Aplicaciones de web mining al análisisdel comportamiento de los usuarios del sitio web del Teatro Solís

Esther Hochsztain, Raúl Ramírez, Andrómaca Tasistro y Carolina Asuaga

II Jornadas Académicas de la Facultad de Ciencias Económicas y de Administración

Agosto de 2011

Resumen

El sitio *web* de una organización cultural proporciona una herramienta fundamental para cumplir con dos de sus objetivos básicos: difundir y democratizar la cultura. Se está iniciando una nueva fase del capitalismo cultural. Antes se generaba riqueza sobre bienes tangibles, pero en el futuro la riqueza se generará esencialmente a través de la producción simbólica.

Una fuerte revolución tecnológica reduce (o elimina) a los intermediarios entre creadores y consumidores culturales. La caída de la intermediación implica romper barreras y democratizar el acceso a la creación.

El sitio *web* del Teatro Solís es una de las principales formas de comunicación con su entorno. Conocer la forma en que los usuarios de una organización utilizan la web es clave para comprender si se está brindando el servicio que se requiere, si los productos o servicios son fácilmente encontrados y, en definitiva, hasta qué punto se cumple con el objetivo que la organización pretende lograr por medio de su presencia en la *web*.

En este trabajo se presenta el análisis de *logfiles* del Teatro Solís, implementado un sessionizer con *timeout* parametrizable, que elimina *software robots* (*crawlers*), asigna un identificador a cada sesión y genera datos de resumen del proceso de sesionalización.

Palabras clave: Internet, Cultura, Teatro, Web Usage Mining, Sessionizer, Web Intelligence, visitantes, clientes

1. Introducción

El sitio *web* de un Teatro constituye un espacio fundamental para la interacción con su público. El objetivo de este trabajo es analizar los *web server logs* de Teatro Solís de Montevideo - Uruguay. Con este propósito se implementó´ en R un *sessionizer* para la identificación de sesiones de usuario en los *logifiles*. Este trabajo se enmarca en el estrecho vínculo que existe actualmente entre Cultura e Internet, y en la forma en que los avances tecnológicos impactan en las instituciones culturales.

Este artículo está estructurado presentando en la sección 2 los aspectos que vinculan la actividad teatral con el sitio web de los teatros, haciendo referencia al entorno internacional y regional, para finalizar tratando en particular el caso del Teatro Solís de Montevideo- Uruguay. A continuación, en la sección 3 se tratan los aspectos relativos al vínculo entre Cultura e Internet y en particular presentando el vínculo entre Internet y Economía de la Cultura, mercados culturales y nueva economía y cultura y uso del tiempo e Internet y teatros. Web Log Mining se aborda en la sección 4, en la que se presentan las principales características de los datos utilizados para construir un sessionador web. A continuación, en la sección 5, se explicitan aspectos relativos a la definición de sesiones realizada, así como se detallan las conclusiones obtenidas en el caso de estudio. Por último, las conclusiones y futuros trabajos se presentan en la sección 6.

2. Teatros y sitios web

El sitio web de un Teatro es un espacio fundamental en donde mostrar no sólo su programación, sino que posibilita la compra de entradas y brinda información cultural que el Teatro entiende de interés.

Por ejemplo, el Royal Opera House[1], ademar de tener una tienda *on line* en donde es posible adquirir, entre otros, los DVDs de sus obras, publica videos por ejemplo con los ensayos del Ballet, entrevistas con el director del cuerpo de baile, coreógrafo, etc. The Sydney Opera House[2] tiene un programa educativo en su sitio web para escolares y liceales, que además de actividades presenciales, brindan formación a través del sitio acorde a los programas escolares.

Al momento de escribir estas páginas, al acceder a la página de la Opera Nacional de Paris[3], el visitante se sorprende con un video a pantalla casi completa de un aria de Sor Angélica, del Trípticos de Giacomo Puccini, que se estrenará el próxima 4 de octubre en la Opera de la Bastilla, mientras que en

[1] http://www.roh.org.uk/video/

[2] 2http://www.sydneyoperahouse.com/

[3] 3www.operadeparis.fr

la página del Lincon Center for the Performins Arts de Nueva York[4] se ofrecen cuatro enlaces dinámicos con los higliths de la a programación del mes.

A nivel regional, el Teatro Colón, de reciente reapertura cuenta con un sitio web acorde a su trayectoria, con un diseño de vanguardia, y en donde es posible adquirir localidades, consultar la programación, así como obtener información sobre convocatorias y otras informaciones de interés para público y artistas [5]. El Teatro Municipal de Santiago muestra una opción más modesta, en donde no es posible adquirir localidades. Sin embargo, muestra un perfil educativo interesante, en la que por medio de youtube, es posible encontrarse con un joven Pavarotti interpretando al Duque de Mantua en Rigoletto, o a Plácido Domingo y Teresa Stratas como Alfredo y Violeta en el Brindis y Dúo de la Traviata. Asimismo, el Teatro Municipal de Río de Janeiro también permite la compra de entradas on line y presenta fotos de los espectáculos realizados.

