(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 371 404**

(51) Int. CI.:
H04L 1/00 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **00980233 .1**
(96) Fecha de presentación: **25.10.2000**
(97) Número de publicación de la solicitud: **1252735**
(97) Fecha de publicación de la solicitud: **30.10.2002**

(54) Título: **SISTEMA Y PROCEDIMIENTO PARA ESTIMAR LA PREVALENCIA DE CONTENIDO DIGITAL EN LA WORLD-WIDE-WEB.**

(30) Prioridad:
12.01.2000 US 175665 P
07.09.2000 US 231195 P

(45) Fecha de publicación de la mención BOPI:
02.01.2012

(45) Fecha de la publicación del folleto de la patente:
02.01.2012

(73) Titular/es:
The Nielsen Company (US), LLC
150 North Martingale Road
Schaumburg, IL 60173, US

(72) Inventor/es:
LAUCKHART, Gregory, J.;
HORMAN, Craig, B.;
KOROL, Christa, L. y
BARTOT, James, T.

(74) Agente: **Sugrañes Moline, Pedro**

DESCRIPCIÓN

Sistema y procedimiento para estimar la prevalencia de contenido digital en la World-Wide-Web

REFERENCIA A SOLICITUDES RELACIONADAS

La presente solicitud reivindica la prioridad respecto a, e incorpora por referencia, la solicitud provisional de patente de invención número 60/175.665, presentada en la Oficina de Patentes y Marcas de los Estados Unidos el 12 de enero de 2000, y la solicitud provisional de patente de invención número 60/231.195, presentada en la Oficina de Patentes y Marcas de los Estados Unidos el 7 de septiembre de 2000.

CAMPO DE LA INVENCIÓN

La presente invención se refiere en general a un sistema, procedimiento y producto de programa informático para rastrear y medir el contenido digital que se distribuye por una red informática como Internet. Más particularmente, la presente invención se refiere a un sistema, procedimiento y producto de programa informático que recopila datos de anuncios en línea, analiza los datos, y usa los datos para calcular mediciones de la prevalencia de esos anuncios.

ANTECEDENTES DE LA INVENCIÓN

El aumento de la popularidad de Internet y la World-Wide-Web ("Web") es debido, en parte, a las tecnologías interactivas que una página web puede emplear. Estas tecnologías interactivas afectan directamente a la Web como medio publicitario porque las tecnologías introdujeron nuevos formatos publicitarios como anuncios de patrocinio de iconos fijos, banners y botones giratorios, y anuncios intersticiales (es decir, anuncios en línea que interrumpen el trabajo del usuario y se adueñan de un porcentaje significativo de la visualización de la pantalla). Aun cuando la creación del anuncio es diferente, el efecto sobre el espectador es similar a la publicidad tradicional. Por ejemplo, un anuncio de banner o un icono de logotipo en una página web crea una impresión del producto para el espectador que es equivalente a un anuncio de valla publicitaria tradicional que promociona un producto presentando la marca o el eslogan. Igualmente, un logotipo del patrocinador en una página web crea una impresión del patrocinador para el espectador que es equivalente a ver un logotipo de patrocinador en el marcador en un partido de baloncesto universitario.

El crecimiento rápido e imprevisible de Internet durante los últimos varios años ha creado una gran demanda de estadísticas de calidad que cuantifiquen su magnitud y tasa de expansión. Varias metodologías de medición tradicionales producen estadísticas útiles acerca de Internet y sus usuarios, pero la complejidad de Internet ha dejado a algunas de estas metodologías incapaces de responder a muchas cuestiones importantes.

La solicitud internacional WO98/59309 desvela un sistema de medición de medios electrónicos cooperativos que usa manipuladores de medios para obtener información de, o si no obtener información acerca de, objetos de medios presentados, incluyendo etiquetas de identificación, de haberlas, para recopilación por agentes de recopilación de datos de investigación y expedición subsiguiente a un controlador de investigación de medios centralizado. El controlador de investigación de medios registra los anuncios y otros medios para una medición subsiguiente y proporciona una etiqueta de identificación única que puede añadirse a, o asociarse con, el objeto de medios existente. Los objetos de medios se presentan a un miembro de panel mediante un dispositivo informático de miembro de panel que puede recibir objetos de medios por medio de una conexión de red, procedentes de una o más fuentes locales, o generar objetos de medios en tiempo real, o una combinación de los mismos. Se asignan uno o más agentes de recopilación de datos de investigación para medir la exposición a, y las interacciones con, medios electrónicos de cada miembro de panel. Los manipuladores de medios cooperativos obtienen automáticamente información de, o si no obtienen información acerca de, objetos de medios presentados incluyendo etiquetas de identificación, de haberlas, y otra información, para recopilación por los agentes de recopilación de datos de investigación. Un agente de recopilación de datos de investigación rastreará un elemento de panel, y recopilará tales transmisiones desde los manipuladores de medios cooperativos, cuando un miembro de panel está en el ámbito del agente de recopilación de datos de investigación.

La patente de EE.UU. 5.995.943 desvela un sistema y procedimiento de agregación y síntesis de información que proporciona agregación y encapsulado de información estructurada y desestructurada procedente de fuentes dispares como las disponibles en una red como Internet. Un dispositivo de interfaz de red compatible/direccionable es manejado por un usuario. El dispositivo de interfaz de red se comunica con almacenes de datos locales o almacenes de datos accesibles por red a través de un esquema de direccionamiento como las direcciones del Localizador Uniforme de Recursos (URLs) utilizadas por Internet. Se accede a los datos que pasan entre el dispositivo de interfaz de red y los almacenes de datos y son escrutados y recuperados a través de un sistema de pasarela intermediaria. Tal información agregada luego se sintetiza, se adapta, se personaliza y se localiza para satisfacer las solicitudes de recursos de información especificadas por el usuario a través del dispositivo de interfaz de red.

La publicidad en línea es un área donde las metodologías tradicionales no se prestan bien a medición. Cada día, miles y miles de anuncios electrónicos aparecen y luego desaparecen de millones de páginas web. La naturaleza transitoria de la actividad publicitaria en línea justifica una metodología novedosa para medir con exactitud la actividad publicitaria.

Los sistemas de rastreo y medición publicitaria existentes automatizan la recopilación de páginas web, pero no automatizan la recopilación de los anuncios en línea. Desde que el contenido de un anuncio en línea cambia o rota a lo largo del tiempo, la reconstrucción precisa de la frecuencia de anuncios específicos requiere el muestreo continuo de páginas web relevantes en las proporciones correctas. Además, debido a la mera magnitud de la Web, los algoritmos de muestreo deben ser afinados con exactitud para optimizar la asignación de recursos (es decir, el ancho de banda de la red, el almacenamiento de bases de datos, el tiempo del procesador, etc.) y permitir simultáneamente la máxima cobertura de Internet. Los sistemas de rastreo y medición publicitaria existentes no satisfacen estas necesidades porque no están optimizados para asignación de recursos y no muestrean continuamente páginas web relevantes en la proporción correcta.

En vista de las deficiencias de los sistemas existentes descritas anteriormente, existe una necesidad de un sistema de rastreo y medición publicitaria que use los recursos más inteligentemente, sea más compatible con los sitios web que visite, sea escalable, y produzca mediciones precisas. La invención desvelada en este documento se ocupa de esta necesidad.

RESUMEN DE LA INVENCIÓN

La presente invención es un sistema, procedimiento, y producto de programa informático para rastrear y medir el contenido digital que se distribuye por una red informática como Internet tal como se define en las reivindicaciones adjuntas. El sistema recopila datos de anuncios en línea, analiza los datos, y usa los datos para calcular mediciones de la prevalencia de esos anuncios.

En la realización preferida, datos de tráfico procedentes de una diversidad de fuentes y metodologías complementarias alimentan el sistema de análisis de tráfico, un agente inteligente (es decir, software que interactúa con, aprende de, y se adapta a un entorno). El sistema de análisis de tráfico procesa los datos de tráfico sin procesar depurando y resumiendo los datos de tráfico antes de almacenar los datos procesados en una base de datos. Cuando el sistema de resumen estadístico calcula la frecuencia de anuncios, las impresiones, y los gastos, se basa en los datos procesados procedentes del sistema de análisis de tráfico.

El sistema de muestreo de anuncios, también conocido como el “sondeador” o “sondeador en la nube”, usa una metodología robusta que busca continuamente los sitios web más significativos e influyentes que hay que sondear (es decir, monitorizar). Por otra parte, la selección y definición de sitio realizada por la presente invención dicta las páginas web que comprenden cada sitio web para asegurar que se informa como tales de entidades completas de marca singular. El sistema de muestreo de anuncios usa tecnología de agente inteligente para recuperar páginas web a diversas frecuencias para obtener una muestra representativa. Esto permite al “sondeador en la nube” evaluar con exactitud con cuánta frecuencia aparece cada anuncio en los datos de tráfico. Después de que el “sondeador en la nube” busca una página web, el sistema de muestreo de anuncios extrae los anuncios de la página web. En la realización preferida, el extractor de anuncios, también conocido como el “extractor”, invoca un proceso de detección automática de anuncios (“AAD”), un proceso de extracción heurística, para extraer automáticamente todos los anuncios de la página web.

Después de la extracción de los anuncios de la página web, el sistema de muestreo de anuncios invoca un motor de clasificación para analizar los fragmentos de anuncios. El clasificador procesa cada fragmento para determinar una clasificación para el fragmento y luego almacena el fragmento y los datos de clasificación en una base de datos. El resultado de los análisis y el procesamiento realizados por el sistema de muestreo de anuncios es un catálogo abundante de actividad publicitaria que puede ser consultado fácilmente por un cliente.

La presente invención usa una parte cliente web y una interfaz de usuario para acceder a y actualizar los datos de la base de datos. La parte cliente web ofrece a un cliente, o usuario, de la presente invención una interfaz de consulta a la base de datos poblada por el análisis de tráfico, el muestreo de anuncios, y los sistemas de resumen estadístico. La interfaz de usuario es una interfaz gráfica de usuario que incluye un componente separado para gestión de cuentas del sistema, administración de sitios, administración de taxonomía, clasificación de contenido publicitario, y recopilación de listas de tarifas. La interfaz de usuario permite a un administrador y operador de cuentas mantener y administrar la presente invención. La interfaz de usuario también permite a un editor de medios revisar los datos de la base de datos para verificar la exactitud e integridad de la vasta cantidad de datos recopilados por la presente invención. Este procedimiento de integridad de datos investiga rutinariamente puntos de datos inusuales o periféricos para calibrar el sistema y adaptarlo a un entorno siempre cambiante.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Las figuras adjuntas ilustran mejor los detalles de la presente invención, tanto en cuanto a su estructura como a su funcionamiento. Los números de referencia y designaciones iguales en estas figuras se refieren a elementos iguales.

La Figura 1 es un diagrama de red que representa el entorno para un sistema de prevalencia publicitaria según la presente invención.

La Figura 2 representa el diagrama de red de la Figura 1, en mayor detalle, para mostrar las relaciones entre el entorno de red y los elementos que comprenden el sistema de prevalencia publicitaria.

La Figura 3 representa el diagrama de red de la Figura 2, en mayor detalle, para mostrar los elementos y subelementos que comprenden el sistema de prevalencia publicitaria y las conexiones al entorno de red.

La Figura 4A es un sitio web de ejemplo que ilustra los valores esperados usados en el cálculo de las estadísticas de prevalencia publicitaria.

La Figura 4B es un sitio web de ejemplo que ilustra los valores observados usados en el cálculo de las estadísticas de prevalencia publicitaria.

La Figura 4C es un sitio web de ejemplo que ilustra los valores ponderados usados en el cálculo de las estadísticas de prevalencia publicitaria.

La Figura 4D es un sitio web de ejemplo que ilustra un procedimiento alternativo para el cálculo de las estadísticas de prevalencia publicitaria.

La Figura 5 ilustra un ejemplo de una estructura de base de datos que puede usar el sistema de prevalencia publicitaria.

La Figura 6 es un diagrama de bloques funcionales del sistema de prevalencia publicitaria que muestra la configuración de los componentes de hardware y software.

La Figura 7A es un organigrama de un proceso en el sistema de prevalencia publicitaria que mide la calidad de la publicidad en línea y la actividad generada por un anuncio en línea.

La Figura 7B es un organigrama que describe, en mayor detalle, el proceso de muestrear datos de tráfico a partir de la Figura 7A.

La Figura 7C es un organigrama que describe, en mayor detalle, el proceso de generar un mapa de sondeo basado en los datos de tráfico muestreados a partir de la Figura 7A.

La Figura 7D es un organigrama que describe, en mayor detalle, el proceso de sondear Internet para reunir datos de muestras a partir de la Figura 7A.

La Figura 7E es un organigrama que describe, en mayor detalle, el proceso de clasificar los datos publicitarios a partir de la Figura 7A.

La Figura 7F es un organigrama que describe, en mayor detalle, el proceso de calcular las estadísticas publicitarias a partir de la Figura 7A.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

La Figura 1 representa el entorno para la realización preferida de la presente invención que incluye la Internet 100, y un sitio web 110, el sistema de muestreo de tráfico 120, el sistema de prevalencia publicitaria 130, y el cliente 140. La presente invención usa tecnología de agente inteligente para reunir datos relacionados con los atributos, la colocación, y la prevalencia de los anuncios en línea. Estos datos proporcionan a un usuario estimaciones actualizadas de estadísticas de anuncios y ayudan al usuario a obtener una ventaja competitiva.

Tal como se muestra en la Figura 1, la Internet 100 es una red de comunicación pública que permite al sistema de muestreo de tráfico 120 y al sistema de prevalencia publicitaria 130 comunicarse con un cliente 140 y un sitio web 110. Aun cuando la realización preferida usa la Internet 100, la presente invención contempla el uso de otras arquitecturas de red públicas o privadas como una intranet o extranet. Una intranet es una red de comunicación privada que funciona de manera similar a la Internet 100. Una organización, como una corporación, crea una intranet para proporcionar un medio seguro para que los miembros de la organización accedan a los recursos por la red de la organización. Una extranet también es una red de comunicación privada que funciona de manera similar a la Internet 100. A diferencia de una intranet, una extranet proporciona un medio seguro para que la organización autorice a los no miembros de la organización a acceder a ciertos recursos por la red de la organización. La presente invención también contempla el uso de un protocolo de red como Ethernet o Token Ring, así como protocolos de red patentados.

El sistema de muestreo de tráfico 120 es un programa que monitoriza y registra la actividad web en la Internet 100. El sistema de muestreo de tráfico 120 es un repositorio intermediario de datos de tráfico entre un internauta (no mostrado) en la Internet 100 y un servidor web 112. El servidor web 112 mostrado en la Figura 1 es un ordenador personal o estación de trabajo informática convencional que incluye el sistema operativo apropiado, hardware, protocolo de comunicaciones (por ejemplo, el Protocolo de Control de Transmisión/Protocolo Internet), y software de servidor web para hospedar una colección de páginas web. El internauta (no mostrado) se comunica con el servidor web 112 solicitando un localizador uniforme de recursos (“URL”) 114, 116, 118 asociado con el sitio web 110, típicamente usando un navegador web. Cualquier programa o dispositivo que pueda registrar una solicitud de un

URL hecha por un internauta (no mostrado) a un servidor web 112 puede realizar las funciones que la presente invención requiere del sistema de muestreo de tráfico 120. El sistema de muestreo de tráfico 120 agrega entonces los datos de tráfico para cada sitio web 110 para uso por el sistema de prevalencia publicitaria 130.

La presente invención puede usar cualquier sistema de muestreo de tráfico disponible comercialmente que proporcione una funcionalidad similar al producto de medición de audiencias Media Metrix. Otros posibles mecanismos para obtener una muestra de datos de tráfico incluyen:

1. El “Muestreo de proxy caché” reúne datos como datos de secuencias de clics del usuario, y solicitudes de páginas web procedentes de una jerarquía distribuida global de servidores proxy caché. Estos datos pasan a través de un mecanismo intermedio que proporciona servicios prebúsqueda y de caché para objetos web. A partir de mayo de 1999, las estadísticas de tráfico calculadas por la presente invención representan la síntesis de datos sin procesar procedentes de nueve cachés de primer nivel y aproximadamente 400 de segundo nivel en los Estados Unidos, así como unas 1100 adicionales por todo el mundo.
2. “Recopilación de paneles del lado del cliente” recupera datos de muestras procedentes de cada panelista a través de un mecanismo del lado del cliente y transfiere los datos a un repositorio de recopilación. El mecanismo del lado del cliente puede monitorizar la barra de direcciones del navegador, el navegador del usuario, un proxy del lado del cliente, o funciones gancho de pilas TCP/IP.
3. Un “transcodificador” es un proxy que reescribe HTML, normalmente con el fin de añadir elementos para generación de ingresos de anuncios o cabeceras/pies de página. Los proveedores de servicios de internet (“ISP”) gratuitos usan típicamente esta técnica.
4. Cualquier mecanismo de distribución de contenido que duplique la página web o el contenido del sitio de una manera pensada para disminuir la congestión de la red o mejorar la experiencia del usuario.
5. Cualquier mecanismo de filtrado de contenido que evalúe las solicitudes de URL y adopte acciones para permitir o denegar tales solicitudes.
6. De registros de servidores mantenidos por proveedores de servicios de Internet (“ISP”) o sitios web individuales.

La Figura 2 amplía el detalle del sistema de prevalencia publicitaria 130 de la Figura 1 para mostrar las relaciones entre el entorno de red y los elementos que comprenden el sistema de prevalencia publicitaria 130. El sistema de prevalencia publicitaria 130 incluye un sistema de análisis de tráfico 210, un sistema de muestreo de anuncios 220, y un sistema de resumen estadístico 230 que comunica datos a la base de datos 200 para su almacenamiento. El administrador de cuentas 260, el operador 262, y el editor de medios 264 pueden acceder a la base de datos 200 a través de la interfaz de usuario 240 para realizar funciones administrativas. El cliente 140 puede acceder a la base de datos 200 a través de la parte cliente web 250.

El sistema de análisis de tráfico 210 recibe datos de tráfico sin procesar procedentes del sistema de muestreo de tráfico 120. El sistema de análisis de tráfico 210 depura los datos de tráfico sin procesar eliminando la información de los datos de tráfico que puede identificar a un usuario particular en la Internet 100 y luego almacena los datos anónimos en la base de datos 200. El sistema de análisis de tráfico 210 estima al tráfico global hacia cada sitio web significativo en la Internet 100. La presente invención usa estos datos no sólo para calcular el número de impresiones publicitarias dada una estimación de la tasa de rotación en esa página, sino también en el sistema de creación de mapas de sondeo 320. En una realización, el sistema de análisis de tráfico 210 recibe datos de tráfico desde un sitio caché en la Internet 100. El objetivo es medir con exactitud el número de vistas de páginas por usuarios individuales, y por lo tanto el número de impresiones publicitarias.

El sistema de muestreo de anuncios 220 usa los datos de tráfico anónimos para determinar qué URLs incluir en la muestra recuperada del servidor web 112. El sistema de muestreo de anuncios 220 contacta con el servidor web 112 a través de la Internet 100 para recuperar un URL 114, 116, 118 y extraer los anuncios del mismo junto con las características acompañantes que describen los anuncios. La tasa de éxito para la recuperación de creativos es elevada. El análisis indica que la presente invención captura más del 95% de los creativos a los que se da servicio. El sistema de muestreo de anuncios 220 almacena estas características de anuncios en la base de datos 200. El sistema de muestreo de anuncios 220, por ejemplo, el Online Media Network Intelligent Agent Collection (“OMNIAC”), o el “sondeador en la nube”, sondea repetidamente sitios web destacados, extrae los anuncios de cada página web devuelta por la sonda, y clasifica los anuncios de cada página web por tipo, tecnología y anunciante.

El sistema de análisis de tráfico 210 y el sistema de muestreo de anuncios 220 también presentan los datos recuperados de la Internet 100 al sistema de resumen estadístico 230 para procesamiento periódico. El sistema de resumen estadístico 230 calcula la frecuencia de anuncios, las impresiones, y los gastos por sitio y por semana.

