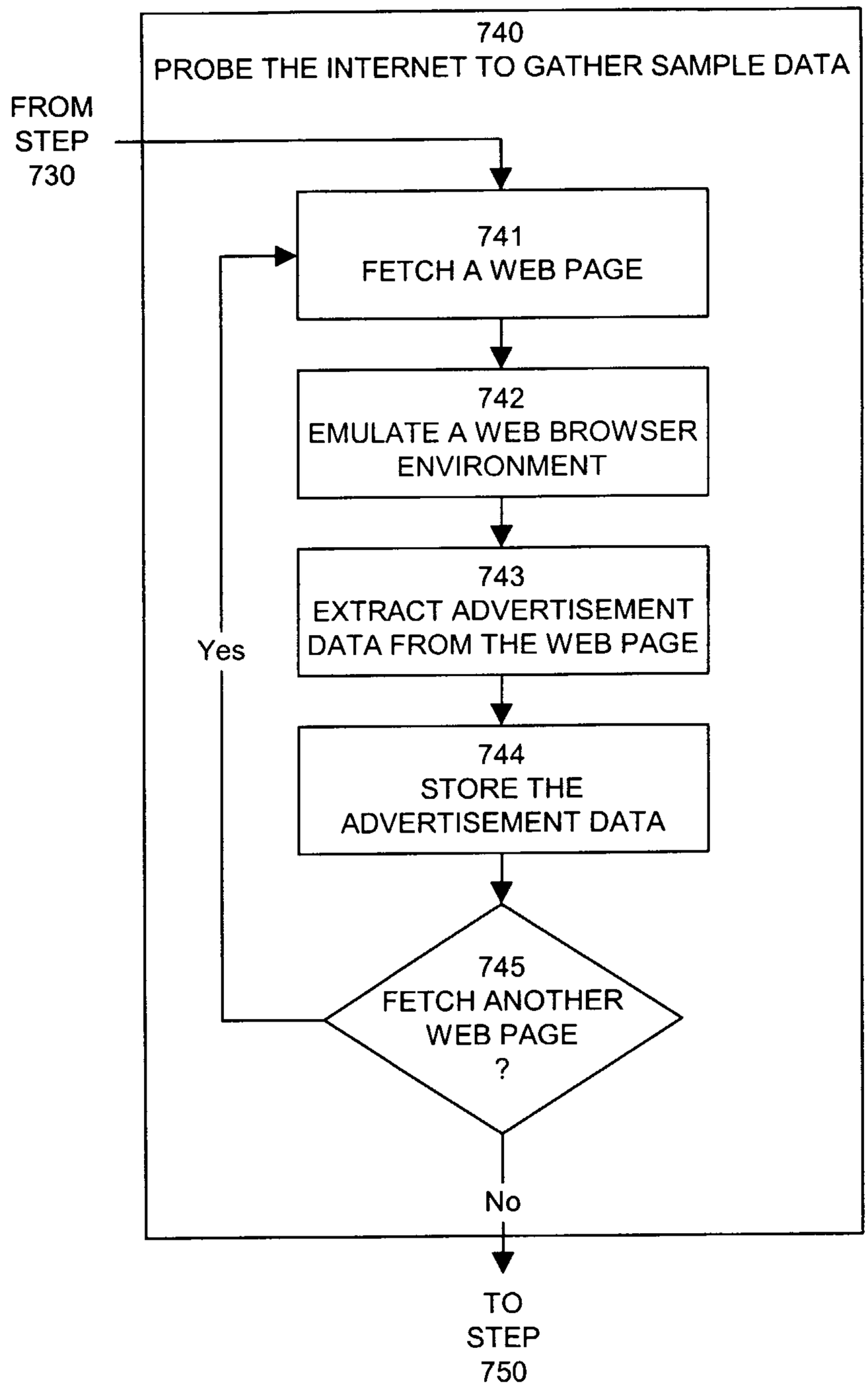

740
PROBE THE INTERNET TO GATHER SAMPLE DATA
FROM STEP 730
741
FETCH A WEB PAGE
742
EMULATE A WEB BROWSER ENVIRONMENT
743
EXTRACT ADVERTISEMENT DATA FROM THE WEB PAGE
744
STORE THE ADVERTISEMENT DATA
745
FETCH ANOTHER WEB PAGE ?
Yes
No
TO STEP 750