El diseño del sitio *web* del Teatro Solís esta´ a la altura de otros países sudamericanos, aunque el gran debe es que no es posible aún comprar las entradas on line como suele ser usual en los grandes teatros.

El sitio web del Teatro Solís, es una herramienta que además de proporcionar información sobre los horarios y tarifas de las distintas actividades, permite que el público visite virtualmente las instalaciones del teatro, posibilita el acceso a su valiosa colección documental; y hasta presenta un programa para docentes en el marco de los distintos programas escolares.

3. Cultura e Internet

No parece quedar dudas que cultura y tecnología siempre han ido de la mano. Quizás Gutenberg y su imprenta con caracteres móviles, es el mayor hito de la historia de la cultura. La imprenta de Gutenberg provoco´ una verdadera revolución cultural; ya que el saber escrito dejo´ de ser patrimonio de una ´elite y se extendió´ a amplias capas de la población, propiciando radicales transformaciones en la política, la religión y las artes.

También la última década del ochocientos y primeras del novecientos deberán recordarse como fundamentales a la hora de vincular cultura y tecnología, debido a la aparición de la radio, el cine y posteriormente la televisión.

Sin embargo, los cambios que trajo consigo la aparición de la imprenta sólo son comparables a los que origino´ la generalización de la informática en el umbral del siglo XXI, y actualmente las computadoras están sustituyendo a los documentos impresos como instrumentos para transmitir y conservar

[4] 4www.lincolncenter.org ultima´

[5] 5http://www.teatrocolon.org.ar/ ultima´

los textos; e Internet es un instrumento optimo´ para obtener información, escuchar música o simplemente comunicarse entre individuos.

El impacto de Internet en el ámbito´ de la cultura ha sido tratado por múltiples autores y diferentes enfoques. Todos ellos destacan los enormes cambios que se han presentado en los últimos´ años, y anticipan que se continuarán registrando cambios.

En este sentido, se plantea (Veltman 2005) que pueden identificarse algunos de los desafíos que habrá´ que abordar en los próximos diez o veinte años. En el campo tecnológico se produce´ el paso de las tecnologías de la información y la comunicación (TIC) a las tecnologías de convergencia universal (TCU). En el ámbito´ de la cultura, estos desafíos conllevarán problemas de almacenamiento, un cambio de políticas sobre el patrimonio cultural, nuevos vínculos entre expresiones nacionales, regionales y locales, y también entre la cultura, el conocimiento y la erudición, sin olvidar aproximaciones a la propiedad intelectual y a los modelos de cultura. Veltman plantea que se perfilan cinco tipos de peligros: un comercialismo excesivamente entusiasta, una actitud tecnofobia por parte de los eruditos, narrativas contrarias a la universalidad, el olvido del pasado y, finalmente, una destrucción sistemática de la memoria. Asimismo, establece la necesidad de una e-culture net permanente, que serviría para hacer frente a estos retos, desarrollar métodos críticos y crear nuevos modelos de cultura.

Por otra parte, en el artículo "Why everything has changed: the recent revolution in cultural economics" (Cowen 2008) se plantea que *Internet*, *iPod*, *Kindle*, *blogs*, juegos de computadora y realidades virtuales indican que la economía de la cultura cambio´ mucho en los últimos´ años, y se explican y analizan algunas de las implicancias de esos cambios.

3.1. Internet y Economía de la Cultura

El valor de la información, el valor de las redes, las motivaciones para la participación en las redes sociales y el impacto de los ciclos de negocio en los sectores culturales se consideran las potenciales áreas´ para desarrollos interdisciplinarios en la Economía de la Cultura en el articulo "The Cultural Economy Moment?" (Flew 2009), destacando el rol fundamental de *Internet*. Asimismo, Flew subraya los siguientes aspectos:

- El crecimiento de Internet y de la economía digital implica un crecimiento exponencial en las cantidades de información disponible y las capacidades para acceder a la misma, lo que conduce a desarrollos relativos al valor de la información.

- Las redes son una de las formas de coordinación entre los agentes a los largo de las jerarquías y de los mercados, su significación aumento´ por razones relacionadas con el desarrollo de Internet y las tecnologías digitales.

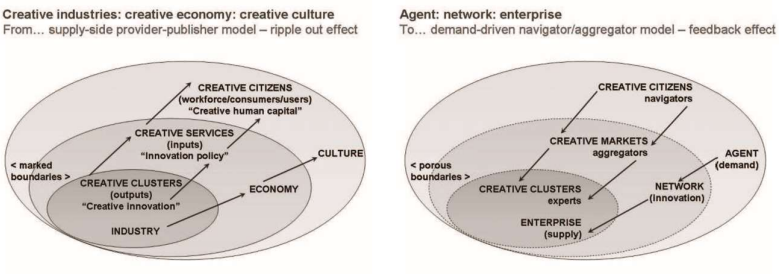
- En cuanto a las motivaciones para la participación y colaboración en las redes sociales on-line, la economía ha subestimado y reducido la significación de las actividades no de mercado. En cambio, actualmente las actividades no de mercado con motivaciones no pecuniarias (hasta altruistas) pasan a ser relevantes.

