Evaluación del éxito del sitio web de una organización

• ESTHER HOCHSZTAIN Y ANDRÓMACA TASISTRO*

RESUMEN

Los sistemas de comercio electrónico no logran incorporar los objetivos de la organización. Se propone un enfoque que denominamos *Web Goal Mining*, que tiene por finalidad facilitar el vínculo entre informáticos y directivos de la empresa. Se incorpora el punto de vista empresarial al análisis de la actividad en el sitio *web*, proponiendo una solución integradora al problema de la desalineación de los objetivos del sitio *web* con los que son estratégicos de la empresa. Para concretarlo se presenta un *framework* de *Web Goal Mining*, que permite incorporar los objetivos de la empresa promotora del sitio *web* al proceso de *Web Mining*.

Palabras clave: *clickstream, web server logs, Web Mining,* comercio electrónico, sitio *web*, objetivos *web*, promotor, usuario.

ABSTRACT

Electronic commerce systems are not able actually to consider organization goals. We propose the Web Goal Mining approach, to make easier communication between computer science and business people. The proposal integrates company point of view to Web logs analysis, providing an integrated solution to the misalignment of Web site objectives with strategic business goals problem. A Web Goal Mining framework is presented in order to implement the proposal, joining Web site sponsor company objectives with Web Mining process.

Keywords: *clickstream, web server logs, Web Mining, e-commerce, web site, web goals, sponsor, user.*

1. INTRODUCCIÓN

Los directivos de una empresa que recientemente ha incorporado el comercio electrónico, no encuentran la forma de relacionar sus objetivos de negocio con los informes que les presenta el Departamento de Informática. Por ejemplo, Informática considera cumplidos los objetivos del sitio *web* presentando la alta disponibilidad del sistema (el sistema no ha permanecido fuera de servicio más de dos minutos en el último mes) y la larga permanencia de los usuarios en el sitio. Sin embargo, los directivos de negocio consideran que las ventas no crecieron en todos los productos en la forma esperada. Luego de varias reuniones se encontró que las diversas áreas de la empresa tienen objetivos diferentes, y que era necesaria una metodología que homogeinice los distintos enfoques.

Como solución al problema, en este trabajo se propone un enfoque que denominamos *Web Goal Mining*, que tiene por finalidad facilitar el vínculo entre informáticos y directivos de la empresa. Este concepto se define formalmente en la Sección 3.

Analizar en qué medida las compañías promotoras de los sitios logran los objetivos que se plantean con su presencia en el *web*, es un tema de reciente interés, que hoy constituye un problema abierto. Las pocas propuestas que se dedican a procurar la satisfacción de los promotores de sitios *web* ofrecen soluciones específicas a problemas aislados [Fayyad, 2007].

La efectividad de un sitio *web* se potencia si se logra que el patrón de comportamiento de los usuarios en

* esther@ccee.edu.uy
tasistro@fing.edu.uy

un sitio *web* se ajuste (en alguna medida) a las metas del negocio y si se alcanza cierto nivel de "consistencia" entre las metas de los promotores y las acciones orientadas a aumentar la satisfacción de los usuarios. Jesús Mena [Mena, 2001], un pionero en *Web Mining*, ya en el año 2001 señalaba la necesidad de que el logro de los objetivos de las compañías en el ámbito *web* se mida en función del retorno de la inversión, y que su éxito y supervivencia dependan de sus beneficios. Como consecuencia, plantea que se necesitan métodos más precisos para medir la eficacia y eficiencia de su sitio *web* desde el punto de vista del logro de las metas de la compañía promotora del mismo.

Son muchas las empresas que tienen presencia en *Internet*. No obstante, esta presencia no siempre es rentable, o no proporciona todos los beneficios esperados. En este artículo sostenemos la hipótesis de que uno de los motivos que lo explican, es la carencia de modelos que incorporen aspectos de negocio. En particular, que tengan en cuenta los objetivos de las empresas y las múltiples unidades organizativas que conforman una organización.

Como consecuencia de la falta de modelos formales de *Web Mining* que integren los objetivos de las compañías promotoras del sitio *web*, surge la necesidad de tener en cuenta explícitamente aspectos empresariales para determinar el éxito de un sitio *web*. En este trabajo se ofrece una solución al problema de incorporación de objetivos comerciales en *Web Mining*. En particular, analizar en qué medida las compañías promotoras de los sitios logran los objetivos que se plantean con su presencia en el *web* es un tema de reciente interés, que constituye hoy un problema abierto.

Este nuevo reto implica descubrir automáticamente información de los enormes volúmenes de datos recolectados en los sitios *web*, con el objetivo de brindar a los directivos de las empresas información que les permita lograr éxito en la consecución de sus metas empresariales. Para integrar los objetivos de la empresa a la identificación de patrones en la *web*, habrá que analizar la estructura de la organización.

En este trabajo se agrega el punto de vista empresarial al análisis de la actividad en el *web*. Se propone una solución integradora al problema de la desalineación de los objetivos del sitio *web* con los estratégicos de la empresa. Pensamos que una de sus causas es la incorrecta consideración de diferentes puntos de vista divergentes. Si bien la desalineación puede no constituir un problema al tratarse cada área por separado, puede dificultar el análisis y la toma de decisiones cuando se considera a la

empresa globalmente como un todo. Los puntos de vista más apartados entre sí generalmente son el de negocio y el informático. Hace falta una investigación en este sentido para reducir y/o sintetizar la distancia entre ellos.

Integramos semántica de negocio al análisis de las actividades desarrolladas en un sitio *web* para lograr la obtención de una visión global del usuario y la organización. También se considera que el significado empresarial de un mismo hecho en una misma organización no es único sino múltiple. Su importancia y alcance pueden diferir para las diversas unidades organizativas, a las que hemos denominado Puntos de Vista (*View Points*) siguiendo a Gordjin [Gordjin, 2003]. Para representar esta situación se modela el vínculo entre objetivos y puntos de vista.

La propuesta ofrece un enfoque global a través del cual se pueden integrar las diversas perspectivas utilizadas para evaluar la actividad en *Internet* (informática, comercial, financiera) asociando los niveles operativo, de procesos de negocio y de tecnología de la información.

2. MOTIVACIÓN

Las formas de comunicarse y de hacer negocios han cambiado mucho a partir del surgimiento de la *World*

Wide Web e *Internet*, dos conceptos estrechamente vinculados. Se denomina *World Wide Web* (*world wide web*, *web*, *Web* o *WWW*) a una red de ordenadores consistente en un conjunto de sitios de *Internet* que ofrecen recursos de texto, gráficos, sonido y animación a través del protocolo de transferencia de hipertexto HTTP.

2.1 Internet

Internet surgió a comienzos de 1969 como ARPANET, la red del Departamento de Defensa de Estados Unidos de América. La red ARPANET surgió con dos objetivos principales. El primero fue desarrollar una arquitectura de red con fines militares de alta disponibilidad (que se mantuviera funcionando ante grandes interrupciones en las comunicaciones). El segundo fue economizar en el uso de recursos computacionales escasos. Pese a no ser exactamente equivalentes, es habitual referirse a *Internet* y *Web* como sinónimos.

La *world wide web* se ha convertido en un nuevo medio de comunicación, más barato y que brinda a sus usuarios más autonomía e independencia para divulgar y consultar que los medios de comunicación tradicionales. Desde el punto de vista de los usuarios, algunas de las ventajas de *Internet*, frente a otros medios de comunicación son su rapidez relativa, su bajo coste, el anonimato "relativo" de sus usuarios, su gran accesibilidad y su facilidad de utilización.

