Sobre la orientación a servicios en las fuentes de datos para la minería de uso de la Web

José Miguel Santos Espino
Dpto. Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
jomis@dis.ulpgc.es

Mario Hernández Tejera
Inst. Univ. de Sistemas Inteligentes y Aplic. Num. en Ingeniería
Universidad de Las Palmas de Gran Canaria
mhernandez@iusiani.ulpgc.es

Javier Lorenzo Navarro
Inst. Univ. de Sistemas Inteligentes y Aplic. Num. en Ingeniería
Universidad de Las Palmas de Gran Canaria
jlorenzo@iusiani.ulpgc.es

Resumen

Los archivos de registro de servidores web (*web server logs*) se han empleado a lo largo de los últimos años como una excelente fuente de datos para descubrir conocimiento acerca de los patrones de comportamiento de los usuarios de sistemas informáticos. Sin embargo, esta fuente de datos presenta limitaciones importantes cuando se pretende realizar un análisis orientado a los servicios del negocio, ya que la información contenida está muy ligada a la implementación de los servicios y no a sus especificaciones. En este trabajo se expone la necesidad, para mejorar la minería de uso de la Web, de disponer de una fuente de datos con información explícita sobre el uso de servicios, se describen las estrategias que se han empleado hasta ahora para lograr este objetivo, se declaran las características que debería cumplir una fuente de datos orientada a servicios y se plantean las oportunidades que la Web Semántica ofrece para satisfacer definitivamente esta necesidad.

1. Introducción

El proceso de descubrimiento resultante de la aplicación de técnicas de minería de datos se ha convertido en los últimos años en una valiosa herramienta multidisciplinar para el análisis de datos masivos [12]. En el ámbito de la Web, las técnicas de minería de datos han mostrado su utilidad, hasta el punto de que la *Minería Web (web mining)* [7] constituye un campo con personalidad propia. Los ámbitos clásicos de estudio de la Minería Web son tres: minería de contenidos, minería de estructura y minería de uso. El presente trabajo está centrado en la minería de uso de la Web (*Web usage mining*), que según Srivastava [23] es "el proceso de aplicar técnicas de minería de datos al descubrimiento de patrones de uso a partir de datos de la Web."

Históricamente, las organizaciones con mayor presencia en la Red se dedicaban a la publicación de contenidos y al comercio electrónico. En esta etapa inicial, el modelo orientado a contenidos propio de la Web ha sido suficiente para desarrollar eficazmente estos negocios. En lo que respecta a la minería de datos, los archivos de registro (*logs*) almacenados en los servidores web constituían una excelente fuente de datos sobre el uso de los sistemas, con el fin de adquirir conocimiento sobre el comportamiento de los usuarios, la utilización (consumo) efectiva de los contenidos y demás elementos de negocio.

Con el paso del tiempo, la Red se ha convertido en un canal genérico que pone en contacto a clientes y proveedores de servicios de todo tipo. En este canal están presentes empresas, administraciones públicas y ONG. Es con esta generalización cuando el modelo orientado a contenidos manifiesta sus limitaciones. En muchos casos, se necesita un análisis de los datos desde una perspectiva orientada a los *servicios* que presta la organización. El problema es que ni el modelo arquitectónico de la Web, ni las fuentes de datos estándares para aplicar procesos de minería, están orientados hacia los servicios, entendidos tales como las interacciones cliente-proveedor.

Este trabajo es una reflexión acerca de la evolución de las fuentes de datos con las que opera la minería de uso de la Web, desde su perspectiva original orientada a la navegación por un grafo de recursos, hacia otra perspectiva orientada a los servicios de la organización.

El resto del documento se organiza así: en la sección 2, se explica la estructura de los archivos

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.279-287
ISBN: 84-9732-449-8

de registro del servidor y sus limitaciones para la minería de uso. La sección 3 insiste en la necesidad de una orientación a servicios en el análisis de datos dentro de una organización. La Sección 4 describe técnicas que se han empleado para alcanzar estos objetivos en la minería de uso de la Web. La Sección 5 propone unos requisitos mínimos que debería cumplir una fuente de datos orientada a servicios. La sección 6 explora las posibilidades que da la Web Semántica para cumplir satisfactoriamente estos objetivos. La sección 7 cuestiona si la minería de uso de la Web sigue teniendo sentido en un marco de análisis orientado a servicios. Finalmente, en la sección 8 se resumen las conclusiones de esta reflexión y las perspectivas de futuro de la orientanción a servicios.