- La relación entre la cultura y el conjunto de la economía conduce a estudiar lo que ocurre con las actividades culturales y las industrias creativas en periodos de recesión.

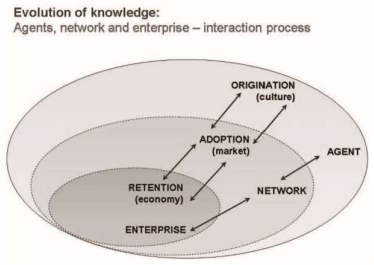
En cuanto a aspectos relativos a la información, Komito (Komito 2008) considera que se pasó de centrarse en la información a poner el foco en la comunicación de la información. La diferencia más visible de este cambio es el desarrollo de sitios de redes sociales que permiten a los individuos incorporar y compartir información. Esta es la punta visible de un iceberg que abarca, entre otros, una creciente dependencia de la tecnología y de las organizaciones que diseñan, producen y dan soporte a tales tecnologías.

Bajo el nombre de Web 2.0 se agrupan los conceptos de contenido generado por los usuarios, publicación web dinámica y grupos sociales on-line. Se paso´ de centrarse en el consumo de información digital, la forma en que los individuos acceden a la información provista por las organizaciones, estas nuevas aplicaciones de Internet permiten compartir la información entre usuarios que son ahora proveedores individuales de información. El discurso se trasladó desde la Sociedad de la Información a la Sociedad de la Comunicación. Esta conclusión esta´ asociada a la evidencia empírica que indica que Internet es menos un medio a través del cual se provee de información a los usuarios, y más un medio mediante el cual la información generada por los usuarios se comparte con otros usuarios de Internet. El interés se desplazó desde el crecimiento en la cantidad de información digital hacia el crecimiento en la comunicación de dicha información digital.

El pasaje de DIY a DIWO (Do-It-Yourself a Do-It-With-Others) ha sido señalado por Hartley como un cambio muy significativo (Hartley 2008). Uno los efectos más característicos provocado por este cambio es que se presenta una nueva forma de productividad asociada a las redes abiertas. Es en este contexto que las innovaciones se transforman en fenómenos 'bottom-up' en lugar de 'top-down', como anteriormente. Asimismo, desaparecen algunos conceptos tradicionales, tales como la distinción entre experto (comercial) y amateur, productor y consumidor, etc. Hartley caracteriza gráficamente el impacto de las tecnologías en las industrias creativas mostrando el pasaje del modelo de oferta al modelo de demanda y al desarrollo interactivo, como muestra la Figura 1.



(a) Era-Industrial-era o modelo del (b) Era-Red o modelo de demanda proveedor



(c) Modelo Interactivo

Figura 1: Modelos de desarrollo en la cultura. Tomado de Hartley (Hartley 2009)

3.2. Mercados culturales y Nueva Economía

Pueden analizarse diversos ámbitos´ de relación entre la nueva economía y la cultura, de acuerdo con el planteo de Rausell (Rausell 2002). Como muestra la Figura 2 la nueva economía puede analizarse desde diversas dimensiones: esta´ vinculada con el proceso de digitalización, con la globalización y con proceso de innovación donde el conocimiento o la producción simbólica constituye un elemento central.

Figura 2: Centralidad de la producción simbólica. Fuente: (Rausell 2002)

Por una parte, Internet y la digitalización han representado una revolución en el transporte de la producción simbólica, entre la que destaca la producción cultural. Como consecuencia, es notoria la pérdida de poder de los sistemas tradicionales de distribución de cultura, tanto en lo que concierne a la distribución propiamente dicha como a la influencia sobre las decisiones del consumidor de cultura, ya que la intermediación ya no es tan necesaria. Además, gracias a su facilidad de almacenaje y capacidad indefinida, Internet permite una cohabitación entre elementos culturales contrapuestos, que se relacionan y se enriquecen. Internet también hace que el consumo de cultura varíe en función de la capacidad de digitalización de la producción cultural, con el consiguiente desplazamiento de la demanda hacia la producción digitalizada.

Por otra parte, el papel central de los bienes culturales en la nueva economía genera un conflicto con el derecho de la propiedad intelectual que hace que sea necesario redefinir este concepto en el nuevo marco, así como analizar el alcance y las consecuencias de la piratería.

Loa motivos conducen a plantear que la revolución tecnológica ha abierto las puertas de la creación y constituye una fuente de recursos creativos, a partir de los cuales es posible generar nuevos productos culturales y crear sin virtuosismo.

