Caracterización de servidores Web de ámbito académico

Tesis de Máster en Ingeniería de Computadores

Departamento de Informática de Sistemas y Computadores
Diciembre de 2008

Directores: Ana Pont Sanjuan y José Antonio Gil Salinas
Autor: Vicente José Castaño Díaz

Índice de contenido

1. Resumen

En este trabajo se presenta un estudio sobre la actividad realizada en los servidores dedicados a soportar sitios Web de perfil académico para poder dar servicio a los clientes, basándonos en los registros de actividad que mantienen estos servidores, con el objetivo de poder caracterizar la carga de trabajo que soportan.

Tomando como base estudios publicados anteriormente de caracterización de servidores como el realizado por Arlitt y Williamson [1] ampliamente aceptados y utilizados como referencia en otros estudios, se pretende establecer si los parámetros (invariantes) que se utilizan para caracterizarlos son aplicables en el entorno Web académico actual. Para ello se realiza el estudio de 2 conjuntos de datos con un gran volumen de información procedentes de servidores de diferentes características que soportan sitios Web del ámbito académico, ayudados para ello de herramientas software específicas presentes en el mercado, comprobando si es posible realizar este estudio haciendo uso de ellas.

2. Introducción y objetivos

La tasa de crecimiento de Internet es muy alta, en los últimos años se ha pasado de casi no existir tráfico WWW en diciembre de 1992 a existir en la actualidad un gran volumen de datos circulando por Internet, siendo cada vez más los sitios Web que ofrecen mayor cantidad de servicios para el usuario, y la tendencia es que este aumento se produzca de manera rápida siguiendo la llamada *Ley de Moore* [21]. A su vez, aparecen elementos nuevos que forman parte de las páginas que se visitan (vídeos, animaciones, etc.), los hábitos de comportamiento de los usuarios van cambiando y todo ello se refleja en el funcionamiento de los elementos que componen el ámbito de Internet. Además, surgen nuevos elementos en la red como los robots/arañas que han de tenerse en cuenta ya que aportan carga de trabajo adicional y no siguen los patrones de comportamiento de los usuarios.

El funcionamiento de la red de redes puede ser analizado partiendo de diversos enfoques: la naturaleza del tráfico existente en la red y su evolución [5], los patrones de acceso desde los clientes Web [13] y la caracterización de éstos [12], la caracterización de los usuarios [15] o del normal funcionamiento de un servidor Web [14], o la evolución de los sitios Web proponiendo métricas de estudio [4], siendo uno de ellos la *caracterización de la carga de trabajo* [1][2][7][16][19]. La caracterización supone determinar y describir las características fundamentales de la carga de trabajo en el tiempo. Este enfoque es utilizado como base para la evaluación de sistemas de computadores y por tanto puede ser aplicado para abordar el estudio de sistemas cliente/servidor como es la Web [6], analizando la Web en base a los

elementos que la componen desde diversas perspectivas, dependiendo del objetivo del estudio: proxies, navegadores o servidores.

El análisis de los servidores ha experimentado un gran avance en los últimos años, ya que un mayor conocimiento de la carga de los servidores permite anticiparse a la demanda de los usuarios, mejorar el funcionamiento del sistema e incluso analizar las preferencias de los usuarios. La caracterización de la carga de trabajo se centra en la distribución del tipo de ficheros, la distribución del tamaño de los ficheros, el comportamiento de los ficheros referenciados, y la distribución de las peticiones al servidor. A través de su estudio, se hace énfasis en encontrar características de la carga de trabajo que sean comunes a todos los conjuntos de datos bajo análisis.

Esta tesina está enfocada a realizar el estudio de la caracterización de la carga de trabajo que soportan los servidores Web, entendiendo dicha carga como la actividad que se produce para proporcionar los recursos que le son solicitados desde los clientes y se basa en el análisis de las medidas recogidas en el sistema durante el procesamiento de las peticiones.

Toda esta actividad soportada por el servidor se registra en unos ficheros (logs) donde de manera detallada queda impresa la información relativa al servicio de las peticiones que se realizan desde los clientes. Se registran datos de la procedencia de cada petición, del momento en el que se realiza, del recurso solicitado, la cantidad de datos transferidos en la transacción, así como el resultado de procesar dicha petición.

De esta manera podemos ver que el análisis de logs es una herramienta para el estudio del tráfico existente en servidores presentes en Internet y la interacción de sus usuarios, por lo que puede tomarse el enfoque tanto hacia la perspectiva del funcionamiento de los elementos que componen la red como hacia el comportamiento que tienen los usuarios en ella.

Se han realizado estudios anteriores con el objetivo de caracterizar los servidores Web, algunos de ellos han intentado establecer unos parámetros de referencia comunes (invariantes) que se cumplan para servidores Web de diferentes ámbitos y de esta manera puedan servir para caracterizarlos [1][19]. En base a estos estudios, se han hecho revisiones posteriores añadiendo nuevos invariantes que completen la caracterización de dichos servidores [3], y también otros que se centran en el estudio de un tipo específico de sitios Web [16] o en aspectos relacionados con las características de la estructura interna de los sitios Web [17].

Los estudios enfocados al estudio de los sitios Web se basan principalmente en los ficheros log que genera el servidor y éste es el método que ha sido utilizado para abordar la tesina que nos ocupa. A partir de estos ficheros se obtienen datos de la actividad llevada a cabo en el servidor como respuesta a las peticiones que le han sido realizadas, y con ellos se pueden cuantificar ciertos parámetros que permiten la caracterización del servidor.

En el presente estudio se han tomado los ficheros log de dos servidores que soportan sitios Web diferentes que hoy en día siguen en funcionamiento, uno perteneciente a la Universidad Católica de Valencia (http://www.ucv.es) y el otro a la Escuela de Informática Aplicada de la Universidad Politécnica de Valencia (http://www.eui.upv.es), ambos sitios

pertenecientes al ámbito académico. El período de tiempo en el que se produjo la actividad analizada está limitado entre las fechas del 1 de febrero de 2007 y el 1 de agosto del mismo año.

Para acometer el procesamiento de la información contenida en los ficheros log de ambos servidores, se ha realizado una búsqueda extensa de las herramientas software disponibles en el mercado, tanto gratuitas como de pago, con el objetivo de comprobar si con ellas es posible abordar un estudio de las características que se pretenden en este.

El objetivo fundamental de este trabajo es **analizar la carga de trabajo en servidores web de ámbito académico**, pero relacionados con éste surgen otros objetivos a su vez necesarios para alcanzar el objetivo principal:

- Comprobar si en la actualidad se cumplen los invariantes establecidos con anterioridad por otros autores en la caracterización de servidores Web de ámbito académico.
- Comprobar si es posible realizar un estudio de estas características haciendo uso de las herramientas software disponibles en la actualidad.

El documento se estructura de la siguiente manera: En primer lugar, se hace una revisión sobre estudios previos que sirven como base para la realización del análisis. A continuación se presentan los ficheros donde residen los datos y las herramientas capaces de su procesamiento. Posteriormente se describen los parámetros objeto de estudio, se citan las herramientas utilizadas y se exponen los resultados obtenidos tras el análisis de los datos. Finalmente se exponen las conclusiones obtenidas.

3. Trabajos relacionados

La caracterización de la carga de trabajo es la base para la evaluación del rendimiento de los sistemas de computadores. Ésta consiste en la descripción de la carga de trabajo mediante parámetros y funciones cuantitativas, con el objetivo de obtener un modelo capaz de mostrar, capturar y reproducir el comportamiento de la carga de trabajo y sus aspectos más importantes.

La caracterización de la carga de trabajo data de los años 70, y ha ido evolucionando siguiendo la evolución de las arquitecturas de computadores, desde los *mainframes* cuya carga de trabajo estaba compuesta básicamente de trabajos por lotes y transacciones hasta los sistemas actuales donde las redes de computadores y los sistemas compartidos han cambiado la visión de los usuarios hacia estos sistemas.

Los nuevos servicios proporcionados en Internet, como el World Wide Web, han introducido el concepto de la carga de trabajo multimedia, que consiste en una combinación

de diferentes tipos de aplicaciones (p.e. transferencia de ficheros, aplicaciones de audio y video en tiempo real) caracterizadas por diferentes requisitos de rendimiento de los recursos, tanto de servidores y clientes como de las redes.

En estos años, la caracterización de la carga de trabajo se ha ocupado de todos los nuevos dominios de aplicación, y las técnicas aplicadas para este propósito han evolucionado, en consecuencia, para hacer frente a la naturaleza de las cargas de trabajo, que se han convertido en más complejas. Para dar una idea de la trascendencia de esta tarea, mencionar que el diseño de los actuales sistemas y de los que van surgiendo están relacionados con una caracterización exacta de éstos. Es más, el diseño y evaluación de las políticas de gestión de recursos, como pueden ser las políticas de *caching* para los servidores WWW, requiere del conocimiento de las características y comportamiento de las peticiones a ser procesadas por el servidor.

En este sentido, se han realizado diversos estudios con el objetivo de caracterizar el funcionamiento de los servidores Web, así en [14] Joseph L. Hellerstein, perteneciente a IBM, junto a Fan Zhang y Perwez Shahabuddin, de la Universidad de Columbia, describen un enfoque estadístico para caracterizar el funcionamiento normal de un sistema durante un período de tiempo variando las cargas de trabajo en un servidor Web. Para ello, se consideran las influencias de la hora del día, el día de la semana, y el mes, así como el tiempo de correlaciones serie. Se aplica dicho enfoque a dos áreas de gestión de capacidad: *la previsión de la carga de trabajo y el problema de detección*. Se muestra que la previsión de la carga de trabajo puede realizarse a partir la extrapolación de los valores de las métricas, datos que pueden obtenerse aplicando un modelo estadístico que caracterice el funcionamiento normal del servidor. Así mismo, esa caracterización puede ser utilizada para filtrar efectos conocidos y permitir la detección de problemas.

Como estudio de referencia en la caracterización de la actividad soportada por servidores web puede considerarse el publicado por Arlitt y Williamson en 1996 [1]. A partir de la idea de obtener información a partir de los accesos realizados a un sitio Web alojado en un servidor, presentan un estudio detallado de *caracterización de la carga de trabajo en servidores Web* a partir de los ficheros log procedentes de seis servidores de diferente ámbito (tres de entornos académicos, dos de organizaciones de investigación científica y uno de un proveedor comercial de Internet) mediante el análisis de sus ficheros de registro (logs). A partir de esos logs, se identifican diez invariantes en la carga de trabajo de los servidores Web: *tasa de aciertos, tipos de fichero, tamaño de transferencia medio, disparidad de peticiones, referencias únicas, distribución de objetos por tamaño, concentración de referencias, tiempo entre referencias, localidad y distribución geográfica de las peticiones*. En el estudio [1] estos invariantes se consideran importantes ya que pueden presentarse como potencialmente ciertos para todos los servidores Web de Internet.

Además, se expone que las características de la carga de trabajo (invariantes) pueden utilizarse para identificar, a partir de los resultados obtenidos del número de ficheros referenciados una sola vez y de que no todos los conjuntos de datos presentan localidad temporal, dos estrategias posibles: para el diseño de sistemas de *caching* y así aumentar el

rendimiento de un servidor Web, y para determinar límites en la posible mejora de rendimiento con cada estrategia.

Un año después estos mismos autores presentaron un documento [19] donde ponen de manifiesto la importancia de la caracterización de la carga de trabajo y sus implicaciones en el rendimiento de los servidores, donde a partir del estudio anteriormente realizado [1], los invariantes anteriormente propuestos y las características observadas se discuten cuestiones de rendimiento y de *caching*, sugirieron mejoras prometedoras para los servidores Web.

En el estudio del rendimiento realizado en [19] se identificó como más efectiva la estrategia de hacer *caching* para reducir el número de peticiones que la de hacer *caching* para reducir los bytes transferidos. Además de que las características de la carga de trabajo observadas sugieren estrategias de gestión de cache, tales como el reemplazo por banda de frecuencia, que pueda mejorar las tasas de acceso a la cache y las tasas de acceso a byte para los servidores Web de Internet.

Diez años después de haberse realizado el primer análisis [1], Arlitt y Williamson revisaron en 2004 [2] el análisis publicado en 1996, abordando el mismo tipo de análisis en 3 de los 6 servidores utilizados en el estudio anterior para comprobar si los invariantes que habían propuesto se cumplían en las condiciones actuales, ya que el tráfico de Internet y el comportamiento de los usuarios había cambiado significativamente, principalmente debido al rápido aumento del volumen de tráfico en los últimos años. En estudios realizados previamente al de 2004 se había observado que los tamaños de las peticiones HTTP habían aumentado, mientras que los tamaños de las respuestas HTTP habían disminuido, aunque los tamaños de las respuestas HTTP más grandes observadas sí que habían aumentado, y además en otros estudios se observaban cambios en los patrones de acceso de los clientes Web. Por ello, en este nuevo estudio [2] se buscaba entender si las características subyacentes a la carga de trabajo de los servidores Web estaban cambiando o evolucionando mientras el volumen de tráfico crecía.

La principal conclusión a la que se llegó es que, a pesar de que el volumen de tráfico había aumentado más de 30 veces entre los años 1994 y 2004, no existían cambios sustanciales en las características de la carga de trabajo soportada por los servidores Web en los últimos 10 años. La mejora de los mecanismos de *caching* Web y otras tecnologías nuevas han cambiado algunas de las características de la carga de trabajo (p.e. porcentaje de peticiones con éxito) observadas en el estudio de 1996, y han influenciado en otras de manera sutil (p.e. tamaño promedio de los ficheros, tamaños de transferencia medios, y localidad temporal más débil). Sin embargo, la mayoría de los invariantes propuestos en 1996 siguen vigentes en este nuevo estudio, incluyendo el comportamiento de las referencias realizadas una sola vez, la alta concentración de referencias, las distribuciones *Heavy-tailed*[1] de ficheros según su tamaño, y el dominio de las peticiones remotas. Por último, se especula con la posibilidad de que los invariantes propuestos continúen cumpliéndose en el futuro debido a

[1] Larga cola: distribución donde una amplia o gran frecuencia es seguida por una baja frecuencia que disminuye gradualmente.

que representan características fundamentales del modo en que las personas organizan, almacenan y acceden a la información en la Web.

En esta misma dirección se publicó en el año 2006 un estudio [3] realizado por A. Faber, M. Gupta y C. Viecco, de la Universidad de Indiana, donde revisaban los invariantes propuestos por Arlitt y Williamson más de una década antes, aplicando el análisis en el contexto de sitios Web científicos. Partiendo de la idea de que el conocimiento de los invariantes de la Web es útil para la mejora del rendimiento y para la generación de carga de trabajo sintética de Web y que además pueden servir como una herramienta útil para detectar anomalías y usos indebidos, los objetivos propuestos fueron: 1) averiguar si los invariantes propuestos por Arlitt y Williamson se cumplían en la Web actual y en particular en el contexto de los sitios Web científicos, 2) analizar cómo habían evolucionado los invariantes en los sitios Web en general y los científicos en particular, y 3) investigar nuevos invariantes, tanto para sitios Web científicos como en general, que podrían ayudar en la detección de anomalías y usos indebidos.

Su conclusión fue que en diversos invariantes se observan diferencias significativas respecto al estudio presentado por Arlitt y Williamson en 1996 [1], por lo que deberían ser modificados para acoplarse a los diferentes tipos de cargas de trabajo y cambios en el tráfico de Internet en general, y especialmente para el contexto científico. Además, proponen tres nuevos invariantes con especial relevancia para los sitios Web científicos: *el tráfico durante la semana frente al fin de semana, el uso durante el día frente a la noche, y la concentración de clientes*.

Otros tipos de análisis se han producido en los últimos años, como el que se publicó en [11], donde se describe el análisis realizado en la empresa Hewlett Packard, ante la necesidad que manifestaban sus clientes, de los patrones de uso de los servidores Web y de qué manera dichos patrones van cambiando en el tiempo, así como los pasos que debían tomar para proporcionar respuesta adecuada a las peticiones recibidas en ese momento y en el futuro. En dicho documento se presentan los resultados de un estudio de una máquina servidora de un sitio Web de alta ocupación durante un período de dos meses. Durante este tiempo el tráfico del sitio se incrementó significativamente en términos de peticiones recibidas y de bytes servidos. En el estudio se procedió al análisis de los tipos de peticiones y respuestas, y se caracterizó la distribución del tráfico en base a la medida de las peticiones, el tiempo de respuesta, y otros factores (rendimiento de elementos http: HTML, imágenes y CGI). Se observa que el tráfico tiende a ser similar a otros sitios Web analizados en la literatura, y en particular se pudo confirmar que varios de los invariantes propuestos por Arlitt y Williamson [1] se cumplían.

Examinando el tráfico diario y semanal, se comprobó que los sistemas bajo una presión severa de peticiones tienen a tener un pronunciado *flat top*[2], un período durante el cual el máximo número de accesos o bytes por período de tiempo permanece coherente, lo que

[2] Cima plana: distribución donde los valores se sitúan en una escala muy elevada sin apenas variación.

puede indicar que bien el tráfico de peticiones o de respuestas excede el ancho de banda disponible de la red o bien que la tasa media de llegada en el servidor es mayor que la tasa de servicio promediada. Además, se constata la tendencia de que durante los fines de semana la carga de trabajo disminuye considerablemente y que el patrón de tráfico durante las horas del día es similar y coincidente con las horas de actividad laboral/académica.

La búsqueda de los patrones del tráfico que circula por Internet ha sido un objetivo perseguido en diversos estudios, introduciendo para ello nuevos conceptos o aplicando nuevas metodologías. Este es el caso del estudio abordado por Mark Crovella y Axer Bestavros [5] de la Universidad de Boston, donde aplican el concepto de *Autosimilitud* (*Self-Similarity)* para caracterizar el tráfico de un sitio Web. Se muestran evidencias de que el subconjunto de tráfico de red debido a transferencias de datos de Web puede mostrar características consistentes con *Self-similarity,* y se presenta una explicación hipotética para esa semejanza propia. Primero, se muestran evidencias de que el tráfico de Web muestra un comportamiento que es consistente con los modelos de tráfico de *Self-similarity*. A continuación, se muestra que esa similitud propia en dicho tráfico puede ser explicada en base a las distribuciones subyacentes de los tamaños de los documentos de la Web, de los efectos del *caching* y de la preferencia de los usuarios en la transferencia de ficheros, el efecto de usuario *thinking time*, y la superimposición de muchas de esas transferencias en una red de área local.