2. Los archivos de registro del servidor

A lo largo del documento, emplearemos el acrónimo WSL (*web server log*) para referirnos al fichero de registro de accesos al servidor en su formato convencional.

2.1. Contenido de los WSL

La mayoría de los servidores web registran los accesos utilizando una variante del formato ECLF (Extended Common Log Format) [11]. En esencia, el ECLF registra todas las peticiones al servidor web, anotando por cada acceso atributos como: fecha y hora del acceso; dirección IP de la máquina cliente; agente (navegador) usado por el cliente; URI del recurso solicitado (incluyendo los argumentos de la consulta, cuando sea el caso); URI desde la cual se ha efectuado la solicitud en el proceso de navegación (*referer*).

Al flujo de peticiones anotadas en un WSL se la conoce como flujo de clics (*clickstream*). A su vez, este flujo se puede abstraer en una secuencia de páginas (*pageviews*), cada una compuesta de varios accesos a URI (para el texto, las imágenes, las hojas de estilo y demás componentes de la página visualizada).

2.2. Los WSL como fuentes de datos

El flujo de clics proporciona información valiosa sobre la interacción del usuario con el sitio web. Lo más distintivo de este flujo es que registra los patrones de *navegación* del usuario a lo largo del hipertexto constituido por los distintos recursos (URI) alojados en el servidor.

Desde el advenimiento de la tecnología web, se ha progresado bastante en la explotación de los WSL como fuente datos a partir de la cual extraer conocimiento. En lo que se refiere a la minería de uso, se han desarrollado técnicas para predecir los futuros accesos de los usuarios, categorizar a los usuarios según el comportamiento observado, etc. El análisis de estos datos se ha aplicado en la personalización de la experiencia de usuario, la mejora del rendimiento (ej. cachés), recomendación de productos y en general todo tipo de actividades de inteligencia de negocio [15].

Para alcanzar estos logros se han tenido que desarrollar técnicas de preparación de datos en el WSL que superen ciertas limitaciones de origen: filtrado de páginas (*pageviews*), extracción de sesiones y eliminación de datos irrevelantes, tales como visitas de robots o recargas consecutivas de páginas. Para una introducción sobre este tipo de técnicas, véanse [8], [25] y [22].

Mediante estas técnicas de preparación de datos, podemos obtener un flujo de acceso a páginas con calidad suficiente para extraer conocimiento útil sobre el comportamiento de los usuarios y la utilización de los servicios del sitio web.

Gracias a los WSL, se ha generado un ingente volumen de datos sobre el comportamiento de usuarios de sistemas informáticos, de un tamaño sin precedentes históricos. A esto hay que añadir que el formato de los WSL se ha mantenido prácticamente sin cambios desde el inicio de la tecnología web. Esto ha dado pie a una fructífera línea de investigación que aún está lejos de acabar, así como al desarrollo de técnicas y herramientas de análisis con alto grado de reutilización, al beneficiarse de la estandarización *de facto* del formato de los WSL.

2.3. Limitaciones del WSL

El éxito del WSL cobra más mérito si tenemos en cuenta que nadie preveía la diversidad de usos que iban a tener los sitios web, incluyendo transacciones económicas *online*. Al ampliarse el tipo de aplicaciones en la Web, se manifestaron

las carencias del WSL, muchas de las cuales se han ido soslayando con mayor o menor fortuna. Algunas de estas carencias son:

- Registro de la navegación: no se registran todas las acciones que suceden en el lado del cliente (botones de retroceso y avance, accesos resueltos a través de cachés…)
- Sesiones: el protocolo de acceso a los servidores HTTPD no maneja estados, así que no existe de forma natural el concepto de "sesión de trabajo" del cliente. No se puede estar completamente seguro de cuándo un usuario finaliza su interacción con el sistema.
- Identificación de usuarios: debe ser resuelta en el nivel de la aplicación, así que no hay un criterio estándar para identificar a un usuario a partir de las entradas en el WSL.
- Semántica de los accesos: las entradas en el WSL registran meros accesos a URI. Una URI es una expresión sintáctica que sólo aporta información acerca de dónde y cómo recuperar un recurso, pero nada acerca de qué contiene o qué se resuelve con ese recurso.