3.3. Cultura y Uso del tiempo

Al igual que el dinero, el tiempo es un recurso escaso, y e las personas que economizan con uno también lo hacen con el otro (Pinnock 2009), lo que tiene en particular tres consecuencias:

1. Dado que los valores culturales se aprenden, y aprender lleva tiempo, a partir de las experiencias de aprendizaje a las que los individuos han dedicado tiempo en el pasado puede nuevas formas de creencias y modificando esencialmente los valores de las inferirse acerca de su futuro comportamiento en relación con el consumo cultural y sus actitudes respecto al arte.

2. Las artes tienen un impacto socio-económico alentando personas. A su vez, las creencias influyen sobre la acción. Por consiguiente, se deriva la importancia del análisis de la generación del valor.

3. El uso de indicadores del tiempo permite a los asesores capturar y cuantificar los impactos en base a escalas monetarias-económica.

El supuesto subyacente en la economía de la cultura "neoclásica" es que el valor cultural puede ser mejor medido en términos de disposición a pagar o disposición a que le paguen. Este concepto puede ser útil´ también para medir impactos.

Por consiguiente, el uso de la economía para mostrar la forma en que los individuos rankean sus preferencias por bienes y servicios, no implica un juicio acerca del valor sino de sus acciones (Peackok and Rizzo 2008). Todo acto de elección implica ponderar los beneficios de la acción en relación a los beneficios de otro curso de acción. No implica decir que un curso de acción es "mejor" que otro. Asimismo, en el uso del tiempo debe tenerse en cuenta el costo de oportunidad de este.

Si planteamos el particular el valor del tiempo libre, en una investigación realizada en USA (Goldfarb and Prince 2008) se encontró´ que las personas con menores ingresos y nivel educativo usuarios de Internet dedicaban ma´s tiempo en Internet, que las de mayores ingresos y nivel educativo.

Se plantean cuatro razones para tal comportamiento: (1) diferencias en el costo de oportunidad del tiempo libre, (2) diferencias en la utilidad de las actividades on-line, (3) diferencias en la cantidad de tiempo libre, y (4) selectividad. Se concluye que existe evidencia para considerar que las diferencias se explican mejor por la primera razo´n, el costo de oportunidad del tiempo libre.

3.4. Internet y Teatros. El Teatros Solís del siglo XXI

Ninguna organización puede obviar hoy en día tener su propio sitio web. Si las organizaciones tienen un fin de lucro, el sitio web es una importante herramienta para mostrar sus productos y efectuar ventas. Pero si la organización tiene un fin social y cultural, la importancia de poder mostrarse a la sociedad es doble, ya que facilitar el acceso a la cultura es un elemento que suele estar estrechamente vinculado a la misión de este tipo de organizaciones.

Desde el ámbito académico, el análisis de los sitios web de las organizaciones culturales es un tema que ha adquirido preponderancia en los museos (ver por ejemplo (Wang et al. 2007), (Marty 2007) (Marty 2008), (Lazarinis, Kanellopoulos, and Lalos 2008) , (Mason and McCarthy 2008), (Mason and McCarthy 2008), etc.), pero no ha tenido un interés mayor en teatros (véase no obstante (Kendall 2006); (Hall 2007) , (A.Abuhamdieh, Kendall, and Kendall 2007), (Hall 2007), (Slack, Rowley, and Coles 2008).

Sin embargo, el sitio *web* de un teatro, en especial de un teatro público, cumple un papel estratégico para dar a conocer el diverso número de actividades culturales que se realizan en la institución intentando alcanzar el objetivo de democratización del acceso a la cultura.

El Teatro Solís no esta´ ajeno a esta realidad, y su misión se define como: ".. servicio público eficiente que brinda una programación abierta a todas las orientaciones estéticas buscando la excelencia artística y promoviendo la accesibilidad democrático de la ciudadanía. Es el mayor referente patrimonial de las artes escénicas del Uruguay y su cuidado debe combinar la capacidad de innovar artística y técnicamente con la preservación del patrimonio".

Cabe destacar que este teatro es sin duda el principal referente de la cultura montevideana. Es una organización cultural pública que depende de la Intendencia Municipal de Montevideo. Esta´ ubicado en el corazón de la Ciudad Vieja y desde su reinauguración ha ofrecido nuevas y variadas opciones culturales a los distintos consumidores. A su vez, junto con otras alternativas de recreación, básicamente nocturnas, ha potenciado la zona logrando acaparar mayor cantidad y más variedad de público.