El crecimiento del *web* explica el gran número de organizaciones de todo tipo que durante la última década han comenzado a usar *Internet* como canal de comunicación, implantando sitios *web* a través de los cuales interactúan con sus clientes.

Estas ventajas tienen como contrapartida la dificultad para establecer relaciones "uno a uno" con los clientes, debido principalmente a la ausencia del contacto físico entre las partes. La naturaleza electrónica (y no física o directa) del contacto a través del *web* impide alcanzar fácilmente la mejor relación con el cliente y establecer efectivos vínculos bidireccionales, considerados fundamentales en el logro de comunicaciones eficaces. Si bien el *web* tiene variados ámbitos de aplicación, muchos de los problemas que se deben enfrentar son comunes a todos ellos y, por consiguiente, independientes del dominio particular del que se trate.

Internet como canal de comunicación no dispone de las características que permiten el establecimiento de una cálida relación con los usuarios. Para dotar al *web* de estas características es necesario explorar técnicas y métodos nuevos. Si bien se han propuesto enfoques tratando de solventar este problema, la mayoría se centran en descubrir la identidad del usuario y analizar su navegación. No obstante, para poder ofertar a cada usuario el servicio/producto apropiado a un costo adecuado es necesario (además de identificarlo) analizar en cada momento sus gustos, preferencias y estado de ánimo.

2.2 Relación con los clientes en el mundo *web*

La búsqueda de soluciones al reto de lograr relaciones "uno a uno" requiere tener en cuenta a los dos actores en el escenario del *web*: los usuarios propiamente dichos y quienes ofertan bienes y servicios a través del *web*:

Los usuarios se conectan a *Internet* para hacer uso de sus servicios. Se caracterizan por su diversidad en edad, en educación, intereses y nivel socioeconómico, entre otros. Como consecuencia, tienen muy variados requisitos de los sitios *web*, buscando obtener el mejor servicio en el menor tiempo y en el momento justo.

Los promotores ofertan bienes y servicios -en sentido amplio- a través del *web*, suministrando información, diversión y productos, entre otros. Como los

sitios *web* tienen diversos dominios de aplicación, los promotores poseen también variados objetivos a alcanzar, cada uno de ellos con una forma específica de medir el éxito. Debido a que el sitio *web* es un medio para que una organización alcance diferentes objetivos, es necesario considerar diferentes puntos de vista de la empresa promotora, que pueden estar asociados con variados significados de "utilidad", "beneficio" y "costo".

En cualquier actividad que se desarrolla en el *web*, al igual que fuera del *web*, el promotor procura lograr el máximo beneficio y el usuario busca obtener el mejor servicio al más bajo costo. En particular, es importante tener en cuenta que tanto promotores como usuarios se desenvuelven en un entorno competitivo, en el cual los promotores buscan identificar y atraer a los "mejores usuarios" desde su punto de vista, así como los usuarios buscan identificar, desde su perspectiva, a los "mejores promotores".

Para progresar en el desarrollo de las posibilidades del *web* y contribuir a que tanto usuarios como promotores satisfagan sus objetivos, es necesario identificar sus demandas, y evaluar en qué medida éstas se logran satisfacer. Lo anterior es capturar conocimiento no explícito del comportamiento de los usuarios y promotores, para obtener ventajas comparativas (respecto a otros sitios *web* y otras formas de relación con los usuarios).

Las fuentes de datos de que se dispone son: ficheros de registros (*log*) de los servidores *web* que contienen todos los accesos de los navegantes (*clickstream*), ficheros de transacciones que contienen la información detallada de las realizadas, información de la estructura del sitio *web* y del contenido de las páginas que lo componen y, para finalizar, toda la información de la operativa de la organización que está desarrollando su actividad en el *web*.

2.3 Web Mining

El objetivo fundamental del *Web Mining* es la detección de patrones desconocidos y potencialmente útiles en los datos del *web*, con el propósito de mejorar la toma de decisiones con relación al diseño, contenido y estructura de los sitios *web*.

La mayoría de las investigaciones se centran en el análisis de la navegación, extrayendo caminos frecuentes, segmentos (*clustering*) de usuarios, asociaciones de páginas visitadas en la misma sesión de navegación. Habitualmente incluyen escasa o nula información de negocio. Por consiguiente es reducida la posibilidad de valorar si una sesión, usuario o tipología influyen (positiva o negativamente) en el logro de los objetivos de la organización.

Si bien los enfoques tradicionales están centrados en el usuario/navegante y procuran satisfacer sus necesidades, existe un reciente interés en la consideración del punto de vista de las compañías. Últimamente ha tomado relevancia la necesidad de mejorar la efectividad de los sitios *web* desde el punto de vista de sus promotores, que fue iniciado por Gomory et al. [Gomory, 1999].

Los primeros proyectos *web* no tuvieron en cuenta que en una sociedad globalizada y competitiva como la contemporánea, el éxito de un sitio *web* depende tanto de la satisfacción de sus usuarios como de la de sus promotores. De esta forma, se fue generando una separación entre el gran desarrollo de la tecnología de la información asociada al *web* y el menor énfasis en los aspectos empresariales. La reseñada disociación de enfoques se refleja en criterios diferentes de éxito de un sitio *web*. Por otra parte, si bien se han planteado requisitos en el sentido de evaluar los sitios *web* desde el punto de vista empresarial o de negocio, no se han propuesto aún formas concretas para dicha evaluación.