Es este último punto en donde queremos centrarnos.

3. Orientación a los servicios

Para la dirección estratégica de la empresa, para un departamento de ventas y para cualquier persona con responsabilidades en el control del negocio y su desarrollo futuro, es crucial contar con información sobre cómo están actuando los usuarios.

Con vistas a este fin, es evidente que hay un salto de calidad entre descubrir que dentro de la web corporativa de una universidad, la tercera URL más utilizada es `/publica/form_inscrip.php` y descubrir que el servicio más utilizado es el de matrícula *online*. En ambos casos se está expresando un hallazgo sobre el comportamiento de los usuarios, pero el primer caso el hallazgo se refiere a la navegación por unos contenidos; mientras que en el segundo caso la información obtenida está en un nivel descriptivo cercano a los intereses de la organización y fácilmente interpretable en ese contexto.

3.1. Navegación, objetivos y servicios

El WSL ofrece información de comportamiento *navegacional*, pero no contiene, al menos de forma manifiesta, información de comportamiento *funcional* e *intencional*.

Cuando hablamos de intencionalidad, nos referimos a los propósitos, objetivos o metas que pueden tener los diferentes participantes en el sistema, especialmente los que actúan como clientes y proveedores del servicio que está suministrando el sitio web.

Planteando un modelo muy simplificado, un sitio web ofrece a los clientes una serie de *servicios* que responden a las intenciones-objetivos del proveedor. Los objetivos de los clientes son *a priori* desconocidos: la misión del proveedor es ofrecer servicios que de alguna forma satisfagan algunos de esos objetivos. En este trabajo sólo trataremos de los objetivos del proveedor, que por ser conocidos e implementados en forma de *software*, son más fáciles de tratar que los objetivos del usuario.

Hay un amplio interés por realizar análisis de la web desde una perspectiva intencional. En particular, los autores de [13] acuñan el término *Web Goal Mining* para referirse a todas aquellas técnicas de minería web relacionadas con los objetivos de negocio del propietario del sitio web.

3.2. Definición de servicio

El vocablo "servicio" es polisémico en la comunidad informática: su significado varía bastante según el autor y el contexto. Por ello es conveniente fijar qué entendemos por servicio en este documento.

Utilizaremos "servicio" de acuerdo con la definición de la antigua norma ISO 8402[1] sobre calidad: *Resultado generado por actividades en la interfaz entre el proveedor y el cliente, y por actividades internas del proveedor, con el fin de responder a las necesidades del cliente [1]*.

Atendiendo a esta definición, una página o una URL no son servicios, sino artefactos que se utilizan para que los usuarios accedan a los

[1] Esta norma fue reemplazada por la ISO 9000:2000, pero en el nuevo documento no se da una definición del concepto "servicio".

servicios del sitio web. Es decir, una URL está relacionada con la *implementación* del servicio.

En lo sucesivo, llamaremos *log orientado a servicios* a una fuente de datos con información explícita sobre el uso de servicios. Diversos autores han propuesto marcos de trabajo para minería web basados en la existencia de *logs* orientados a servicios (ej. [2],[3],[13],[9]). En muchos casos, se trata de *logs* enriquecidos no sólo con información basada en servicios, sino también con otros metadatos acerca del contenido o la estructura de las páginas.

3.3. Ventajas de la orientación a servicios

Las ventajas de trabajar con una fuente de datos orientada a servicios nacen de estas dos características de los datos en ella contenidos:

- Perspectiva de la organización, no de la tecnología.
- Perspectiva intencional (basada en los objetivos funcionales), no de ejecución (basada en cómo se resuelven los objetivos).

En primer lugar, los datos sobre uso de servicios son más fáciles de interpretar por el consumidor final de los resultados, que será un miembro de la organización.