La interfaz gráfica de usuario para la presente invención incluye la interfaz de usuario 240 y la parte cliente web 250. El administrador de cuentas 260, el operador 262, y el editor de medios 264 acceden a la interfaz de usuario 240 para administrar el acceso por parte del cliente 140 la parte cliente web 250 (por ejemplo, gestión de cuentas y contraseñas), definen sitios e instrucciones de sondeo, y gestionan la taxonomía publicitaria, la clasificación de

contenidos, y la recopilación de listas de tarifas para el sistema de prevalencia publicitaria 130. La parte cliente web 250 es la interfaz del navegador web que un cliente 140 usa para recuperar los resultados de medición de anuncios de la base de datos 200 tal como son generados por el sistema de análisis de tráfico 210, el sistema de muestreo de anuncios 220, y el sistema de resumen estadístico 230.

La Figura 3 amplía más el detalle del sistema de prevalencia publicitaria 130 para representar los componentes lógicos que comprenden los elementos del sistema de prevalencia publicitaria 130 mostrado en la Figura 2. La Figura 3 también representa las relaciones entre el entorno de red y esos componentes lógicos.

El sistema de análisis de tráfico 210 incluye un proceso de sistema de anonimato 310 y resumen de tráfico 312.

El sistema de anonimato 310 depura los datos recibidos desde el sistema de muestreo de tráfico 120 eliminando la información que identifica a un usuario particular en la Internet. Se hace que los datos resulten anónimos pasando toda la información del usuario (por ejemplo, el número de protocolo de internet ("IP") de origen o las cookies) a través de una función hash unidireccional criptográficamente segura; esto asegura la mayor privacidad para los usuarios web sin devaluar los datos resultantes. El sistema de anonimato 310 presenta los datos depurados al sistema de resumen de tráfico 312 que a su vez almacena la información de recuento de URL agregada en la base de datos 200.

El proceso de resumen de tráfico 312 recibe los datos depurados procedentes del sistema de anonimato 310. Los datos de tráfico anónimo se resumen para producir totales de tráfico por semana o mes para URLs, dominios y sitios web individuales. El proceso de resumen de tráfico 312 escala los datos por factores de ponderación para extrapolar el tráfico global total a partir de la muestra.

El sistema de muestreo de anuncios 220 de la Figura 3 incluye un sistema de creación de mapas de sondeo 320, un sistema de recuperación de páginas web 322, un entorno de emulación de navegador web 324, un extractor de anuncios 326, y un clasificador estructural 328.

El sistema de creación de mapas de sondeo 320 genera un mapa de sondeo, es decir, los URLs 114, 116, 118 que visitará el sistema de muestreo de anuncios 220. Este mapa de sondeo ayuda al sistema de muestreo de anuncios 220 con la medición de la rotación de anuncios en sitios web individuales. La realización preferida de la presente invención busca continuamente diversas páginas web en el mapa de sondeo. En una realización alternativa, la presente invención visita cada URL del mapa de sondeo aproximadamente cada 6 minutos. Otra realización puede variar la frecuencia de búsqueda considerando varios factores incluyendo la cantidad de tráfico que visita el sitio web como un todo y la página web individual en cuestión, el número de anuncios vistos históricamente en la página web, y la similitud de la rotación de anuncios observada históricamente hacia otras páginas muestreadas.

El sistema de recuperación de páginas web 322 usa el mapa de sondeo generado por el sistema de creación de mapas de sondeo 320 para determinar qué páginas web tiene que muestrear y la frecuencia del muestreo. Para cada URL del mapa de sondeo generado por el sistema de creación de mapas de sondeo 320, el sistema de recuperación de páginas web 322 busca una página web, extrae cada anuncio de la página web, y almacena los atributos del anuncio en la base de datos 200. Los datos recuperados de cada URL del mapa de sondeo se usan para calcular la frecuencia con la que se muestra cada anuncio en un sitio web particular.

Para cada página web, el entorno de emulación de navegador web 324 simula la visualización de la página web en un navegador. Esta simulación garantiza que la presente invención detectará no sólo anuncios estáticos, sino también anuncios dinámicos generados por programas de software escritos en un lenguaje como JavaScript, Perl, Java, C, C++, o HTML que pueden estar incrustados en una página web.

El extractor de anuncios 326 extrae los anuncios en línea del resultado de la simulación realizada por el entorno de emulación de navegador web 324. El extractor de anuncios 326 identifica características del contenido publicitario (es decir, "fragmentos") extraídas de las páginas web devueltas por el sistema de creación de mapas de sondeo 320 que son de particular interés. Los anuncios son la característica dinámica más interesante de extraer, sin embargo, una realización alternativa de la presente invención puede usar la tecnología de extracción para recopilar cualquier tipo de contenido digital incluyendo promociones, encuestas, y artículos de noticias. El extractor de anuncios 326 puede usar diversos procedimientos de extracción de anuncios, incluyendo extracción basada en reglas, extracción heurística, y extracción por comparación.

La extracción basada en reglas se basa en un editor de medios 264 para usar la interfaz de usuario 240 para crear reglas. La interfaz de usuario 240 almacena las reglas en la base de datos 200 y el extractor de anuncios 326 aplica las reglas a cada página web que recupera el sistema de recuperación de páginas web 322. El efecto de ejecutar una regla es identificar y extraer un fragmento de HTML de la página web (es decir, la parte de la página que contiene el anuncio). El extractor de anuncios 326 primero convierte la representación HTML de la página web buscada en una representación XML perfectamente formada. Después de esta conversión, se aplican las reglas al

árbol de análisis sintáctico de la representación XML de la página web.

La extracción heurística se basa en la similitud de los anuncios a nivel de código fuente HTML o XML porque los anuncios son insertados típicamente por un servidor de anuncios cuando se genera la página web en respuesta a la solicitud del entorno de emulación de navegador web 324 de visualizar la página web. La extracción heurística analiza pistas del código fuente (por ejemplo, referencias a los nombres de servidores de anuncios conocidos) y extrae los fragmentos que rodean a esas pistas. La ventaja de este procedimiento es que la extracción es automática y el editor de medios no tiene que crear las reglas.

La extracción por comparación busca repetidamente la misma página web. Este procedimiento de extracción compara las diferentes versiones de la página web para determinar si el contenido varía de versión a versión. La porción de la página web que varía con cierto grado de frecuencia es normalmente un anuncio y es extraída.

El clasificador estructural 328 analiza sintácticamente cada anuncio y almacena los componentes estructurales en la base de datos 200 y pasa esos componentes al sistema de resumen estadístico 230. Cada fragmento de anuncio extraído por el extractor de anuncios 326 es analizado por el clasificador estructural 328. El proceso realizado por el clasificador estructural 328 comprende la eliminación de fragmentos duplicados, el análisis de fragmentos estructurales, y la detección de anuncios duplicados.

El clasificador estructural 328 realiza la eliminación de fragmentos duplicados comparado el fragmento de anuncio actual con otros fragmentos de la base de datos 200. Dos fragmentos de anuncio están duplicados si los fragmentos son idénticos (por ejemplo, cada fragmento tiene exactamente el mismo contenido HTML). Si el clasificador estructural 328 determina que el fragmento actual es un duplicado de un fragmento de la base de datos, el sistema de muestro de anuncios 220 registra otra observación del fragmento y sigue procesando fragmentos.

El clasificador estructural 328 realiza análisis de fragmentos estructurales sobre la representación XML de la página web determinando el “tipo físico” del fragmento (es decir, el código fuente HTML usado para construir el anuncio). Los tipos físicos que la presente invención reconoce incluyen banner, formulario, enlace individual, y contenido incrustado. Los fragmentos de anuncio de banner incluyen un enlace HTML individual que tiene una o dos imágenes adjuntas y ninguna etiqueta FORM o IFRAME. Loa fragmentos de anuncio de formulario incluyen un formulario HTML individual que no tiene etiqueta IFRAME. Los fragmentos de anuncio de enlace individual incluyen un enlace con etiqueta textual, pero no etiquetas IMG, FORM o IFRAME. Los fragmentos de anuncio de contenido incrustado hacen referencia a una entidad externa usando una etiqueta IFRAME. Después de realizar este análisis, el clasificador estructural 328 actualiza el fragmento de anuncio de la base de datos. Para un fragmento de anuncio de banner, el clasificador estructural 328 almacena los URLs de enlace e imagen en la base de datos 200. Un fragmento de anuncio de formulario requiere la creación de un URL simulando una presentación de usuario que pone cada control HTML en su valor por defecto. El clasificador estructural 328 almacena este URL y la “firma de formulario” (es decir, una cadena que describe unívocamente el contenido de todos los controles del formulario) en la base de datos 200. Para un fragmento de anuncio de texto individual, el clasificador estructural 328 almacena el URL para el enlace y todo el texto contenido dentro del enlace en la base de datos 200. Para fragmentos de anuncio de contenido incrustado, el clasificador estructural 328 almacena el URL asociado con la referencia externa en la base de datos 200. Este URL es cargado por el sistema, y se carga el documento al que se hace referencia. Una vez que el documento cargado ha sido analizado estructuralmente, el fragmento original hereda cualquier atributo que resulte del análisis del nuevo fragmento.

El clasificador estructural 328 realiza la detección de anuncios duplicados sobre cada fragmento de anuncio que tenga un tipo físico conocido porque estos fragmentos representan anuncios. Cada anuncio único tiene información, incluyendo qué definiciones de sitios están asociadas con el fragmento, almacenada en la base de datos 200. La determinación de unicidad del clasificador estructural 328 depende de diferentes criterios para cada tipo de fragmento. La primera etapa para cada tipo de definición es resolver todos los URLs asociados con el registro. Se cargan los URLs que hacen referencia a imágenes, y se anotan las imágenes duplicadas. Los URLs de enlace HTML, también conocidos como “URLs de clic”, se siguen cada vez que se crea un nuevo anuncio. Se anota el destino final para un URL de clic, después de seguir todas las redirecciones HTTP. Esto también se hace para URLs de presentación de enlace simulado asociados con definiciones de formularios. Una vez que se han resuelto todos los URLs, el clasificador estructural 328 determina si el anuncio es único. Los fragmentos de anuncio de banner se consideran únicos si tienen el mismo número de imágenes, si las imágenes son idénticas, y si el URL de destino es idéntico. Los fragmentos de anuncio de formulario se consideran únicos si tienen la misma firma, y el mismo URL de destino. Los fragmentos de anuncio de enlace individual se consideran únicos si tienen el mismo contenido textual y el mismo URL de destino.

El sistema de resumen estadístico 230 calcula las estadísticas de anuncios para cada anuncio único de la base de datos 200. La presente invención calcula, para cada sitio web, las impresiones publicitarias (es decir, el número de veces que un ser humano ve un anuncio). La presente invención calcula las impresiones publicitarias, I, usando la fórmula I = T x R, donde T es el tráfico que va al sitio, y R es la rotación de anuncios en ese sitio. La presente

invención también calcula los gastos S, usando la fórmula S = I x RC, donde I son las impresiones publicitarias para un sitio web, y RC es el código de tarifas para el sitio web. La mayoría de las compras publicitarias son acuerdos complicados con descuentos por volumen de compra, así que nuestros números no representan necesariamente el coste real de la compra total.

La parte cliente web 250 es una interfaz gráfica de usuario que proporciona a un cliente 140 una interfaz de consulta a la base de datos 200 poblada por el sistema de análisis de tráfico 210, el sistema de muestreo de anuncios 220, y el sistema de resumen estadístico 230. El cliente 140 puede usar la parte cliente web 250 para crear, almacenar, editar y descargar informes gráficos y tabulares para una o más categorías industriales dependiendo del nivel de servicio que el cliente 140 selecciona.

La interfaz de usuario 240 de la Figura 3 incluye un componente separado para gestión de cuentas del sistema 340, administración de sitios 342, administración de taxonomía 344, clasificación de contenido publicitario 346, y recopilación de listas de tarifas 348.

El administrador de cuentas 260 usa el módulo de gestión de cuentas del sistema 340 de la interfaz de usuario 240 para simplificar la administración de la parte cliente web 250. El administrador de cuentas 260 usa el módulo de gestión de cuentas del sistema 340 para crear y eliminar cuentas de usuario, gestionar contraseñas de cuentas de usuario, y comprobar la salud general de la parte cliente web 250.

El operador 262 usa el módulo de administración de sitios 342 de la interfaz de usuario 240 para simplificar la administración de las definiciones de sitios. Los analistas del Internet Advertising Bureau estiman que más del 90% de todos los dólares de publicidad web se gastan en los cincuenta primeros sitios web. La selección de sitios comienza escogiendo los 100 primeros anuncios considerando los datos procedentes de Media Metrix, Nielsen/Net Ratings, y los datos de tráfico proxy de la base de datos 200. Estas listas son actualizadas periódicamente para bajar de categoría los sitios web con bajos niveles de tráfico y promover nuevos sitios con elevados niveles de tráfico. La presente invención también incluye sitios web que proporcionan contenido significativo en industrias clave. Un sitio escogido para su inclusión en las definiciones de sitios debe tener la estructura del sitio analizado para eliminar las secciones que no sirven anuncios, proceden de países extranjeros, o son parte de un conjunto de marcos. Los sitios procedentes de un país extranjero, como yahoo.co.jp, venden publicidad en el país anfitrión, y por lo tanto no son aplicables a las mediciones calculadas por la presente invención. Los sitios web que usan un conjunto de marcos HTML son tratados con mucho cuidado para sólo aplicar frecuencias de rotación al tráfico procedente de las secciones del conjunto de marcos que contienen el anuncio. Estas exclusiones combinadas son clave para hacer estimaciones exactas de impresiones publicitarias. Las presente invención también etiqueta las secciones que no pueden medirse directamente, debido a requisitos de registro (por ejemplo, páginas de correo). Como los sitios web cambian la frecuencia, este análisis estructural se repite periódicamente. Finalmente, la fase de análisis marcará automáticamente los sitios alterados para permitir actualizaciones aún más oportunas.

El editor de medios 264 usa los módulos de administración de taxonomía 344, clasificación de contenido publicitario 346, y recopilación de listas de tarifas 348 de la interfaz de usuario 240. El módulo de administración de taxonomía 344 simplifica la creación y el mantenimiento de los atributos asignados a anuncios durante la clasificación de contenido incluyendo la industria, la compañía y los productos del anuncio. La taxonomía nombra cada atributo y especifica su tipo, ascendencia y segmento de pertenencia. Por ejemplo, una compañía Honda, podría estar emparentada con la industria de la Automoción y pertenecer al segmento industrial Fabricantes de automóviles. El componente de clasificación de contenido publicitario 346 ayuda al editor de medios 264 a realizar la clasificación de contenido.

El clasificador estructural 328 realiza la asignación automatizada de productos publicitarios para determinar qué anuncio se está publicando. Este proceso incluye asignar “productos publicitarios” (es decir, atributos que describen cada “cosa” de la que el anuncio está haciendo publicidad) a cada fragmento de anuncio. En otra realización de la presente invención, el sistema de muestreo de anuncios 220 usa un conjunto ampliable de heurísticas para asignar productos publicitarios a cada anuncio. En la realización preferida, sin embargo, el único procedimiento automático empleado es la clasificación de dirección. La clasificación de dirección se basa en el URL de destino para asignar un conjunto de productos publicitarios a un anuncio. Un editor de medios 264 usa la interfaz de usuario 240 para mantener el conjunto de direcciones clasificadas. Por ejemplo, la primera vez que un editor de medios observa un anuncio en el que el URL de cliqueo es www.honda.com, puede introducir este URL como perteneciente al anunciante “Honda Motors”. Cualquier anuncio subsiguiente que incluya el mismo URL de cliqueo también será reconocido como un anuncio de Honda. Una dirección clasificada comprende un anfitrión, un prefijo de ruta URL, y un conjunto de productos publicitarios. La clasificación de dirección asigna un producto publicitario de dirección clasificada a un anuncio si el anfitrión en el URL de destino coincide con el anfitrión de la dirección clasificada y el prefijo de ruta en la dirección clasificada coincide con el comienzo de la ruta en el URL de destino.

El clasificador estructural 328 realiza la asignación y verificación humana de productos publicitarios como comprobación de calidad de los datos de productos publicitarios. Esta fase es la que requiere más trabajo humano.

Un editor de medios 264 usa un módulo de interfaz gráfica de usuario en la interfaz de usuario 240 para visualizar cada anuncio, verificar las asignaciones automáticas de productos publicitarios, y asignar cualquier otro producto publicitario que parezca apropiado después de la inspección del anuncio y el destino del anuncio. La base de datos de clasificación de dirección también es mantenida típicamente en este momento.

El editor de medios 264 usa el módulo de recopilación de listas de tarifas 348 para introducir la información de contacto y lista de tarifas para un sitio web identificado por el sistema de análisis de tráfico 210, así como los anunciantes designados. La introducción de la lista de tarifas incluye el trimestre aplicable (por ejemplo, Q4 2000), las dimensiones del anuncio en píxeles, la estructura de tarifas (por ejemplo, CPM, tarifa plana, o por clic), el baremo de costes para compras de diversas cantidades y duración. El editor de medios 264 también registra la dirección URL del kit de medios en línea y si están publicadas tarifas en el mismo. La información de contacto para un sitio web o anunciante incluye la página de inicio, el nombre, los números de teléfono y fax, la dirección de correo electrónico, y la dirección física.

Las Figuras 4A a 4C ilustran el procedimiento preferido para calcular las estadísticas de prevalencia publicitaria. El cálculo de las estadísticas de prevalencia publicitaria es un proceso iterativo que usa valores esperados deducidos por el sistema de análisis de tráfico 210 y valores observados deducidos por el sistema de prevalencia publicitaria 220 para calcular los valores ponderados y las estadísticas de prevalencia publicitaria. Cada una de las Figuras 4A a 4C representa una red en la Internet 100 que incluye dos sitios web servidos por el servidor web P 410 y el servidor web Q 420. La Figura 4A ilustra valores de tráfico esperados de ejemplo para la red. La Figura 4B ilustra valores de tráfico observados de ejemplo para la red. La Figura 4C ilustra valores de tráfico ponderados de ejemplo para la red.

La primera etapa en el proceso es normalizar los resultados procedentes del sistema de análisis de tráfico 210. El sistema de análisis de tráfico 210 proporciona el tráfico recibido por cada página web en la muestra de datos de tráfico. La Figura 4A representa el tráfico de ejemplo recibido en cada página web 411-416, 421-424 en la Internet 100 con la etiqueta “Tráfico =”. El mapa de sondeo generado por el sistema de creación de mapas de sondeo 320 incluye una entrada para cada página web 411-416, 421-424. El mapa de sondeo también incluye un “área” que cada página web 411-416, 421-424 consume en el mapa de sondeo con la etiqueta “Área =”. Los resultados normalizados son calculadas dividiendo el área que una página web consume en el mapa de sondeo por la suma del área para cada página web en la muestra de tráfico. En la Figura 4A, el valor normalizado, o probabilidad, para la página web P1 411 es el área para la página web P1 (es decir, 15) dividida por la suma del área para la página web P1, P2, P3, P4, P5, P6, Q1, Q2, Q3 y Q4 (es decir, 120). El valor normalizado es, por lo tanto, 0,125, o el 12,5%. Además del valor normalizado, el sistema también determina la escala dividiendo el tráfico para una página web por el área para la página web. En la Figura 4A, la escala para la página web P1 411 es el tráfico para la página web P1 (es decir, 150) dividido por el área para la página web P1 (es decir, 15), por lo tanto, la escala para la página web P1 es 10. La Tabla 1 resume los valores de escala y probabilidad para la página web restante en la Figura 4A.

Tabla 1

Página web	Área	Escala	Probabilidad
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%

La Figura 4B representa las búsquedas de páginas web de ejemplo en cada página web 411-416, 421-424 en la Internet 100 con la etiqueta “Búsquedas =”. La Figura 4B también representa el número de vistas de ejemplo de cada anuncio en una página web 411-416, 421-424 con una etiqueta como “Vistas de A1 =” para indicar el número de vistas del anuncio A1, “Vistas de A2 =” para indicar el número de vistas del anuncio A2, etc.