Para ello, en dicho estudio primero se considera la posibilidad de que haya *autosimilitud* del tráfico Web durante las horas de mayor ocupación que son medidas. Estos análisis soportan la noción de que el tráfico Web puede mostrar características de similitud propias, al menos cuando la demanda es suficientemente alta. En segundo lugar, haciendo uso del tráfico Web, las preferencias de usuario, y el tamaño de los ficheros de datos, se comentan las razones por las que los períodos de transmisión y los períodos tranquilos para cualquier sesión Web particular son *heavy-tailed*, lo cual es una característica esencial de los mecanismos propuestos para la *autosimilitud* de tráfico. En particular, se argumenta que muchas características del uso de la Web pueden ser modeladas usando distribuciones *heavy-tailed*, incluyendo la distribución de los períodos de transferencia, la distribución de las peticiones de usuarios para documentos, y la distribución subyacente de los tamaños de documentos disponibles en la Web.

El método de la caracterización ha sido aplicado en diversos ámbitos relacionados con las redes de computadores y sus elementos, siendo uno de ellos los servidores Web y los sitios soportados por ellos. Con el objetivo de obtener pautas indicativas del funcionamiento de un sitio Web en la literatura se han tomado diversos enfoques, uno de ellos es la *caracterización de los clientes,* utilizado en el estudio realizado por Balachander Khrisnamurthy y Craig Wills [12], donde a partir de la funcionalidad del servidor se caracteriza a los clientes. En él, se establecen categorías para grupos de clientes usando como referencia el nivel de conectividad con respecto a un servidor Web, tomando como base la información que puede ser determinada por el servidor. Los usuarios con conectividad “pobre” pueden elegir no permanecer en un sitio Web si se tarda mucho tiempo en recibir una página, incluso aunque el cuello de botella no esté provocado por el servidor Web del sitio. Se exploran varias consideraciones que podrían ser utilizadas por un servidor Web para caracterizar a un cliente: conectividad a la red, tiempo de respuesta, datos de la cabecera de la página (tipo de

contenido, versión del protocolo http), o el propio software del cliente. Una vez caracterizado el cliente por su conectividad como *pobre o rico*, el servidor puede entregar el contenido modificado, conocer como modificar el contenido, modificar políticas y decisiones de *caching*, o decidir cuándo redirigir al cliente a un sitio espejo.

Otro de los enfoques utilizados para establecer pautas de comportamiento del tráfico en Internet es a partir de su utilización por parte de los usuarios. Un ejemplo de ello es el estudio de Giancarlo Ruffo [15] donde se expone que la caracterización del comportamiento de los usuarios proporciona una interesante perspectiva para entender la carga de trabajo impuesta en un sitio Web y puede ser utilizada para dirigir aspectos como la distribución de la carga, el *caching* de contenido o la distribución y replicación de datos. Utiliza para ello la estrategia de *Web Mining*.

El *Web Mining* se refiere a la aplicación de técnicas en los repositorios de datos Web para realzar las capacidades analíticas de las herramientas estadísticas conocidas. En el proceso de realización de Web Mining se distinguen tres fases diferentes: *pre-procesado, descubrimiento de patrones y análisis de patrones*. Existen tres actividades diferentes de *Web Mining: Web structure mining, Web content mining y Web usage mining*. Este último consiste en la investigación de cómo accede la gente a páginas web y cuál es su comportamiento en la navegación, lo que supone el descubrimiento automático de patrones de acceso de los usuarios a partir de uno o más servidores Web. Los objetivos del *Web usage mining* pueden clasificarse en dos categorías: la dedicada a entender el comportamiento del usuario y personalización consecuente, y aquella dedicada a un progreso estructural de la efectividad del sitio Web.

La minería de uso de Web también puede proporcionar una nueva percepción del comportamiento del tráfico en la Web, permitiendo desarrollar políticas adecuadas para el *caching* y la distribución de contenido, la distribución de la carga, la transmisión en la red y la gestión de seguridad. La reconstrucción de los patrones de navegación de los usuarios es relevante tanto para personalizar el contenido presentado al usuario como para mejorar la estructura del sitio Web.

Otros estudios se han centrado en caracterizar la evolución de los sitios Web, intentando medir la tasa de cambios. Este es el caso del presentado por Ludmila Cherkasova y Magnus Karlsson [4], cuyo objetivo es el desarrollo de una herramienta de análisis de logs que produzca un perfil del sitio Web y del uso de los recursos de su sistema, de manera que sea útil para los proveedores de dicho servicio. Para ello, exponen que entender la naturaleza del tráfico del sitio Web es crucial para acometer un diseño apropiado de la infraestructura que soporte el sitio, especialmente para los sitios de tamaño y carga considerables, y con ese objetivo pretenden caracterizar la evolución de los sitios Web a partir de los patrones de la Web del momento, además de medir la tasa de cambios.

En el análisis realizado exponen las tendencias detectadas respecto a los resultados obtenidos por Arlitt y Williamson [1] en referencia a la existencia de una mayor concentración de las peticiones en un reducido porcentaje de ficheros, el aumento de contenido gráfico por página HTML, y el aumento en promedio del tamaño (bytes) de transferencia por petición.

En ese documento [4] principalmente se analizan tres cuestiones: los nuevos patrones de acceso de WWW, cómo caracterizar los cambios o evolución de los sitios Web, y cómo medir la tasa de cambios. Se evalúan, además, las características de conjunto de trabajo, definido como el tamaño (en bytes) combinado de todos los ficheros accedidos durante el período observado, y además se tiene en cuenta el número de bytes transferidos desde el sitio Web durante dicho período.

Pero la principal contribución del estudio es la propuesta de nuevas métricas. Se expone que hay dos factores principales que influyen en el funcionamiento de los servidores Web: *el número de peticiones que el servidor debe procesar y la cantidad bytes de respuesta* correspondientes que el servidor debe transferir (del disco o de la memoria, a la red). Se introduce el concepto de *núcleo*, definido como el conjunto de ficheros que reciben el 90% de los accesos del sitio Web, y cuyo comportamiento proporciona elementos para medir la evolución del sitio Web. A partir de ello, por un lado se establece una *localidad de conjunto de trabajo*, definido como el porcentaje del conjunto de trabajo que los ficheros más frecuentemente accedidos ocupan, obteniendo un porcentaje específico del número total de peticiones. Así mismo se propone una *localidad de bytes transferidos* para caracterizar la cantidad de bytes transferidos debido a las peticiones de los ficheros más frecuentemente accedidos.

Otra métrica importante expuesta en el citado estudio es la *calidad de servicio*, y se propone la cuantificación de las conexiones abortadas como posible manera de medirla. Dicho parámetro se establece mediante la diferencia tras comparar el tamaño del fichero transferido respecto del tamaño “percibido” (ver en una misma fila del log los tamaños de 2 ficheros iguales con el mismo tamaño). Estas métricas permiten una fácil comparación con respecto al conjunto de trabajo, identificando similitudes y diferencias en las cargas de trabajo de los servidores.

Dentro de las tendencias observadas en los últimos años en los sitios Web, una de las más importantes es la progresiva introducción de contenido dinámico, así como la personalización cada vez mayor de los sitos Web presentes en Internet. Por este motivo también han surgido estudios en relación a ello para determinar las características propias de éstos, como el realizado por W. Shi, R. Wright, E. Collins y V. Karamcheti, de la Universidad de Nueva York [16] en el que parten del hecho de que las peticiones de contenido dinámico y personalizado se han convertido en una importante parte del tráfico de Internet, pero sin embargo, las arquitecturas de *caching* tradicionales no son apropiadas para ese tipo de contenido.

En estudios previos al mencionado [16] se realiza el modelado de los objetos de las Webs dinámicas, pero es necesario complementarlo con una caracterización de la carga de trabajo de esos sitios. Dicha caracterización es relevante para servir a dos propósitos: primero, proporcionar pruebas para saber si se necesitan técnicas de composición de objetos, y si pueden resultar beneficiosas (dar el cliente específico y las características de contenido); y segundo, que esa caracterización puede dar nuevas ideas en la mejora del servicio de contenido dinámico y personalizado. En el estudio se verifica tanto la necesidad como el beneficio probable a partir del caching de contenido. Los beneficios sustanciales se pueden

obtener de aplicar técnicas de composición de objetos para contenido personalizado (p.e. reusar contenido). Tanto la carga de servidor como la latencia percibida por el cliente pueden ser reducidas mediante la aplicación de *prefetching (mover datos de la memoria a la cache para anticiparse ante futuros accesos para el procesamiento de datos)* del contenido. Dicha optimización puede ser lograda en la práctica con la identificación de un pequeño grupo de clientes debido a que el comportamiento de la popularidad de los clientes y personalización sigue una distribución Zipf[3]. Además, puede conseguirse que las latencias de las peticiones percibidas por el cliente sean más uniformes mediante la especialización de plantillas de documentos y contenidos, usando transcodificación, a la conexión de red empleada por el cliente.

Con el propósito de caracterizar los sitios Web con contenido dinámico, tres de los autores del anterior estudio mencionado, un año después publican un estudio [17] en el que se pretende ofrecer un conjunto de modelos que capturen las características propias de los contenidos dinámicos de las Webs tanto en términos de parámetros independientes tales como la distribución de los tamaños de los objetos y sus tiempos de "frescura", como también parámetros derivados tales como la reusabilidad de contenido a través del tiempo y los documentos enlazados. Las métricas de interés estudiadas fueron:

- *El número de objetos y su distribución de tamaños:* la primera indica la oportunidad de reutilización, mientras que la segunda determina si esa oportunidad puede ser o no utilizada contra las sobrecargas de gestión de objetos.

- *La reusabilidad de contenido:* cuanto mayor sea el contenido de reutilización, mayores serán los posibles beneficios de las técnicas de reutilización o de caché de los objetos individuales que la componen.

Los principales hallazgos encontrados son: por una parte, que los tamaños de los objetos de documentos dinámicos pueden ser modelados usando una distribución Exponencial o una distribución de Weibull[4]; por otro lado, que para los documentos que cambian durante el tiempo, sus tiempos de *frescura* pueden ser modelados también en términos de una distribución de Weibull, que a veces degenera en un caso donde todos los objetos cambian en cada acceso; y además, que existe una oportunidad significativa para la reutilización tanto para múltiples accesos del mismo documento como para accesos de documentos relacionados. Los resultados ponen también de relieve la dependencia entre el tamaño en el que un objeto es gestionado y el correspondiente potencial de reutilización.

A partir de la estrategia de la caracterización de los sitios Web ha quedado patente que se pueden tomar diversos enfoques, bien desde la perspectiva del contenido de la información proporcionada en el sitio (oficial, personal, educativa, etc) o desde la perspectiva del tipo de contenido ofrecido (estático, dinámico) para abordar el análisis de su carga de

[3] Ley de Zipf: ver Anexo

[4] La distribución de Weibull es una función de distribución continua que se utiliza en los análisis de fiabilidad, para establecer, por ejemplo, el período de vida de un componente hasta que presenta un fallo

trabajo. En [20] J. Almeida, J. Krueger, D. Eager y M. Vernon presentan un análisis extenso de las cargas de trabajo de servidores educativos en dos importantes universidades de Estados Unidos. Los objetivos del análisis incluyen proporcionar datos para la generación de cargas de trabajo sintéticas, obteniendo una nueva percepción dentro del diseño de las redes de distribución de contenidos de streaming y cuantificando qué cantidad de ancho de banda de servidor puede ser reservada en entornos educativos interactivos usando métodos de streaming multicast desarrollados para el contenido almacenado.

En los servidores analizados en [20], los tiempos entre peticiones recibidas representan una distribución heavy-tailed como Pareto[5], de la misma manera que en servidores con carga de trabajo de características tradicionales. Además, los períodos de acceso a ficheros de frecuencia estable pueden ser caracterizados mediante distribuciones del tipo Zipf, ya observado previamente en otros servidores Web con cargas de trabajo de medios streaming.

En cada servidor analizado, se detectó que una parte significativa de los ficheros nuevos accedidos cada hora nos son solicitados de nuevo hasta haber transcurrido más de 8 horas, lo que motiva la necesidad de reevaluar la estrategia de caching *cache-on-first-miss* para su uso con el contenido streaming. Dentro de cada sesión se observa que existe una significativa fracción de peticiones de corta duración (menos de 3 minutos de vídeo), igual que en otras cargas de trabajo. La distribución de la frecuencia de acceso a todos los ficheros multimedia, o a todos los ficheros multimedia con una duración determinada (p.e. menor de 5 minutos, o entre 50-55 minutos) puede ser aproximada para ser representada muy cercana a una distribución Zipf.

Otro aspecto importante resaltado en el estudio a tener en cuenta para aplicar el caching de ficheros multimedia es que la inserción de un nuevo fichero en la caché significa una sobrecarga de escritura en disco. Tras el análisis realizado se obtuvo que en muchas ocasiones un fichero multimedia no es solicitado hasta varias horas después de haberlo sido, por lo que la estrategia de caching *cache-on-first-access* no parece indicada y ha de ser reevaluada. Se comprueba además que los ficheros más populares, que representan un 90% de los streams multimedia, tienen una duración menor de 3 minutos y que la aplicación de técnicas de entrega multicast podrían reducir el ancho de banda utilizado para servirlos hasta un 40-60%.

En los últimos tiempos han ido ganando relevancia otros elementos presentes en Internet como son los motores de búsqueda en la Web, más conocidos como *robots, spiders o crawlers.* Estos elementos introducen una cantidad de tráfico adicional y no siguen las mismas pautas de comportamiento que los usuarios “normales”, por lo que merece ser objeto de estudio para analizar sus características propias y la influencia que puede tener en el

[5] Distribución de Pareto: ver Anexo

rendimiento de los servidores Web. Más cuando el número y la variedad de *robots* activos que operan en Internet aumentan continuamente, lo que puede resultar un impacto notable en el tráfico Web y en la actividad de los servidores Web. En [18] M. Dikaiakos, A. Stassolpoulou y L. Papageorgiou presentan un estudio de caracterización de motores de búsqueda en que analizan las peticiones de los *crawler* que deriva en la percepción de su comportamiento y estrategia. Se proponen una serie de métricas sencillas que describen características cualitativas de su comportamiento*: la preferencia de los crawlers en los recursos de un determinado formato, su frecuencia de visitas a un sitio Web, y la omnipresencia de sus visitas a un sitio determinado*.

La caracterización de la actividad de un crawler es importante ya que permite a los investigadores obtener información relevante para sus propósitos: 1) estimar el impacto de los robots en la carga de trabajo y rendimiento de los servidores Web, 2) investigar la contribución de los crawlers al tráfico en Internet, 3) descubrir y comparar las estrategias utilizadas por diferentes crawlers para recopilar recursos de la Web, y 4) modelar la actividad de los robots para producir cargas de trabajo sintéticas de crawlers para estudios de simulación, además de que la caracterización puede ser la base para la detección automática de los robots.

Para caracterizar el comportamiento de los crawlers objeto del estudio (Google, Altavista, Inktomi, FastSearch y CiteSeer) se examinaron diversos aspectos: 1) las características de tráfico http inducidas por los crawlers (distribución de peticiones http y códigos de respuesta), 2) aspectos de los recursos Web descubiertos (formato, tamaño y porcentaje sobre otros tipos de recursos), y 3) propiedades temporales que desvelan los momentos de las peticiones de los crawlers (tasa de llegada de las peticiones y distribución de los tiempos entre llegadas).

Los resultados obtenidos en [18] ponen de manifiesto que la actividad de un *crawler* tiene un impacto notable en la carga de trabajo de un servidor Web, aunque en otros estudios se mantiene lo contrario. Así mismo, se afirma que las peticiones GET inducidas por ellos son mucho mayores que la del resto de clientes, siendo también mayores sus peticiones a los recursos de texto y páginas HTML. Además, al contrario de lo que ocurre con los clientes "normales" sus peticiones no se concentran en un reducido número de recursos y los tamaños promedios de las respuestas ante peticiones de los *crawlers* son menores que con el resto de clientes. Por tanto, su actividad debe ser tenida en cuenta en el momento de realizar estudios relacionados con el tráfico web por el impacto que pueden tener en sobre él.

4. Los ficheros Log

Los estudios de caracterización de servidores Web, y más específicamente los enfocados a obtener una caracterización de sus cargas de trabajo, se realizan a partir del análisis de la información relativa al tráfico que soportan, cuyos detalles quedan registrados en sus ficheros Log correspondientes.

4.1. Concepto de Log

Un ***Log*** es un fichero en el que se realiza el registro de eventos durante un periodo de tiempo en particular, para registrar datos o información detallada sobre un evento.

La mayoría de los logs son almacenados siguiendo un formato estándar determinado, que es un conjunto de caracteres que tiene un significado explícito para dispositivos comunes y aplicaciones. De esta forma cada log generado por un dispositivo en particular puede ser leído y desplegado en otro diferente.

En el caso de los servidores Web, la información que se registra en estos ficheros detalla la actividad que se realiza sobre éste cada vez que se realiza una petición desde una máquina cliente a los objetos del servidor. Estos ficheros suelen crearse diariamente, teniendo almacenada la información de las peticiones realizadas al servidor distribuida por días y juegan un importante papel midiendo la sobrecarga del servidor y de la red, además de evaluar el funcionamiento del protocolo HTTP.

Un problema del análisis de logs es que no existe un estándar, con lo cual el formato depende del servidor instalado [8][9]. W3C intentó, con poco éxito, estandarizar el formato de los logs [10]. Además, la mayoría de los servidores Web permiten personalizar el formato de los logs añadiendo o eliminando parámetros nuevos.

4.2. Formato de los ficheros log

Existen diferentes formatos para el registro de los eventos en los servidores web, dependiendo principalmente del software que soporta el sitio Web, aunque la mayoría de los servidores Web registran al menos los siguientes campos:

- Fecha y hora de la petición
- Dirección IP del cliente
- Pagina solicitada
- Código http
- Bytes transferidos
- User Agent: información del usuario (aplicación, idioma, etc.)
- HTTP referer: información desde donde se accedió a la web.