Bouret (Bouret 2004) señala que el Teatro Solís es considerado documento, patrimonio y monumento. Por eso que en cualquiera de los posibles caminos a recorrer para esa construcción histórica, el documento/monumento Teatro Solís debe ser entendido mostrando la diversidad de sus usos (desde galas de óperas´ a bailes de carnaval; desde entregas de títulos universitarios a velatorios); la diversidad de sus público (desde la elite política y grandes propietarios hasta los sectores populares); la diversidad de gestión (desde empresa de acciones a patrimonio público); y las modalidades de uso (desde arrendamiento a producciones); y el diálogo con el entorno y los cambios urbanísticos. Bouret (op cit) también sostiene que considerar como "´éxito" o "fracaso" su gestión, esta´ en relación directa

con los objetivos programados por cada administración, con la calidad de los espectáculos ofrecidos, con el público que accedió´, con las formas de promoción, y fundamentalmente, con el grado de inserción en su territorio. Porque un territorio es un recurso con valor económico, ambiental y cultural, inmerso en proyectos de desarrollo sectorial.

4. Web Log Mining

Conocer la forma en que los usuarios de una organización utilizan la *web* es clave para comprender si se esta´ brindando el servicio que se requiere, si los productos o servicios son fácilmente encontrados y, en definitiva, hasta que punto se cumple con el objetivo que la organización pretende lograr por medio de su presencia en la *web*. Uno de los métodos habituales para conseguir este conocimiento es el análisis de *logfiles*, un sendero útil´ pero no exento de problemas. Internet esta´ plagada de ejemplos de sitios *web* cuidadosamente diseñados en los que, sin embargo, los usuarios se pierden, no encuentran aquello que buscan, existiendo en la web o peor aún, buscan algo que debería estar y no esta´. Por otra parte, los gestores del sitio web desconocen en muchos casos lo que hacen sus usuarios dentro de la web así como si encuentran lo que buscan, si buscan conceptos que no están pero que podrían estar o si simplemente se pierden y se aburren abandonando el sitio. Sin conocer el impacto que tienen las acciones de la organización en su *website*, difícilmente se progresara´ en la dirección adecuada.

4.1. Logfiles

La estructura de un *logfile* es extremadamente simple. Cada vez que alguien descarga un elemento de la *web*, como por ejemplo una página o una imagen, el servidor escribe una línea en el archivo histórico o logfile. Lo importante es que a pesar de lo elemental que es, el estudio estadístico de la agregación de las muchas peticiones que hacen los navegadores de los usuarios al servidor permite conocer una gran cantidad de información derivada. Se puede conocer el número de páginas visitadas por día, semana, mes o unidad de tiempo que se desea, los sitios que apuntan al sitio *web* analizado y redirigen tráfico (referrer), las palabras que se buscan más habitualmente en el sitio *web* analizado y otros numerosos aspectos. Gran parte de los analizadores de *logfiles* de un servidor *web* muestran el número de visitas únicas, de visitantes, el tiempo de visita a una pagina, entre otros. Sin embargo, se presentan varias dificultades para extraer información de buena calidad, tal y como se presenta en el apartado 4.2.

Los enfoques de las investigaciones actuales realizadas en *Web Usage Mining* son muy variados, pero la mayoría se centran en las sesiones como los presentados por (Velásquez and Jain 2010) (u¨rnkranz2010), (Dimopoulosa et al. 2010), (Suneetha and Krishnamoorthi 2009), (Baohua

2009), (Jin-hua Zhu and Jun-jie Chen 2008), (Liu and Keelj 2007), (Andrejko et al. 2007) y (Castellano, Fanelli, and Torsello 2007).

4.2. Datos utilizados en la construcción de un Sessionizer

Uno de los principales problemas en *Web Usage Mining* consiste en la identificación y el procesamiento de grandes cantidades de datos. Los datos de uso utilizados pueden ser obtenidos a nivel del servidor (*server level collection*), a nivel del cliente (*client level collection*) o a nivel proxy (*proxy level collection*).

Sin embargo, gran parte de la información obtenida mediante el análisis de logfiles es de fiabilidad reducida por los siguientes motivos:

- HTML es un *stateless protocol*. Cada petición resulta en una nueva conexión independiente que se abre y se cierra para la ocasión y no se puede relacionar de un modo fiel con otra hecha por la misma dirección IP. Aun más, si la IP es dinámica, es decir si la pueden usar distintos usuarios.
- Muchas páginas se reciben desde caches de servidores intermedios, sin que el servidor de la organización llegue a enterarse nunca de que alguien ha visto esa página guardada en otro servidor. El uso de caches en Internet no solo es conveniente sino la única´ manera de no colapsar ante un trafico creciente, pero limita el conocimiento del uso real del sitio web bajo análisis. Así pues, es imposible conocer de verdad cuantas pa´ginas han sido vistas.

En resumen. Atendiendo simplemente al análisis de un *logfile*, no se puede conocer el número de visitantes, no se puede determinar cuántas visitas ha habido, no se puede conocer la identidad de los visitantes ni se pueden establecer de forma fidedigna las rutas que han seguido. Tampoco se puede saber cuánto tiempo han estado usando nuestra *web*.

Sin embargo, ello no significa que las informaciones que se derivan del análisis de *logfiles*, aunque incompletas, no sean valiosas.