La Figura 4C representa las búsquedas ponderadas de páginas de web de ejemplo en cada página web 411-416, 421-424 en la Internet 100 con la etiqueta “Búsquedas =”. La Figura 4C también representa el número de vistas de ejemplo de cada anuncio en una página web 411-416, 421-424 con una etiqueta como “Vistas de A1 =” para indicar el número de vistas del anuncio A1, “Vistas de A2 =” para indicar el número de vistas del anuncio A2, etc. La siguiente etapa en el proceso de cálculo es calcular las búsquedas a escala para cada sitio web 410, 420 sumando el producto de las búsquedas observadas de la Figura 4B y la escala de la Figura 4A, para cada página web 411-416, 421-424 en el sitio web. A continuación, el cálculo calcula el tráfico para cada sitio web 410, 420 sumando el tráfico de la Figura 4A para cada página web 411-416, 421-424 en el sitio web. La lista de tarifas, o CPM, es un valor

asignado por el editor de medios 264 para cada sitio web 410, 420. La Tabla 2 resume las búsquedas a escala, el tráfico, y el CPM para las Figuras 4A a 4C.

Tabla 2

Sitio	Búsquedas a escala	Tráfico	CPM
P	193,5	185	$35,00
Q	43	51	$50,00

Lo siguiente en el proceso de cálculo es calcular las Observaciones a Escala para cada anuncio en cada sitio web 410, 420 sumando el producto de las vistas del anuncio de la Figura 4B y la escala de la Figura 4A, para cada página web 411-416, 421-424 en el sitio web 410, 420. La etapa final en el cálculo es calcular las estadísticas de prevalencia publicitaria (es decir, frecuencia, impresiones, y gastos) para cada anuncio en cada sitio web 410, 420. La frecuencia se calcula dividiendo las observaciones a escala por las búsquedas a escala para cada anuncio en cada sitio web 410, 420. Las impresiones se calculan multiplicando la frecuencia por el tráfico de la Tabla 2 anterior para cada anuncio en cada sitio web 410, 420. Los gastos se calculan multiplicando las impresiones por el CPM de la Tabla 2 anterior para cada anuncio en el sitio web 410, 420. La Tabla 3 resume las Observaciones a Escala, la Frecuencia, las Impresiones, y los Gastos para el sitio web Q 420 usando los datos de las Figuras 4A a 4C.

Tabla 3

	Observaciones a escala	Frecuencia	Impresiones	Gastos
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				

Tabla 4

	Observaciones a escala	Frecuencia	Impresiones	Gastos
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

La Figura 4D ilustra una realización alternativa para calcular las estadísticas de prevalencia publicitaria. En esta realización, el sondeador se afina para optimizar la exactitud de medición de rotación. Las estimaciones estadísticas de exactitud en el campo son difíciles de realizar, debido a la naturaleza no estacionaria de los servidores publicitarios. Cuando se sondea cada 6 minutos, tiene una resolución del 0,06% en tasa de rotación a lo largo de un periodo de medición de una semana.

También en la realización alternativa de la Figura 4D, las sondas se distribuyen entre los sitios para medir con exactitud la rotación de anuncios en cada sitio. El número de URLs de sondeo asignadas a un sitio se determina a partir de tres variables. La primera es una constante a través de todos los sitios; se requiere un cierto número de URLs de sondeo para medir con exactitud la rotación incluso en el sitio más pequeño. La mitad de las sondas se asignan con esta variable. La segunda variable, ponderada al 40%, es la cantidad de tráfico que va a un sitio, ya que cada URL de sondeo representa una proporción del tráfico total de Internet. Los veinte sitios más grandes reciben más del 75% de estas sondas. Por último, se tiene en cuenta la complejidad del sitio, tal como se mide por el número total de URLs únicos encontrados en nuestros datos de tráfico proxy, con los sitios más complicados recibiendo URLs de sondeo extra. Esto representa el 10% restante de la distribución de sondas. Los URLs de sondeo pueden escogerse usando un algoritmo triturador de sitios para romper el sitio en zonas (es decir, conjuntos de páginas cuyas características de rotación de anuncios es probable que sean similares) para sondeo. La distribución de zonas está diseñada matemáticamente para maximizar la cobertura del sitio y, por lo tanto, la exactitud de rotación de anuncios. Se escoge un único URL para representar la rotación publicitaria de cada zona. Este URL se escoge como la página con tráfico más denso que contiene anuncios en esa zona. El algoritmo evita páginas de fechas específicas o páginas que hacen referencia a un evento limitado en el tiempo como el eclipse lunar total de agosto de 1999.

La realización alternativa de la Figura 4D calcula las impresiones de anuncios combinando las estimaciones de rotación y tráfico para cada sitio web 430. Para hacer esto, el sistema descompone el sitio en sus raíces constituyentes usando el algoritmo triturador de sitios. Se calcula la rotación de anuncios en cada espacio de anuncio y se aplica para estimar impresiones publicitarias en su raíz asociada. La rotación de anuncios en raíces sin sondas se estima a partir de una media, ponderada por el tráfico, de la rotación de anuncios de sondas en un nivel similar.

Por ejemplo, en la Figura 4D, el árbol del sitio de muestra tiene cinco URLs de sondeo 431-435, P1-5, colocados en cinco ramas principales de una página principal y 14 ramas secundarias. El número de cada página es el tráfico de muestra que va a esa página. La sonda P1 en la página de inicio, "www.testsite.com", mide la rotación, R, que ha de aplicarse al tráfico que va a esa página principal, con tráfico de 88 vistas de página. La rama A tiene una única sonda, P2, colocada en la página de nivel superior de esa rama con un URL de sondeo "www.testsite.com/A/". La rotación de este único URL de sondeo se estima como RA y se aplica al tráfico para esa raíz entera, un total de 21 vistas de página. La rama C tiene una sonda, P3, en una página de rama secundaria con tráfico denso, con un URL de sondeo "www.testsite.com/C/third.html". La rotación, RC, de esta página se aplica a todas las páginas de ramas secundarias en esa raíz y también un nivel arriba en el árbol, a lo largo de un total de 25 vistas de página. La rama E recibe una gran parte del tráfico para el sitio, un total de 61 vistas de página, y por lo tanto se le asignan dos sondas, P4 y P5. Éstas están en dos páginas de rama secundaria, "www.testsite.com/E/first.html" y "www.testsite.com/E/third.html". A la rotación de cada una se aplica el tráfico a esas páginas individuales. Para las restantes 18 vistas de página en esa rama (diez vistas de página de dos páginas secundarias y ocho de la página de nivel superior de esa rama) se calcula una rotación ponderada, RE = ((13 x RE1) + (30 x RE3))/(13+30). El análisis de la rotación de raíces tiene como resultado impresiones publicitarias para más del 96% del sitio. Las impresiones para las dos ramas finales, B y D, se calculan con una rotación media a partir de ramas adyacentes, ponderada por tráfico,

RB = RD = ((21 x RA) + (25 x RC) + (61 x RE)) ÷ (21 + 25 + 61).

Este análisis tiene como resultado las impresiones totales a través del sitio para cada anuncio único. El cálculo final realizado por la realización alternativa de la Figura 4D son los gastos, el producto de las impresiones y la lista de tarifas.

La Figura 5 ilustra una estructura de base de datos que puede usar el sistema de prevalencia publicitaria 130 para almacenar información recuperada por el sistema de muestreo de tráfico 120 y el sistema de recuperación de páginas web 322. La realización preferida segmenta la base de datos 200 en particiones. Cada partición puede realizar funciones similares a una base de datos independiente como la base de datos 200. Además, una base de datos particionada simplifica la administración de los datos de la partición. Aun cuando la realización preferida usa particiones de bases de datos, la presente invención contempla la consolidación de estas particiones en una sola base de datos, así como hacer cada partición una base de datos independiente y distribuir cada base de datos a una estación de trabajo informática o servidor de propósito general separados. Las particiones para la base de datos 200 de la presente invención incluyen registros de muestreo 510, definiciones de sondeo 520, datos de soporte publicitario 530, y resumen publicitario 540. La realización preferida de la presente invención usa un sistema de gestión de bases de datos relacionales, como el producto Oracle8i de Oracle Corporation, para crear y gestionar la base de datos y las particiones. Aun cuando la realización preferida usa una base de datos relacional, la presente invención contempla el uso de otras arquitecturas de bases de datos como un sistema de gestión de bases de datos orientadas a objetos.

La partición de registros de muestreo 510 de la base de datos 200 comprende tablas de bases de datos que están segmentadas lógicamente en áreas de datos de tráfico 512, registro de vistas de anuncios 514, y estructura de anuncio 516.

El área de datos de tráfico 512 contiene datos procesados por el sistema de muestreo de tráfico 120, el sistema de anonimato 310, y el sistema de resumen estadístico 230. Los datos almacenados en este esquema incluyen un URL "alterado", y el recuento de tráfico que cada URL recibe por fuente de tráfico a lo largo de un periodo de tiempo. Un URL "alterado" es un URL ordinario con el campo de protocolo eliminado y el orden de los componentes con puntos en el nombre de equipo invertidos. Por ejemplo, la presente invención transforma un URL ordinario, como http://www.somesite.com/food, en un URL alterado eliminando el campo de protocolo (es decir, "http:/") e invirtiendo el orden los componentes con puntos en el nombre de equipo (es decir, www.somesite.com). El URL alterado resultante en este ejemplo es "com.somesite.www/food". La presente invención usa este formato URL patentado para mejorar en gran medida el proceso de análisis de datos de tráfico. El sistema de muestreo de tráfico 120 puebla el área de datos de tráfico 512 de la base de datos 200. El sistema de creación de mapas de sondeo 320 accede a los datos del área de datos de tráfico 512 para ayudar al sistema de recuperación de páginas web 322 y al sistema de resumen estadístico 230 con el cálculo de la impresión publicitaria y las estadísticas de gastos.

El área de registro de vistas de anuncios 514 registra el tiempo, el URL y el identificador de anuncio para cada anuncio encontrado en la Internet 100. Esta área también registra cada vez que el sistema no detecta un anuncio en una página web que previamente incluía el anuncio. Además, el sistema registra cada vez que el sistema detecta un anuncio potencial, pero no reconoce el anuncio durante la clasificación estructural. El clasificador estructural 328 y el sistema de recuperación de páginas web 322 del sistema de muestreo de anuncios 222 puebla el área de registro de vistas de anuncios 514 de la base de datos 200. El sistema de resumen estadístico 230 accede a los datos del área de registro de vistas de anuncios 514 para determinar la frecuencia con la que cada anuncio aparece en cada sitio.

El área de estructura de anuncio 516 contiene datos que caracterizan a cada anuncio único localizado por el sistema. Estos datos incluyen el contenido del anuncio, el tipo de anuncio (por ejemplo, imagen, formulario HTML, Flash, etc.), el URL de destino vinculado al anuncio, y varios elementos usados durante la clasificación de contenido y la diagnosis, incluyendo dónde se vio por primera vez el anuncio, y qué definición de anuncio produjo originalmente el anuncio. El componente clasificador estructural 328 del sistema de muestreo de anuncios 220 puebla el área de estructura de anuncio 516 de la base de datos 200. La interfaz de usuario 240 accede a los datos del área de estructura de anuncio 516 para presentar cada anuncio al editor de medios 264 durante la edición de clasificación. La parte cliente web 250 también accede a los datos del área de estructura de anuncio 516 para presentar los anuncios al cliente 140.

La partición de definiciones de sondeo 520 de la base de datos 200 comprende tablas de bases de datos que están segmentadas lógicamente en áreas de definición de sitio 522, mapa de sondeo 524, y definición de reglas de extracción de anuncios 526.

El área de definición de sitio 522 divide en zonas la parte de la Internet 100 que el sistema sondea. La definición de zona primaria es un “sitio”, una entidad coherente que el sistema tiene que analizar, muestrear y resumir. El sistema define cada sitio en términos de prefijos URL alterados tanto inclusivos como exclusivos. Un “prefijo URL alterado” es un URL alterado que representa la zona de todos los URLs alterados para los que es un prefijo. Un “prefijo URL alterado inclusivo” especifica que un URL es parte de alguna entidad. Un “prefijo URL alterado exclusivo” especifica que un URL no es parte de alguna entidad, anulando las partes de la entidad incluidas por un prefijo inclusivo. Como ilustración, lo que viene a continuación es la lista de URLs alterados que puede resultar del procesamiento de un conjunto de URLs en una muestra de tráfico.

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

Si la definición de sitio para “somesite” incluye el prefijo URL inclusivo “com.somesite/” y el prefijo URL exclusivo “com.somesite/foo/bar”, la aplicación de esta definición de sitio a los URLs de muestra enumerados anteriormente produce un sistema que incluye los URL 1, 2 y 4. El URL 3 no es parte de la definición de sitio debido a la exclusión explícita de “com.somesite/foo/bar”. El URL 5 no es parte de la definición de sitio porque nunca estuvo incluido en el prefijo URL inclusivo “com.somesite/”. La interfaz de usuario 240 puebla el área de definición de sitio 522 de la base de datos 200. El sistema de creación de mapas de sondeo 320 accede a los datos del área de definición de sitio 522 para determinar qué URLs sondear. El sistema de resumen estadístico 230 accede a los datos del área de definición de sitio 522 para determinar los niveles de tráfico hacia los sitios sumando el tráfico hacia los URLs incluidos en un sitio.

El área de mapa de sondeo 524 contiene un peso para cada URL de cada sitio que el sistema está midiendo. Este peso determina la probabilidad de que el sistema escoja un URL para cada sonda. El sistema genera los pesos ejecutando complejos algoritmos iterativos frente a los datos de tráfico y los registros de sondeo de la base de datos 200. Un análisis de los datos de tráfico puede discernir qué URLs han sido visitados, cada cuánto tiempo los usuarios han visitado esos URLs. El resultado del análisis garantiza que el sistema realiza el muestreo de anuncios de estos URLs en proporciones similares, dadas ciertas limitaciones como un número máximo de sondas que asignar a cualquier URL individual. Los datos de la partición de registros de muestreo 510 de la base de datos 200 es útil para determinar qué URLs tienen necesidad de tratamiento especial debido al comportamiento pasado (por ejemplo, un URL es muestreado menos infrecuentemente si el sistema nunca ha detectado un anuncio en el URL). El sistema de creación de mapas de sondeo 320 puebla las áreas de mapa de sondeo 524 de la base de datos 200. El sistema de creación de mapas de sondeo 320 accede a los datos del área de mapa de sondeo 524 para asignar las sondas. El sistema de resumen estadístico 230 accede a los datos del área del mapa de sondeo 524 para determinar qué URLs deberían haber modificado a escala sus rotaciones para contrarrestar el efecto de la aplicación forzosa de las limitaciones del mapa de sondeo.

El área de definición de reglas de extracción de anuncios 526 describe etiquetas de Lenguaje de Marcas Extensible (“XML”), que representan típicamente un documento HTML normalizado, que indican aquellas partes del contenido que el sistema considera que son anuncios. El sistema define una regla de extracción en términos de “estructura XML” y “características XML”. La “estructura XML” se refiere a la colocación de diversos nodos XML en relación con otros nodos XML. Por ejemplo, un nodo ancla (“A”) que contiene un nodo de imagen (“IMG”) es probablemente un anuncio. Después de usar este proceso de detección estructural para concordar con el contenido del anuncio, el sistema examina las características del contenido para determinar si el contenido es un anuncio. Para continuar el ejemplo previo, si el nodo de imagen contiene una característica de enlace (“href”) que contiene la sub-cadena “adserver”, con mucha probabilidad es un anuncio. Las características pueden concordar basadas en una simple

sub-cadena, como en el ejemplo, o una expresión regular más complicada. Otra forma de regla de extracción puede apuntar a un nodo específico en una estructura XML usando alguna forma de especificación de ruta XML, como un "Xpointer". El editor de medios 264 puebla el área de definición de reglas de extracción 526 de la base de datos 200. El extractor de anuncios 526 del sistema de muestreo de anuncios 220 accede a los datos del área de definición de reglas de extracción de anuncios 326 para determinar qué partes de cada página sondeada representan un anuncio.

La partición de datos de soporte publicitario 530 de la base de datos 200 comprende tablas de bases de datos que están segmentadas lógicamente en área de taxonomía de productos publicitarios 532, información publicitaria 534, lista de tarifas 536, e información de productos publicitarios ampliada 538.

El área de taxonomía de productos publicitarios 532 contiene una taxonomía jerárquica de productos publicitarios, atributos que describen qué anuncio se está anunciando. Esta taxonomía incluye industrias, compañías, productos, sitios web, subsitios web, mensajes, etc. Cada nodo en la jerarquía tiene un tipo que especifica qué clase de entidad representa y un nodo padre. Por ejemplo, la jerarquía puede especificar qué productos existen dentro de las compañías, que a su vez existen dentro de las industrias. El editor de medios 264 puebla el área de taxonomía de productos publicitarios 532 de la base de datos 200. La interfaz de usuario 240 accede a los datos del área de taxonomía de productos publicitarios 532 para generar datos estadísticos y registrar dónde tienden a anunciar las compañías, industrias, etc. La parte cliente web 250 también accede a los datos del área de taxonomía de productos publicitarios 532 para presentar esta información al cliente 140.

El área de información publicitaria 534 contiene los datos que describen qué anuncia cada anuncio único registrado por el sistema. Las tablas del área de información publicitaria 534 asocian los productos publicitarios con los anuncios. Por ejemplo, el sistema puede asociar un tipo de compañía de productos publicitarios con un anuncio específico para indicar que el anuncio está anunciando la compañía. El sistema usa los siguientes procedimientos para asociar un producto publicitario con un anuncio:

1. Una "clasificación directa" asigna un producto publicitario directamente al anuncio. Por ejemplo, un editor de medios 264 crea una clasificación directa especificando que un anuncio particular anuncia el producto publicitario "Honda".
2. Una "clasificación de dirección" asigna un producto publicitario a un prefijo de dirección que el sistema usa para concordar con el destino del anuncio. Por ejemplo, un editor de medios 264 crea una clasificación de dirección especificando que la dirección "com.honda" indica un anuncio para Honda. Un anuncio que apunta a "com.honda.www/cars", por lo tanto, asocia el anuncio con Honda.
3. Una "clasificación de ancestro" asigna un ancestro del producto publicitario a un anuncio. Por ejemplo, si una clasificación directa asigna Honda a un anuncio, el producto publicitario de industria "automoción" es un predecesor de Honda. La clasificación de ancestros usa esta relación para asociar automoción al anuncio.

El editor de medios 264 puebla el área de información publicitaria 534 de la base de datos 200. La interfaz de usuario 240 accede a los datos del área de información publicitaria 534 para generar datos estadísticos.

El área de lista de tarifas 536 contiene datos que describen el coste de los anuncios en un sitio web. Estos costes incluyen valores monetarios para cada forma, tamaño, o duración de ejecución específicos que los anunciantes de la Internet 100 usan para determinar el coste de las compras de anuncios. El sistema almacena datos de lista de tarifas para cada sitio web que el sistema sondea. El editor de medios 264 puede el área de lista de tarifas 536 en la base de datos 200. La interfaz de usuario 240 accede a los datos del área de lista de tarifas 536 para generar datos estadísticos.

El área de información de productos publicitarios ampliada 538 contiene información adicional acerca de tipos específicos de productos publicitarios no captados inmediatamente en la jerarquía de taxonomía. Específicamente, esto incluye información adicional relacionada con sitios web y compañías, como URLs de información de contacto de la compañía, sitio web y kit de medios. Esta información amplía la utilidad del sistema proporcionando información adicional al cliente 140 acerca de las entidades sondeadas. Por ejemplo, un cliente 140 puede seguir un hiperenlace a la información de contacto de la compañía directamente a partir de un informe del sistema. El editor de medios 264 puebla el área de información de productos publicitarios ampliada 538 de la base de datos 200. La parte cliente web 250 accede a los datos del área de información de productos publicitarios ampliada 538 para suministrar información de valor añadido a un cliente 140.