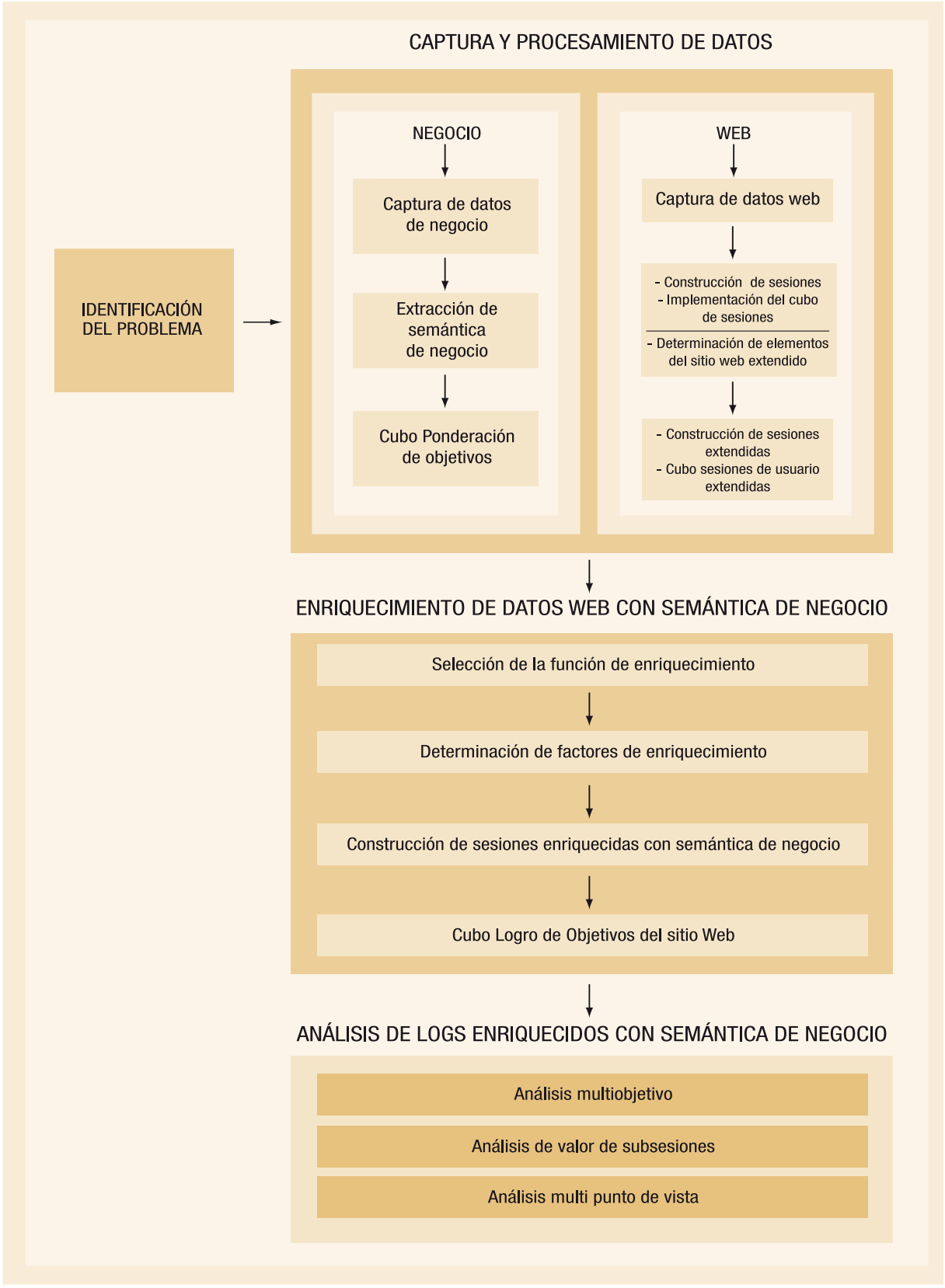
Las grandes cantidades que en los últimos años fueron invertidas en *software* de comercio electrónico, no prestaron demasiada atención al Retorno de la Inversión (ROI - *Return On Investment*). Esto fue debido, fundamentalmente, a que la mayoría de los criterios de éxito utilizados en la evaluación de sitios *web* se centran en el usuario-navegante y se basan en la información de los registros de los servidores *web* (*web server logs*). Sin embargo, los datos ofrecidos por los registros de los servidores *web* no son suficientes para:

- Evaluar el éxito de un sitio *web*.
- Descubrir conocimiento para cuantificar el retorno de la inversión.
- Adoptar decisiones para detectar eventos críticos del punto de vista de negocio.
- Incentivar a los usuarios para que se conviertan en clientes.
- Determinar el valor de los usuarios para el sitio *web*.

Kohavi et al. [Kohavi, 2004] consideran que puede resultar llamativo que no se hayan incorporado explícitamente, en la evaluación de los sitios *web*, los objetivos comerciales de las empresas promotoras, tales como aumentar el volumen de ventas, reducir los

FIGURA 1

Módulos de Web Goal Mining



fraudes, retener clientes o fijar precios competitivos. En el mundo actual no es posible prescindir de aspectos económicos y financieros para evaluar cualquier tipo de actividad, y en particular las desarrolladas en el ámbito de *Internet*.

Actualmente la mayoría de los trabajos evalúan el éxito de un sitio *web* midiendo su eficiencia y calidad. La eficiencia se mide por el número de páginas a las que accede un usuario durante una sesión, la duración de las sesiones, las acciones desarrolladas por los usuarios (comprar, consultar, salir) entre otros. La calidad del sitio se mide teniendo en cuenta el tiempo de respuesta del sitio *web* ante las acciones del usuario, la accesibilidad de las páginas *web* o del propio sitio *web*, la cantidad de páginas promedio a que accede un usuario/navegante y el número de visitantes por página, entre otros. Por otra parte, el éxito de una compañía se evalúa mediante indicadores como ingresos, egresos, rentabilidad, retorno de la inversión, volumen de ventas y prestigio en el mercado, entre otros.

La diferencia existente entre los criterios utilizados para determinar si la inversión en un sitio *web* es exitosa y los utilizados para evaluar los logros de los proyectos de inversión en otros ámbitos (fuera del *web*) son fuente de sorpresa, desilusión y desencanto por parte de ejecutivos de divisiones comerciales de las empresas. Las propuestas recientes tienden a establecer un nexo (o puente) entre el éxito evaluado desde la perspectiva de uso, contenido y estructura del sitio *web* y el éxito considerado como logro de los objetivos empresariales de la compañía que promueve el sitio *web*.

2.4 Enfoque de negocio de un sitio web

Si bien la satisfacción de usuarios/navegantes planteada anteriormente constituye un prerrequisito para el éxito del sitio *web* desde el punto de vista empresarial, no lo asegura por sí sola. Coincidimos con Mobasher et al. [Mobasher, 2001] al plantear que "el objetivo del sitio *web* no es hacer felices a los usuarios, es contribuir al éxito de la empresa". Podría ocurrir que los usuarios se sientan a gusto al navegar en un sitio de comercio electrónico, permanezcan largo tiempo, consulten por muchos productos, utilicen las recomendaciones, etc. y sin embargo no necesariamente compren. Entonces, podría concluirse que lo utilizan esencialmente como medio de información sobre los productos que posiblemente compren después en otros sitios *web*. En este caso, el sitio no debería ser considerado exitoso, a pesar de contar con usuarios satisfechos.

Tradicionalmente se han utilizado los datos de *clickstream* para evaluar los logros de los sitios *web*. Schmitt et al. [Schmitt, 1999] plantean que usar solamente hits y vistas de páginas para juzgar el éxito de un sitio es como "evaluar la calidad musical a través de su volumen".

Como consecuencia de utilizar exclusivamente la métrica de *clickstream*, los análisis de sitios *web* brindan información de las acciones de los navegantes. No obstante, no distinguen entre navegantes y clientes (compradores), ni analizan y caracterizan explícitamente a los compradores o el importe de las compras efectuadas. Kohavi et al. [Kohavi, 2001] plantean que "si el objetivo de la compañía es aumentar las ventas, el sitio *web* requiere atraer a compradores más que a navegantes".

A pesar de la enorme cantidad de datos relativos a la red almacenados por los servidores vinculados a su uso, contenido y estructura, es difícil comprender las relaciones entre datos de navegación de usuarios y la efectividad de los sitios. Tampoco es clara la forma de