- Para empezar, a falta de un sitio web en el que se obligue a los usuarios a identificarse mediante un “login”y un “password”, la información de los logfiles es probablemente todo lo que tenemos.
- Aunque la información sea incompleta, se puede llegar a una gran cantidad de conclusiones estudiando un logfile. Por ejemplo:
  - Qué conceptos buscan nuestros usuarios que no están en la web
  - Qué conceptos que sí que están no son encontrados.
  - Qué zonas de nuestra web registran más actividad.

- La aparición de patrones regulares y repetitivos en los caminos que encuentran los ”sesionizadores” suele corresponder a patrones reales de comportamiento.

5. Sesionizer para análisis de sesiones en sitios web culturales

A continuación, se presenta el trabajo realizado en la construcción de un sessionizer implementado en R (www.r-project.org) y su aplicación al análisis de los *web server logs* del Teatro Solís.

5.1. Identification de sesiones

Las principales formas de identificar sesiones se basan en uno de los siguientes aspectos: IP + Agente, Identificadores de sesiones embebidos, Registro, Cookie o Agente de Software.

En nuestro trabajo, utilizamos el primero (IP + Agente) dado que siempre esta´ disponible y no requiere tecnología adicional. Sin embargo es de destacar que el método seleccionado presenta como desventaja que no garantiza que el usuario sea único. ´

Sin embargo, esta limitación es menos relevante que las limitaciones de las restantes alternativas: generar una sobrecarga adicional en páginas dinámicas (identificadores de sesiones embebidos), que muchos usuarios no se registraran (registro), que puede ser eliminado por el usuario (cookie) o que puede desagradar a los usuarios (agente de software).

5.2. Identification de crawlers

Un *crawler* es un programa que realiza búsquedas en la web, por tanto deben distinguirse los usuarios “humanos” de los usuarios *crawlers*. Los *crawlers* trabajan de una forma metódica y automatizada (Giles, Sun, and Councill 2010), creando así una base de datos donde va guardando los resultados de sus búsquedas consecutivas, para posteriormente poder analizar dicha información y realizar tareas como la indexación y búsquedas más eficientes en la web (técnicas usadas por ejemplo por motores de búsqueda como Google, Yahoo, etc.).

El comportamiento de un crawler es cíclico, realiza búsquedas en profundidad, accediendo internamente a los *links* que se le dan y así sucesivamente dentro de los *links* que va encontrando a su paso.

La remoción de *crawlers* resulta fundamental al analizar *web server logs*. En muchas ocasiones se piensa que se cuenta con gran cantidad de accesos del exterior en un sitio *web*, y al analizar en detalle los *logfiles* se encuentra que los mismos no son usuarios propiamente dichos, sino que corresponden a *robots*.

5.3. Proceso de análisis

El estudio y exploración comienza con la obtención de los datos, lo que en muchos casos puede no resultar fácil para usuarios 'de negocio'. Posteriormente, los datos deben preprocesarse para ajustar su formato a los requerimientos de etapas posteriores. A continuación, se procede a la identificación y eliminación de *crawlers*. Posteriormente se requiere identificar el *timeout* considerado. Por último´ se pasa a la determinación de sesiones propiamente dicha, asignando a cada una de ellas un identificador. Por último se obtienen medidas de resumen, tales como cantidad de sesiones, duración de cada una de ellas, cantidad de páginas accedidas, bytes transferidos, etc. Las etapas requeridas para el análisis de *logiles* se resumen en la Figura 3.

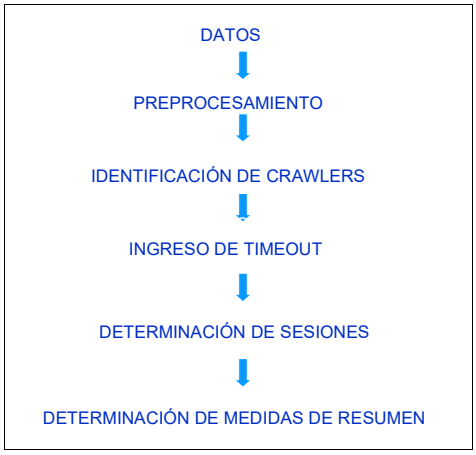


Figura 3: Etapas de la construcción de un *sessionizer*

5.4. Caso de estudio

El análisis de los datos de los *web server logs* del Teatro Solís, permitió´ constatar los siguientes aspectos, acordes a lo señalado por gran parte de la bibliografía:

- la mayoría de las sesiones tiene poca duración, con una distribución con marcada asimetría positiva (cola a la derecha)
- existe correlación positiva entre la duración de las sesiones, la cantidad de páginas accedidas en la sesión y la cantidad de *bytes* transferidos
- los datos de duración, tamaño y cantidad de páginas accedidas se encuentran concentrados en valores reducidos, coincidiendo mínimo, primera cuartila y mediana. Sin embargo, el 50% superior de los datos presenta grandes posibilidades de análisis.

En el gráfico 4 se resumen varios de los hallazgos antes reseñados.