La partición de resumen publicitario 540 de la base de datos 200 comprende tablas de bases de datos que están segmentadas lógicamente en áreas de estadísticas publicitarias 542, integridad de datos 544 y resumen de información publicitaria 546.

El área de estadísticas publicitarias 542 describe cada cuánto tiempo aparece un anuncio en cada sitio web. El sistema calcula y almacena las siguientes estadísticas en esta área.

1. La proporción de vistas de páginas que presentan un anuncio respecto al número total de vistas de páginas. El sistema determina esta estadística analizando los registros de sondeo.
2. El número de impresiones que recibió un anuncio. El sistema determina esta estadística midiendo niveles de tráfico para el sitio web usando la definición de sitio y los datos de tráfico, y multiplicando esa medición por la proporción de vistas de páginas calculada anteriormente.
3. La cantidad de gastos que recibió un anuncio. El sistema determina esta estadística aplicando la información de lista de tarifas al número de impresiones que el anuncio recibe calculado anteriormente.

El sistema de resumen estadístico 230 puebla el área de estadísticas publicitarias 542 de la base de datos 200. La parte cliente web 250 accede a los datos del área de estadísticas publicitarias 542 para informar de los gastos, las impresiones y la rotación de anuncios al cliente 140.

El área de integridad de datos 544 contiene información en profundidad acerca de valores estadísticos atípicos y otras anomalías potenciales que resultan de análisis de tendencias y segmentos de tiempo. Esta monitorización y análisis automatizado garantiza que el sistema contendrá datos de análisis exactos. Además, el sistema usa información publicitaria del mundo real, como entrada al sistema, para verificar la exactitud de los datos de análisis. El sistema de análisis de integridad de datos, realizado por el sistema de resumen estadístico 230, puebla el área de integridad de datos 544 de la base de datos 200. El operador 262 accede al área de integridad de datos 544 para detectar errores potenciales y monitorizar la salud del sistema general.

El área de resumen de información publicitaria 546 resume la información publicitaria en un formato que sea compacto y fácil de distribuir. El sistema extrae los datos de esta área de la partición de datos de soporte publicitario 530. Aunque los datos no son tan descriptivos como los datos de la partición de datos de soporte publicitario 530, proporcionan la capacidad de realizar rápidamente una consulta precisa. La partición de datos de soporte publicitario 530 asocia cada anuncio con una compañía, producto, o industria. Si el sistema asocia múltiples productos publicitarios del mismo tipo con un anuncio, se escoge un único productos publicitarios para resumir esos asociados usando un sistema de prioridad de asignación, de la siguiente manera:

1. Los productos publicitarios asociados con un anuncio usando clasificación directa reciben la prioridad más alta posible, “M”.
2. Los productos publicitarios asociados con un anuncio usando clasificación de dirección reciben una prioridad igual a la longitud de la cadena del prefijo de dirección al que están asignados, por lo tanto, una cadena de prefijo de dirección larga recibirá una prioridad más alta que una cadena de prefijo de dirección corta.
3. Los productos publicitarios asociados con un anuncio usando clasificación de ancestros reciben la prioridad del ancestro asignado.
4. El anuncio recibe el producto publicitario de prioridad más alta en cada tipo.
5. Cuando se asignan a un anuncio dos ancestros que tienen el mismo tipo y prioridad, se produce un conflicto y debe ser corregido por el editor de medios 264.

El sistema de resumen estadístico 230 puebla el área de resumen de información publicitaria 546 de la base de datos 200. La parte cliente web 250 accede al área de resumen de información publicitaria 546 para generar informes para el cliente 140.

La siguiente descripción trata de una realización de la estructura de base de datos ilustrada en la Figura 5. Este modelo de datos se codifica en una base de datos Oracle. La estructura de la tabla comprende tres entornos, el esquema básico, el esquema de análisis, y la parte cliente. El esquema básico describe el entorno de la parte servidor que permite al “sondeador en la nube” dirigir procesos autónomos en directo que buscan continuamente por toda la Web anotando la actividad publicitaria y los operadores y los editores de medios para la presente invención para dirigir, monitorizar y aumentar la información proporcionada por el “sondeador en la nube”. El esquema de análisis es el entorno de la parte servidor que permite al sistema de muestreo de anuncios, también conocido como OMNIAC, aplicar rigurosos procedimientos de análisis de datos a la información reunida desde la Web. El esquema de la parte cliente ayuda a un cliente de la presente invención a acceder a datos, construir cadenas de consulta de bases de datos, y generar informes.

Los objetos de base de datos que comprenden el “esquema básico” son usados con más frecuencia por diversos componentes del sistema OMNIAC. Las bases de código que se basan en este esquema incluyen la implementación de los procesos de la parte servidor que sacan anuncios de la Web. Además, los esquemas de base de datos utilizados por otros componentes asociados con el OMNIAC están compuestos de algunas o todas las tablas del esquema básico. El esquema básico está compuesto conceptualmente de cuatro subesquemas que incluyen publicidad, anuncios, sondeo y sitios. El subesquema de publicidad contiene información acerca de entidades de “productos publicitarios” junto con las cuales se está anunciando cada anuncio. El subesquema de anuncios describe los anuncios que el sistema ha localizado y analizado. El subesquema de sondeo define “cuándo”, “dónde”, y “cómo” para el proceso de sondeo. El subesquema de sitios describe sitios web, incluyendo definiciones de sitios estructurales e información de lista de tarifas.

De los cuatro subesquemas, Publicidad sirve para el propósito más general y es, por lo tanto, al que se hace referencia con más frecuencia. La tabla primaria en este subesquema es ADVERTISABLE, la cual define productos publicitarios. Muchas de las entidades conceptuales en el universo de OMNIAC son productos publicitarios: industrias, compañías, productos, servicios y sitios web están todos definidos aquí. El campo de tipo, que hace referencia a la tabla ADVERTISABLE_TYPE, distingue entre diferentes tipos de productos publicitarios, y el campo padre organiza los registros jerárquicamente, estableciendo tales relaciones como industria-contiene-compañía y compañía-produce-producto.

Además de la agrupación inherente implicada por la relación padre-hijo definida en ADVERTISABLE, se usa ADVERTISABLE_GROUP_MEMBER para agrupar más los productos publicitarios. Ejemplos de grupos definidos de este modo incluyen clases de automóviles, segmentos de la industria de viajes, y tipos de hardware informático.

Otras tablas del subesquema de publicidad sirven para definir qué es anunciado por cada anuncio. ADVERTISES se usa para asociar productos publicitarios directamente con anuncios. LOCATION ADVERTISES, CLASSIFIED LOCATION y LOCATION_MATCHES también asocian indirectamente productos publicitarios con anuncios a través de la dirección de destino del anuncio.

Los “anuncios” a los que se hace referencia anteriormente son referencias a registros en AD, la tabla primaria en el subesquema de anuncios. El subesquema de anuncios sirve para definir cada anuncio en el universo de OMNIAC. Cada anuncio único tiene un registro en AD, junto con uno o más registros en AD_DEFINITION. Las definiciones de anuncios son fragmentos únicos de XML que OMNIAC ha recuperado de la Web. Ads son anuncios únicos definidos por conjuntos de definiciones de anuncios determinados para que sean equivalentes durante la clasificación automatizada.

Otras tablas en Anuncios contienen atributos de anuncios, a los que se hace referencia por AD y AD_DEFINITION. AD_TECHNOLOGY describe tecnologías web conocidas usadas para producir anuncios, mientras que TEXT describe contenido textual para ciertos anuncios. FUZZY_WEB_LOCATION contiene direcciones difusas encontradas en los anuncios. Una dirección difusa es un URL que tiene que ser procesado por el sistema, como un ancla o una imagen. Una vez que el OMNIAC ha cargado una dirección difusa, se hace una referencia a MIME_CONTENT si el URL hace referencia a una imagen, o DEST_WEB_LOCATION si el URL hace referencia a otra página HTML.

Siguiendo adelante, el subesquema de sondeo controla el comportamiento del sondeo del OMNIAC y los componentes de extracción de anuncios. El propósito primario de este esquema es definir conjuntos de objetivos. Un conjunto de objetivos es una construcción conceptual que ordena al OMNIAC buscar en un conjunto de páginas en ciertos intervalos, extrayendo los anuncios usando un conjunto de reglas llamadas reglas de extracción. Cada conjunto de objetivos está definido por una fila en TARGET_SET.

Las frecuencias, direcciones, y reglas de extracción que constituyen cada conjunto de objetivos están definidas en STROBE, AD_WEB-LOCATION, y EXTRACTION_RULE, respectivamente. Las relaciones muchos-a-muchos entre filas de estas tablas se definen en TS_RUNS_AT, TS_PROBES, y TS_APPLIES.

El subesquema cuarto y final es Sitios, que simplemente registra información acerca de sitios web. Cada sitio o subsitio definido en la jerarquía de productos publicitarios tiene un registro correspondiente en SITE_INFO, junto con un número de filas en SITE_DOMAIN y SITE_MONTHLY_DATA. SITE_DOMAIN describe la estructura física de un sitio en términos de raíces URL inclusivas y exclusivas. SITE_MONTHLY_DATA registra listas de tarifas de anuncios, estimaciones de tráfico de terceros, y estadísticas de caché para cada sitio según una base mensual.

El esquema de análisis es una prolongación al esquema básico que incluye varias tablas adicionales pobladas por el módulo de análisis del OMNIAC. El módulo de análisis es la unidad a cargo del procesamiento de la información contenida en el esquema básico, produciendo un conjunto de datos equilibrado que describe con exactitud la actividad publicitaria.

Como el esquema básico, el esquema de análisis está compuesto de cuatro subesquemas conceptuales compuestos de tablas que implementan la funcionalidad común. Estos subesquemas incluyen descomposición publicitaria, resumen de vistas de anuncios, estadísticas de espacios, y estadísticas de sitios. El subesquema de descomposición publicitaria contiene información acerca de cada anuncio del sistema, incluyendo atributos y qué está anunciando el anuncio. El subesquema de resumen de vistas de anuncios resume las vistas de anuncios, registrando cuántas veces fue visto cada anuncio en cada espacio en el transcurso de un día. El subesquema de estadísticas de espacios describe la rotación de anuncios para cada espacio durante cada periodo de tiempo. El subesquema de estadísticas de sitios describe información de sitios, incluyendo rotación de anuncios para cada periodo de tiempo.

La tabla primaria en el subesquema de descomposición publicitaria es AD_INFO, que contiene registros desnormalizados que describen atributos de anuncios. Los registros de AD_INFO se toman como datos de entrada de control de ID's en la tabla AD; existe un registro de AD_INFO por cada registro de AD que ha sido clasificado completamente y representa un anuncio válido. AD_INFO es poblado por el módulo de análisis de las relaciones publicitarias descritas en las tablas del esquema básico ADVERTISES y LOCATION_ADVERTISES.

Los campos AD_INFO que especifican qué es anunciado por un anuncio son: CATEGORY (industria), ORGANIZATION (compañía), ORGANIZATION_GROUP (segmento industrial), ORGANIZATION_OVERGROUP, COMMODITY (producto/servicio), COMMODITY_GROUP (segmento de productos/servicios), COMMODITY_OVERGROUP, y MESSAGE.

AD_INFO también incluye campos que describen varios atributos no publicitarios. FORMAT, que hace referencia a AD_SLOT_TYPE.ID, especifica el factor de forma de un anuncio. TECHNOLOGY, que hace referencia a AD_TECHNOLOGY2.ID, especifica la tecnología usada para implementar el anuncio. DEFINITION, IMAGE y DESTINATION especifican los registros de AD_DEFINITION, IMAGE y DEST_WEB_LOCATION asociados con el anuncio. Estos tres campos reflejan los campos de la tabla AD.

El esquema de descomposición publicitaria contiene unas pocas tablas además de AD_INFO. ADV_IMPLICATION es una caché de implicaciones de productos publicitarios derivadas de la jerarquía en ADVERTISABLE. Esto se usa para acelerar el funcionamiento del módulo de análisis. AD_INFO_FLATTENED es una versión consultada más fácilmente de AD_INFO que contiene pares anuncio/producto publicitario para cada uno de los campos en AD_INFO que hacen referencia a ADVERTISABLE. Por último, AD_TECHNOLOGY2 describe tecnologías de anuncios interpretadas por el módulo de análisis que son presentables al usuario en la parte cliente.

El subesquema de resumen de vistas de anuncios abarca la única tabla PLACEMENT_SUMMARY. PLACEMENT_SUMMARY se toma como dato de entrada de control del día, el anuncio y el espacio, y contiene, en el campo CNT, el número de veces que fue visto un anuncio en un espacio en un día particular.

El módulo de análisis puebla PLACEMENT_SUMMARY agregando los accesos registrados en las tablas APD_n, de las cuales existe una para cada día, siendo n la ID del día en cuestión. Estas tablas son creadas y pobladas por la parte servidor como flujo de accesos a anuncios dentro del sistema.

El tercer subesquema en el esquema de Análisis es Estadísticas de espacios. Este subesquema describe el comportamiento de los anuncios en el contexto de los espacios además de información acerca de los propios espacios. Un espacio es una dirección en la Web en la que rotan anuncios, actualmente definidos en términos de ID de dirección (una referencia a AD_WEB_LOCATION.ID) e ID de regla de extracción (una referencia a EXTRACTION_RULE.ID).

La tabla primaria en las Estadísticas de espacios es SLOT_AD_VIEWS, que registra las vistas totales y la frecuencia relativa para cada anuncio en cada espacio durante cada periodo de tiempo. La clave primaria de esta tabla está compuesta de los campos PERIOD_TYPE, PERIOD, LOCATION_ID, RULE_ID y AD_ID. Existen dos campos fuera de la clave primaria: CNT contiene el número total de vistas de anuncios, y FREQUENCY contiene la frecuencia relativa.

En este subesquema también está SLOT_SUMMARY, que registra información general de espacios exteriores al contexto de los anuncios individuales. Por consiguiente, esta tabla se toma como dato de entrada de control de los campos PERIOD_TYPE, PERIOD, LOCATION_ID y RULE_ID. El campo CNT registra las vistas de anuncios totales en el espacio; este campo se divide en el SLOT_AD_VIEWS.CNT para determinar la frecuencia relativa. También en SLOT_SUMMARY está un campo SLOT_TYPE que especifica el tipo de anuncio visto con más frecuencia en el espacio, y SITE_ID, que especifica dentro de qué sitio reside el espacio.

La tabla final en el subesquema de estadísticas de espacios es SLOT_TYPE_COUNT. Esta tabla se usa para determinar qué valor usar en SLOT_SUMMARY.SLOT_TYPE. Se registra el número de veces que fue visto cada formato de anuncio, y el tipo de espacio que recibe la mayoría de las vistas se pone dentro de SLOT_SUMMARY.SLOT_TYPE.

La Figura 6 es un diagrama de bloques funcionales del sistema de prevalencia publicitaria 130. La memoria 610 del sistema de prevalencia publicitaria 130 almacena los componentes de software, de acuerdo con la presente invención, que analizan los datos de tráfico por la Internet 100, muestrean los datos publicitarios a partir de esos datos de tráfico, y generan datos de resumen que caracterizan los datos publicitarios. El bus de sistema 612 conecta la memoria 610 del sistema de prevalencia publicitaria 130 al adaptador de red del protocolo de control de transmisión/protocolo internet ("TCP/IP") 614, la base de datos 200 y el procesador central 616. El adaptador de red TCP/IP 614 es el mecanismo que facilita el paso del tráfico de red entre el sistema de prevalencia publicitaria 130 y la Internet 100. El procesador central 616 ejecuta las instrucciones programadas almacenadas en la memoria 610.

La Figura 6 muestra los módulos funcionales del sistema de prevalencia publicitaria 130 dispuestos como un modelo de objeto. El modelo de objeto agrupa los programas de software orientado a objetos en componentes que realizan las funciones y aplicaciones fundamentales en el sistema de prevalencia publicitaria 130. Una implementación adecuada de los componentes del programa de software orientado a objetos de la Figura 6 puede usar la especificación Enterprise JavaBeans. El libro de Paul J. Perrone y col, titulado “Building Java Enterprise Systems with J2EE” (Sams Publishing, junio de 2000) proporciona una descripción de una aplicación empresarial Java desarrollada usando la especificación Enterprise JavaBeans. El libro de Matthew Reynolds, titulado “Beginning E-Commerce” (Wrox Press Inc., 2000) proporciona una descripción del uso de un modelo de objeto en el diseño de un servidor web para una aplicación de comercio electrónico.

El modelo de objeto para la memoria 610 del sistema de prevalencia publicitaria 130 emplea una arquitectura de tres capas que incluye la capa de presentación 620, la partición de objetos de infraestructura 630, y la capa de lógica de negocios 640. El modelo de objeto además divide la capa de lógica de negocios 640 en dos particiones, la partición de objetos de servicio de aplicación 650 y la partición de objetos de datos 660.

La capa de presentación 620 contiene los programas que gestionan la interfaz entre el sistema de prevalencia publicitaria 130 y el cliente 140, el administrador de cuentas 260, el operador 262, y el editor de medios 264. En la Figura 6, la capa de presentación 620 incluye la interfaz TCP/IP 622, la parte cliente web 624, y la interfaz de usuario 626. Una implementación adecuada de la capa de presentación 620 puede usar servlets Java para interactuar con el cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264 de la presente invención a través del protocolo de transferencia de hipertexto (“HTTP”). Los servlets Java se ejecutan dentro de un servidor de solicitud/respuesta que se ocupa de mensajes de solicitud procedentes del cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264 y devuelven mensajes de respuesta al cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264. Un servlet Java es un programa de Java que se ejecuta dentro de un entorno de servidor web. Un servlet Java toma una solicitud como entrada, analiza sintácticamente los datos, realiza operaciones lógicas, y emite una respuesta de vuelta al cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264. La plataforma de ejecución Java combina los servlets Java para dar servicio simultáneamente a muchas solicitudes. Una interfaz TCP/IP 622 que usa servlets Java funciona como un servidor web que se comunica con el cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264 usando el protocolo HTTP. La interfaz TCP/IP 622 acepta solicitudes HTTP del cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264 y pasa la información de la solicitud al objeto de visita 642 en la capa de lógica de negocios 640. El objeto de visita 642 pasa la información de resultados devuelta desde la capa de lógica de negocios 640 a la interfaz TCP/IP 622. La interfaz TCP/IP 622 envía estos resultados de vuelta al cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264 en una respuesta HTTP. La interfaz TCP/IP 622 intercambia datos con la Internet 100 a través del adaptador de red TCP/IP 614.

La partición de objetos de infraestructura 630 contiene los programas que realizan funciones administrativas y de sistema en nombre de la capa de lógica de negocios 640. La partición de objetos de infraestructura 630 incluye el sistema operativo 636, y un componente de programa de software orientado a objetos para la interfaz de sistema de gestión de bases de datos (“DBMS”) 632, la interfaz de administrador 634, y la plataforma de ejecución Java 638.

La capa de lógica de negocios 640 contiene los programas que realizan lo esencial de la presente invención. La capa de lógica de negocios 640 en la Figura 6 incluye múltiples instancias del objeto de visita 642. Existe una instancia separada del objeto de visita 642 para cada sesión de cliente iniciada por la parte cliente web 624 o la interfaz de usuario 626 a través de la interfaz TCP/IP 622. Cada objeto de visita 642 es una vaina de sesión dinámica que incluye un área de almacenamiento persistente desde el inicio hasta la terminación de la sesión del cliente, no sólo durante una única interacción o llamada de procedimiento. El área de almacenamiento persistente contiene información asociada con el URL 114, 116, 118 o el cliente 140, el administrador de cuentas 260, el operador 262 y el editor de medios 264. Además, el área de almacenamiento persistente contiene datos intercambiados entre el sistema de prevalencia publicitaria 130 y el sistema de muestreo de tráfico 120 a través de la interfaz TCP/IP 622 como los conjuntos de resultados de consultas procedentes de una consulta de la base de datos 200.