FIGURA 2
Ponderación de objetivos

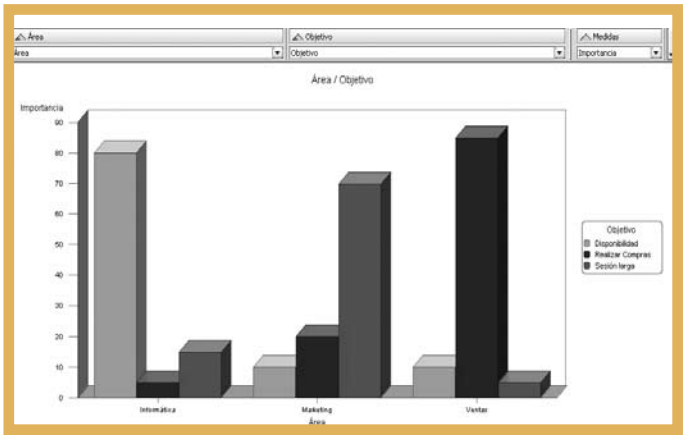
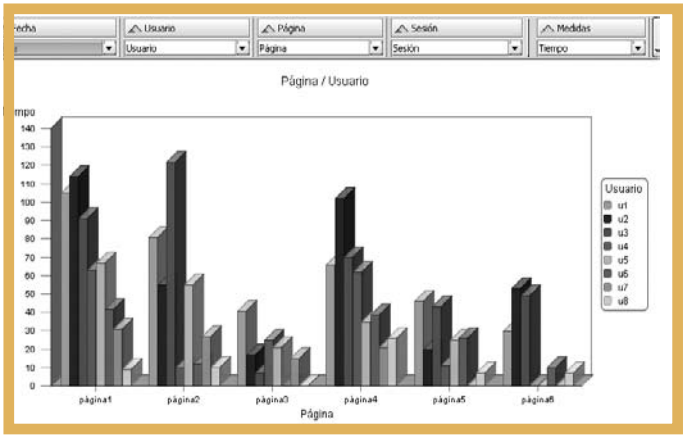


FIGURA 3
Cubo de sesiones



utilizarlos para el diseño de "buenas páginas" en términos del logro de los objetivos de los usuarios y de los promotores de los sitios.

Las principales tendencias en técnicas de análisis de negocio (*business analytics*) se centran en reducir la brecha entre *Data Mining* y los usuarios de sus resultados. El consumidor clave de las técnicas de análisis de negocio es el usuario de negocio (*business user*), caracterizado como una persona cuyo trabajo no está directamente relacionado con el análisis per-se (de las disciplinas de administración, contabilidad, *marketing*), que debe utilizar herramientas analíticas para mejorar los resultados de procesos de negocio a lo largo de una o varias dimensiones, tales como beneficios y permanencia en el mercado. Se espera que utilice la información extraída para mejorar el desempeño en base a múltiples métricas. Sin embargo, existe una brecha entre los análisis que se les proveen y las necesidades de los usuarios de negocios. Se requiere, por otra parte, una clara definición de objetivos de negocio y métricas, porque en el pasado existieron expectativas desmedidas relativas a un *Data Mining* mágico orientado a esfuerzos no guiados, sin claros objetivos y métricas.

Para reducir dicha brecha generalmente se identifica claramente "qué" se debe hacer, pero no se presenta una metodología clara que muestre "cómo" hacerlo, y en definitiva no se brinda una alternativa para la resolución del problema tan claramente explicitado.

3. WEB GOAL MINING

3.1 Conceptos básicos

Web Goal Mining es el proceso no trivial de descubrimiento de conocimiento válido, novedoso, comprensible y potencialmente útil que integra los objetivos de una organización, datos de navegación del sitio *web* y datos adicionales (de usuario, estructura y contenido del sitio) de tal manera que se puedan evaluar los logros que la empresa se propone alcanzar a través de su presencia en el *web*.

Web Goal Mining se subdivide en tres categorías en función de sus objetivos:

- *Web Usage Goal Mining* es el proceso de extracción de patrones de uso del *web* que muestra el vínculo entre la navegación y las metas de la empresa promotora del sitio *web*.
- *Web Content Goal Mining* es el proceso de identificación de patrones que vinculan el contenido de un sitio *web* y su entorno con las metas de la empresa promotora del mismo.
- *Web Structure Goal Mining* es el proceso de extracción de patrones de la topología del *web* que vinculan la estructura del *web* y las metas de la empresa promotora de un sitio *web*.

3.2 Proyecto de Web Goal Mining

Un proyecto de *Web Goal Mining* integra los objetivos de las unidades administrativas y funcionales de la empresa al descubrimiento y análisis de información útil del *web* para proporcionar conocimiento en términos del negocio. En el caso de estudio de este trabajo, *Marketing*, Informática y Ventas constituyen las tres unidades administrativas de la empresa vinculadas con el proyecto.

FIGURA 4A

Cubo sesiones extendidas

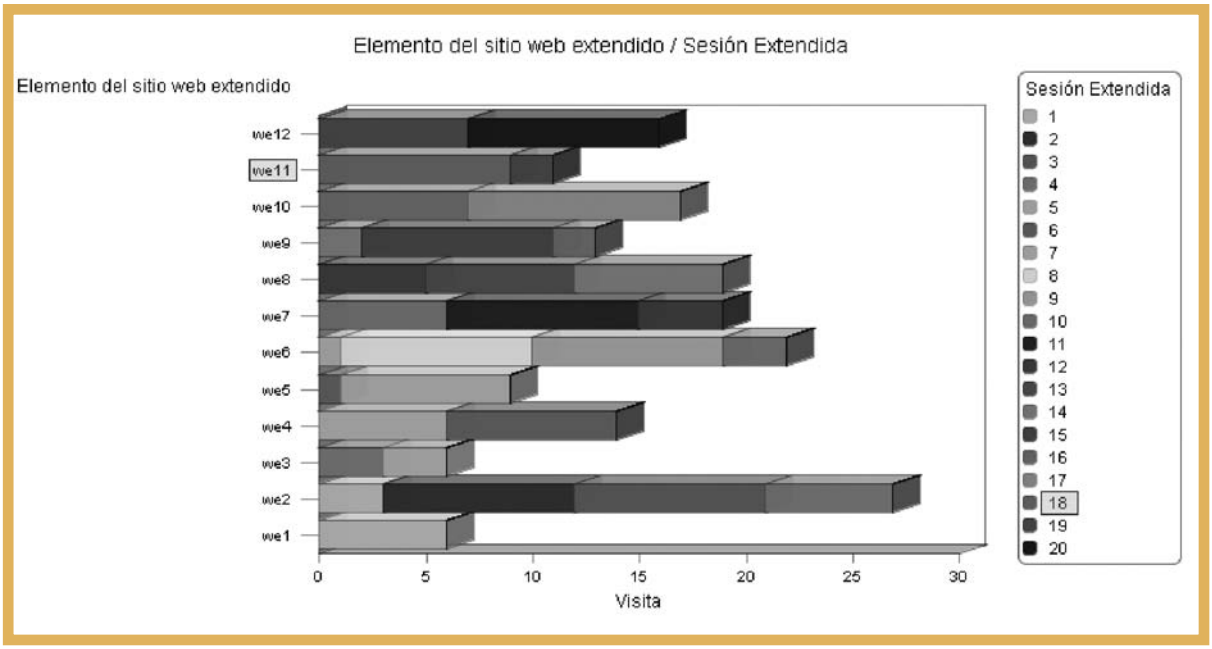


FIGURA 4B

Diseño de la tabla de hechos de sesiones extendidas

Un proyecto de *Web Goal Mining* integra los objetivos de la empresa al proceso de *Web Mining* y proporciona conocimiento del *web* adaptado específicamente al negocio de la organización. Su principal característica es que genera resultados comprensibles y directamente aplicables por los usuarios, es decir, los directivos de todas las áreas de negocios de la organización.

Para lograrlo, en el proceso propuesto se considera que los objetivos de la empresa son parte de los datos, y por tanto se los captura, analiza e integra al proceso de descubrimiento de información del *web*. Por lo tanto, un proyecto de *Web Goal Mining* se diseña para generar información útil en términos de negocio integrando los objetivos de las empresas promotoras de sitios *web*.

4. MARCO CONCEPTUAL DE WEB GOAL MINING

El marco conceptual (*Framework*) consta de los módulos que se observan en la Figura 1.