FIG. 7D

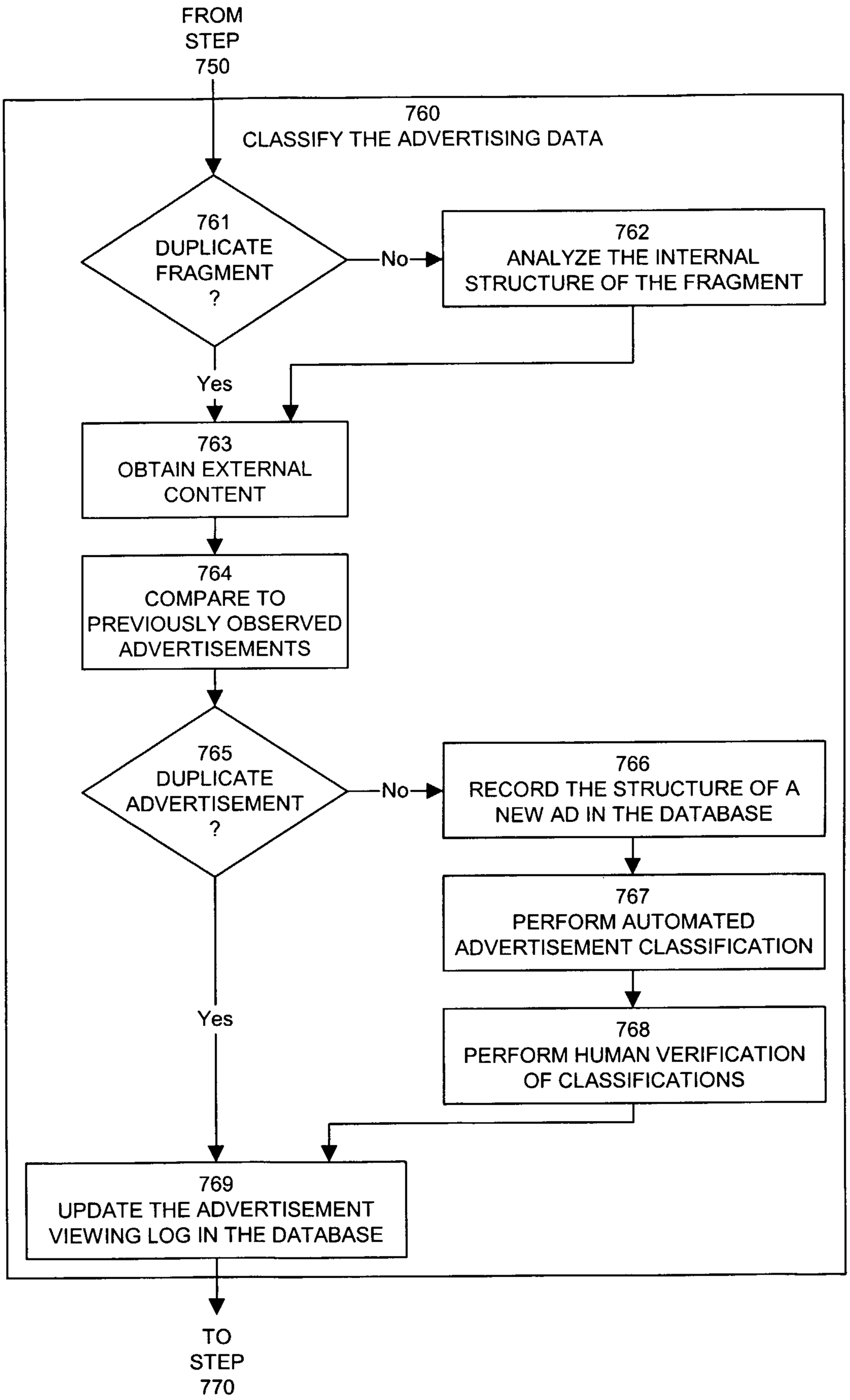

FROM STEP 750
760 CLASSIFY THE ADVERTISING DATA
761 DUPLICATE FRAGMENT ?
No
762 ANALYZE THE INTERNAL STRUCTURE OF THE FRAGMENT
Yes
763 OBTAIN EXTERNAL CONTENT
764 COMPARE TO PREVIOUSLY OBSERVED ADVERTISEMENTS
765 DUPLICATE ADVERTISEMENT ?
No
766 RECORD THE STRUCTURE OF A NEW AD IN THE DATABASE
767 PERFORM AUTOMATED ADVERTISEMENT CLASSIFICATION
768 PERFORM HUMAN VERIFICATION OF CLASSIFICATIONS
Yes
769 UPDATE THE ADVERTISEMENT VIEWING LOG IN THE DATABASE
TO STEP 770