Cuando el sistema de muestreo de tráfico 120 termina de recopilar información acerca de un URL 114, 116, 118, envía un mensaje a la interfaz TCP/IP 622 que invoca un procedimiento para crear un objeto de visita 642 y almacena la información acerca de la conexión en el estado del objeto de visita 642. El objeto de visita 642, a su vez, invoca un procedimiento en la aplicación de análisis de tráfico 652 para procesar la información recuperada por el sistema de muestreo de tráfico 120. La aplicación de análisis de tráfico 652 almacena los datos procesados procedentes del sistema de anonimato 310 y el sistema de creación de mapas de sondeo 320 en el estado de los datos de análisis de tráfico 662 y la base de datos 200. Las Figuras 7A y 7B describen, con mayor detalle, el proceso que sigue la aplicación de análisis de tráfico 652 para cada URL 114, 116, 118 obtenido del sistema de muestreo de tráfico 120. Aun cuando la Figura 6 representa el procesador central 616 controlando la aplicación de análisis de

tráfico 652, ha de entenderse que la función realizada por la aplicación de análisis de tráfico 652 puede ser distribuida a un sistema separado configurado de manera similar al sistema de prevalencia publicitaria 130.

Después de que la aplicación de análisis de tráfico 652 procesa un URL 114, 116, 118 identificado por el sistema de muestreo de tráfico 120, el objeto de visita 642 invoca un procedimiento en la aplicación de muestreo de anuncios 654 para recuperar el URL 114, 116, 118 del sitio web 110. La aplicación de muestreo de anuncios 654 procesa la página web recuperada extrayendo los anuncios incrustados y clasificando esos anuncios. La aplicación de muestreo de anuncios 654 almacena los datos recuperados por el sistema de recuperación de páginas web 322 y procesados por el entorno de emulación de navegador web 324, el extractor de anuncios 326 y el clasificador estructural 328 en el estado de los datos de muestreo de anuncios 664 y la base de datos 200. Las Figuras 7A, 7C, 7D y 7E describen, con mayor detalle, el proceso que sigue la aplicación de muestreo de anuncios 654 para cada URL 114, 116, 118 identificado por el sistema de muestreo de tráfico 120. Aun cuando la Figura 6 representa el procesador central 616 controlando la aplicación de muestreo de anuncios 654, una persona experta en la materia se dará cuenta de que el procesamiento realizado por la aplicación de muestreo de anuncios 654 puede ser distribuida a un sistema separado configurado de manera similar al sistema de prevalencia publicitaria 130.

Después de que la aplicación de análisis de tráfico 652 y el sistema de muestreo de anuncios 654 procesan el URL 114, 116, 118 identificado por el sistema de muestreo de tráfico 120, el objeto de visita 642 invoca un procedimiento en la aplicación de resumen estadístico 656 para calcular estadísticas de resumen para los datos. La aplicación de resumen estadístico 656 calcula estadísticas de impresión publicitaria, gastos y estadísticos de valoración para cada anuncio incrustado en el URL 114, 116, 118. La aplicación de resumen estadístico 656 almacena los datos estadísticos en el estado de los datos de resumen estadístico 666 y la base de datos 200. La Figura 7F describe, con mayor detalle, el proceso que sigue la aplicación de resumen estadístico 656 para cada URL 114, 116, 118 identificado por el sistema de muestreo de tráfico 120. Aun cuando la Figura 6 representa el procesador central 616 controlando la aplicación de resumen estadístico 656, una persona experta en la materia se da cuenta de que la función realizada por la aplicación de resumen estadístico 656 puede ser distribuida a un sistema separado configurado de manera similar al sistema de prevalencia publicitaria 130.

La Figura 7A es un organigrama de un proceso en el sistema de prevalencia publicitaria 130 que mide el valor de los anuncios en línea rastreando y comparando la actividad publicitaria en línea a través de todas las principales industrias, canales, formatos publicitarios y tipos. El proceso 700 comienza, en la etapa 710, muestreando los datos de tráfico procedentes de la Internet 100. La Figura 7B describe la etapa 710 con mayor detalle. La etapa 720 usa los datos de tráfico muestreados de la etapa 710 para realizar la selección de sitios, y definir y refinar las definiciones de sitios para el sistema de prevalencia publicitaria 130. La etapa 730 usa el resultado del proceso de selección y definición de sitios para generar un mapa de sondeo basado en los datos de tráfico muestreados. La Figura 7C describe la etapa 730 con mayor detalle. La etapa 740 usa el mapa de sondeo de la etapa 730 para visitar la Internet 100 para reunir datos de muestras procedentes de los sitios de sondeo identificados en la etapa 730. La Figura 7D describe la etapa 740 con mayor detalle. Para cada URL recuperado en la etapa 740, la etapa 750 extrae los anuncios del URL, la etapa 760 clasifica cada anuncio, y la etapa 770 calcula las estadísticas para cada anuncio. Las Figuras 7E y 7F describen, respectivamente, las etapas 760 y 770 con mayor detalle. Por último, el proceso 700 realiza comprobaciones de integridad de datos en la etapa 780 para verificar la integridad de los datos y los resultados del análisis en el sistema.

La Figura 7B es un organigrama que describe, con mayor detalle, el proceso de muestreo de datos de tráfico de la Figura 7A, etapa 710. El proceso 710 comienza en la etapa 711 reuniendo datos procedentes de un monitor de tráfico web como el sistema de muestreo de tráfico 120. El proceso 710 vacía la información de usuario de los datos recuperados por el monitor de tráfico web en la etapa 712 para depurar los datos y garantizar el anonimato de la muestra. Para cada URL de la muestra depurada, la etapa 713 mide el número de vistas de páginas web observadas en los datos de tráfico. La etapa 714 completa el proceso 710 extrapolando estadísticamente el número medido de vistas de páginas web en la muestra a todo el universo de la Internet 100.

La Figura 7C es un organigrama que describe, con mayor detalle, el proceso de generación de un mapa de sondeo basado en los datos de tráfico muestreados de la Figura 7A, etapa 730. El proceso 730 comienza en la etapa 731 analizando un subconjunto de los datos de tráfico muestreados que entra dentro de las definiciones de sitio elegibles. Después del análisis de la etapa 731, la etapa 732 construye un mapa de sondeo inicial basado en los datos de tráfico de muestra. La etapa 733 analiza los resultados históricos de mediciones de anuncios de la base de datos 200 para los URLs del mapa de sondeo inicial. La etapa 734 usa estos resultados históricos así como parámetros del sistema para optimizar el plan de muestreo. La etapa 735 completa el proceso 730 monitorizando los resultados de las muestras y ajustando el sistema según sea necesario.

La Figura 7D es un organigrama que describe, con mayor detalle, el proceso de sondeo de la Internet 100 para reunir datos de muestra de la Figura 7A, etapa 740. El proceso 740 comienza en la etapa 741, buscando una página web de la Internet 100. La página web de la etapa 741 se pasa a un entorno de emulación de navegador web en la etapa 742 para simular la presentación de la página web en un navegador. Esta simulación permite al sistema de

prevalencia publicitaria 130 detectar los anuncios incrustados en la página web. Estos anuncios pueden estar incrustados en código JavaScript, applet Java o código servlet, o código de interfaz de pasarela común como Perl script. Además, la simulación en la etapa 742 permite al sistema de prevalencia publicitaria 130 detectar anuncios dinámicos e interactivos en la página web. Después de la simulación en la etapa 742, la etapa 743 extrae los datos de anuncios de la página web y la etapa 744 almacena los datos de anuncios en la base de datos 200. La etapa 745 determina si el proceso 740 tiene que buscar otra página web para reunir más datos de muestras. En la realización preferida, el proceso 740 muestrea continuamente páginas web de la Internet 100. Una persona experta en la materia se da cuenta de que la funcionalidad llevada a cabo por la etapa 745 puede asociarse con un sistema de planificación que planificará el sondeo de la Internet 100 para reunir los datos publicitarios de muestra.

La Figura 7E es un organigrama que describe, con mayor detalle, el proceso de clasificación de los datos publicitarios de la Figura 7A, etapa 760. El proceso 760 comienza el análisis de fragmentos de anuncios en la etapa 761 determinando si el fragmento es un duplicado. Cuando el sistema de prevalencia publicitaria 130 encuentra un fragmento de anuncio por primera vez, la etapa 762 analiza la estructura interna del fragmento. Después de la etapa 762, o cuando la etapa 761 determina que el fragmento de anuncio es un duplicado, la etapa 763 recupera el contenido externo del anuncio de la página web. La etapa 764 compara entonces el contenido externo con los anuncios observados previamente. La etapa 765 analiza el resultado de la comparación de la etapa 764 para determinar si el anuncio es un duplicado. Cuando el sistema de prevalencia publicitaria 130 encuentra un anuncio por primera vez, la etapa 766 comienza a procesar el nuevo anuncio registrando la estructura del nuevo anuncio en la base de datos 200. La etapa 767 realiza entonces la clasificación automatizada de anuncios y almacena los tipos de clasificación en la base de datos 200. La etapa 768 completa el procesamiento de un nuevo anuncio realizando la verificación humana de las clasificaciones de anuncios. Después de la etapa 768, o cuando la etapa 765 determina que el anuncio es un duplicado, la etapa 769 actualiza el registro de vistas de anuncios en la base de datos 200 para indicar la observación del anuncio.

La Figura 7F es un organigrama que describe, con mayor detalle, el proceso de cálculo de estadísticas publicitarias de la Figura 7A, etapa 770. El proceso 770 comienza el cálculo de las estadísticas publicitarias en la etapa 771 resumiendo los resultados de las mediciones publicitarias. En la etapa 772, el proceso 770 usa el mapa de sondeo generado en la etapa 730 para ponderar los resultados de las mediciones publicitarias. La frecuencia de anuncios se calcula en la etapa 773 para cada solicitud de página web. La etapa 774 usa los datos de tráfico de muestra de la etapa 710 y la frecuencia de anuncios de la etapa 773 para calcular las impresiones publicitarias para cada anuncio. La etapa 775 completa el proceso 770 calculando los gastos de anuncios combinando las impresiones publicitarias de la etapa 774 y los datos de lista de tarifas introducidos por el editor de medios 264 con el módulo de recopilación de listas de tarifas 348 de la interfaz de usuario 240.

Aunque las realizaciones desveladas en la presente invención describen un sistema totalmente funcional, ha de entenderse que existen otras realizaciones que son equivalentes a las realizaciones desveladas en este documento. Como a quienes revisen la presente solicitud se les ocurrirán numerosas modificaciones y variaciones, la presente invención no está limitada a la construcción y el funcionamiento exactos ilustrados y descritos en este documento. Por consiguiente, se pretende que todas las modificaciones y equivalentes adecuados a los que pueda recurrirse entren dentro del alcance de las reivindicaciones.

REIVINDICACIONES

1. Un procedimiento para estimar la prevalencia de contenido digital en una red, comprendiendo el procedimiento:

recibir (710) una estimación de un número de veces que se ha accedido a una página web (110; 110'; 110"; 411-416; 421-424);
solicitar repetidamente (740) la página web (110; 110'; 110"; 411-416; 421-424) y, en respuesta, recibir archivos de contenido;
determinando (750) un número de veces que un primer objeto de contenido (A1; A2; A3) está incluido en los archivos de contenido;
determinando (770) un número total de veces que la página web (110; 110'; 110"; 411-416; 421-424) ha sido solicitada; y
estimando (773, 774) el número de veces que el primer objeto de contenido (A1; A2; A3) ha sido presentado a los visitantes de la página web basado en el número de veces que el primer objeto de contenido (A1; A2; A3) fue incluido en los archivos de contenido, el número total de veces que la página web (110; 110'; 110"; 411-416; 421-424) fue solicitada, y la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424).

2. Un procedimiento según la reivindicación 1, en el que al menos una parte de la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424) es recibida desde un proxy.

3. Un procedimiento según la reivindicación 1, en el que el procedimiento es realizado por un sistema de prevalencia publicitaria (130).

4. Un procedimiento según la reivindicación 1, en el que al menos una parte de la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424) es recibida desde al menos un ordenador de panelista.

5. Un procedimiento según la reivindicación 1, en el que el objeto de contenido (A1; A2; A3) es un anuncio.

6. Un procedimiento según la reivindicación 1, en el que estimar el número de veces que el primer objeto de contenido ha sido presentado a los visitantes comprende:

determinando (773) una tasa de rotación para el objeto de contenido dividiendo el número total de veces que la página web fue solicitada por el número de veces que el primer objeto de contenido fue incluido en los archivos de contenido;
determinando (774) el número de veces que el primer objeto de contenido ha sido presentado a los visitantes multiplicando la estimación del número de veces que se ha accedido a la página web por la tasa de rotación.

7. Un sistema para estimar la prevalencia de contenido digital en una red, comprendiendo el sistema:

un dispositivo de estimación (210) para determinar una estimación de un número de veces que se ha accedido a una página web (110; 110'; 110"; 411-416; 421-424);
un sondeador (220) para solicitar repetidamente la página web (110; 110'; 110"; 411-416; 421-424) y, en respuesta, recibir archivos de contenido;
un sistema de resumen estadístico (230) para determinar un número de veces que un primer objeto de contenido (A1; A2; A3) está incluido en los archivos de contenido, determinar un número total de veces que la página web (110; 110'; 110"; 411-416; 421-424) ha sido solicitada, y estimar el número de veces que el primer objeto de contenido (A1; A2; A3) ha sido presentado a los visitantes de la página web basado en el número de veces que el primer objeto de contenido (A1; A2; A3) fue incluido en los archivos de contenido, el número total de veces que la página web (110; 110'; 110"; 411-416; 421-424) fue solicitada, y la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424) .

8. Un sistema según la reivindicación 7, en el que al menos una parte de la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424) es recibida desde un proxy.

9. Un sistema según la reivindicación 7, en el que el sistema es un sistema de prevalencia publicitaria (130).

10. Un sistema según la reivindicación 7, en el que al menos una parte de la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424) es recibida desde al menos un ordenador de panelista.

11. Un sistema según la reivindicación 7, en el que el objeto de contenido (A1; A2; A3) es un anuncio.

12. Un sistema según la reivindicación 7, en el que el sistema de resumen estadístico estima el número de veces que el primer objeto de contenido ha sido presentado a los visitantes:

determinando (773) una tasa de rotación para el objeto de contenido dividiendo el número total de veces que la página web fue solicitada por el número de veces que el primer objeto de contenido fue incluido con los archivos de contenido;
determinando (774) el número de veces que el primer objeto de contenido ha sido presentado a los visitantes multiplicando la estimación del número de veces que se ha accedido a la página web por la tasa de rotación.

13. Un medio legible por una máquina que almacena instrucciones que, cuando son ejecutadas hacen que una máquina al menos:

reciba (710) una estimación de un número de veces que se ha accedido a una página web (110; 110'; 110"; 411-416; 421-424);
solicite repetidamente (740) la página web y, en respuesta, reciba archivos de contenido;
determine (750) un número de veces que un primer objeto de contenido (A1; A2; A3) está incluido en los archivos de contenido;
determine (770) un número total de veces que la página web (110; 110'; 110"; 411-416; 421-424) ha sido solicitada; y
estime (773, 774) el número de veces que el primer objeto de contenido (A1; A2; A3) ha sido presentado a los visitantes de la página web basado en el número de veces que el primer objeto de contenido (A1; A2; A3) fue incluido en los archivos de contenido, el número total de veces que la página web (110; 110'; 110"; 411-416; 421-424) fue solicitada, y la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424).

14. Un medio legible por una máquina según la reivindicación 13, en el que al menos una parte de la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424) es recibida desde un proxy.

15. Un medio legible por una máquina según la reivindicación 13, en el que las instrucciones almacenadas en el medio legible por una máquina son ejecutadas por un sistema de prevalencia publicitaria (130).

16. Un medio legible por una máquina según la reivindicación 13, en el que al menos una parte de la estimación del número de veces que se ha accedido a la página web (110; 110'; 110"; 411-416; 421-424) es recibida desde al menos un ordenador de panelista.

17. Un medio legible por una máquina según la reivindicación 13, en el que el objeto de contenido (A1; A2; A3) es un anuncio.

18. Un medio legible por una máquina según la reivindicación 13, en el que las instrucciones almacenadas en el medio legible por una máquina estiman el número de veces que el primer objeto de contenido ha sido presentado a los visitantes:

determinando (773) una tasa de rotación para el objeto de contenido dividiendo el número total de veces que la página web fue solicitada por el número de veces que el primer objeto de contenido fue incluido en los archivos de contenido;
determinando (774) el número de veces que el primer objeto de contenido ha sido presentado a los visitantes multiplicando la estimación del número de veces que se ha accedido a la página web por la tasa de rotación.

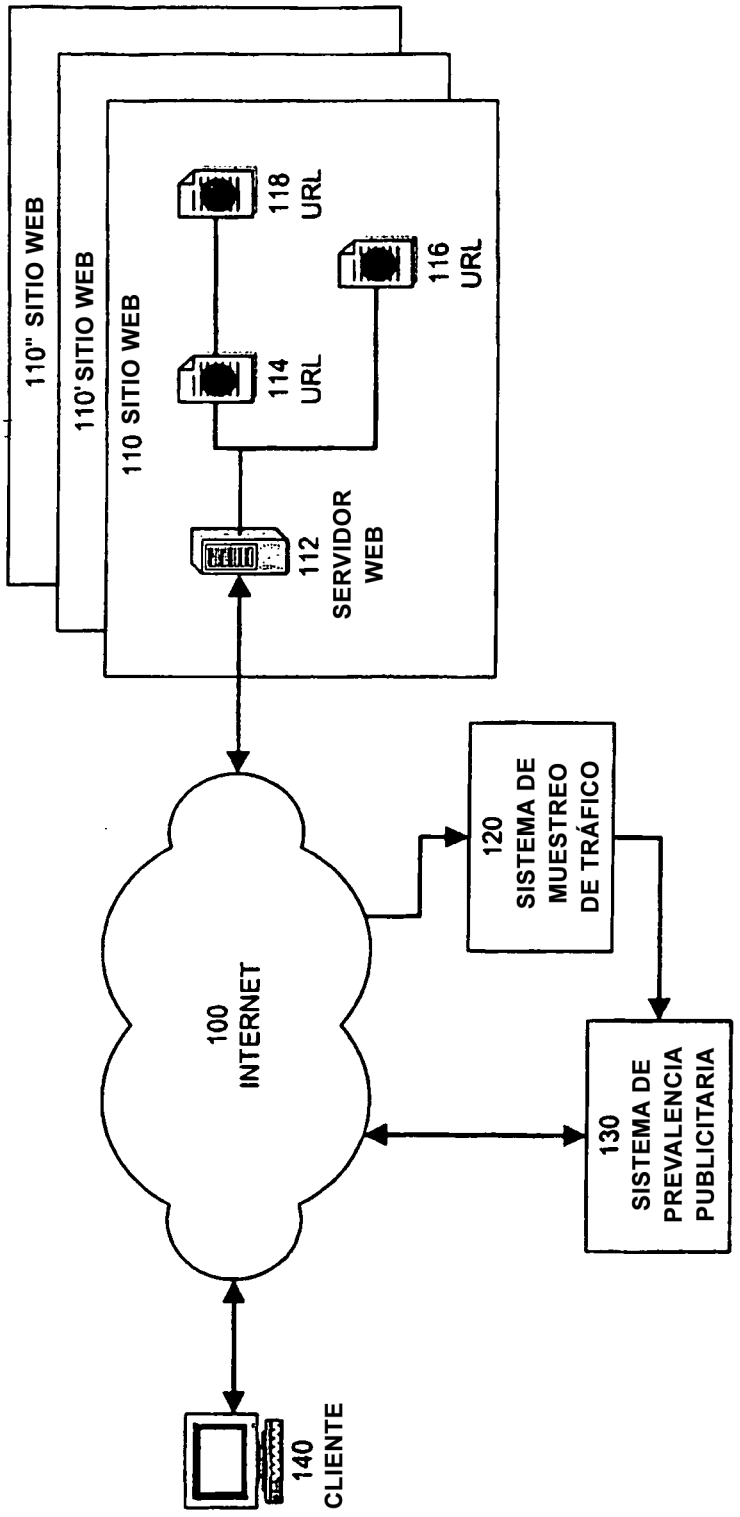
110'' SITIO WEB
110' SITIO WEB
110 SITIO WEB
118 URL
116 URL
114 URL
112 SERVIDOR WEB
100 INTERNET
120 SISTEMA DE MUESTREO DE TRÁFICO
130 SISTEMA DE PREVALENCIA PUBLICITARIA
140 CLIENTE