En segundo lugar, la tecnología con la que se implementan los servicios suele evolucionar a un ritmo más rápido que los propios servicios. Por ejemplo, si la organización decide realizar un cambio tecnológico en su sitio web (ej. evolucionar de PHP a Java), todos aquellos resultados de minería web que se hubieran realizado usando la URL como elemento de análisis perderían vigencia. Sería necesario establecer la trazabilidad entre las URL antiguas y las nuevas que sean equivalentes. Esta preocupación es innecesaria si se opera en el nivel de los servicios de la organización, que en principio no varían tras un cambio meramente tecnológico.

Por tanto, la independencia de la tecnología facilita el análisis temporal del sistema durante un periodo de tiempo largo. Los datos son más duraderos. Por otro lado, al trabajar con un nivel descriptivo organizacional, los resultados de la minería serán más fáciles de comparar con los de otras organizaciones similares, siempre que trabajen con taxonomías de servicios que sean comparables.

4. Incorporando semántica de servicios al WLS

Como se ha indicado, la incorporación al WLS de significado funcional o intencional se ha estado practicando desde etapas muy tempranas de la tecnología web. Con este fin, cabe plantear varias clases de estrategias:

1. Asociar a cada URL, o a un subconjunto de ellas, el servicio que está ayudando a resolver. La asociación puede ser manual o automática.
2. Registrar directamente eventos en el nivel de aplicación, bien insertando entradas en el WLS, bien descartando el uso del WLS en favor de otro tipo de archivo.
3. Identificar secuencias de accesos en el WLS que se correspondan con escenarios típicos de algún servicio y asociarlos al servicio correspondiente.

4.1. Asociación de URL a servicios

Una forma muy sencilla de asociar URL con servicios consiste en vincular a cada tipo de servicio una expresión regular de las URL que están asociadas a él (este es el mecanismo empleado dentro de la herramienta WUM [21]). Otra sencilla técnica consiste en etiquetar directamente aquellas URL cuyo acceso implica el cumplimiento de algún objetivo de negocio de interés (ej. páginas de confirmación de compras). Esta asociación se puede realizar en cualquier momento, tanto en tiempo de ejecución como una vez generado el archivo de *log*.

En muchos sitios web es posible asociar un objetivo de negocio con una página concreta. Por ejemplo, en sitios de comercio electrónico, es fácil asociar la compra (objetivo de negocio) con aquella(s) URL en las que el usuario cierra la operación. Herramientas comerciales como Webtrends [24] o Clicktracks [6] permiten además hacer minería sobre el cumplimiento de objetivos de negocio, estableciendo relaciones con el perfil de los usuarios o las páginas desde las que proceden. Un uso habitual es el cálculo de rentabilidad de campañas de publicidad.

Otra opción es descubrir el tipo de servicio a partir del contenido de la URL, por ejemplo a partir de palabras claves dentro del texto. Numerosos sitios web están orientados al suministro de contenidos: prensa electrónica, boletines digitales, bases de conocimiento… En este tipo de organizaciones, los tipos de servicios casi siempre se pueden modelar como *tipos de contenidos*. Por ello, son organizaciones susceptibles de emplear esta estrategia.

4.2. Registrar eventos de negocio

Esta estrategia, de la que pueden verse ejemplos en [2] y [9], es la que puede dar más fidelidad en el resultado, ya que las anotaciones semánticas las realiza la misma aplicación que está ejecutando el servicio. En su contra tiene el inconveniente de que exige un componente en tiempo de ejecución que registre activamente los eventos de negocio; así como el riesgo de que ese registro se aparte de los estándares y no se puedan aprovechar las técnicas y herramientas disponibles. Todo ello eleva el coste total de implantación de esta estrategia.

4.3. Detectar escenarios de uso

La tercera estrategia serviría para situaciones en las que la ejecución de un servicio no pueda identificarse señalando a una única página, sino a una interacción compleja. La identificación de patrones de uso es un objetivo habitual de la minería de uso web, aunque no se suele aplicar para el etiquetado de servicios. Esto se debe a que las dos técnicas anteriormente descritas son suficientes en la inmensa mayoría de los casos. La técnica de detección de escenarios tiene más utilidad para descubrir servicios implícitos o emergentes (servicios de los que la organización *no es consciente* que proporciona), o bien para inferir objetivos de usuario. Estas técnicas quedan fuera del alcance de la presente exposición.