4.1 Identificación del problema

Esta etapa consiste en definir los objetivos del negocio, cuáles son las metas para definir el sitio *web*. Se centra en comprender el problema de negocio y transformarlo en un problema de *Web Mining*. Dicho en otros términos, implica evaluar la presencia de la organización en el *web* y el impacto de los navegantes/clientes sobre el negocio de la empresa.

Entrevistando a los responsables de las tres áreas involucradas se encontró que tenían diferentes criterios para caracterizar el éxito de una sesión:

- Para Ventas: se considera que una sesión es exitosa si se realizó alguna compra.
- Para Informática: se considera que una sesión es exitosa si todas las páginas del sitio *web* requeridas estaban disponibles.
- Para Marketing: se considera que una sesión es exitosa si la duración de la sesión fue mayor a cinco minutos.

El problema identificado es que diferentes áreas de la empresa tienen distintos objetivos, y que se necesita una metodología que los reúna manteniendo a su vez su independencia.

4.2 Captura y preprocesamiento de datos

Consiste en recolectar los datos y ponerlos en un formato adecuado. Se divide en dos procesos que se describen a continuación:

- Captura y preprocesamiento de datos de negocio.
- Captura y preprocesamiento de datos *web*.

4.2.1 Captura y preprocesamiento de datos de negocio

Brinda herramientas formales para comprender el problema de negocio planteado con la presencia de la compañía en el *web* y transformarlo en un problema de *Web Mining*. Su fin es comprender, identificar y ponderar los objetivos que busca la compañía con su presencia en el *web* así como identificar las unidades y procesos de negocio de la empresa vinculados con la presencia en el *web*. En nuestro caso de estudio el objetivo es incrementar las ventas globales de la empresa, al incorporar el comercio electrónico como un nuevo canal de ventas.

Consta de las siguientes etapas:

• Captura de datos de negocio: La captura de datos de negocio abarca la identificación de puntos de vista (*view points*) que comprenden tanto a la estructura (formal e informal) de la organización como a los procesos de negocio. Asimismo se identifican los objetivos que persiguen en cada uno de los puntos de vista. Como ya se mencionó, en nuestro caso de estudio los puntos de vista identificados fueron: Ventas, Marketing e Informática. Cada uno de ellos otorga cierta importancia a cada objetivo de la organización.

• Extracción de semántica de negocio: La extracción de semántica de negocio abarca la determinación de criterios de éxito utilizados para evaluar el logro de cada objetivo y el establecimiento del peso que posee cada objetivo de la organización en cada una de las unidades organizativas. Los criterios de éxito son los siguientes:

• Para Ventas, que el resultado de la sesión termine en la compra de un producto.

• Para Informática, que el sistema esté disponible.

• Para Marketing, que las sesiones sean largas, de modo que demuestren interés en el sitio por parte de los usuarios.

En conversaciones con los directivos de la empresa, se planteó que si bien todos los objetivos de los distintos puntos de vista eran importantes, debería asignarse en el estudio un peso mayor al objetivo propuesto por el departamento de Ventas.

• Implementación del cubo de ponderación de objetivos: Como resultado de esta fase se construye un cubo, como se muestra en la Figura 2, que permite observar la importancia asignada a cada objetivo por las diversas áreas de la empresa.

4.2.2 Captura y preprocesamiento de datos *web*

Su propósito es comprender, identificar, obtener, representar y preparar los datos del sitio *web* de la organización y de su entorno. Se divide en captura de datos *web* y construcción de sesiones.

Captura de datos *web*: Los datos del sitio *web* están conformados por datos de uso del sitio *web*, del contenido de las páginas, la estructura de los enlaces y de la semántica (es decir, los conceptos asociados).

El enfoque semántico no se interesa en descubrir patrones de acceso a URLs, sino patrones de acciones tales como comprar o mostrar interés. La diferencia entre ambos queda clara al considerar que en el enfoque centrado en patrones interesan patrones del esti-

lo If http://www.theshop.com/show.html-?item=123, then http://www.theshop.com/show.html?item=456'. Por el contrario, los anteriores no resultan de interés en el enfoque semántico, sino que se busca transformarlos en patrones del tipo "Los usuarios que compraron 'Hamlet' también compraron 'Cómo dejar de preocuparse y comenzar a vivir'" [Berendt, 2003].

Construcción de sesiones, que se basa en las siguientes definiciones:

Sesión de usuario: Una sesión de usuario es el *clickstream* de vistas de páginas para un único usuario en el recorrido del *web*.

Clickstream es una secuencia de solicitudes de vistas de página. No siempre el servidor brinda información suficiente para reconstruir el *clickstream* completo de un sitio *web*, porque las vistas de página a las que se accede desde un servidor del cliente o *proxy* no se ven desde el servidor.

Usuario: Un usuario se define como un único individuo que accede a ficheros desde uno o más servidores *web* a través de un navegador. Si bien esta definición puede parecer trivial, en la práctica resulta difícil identificar usuarios que acceden desde diferentes máquinas, o utilizan más de un agente en una única máquina.

Vista de página: Una vista de página (*page view*) está formada por cada uno de los ficheros que integran la presentación en el navegador del usuario en un cierto momento. Las vistas de página habitualmente se asocian con una única acción del usuario, la cual puede estar integrada por varios ficheros, tales como marcos, gráficos y ejecución de programas (*scripts*). Toda la información requerida para determinar los ficheros específicos que constituyen una vista de página se encuentra en el servidor *web*.

- *Implementar el cubo de sesiones:*
  En el cubo de sesiones, que se presenta en la Figura 3, se almacenan la fecha, usuario, páginas requeridas, características de la sesión y tiempo (duración) de la sesión.
- *Determinación del sitio web extendido:*
  La definición tradicional de sitio *web* no es suficiente