Los formatos más conocidos son:

- **NCSA (Common or Access, o Combined)**

Los formatos de log NCSA están basados en NCSA httpd y están ampliamente aceptados entre los vendedores de servidores HTTP. El formato ofrece un alto grado de configuración, usándose formato de texto. Existen dos tipos de formato:

Formato Común:

Contiene solamente información básica de los accesos, del recurso solicitado y algunas partes de información, pero no contiene la referencia ni, el cliente.

Campos que componen el formato:

host rfc931 username date:time request statuscode bytes

Un ejemplo de registro usando este formato:

125.125.125.125 - dsmith [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043

Formato Combinado:

Este formato es una extensión del NCSA Común, contiene la misma información que el anterior y además añade 2 campos adicionales: la referencia y el cliente

host rfc931 username date:time request statuscode bytes referrer user_agent

Un ejemplo de formato combinado:

125.125.125.125 - dsmith [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043 "http://www.ibm.com/" "Mozilla/4.05 [en] (WinNT; I)" "USERID=CustomerA;IMPID=01234"

- **W3C Extended (used by Microsoft IIS 4.0 and 5.0)**

Este formato de ficheros log es utilizado en por Microsoft Internet Information Server. Los campos están separados por espacios en blanco, si alguno de ellos no se utiliza se registra el símbolo "-"como marca para omitirse. Los campos lo forman un prefijo y un identificador, separados por "-".

Campos:

date time c-ip cs-username s-ip cs-method cs-uri-stem cs-uri-query sc-status sc-bytes cs-bytes time-taken cs-version cs(User-Agent) cs(Cookie) cs(Referrer)

Ejemplo:

1998-11-19 22:48:39 206.175.82.5 - 208.201.133.173 GET /global/images/navlineboards.gif - 200 540 324 157 HTTP/1.0 Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+95) USERID=CustomerA;+IMPID=01234 http://yourturn.rollingstone.com/webx?98@@webx1.html

- **SunTM ONE Web Server (iPlanet)**

Cada archivo log en este formato contiene una secuencia de líneas que contienen caracteres ASCII. Cada línea puede contener una directiva o una entrada.

Las entradas consisten en una secuencia de campos relacionadas con una transacción http, separados por espacios en blanco. Si un campo no se utiliza en una entrada se registra un guion "-". Las directivas se sitúan al principio del log, definiendo la información y el formato contenido en las entradas.

Campos:

format=%Ses->client.ip% - %Req->vars.auth-user% [%SYSDATE%] "%Req->reqpb.clf-request%" %Req->srvhdrs.clf-status% %Req->srvhdrs.content-length%

Ejemplo:

10.101.128.66 - - [03/Apr/2002:00:42:11 -0500] "GET /multimedia/flash/dis.swf HTTP/1.1" 304

- **WebSphere Application Server (WAS) Logs**

Proporciona una API para uso de niveles de registro de eventos o *logging*. Esta API soporta tres métodos de logging: http, base de datos, y fichero. El formato de registro difiere ligeramente dependiendo del método utilizado, aunque cada registro contiene los siguientes 5 campos:

appid reqid httpdata cookie appdata

Ejemplo:

<custCareID> <01010123-1234> <-> <-> <target=TP600E;crosssells=TP770,TP380;ads=pIII>

5. Aplicaciones disponibles para el estudio

Actualmente existen gran cantidad de herramientas desarrolladas para realizar el análisis de actividad de un sitio Web, pudiéndose diferenciar principalmente por el modo en el que realizan la obtención de la información. Por un lado están las aplicaciones que han de estar implantadas en el propio servidor Web que da soporte al sitio Web para obtener información relativa al tráfico que se produce durante la actividad del sitio, y por tanto obtienen la información de manera instantánea. Por otro lado están las aplicaciones que procesan la información registrada por los propios servidores en los ficheros log correspondientes, por lo que dicho procesamiento se realiza posteriormente a haberse producido la actividad en el servidor y de forma externa a éste.

Debido al tipo de información de la que se dispone para proceder al análisis de la actividad de los servidores objeto de estudio, que no es otra que los ficheros log donde queda registrada la actividad de los sitios Web, serán las aplicaciones capaces de procesar dichos ficheros las utilizadas en el presente estudio.

Muchas de estas herramientas, en los dos tipos mencionados, además de servir para realizar el análisis de sitios Web y pueden ser utilizadas para el análisis de los accesos a servidores dedicados a proporcionar otros servicios: ftp, correo, etc., y están desarrolladas para poder ser utilizadas para el análisis de actividad en sitios Web soportados tanto por servidores basados en sistemas operativos Microsoft que utilizan para ello Internet Information Server (IIS) [14], así como por servidores basados en distribuciones Linux donde principalmente se utiliza Apache [15], aceptando el formato de los ficheros log generados por ambos sistemas.

Las aplicaciones ofrecen diversas opciones que permiten configurar ciertos parámetros, entre ellos se encuentran: especificar los tipos (extensiones) de ficheros para su clasificación como páginas (p.e. htm, html, php, asp, etc.), imágenes (p.e. jpg, bmp, png, gif, etc.), o ficheros de descarga (p.e. txt, arj, bin, dll, exe, etc.). También pueden establecerse filtros sobre diversos parámetros procedentes de los accesos registrados (p.e. tipo de navegador, sistema operativo, país de procedencia, etc.) y determinar la información que ha de mostrarse en los informes generados como consecuencia del procesado de los ficheros log (p.e. estadísticas generales, de actividad diaria, de navegadores utilizados, etc.).

Se pueden obtener gráficas e informes con los diversos parámetros que se pueden procesar a partir de los logs que además cumplan con las características especificadas en las opciones elegidas en la aplicación. Los parámetros recogidos por la mayoría de los analizadores son los siguientes:

Parámetros relativos a visitas y usuarios

- Número de visitas y número de visitantes únicos
- Duración de las visitas/sesiones y últimas visitas
- Páginas más/menos visitadas, accedidas y abandonadas
- Características del usuario:
    - Navegadores utilizados por el cliente
    - Sistemas operativos utilizados
    - Usuarios autentificados y las últimas visitas autentificadas: permiten
    - distinguir entre los usuarios internos y externos.
- Origen de las peticiones:

- Lista de hosts desde donde se ha accedido: esta es la información sobre desde donde se accedió a la web. Esta información es sensible, ya que contiene información de carácter sensible.
- Dominios y/o países de conexión de los visitantes
- Motores de búsqueda, palabras y frases usadas para encontrar el sitio web: esta información es de vital importancia si con el análisis lo que pretendemos extraer es información de cómo accedieron los usuarios a la web, con el objetivo de mejorar la visibilidad de la misma.
- Robots: este es uno de los parámetros más interesantes, ya que conviene descartar estos accesos, pues no pertenecen a usuarios reales.
- Rutas donde acceden las visitas: las rutas permiten estudiar la visibilidad de la web.

Parámetros relativos a tráfico

- Bytes transferidos durante una visita
- Tipos de ficheros accedidos
- Descargas de ficheros realizadas
- Puntos de entrada y salida de las visitas en la web
- Días de la semana y horas de accesos punta.
- Accesos por día, semana, mes
- Códigos de estado
- Errores http producidos en el sitio web en los intentos de acceso.

A continuación se presenta una relación de aplicaciones encontradas capaces de realizar el procesamiento de ficheros log procedentes de servidores Web y generar informes de datos a partir de ello.

Affinium NetInsight

- Descripción: aplicación comercial que analiza los datos acceso de servidores webs, principalmente para estudios de marketing. Permite estudiar Logs y Page Tags, con servidores corriendo en cualquier sistema operativo y servidor web.
- Url: http://netinsight.unica.com/

AlterWind Log Analyzer

- Descripción: aplicación comercial disponible tanto en versión Estándar como Profesional, enfocada a proporcionar datos para la optimización de los motores de búsqueda en el sitio web (SEO), la promoción del sitio y las aplicaciones de pago por click (PPC)[6]. Permite personalizar los informes resultantes del procesamiento de los Logs.
- Url: http://www.alterwind.com/

Analog

- Descripción: aplicación gratuita que analiza los datos acceso de servidores web. Permite analizar ficheros Log y admite diversos formatos, tanto de servidores Linux como Microsoft, y es muy configurable aunque no sea de manera trivial.
- Url: http://www.analog.cx/

AWStats

- Descripción: es una herramienta gratuita, GNU, que permite realizar análisis de servidores Web, Streaming, FTP o Email, mediante el análisis de archivos de logs. Esta aplicación está desarrollada en Perl. Es la aplicación gratuita con un mayor número de funcionalidades (16) y una de las más activas en sourceforge.
- Url: http://awstats.sourceforge.net/

ClickTracks

- Descripción: ClickTrack es una aplicación comercial destinada al estudio de los accesos a contenidos webs que obtuvo en 2006 el premio a la mejor aplicación en este ámbito, Product of the Year 2006, 2005, Best Web Analytics . Esta aplicación está diseñada para el análisis la visibilidad de la web en motores de búsqueda y otros parámetros relacionados con el marketing (como ROI, return of inversión in webs). Entre los idiomas soportados se incluye el español.
- Url: http://www.clicktracks.com/

[6] Método de marketing que consiste en colocar un banner en otro sitio web que enlace a tu sitio web. Se paga al propietario del otro sitio web por cada usuario que accede a tu Web desde ese sitio.

FastStats Log Analyzer

- Descripción: es una aplicación para el análisis de los patrones de tráfico en los accesos a la Web. La aplicación está diseñada para el estudio del comportamiento de los usuarios, potenciales clientes. Dicha aplicación es la más rápida de las existentes en el mercado, según datos del fabricante.
- Url: http://www.mach5.com/

Sawmill

- Descripción: es una aplicación comercial que funciona tanto para Pc como para Mac dirigida a proporcionar información a las empresas acerca del tráfico de la red para que sirvan de ayuda en la toma de decisiones en aspectos como pueden ser la seguridad y el crecimiento de la empresa. Permite su uso para el análisis de servidores web y de correo, así como de monitorización de seguridad gestión. Dispone de una interfaz en modo de página web de fácil manejo y escalabilidad.
- Url: http://www.sawmill.net/

WebAlizer

- Descripción: es una herramienta gratuita, GNU, que permite la realización de múltiples análisis de Logs. Además, incorpora multitud de lenguajes en los distintos informes que genera (entre ellos español). Un aspecto destacable de la aplicación es que el código está escrito en C y el procesamiento es bastante rápido (unos 70.000 registros por segundo según sus desarrolladores). Soporta varios formatos de Log: el de Apache, Common Logfile Format, diversas variaciones del NCSA Combined Logfile Format, y W3C Extended Log Format. No está adaptada para IIS de Microsoft.
- Url: http:// www.mrunix.net/webalizer/

WebLog Expert

- Descripción: es una aplicación comercial que proporciona información acerca de las visitas del sitio web disponible en tres versiones diferentes: *Reducida, Estándar y Profesional*. Acepta tanto ficheros Log procedentes de servidores Apache como de IIS, pudiendo leerlos incluso si están dentro de un fichero comprimido GZ o ZIP. Su interfaz es de uso sencillo e intuitivo y permite una configuración flexible de diversos parámetros.
- Url: http://www.weblogexpert.com/

Web Log Explorer

- Descripción: es una aplicación comercial que funciona bajo Windows que proporciona información acerca del tráfico de un sitio web. Dispone de una interfaz de fácil manejo desde donde permite la creación de informes en el momento de forma dinámica y por tanto ofrece gran flexibilidad en la obtención de información.
- Url: http://www.exacttrend.com/WebLogExplorer/

WebTrends Analytics

- Descripción: WebTrends es una aplicación enfocada a obtener información recogida en el servidor Web para descubrir las preferencias de los visitantes con el propósito de que permita a las empresas medir aspectos que ofrezcan datos de utilidad para mejorar su actividad y su relación con los clientes.
- Url: http://www.webtrends.com/

Dentro de las aplicaciones presentadas capaces de procesar los ficheros log y obtener información estadística a partir de ello, pueden diferenciarse dos grupos: *las aplicaciones comerciales y las aplicaciones gratuitas*.

5.1. Aplicaciones comerciales

Son herramientas que están disponibles para su uso gratuito durante un determinado período de tiempo, pasado el cual ha de procederse a su compra para poder seguir utilizándose. Están enfocadas a proporcionar información estadística de utilidad principalmente comercial, que pueda servir a las empresas poseedoras del sitio Web el disponer de datos adicionales con los que planificar su actividad.

Son bastante completas en cuanto a los parámetros que mide procedentes de los ficheros log de un sitio web, así como en el nivel de detalle que muestran en los informes relativos a estos parámetros, pero están dirigidas a ofrecer datos útiles desde la perspectiva empresarial. En los informes generados se muestran los datos correspondientes, el porcentaje de éstos y la representación gráfica según sea el caso, así como los totales. Además permite realizar filtrado de los datos de los ficheros log que se quieren tener en cuenta en los informes correspondientes para poder obtener una visión más acertada de los parámetros que más interesan cuando se realiza su análisis.

Dispone de un sencillo asistente de instalación que facilita esta tarea, además de otros asistentes para el uso de las diferentes funcionalidades que ofrece la aplicación. La interfaz de las herramientas es de fácil utilización y bastante intuitiva para realizar la configuración y personalización relativa a sitios Web, el filtrado de parámetros si se desea y la carga del/los fichero/s log. La precarga de ficheros log de un tamaño considerable (por ej. 180 Mb) suele resultar rápida y la interfaz permite una sencilla configuración de opciones. Pueden cargarse varios ficheros para proceder posteriormente al análisis simultáneo de todos ellos, en algunos casos incluso si se encuentran dentro de un fichero comprimido (p.e. zip, rar).

Los informes se representan en formato de páginas HTML y en muchos casos también en pdf (dependiendo de la herramienta), formatos prácticos para su uso a la vez que muestra de manera clara los datos tanto en tablas como en gráficas si se da el caso.

Ha de comentarse sin embargo que no se generan algunos datos importantes cuando se lleva a cabo el análisis de actividad de un sitio Web enfocado a medir su rendimiento, como por ejemplo los datos detallados de errores producidos ante peticiones y los tamaños de los ficheros transferidos.

5.2. Aplicaciones gratuitas

Son herramientas englobadas dentro de las denominadas de software libre, GNU, y por tanto gratuitas. Son aplicaciones que proporcionan información estadística general y también información más detallada de diversos parámetros registrados en los ficheros log, sin tener dicha información un enfoque determinado. Pueden funcionar tanto en sistemas operativos Microsoft como Linux y admiten diversos formatos de log.

Algunas pueden utilizarse tras realizarse instalación, pero en otras no se realiza una instalación propiamente dicha para poner en funcionamiento la aplicación en el ordenador, por tanto carece de asistente para ello. La implantación se realiza mediante la inclusión de los ficheros que componen la herramienta en una carpeta del disco duro, y en algunos casos se necesita de disponer de software específico instalado para su ejecución (p.e. Awstats necesita soporte Perl para su ejecución).

Tanto la ejecución como la carga de los ficheros log no son triviales, y han de aplicarse en modo texto a través de una serie de comandos en su fichero de configuración, tras lo cual se ha de ejecutar el fichero que lanza la aplicación. Permiten la configuración de diversas opciones y filtros, pero tampoco estas actividades son triviales, ya que se han de realizar mediante la modificación de diversos parámetros en determinados ficheros de texto incluidos en el directorio donde reside la aplicación.

Para la obtención de los correspondientes informes de los ficheros log cargados, ha de ejecutarse el proceso que inicia la generación del procesado de los log, y posteriormente ha de abrirse el correspondiente fichero de informe designado en la configuración (p.e. html). Los informes que se muestran son amplios y detallados, cuantificados en tablas y representados también en gráficas. Los datos que se muestran hacen referencia tanto a páginas como a

accesos, recursos accedidos, errores, aciertos, sistemas operativos, navegadores, rutas de acceso, etc. El formato de estos informes suele ser poco atractivo y práctico, y en algunas ocasiones (p.e. AwStats) puede ser favorecido mediante el uso de ciertas herramientas auxiliares.

5.3. Aplicaciones comerciales vs gratuitas

Las herramientas de análisis de logs en general no están enfocadas a proporcionar datos que analizar en el ámbito científico, pero generan una considerable cantidad de información relativa a las peticiones, visitas, ancho de banda, ficheros, etc. muy útil a este fin.

La herramienta comercial es más sencilla tanto de instalar, como de configurar y utilizar, y su interfaz facilita el manejo de la aplicación, al contrario que las herramientas gratuitas, donde ha de accederse a los ficheros de ayuda para comprender la mecánica de su funcionamiento. Además la aplicación de filtros es también más sencilla y completa en las primeras, ya que en las gratuitas ha de hacerse mediante la modificación de ciertos parámetros dentro de ficheros de texto. Por otra parte, algunas de las herramientas comerciales presentan limitaciones en el volumen de información a ser procesado de forma simultánea. En algunos casos esta limitación está impuesta en las versiones de prueba por la empresa propietaria, pero en otros casos dicha limitación, aunque elevada, está presente en las versiones adquiridas previo pago.

En lo referente a la generación de informes, lo más destacado es que las herramientas comerciales son más configurables en cuanto a los parámetros a obtener y la forma de presentarlos. Por otra parte ambos tipos de herramientas generan gran cantidad de informes cuantificados en tablas con alto nivel de detalle y teniendo en cuenta un amplio abanico de parámetros. La velocidad de procesamiento de los ficheros logs es mayor en el caso de las aplicaciones no comerciales, dato importante cuando es analizada gran cantidad de información.

Por tanto para usuarios poco expertos en el uso de aplicaciones sin instalación guiada ni interfaz convencional, es aconsejable el uso de la herramienta comercial ya que la otra puede resultarles muy incómoda de utilización, incluso imposible de conseguir que funcione correctamente. En caso de ser usuarios más experimentados en el uso de aplicaciones no comerciales, es más indiferente el uso de una u otra herramienta aunque resulta más cómodo y rápido el uso de las comerciales que el de las libres.

6. Parámetros objeto de análisis

Los logs ofrecen gran cantidad de información que puede interpretarse desde diferentes perspectivas dependiendo del objetivo que se persiga al realizar su análisis. En

nuestro caso, el objetivo está enfocado hacia el trabajo realizado por los servidores Web ante las peticiones de los clientes y a partir de ello caracterizar la carga de trabajo que soportan.