4.4. Inconvenientes

Las distintas soluciones que hemos expuesto son eficaces en su ámbito de aplicación, pero comparten el inconveniente de ser técnicas *ad hoc*, para resolver problemas concretos de un dominio o de un sitio web específico. La especificidad de las soluciones presenta estos inconvenientes:

- Se reduce la posibilidad de reutilización de la tecnología existente de minería web (técnicas y productos).
- Los resultados son más difíciles de comparar entre distintos sitios.

Una política que atenúa los problemas de reusabilidad y comparabilidad consiste en crear formatos de *log* estándares dentro de ciertos dominios, como la propuesta para el ámbito de las bibliotecas digitales de [10]. De esta forma, se garantiza una mayor calidad semántica en los datos al mismo tiempo que se mantiene una comunidad de participantes que utiliza las mismas herramientas y técnicas. Más adelante retomaremos la cuestión de los estándares.

5. Requisitos de un *log* orientado a servicios

En este apartado se exponen las características deseables en una fuente de datos (*log* o similar), procedente del uso de una web, a la que se incorpore explícitamente información sobre el uso de servicios. Consideramos que estas características son requisitos para que los futuros *logs* orientados a servicios no sólo enriquezcan semánticamente al WLS, sino que mantengan las virtudes que han dado éxito a éste.

- La fuente de datos debe reflejar la ejecución de los servicios de forma explícita.
- Debe tenderse a la estandarización de los catálogos de servicios.
- Los catálogos de servicios deben ser jerárquicos.
- La información sobre navegación también es útil y por tanto debe preservarse.

El primer requisito se sigue de la definición de *log* orientado a servicios. La forma en que se haya obtenido la información sobre el acceso a servicios es irrelevante, lo importante es que se encuentre ya anotada en los *log* antes de empezar el proceso de minería.
En los próximos apartados describimos con más detalle los restantes.

5.1. Catálogo de servicios estandarizado

La catalogación de servicios debería ser lo más estandarizada posible, para favorecer la durabilidad de los resultados de la minería y también la comparabilidad entre distintas organizaciones.

En tanto las necesidades de comparabilidad de datos dentro de un dominio sean altas, es probable que los propios actores adquieran cierta estandarización. Ya hay ejemplos de esta tendencia. La Organización Internacional de Auditores de Medios de Comunicación (IFABC) definió unos estándares para la métrica de medios de comunicación electrónicos [14] con el fin de normalizar la auditoría de los sitios y establecer comparaciones entre distintos medios. Este estándar incluye un vocabulario normalizado sobre métricas de acceso a la Web, que puede ser un punto de partida para una ontología de servicios en este dominio.

En el ámbito del comercio electrónico, Blue Martini definió un conjunto de treinta métricas orientadas a los servicios de este dominio [17], que facultaban la realización de comparaciones entre distintos comercios *online*. Estas métricas se podían obtener con un mínimo esfuerzo a partir de los *logs* propietarios de Blue Martini, que registran eventos de negocio orientados específicamente al comercio electrónico.

El reto de la estandarización es lograr una homogeneidad que permita la reutilización de datos y herramientas, pero manteniendo una diversidad suficiente para que cada organización elabore un catálogo de servicio conveniente para sus intereses.

La definición de estándares de servicios llega a su modalidad óptima si se realiza mediante ontologías formales. Más adelante veremos los avances que en esta línea puede ofrecer la Web Semántica.

5.2. Catálogo de servicios jerárquico

De acuerdo con la definición de servicio empleada en este trabajo, un servicio puede consistir en un conjunto de servicios de menor entidad. Esto lleva a la posibilidad de establecer una jerarquía de servicios con distintos niveles de detalle. Un sitio web en su conjunto puede considerarse como un servicio que se desglosa en unos servicios primarios (pueden ser procesos de negocio) y así hasta llegar a un nivel elemental. Las jerarquías de servicios se han empleado con este fin desde hace tiempo en minería web [3],[20].

En nuestra opinión, es imprescindible que el catálogo de servicios sea jerárquico. De esta forma, cada miembro de la organización trabajará con el nivel de detalle que considere apropiado para sus objetivos de extracción de conocimiento. Por otra parte, pueden existir diferentes tipos de objetivos dentro de la organización [19], que aconsejan modelar los servicios siguiendo jerarquías paralelas.