FIGURA 5

Proceso de enriquecimiento de logs con semántica de negocio

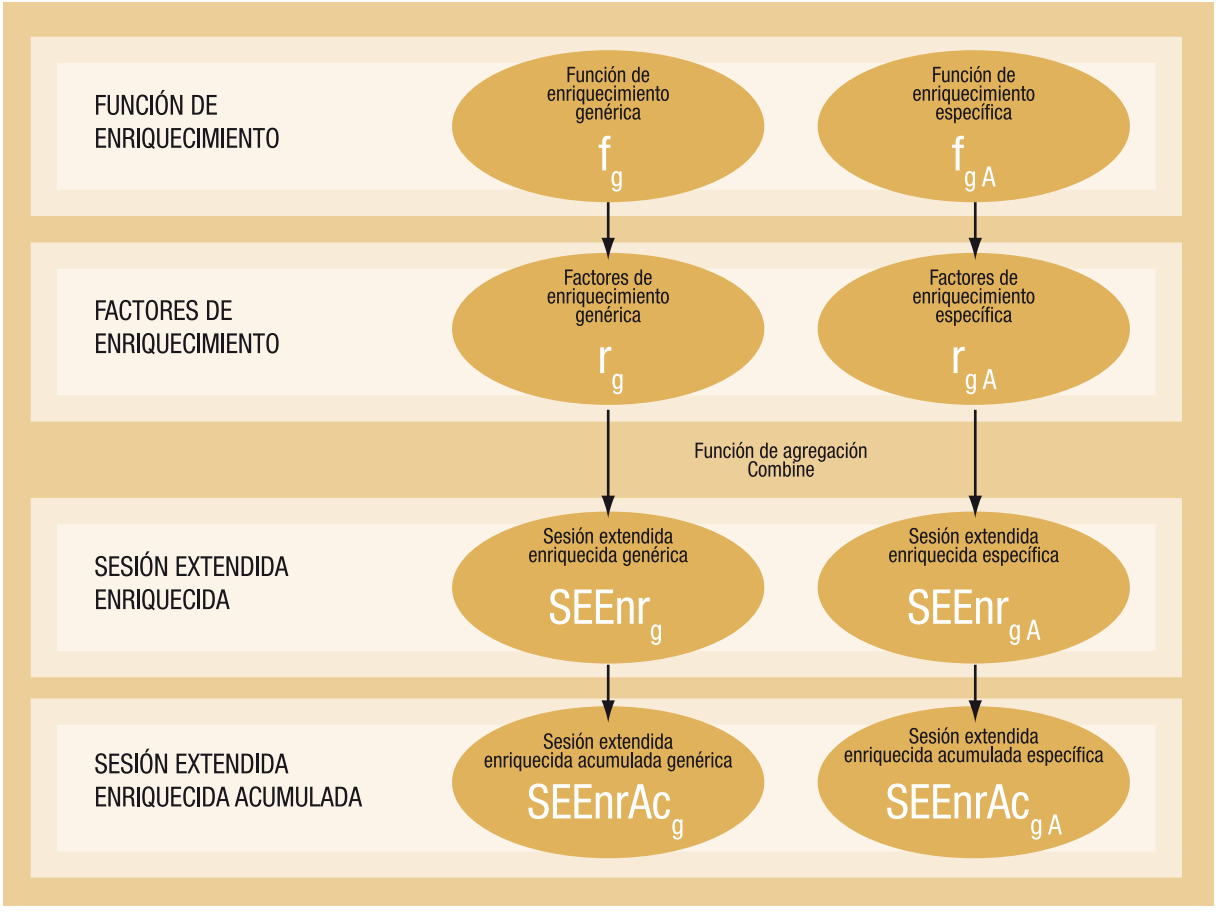


FIGURA 6
Ponderación de objetivos

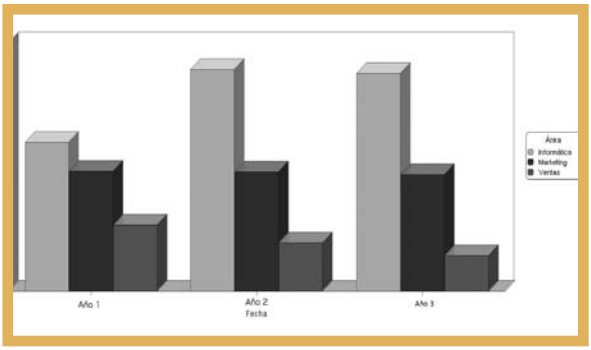
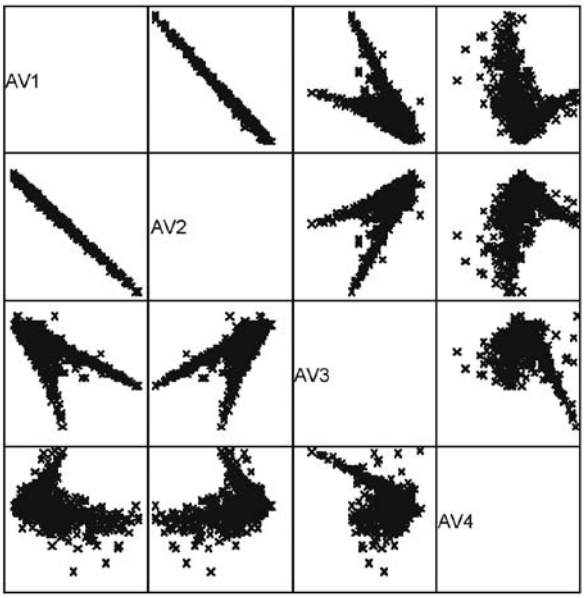


FIGURA 7
Matriz de dispersión



para analizar el desempeño de un sitio *web* en términos de negocio. Por tal motivo se incorpora en este trabajo el concepto de sitio *web* extendido. Un sitio *web* extendido es el conjunto formado por las páginas del sitio (estáticas y dinámicas), enlaces y los conceptos asociados a los elementos anteriores, así como cualquier otro aspecto relevante para representar el uso, contenido y estructura del sitio *web*.

- *Construcción de sesiones extendidas:*
  Con el objetivo de analizar el comportamiento del usuario, se considera que el nivel básico de abstracción es la sesión de usuario extendida (o simplemente sesión extendida), que se define en base a los elementos del sitio *web* extendido y a los datos del usuario. Una sesión extendida es una secuencia de elementos del sitio *web* extendido asociados a los elementos visitados por el usuario durante el transcurso de una sesión.
- *Implementación del cubo de sesiones extendidas:*
  En la Figura 4A se presenta el resultado de una consulta OLAP sobre el cubo de sesiones de usuario extendidas, en que se observan los elementos del sitio *web* extendido (*we*) contenidos en cada sesión. En la Figura 4B se presenta el diseño de la tabla de hechos, que corresponde al tipo denominado tabla de hechos sin hechos (*factless fact table*), [Tasistro, 2005].

4.3 Enriquecimiento de datos web con semántica de negocio

Pueden plantearse dos alternativas para establecer el enriquecimiento de registros:

- Que sean fijados directamente por los gerentes de la compañía, que están familiarizados con el dominio (y se asume que poseen el conocimiento necesario para determinarlos).
- Utilizar técnicas de *Data Mining* para establecerlos.

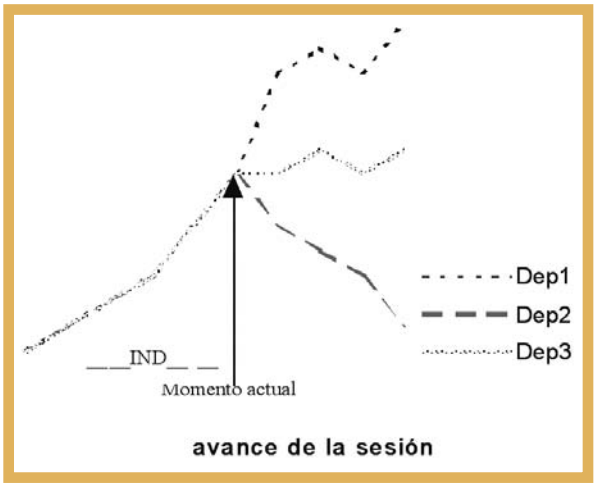
La primera, si bien es viable requiere que se determine en forma manual la función que desempeña cada una de las páginas del sitio *web* en la consecución de cada objetivo de la empresa promotora. Esto ha sido usado para determinar páginas objetivo y de acción considerando una única meta de la empresa. Sin embargo, hasta en el caso más simple planteado (un único objetivo) constituye una tarea larga, difícil aun para gerentes experimentados y en cierta forma subjetiva porque podrían obtenerse resultados diferentes si dos personas distintas realizan la tarea.

Se propone en este trabajo aplicar técnicas de *Data Mining* para establecer el enriquecimiento de registros en forma sencilla, rápida y basada en criterios objetivos. Asimismo, resulta de interés señalar que ante un problema análogo al planteado aquí, consistente en la determinación de probabilidades en cadenas de Markov, en un trabajo de Etzion et al. [Etzion, 2004] al igual que en el nuestro, se usan técnicas de *Data Mining* y se descarta la determinación manual.