6. Conclusiones y trabajos futuros

Desde hace varios años nuestro grupo esta´ trabajando en temas de *Web Intelligence*. Se han desarrollado metodologías y herramientas para cuantificar en qué medida un sitio web satisface los objetivos previstos. Una característica particular de *Web Mining* es que la gran cantidad de datos a

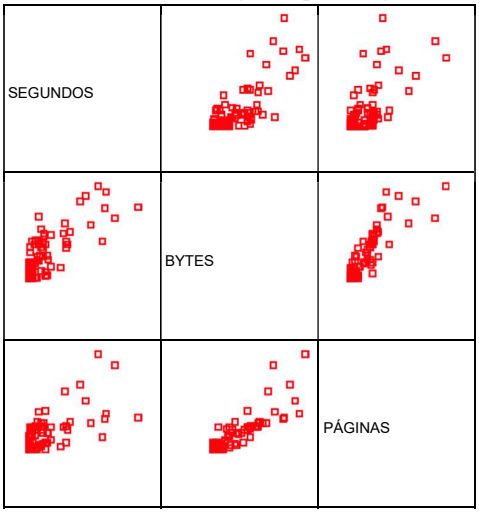


Figura 4: Diagramas de dispersión

procesar constituye en muchas ocasiones una limitante, si no se cuenta con el *hardware* adecuado. Por este motivo, recientemente logramos adquirir un servidor para el análisis de *web logs*. En el pasado los trabajos de demoraban por la inadecuación de los recursos informáticos disponibles.

Actualmente se está´ trabajando en la mejora de la interfaz gráfica que facilite el uso del sistema, disponer de una galería de gráficos, ampliar los datos de resumen, establecer un método de consulta interactivo, contar con salida a una base de datos MySQL y el modelamiento multidimensional con vistas a realizar consultas OLAP.

Como caso de estudio se han aplicado las técnicas y metodologías desarrolladas al análisis de los web server logs del Teatro Solís. Con los resultados preliminares se pudo apreciar que la página principal esta´ cumpliendo el objetivo de brindar los datos necesarios para la mayoría de los usuarios, las sesiones generadas presentan grandes potencialidades para conocer los intereses de los usuarios y proponer ajustes al diseño de su sitio web. Asimismo, se están testeando las nuevas funcionalidades con los web server logs del Teatro Solís, y ajustándolos en base a los requerimientos planteados.

El teatro Solís es patrimonio de todos los uruguayos. Una correcta gestión de su sitio *web*, en el que además de promover y difundir los diversos espectáculos artísticos, se propicie la cohesión social, así como la generación y reafirmación de valores simbólicos compartidos, es un tema no menor tanto en los objetivos del Teatro como en el beneficio de la sociedad toda.

Agradecimientos

Quisiéramos expresar nuestra profunda gratitud a las autoridades del Teatro Solís, así como a los alumnos de los diversos cursos vinculados a la temática del artículo. El trabajo fue financiado parcialmente por el proyecto RedClara - AECI, a través del cual se financio´ la compra del servidor.

Referencias

A. Abuhamdieh, J.E. Kendall, and K. E Kendall. 2007. "E -Commerce Opportunities in the Nonprofit Sector: The Case of New York Theatre Group." *International Journal of Cases on Electronic Commerce, Vol. 3, Issue 1*.

Andrejko, Anton, Michal Barla, Ma´ria Bielikova´, and Michal Tvarozek. 2007. "User Characteristics Acquisition from Logs with Semantics." *ISIM*.

Asuaga, C., Cambeiro, P., & Cami, M. 2007. Gestión de Teatros Públicos Quantum, Revista de Contabilidad, Economía y Administración, 2(1).

Asuaga, C., & Rausell P. 2006. Un análisis de la gestión de Instituciones Culturales: El caso específico de los Museos *Revista iberoamericana de contabilidad de gestión*, *4*(8), 83-104.

Baohua, Zhao. 2009. "A Web Mining Based Courseware Access Pattern Analysing System For Distance Education." *Computer Applications and Software*.

Bouret, Daniela. 2004. "Teatro Solís, Historias y Documentos." *Editorial EBO. Montevideo p.43*.

Budiño, G. 2004. Sistemas de información para la satisfacción de clientes. Universitario Autónomo del Sur-Uruguay

Budiño, G., Correa, N & Pintos, G (2011). Nuevas tendencias, tecnología e impacto en las organizaciones. Facultad de Ciencias Económicas y de Administración.

Castellano, G., A. M. Fanelli, and M. A. Torsello. 2007. "LODAP: a log data preprocessor for mining web browsing patterns." *AIKED'07: Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 12–17.

Cowen, Tyler. 2008. "Why everything has changed: the recent revolution in cultural economics." *Journal of Cultural Economics. Publisher Springer Netherlands* 32 (4): 261–273 (diciembre).