FIG. 1

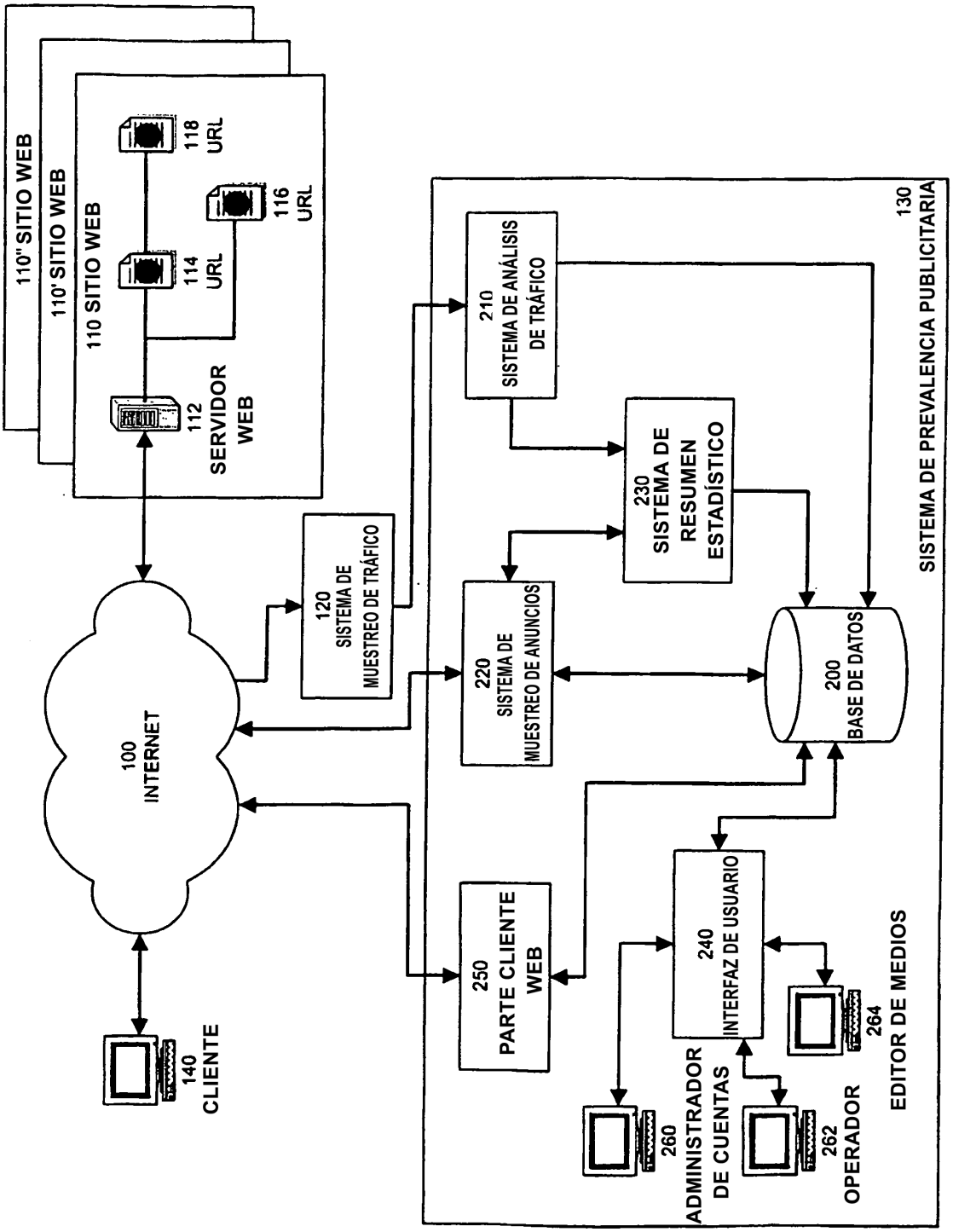
110" SITIO WEB
110' SITIO WEB
110 SITIO WEB
118 URL
116 URL
114 URL
112 SERVIDOR WEB
100 INTERNET
120 SISTEMA DE MUESTREO DE TRÁFICO
210 SISTEMA DE ANÁLISIS DE TRÁFICO
230 SISTEMA DE RESUMEN ESTADÍSTICO
220 SISTEMA DE MUESTREO DE ANUNCIOS
200 BASE DE DATOS
130 SISTEMA DE PREVALENCIA PUBLICITARIA
250 PARTE CLIENTE WEB
240 INTERFAZ DE USUARIO
140 CLIENTE
260 ADMINISTRADOR DE CUENTAS
262 OPERADOR
264 EDITOR DE MEDIOS

FIG. 2

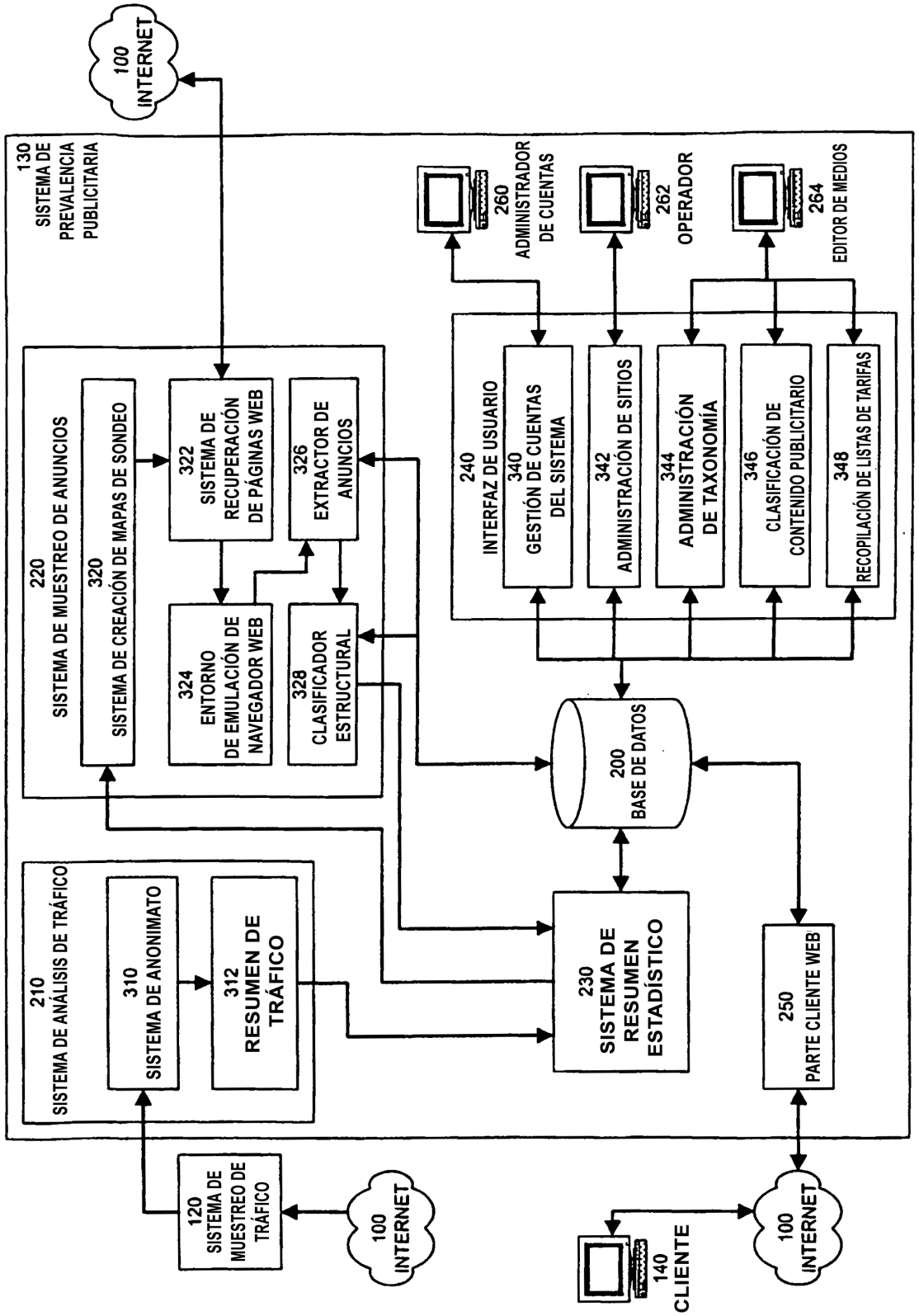
100
INTERNET
130
SISTEMA DE
PREVALENCIA
PUBLICITARIA
260
ADMINISTRADOR
DE CUENTAS
262
OPERADOR
264
EDITOR DE MEDIOS
220
SISTEMA DE MUESTREO DE ANUNCIOS
320
SISTEMA DE CREACIÓN DE MAPAS DE SONDEO
322
SISTEMA DE
RECUPERACIÓN
DE PÁGINAS WEB
326
EXTRACTOR DE
ANUNCIOS
324
ENTORNO
DE EMULACIÓN DE
NAVEGADOR WEB
328
CLASIFICADOR
ESTRUCTURAL
240
INTERFAZ DE USUARIO
340
GESTIÓN DE CUENTAS
DEL SISTEMA
342
ADMINISTRACIÓN DE SITIOS
344
ADMINISTRACIÓN
DE TAXONOMÍA
346
CLASIFICACIÓN DE
CONTENIDO PUBLICITARIO
348
RECOPILACIÓN DE LISTAS DE TARIFAS
200
BASE DE DATOS
210
SISTEMA DE ANÁLISIS DE TRÁFICO
310
SISTEMA DE ANONIMATO
312
RESUMEN DE
TRÁFICO
230
SISTEMA DE
RESUMEN
ESTADÍSTICO
250
PARTE CLIENTE WEB
120
SISTEMA DE
MUESTREO DE
TRÁFICO
100
INTERNET
100
INTERNET
140
CLIENTE
100
INTERNET

FIG. 3

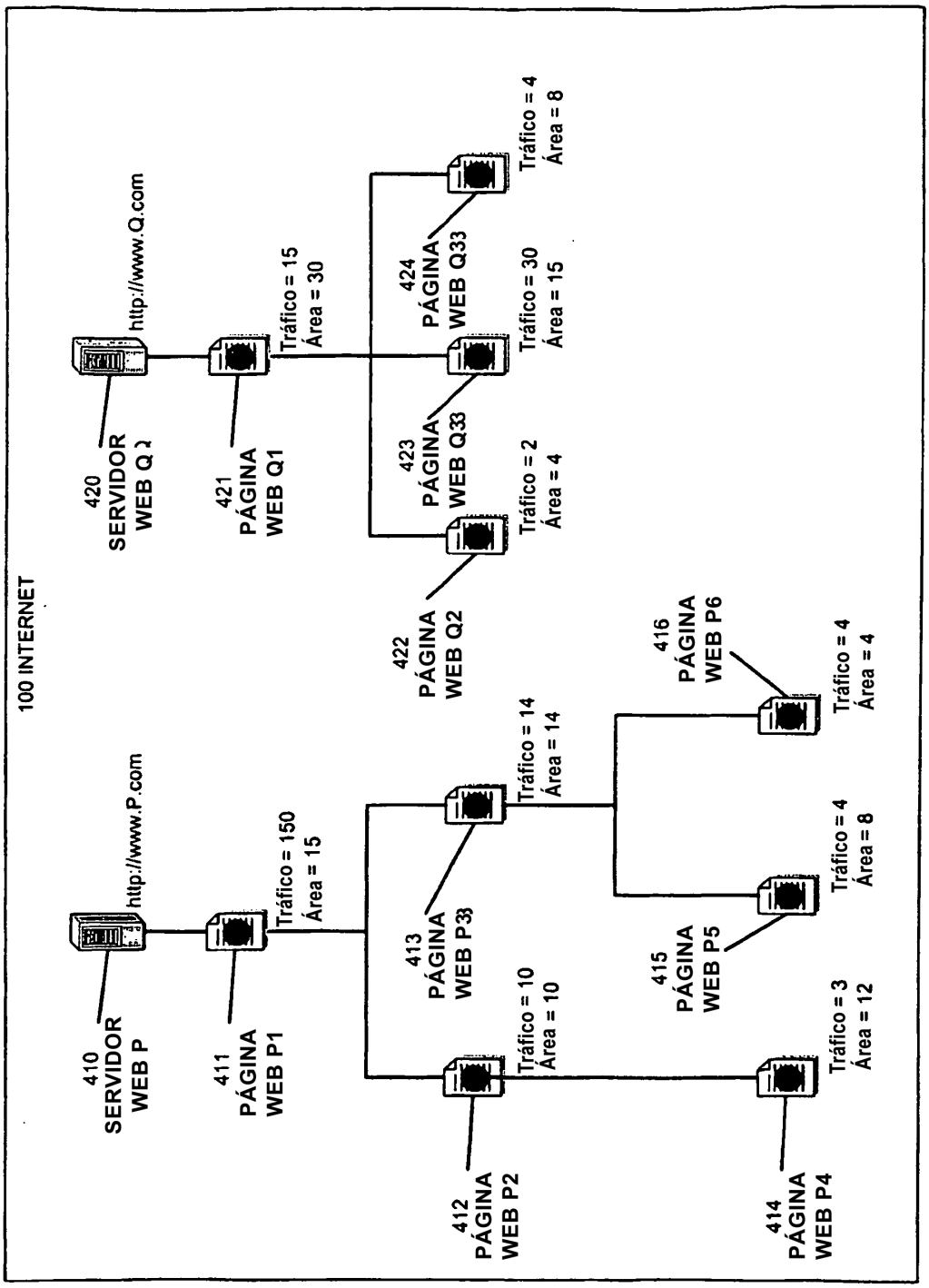
100 INTERNET
http://www.P.com
410 SERVIDOR WEB P
411 PÁGINA WEB P1
Tráfico = 150
Área = 15
412 PÁGINA WEB P2
Tráfico = 10
Área = 10
413 PÁGINA WEB P3
Tráfico = 14
Área = 14
414 PÁGINA WEB P4
Tráfico = 3
Área = 12
415 PÁGINA WEB P5
Tráfico = 4
Área = 8
416 PÁGINA WEB P6
Tráfico = 4
Área = 4
http://www.Q.com
420 SERVIDOR WEB Q
421 PÁGINA WEB Q1
Tráfico = 15
Área = 30
422 PÁGINA WEB Q2
Tráfico = 2
Área = 4
423 PÁGINA WEB Q33
Tráfico = 30
Área = 15
424 PÁGINA WEB Q33
Tráfico = 4
Área = 8

FIG. 4A

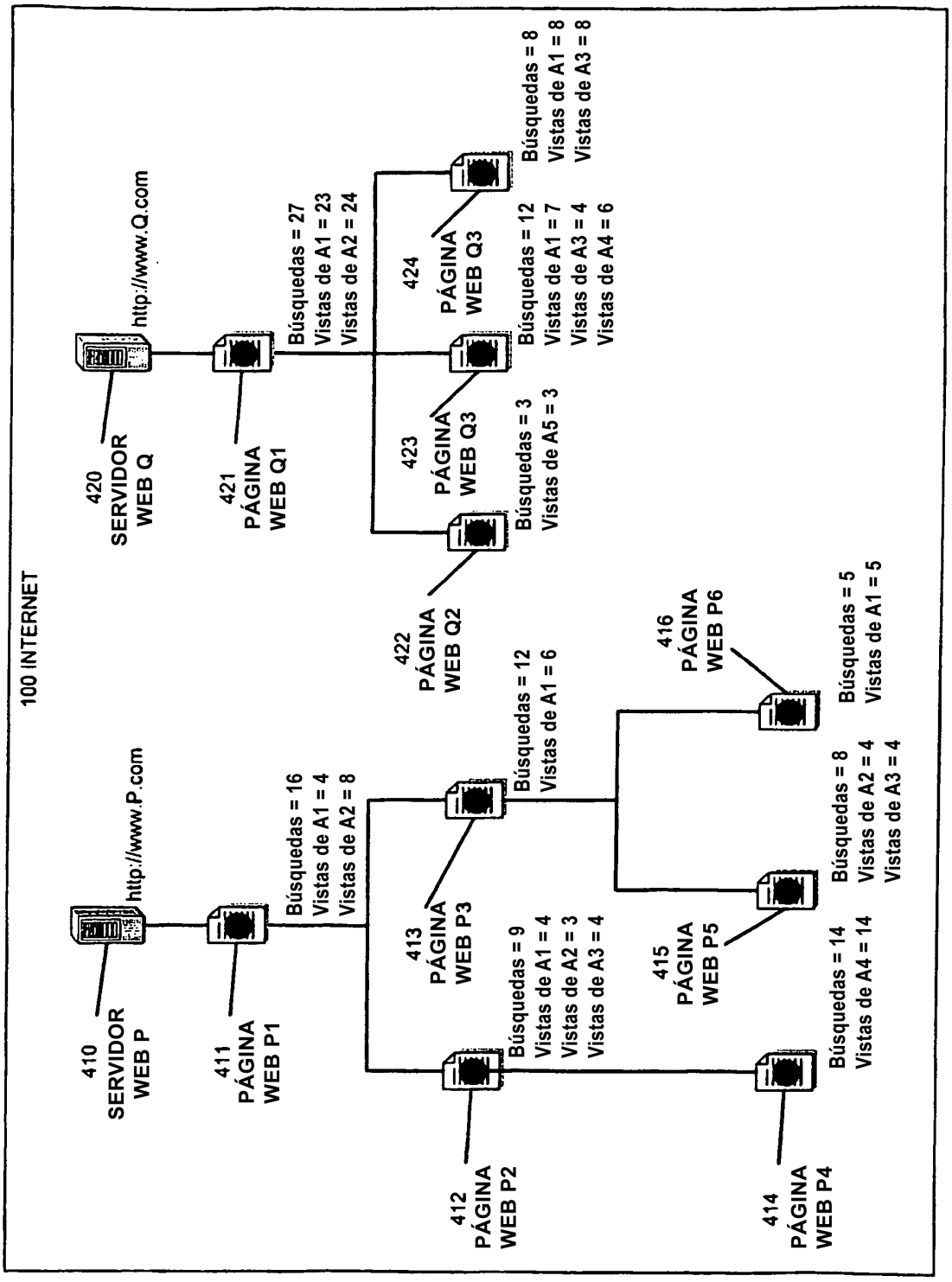
100 INTERNET
http://www.P.com
410 SERVIDOR WEB P
411 PÁGINA WEB P1
Búsquedas = 16
Vistas de A1 = 4
Vistas de A2 = 8
413 PÁGINA WEB P3
Búsquedas = 12
Vistas de A1 = 6
412 PÁGINA WEB P2
Búsquedas = 9
Vistas de A1 = 4
Vistas de A2 = 3
Vistas de A3 = 4
416 PÁGINA WEB P6
Búsquedas = 5
Vistas de A1 = 5
415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