El nivel de detalle máximo aconsejable es el de "caso de uso del sistema", entendido éste como el conjunto mínimo de acciones que proporcionan valor a un usuario concreto. Por debajo de ese nivel de detalle, entramos en el territorio de las operaciones del sistema, métodos de objetos, etc., que pueden tener cierta utilidad pero que son dependientes de la tecnología y se alejan del ámbito del análisis de los servicios de la organización.

5.3. Preservar la información sobre navegación

La necesidad de trabajar con *logs* orientados a servicios no implica la desaparición de la información sobre navegación. Precisamente uno de los éxitos de los WSL ha sido el registrar esta información, como hemos señalado previamente.

El registro de acceso a servicios nos informa sobre qué funciones o contenidos emplean los usuarios, mientras que el registro de acceso a navegación nos da una información más rica sobre la conducta de los usuarios: estrategias de búsqueda de información, focalidad o dispersión en su exploración, tiempos de lectura de cada contenido, etc. Este conocimiento no es imposible de extraer en un registro puramente orientado a servicios, pero podría resultar más difícil que en un WSL convencional.

En definitiva, la información sobre servicios y la información sobre navegación han de considerase distintas y complementarias.

6. La Web Semántica

La Web Semántica es la propuesta de nueva generación para la Web [5]. La idea clave de la Web Semántica es la construcción de una red

semántica o red de conocimiento cuyos predicados están escritos en RDF, un lenguaje basado en XML. Estos predicados son URI y se distribuyen por la Red como cualquier otro recurso, siendo accesibles por los mecanismos convencionales de la Web (por ejemplo, http). Si la Web clásica es una red de objetos conectados por relaciones de navegabilidad, la Web Semántica forma una red de objetos conectados por relaciones semánticas arbitrarias.

6.1. Ontologías estándares

Para categorizar el conocimiento, la Web Semántica emplea *ontologías.* Una típica ontología consiste en una descripción jerárquica de los conceptos manejados en un dominio determinado (ej. comercio al por menor, educación universitaria, etc.) Dentro de la Web Semántica, OWL (*Web Ontology Language*) es el lenguaje propuesto para la definición de ontologías [18]. OWL está basado en RDF y opera con clases, propiedades y restricciones entre ellas.

Las posibilidades de confluencia entre la minería web y la Web Semántica están siendo exploradas. En [4] los autores prevén el empleo de ontologías de la Web Semántica para la descripción semántica de los actuales sitios web y la consiguiente mejora en la minería de los datos, tanto en los contenidos como en el uso de ellos.

OWL y la Web Semántica se pueden aprovechar como la tecnología con la que asociar servicios a las acciones de los usuarios en una web. Los beneficios de usar OWL son enormes, ya que se emplearía una tecnología totalmente compatible con la infraestructura de la Web y con sus mismas características de escalabilidad y distribución de contenidos.

En noviembre de 2004 ha sido entregado al W3C un documento para discusión sobre OWL-S (*Ontology Language for Services*)[16], un *framework* basado en OWL y orientado a la descripción de Servicios Web. La finalidad de OWL-S se centra sobre todo en facilitar la interacción con componentes en arquitecturas de Servicios Web: descubrimiento de servicios, composición, monitorización, validación, etc. Aunque la minería de uso no es uno de los propósitos fundacionales de OWL-S, aquellas aplicaciones web que se apoyen en esta tecnología podrán generar *logs* de uso enriquecidos con la información de servicios que se están ejecutando, anotada mediante URL a recursos OWL-S. Dado que este *framework* u otro similar puede convertirse en un estándar[2] en un futuro próximo, se sentarían las bases para llegar a un *log* orientado a servicios ontológico y estándar.

6.2. Adecuación de las ontologías de servicios

La dificultad principal para que esta información semántica tenga valor en la minería de datos estriba en que las ontologías de OWL-S están concebidas para el diálogo entre componentes de software (agentes), así que el tipo de servicios descritos puede estar en un nivel demasiado alejado de la perspectiva de la organización (procesos de negocio). Aquí puede ayudar la tendencia moderna a construir el software en un nivel descriptivo próximo al del negocio, como en las aproximaciones MDA (*Model Driven Architecture*) y BPM (*Business Process Modeling*). La confluencia entre estos modelos de construcción de software (MDA, BPM), la Web Semántica y la minería de datos abre unas posibilidades que hoy sólo podemos sospechar.