Dimopoulosa, Costantinos, Christos Makrisa, Yannis Panagisa, Evangelos Theodoridis, and Athan. 2010. "A web page usage prediction scheme

using sequence indexing and clustering techniques." *Data Knowledge Engineering*.

Flew, Terry. 2009. "The Cultural Economy Moment?" *Cultural Science* 2, no. 1.

Giles, C. Lee, Yang Sun, and Isaac G. Councill. 2010. "Measuring the web crawler ethics." *WWW '10: Proceedings of the 19th international conference on World wide web*. New York, NY, USA: ACM, 1101–1102.

Goldfarb, Avi, and Jeff Prince. 2008. "Internet adoption and usage patterns are different: Implications for the digital divide." *Information Economics and Policy* 20 (1): 2–15 (March).

Hall, John C. 2007. "The Old Operating Theatre, Museum and Herb Garret website." *Journal of Surgery. Vol. 77 Issue 12*.

Hartley, John. 2008. "The future is an open future. cultural studies at the end of the 'long twentieth century'and the beginning of the 'chinese century'." *CCI/ Feast Joint Research Workshop: Creative Destruction: Lessons for Science andInnovation Policy from the Rise of the Creative Industries*.

_____ . 2009. Chapter From the Consciousness Industry to Creative Industries: Consumer-created content, social network markets and the growth of knowledge of *Media Industries: History, Theory and Methods*, edited by Jennifer Holt and Alisa Perren, 231–244. Oxford: Blackwell.

Hochsztain, E., & Asuaga, C. (2008). Cultura e Internet: Actividad de extensión vinculada al análisis del sitio web del Teatro Solís. *I Jornadas de Extensión del Area Social*

Hochsztain, E., Transistro, A., & Asuaga, C. (2008). Sesionador Web dirigido al estudio de sitios web culturales: Diseño e Implementación del paquete RWeb Sessionizer. *VIII Congreso Latinoamericano de Sociedades de Estadística*

Jin-hua Zhu, and Jun-jie Chen. 2008. "Research on Method for Session Identification in Web Log Mining." *Journal of Taiyuan University of Technology*.

Kendall, J. E. 2006. "Theatres, Metaphors, and E-Collaboration: An Examination of Web-Based Cooperation of Regional Nonprofit Theatres." *Theatres.International Journal of e-Collaboration, Vol. 2, Issue 1*.

Komito, Lee. 2008. Chapter Information Society Policy of *Knowledge policy: challenges for de 21st century*, edited by Greg Hearn and David Rooney, 83–97. Edward Elgar Publishing Limited.

Lazarinis, Fotis, Dimitris Kanellopoulos, and Petros Lalos. 2008. "Heuristically Evaluating Greek e-Tourism and e-Museum Websites." *Electronic Journal Information Systems Evaluation Volume 2008*.

Liu, Haibin, and Vlado Keelj. 2007. "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users future requests." *Information Processing Management Volume 61, Issue 2*.

Marty, P. 2007. "Museum Websites and Museum Visitors: Before and After the Museum Visit." *Museum Management and Curatorship, Volume 22, Issue 4*.

_____. 2008. "Museum websites and museum visitors: digital museum resources and their use." *Museum Management and Curatorship, Volume 23, Issue 1*.

Mason, D, and C McCarthy. 2008. "Museums and the culture of new media: an empirical model of New Zealand museum websites." *Museum Management and Curatorship, Volume 23, Issue 1*.

Peackok, Alan, and Ilde Rizzo. 2008. *The Heritage Game. Economics, Policy, and Practice*. Edited by Oxford University Press. Oxford University Press.

Pinnock, Andrew. 2009. "The measure of all things: on the use of time as a value indicator in arts impact assessment." *Cultural Trends, Volume 18, Issue 1 March 2009 , pages 47 - 74* 18 (1): 47 – 74 (March).

Rausell, Pau. 2002. "Los mercados culturales y el desarrollo de la nueva economía." *Debate Culturales. Universitat Oberta de Catalunya*.

Slack, F., J. Rowley, and S. Coles. 2008. "Consumer behaviour in multichannel contexts: the case of a theatre festival." *Internet Research. Volume 18 Issue:1*.

Suneetha, K. R., and R. Krishnamoorthi. 2009. "Identifying User Behavior by Analyzing Web Server Access Log File." *IJCSNS International Journal of Computer Science and Network Security*.

Velazquez, Juan D., and Lakhmi C. Jain. 2010. *Advanced Techniques in Web Intelligence-1*. Springer.

Veltman, Kim H. 2005. "Desaf´ıos en las aplicaciones de las TIC/TCU en el patrimonio cultural." *Digithum: Las humanidades en la era digital*, vol. 7.

Wang, Yiwen, Lora M. Aroyo, Natalia Stash, and Lloyd Rutledge. 2007. "Interactive User Modeling for Personalized Access to Museum Collections: The Rijksmuseum Case." *Study Lecture Notes in Computer Science. Volume 4511*.