Existen estudios anteriores que a través del análisis de logs han pretendido caracterizar el comportamiento de los servidores mediante el establecimiento de parámetros ***invariantes*** que se cumplan en los servidores Web independientemente del ámbito en el que funcionen. Fruto de las investigaciones realizadas, a finales de los años 90, Arlitt y Williamson propusieron una serie de invariantes [1] que han sido los más ampliamente aceptados y utilizados por otros autores para llevar a cabo estudios relacionados con la materia.

El estudio que se describe en el presente documento se ha basado principalmente en revisar los invariantes establecidos por Arlitt y Williamson [1], comprobando además los invariantes adicionales propuestos por Faber, Gupta y Viecco [3], y revisando además las métricas propuestas por Cherkasova y Karlsson [4].

6.1. Invariantes de Arlitt y Williamson [1]

Estos autores en el año 1996 realizaron el estudio ***Web Server Workload Characterization: The Search for Invariants*** en el que establecieron una serie de características o parámetros que potencialmente se podrían cumplir en todos los servidores Web presentes en Internet. Los invariantes detectados fueron los siguientes:

Tasa de éxito

Este invariante, también llamado *Successful Request,* contiene información relativa a los accesos: duración, transferencia, códigos de respuesta, errores, etc. Se basa en la información de los códigos de estado generados durante el proceso de servir un recurso solicitado. Existen diversos estados, los más relevantes son: *éxito, no modificado, encontrado, no encontrado, y fallo.*

Estos parámetros permiten obtener una visión general del tipo de tráfico que tiene que atender el servidor, así como de su comportamiento con diferentes cargas de trabajo.

Tipos de fichero

El análisis de los tipos de ficheros trasferidos permite obtener una idea sobre el tipo de contenidos accedidos en el servidor. Analizando el tipo de documentos se puede tener una idea de la evolución del sitio Web y del tipo de ficheros que estructuran el sitio, si el contenido es más estático o dinámico, si consta de muchos elementos multimedia, etc.

Tamaño medio de transferencia

Este invariante hace referencia a la cantidad de datos transferidos en el proceso de servir la petición realizada. Con ello se obtiene el volumen de la carga que puede estar soportando el servidor. El análisis por tipo de documento posibilita conocer cuales de ellos son los que provocan mayor carga de trabajo.

Disparidad de las peticiones

Se trata de obtener una medida de las peticiones a ficheros distintos en relación con el total. Si las peticiones están muy concentradas los ficheros mas solicitados por los clientes deberían ser *cacheados* en el servidor para reducir la latencia del cliente.

Referencias únicas

Estas referencias son ficheros que en la traza de logs analizados han sido accedidos solamente una vez. Este parámetro puede dar un indicativo de la conveniencia o no de realizar *caching*, ya que si la cantidad es importante, no es oportuno hacer caching de esos ficheros porque ralentizaría el tiempo de servicio.

Distribución de los objetos dependiendo de su tamaño

Hace mención a los tamaños de los ficheros accedidos, siendo agrupados por dicho parámetro. Se establece que un conjunto de ficheros comprendidos en un rango de tamaño específico acumulan la mayor parte de las peticiones que procesa el servidor, siguiendo en general una distribución de Pareto.

Concentración de referencias

Determina si la distribución de peticiones a los ficheros accedidos es uniforme o no. Se trata de obtener una clasificación de los ficheros accedidos, agrupándolos por el número de de veces que se ha producido una petición de este con respecto al total, en otras palabras, se mide la popularidad de los objetos del servidor.

Tiempo medio transcurrido entre referencias

Este invariante estudia el tiempo entre diversas peticiones de una misma referencia. Con ello puede observarse la frecuencia con la que es pedido un objeto o la probabilidad de haber sido modificado, ya que cuanto mayor sea ese tiempo más probable es que el contenido haya sido modificado y con ello se tenga que refrescar la caché del proxy. En el estudio se estableció que el tiempo entre referencias esta distribuido exponencialmente e independiente entre servidores.

Localidad

Establece la proporción entre las peticiones realizadas dentro de la propia red local y el resto de peticiones realizadas desde el exterior. Para ello es necesario tener conocimientos de las ip's de la red local y de esta manera poder filtrarlas.

Distribución geográfica

El objetivo de este invariante es determinar el origen geográfico de las peticiones realizadas al servidor, entendiendo como tal el país/estado desde la que se producen, además se toman en cuenta los dominios des procedencia de las peticiones. Se comprueba si existe concentración de peticiones desde un país o desde un reducido grupo de dominios.

6.2. Invariantes de Faber, Gupta y Viecco [3]

En el año 2006 estos autores realizaron una revisión de los invariantes propuestos por Arlitt y Williamson en el documento ***Revisiting Web Server Workload Invariants in the Context of Scientific Web Sites*** con el objetivo de comprobar si seguían vigentes en la actualidad en su aplicación a los sitios web científicos. Además propusieron tres nuevos invariantes que completaran a los anteriores para llevar a cabo la caracterización de servidores. Los invariantes propuestos fueron:

Acceso en los días de continuo frente a los fines de semana.

Trata de determinar si existen diferencias de utilización de la Web dentro de una semana entre los días que van de lunes a viernes y los fines de semana. Se establece que las diferencias son significativas, siendo menor los fines de semana.

Uso diurno de la Web.

Relaciona los accesos realizados a la Web con la hora del día en la que se realizan por parte del cliente. Se trata de determinar si existe diferencia de uso entre las horas establecidas como diurnas y las nocturnas. Es fundamental por tanto en el caso de peticiones procedentes de otros países saber en que franja horaria tiene cabida la hora registrada.

La separación de estas franjas horarias resulta arbitraria, intentando que coincidan de la forma más general y fiel posible con la separación de horas que corresponden a la jornada laboral/académica y las que no, por lo que ni siquiera puede establecerse un número de horas fijo para cada una de las franjas.

Concentración de clientes.

Establece la relación entre las solicitudes realizadas y la cantidad de clientes que las realizan, determinando si existe una distribución uniforme o por el contrario existe una concentración de peticiones entre un reducido porcentaje de clientes.

6.3. Invariantes de Cherkasova y Karlsson [4]

En el año 2001, Luzmila Cherkasova y Magnus Karlsson revisaron los invariantes propuestos por Arlitt y Williamson en el documento ***Dynamics and Evolution of Web Sites: Analysis, Metrics and Design Issues***, mostrando varias tendencias nuevas en la carga de trabajo soportada por los servidores Web. Con el objetivo de realizar una caracterización de las dinámicas y evolución de los sitios Web y medir la tasa de cambios se proponen unas nuevas métricas centradas en la aplicación de localidad de los ficheros accedidos. A partir de ello, exponen que partiendo de un conjunto de ficheros que son los que reciben las peticiones durante un determinado periodo de tiempo, pueden establecerse unas líneas de tendencia en la actividad del servidor.

Cambios de las estadísticas absolutas en el tiempo

Se expone el hecho de que el volumen de tráfico del servidor Web, medido en cantidad de peticiones, puede ser bastante predecible para algunos sitios cuando ha sido medido durante el tiempo suficiente. El tiempo en cuestión es relativo y no se expone la manera de estimarlo. Se puede realizar la comparación entre semanas, días u horas.

Las dinámicas de los ficheros nuevos

Es importante saber si los ficheros del servidor de un sitio Web serán replicados en otros muchos sitios, descargando de esta manera el trabajo a la localización original. Si el volumen de ficheros nuevos es alto, el rendimiento del sistema se vería considerablemente afectado, por ello se pretende conocer la tendencia que sigue el volumen de ficheros nuevos. Por ello, conocer si los nuevos ficheros son populares es muy importante, ya que deberían ser replicados para mejorar los accesos a ellos.

El núcleo

Se define como el conjunto de los ficheros más accedidos que representan el 90%, o incluso más, de las peticiones que se reciben. Partiendo del hecho de que un pequeño conjunto de ficheros (2-4%) concentran la gran mayoría de las peticiones (90%), se expone que la frecuente variación de este *núcleo* puede suponer un problema en el rendimiento del sistema por lo que ha de observarse el porcentaje de cambios de ficheros nuevos por día o por semana. Lo que no se muestra es el tiempo de permanencia de un fichero en el *núcleo*, o si un fichero entra y sale frecuentemente.

7. Caso de estudio

Para realizar el estudio se ha podido disponer de los ficheros Log correspondientes a dos servidores del ámbito académico, concretamente pertenecientes a dos universidades situadas en la ciudad de Valencia con el objetivo de caracterizar sus cargas de trabajo en base a los invariantes de la literatura.

7.1. Contexto

Los conjuntos de datos utilizados para la realización del presente estudio provienen de dos servidores Web actualmente en uso ubicados en centros universitarios diferentes en Valencia. Uno de ellos pertenece a la Universidad Católica de Valencia San Vicente Mártir, y es el servidor que soporta todo el sitio Web (http://www.ucv.es) sobre un sistema operativo Windows con Internet Information Server (IIS). El otro conjunto de datos pertenece a la Escuela Técnica Superior de Informática Aplicada de la UPV, soportando su sitio web (http://www.eui.upv.es) sobre un sistema operativo Linux con Apache Web Server.

7.2. Conjuntos de datos utilizados

Los logs utilizados pertenecen a diferentes periodos de tiempo dependiendo del servidor, teniendo en común la fecha de inicio de dichos periodos. Debido a que el volumen de peticiones mensuales es significativamente mayor en el caso de la UCV, se decide aumentar el periodo a estudiar en el caso de la EUI. De esta manera los datos recogidos en la UCV están entre el 1 de febrero de 2007 y el 30 de abril del mismo a; o, en el caso de la EUI, el periodo analizado se sitúa entre el 1 de febrero y el 30 de junio de 2007.

7.3. Tipos de ficheros log utilizados

Para llevar a cabo el estudio se ha realizado el análisis de 2 sitios web diferentes, siendo cada uno de estos sitios soportado por servidores Web con diferentes características software. Uno de estos sitios está funcionando sobre un servidor con IIS (Internet Information Software), mientras que el otro está operando sobre un servidor Apache. Debido a esto, presentan diferente formato de ficheros log, el primero de ellos es de formato W3C y el segundo NCSA-Apache. A continuación se hace una descripción del formato de los logs utilizados en ambos casos.

1. Servidor web de la Universidad Católica de Valencia (http://www.ucv.es)

La web de esta soportada por un servidor con IIS. Como se ha mencionado anteriormente, los datos relativos al ancho de banda consumido no han sido registrados en todos los logs del periodo analizado.

IIS – W3C Extended Log File Format

#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie) cs(Referer) cs-host sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken

http://www.w3.org/TR/WD-logfile.html

Ejemplo**:**

2007-04-05 23:59:59 W3SVC1 ISANLO 10.58.1.6 GET /ban1_oct2006.swf - 80 - 10.58.1.1 HTTP/1.1
Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322;+InfoPath.2;+.NET+CLR+2.0.50727)
__utmc=239886087;+__utma=239886087.1133225009.1172062292.1175685445.1175687864.49;+__ut

mz=239886087.1172062292.1.1.utmccn=(direct)|utmcsr=(direct)|utmcmd=(none);+ASPSESSIONIDCQTBCQAA=IGHEBIJDAKBNEDGKGMKCHJPA;+ASP.NET_SessionId=mg2ep0bmbfg54p45lu1wkazc;+ASPSESSIONIDAQTCCRBA=AJMJEEGABOMGDEKPOHIECFGC http://www.ucv.es/ban1_oct2006.swf www.ucv.es 304 0 0 92 714 203

Campos que componen el formato

- ***date time***: 2007-04-05 23:59:59

Momento en el que se produce el acceso al sitio web desde el cliente.

- **s-sitename**: W3SVC1

Carpeta donde se almacenan los ficheros log en el servidor

- **s-computername**: ISANLO

Nombre del servidor web donde se aloja el sitio web

- **s-ip:** 10.58.1.6

Dirección de red del servidor web

- **cs-method**: GET

Método utilizado para la petición del objeto por parte del cliente

- ***cs-uri-stem***: /ban1_oct2006.swf

Objeto (fichero) solicitado por el cliente

- **cs-uri-query**: -

Parámetros adicionales en la solicitud del objeto por parte del cliente

- **s-port**: 80

Puerto del servidor por el que se reciben y sirven las solicitudes.

- **cs-username**: -

Usuario autentificado que realiza la petición del objeto

- ***c-ip***: 10.58.1.1

Dirección ip de la máquina que realiza la petición (ip que se recibe del cliente)

- ***cs-version***: HTTP/1.1

Versión del protocolo utilizado por el cliente

- ***cs(User-Agent)***: Mozilla/4.0+(compatible;+MSIE+7.0;+Windows+NT+5.1;+.NET+CLR+1.1.4322;+InfoPath.2;+.NET+CLR+2.0.50727)

Agente utilizado por el cliente en la petición del objeto.

- ***cs(Cookie)***: __utmc=239886087;+__utma=239886087.1133225009.1172062292.1175685445.1175687864.49;+__utmz=239886087.1172062292.1.1.utmccn=(direct)|utmcsr=(direct)|utmcmd=(none);+ASPSESSIONIDCQTBCQAA=IGHEBIJDAKBNEDGKGMKCHJPA;+ASP.NET_SessionId=mg2ep0bmbfg54p45lu1wkazc;+ASPSESSIONIDAQTCCRBA=AJMJEEGABOMGDEKPOHIECFGC

Información de la cookie generada en la petición del objeto

- ***cs(Referer)***: http://www.ucv.es/ban1_oct2006.swf

Dirección (referencia) completa del objeto solicitado

- ***cs-host***: www.ucv.es

Host cliente

- **sc-status**: 304

Código de estado obtenido en el intento de acceder al objeto solicitado

- **sc-substatus**: 0

Código de estado adicional obtenido en el intento de acceder al objeto

- **sc-win32-status**: 0

Estado de la acción, en términos de Microsoft

- **sc-bytes**: 92

Bytes transferidos desde el cliente al servidor en el proceso de solicitud del objeto

- **cs-bytes**: 714

Bytes transferidos desde el servidor al cliente en el proceso de servir el objeto al cliente

- **time-taken**: 203

Tiempo (en milisegundos) que ha tardado en servirse el objeto desde que se realiza la petición.

2. Servidor Web Escuela Técnica Superior de Informática de la UPV. (http://www.eui.upv.es)

La web esta soportada por un servidor con Apache HTTP Server. Existen dos formatos utilizados por los servidores Apache para registrar los eventos, el formato Común (Common) y el formato Combinado (Combined). La diferencia entre ambos estriba principalmente en que el formato Combinado registra más información en relación a cada evento que el formato Común.

El formato utilizado en los ficheros log de la EUI, es el formato Combined y además añaden un campo a dicho formato donde se registra el tiempo que se tarda en servir la petición del recurso al servidor.

Combined Log Format Apache HTTP Server

LogFormat "%h %l %u %t %T \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\""

http://httpd.apache.org

http://www.loganalyzer.net/log-analyzer/apache-custom-log.html

Ejemplos**:**

74.6.74.205 - - [01/Feb/2007:05:00:23 +0100] 15625 "GET /robots.txt HTTP/1.0" 404 2397 "-" "Mozilla/5.0 (compatible; Yahoo! Slurp;

158.37.73.129 - - [05/Feb/2007:05:23:09 +0100] 15626 "GET /webei/imagenes/plantillas/logo_upv.jpg HTTP/1.1" 200 3265 "http://www.ei.upv.es/webei/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; es-ES; rv:1.8.0.9) Gecko/20061206 Firefox/1.5.0.9"

http://help.yahoo.com/help/us/ysearch/slurp)"

Campos que forman el formato

- ***%h*****:** 158.37.73.129

Dirección ip del cliente

- ***%l***: -
- ***%u***: -

Identificador del usuario que ha hecho la petición del recurso.

- ***%t***: [05/Feb/2007:05:23:09 +0100]

Momento en que se ha terminado de procesar la petición. El formato interno de este campo es: `[day/month/year:hour:minute:second zone]`

day = 2 números: 05

month = 3 caracteres: Feb

year = 4 números: 2007

hour = 2 números: 05

minute = 2 números: 23

second = 2 números: 09

zone = (`+' | `-') 4 números: +0100

- ***%T***: 15625

  Tiempo que se tarda en servir la petición

- ***\"%r\"***: "GET /webei/imagenes/plantillas/logo_upv.jpg HTTP/1.1"

Este campo esta compuesto a su vez de varios campos que especifican los parámetros de la petición por parte del cliente. Los parámetros son:

  - Método utilizado para realizar la petición: GET
  - Recurso solicitado por el cliente: /webei/imagenes/plantillas/logo_upv.jpg
  - Protocolo utilizado por el cliente para hacer la petición: HTTP/1.1

- ***%>s***: 200

Código de estado que el servidor devuelve al cliente

- ***%b:*** *3265*

Tamaño del objeto servido al cliente

- ***"%{Referer}i\":*** *http://www.ei.upv.es/webei/*

Referencia (página) donde ha sido localizado el cliente al hacer la petición del objeto

- ***\"%{User-agent}i\":*** *"Mozilla/5.0 (Windows; U; Windows NT 5.1; es-ES; rv:1.8.0.9) Gecko/20061206 Firefox/1.5.0.9"*

Agente (navegador, versión y sistema ogperativo) utilizado por el usuario para realizar la petición

7.4. Herramientas utilizadas

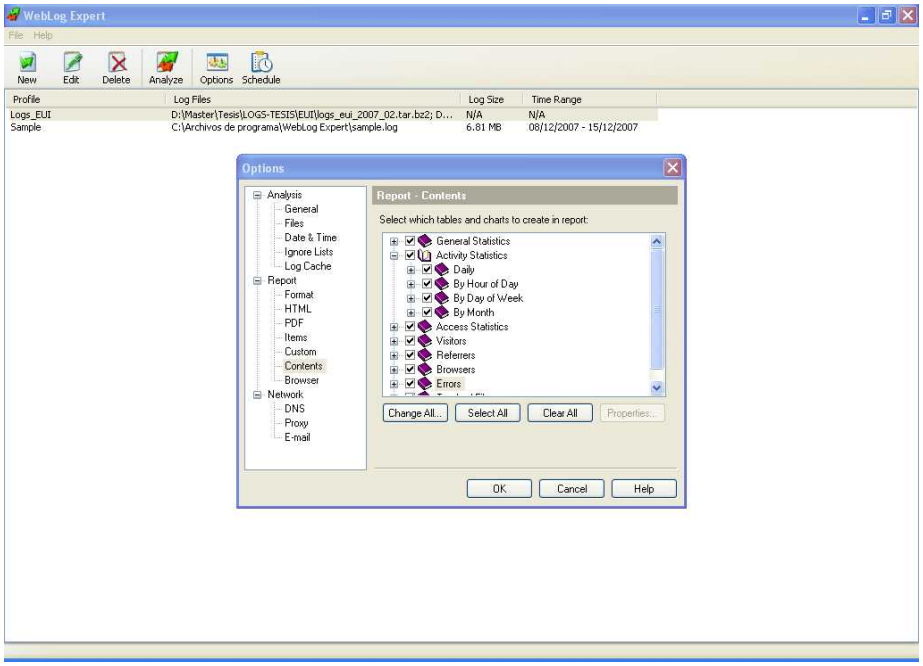
En los estudios previos consultados relacionados con la caracterización de servidores Web es habitual que se desarrollen herramientas software ad-hoc con el objetivo de obtener los datos necesarios para realizar el análisis correspondiente. En nuestro caso, uno de los objetivos era comprobar si haciendo uso de las herramientas software disponibles en la actualidad es posible realizar un estudio de tales características, sin la necesidad de tener que desarrollar específicamente ninguna aplicación para ello.