7. ¿Tiene sentido hablar de *minería web* en un entorno orientado a servicios?

A lo largo de este documento, hemos planteado el objetivo de trabajar con fuentes de datos de uso de sistemas web con una perspectiva funcional, orientada a servicios. En breve, trabajar con *logs* que describan el acceso a servicios del negocio. Ante este escenario, es lícito preguntarse qué diferencia habría entre la minería de uso web y la minería del uso de cualquier otro tipo de aplicación informática. Al fin y al cabo, una aplicación basada en web no es más que una aplicación cliente-servidor con una interfaz de usuario determinada. Esto es evidente en el caso de sistemas de información preexistentes a los que se les dota de una interfaz web, o a los sistemas multicanal que, aparte de una interfaz web, poseen otros mecanismos de acceso (ej. mensajes SMS). En un mundo de Servicios Web y en un marco de

[2] Aunque hay otros contendientes, como WSMO (*Web Services Modeling Ontology*), la cuestión clave es la adopción de un estándar.

análisis orientado a los servicios, podemos cuestionarnos si la minería de uso de la Web pierde sentido y se disuelve en el campo más amplio de la minería de datos a secas.

Esta última afirmación tiene validez si la tecnología web se emplea en un entorno cerrado, tal como una Intranet o una red privada virtual. Pero, desde el momento en que un sistema informático se hace presente en la Web pública, surgen unos fenómenos característicos y exclusivos de la Web, en lo que se refiere al comportamiento de los usuarios y el uso de la aplicación. Estos comportamientos característicos de la Web derivan de la naturaleza hipertextual y ubicua de la Red, y son:

- Aumenta la variedad tipológica de usuarios, incluyendo usuarios inesperados y usuarios no deseados (intrusos, fisgones).
- Surgen accesos por parte de agentes no humanos (robots, *crawlers*), que distorsionan el registro de acceso si no son convenientemente filtrados.
- El acceso al sistema puede ocurrir desde múltiples puntos de partida, algunas de ellas con relevancia en el análisis del negocio (ej. accesos desde buscadores).
- Los usuarios acceden al sistema desde diferentes tipos de artefactos (distintos navegadores, distintas plataformas *software* y *hardware*).

Todos estos fenómenos influyen en las técnicas de extracción de conocimiento sobre uso de la aplicación, bien generando nuevos problemas en la preparación de los datos, bien generando nuevas posibilidades de análisis.

Por último, señalemos que al insertarse en la Web, la aplicación entra a formar parte de un gigantesco grafo cuya dinámica todavía estamos empezando a entender (leyes potenciales, autosimilaridad, etc.) Las técnicas de minería de uso web introducen heurísticas obtenidas de la experiencia dentro de la Web que posiblemente tienen una alta especificidad, aunque esta afirmación es algo que merece una investigación más profunda.

8. Conclusiones

En este documento se ha expuesto la necesidad de enriquecer los WLS (*web server logs*) con información sobre la utilización de servicios. El objetivo es conseguir una fuente de datos con la que se puedan realizar análisis de comportamiento más valiosos para la organización, duraderos en el tiempo, independientes de la implementación de los servicios y comparables entre distintas organizaciones. Este objetivo sólo se puede lograr plenamente si hay algún tipo de estandarización sobre las tipologías de servicios (a través de ontologías) y sobre la representación de esta información en las fuentes de datos.

La aparición de la Web Semántica es una muy buena oportunidad para realizar estos objetivos, especialmente en lo que se refiere a la estandarización.

No hay que olvidar que la Web fue concebida como un conjunto de recursos alojados en distintas máquinas y conectados mediante hiperenlaces. El WSL es coherente con ese modelo: es una bitácora de la navegación que se ha observado dentro de una determinada máquina. Por ello, si pretendemos obtener conocimiento sobre el uso de servicios y sobre el alcance de metas funcionales, debemos tener presente que se trata de elementos ajenos al modelo subyacente de la Web y que su encaje no siempre será fácil.