FIG. 4B

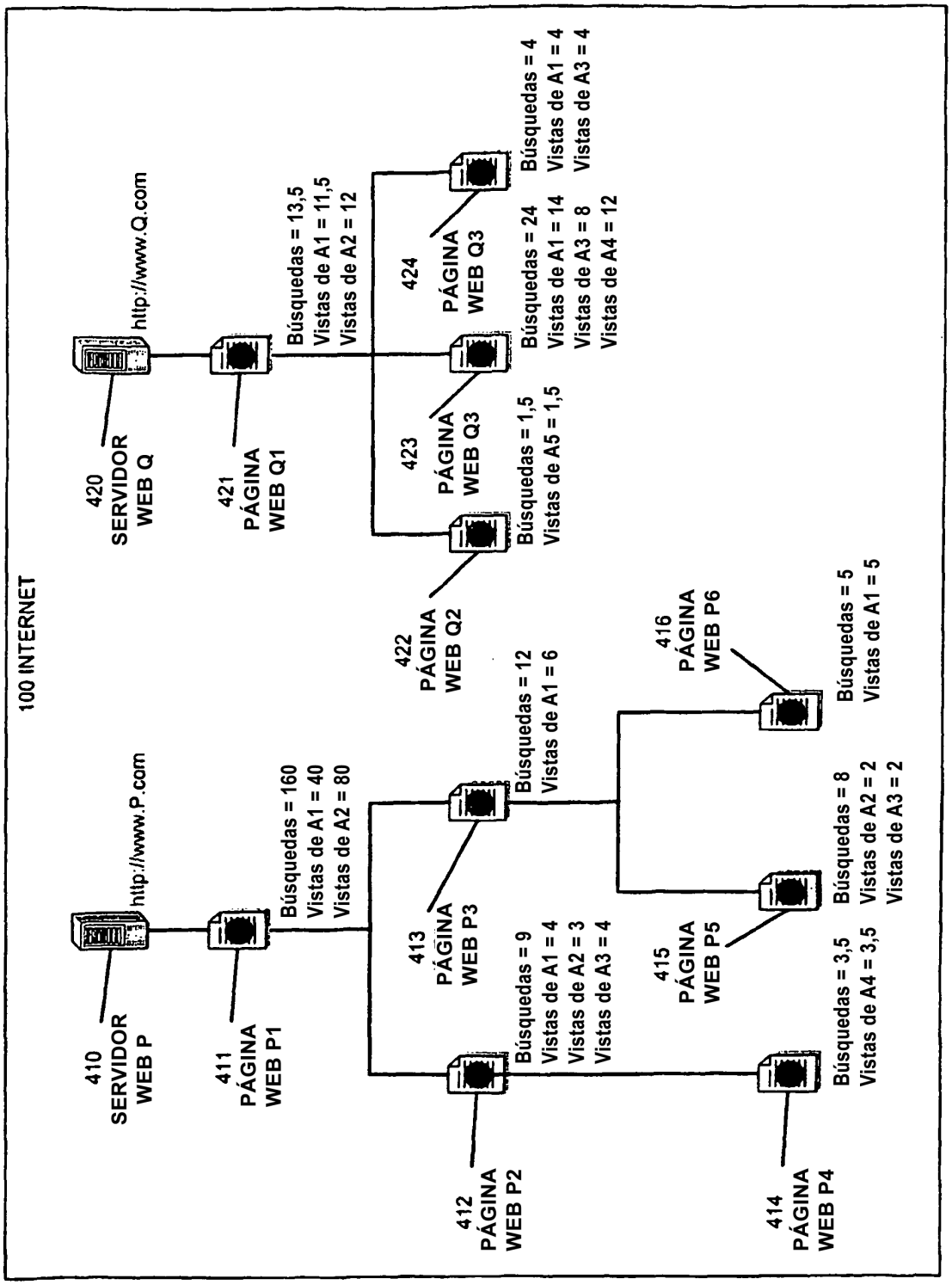
100 INTERNET
410 SERVIDOR WEB P
http://www.P.com
411 PÁGINA WEB P1
Búsquedas = 160
Vistas de A1 = 40
Vistas de A2 = 80
412 PÁGINA WEB P2
Búsquedas = 9
Vistas de A1 = 4
Vistas de A2 = 3
Vistas de A3 = 4
413 PÁGINA WEB P3
Búsquedas = 12
Vistas de A1 = 6
414 PÁGINA WEB P4
Búsquedas = 3,5
Vistas de A4 = 3,5
415 PÁGINA WEB P5
Búsquedas = 8
Vistas de A2 = 2
Vistas de A3 = 2
416 PÁGINA WEB P6
Búsquedas = 5
Vistas de A1 = 5
420 SERVIDOR WEB Q
http://www.Q.com
421 PÁGINA WEB Q1
Búsquedas = 13,5
Vistas de A1 = 11,5
Vistas de A2 = 12
422 PÁGINA WEB Q2
Búsquedas = 1,5
Vistas de A5 = 1,5
423 PÁGINA WEB Q3
Búsquedas = 24
Vistas de A1 = 14
Vistas de A3 = 8
Vistas de A4 = 12
424 PÁGINA WEB Q3
Búsquedas = 4
Vistas de A1 = 4
Vistas de A3 = 4

FIG. 4C

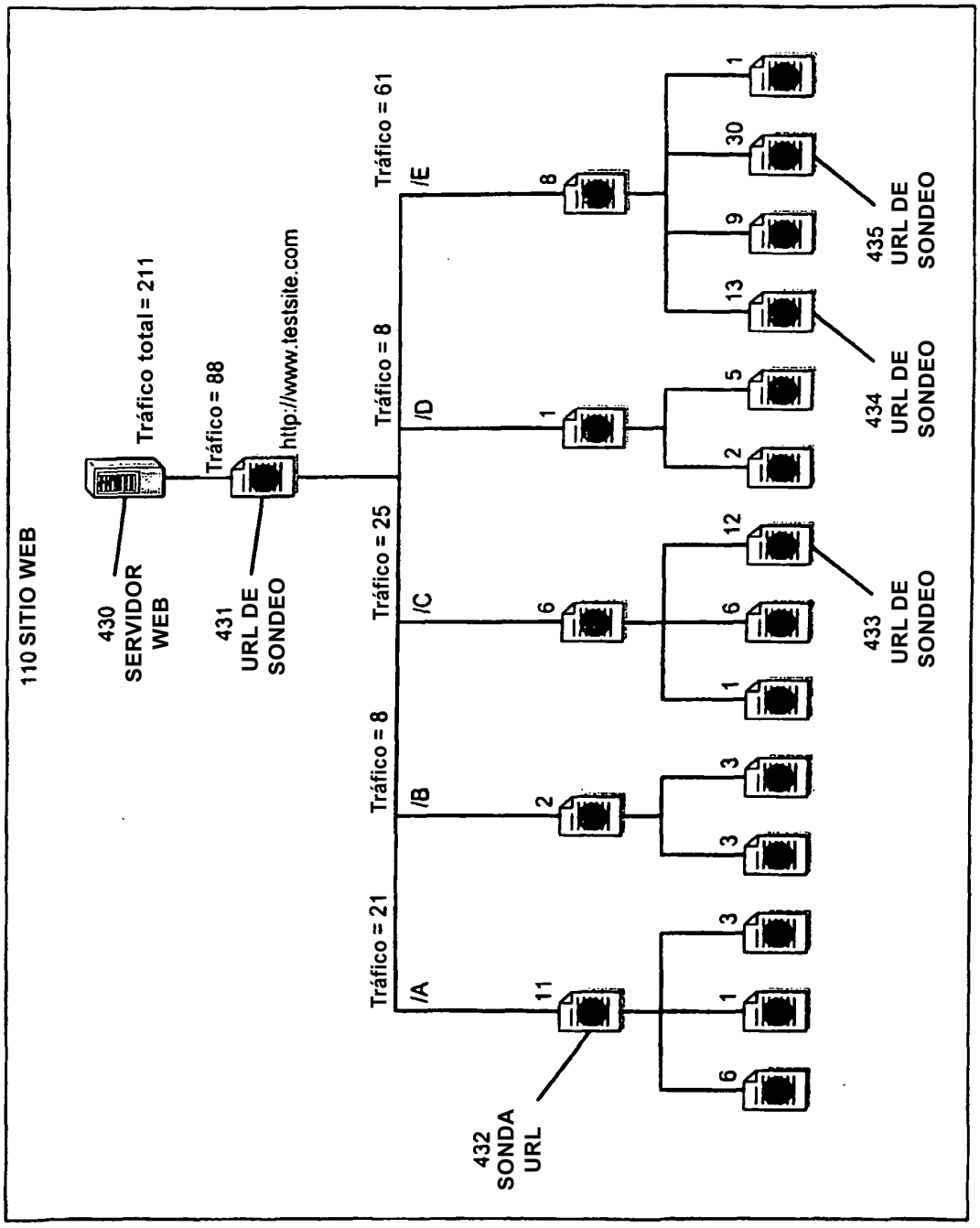
110 SITIO WEB
Tráfico total = 211
Tráfico = 88
http://www.testsite.com
430 SERVIDOR WEB
431 URL DE SONDEO
Tráfico = 21
/A
11
Tráfico = 8
/B
2
Tráfico = 25
/C
6
Tráfico = 8
/D
1
Tráfico = 61
/E
8
432 SONDA URL
435 URL DE SONDEO
434 URL DE SONDEO
433 URL DE SONDEO

FIG. 4D

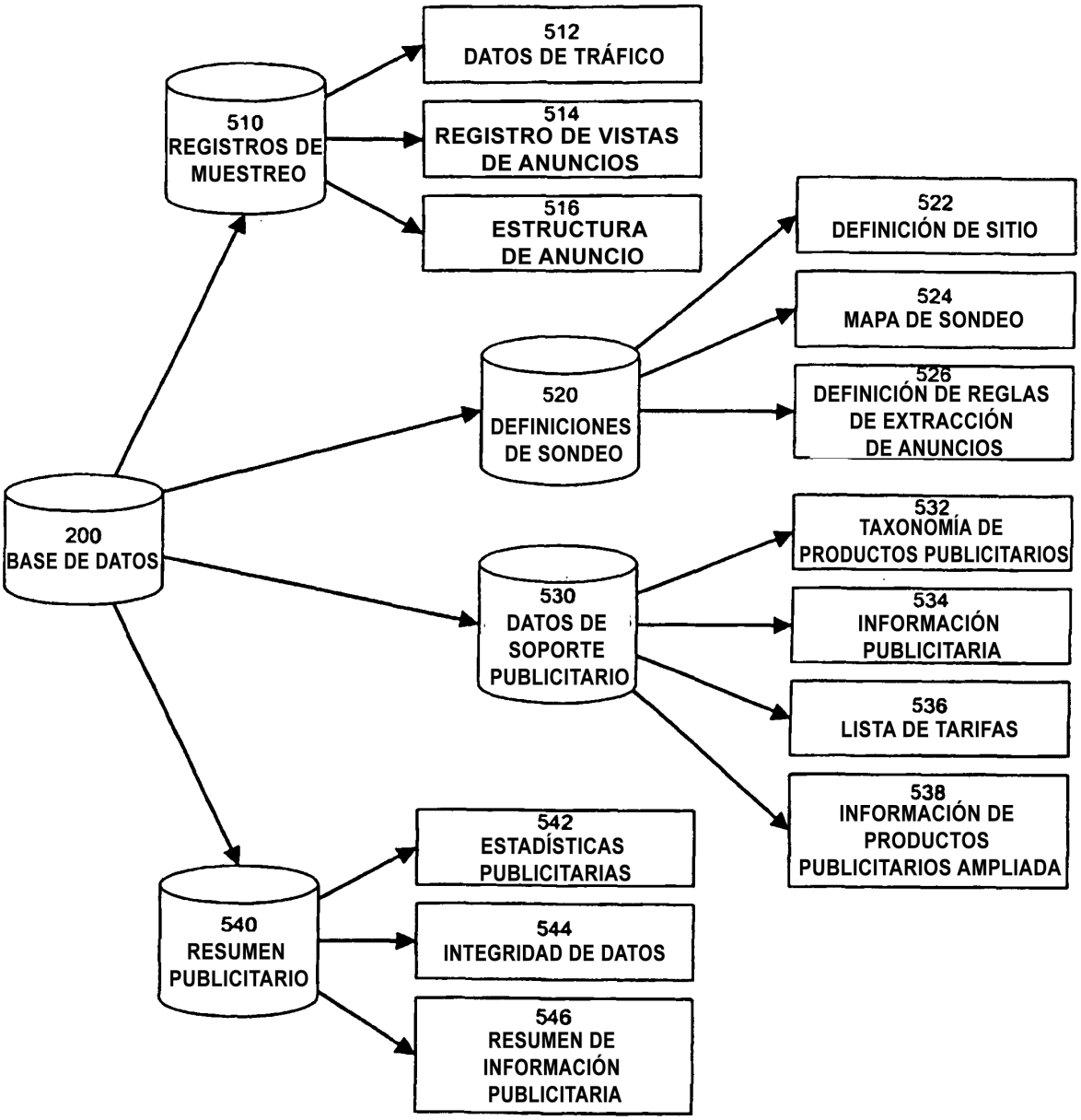
512
DATOS DE TRÁFICO
510
REGISTROS DE MUESTREO
514
REGISTRO DE VISTAS DE ANUNCIOS
516
ESTRUCTURA DE ANUNCIO
522
DEFINICIÓN DE SITIO
524
MAPA DE SONDEO
520
DEFINICIONES DE SONDEO
526
DEFINICIÓN DE REGLAS DE EXTRACCIÓN DE ANUNCIOS
200
BASE DE DATOS
532
TAXONOMÍA DE PRODUCTOS PUBLICITARIOS
530
DATOS DE SOPORTE PUBLICITARIO
534
INFORMACIÓN PUBLICITARIA
536
LISTA DE TARIFAS
538
INFORMACIÓN DE PRODUCTOS PUBLICITARIOS AMPLIADA
542
ESTADÍSTICAS PUBLICITARIAS
540
RESUMEN PUBLICITARIO
544
INTEGRIDAD DE DATOS
546
RESUMEN DE INFORMACIÓN PUBLICITARIA

FIG. 5

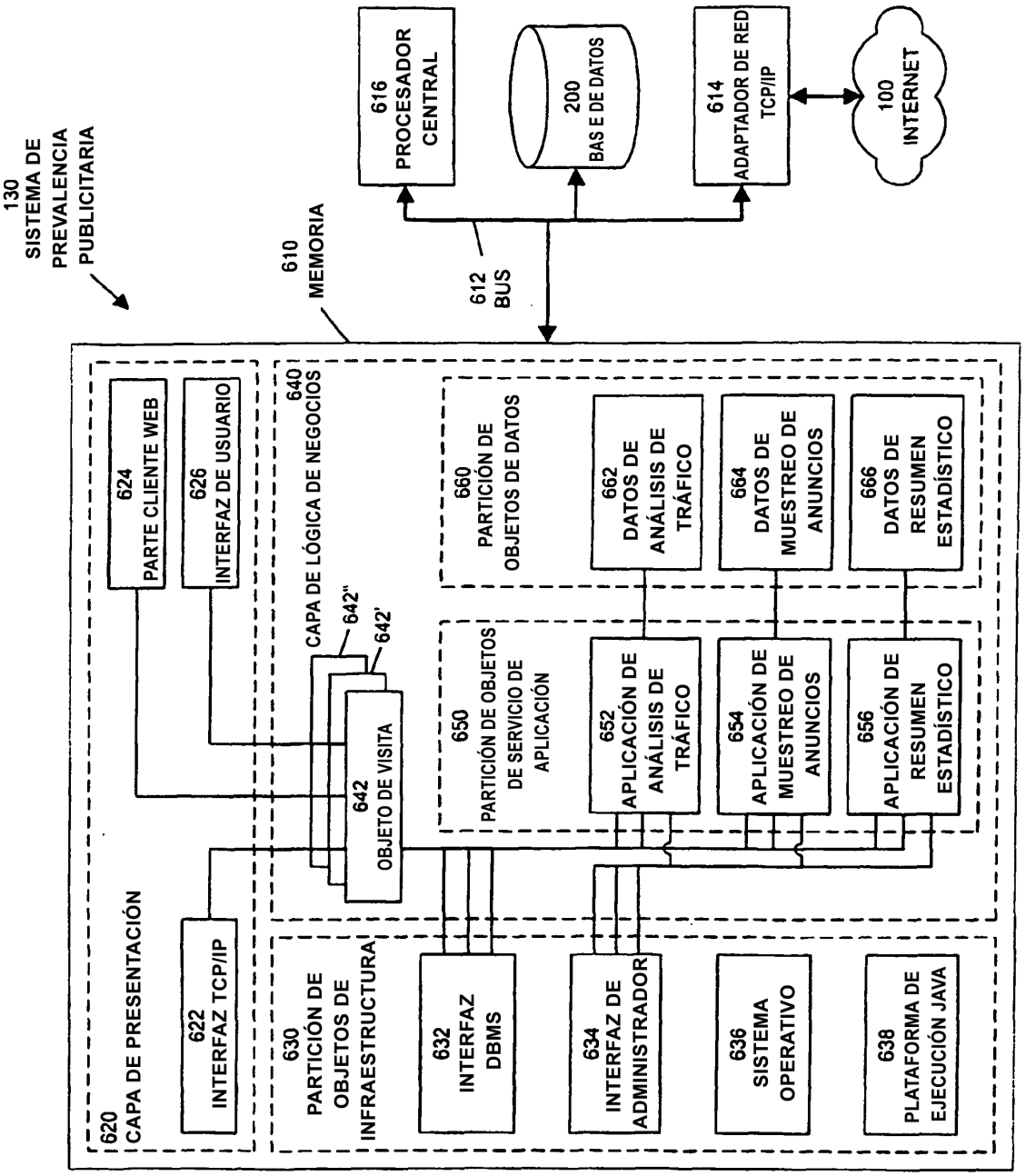
130
SISTEMA DE PREVALENCIA PUBLICITARIA
610
MEMORIA
612
BUS
616
PROCESADOR CENTRAL
200
BAS E DE DATOS
614
ADAPTADOR DE RED TCP/IP
100
INTERNET
620
CAPA DE PRESENTACIÓN
624
PARTE CLIENTE WEB
626
INTERFAZ DE USUARIO
622
INTERFAZ TCP/IP
640
CAPA DE LÓGICA DE NEGOCIOS
642"
642'
642
OBJETO DE VISITA
650
PARTICIÓN DE OBJETOS DE SERVICIO DE APLICACIÓN
652
APLICACIÓN DE ANÁLISIS DE TRÁFICO
654
APLICACIÓN DE MUESTREO DE ANUNCIOS
656
APLICACIÓN DE RESUMEN ESTADÍSTICO
660
PARTICIÓN DE OBJETOS DE DATOS
662
DATOS DE ANÁLISIS DE TRÁFICO
664
DATOS DE MUESTREO DE ANUNCIOS
666
DATOS DE RESUMEN ESTADÍSTICO
630
PARTICIÓN DE OBJETOS DE INFRAESTRUCTURA
632
INTERFAZ DBMS
634
INTERFAZ DE ADMINISTRADOR
636
SISTEMA OPERATIVO
638
PLATAFORMA DE EJECUCIÓN JAVA