De las aplicaciones que se han probado, para que se consideraran útiles se buscaba que cumplieran varios objetivos. En primer lugar, que fueran capaces de procesar los ficheros log de que se dispone, pudiendo leer por tanto los datos en los correspondientes formatos en los que están estructurados, que en nuestro caso son dos: *IIS – W3C Extended Log File (UCV) Format y Combined Log Format Apache (EUI)*. En segundo lugar, que fueran capaces de procesar grandes cantidades de información, ya que el volumen de datos del que se disponía para realizar este estudio era considerable, en total ocupan en torno a los 7,5 Gb de disco (1,75 Gb de la EUI y 5,8 de la UCV). Por último, se buscaban aplicaciones software que ofrecieran datos fiables que fueran útiles para el principal objetivo del estudio planteado, es decir, ***la caracterización de servidores Web a partir de ciertos invariantes***.

Tras la realización de pruebas de procesamiento con varias de ellas, se decidió que era necesaria la utilización tanto de aplicaciones comerciales como gratuitas, puesto que ninguna de ellas por si sola ofrecía los datos necesarios, y por tanto se procedió al uso combinado de algunas. En definitiva, han sido utilizadas las aplicaciones ***WebLog Expert, Analog, WebLog Explorer y AlterWind***, unas en mayor medida que otras. Las aplicaciones comerciales solamente pudieron ser utilizadas durante el período de prueba que tienen establecido antes de tener que proceder a su compra (normalmente 30 días a partir de su instalación), y aunque en algún caso se intentó solicitar al fabricante una ampliación del período con motivo de utilizarse en un trabajo de investigación, éstas fueron ignoradas.

WebLog Expert 5.7

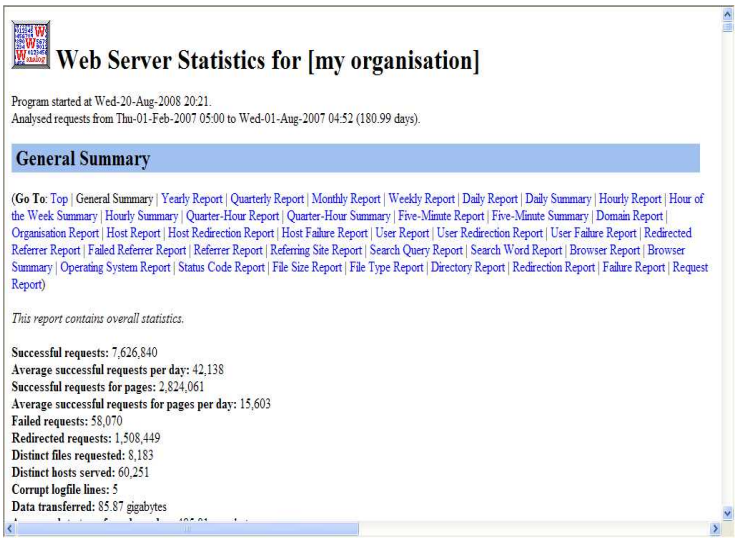
Esta aplicación comercial es capaz de procesar la información contenida Logs de la UCV, pero para procesar los de la EUI ha tenido que realizarse un preprocesado de sus Logs para que su formato fuera compatible con los admitidos en la aplicación. Proporciona una interfaz de manejo sencillo que permite configurar diversas opciones en relación al sitio Web del que proceden los ficheros Log, el período de tiempo objeto de análisis, filtros a aplicarse, etc, así como definición del tiempo de sesión. Los informes que genera son extensos y detallados, permitiendo configurar su formato (HTML o pdf), idioma, y contenidos de los mismos a partir de los datos que se pueden obtener.



Esta herramienta ha sido una de las más utilizadas para la obtención de los datos que se necesitaban para la realización del presente estudio por la cantidad de información útil que proporciona y por la posibilidad de establecer diversos tipos de filtros y de personalizar los datos presentados en los informes generados.

Analog 6.0

Esta aplicación es de software libre y su última *release* data de diciembre de 2004. Para poder utilizarla, no se realiza una instalación propiamente dicha, si no que se produce una extracción de ficheros en un determinado directorio. Para su utilización no hay una interfaz sobre la cual operar, si no que han de manejarse diversos ficheros, tanto para su configuración como para iniciar el proceso de análisis de los ficheros Log, el cual termina con la generación de un documento de informe en formato HTML.



Web Server Statistics for [my organisation]

Program started at Wed-20-Aug-2008 20:21.
Analysed requests from Thu-01-Feb-2007 05:00 to Wed-01-Aug-2007 04:52 (180.99 days).

General Summary

(Go To: Top | General Summary | Yearly Report | Quarterly Report | Monthly Report | Weekly Report | Daily Report | Daily Summary | Hourly Report | Hour of the Week Summary | Hourly Summary | Quarter-Hour Report | Quarter-Hour Summary | Five-Minute Report | Five-Minute Summary | Domain Report | Organisation Report | Host Report | Host Redirection Report | Host Failure Report | User Report | User Redirection Report | User Failure Report | Redirected Referrer Report | Failed Referrer Report | Referrer Report | Referring Site Report | Search Query Report | Search Word Report | Browser Report | Browser Summary | Operating System Report | Status Code Report | File Size Report | File Type Report | Directory Report | Redirection Report | Failure Report | Request Report)

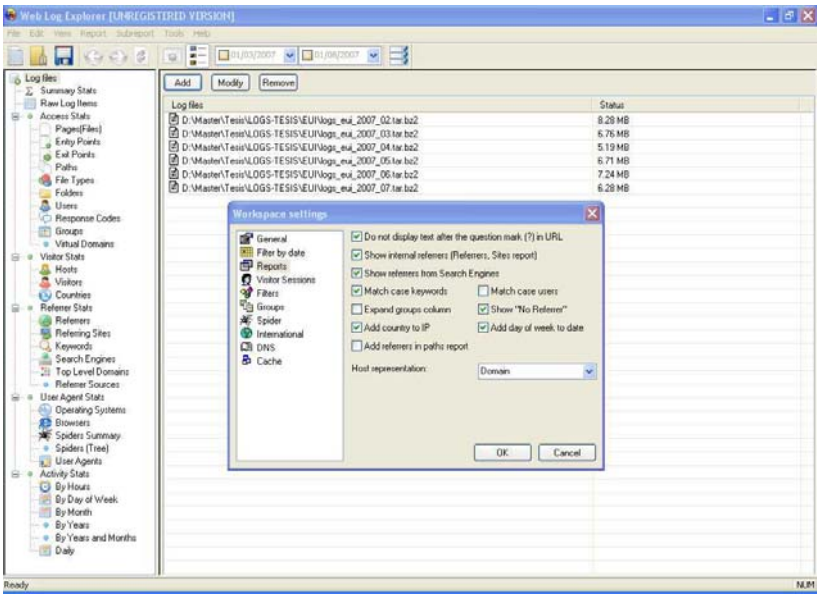
This report contains overall statistics.

Successful requests: 7,626,840
Average successful requests per day: 42,138
Successful requests for pages: 2,824,061
Average successful requests for pages per day: 15,603
Failed requests: 58,070
Redirected requests: 1,508,449
Distinct files requested: 8,183
Distinct hosts served: 60,251
Corrupt logfile lines: 5
Data transferred: 85.87 gigabytes

Aunque se trata de una aplicación desactualizada y cuyo manejo no es sencillo, es capaz de procesar los formatos de ficheros disponibles y permite un alto grado de personalización en los parámetros de configuración. Ha sido de utilidad debido a que proporciona algunos datos interesantes que no se han podido obtener a partir del resto de aplicaciones utilizadas para realizar el estudio.

Weblog Explorer 3.62

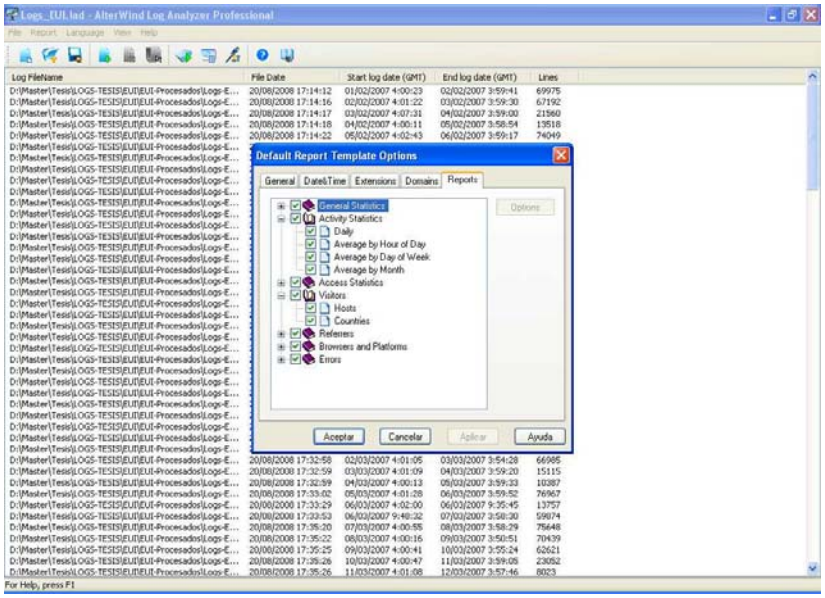
Esta aplicación comercial dispone de una interfaz sobre la que interactuar de manera sencilla, y es capaz de procesar tanto los Logs de la UCV como los de la EUI, incluso si forman parte de un fichero comprimido. Permite configurar diversas opciones y filtros y proporciona gran cantidad de información a partir de los Logs que procesa.



La aplicación no proporciona tanta flexibilidad de configuración como otras de las herramientas utilizadas, pero ha permitido comprobar la fiabilidad de los datos proporcionados ellas en diversos parámetros.

AlterWind Log Analyzer Professional 3.11

Esta aplicación también comercial, proporciona una interfaz de fácil manejo y diversas opciones de configuración de adquisición de datos y establecimiento de filtros en la línea de las otras aplicaciones comerciales mencionadas. Es capaz de procesar los Logs tanto de la EUI como de la UCV sin necesidad de procesado previo, pero no es capaz de proporcionar toda la información configurable en la aplicación si el volumen de datos de los Logs es muy elevado. La aplicación se bloqueaba cuando se le pedía que analizara todos los ficheros Log disponibles de la UCV (5,8 Gb de datos) y mostrara todos los informes posibles sobre ello.



A pesar de las limitaciones mencionadas, esta aplicación ha sido de utilidad para obtener información relativa a diversos parámetros y comprobar la fiabilidad de éstos respecto a los obtenidos por otras aplicaciones utilizadas.

Ninguna de las herramientas analizadas proporciona por separado los datos estadísticos necesarios para llevar a cabo el análisis de actividad de un sitio Web, sino que hemos tenido que obtener los datos a partir de unas y otras para poder completar el estudio, ya que no están ideadas para el análisis de ficheros Log desde la perspectiva académica o científica.

A partir de ellas la gran mayoría de los datos requeridos se han podido obtener, pero aun combinando el uso de varias, ha habido datos que pretendían ser estudiados de los que no

se ha podido extraer información, principalmente los relacionados con los tiempos de procesado asociados a las peticiones. Para la mayoría de los datos, ha sido necesaria también la utilización de otras aplicaciones ofimáticas para poder hacer un procesado posterior de los datos obtenidos y extraer la información necesaria para analizarla adecuadamente.

8. Metodología

Lo primero que se tuvo que hacer es conseguir ficheros logs actuales de servidores que soporten un volumen de tráfico suficientemente importante para que los datos obtenidos a partir de ellos se puedan considerar significativos. Finalmente se ha podido disponer de Logs procedentes de 2 servidores diferentes, el servidor Web de la Universidad Católica de Valencia San Vicente Mártir y el de la Escuela Técnica de Informática Aplicada de la Universidad Politécnica de Valencia.

A continuación se ha estudiado la documentación encontrada relacionada con la caracterización de los elementos que componen Internet, especialmente los referidos a servidores Web y su caracterización, extrayendo los parámetros de referencia (invariantes) sobre los que basar el presente estudio. Una vez hecho esto, se ha realizado la búsqueda de herramientas software desarrolladas para procesar ficheros logs y se ha comprobado cuales de ellas proporcionan datos resultantes útiles para el propósito de valorar los invariantes presentados.

Posteriormente al estudio de la documentación se procede al procesamiento de los logs mediante las herramientas software expuestas, y con los datos obtenidos, se realiza la caracterización de la carga de trabajo de los servidores de los que proceden, en base a los invariantes seleccionados de la bibliografía.

Finalmente se analizan los resultados obtenidos y se determinan una serie de conclusiones con respecto a los objetivos inicialmente propuestos.

8.1. Selección de los ficheros log

De los ficheros conseguidos de ambos servidores se han seleccionado los coincidentes en el mismo período, por lo que en ambos casos la primera fecha a tener en cuenta es el 1 de Febrero de 2007. Debido a que el volumen de información diaria es superior en el caso de la UCV, se ha optado por procesar más días de logs en el caso de EUI que de UCV.

Sitio web UCV

Accesos realizados en el período de tiempo transcurrido entre el 1 de febrero de 2007 y el 30 de abril de ese mismo año, en total 3 meses, 88 días. El total de peticiones de acceso a objetos procesados en ese período es de más de 45 millones y un ancho de banda registrado de más de 85 Gb. Ha de mencionarse que los datos del ancho de banda no representan el total de las transferencias de bytes realizadas durante los 3 meses de los logs analizados, ya que durante más de la primera mitad de este tiempo no se recogía dicha información en los ficheros. El espacio que ocupan en disco los ficheros es de 11,2 Gb.

Sitio web: EUI

Accesos realizados en el período de tiempo transcurrido entre el 1 de febrero de 2007 y el 31 de julio de ese mismo año, en total 6 meses completos, que comprende 181 días. El total de peticiones de acceso a objetos procesados en ese período es de más de 9 millones y el ancho de banda consumido es de más de 85 Gb. El espacio que ocupan en disco es de 1,74 Gb.

Peticiones	**UCV**	**EUI**
Peticiones totals	45,827,480	9,107,804
Promedio de peticiones por día	514,915	50,042
Peticiones de caché	23,130,022	3,442,999
Peticiones fallidas	624,523	57,775
Ancho de banda [7]		
Total de ancho de banda	85.48 GB	85.05 GB
Promedio de ancho de banda por día	983.46 MB	478.55 MB
Promedio de ancho de banda por hit	1.96 KB	9.79 KB
Promedio de ancho de banda por visitante	245.32 KB	307.20

Tabla 1. Resumen de datos

8.2. Preprocesado de los logs

Tras la selección de los ficheros que van a ser objeto de estudio, se ha de realizar un preprocesado respecto a los datos proporcionados en dichos ficheros para filtrar la información que nos interesa obtener y de esta manera excluir la que no nos aporta información útil en nuestro estudio o la que nos puede llevar a obtener conclusiones erróneas.

[7] Los datos del ancho de banda de UCV corresponden al mes de abril.

En primer lugar se ha realizado un filtrado para obtener los robots/arañas detectados/as en los ficheros log por las herramientas utilizadas en el procesado de los ficheros, excluyendo a éstos de los datos obtenidos con los que se posteriormente se procede a realizar el análisis.

Posteriormente y debido al hecho de que los datos de ancho de banda consumido en la UCV no han sido registrados durante todo el periodo de sus logs disponibles, se filtran para obtener los correspondientes solamente al mes de abril que es en el que se dispone de toda esa información.

8.2.1. Arañas web

Una **araña web** - también conocida como **spider, web crawler, robot, worm, walker o wanderer** - es un programa que inspecciona las páginas de los sitios web de forma metódica y automatizada, "atraviesa" la estructura de hipertexto del sitio Web recuperando un documento y de manera recursiva todos los documentos referenciados desde dicho documento. Las arañas web son uno de los componentes fundamentales de los buscadores web. Más recientemente, los sistemas de arañas se han utilizado también como herramientas para el rastreo dirigido, por aplicaciones *shopbot*[8], y como apoyo a los servicios de valor añadido en la web (portales, personalizada y los servicios móviles, etc.).

Diferentes estudios han sido realizados con el objetivo de caracterizar el comportamiento las arañas Web, y basándose en los invariantes propuestos para la caracterización de los servidores han evaluado diferentes aspectos [18], todos ellos aplicados específicamente a la actividad de las arañas:

- Distribución del tipo de recursos.
- Distribución del tamaño del recurso.
- Peticiones distintas.
- Popularidad del recurso solicitado por la araña.
- Concentración de peticiones.
- Distribución del tiempo entre peticiones.
- Periodicidad de la araña.