En la fase de enriquecimiento de registros se construye un conjunto de datos, que denominamos registros enriquecidos, que permite predecir el logro de los objetivos en función de los elementos que conforman el sitio *web*. Se aplican técnicas supervisadas de *Data Mining* considerando como variable explicada la consecución de un objetivo de la empresa (evaluado con una métrica específica). Los elementos del sitio *web*

FIGURA 8

Valor de subsesiones futuras



extendido conforman las variables explicativas del modelo [Hochsztain, 2003].

En términos formales se plantea que el proceso de enriquecimiento de registros consiste en predecir la variable gA en función de los elementos we del sitio *web* extendido, así como determinar el aporte de cada elemento we del sitio *web* extendido al valor global de gA.

Como muestra la Figura 5, el proceso de enriquecimiento de registros (*logs*) está conformado por las etapas: selección de la función de enriquecimiento, determinación de factores de enriquecimiento, determinación de sesión extendida enriquecida y por último determinación de sesión extendida enriquecida acumulada.

Con los datos obtenidos como resultado de esta etapa se construye el cubo de logro de objetivos, que se presenta en la Figura 6. En el mismo se observa que Informática presenta los mayores logros, seguida de Marketing. En cambio, los menores logros se registran en Ventas.

4.4 Análisis de logs enriquecidos con semántica de negocio

Utilizando como datos de entrada los *logs* enriquecidos con semántica de negocio generados, es posible desarrollar un análisis multiobjetivo del valor de subsesiones y de valor global.

4.4.1 Análisis multiobjetivo

El análisis conjunto del logro de múltiples objetivos se realiza mediante una matriz de diagramas de dispersión, que muestra conjuntamente varios diagramas de dispersión. Considerando cuatro objetivos, se

FIGURA 9

Histograma del valor global de las sesiones para el punto de vista 1

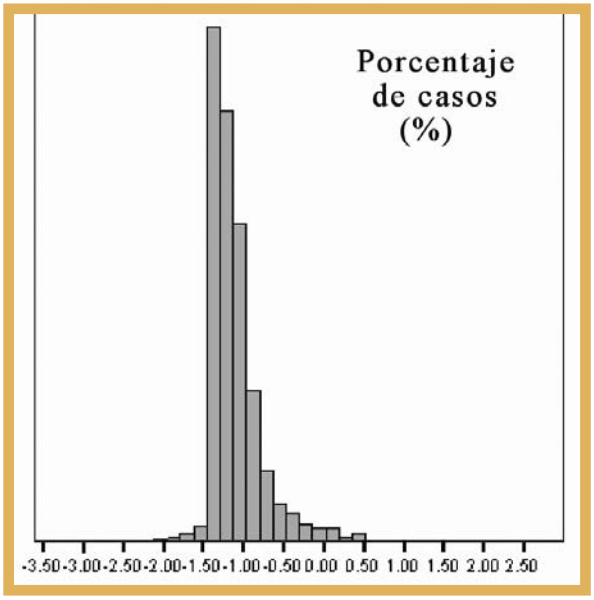
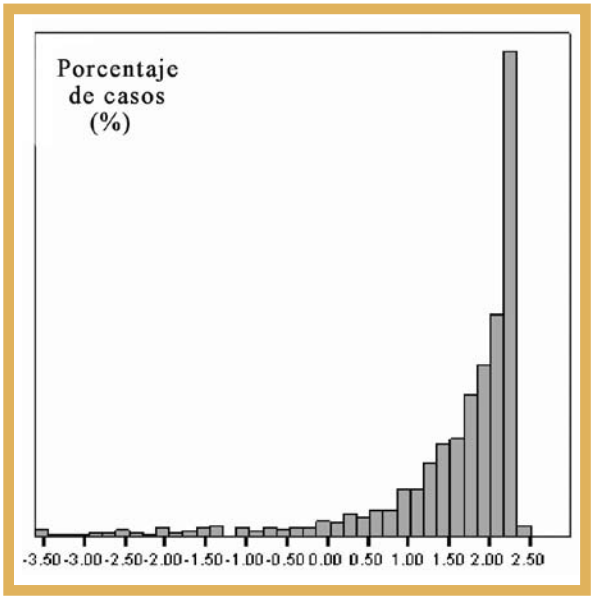


FIGURA 10

Histograma del valor global de las sesiones para el punto de vista 2



denota por AVi (i = 1, . . . , 4) al valor promedio de la sesión para el objetivo i, de modo que AVi representa el valor promedio de las sesiones para el objetivo 1, AV2 para el objetivo 2, y así sucesivamente.

Para comparar el valor en relación a cuatro objetivos son necesarios seis diagramas de dispersión. Para mos-

trarlos en forma conjunta se construye una matriz de diagramas de dispersión, que es una matriz simétrica que permite representar varios diagramas de dispersión a la vez entre diferentes pares de variables. Cada celda de la matriz es un plot con dos variables y en la diagonal tenemos identificadas las mismas.

Por lo tanto, una matriz de diagramas de dispersión permite contrastar el valor promedio de las sesiones para diferentes objetivos (ver Figura 8). Por ejemplo, la primera fila y la primera columna cruzan el valor de las sesiones para el objetivo 1 con el valor promedio para los restantes. De este modo, en el cruce de la primera fila y la segunda columna se presenta el vínculo entre los valores promedio de las sesiones para los objetivos 1 y 2, y en el cruce de la segunda fila y la primera columna se muestran (con los ejes invertidos) los mismos valores.

El análisis del gráfico permite observar que los valores de las sesiones para los objetivos 1 y 2 son totalmente opuestos. Asimismo los valores promedio de las sesiones para los objetivos 2 y 3 muestran que cuando las sesiones presentan valor promedio alto para el 2 también tienen valor promedio alto para el objetivo 3 y las sesiones con bajo valor para el objetivo 2 tienen también valores bajos para el objetivo 3. Por consiguiente los valores de las sesiones de los objetivos 2 y 3 son similares, y los valores de las sesiones para los objetivos 1 y 2 son muy distintos.

4.4.2 Análisis de valor de subsesiones

En el algoritmo de valor esperado de subsesiones se combinan las reglas de comportamiento frecuente planteadas con el valor de una sesión propuesto en este trabajo [Hochsztain, 2005]. Como ejemplo en la Figura 9 se observa que en determinado momento el usuario visitó la subsesión IND y se encuentra en una página de decisión. Se presentan tres caminos frecuentes representados por Dep 1, Dep 2 y Dep 3, de los cuales Dep 2 reduce el valor y los dos restantes lo aumentan. El valor esperado de las subsesiones posteriores a IND permite reunirlos y ponderarlos por sus respectivas probabilidades.

El algoritmo de valor esperado de una subsesión debe ejecutarse en línea cada vez que el navegante visita una nueva página. La aplicación del algoritmo de reglas de comportamiento frecuente permite determinar los caminos más probables que tomará el usuario a partir de dicha página. Posteriormente se determina el valor de los caminos futuros que puede tomar el usuario, que se denominan subsesiones frecuentes. Se consideran subsesiones frecuentes a los caminos posteriores que pueden ocurrir con una probabilidad que supera cierto umbral. El valor esperado de las subsesiones futuras es una medida de resumen del valor futuro de la sesión a partir del momento presente.