FIG. 6

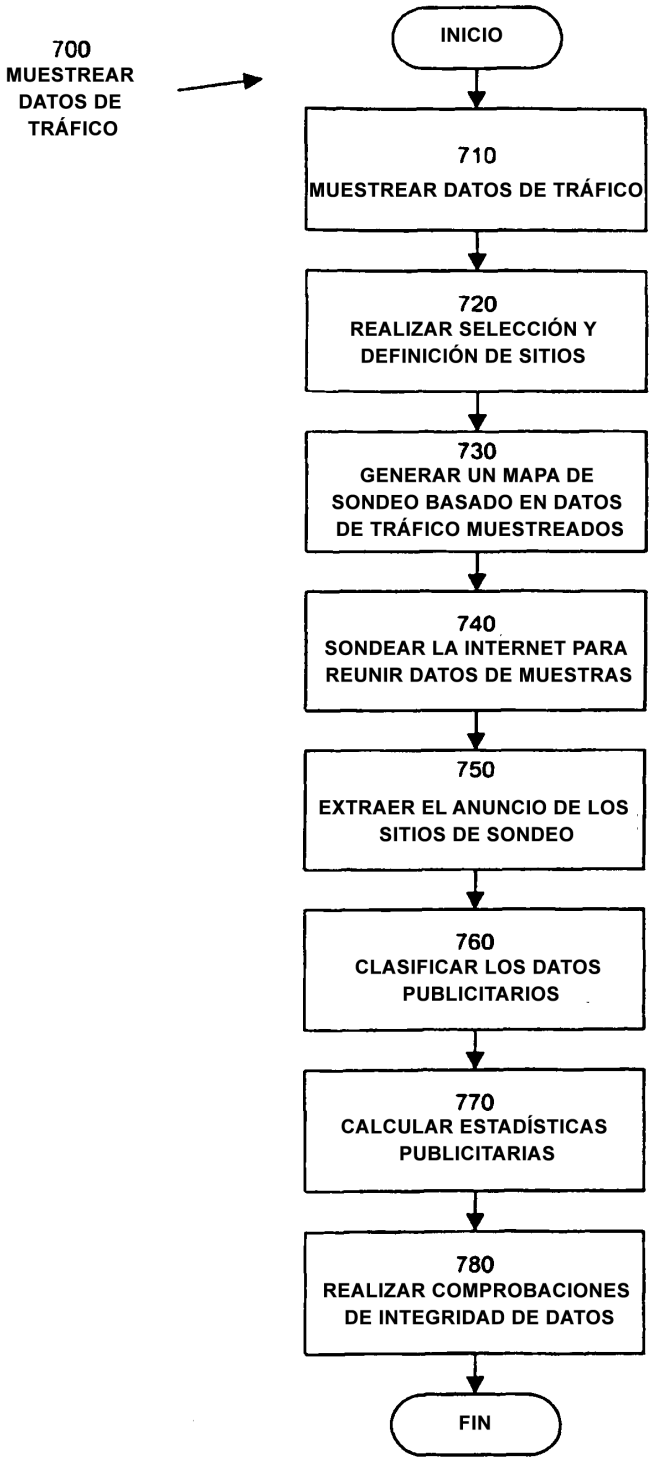
700
MUESTREAR DATOS DE TRÁFICO
INICIO
710
MUESTREAR DATOS DE TRÁFICO
720
REALIZAR SELECCIÓN Y DEFINICIÓN DE SITIOS
730
GENERAR UN MAPA DE SONDEO BASADO EN DATOS DE TRÁFICO MUESTREADOS
740
SONDEAR LA INTERNET PARA REUNIR DATOS DE MUESTRAS
750
EXTRAER EL ANUNCIO DE LOS SITIOS DE SONDEO
760
CLASIFICAR LOS DATOS PUBLICITARIOS
770
CALCULAR ESTADÍSTICAS PUBLICITARIAS
780
REALIZAR COMPROBACIONES DE INTEGRIDAD DE DATOS
FIN

FIG. 7A

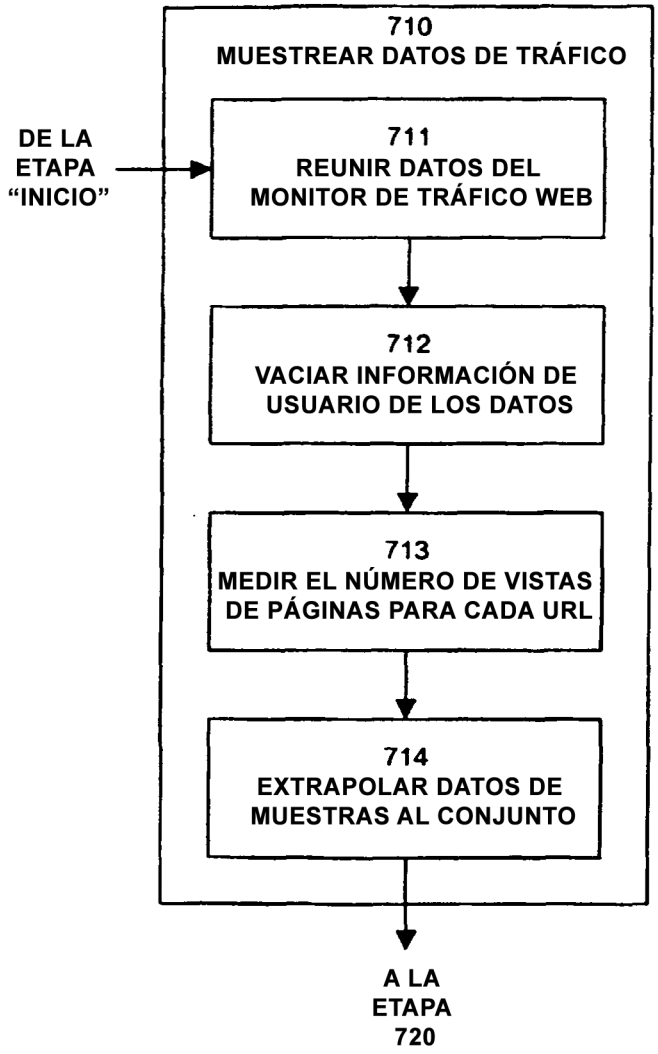
710
MUESTREAR DATOS DE TRÁFICO
DE LA ETAPA "INICIO"
711
REUNIR DATOS DEL MONITOR DE TRÁFICO WEB
712
VACIAR INFORMACIÓN DE USUARIO DE LOS DATOS
713
MEDIR EL NÚMERO DE VISTAS DE PÁGINAS PARA CADA URL
714
EXTRAPOLAR DATOS DE MUESTRAS AL CONJUNTO
A LA ETAPA 720

FIG. 7B

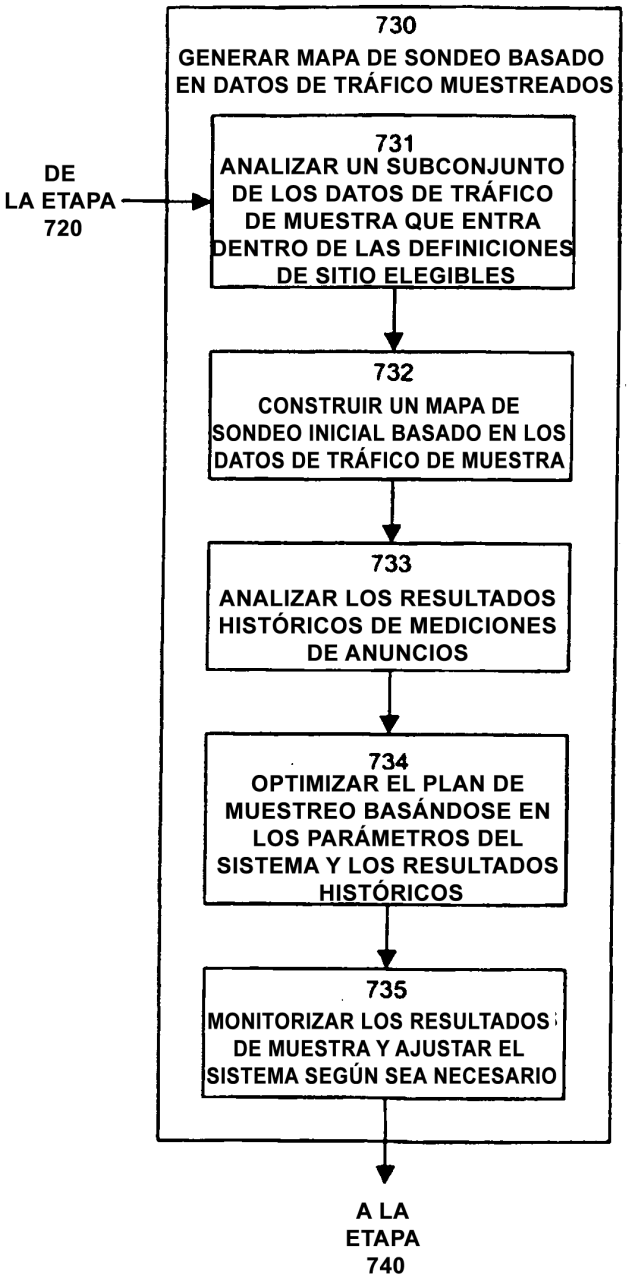
730
GENERAR MAPA DE SONDEO BASADO EN DATOS DE TRÁFICO MUESTREADOS
DE LA ETAPA 720
731
ANALIZAR UN SUBCONJUNTO DE LOS DATOS DE TRÁFICO DE MUESTRA QUE ENTRA DENTRO DE LAS DEFINICIONES DE SITIO ELEGIBLES
732
CONSTRUIR UN MAPA DE SONDEO INICIAL BASADO EN LOS DATOS DE TRÁFICO DE MUESTRA
733
ANALIZAR LOS RESULTADOS HISTÓRICOS DE MEDICIONES DE ANUNCIOS
734
OPTIMIZAR EL PLAN DE MUESTREO BASÁNDOSE EN LOS PARÁMETROS DEL SISTEMA Y LOS RESULTADOS HISTÓRICOS
735
MONITORIZAR LOS RESULTADOS DE MUESTRA Y AJUSTAR EL SISTEMA SEGÚN SEA NECESARIO
A LA ETAPA 740

FIG. 7C

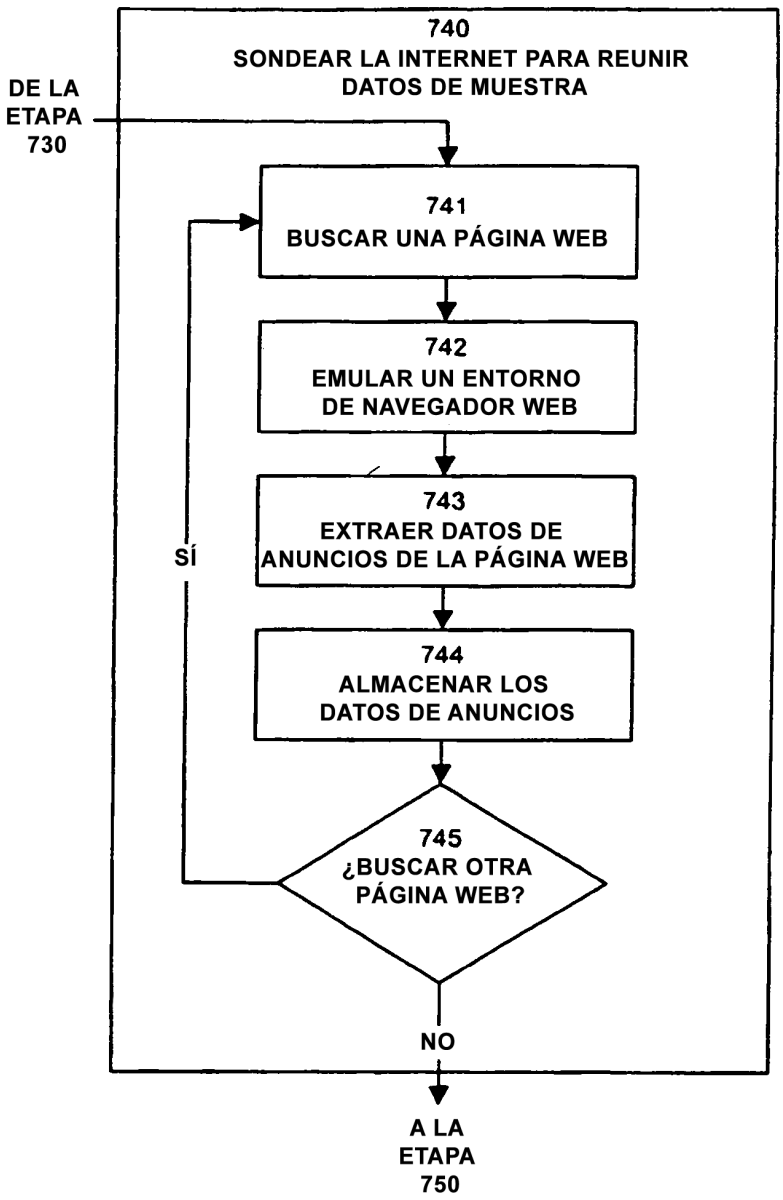
740
SONDEAR LA INTERNET PARA REUNIR DATOS DE MUESTRA
DE LA ETAPA 730
741
BUSCAR UNA PÁGINA WEB
742
EMULAR UN ENTORNO DE NAVEGADOR WEB
743
EXTRAER DATOS DE ANUNCIOS DE LA PÁGINA WEB
SÍ
744
ALMACENAR LOS DATOS DE ANUNCIOS
745
¿BUSCAR OTRA PÁGINA WEB?
NO
A LA ETAPA 750

FIG. 7D

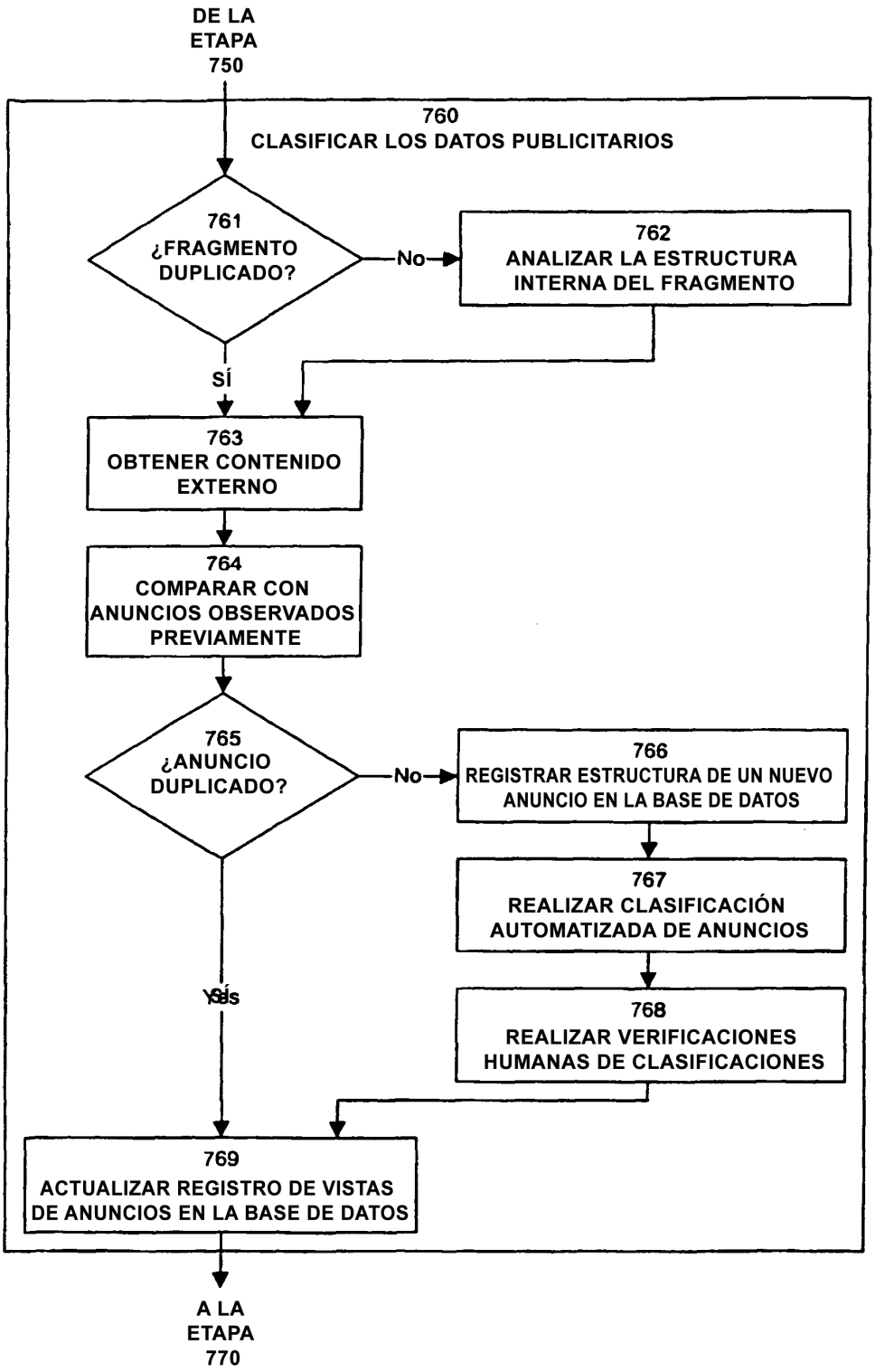
DE LA ETAPA 750
760
CLASIFICAR LOS DATOS PUBLICITARIOS
761 ¿FRAGMENTO DUPLICADO?
No
762 ANALIZAR LA ESTRUCTURA INTERNA DEL FRAGMENTO
SÍ
763 OBTENER CONTENIDO EXTERNO
764 COMPARAR CON ANUNCIOS OBSERVADOS PREVIAMENTE
765 ¿ANUNCIO DUPLICADO?
No
766 REGISTRAR ESTRUCTURA DE UN NUEVO ANUNCIO EN LA BASE DE DATOS
767 REALIZAR CLASIFICACIÓN AUTOMATIZADA DE ANUNCIOS
Sí
768 REALIZAR VERIFICACIONES HUMANAS DE CLASIFICACIONES
769 ACTUALIZAR REGISTRO DE VISTAS DE ANUNCIOS EN LA BASE DE DATOS
A LA ETAPA 770

FIG. 7E

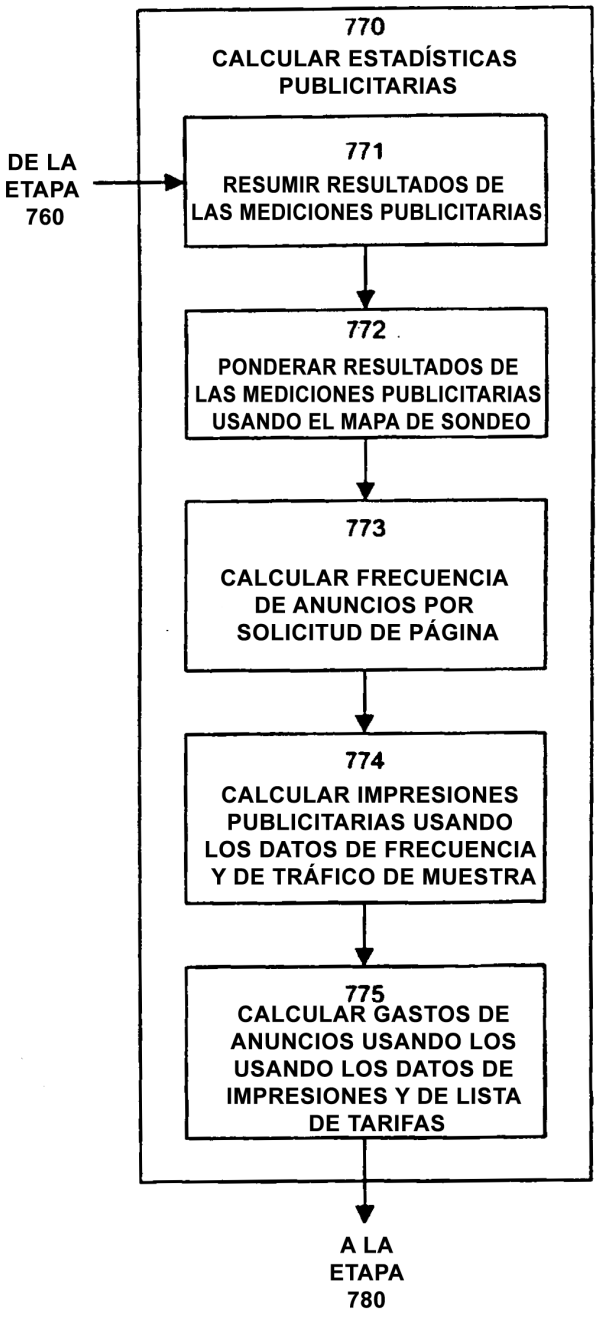
770
CALCULAR ESTADÍSTICAS PUBLICITARIAS
DE LA ETAPA 760
771
RESUMIR RESULTADOS DE LAS MEDICIONES PUBLICITARIAS
772
PONDERAR RESULTADOS DE LAS MEDICIONES PUBLICITARIAS USANDO EL MAPA DE SONDEO
773
CALCULAR FRECUENCIA DE ANUNCIOS POR SOLICITUD DE PÁGINA
774
CALCULAR IMPRESIONES PUBLICITARIAS USANDO LOS DATOS DE FRECUENCIA Y DE TRÁFICO DE MUESTRA
775
CALCULAR GASTOS DE ANUNCIOS USANDO LOS USANDO LOS DATOS DE IMPRESIONES Y DE LISTA DE TARIFAS
A LA ETAPA 780

FIG. 7F