[8] Shopbots: agentes de compra que se dedican a comparar las características y precios de los distintos productos que ofrecen las tiendas en línea

Además, han sido propuestas algunas métricas para la evaluación del comportamiento de las arañas de manera individualizada y así poder realizar las comparativas oportunas: *el formato de preferencia, la frecuencia de las visitas y la cobertura* [18].

El análisis del comportamiento de las arañas y su caracterización no es objeto de estudio del presente documento, por lo cual no profundizaremos más en el mismo. Sin embargo sí que es de nuestro interés conocer el grado de influencia en la actividad soportada por el servidor, por lo que se realiza un breve análisis para cuantificar su relevancia.

Dado que las peticiones de acceso por parte de las arañas/robots no representan el acceso por parte de ningún usuario en concreto, no debe ser un dato a tomarse en cuenta al llevar a cabo el análisis de los ficheros log.

Para obtener de la manera más aproximada posible la cantidad de arañas utilizadas en los accesos a los sitios web objeto del estudio, se realizó una comparación entre el filtrado realizado mediante el uso de Analog 6.0, añadiendo listas de robots actualizadas obtenidas de diversas fuentes (http://www.robotstxt.org, http://user-agents.org/, http://www2.owen.vanderbilt.edu/mike.shor/diversions/analog/) y el filtrado de robots realizado mediante las otras herramientas software que no permiten la inclusión de nuevas listas por parte del usuario (Web log Expert, Web log Explorer y AlterWind).

Como resultado se obtuvo que se detectaban mayor cantidad de robots por parte de estos últimos, por lo que se ha optado por realizar el filtrado de los robots mediante el uso de las herramientas software utilizadas en el procesado de los ficheros log (Web log Expert, Web log Explorer y AlterWind).

Peticiones debidas a arañas	**UCV**	**EUI**
Peticiones totales	260.081	161.258
Promedio de peticiones por día	2.922	886
Peticiones de cache	5.801	11.467
Peticiones fallidas	27.015	41.911
Ancho de banda consumido por arañas		
Total de ancho de banda	3.44 GB	9.68 GB
Promedio de ancho de banda por día	39.63 MB	54.49 GB
Promedio de ancho de banda por hit	13.89 KB	62.97 KB
Promedio de ancho de banda por visitante	33.94 KB	164.96 KB
Porcentajes	**UCV**	**EUI**
Porcentaje de peticiones de arañas respecto al total	0.57%	1.77%
Porcentaje ancho de banda de arañas respecto al total	4%	11.3%

Tabla 2. Resumen de datos

Como puede observarse en la Tabla 2, las peticiones realizadas por parte de arañas representan entre el 0.57% y el 1.77% del total de peticiones a los sitios web objeto de estudio. Son porcentajes muy bajos respecto del total, lo que indica que no influyen de manera significativa en la carga de trabajo que han de soportar los servidores Web al servir los objetos que le son solicitados por parte de estos elementos. Sin embargo, el ancho de banda consumido por las arañas oscila entre el 4% y el 11.3%, que son valores que han de tenerse en cuenta.

Códigos de respuesta	UCV	EUI
200 OK	86,48%	64,36%
304 No modificado desde la última petición	9,11%	7,11%
302 Objeto encontrado con diferente URI	0,02%	2,5%
206 Petición atendida parcialmente	0,002%	0,01%
404 Objeto no encontrado	0,001%	22,17%
403 Acceso prohibido (el servidor no ejecuta la petición)	0,056%	3,82%
406 El método utilizado no es el permitido para el objeto	0,046%	0
301 Objeto movido permanentemente	0,87%	0,02%
500 Error interno del servidor	1,17%	0

Tabla 3. Códigos de estado ante peticiones de arañas

Las peticiones realizadas por arañas provocan en su gran mayoría la devolución del objeto solicitado, entre el 64,4% y el 86,5% como se muestra en la Tabla 3, el estado *No modificado* es generado entre el 7,1 y el 9,1% de las ocasiones. Existe gran diferencia en el caso de los objetos no encontrados, ya que en UCV solamente ocurre en un ínfimo 0,001%, mientras que en EUI se produce en más del 22%.

Como puede verse en las Figuras 1 y 2, las arañas con más presencia en las peticiones son las procedentes de Google, Yahoo y MSN, abarcando entre ellas el 70 – 80% del total. En lo relativo a la procedencia geográfica, en torno al 89% de las peticiones y al 95% de las visitas son de Estados Unidos. Este es un dato esperado, ya que las arañas anteriormente citadas tienen su origen en ese país.

Como se ha mostrado anteriormente, las arañas no suponen una importante influencia sobre el número de peticiones totales en general, pero sí en los datos de procedencia geográfica. Antes de filtrar las arañas, los accesos desde Estados Unidos representan entre 0,7% y un 2% del total, y una vez filtradas este dato está entre 0,22% y 0,35%, ligeramente inferior, pero en cambio pueden variar significativamente los datos de procedencia de los visitantes, ya que en nuestro caso con la presencia de arañas la procedencia desde Estados Unidos representa entre un 23 y un 28% de las visitas y una vez excluidas este valor se sitúa entre 1,2% y 2,8%.

Este puede indicar que cada visitante que es una araña supone pocas peticiones, o que la araña es detectada como visitante nuevo en muchas de las ocasiones que accede a los sitios Web debido a su modo de operar.

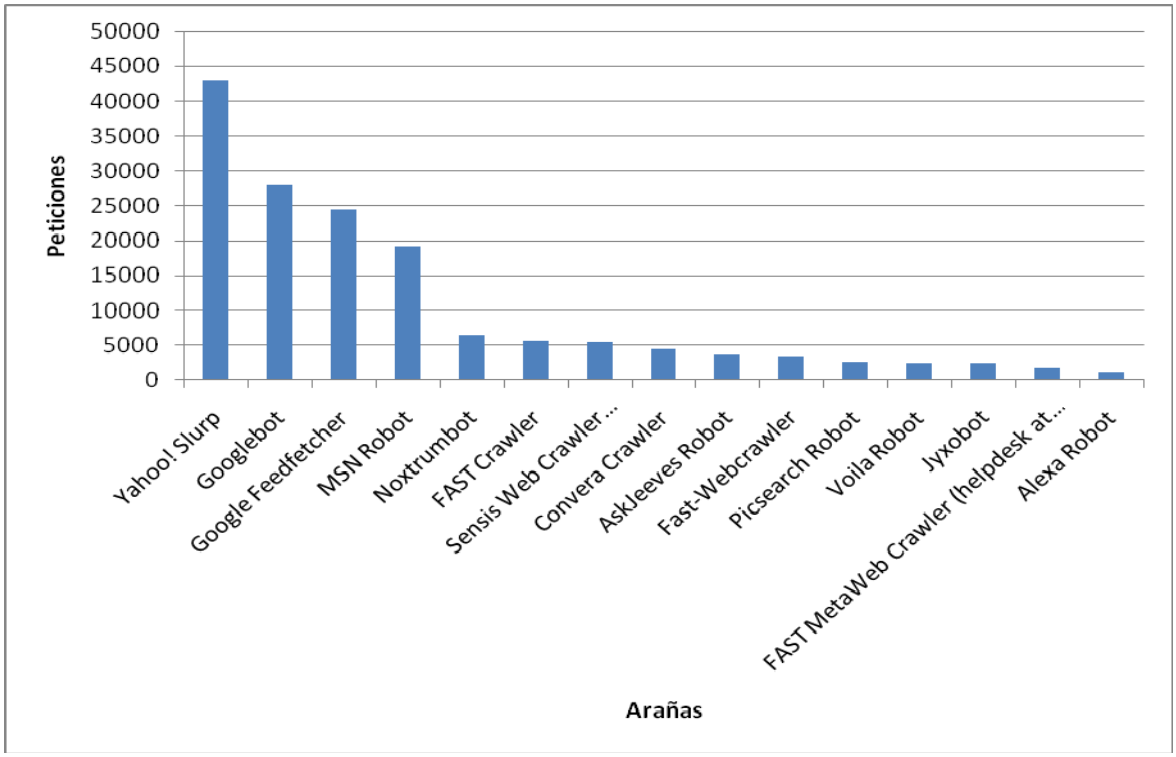


Figura 1. Arañas UCV

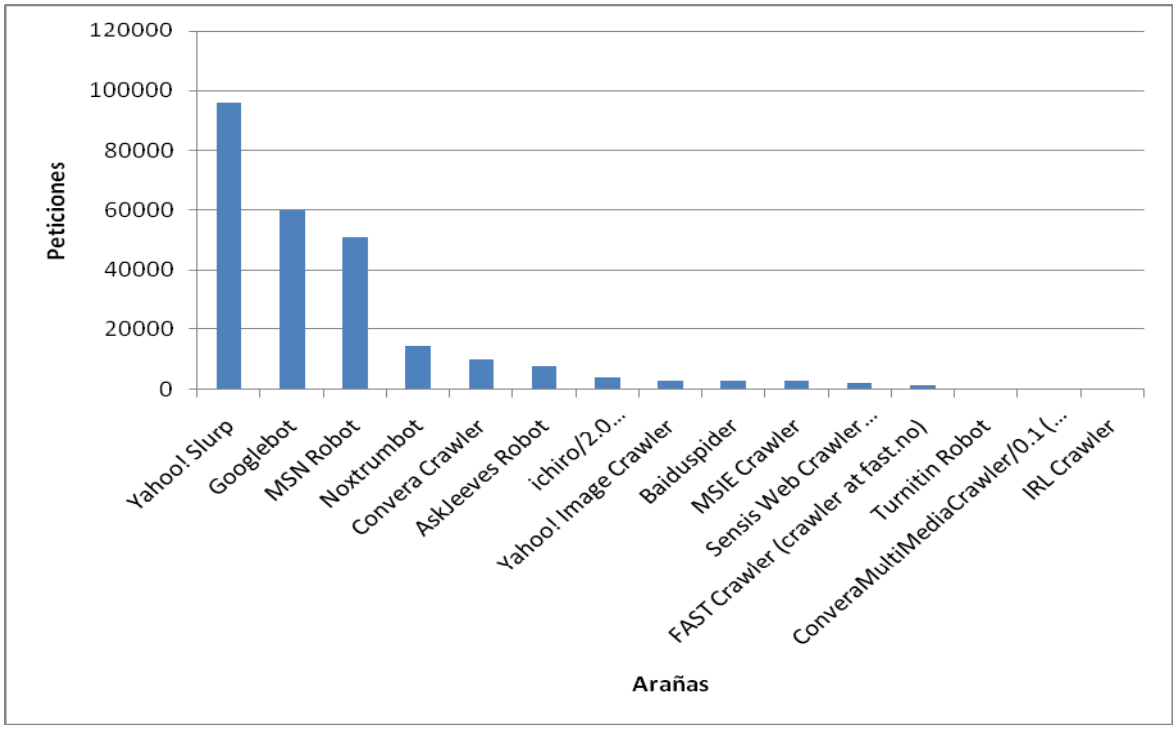


Figura 2. Arañas EUI

8.3. Análisis

El análisis de la carga de trabajo se basa fundamentalmente en la obtención de dos tipos de datos: las peticiones realizadas y la transferencia de bytes asociada.

Debido a que del conjunto de datos disponible de la UCV solamente durante uno de los meses (abril) quedaron registrados los datos relativos a los bytes transferidos durante el procesado de las peticiones, y como el volumen de información de ese mes es suficientemente importante para ser analizado individualmente, serán utilizados solamente datos de ese mes cuando haya que establecer datos comparativos con respecto a otras métricas como las peticiones.

Una vez filtrados los registros de información correspondientes a las peticiones realizadas por los robots/arañas, se estudian los datos obtenidos tras el procesado de los logs.

En base a los invariantes mencionados anteriormente y a los datos resultantes del procesado de los ficheros log, se procede a realizar el análisis para obtener parámetros que permitan caracterizar la carga de servidores que soportan sitios Web académicos.

8.3.1. Volumen y distribución de las peticiones

El número de peticiones registradas en los logs que se realizaron durante el período analizado, en el caso de la UCV suman un total de 45.567.399, y en el caso de los datos obtenidos a partir de los ficheros log procedentes de la EUI suman 8.947.056 peticiones de recursos. La distribución de las peticiones en el período de tiempo estudiado se produce tal como se muestra en las tablas 5 y 6, donde puede observarse que dicha distribución por mes es homogénea en el caso de la EUI, que oscila entre el 13.64% y el 19.37%, no siendo así en el caso de la UCV ya que el mes de febrero concentra el 44% del total de peticiones.

El ancho de banda total consumido en la EUI es de más de 76 Gb, mientras que en la UCV solamente durante el mes de abril se transfirieron más de 78 Gb. La distribución presentada por meses puede verse en las Tablas 5 y 6.

Datos de peticiones	UCV	EUI
Peticiones a ficheros distintos	119.036	8.183
Peticiones totales	45.567.399	9.032.101
Promedio de peticiones por día	511.993	49.159
Promedio de peticiones por visitante	175,95	39,04
Peticiones de cache	23.124.221	3.431.559
Peticiones fallidas	597.508	16.035

Tabla 4 .1. Resumen de datos de peticiones de los servidores

Ancho de banda*	UCV (abril)	EUI
Total ancho de banda*	75.53 GB	76.30 GB
Promedio de ancho de banda por día	2.52 GB	429.30 MB
Promedio de ancho de banda por petición	6.65 KB	8.86 KB

Tabla 4.2 Resumen de datos de transferencia de bytes de los servidores

Mes	Peticiones	Páginas vistas	Visitantes
Febrero 2007	20.063.953	3.880.002	103.241
Marzo 2007	13.595.212	2,315,954	85.921
Abril 2007	11.908.234	1.784.423	69.818

Tabla 5. Datos del servidor UCV por meses

Mes	Peticiones	Páginas vistas	Visitantes
Febrero 2007	1,685,373	761,771	43,174
Marzo 2007	1,475,691	696,945	39,393
Abril 2007	1,220,670	615,366	30,605
Mayo 2007	1,518,237	723,469	40,803
Junio 2007	1,732,729	877,775	43,005
Julio 2007	1,312,791	549,695	32,070

Tabla 6. Datos del servidor EUI por meses

Del total de las peticiones realizadas, la tasa de acierto se sitúa alrededor del 97.4%, siendo las peticiones que han fallado el 1.3% del total, y también el 1.3% las que no consiguieron completarse por diferentes motivos. Alrededor del 50% de las solicitudes fueron resueltas a través del uso de la cache.

En lo relativo a los códigos de estado de las peticiones, puede observarse en la Tabla 7 que la inmensa mayoría de las peticiones realizadas generan un código de respuesta de tipo 200 (ok), 304 (no modificado) o 302 (movido temporalmente), que comprenden entre un 98% (UCV) y un 99% (EUI) de las peticiones realizadas. Los accesos con éxito (ok) en los que se encuentra el objeto y es servido al cliente representan un gran número, situándose entre el 44,1% - 45,1%. En el caso de los códigos 304, donde se interpreta que se ha encontrado el objeto sin ser modificado pero no se sirve en ese momento existen diferencias entre los servidores analizados, ya que en el caso de la UCV este hecho se produce en un 50,75% mientras que en la EUI se da en un 38,35%. Por lo tanto, la diferencia que se aprecia es el modo en que se reparten las peticiones cuyo procesamiento resultante no genera una devolución satisfactoria del recurso solicitado.

La principal diferencia observada entre ambos servidores radica en el dato relativo a los objetos encontrados con diferente URI, que varia entre el 2.44% de la UCV y el 16.65% de la EUI.

En estudios anteriores la tasa correspondiente a los códigos de éxito (200) es significativamente mayor, si bien esta diferencia es menor cuanto más reciente es el estudio, pasando de un 90% en 1996 por Arlitt y Williamson a aproximadamente un 69-70% en 2004 - 2005. Sin embargo el código 304 es significativamente mayor, tendencia que también se aprecia en los estudios anteriores, que varia entre un 4-13% en 1996 y un 22,9% en 2004 en los mismos servidores. Faber y Gupta obtuvieron entre 6-16,5% en 2006. Otro dato relevante es que la tasa de objetos no encontrados es muy inferior a la obtenida con anterioridad, rompiendo con la tendencia que había de aumentar su valor, en 1996 era del 1,3%, en 2004 del 4,2% y en 2006 del 9,5%.

Códigos de respuesta	UCV	EUI
200 OK	45,12%	44,14%
304 No modificado desde la última petición	50,75%	38,35%
302 Objeto encontrado con diferente URI	2,44%	16,65%
206 Petición atendida parcialmente	0,0003%	0,0006%
404 Objeto no encontrado	0,001%	0,0001%
403 Acceso prohibido (el servidor no ejecuta la petición)		
405 El método utilizado no es el permitido para el objeto		
301 Objeto movido permanentemente		
401 La petición necesita autentificación de usuario		
400 Petición mal realizada (no entendida)		

Tabla 7. Códigos de estado obtenidos

8.3.2. Tipos de ficheros

Se ha decidido realizar una agrupación de los tipos de fichero utilizados en las peticiones para facilitar su análisis. Esta agrupación se ha establecido en función de la utilidad de cada uno de los tipos de fichero, resultando de la siguiente manera:

- Documentos: .dot, .ttf, .txt, .xls, .rtf, .pps, .ppt, .doc, .pdf
- Imágenes: .ico, .tif, .png, .bmp, .gif, .jpg
- Multimedia: .mp3, .wma, .wmz, .mdi, .swf, .mpg
- Dinámicas: .js, .asp, .aspx, .axd, .php
- Estáticas/HTML: .htm, .hmt, .css, .xml, HTML

Tipo de fichero	Peticiones	Incompletas	Transferencia (KB)
DOCUMENTOS	0,92%	85,86%	40,90%
IMÁGENES	60,10%	1,87%	16,32%
HTML	5,12%	6,90%	0,86%
MULTIMEDIA	17,45%	5,12%	6,47%
DINAMICAS	16,41%	0,17%	32,87%

Tabla 8. Actividad en UCV por tipo de fichero solicitado

Tipo de fichero	Peticiones	Incompletas	Transferencia (KB)
DOCUMENTOS	1,27%	97,83%	18,48%
IMÁGENES	49,94%	1,62%	10,25%
HTML	1,38%	0,06%	0,18%
MULTIMEDIA	0,01%	0,45%	17,46%
DINAMICAS	47,4%	0,04%	53,63%

Tabla 9. Actividad en EUI por tipo de fichero solicitado

En los datos mostrados en las Tablas 8 y 9 puede verse de manera clara que los ficheros que mayor número de peticiones reciben son las imágenes, oscilando entre el casi 50% en el servidor de la EUI y el 60% en el caso de la UCV. Siguiendo esta clasificación se sitúan los ficheros relacionados con las páginas dinámicas, aunque en este caso existen diferencias significativas entre ambos servidores. En la EUI las peticiones a estos ficheros suponen más de un 47%, mientras que en la UCV solamente un 17,45%. Por otra parte la tasa de ficheros menos solicitados también difiere en ambos servidores, ya que en la EUI se corresponde con los ficheros multimedia (0,01%) y en la UCV son los documentos (0,92%).