En el transcurso de una sesión de usuario se activan reglas de comportamiento frecuente que superan umbrales prefijados de soporte y confianza. El valor esperado de una subsesión se calcula como la esperanza matemática del valor de las subsesiones dependientes. Permite predecir el impacto en el logro de cierto objetivo. Dada una secuencia de páginas ya visitada o independiente el valor esperado se determina ponderando el valor de cada subsesión dependiente por la probabilidad condicionada de ocurrencia de la misma. Se asume que el número de probables caminos es finito, motivo por el cual los cálculos de probabilidades se basan en sumatorias.

4.4.3 Análisis multi-punto de vista

El análisis multi-punto de vista permite analizar el logro de todos los objetivos desde la óptica de varias unidades organizativas. Para aplicarlo se necesita la determinación del valor global de las sesiones para cada punto de vista, ponderando los objetivos de acuerdo a su peso para cada punto de vista.

En el caso de estudio el valor cero representa un ajuste neutro de la sesión con los objetivos de la empresa. En base a los conceptos de extracción de semántica del negocio se establecieron los pesos para los cuatro objetivos y dos puntos de vista, que dieron lugar a las ecuaciones, en donde GV1 y GV2 representan el valor global de las sesiones para los puntos de vista 1 y 2 respectivamente.

$$GV1 = AV1 \cdot 0{,}4 + AV2 \cdot 0{,}1 + AV3 \cdot 0{,}3 + AV4 \cdot 0{,}2$$

$$GV2 = AV1 \cdot 0{,}1 + AV2 \cdot 0{,}4 + AV3 \cdot 0{,}1 + AV4 \cdot 0{,}4$$

En la Figura 10 se presenta el histograma del valor global de las sesiones para el punto de vista 1, que muestra valores marcadamente negativos. En cambio, el histograma del valor global para el punto de vista 2 que se presenta en la Figura 11 muestra valores básicamente positivos. Por consiguiente, las sesiones analizadas resultan más favorables para el punto de vista 2 que para el punto de vista 1.

5. CONCLUSIONES Y TRABAJOS FUTUROS

En el presente trabajo se ha presentado el enfoque *Web Goal Mining*, que permite incorporar los objetivos de la compañía promotora del sitio *web* al proceso de *Web Mining*. Para ello es necesario integrar

semántica de negocio al proceso de *Web Mining*.

Para integrar los objetivos que se proponen las empresas con su presencia en el *web*, se ha desarrollado un *framework* o marco conceptual, cuyos módulos permiten observar las distintas fases del proceso.

De esta manera en el establecimiento de la semántica de negocio se ha tenido en cuenta la estructura y objetivos de las empresas y ha quedado clara la necesidad de disponer de un enfoque multicriterio y multiobjetivo. Asimismo, se ha tenido en cuenta el tipo de patrones que la organización quiere obtener, la naturaleza de los objetivos y requisitos subyacentes en los datos. Todo ello ha permitido constatar los aspectos diferenciadores de un proyecto de *Web Goal Mining* (orientado a satisfacer a los usuarios-promotores) con respecto a un proyecto de *Web Mining* (orientado a agradar a los usuarios/navegantes).

La incorporación de semántica se refleja en el enriquecimiento de *logs* con semántica de negocio. Los mismos permiten determinar la contribución de cada sesión de usuario en el sitio *web* al logro de los objetivos de la empresa promotora de ese sitio.

Si bien *Web Goal Mining* se tendrá que refinar con la incorporación de más técnicas de *Data Mining* a los datos del *web*, se puede afirmar que se dispone de un primer modelo de incorporación de semántica de negocio al proceso de *Web Mining*.

A partir de este trabajo se abren varios temas de investigación. Una línea interesante es lograr que automáticamente se propongan mejoras en la estructura del sitio de forma de contribuir al logro de los objetivos. Otra interesante línea de investigación que se abre a partir de este trabajo es el enfoque multicanal de comercio electrónico. Consiste en extender la propuesta para determinar técnicas de captura de datos de comunicación de la organización (cualquiera sea el medio de comunicación usado) e integrarlos con los datos de negocio para obtener patrones conjuntos de comportamiento.

De esta forma se podría comparar la eficacia y eficiencia de cada uno de los medios de comunicación para el logro de los distintos objetivos de la organización. También sería posible conocer el efecto del uso conjunto de dos o más medios de comunicación.

BIBLIOGRAFÍA CITADA

- [Berendt, 2003] B. Berendt, D. Oberle, A. Otto, J. González. Conceptual User Tracking. *AtlanticWeb Intelligence Conference Proceedings Springer AWIC 2003 LNAI 2663,* 2003.
- [Etzion, 2004] O. Etzion, A. Fisher, S. Wasserkrug. e-CLV: A Modelling Approach for Customer Lifetime Evaluation in e-Commerce Domains, with and Application and Case Study for Online Auctions. *IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04),* 2004.
- [Fayyad, 2007] U. Fayyad. The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007.
- [Gomory, 1999] S. Gomory, R. Hoch, J. Lee, M. Podlaseck, E. Schonger. E-commerce Intelligence: Measuring, Analyzing, and Reporting on Merchandising Effectiveness of Online Stores. Technical report, IBM Institute of Advanced Commerce, 1999.
- [Gordjin, 2003] J. Gordjin. Why visualization of e-business models matters. 16th e-Commerce Conference eTransformation panel Business models & the mobile industry: Concepts, Metrics, Visualization and Cases, 2003.
- [Hochsztain, 2003] E. Hochsztain, A. Tasistro, E. Menasalvas, S. Millán, M. S. Pérez. An Approach to Estimate the Value of User Sessions Using Multiple Viewpoint and Goals. Web Mining: From Web to Semantic Web. First European Web Mining Forum, EWMF 2003, Cavtat-Dubrovnik, Croatia.
- Revised Selected and Invited Papers (Lecture Notes in Computer Science) Springer-Verlag Serie Lectures Notes in Artificial Intelligence, 2003.
- [Hochsztain, 2005] E. Hochsztain, S. Millán, E. Menasalvas, M. Hadjimichael. Foundations of Data Mining and Knowledge Discovery: Studies in Computational Intelligence. Volume 6/2005. 2005.
- [Kohavi, 2001] R. Kohavi, S. Ansari, L. Mason, Z. Zheng. Integrating E-Commerce and Data Mining: Architecture and Challenges. In *ICDM'01: The 2001 IEEE International Conference on DataMining,* 2001.
- [Kohavi, 2004] R. Kohavi, L. Mason, R. Parekh, Z. Zheng. Lessons and Challenges from Mining Retail E-commerce Data. *Machine Learning Journal, Special Issue on Data Mining Lessons Learned,* 2004.
- [Mena, 2001] Interview with Jesús Mena. *http://www.kdnuggets.com/news/2001/n13/13i.html..*
- [Mobasher, 2001] B. Mobasher, B. Berendt, M. Spiliopoulou. KDD for Personalization. *Tutorial at the 12th European Conference on Machine Learning (ECML'01) / 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*
- [Schmitt, 1999] E. Schmitt, H. Manning, Y. Paul, J. Tong. *Measuring Web Success.* Forrester Report, 1999.
- [Tasistro, 2005] A. Tasistro, E Hochsztain. Diseño e Implementación de Bases de Datos y Data Warehouses. Comisión de Educación Permanente de la Universidad de la República, 2005.

AGRADECIMIENTO

A los responsables del sitio de comercio electrónico en base a cuyos datos se hizo este estudio, aunque por obvias razones no es posible divulgar datos que identifiquen a la empresa.