FIG. 7E

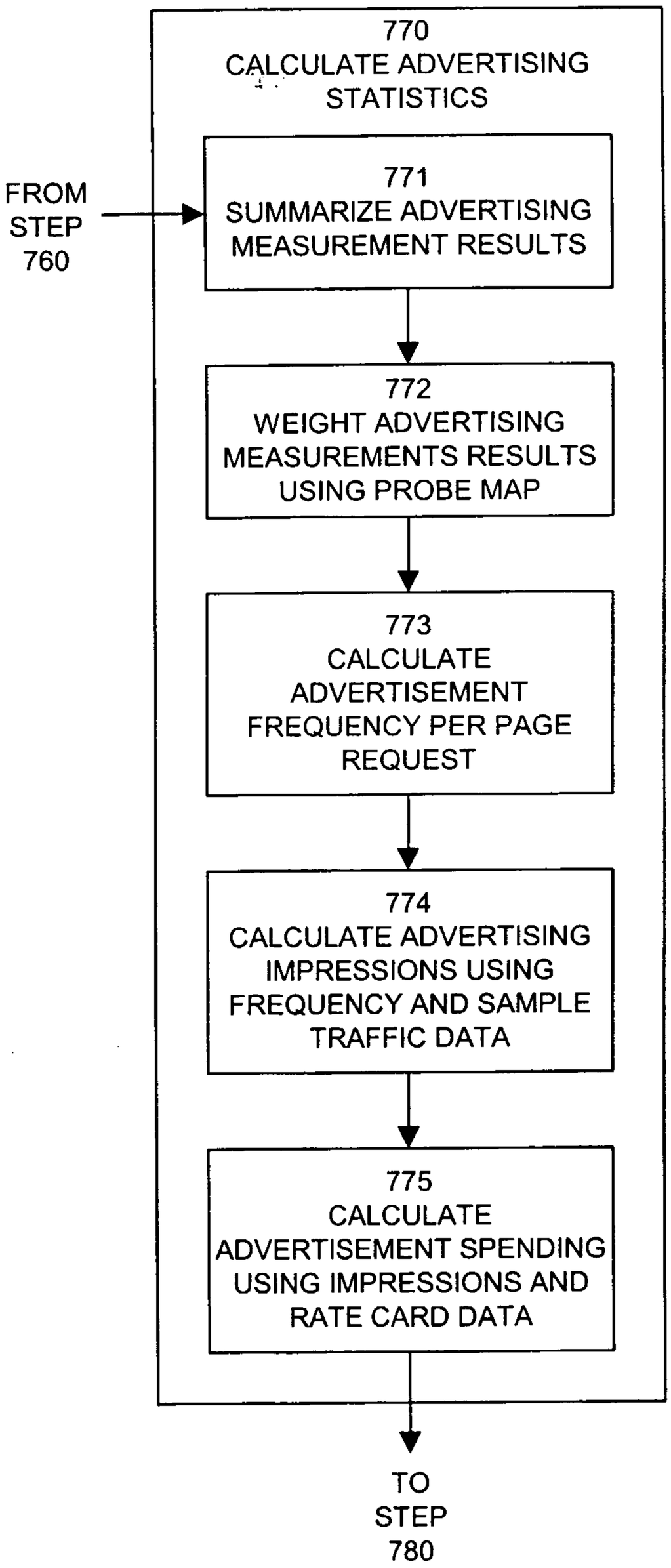
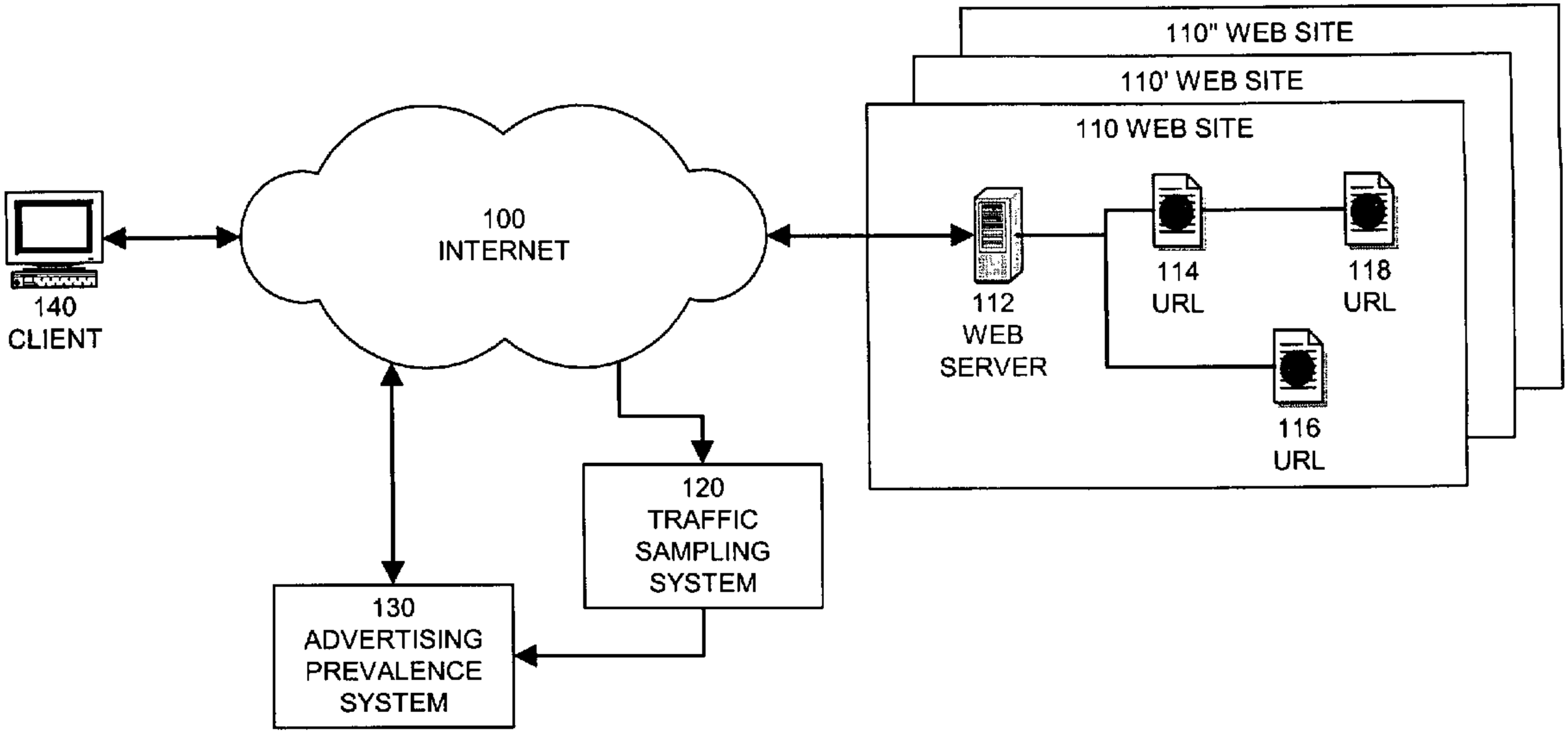

770
CALCULATE ADVERTISING STATISTICS
FROM STEP 760
771
SUMMARIZE ADVERTISING MEASUREMENT RESULTS
772
WEIGHT ADVERTISING MEASUREMENTS RESULTS USING PROBE MAP
773
CALCULATE ADVERTISEMENT FREQUENCY PER PAGE REQUEST
774
CALCULATE ADVERTISING IMPRESSIONS USING FREQUENCY AND SAMPLE TRAFFIC DATA
775
CALCULATE ADVERTISEMENT SPENDING USING IMPRESSIONS AND RATE CARD DATA
TO STEP 780

FIG. 7F


110" WEB SITE
110' WEB SITE
110 WEB SITE
100
INTERNET
140
CLIENT
112
WEB
SERVER
114
URL
118
URL
116
URL
120
TRAFFIC
SAMPLING
SYSTEM
130
ADVERTISING
PREVALENCE
SYSTEM