De estos datos puede extraerse que existen diferencias en la estructura de ambos sitios web, ya que, aunque en ambos casos se coincide en que las páginas contienen una importante cantidad de imágenes, en el caso de la EUI está compuesta de una gran mayoría de páginas dinámicas y prácticamente ningún contenido multimedia. Por el contrario, en el caso de la UCV las páginas dinámicas tienen una proporción menor y además existe una importante cantidad de contenido multimedia (17,5%), dato muy diferente al obtenido en estudios previos, donde es prácticamente inexistente.

En estudios anteriores, cuanto mas actual es este, se pasa de producirse una mayoría de las peticiones a los ficheros HTML y las imágenes, llegando a concentrar entre ambos tipos el 89-99%, a existir un porcentaje significativamente menor, entre el 58-78%. Por los datos obtenidos, se confirma la tendencia apreciada en estudios anteriores de que existe un aumento progresivo del contenido dinámico en detrimento de los ficheros HTML.

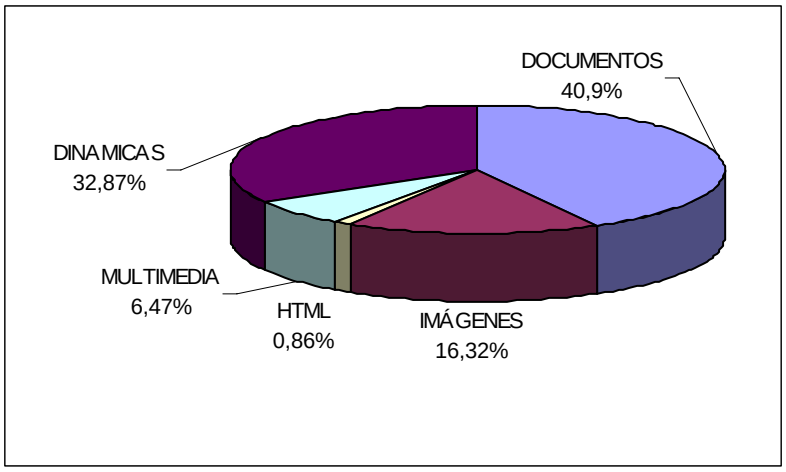


Figura 3. Transferencia de bytes por tipo de ficheros en UCV

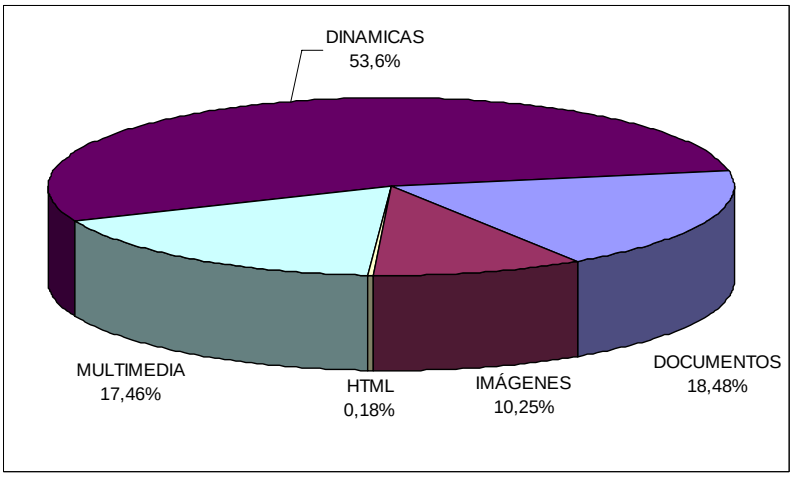


Figura 4. Transferencia de bytes por tipo de ficheros en EUI

En lo referente a la tasa de transferencia, puede observarse en la Figura 3 que en el caso de la UCV los documentos son los que consumen un mayor ancho de banda (40,9%), seguidos de los ficheros dinámicos (32,8%). Es sorprendente que la tasa de transferencia de los documentos sea tan alta siendo su tasa de peticiones tan baja, podemos deducir que esto viene determinado por el tamaño de los ficheros solicitados, principalmente pdf. En la Figura 4 se aprecia que en la EUI la mayor transferencia la realizan los ficheros dinámicos (53,6%), seguidos de los documentos (18,4%), mientras que los documentos estáticos (HTML) conllevan la transferencia de bytes más baja, como ocurre en el caso de la UCV.

8.3.3. Distribución del tamaño de transferencia de los objetos

Dado el total de bytes transferidos y el número de peticiones realizadas que se muestran en las Tablas 4.1 y 4.2, se obtiene que el promedio de bytes transferidos por petición

en términos generales se sitúa aproximadamente entre los 6,65 kb en el caso de UCV[9] y los casi 8,86 kb en el caso de EUI, por lo que en general situaremos el ancho de banda consumido por petición inferior a los 9 kb. Este dato difiere con los obtenidos en [1], donde se establecía el rango entre los 14 y los 38 kb, y los de [3], donde la cota inferior era de 6 kb y la superior de 21 kb. Puede apreciarse por tanto que se sigue una tendencia a que los tamaños de transferencia vayan disminuyendo, donde la principal causa hay que buscarla en el aumento progresivo de contenido dinámico, cuya transferencia no consume gran cantidad de ancho de banda.

Tipo archivo	Peticiones	%	Ancho de banda (KB)	%
Jpg	3,729,930	31,73%	10,011,380	12,68%
Gif	3,506,918	29,83%	2,846,668	3,60%
Swf	2,235,383	19,01%	4,737,936	6,00%
Asp	1,087,245	9,25%	20,314,813	25,74%
Js	218,991	1,86%	819,637	1,03%
Css	213,097	1,81%	177,614	0,22%
Htm	202,491	1,72%	94,807	0,12%
Axd	165,923	1,41%	972,282	1,23%
Html	149,304	1,27%	302,406	0,38%
Aspx	113,999	0,96%	3,678,493	4,66%
Pdf	60,232	0,51%	18,793,575	23,81%
Doc	38,847	0,33%	7,915,084	10,02%
Ppt	17,304	0,14%	5,686,402	7,20%
Ico	9,304	0,07%	6,760	0,00%
Mht	1,404	0,01%	105,980	0,13%

Tabla 10. Datos de los tipos de fichero más solicitados de UCV en el mes de abril

Tipo archivo	Peticiones	%	Ancho de banda (KB)	%
Php	4,198,758	46,48%	42,405,309	53,73%
Gif	2,619,358	29%	6,075,967	7,69%
Jpg	1,769,999	19,6%	1,995,605	2,52%
Css	118,566	1,3%	131,787	0,16%
Pdf	111,952	1,24%	14,606,311	18,50%
Ico	93,164	1%	124,982	0,15%
Js	14,035	0,15%	25,168	0,03%
Htm	2,631	0,03%	9,021	0,01%
Rtf	808	0,009%	15,973	0,02%

[9] Recordar que los datos de ancho de banda respecto a las peticiones se refieren a los del mes de abril.

Wmz	501	0,005%	697	0,0009%
Avi	408	0,0045%	13,348,689	16,91%
Doc	360	0,0040%	23,306	0,029%
Class	229	0,0025%	510	0,0006%
Mpg	101	0,0011%	126,497	0,16%
Png	67	0,0007%	3,773	0,004%

Tabla 11. Datos de los tipos de fichero más solicitados de EUI

	UCV		EUI	
Tamaño objetos	**Peticiones (%)**	**Transferencia (%bytes)**	**Peticiones (%)**	**Transferencia (%bytes)**
0 – 1B	0,42901%	Desconocido	45,57%	Desconocido
1B – 10B	0,00000%	0	0,0001%	Desconocido
11B – 100B	40,34029%	0,53	0,74%	Desconocido
101B – 1KB	20,44649%	1,15	1,95%	0,09%
1KB – 10KB	31,01522%	15,79	21,79%	10,27%
10KB – 100KB	6,48853%	24,76	29,37%	50,35%
100KB – 1MB	1,20792%	29,87	0,5%	12,17%
1MB – 10 MB	0,07152%	26,31	0,04%	5,80%
10MB – 100MB	0,00101%	1,59	0,007%	21,32%

Tabla 12. Distribución de actividad por tamaños de los ficheros

Puede observarse en las Tablas 10 y 11 que los tipos de fichero más solicitados varían de un servidor a otro, pero es común que los de imágenes (jpg, gif) están entre los predominantes, al igual que los de tipo dinámico (php, asp).

Desde la perspectiva del tamaño de los objetos, los más demandados se sitúan en rangos diferentes dependiendo del servidor analizado. Por un lado, en el servidor UCV la gran mayoría de las peticiones se realizan a ficheros que ocupan entre 10 bytes y 10 kb (91,8%), mientras que en la EUI los más solicitados se encuentran distribuidos en dos rangos diferentes, uno que comprende tamaños entre 1 kb y 100 kb (51%) y el otro entre 0 bytes y 1byte (45,57%). No se observa una regla común entre ambos servidores, ya que sitúan sus porcentajes significativos de peticiones en rangos diferentes, como puede verse en la Tabla 12.

Resultan sorprendentes los datos obtenidos en los que la transferencia de bytes es “0” siendo que existen peticiones para esos tamaños de fichero, es especialmente relevante en el caso del servidor de la EUI, ya que el mayor porcentaje de peticiones se concentra en ficheros que no superan el byte. Se concluye que este dato se corresponde a que no se ha podido determinar el tamaño de las transferencias. Una posible explicación de este hecho es que se produjera una interrupción durante la transferencia de objetos que se realiza mediante la transferencia del objeto por partes de reducido tamaño.

8.3.4. Distribución de peticiones distintas

Es importante conocer la cantidad de ficheros distintos que son solicitados al servidor y la concentración de peticiones en ellos, ya que si un mismo fichero es demandado por multitud de clientes, se le puede aplicar "caching" para reducir notablemente la latencia del cliente. Partiendo de los datos de la Tabla 13, se obtiene que entre el 0,0009% y 0,0026% de las peticiones se produce para ficheros distintos, de lo que puede deducirse que los objetos que forman el sitio Web tienen un alto grado de reutilización, por lo cual cobra gran importancia la aplicación del caching para mejorar el funcionamiento del sistema.

Estudios previos han cifrado este porcentaje en valores considerablemente superiores, Arlitt y Williamson [1][2] obtuvieron que este dato era menor del 3%, mientras que Faber y Gupta establecen la cota superior en el 20%. Estos datos son indicativos de que actualmente existe una menor dispersión de las peticiones respecto a los ficheros, o lo que es lo mismo, una mayor concentración de éstas en un conjunto de ficheros. Esto puede tener su explicación en el hecho de que actualmente la proporción de contenido dinámico sea mayor y por tanto las peticiones muestren un mayor grado de concentración en determinados objetos. Estas observaciones concuerdan con las expuestas en el estudio realizado por Cherkasova y Karlsson [4] donde se aprecia una mayor concentración de las peticiones.

	UCV	EUI
Peticiones distintas / Peticiones totales	0,0026%	0,0009%
Ficheros accedidos solamente una vez	6,46%	13,3%

Tabla 13. Datos estadísticos sobre peticiones distintas

8.3.5. Ficheros únicos

Los datos relativos a los ficheros que solamente han sido solicitados una sola vez tal como se muestra en la Tabla 13, determinan que este hecho se produce entre el 6,46% y el 13,3% del total de ficheros distintos solicitados, siendo el volumen sensiblemente inferior al de otros obtenidos en estudios realizados años atrás, como Arlitt y Williamson que en el año 1996 los situaban entre el 23% y el 42% [1]. En estudios más actuales, como la revisión realizada del mencionado por sus propios autores en el 2004 y el realizado por Faber y Gupta en el 2006, los datos son más próximos a los obtenidos, estando entre el (15-26%) de Arlitt y Williamson y un 33% de Faber y Gupta [2][3]. Por lo tanto se observa un seguimiento en la tendencia mostrada de ir disminuyendo la tasa de peticiones de este tipo.

Este invariante esta estrechamente relacionado con el anterior, y del mismo modo que se puede apreciar una leve tendencia a ir disminuyendo el porcentaje de los ficheros que reciben una sola petición, también ha ido aumentando la concentración de peticiones distintas, hecho indicativo de la presencia de mayor porcentaje de contenido dinámico.

8.3.6. Concentración de referencias: popularidad

La distribución de las peticiones realizadas en relación a los ficheros accedidos no sigue una norma equitativa, produciéndose una concentración importante de las peticiones sobre un grupo reducido de ficheros. Puede verse en la Tabla 14 que en torno al 0,25% de los ficheros concentran un 92 – 94% de las peticiones realizadas, en el 1% se llegan a tener entre el 97,3% y el 98%, y llegando al 5% la concentración está sobre el 99%.

Estudios realizados años atrás indicaban que el 10% de los ficheros concentraban el 90% de las peticiones [1] (1996), en otros más recientes se aprecian diferencias significativas. Por un lado hay unos que mantienen los mismos valores de concentración (Arlitt [2], Faber [3]), sin embargo otros (Cherkasova [4]) determinan que la concentración es más acentuada y obtienen que entre el 2% - 4% de los ficheros concentran el 90% de las peticiones. Este hecho es digno de mención principalmente porque podría pensarse por los datos obtenidos que existe una tendencia a que la concentración de peticiones sea cada vez mayor, pero el estudio realizado por Cherkasova es anterior a los de [2] y [3].

En cuanto a la tasa de transferencia, la tendencia a producirse un mayor grado de concentración también se cumple, y se obtiene que más del 99,8% del ancho de banda consumido se concentra en un 10% de los ficheros solicitados. Fijándonos en valores más bajos, se obtiene que el 1% de los ficheros concentran entre el 89 y el 94%. En los estudios previos, el grado de concentración de bytes se mantenía estable, siendo un 10% el que aglutinaba un 90% de las transferencias. En la Figura 5 queda patente la concentración de las peticiones de manera que sigue una distribución similar a la de Zipf, donde se observa que un pequeño porcentaje de los ficheros concentra un gran porcentaje de las peticiones.

Concentración de ficheros más populares	UCV	EUI
0.25%	93.8%	92%
1%	97.3%	98.1%
5%	98.9%	99.5%
10%	99.7%	99.8%

Tabla 14. Concentración de peticiones en los objetos más populares

Concentración de ficheros más populares	UCV	EUI
0.25%	69.78%	83.2%
1%	89%	93.9%
5%	98.97%	99.38%
10%	99.8%	99.86%

Tabla 15. Concentración de transferencias en los objetos más populares

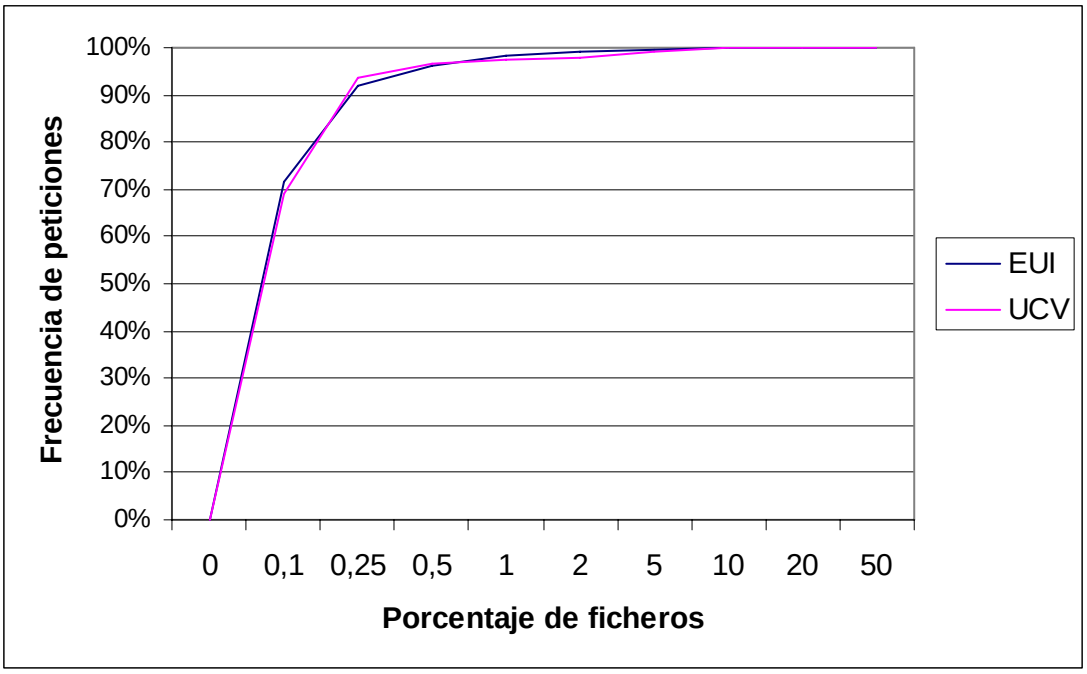


Figura 5. Concentración de referencias

8.3.7. Distribución geográfica de las peticiones

El origen geográfico de las peticiones se concentra principalmente en el país donde está situado el sitio Web (España), aunque existen diferencias en el grado de concentración entre los datos obtenidos en ambos servidores. En la UCV el 87,6% de las peticiones tiene su origen en nuestro país, mientras que en la EUI este porcentaje se eleva hasta el 98,7%. Las tasas de transferencia también están concentradas en la misma zona geográfica, siendo similares en ambos servidores, que acumulan entre el 89,4% y el 93,1% del ancho de banda total.

En estudios anteriores se realiza una diferenciación entre accesos locales y remotos, pero en nuestro caso no disponemos de la información pertinente (rangos de red correspondientes a cada centro universitario) para hacer dicha diferenciación.