Al incorporarse la semántica de servicios a las fuentes de datos procedentes de los servidores web, se diluye la frontera entre la minería de uso web y la minería de uso de sistemas, al menos en aquellos sistemas no orientados a documentos. No obstante, todas las aplicaciones presentes en la Web pública experimentan condiciones de comportamiento específicas de la Web que seguirán influyendo en las técnicas y herramientas para la obtención de conocimiento de estos sistemas.

Una meta más ambiciosa relacionada con la orientación a los servicios consiste en el descubrimiento de los objetivos de los usuarios a través del uso que hacen de los servicios. Este reto será menos costoso si el punto de partida son fuentes de datos orientadas a servicios.

Referencias

[1] AENOR, UNE-EN-ISO 8402 Gestión de la calidad y aseguramiento de calidad. Vocabulario (ISO 8402:1994), AENOR, 1995.

[2] Ansari, S. et al. Integrating E-Commerce and Data Mining: Architecture and Challenges. ICDM'01: The 2001 IEEE International Conference on Data Mining, 2001.

[3] Berendt, B.; Spiliopoulou, M. Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB Journal 9, 2000.

[4] Berendt, B.; Hotho, A.; Stumme, G. Towards Semantic Web Mining. International Semantic Web Conference (ISWC02), 2002.

[5] Berners-Lee, T. Semantic Web Road Map. W3C Consortium (Technical Report), 1998.

[6] Clicktracks. **www.clicktracks.com**

[7] Cooley, R.; Mobasher, B.; Srivastava, J. Web Mining: Information and Pattern Discovery on the World Wide Web. International Conference on Tools with Artificial Intelligence. IEEE, 1998.

[8] Cooley, R; Mobasher, B.; Srivastava, J. Data Preparation for Mining World Wide Web Usage Patterns. Knowledge and Information Systems 1, Springer-Verlag, 1999.

[9] Fraternali, P.; Matera, M.; Maurino, A. Conceptual-Level Log Analysis for the Evaluation of Web Application Quality. First Latin American Web Congress (LA-WEB'03), 2003.

[10] Gonçalves, M.A. et al. An XML Log Standard and Tool for Digital Library Logging Analysis. ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, 2002.

[11] Hallam-Baker, P.; Behlendorf, B. Extended Log File Format. W3C Consortium website, 1996.

[12] Hernández, J.; Ramírez, M.J.; Ferri, C. Introducción a la Minería de Datos. Pearson Educación, 2004.

[13] Hochsztain, E.; Millán, S.; Pardo, B.; Peña, J. M.; Menasalvas, E. A Framework to Integrate Business Goals in Web Usage Mining. First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, 2003.

[14] Internacional International Federation of Audit Bureaux Circulation (IFABC). **www.ifabc.org/standards.htm**

[15] Lee, L.L. Web Mining. Technical Report. CSC Consulting, 2004.

[16] Martin, D. et al. OWL-S: A Semantic Markup for Web Services. Propuesta remitida al W3C Consortium, 2004.

[17] Mason, L.; Zheng, Z.; Kohavi, B.; Frasca, B. eMetrics Study. Blue Martini, 2001.

[18] McGuinness, D.L.; van Harmelen, F. Web Ontology Language. W3C Consortium Recommendation, 2004.

[19] Menasalvas, E.; Millán, S.; Pérez, M.; Hochsztain, E. et al. Beyond user clicks: an algorithm and agent-based architecture to discover user behavior. 1st European Web Mining Forum. Dubrovnik, 2003.

[20] Pohle, C.; Spiliopoulou, M. Building and Exploiting Ad Hoc Concept Hierarchies for Web Log Analysis. Data Warehousing and Knowledge Discovery, 4th International Conference, DaWaK 2002. Springer-Verlag, 2002.

[21] Spiliopoulou, M.; Faulstich, L. C. WUM: A web utilization miner. EDBT Workshop WebDB98. .Springer Verlag, 1998.

[22] Spiliopoulou, M; Mobasher, B. Berendt, B.; Nakagawa, M. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. INFORMS Journal on Computing, Vol. 15 Nr. 2, 2003.

[23] Srivastava, J.; Cooley, R. et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000.

[24] Webtrends. **www.webtrends.com**

[25] Zhang, F; Chang, H-Y. Research and Development in Web Usage Mining System-Key Issues and Proposed Solutions:
A Survey. 2002 International Conference on Machine Learning and Cybernetics, 2002.