8.3.8. Distribución semanal de las peticiones

En el contexto de una semana, la distribución de las peticiones se distribuye de manera que se aprecian dos grupos claramente diferenciados, los días de continuo (de lunes a viernes) y los días del fin de semana (sábado y domingo), siendo mucho menor la actividad en estos dos días como puede apreciarse en las Tablas 16 y 17. De lunes a viernes, la tasa de peticiones media se sitúa entre el 16,9% y el 18,4%, mientras que durante los fines de semana el volumen es menor de la mitad, entre el 4% y el 7,6%. La distribución que sigue el ancho de banda

consumido es similar al de las peticiones, durante la los primeros días de la semana la tasa de transferencia esta en el rango de 16,5% - 17,4%, siendo los fines de semana de 6,5% - 8,8%.

Día	Peticiones	Porcentaje	Visitantes	Ancho de banda
Lunes	7,792,075	17,10%	13,172,938	15,31%
Martes	8,446,121	18,54%	16,221,738	18,86%
Miércoles	7,887,330	17,31%	14,629,200	17,01%
Jueves	8,805,383	19,32%	15,107,548	17,56%
Viernes	5,687,248	12,48%	11,777,378	13,69%
Sábado	3,124,813	6,86%	6,740,341	7,84%
Domingo	3,824,429	8,39%	8,367,583	9,73%

Tabla 16. Peticiones semanales en UCV

Día	Peticiones	Porcentaje	Visitantes	Ancho de banda
Lunes	1,530,035	17,10%	12,376,911	15,68%
Martes	1,648,007	18,42%	14,207,653	17,99%
Miércoles	1,799,015	20,11%	15,129,943	19,16%
Jueves	1,781,746	19,91%	15,068,332	19,08%
Viernes	1,468,223	16,41%	11,988,671	15,18%
Sábado	491,410	5,49%	6,177,042	7,82%
Domingo	228,620	2,56%	4,006,101	5,07%

Tabla 17. Peticiones semanales en EUI

8.3.9. Distribución horaria

Se ha optado por agrupar las horas del día en dos franjas horarias, intentado diferenciar de manera aproximada las horas del día que se corresponden con las que se realiza actividad laboral o académica (diurna) y las que no (nocturna). La franja diurna ha sido establecida entre las 8:00 de la mañana y las 19:00 de la tarde. Dado que esta separación se ha realizado en base al horario utilizado en España, se toman en cuenta solamente los datos obtenidos como consecuencia de las peticiones realizadas desde esta región.

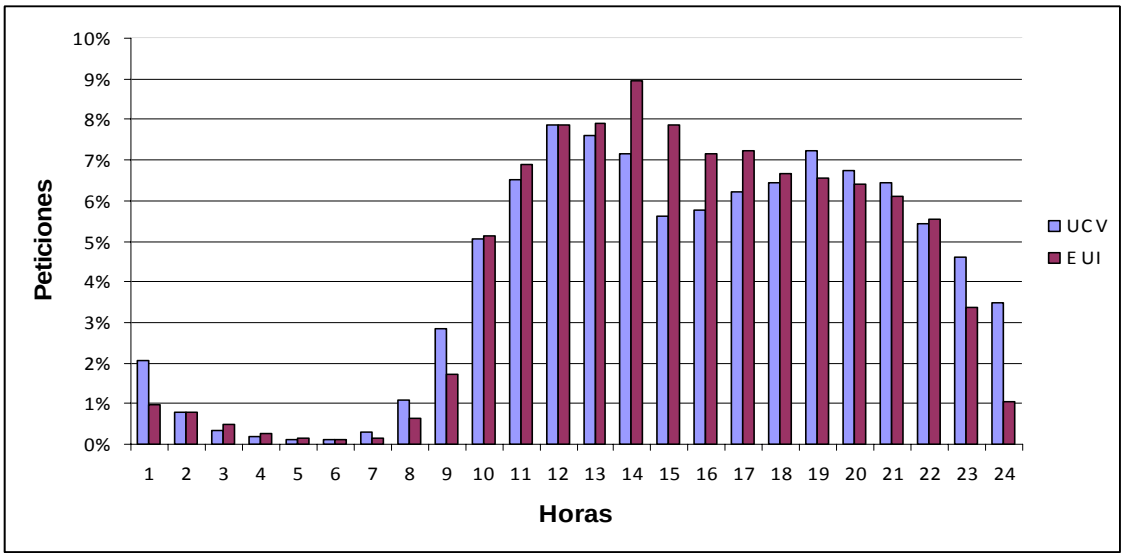


Figura 6. Comparativa del número de peticiones recibidas en ambos servidores durante el día

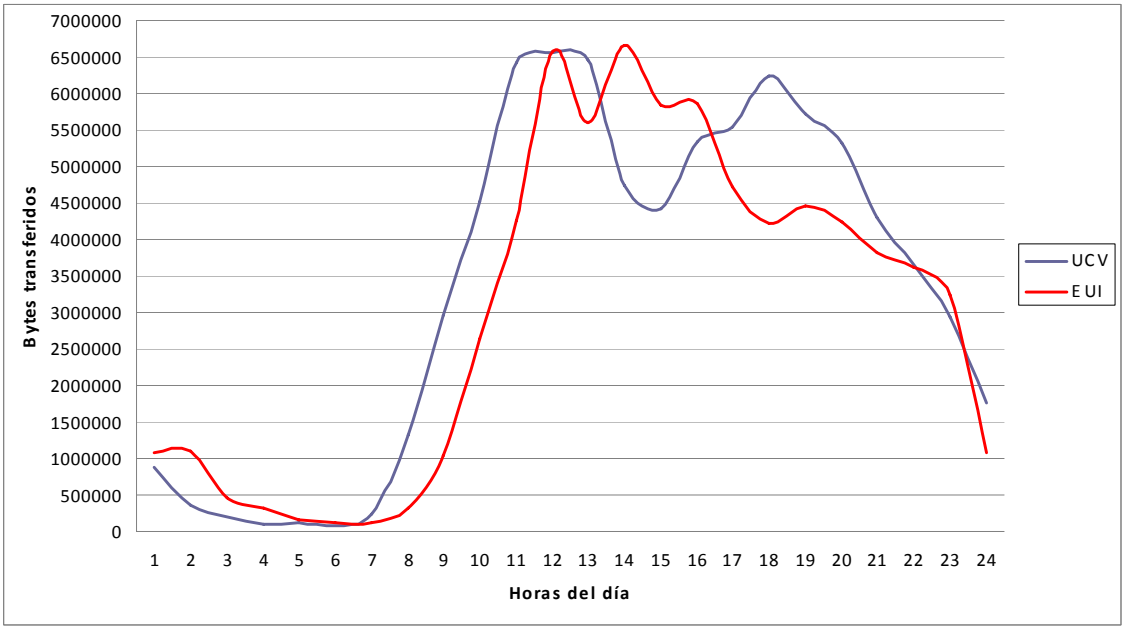


Figura 7. Comparativa del ancho de banda transferido en ambos servidores durante el día

La gran mayoría de los accesos se producen durante las horas establecidas como jornada diurna, como queda patente en la Figura 6, entre un 68,3% en el caso de la UCV y un 74% en la EUI, aun siendo menor el número de horas (11) que en la otra franja horaria.

Del mismo modo, puede verse en la Figura 7 que la mayor parte de la transferencia de bytes se produce también en la franja diurna, concretamente entre un 72,5 y un 73,5%. Este dato no es posible compararlo con los obtenidos en el de Faber, Gupta y Viecco [3] ya que las regiones son diferentes y las franjas horarias establecidas también lo son, pero es digno de mención el hecho de que en el presente estudio la actividad muestra una tendencia descendente clara durante la franja horaria nocturna a partir de las 22:00, mientras que en [3] la actividad mostraba picos irregulares de actividad durante toda la jornada.

9. Conclusiones

El objetivo fundamental propuesto a la hora de realizar el presente estudio era caracterizar la carga de trabajo de servidores web presentes en el ámbito académico, concretamente de dos universidades, que hoy en día siguen operando con normalidad.

Para alcanzar dicho objetivo se decidió consultar diversos estudios previamente publicados relacionados con la materia, de donde se extrajeron diversos parámetros (invariantes) sobre los que analizar los datos procedentes de los servidores objeto de estudio. Este análisis de los datos se ha realizado mediante el procesamiento de los ficheros Log de los servidores haciendo uso para ello de herramientas software destinadas al análisis de ficheros log en general.

Por tanto, además del objetivo principal, puede determinarse que existen otros dos objetivos secundarios, que son: por un lado comprobar que los invariantes de la literatura sirven hoy en día para caracterizar la carga de trabajo de los servidores web; y por otra parte, comprobar que se puede realizar un análisis de estas características sin hacer uso de aplicaciones *ad*-hoc, sino únicamente utilizando aplicaciones software disponibles en la actualidad destinadas al procesamiento de ficheros Log.

Invariantes estudiados

Se está produciendo una evolución constante en la estructura de los sitios Web, que conlleva principalmente a una predominante presencia de contenido dinámico y de una mayor utilización de elementos visuales (imágenes) e introducción de otros nuevos (multimedia). Esto se ve reflejado en varios de los invariantes estudiados, de manera que en algunos de ellos se confirma la tendencia de que cada vez se produce una mayor concentración de las peticiones sobre un reducido conjunto de ficheros. Este dato es interesante y está muy relacionado con lo expuesto en [4] y la perspectiva de centrar los análisis en un determinado conjunto de datos y las variaciones sufridas en este *núcleo* para evaluar el comportamiento y evolución de los sitios Web. Este hecho eleva la importancia de hacer un uso efectivo del caching para mejorar el rendimiento del sistema. Se ha apreciado también un aumento significativo en el código de estado *No modificado* con respecto a los estudios anteriores, que conlleva a su vez una disminución en el código 200 (acierto).

Los datos de otros invariantes reflejan diferencias significativas con respecto a estudios anteriores, algunos de ellos recientes. Habría que hacer una revisión de los invariantes para poder caracterizar de manera efectiva las nuevas tendencias en Internet, como por ejemplo portales de tipo YouTube o aplicaciones Second Life.

En cuanto a los invariantes de distribución temporal estudiados (horaria o diaria), resultan parciales debido a que la actividad de las personas en los horarios y las fechas en cada

región geográfica son en ocasiones muy dispares y por tanto no se puede definir una pauta general para todos los casos.

Importancia de las arañas

En las referencias bibliográficas se pone de manifiesto la importancia que pueden tener las arañas en el tráfico web y por tanto en el rendimiento de los servidores que soportan los sitios web. En los conjuntos de datos estudiados, la presencia de actividad de arañas detectadas es muy baja, por lo que no representa una influencia significativa sobre los datos de las peticiones o del ancho de banda consumido. Sin embargo sí que han de tenerse en cuenta en lo relativo a las visitas detectadas y el origen de estas, ya que la gran mayoría proceden de Estados Unidos, introduciendo un porcentaje de visitas importante desde esta zona. Además la actividad de estas aplicaciones no sigue los patrones de acceso habituales, ni durante la semana ni durante las horas del día, por lo que pueden introducir datos que varíen las pautas habituales.

Utilidad de las herramientas software utilizadas

Las aplicaciones de análisis utilizadas permiten obtener gran cantidad de datos de los logs procesados, pero no todos los parámetros necesarios para la realización de un estudio detallado de la actividad del servidor. No se han encontrado aplicaciones comerciales enfocadas al análisis desde la perspectiva académica o científica, y en los estudios de caracterización de la literatura normalmente se desarrollan aplicaciones ad-hoc para ello.

Aun así, con los datos obtenidos de estas herramientas (no ad-hoc) se han podido analizar la mayor parte de los invariantes, pero ha sido necesaria la utilización de más de una de ellas, además del uso de software de adicional para obtener una presentación útil de los datos estadísticos.

Diferencias entre los conjuntos de datos estudiados

En términos generales, los datos obtenidos en los invariantes de ambos servidores indican que la estructura de contenido de ambos sitios es similar, y el comportamiento del servidor ante las peticiones también, aun siendo servidores son soporte software diferente, aunque es cierto que se aprecia algunas diferencias significativas.

Estos datos diferenciadores entre ambos conjuntos se ponen de manifiesto en algunas características: la diferencia en la distribución de los códigos de acceso (diferencia en los 304), los tipos de ficheros accedidos (el porcentaje de páginas dinámicas es superior en la EUI), los tamaños de los ficheros más demandados (se mueven en rangos de tamaño diferentes), y la distribución geográfica de las peticiones (en la EUI están mucho más localizadas en España, mientras que en la UCV hay más accesos desde el exterior).

10. ANEXO. Distribuciones estadísticas

Distribución de Zipf

George Kingsley Zipf, (1902-1950)), fue un lingüista y filólogo estadounidense que aplicó el análisis estadístico al estudio de diferentes lenguas. A él se debe la llamada Ley de Zipf, que afirma que un pequeño número de palabras son utilizadas con mucha frecuencia, mientras que frecuentemente ocurre que un gran número de palabras son poco empleadas. Esta afirmación, expresada matemáticamente quedaría de la siguiente forma:

$$P_n \sim 1/n^a$$

Diversos investigadores observaron años atrás en sus estudios que la frecuencia relativa con la que las páginas Web eran solicitadas seguían la mencionada Ley de Zipf. La aplicación de esta ley en este aspecto determina que la probabilidad relativa de que una petición para la i-ésima página más popular es proporcional a 1/i.

En otros estudios se ha encontrado que la distribución de las peticiones de acceso tanto a servidores Web como a Proxies no siguen la Ley de Zipf, pero sin embargo, sí que siguen una distribución cercana a ésta, con un coeficiente de aproximación que varía de unas trazas a otras.

Distribución de Pareto

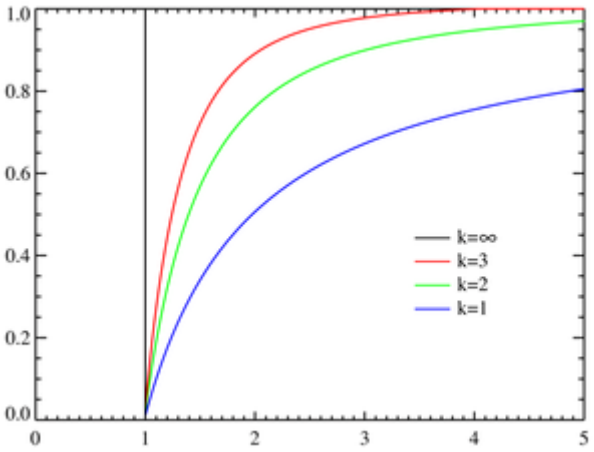
La distribución de Pareto es una distribución de las llamadas *heavy-tailed* (cola larga) que se caracterizan porque una amplia frecuencia es seguida por una baja frecuencia que disminuye gradualmente.

En estadística la distribución Pareto es una distribución de probabilidad continua con dos parámetros a y b cuya función de densidad para valores $x \geq b$ es:

$$f(x) = \frac{ab^a}{x^{a+1}}$$

Su función de distribución es

$$F(x) = 1 - \left(\frac{b}{x}\right)^{a}$$



Distribución de Pareto

En estudios de la literatura, diversos autores han observado que la distribución de los tamaños de los ficheros accedidos en los servidores Web presenta un buen ajuste con la distribución de Pareto.

11. Bibliografía

[1] Web Server Workload Characterization: The Search of Invariants. Martin F. Arlitt, Carey L. Williamson. 1996.

[2] *Web Workload Characterization: Ten Years Later.* AdepeleWilliams, Martin Arlitt, CareyWilliamson, and Ken Barker. 2004.

[3] Revisiting Web Server Workload Invariants in the Context of Scientific Web Sites. Anne M. Faber, Minaxi Gupta, and Camilo H. Viecco. 2006.

[4] Dynamics and Evolution of Web Sites:Analysis, Metrics and Design Issues. Ludmila Cherkasova and Magnus Karlsson. 2001.

[5] Self Similarity in World Wide Web Traffic: Evidence and Possible Causes. Mark E. Crovella and Azer Bestavros.1997.

[6] *Workload Characterization: Issues and Metodologies*. Maria Calzarosa, Luisa Massari, and Daniele Tessera. 2000.

[7] Workload Characterization of the 1998 World Cup Web Site. Martin Arlitt, Tai Jin.

[8] Microsoft. Logs de Internet Information Server [Online]

http://www.microsoft.com/technet/prodtechnol/WindowsServer2003/Library/IIS/3e27a577-a6e3-4b0b-9379-68efb5d52ee9.mspx?mfr=true.

[9] *Apache Log Files*. [Online] http://httpd.apache.org/docs/1.3/logs.html

[10] *W3C Extended Log File Format*. [Online] http://www.w3.org/TR/WD-logfile.html

[11] Web Server Workload Characterization. John Dilley. 1996.

[12] Improving Web Performance by Client Characterization Driven Server Adaptation. Balachander Krishnamurthy and Craig E. Wills. 2002.

[13] Changes in Web Client Access Patterns.Characteristics and Caching Implications. Paul Bardford, Azer Bestavros, Adam Bradley, Mark Crovella. 1999.

[14] Characterizing Normal Operation of a Web Server: Application to Workload Forecasting and Problem Detection. Joseph L. Hellerstein, Fan Zhang, and Perwez Shahabuddin. 1998.

[15] Web User Behavior Characterization:Techniques, Applications and Research Directions. Giancarlo Ruffo. 2002.

[16] Workload Characterization of a Personalized Web Site And Its Implications for Dynamic Content Caching. Wisong Shi, Randy Wright, Eli Collins and Vijay Karamcheti. 2002.

[17] Modeling Object Characterisitics of Dynamic Web Content. W. Shi, E. Collins, V. Karamcheti. 2003.

[18] *An investigation of web crawler behavior: characterization and metrics.* Marios D. Dikaiakosa, Athena Stassopouloub, Loizos Papageorgioua. 2005.

[19] Internet Web Servers: Workload Characterization and Performance Implications. Martin F. Arlitt and Carey Williamson. 1997.

[20] *Analysis of Educational Media Server Workloads.* Jussara M. Almeida, Jeffrey Krueger, Derek L. Eager, Mary K. Vernon. 2001.

[21] Ley de Moore [Online] *http://es.wikipedia.org/wiki/Ley_de_Moore*