Resumen Tema 5: Minería de uso

José Alberto Benítez Andrades

Febrero 2011

En este trabajo se resumen las conclusiones obtenidas después de haber realizado la lectura de los artículos propuestos R. Cooley, B. Mobasher, and J. Srivastava. *Web mining and Pattern Discovery on the World Wide Web* ; J. Srivastava, R. Cooley, M. Deshpande, P. Tan. *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*; B. Mobasher. *Web Usage Mining and Personalization* ; Mike Perkowitz and Oren Etzioni. *Adaptive Web Sites: an AI Challenge* ; Thorsten Joachims, Dayne Freitag and Tom M. Mitchell. *Web Watcher: A Tour Guide for the World Wide Web.*

1. Definición y objetivos de minería de uso de la web.

Teniendo en cuenta las lecturas realizadas, como bien dicen *Jaideep Srivastava, Robert Cooley, Mukund Deshpande y Pang-Ning Tan,* podemos decir que la minería de uso de la web es el **proceso de aplicación de técnicas de minería de datos para el descubrimiento de uso de patrones desde datos Web**.

La velocidad con la que las transacciones son realizadas sobre una Web, se ha convertido en la llave principal en el crecimiento del comercio electrónico. La actividad de comercio electrónico se ha convertido en una revolución importante. Es increible la capacidad de poder comprar en una tienda sin necesidad de tener una persona que te atienda ni una tienda física que tenga que estar abierta, pudiendo comprar las 24 horas del día. También, el vendedor gracias al ecommerce, tiene la capacidad de mandar mensajes masivos o personalizados a sus compradores.

Son muchos los sitios dedicados al comercio electrónico o a proveer información. Estos sitios necesitan aprender cada día sobre los clientes o usuarios que navegan en sus sitios. Sólo de esta manera podrán dirigir adecuadamente los esfuerzos para mejorar los servicios de marketing y la personalización del sitio.

El descubrimiento de patronces de actividad y comportamiento relacionado con la navegación Web requiere el desarrolllo de algoritmos de Minería de Datos capaces de descubrir patrones de accesos secuenciales de ficheros log.

La Minería de Uso es una de las tres partes en las que se divide la Minería Web. Las otras dos restantes son Minería de Estructura y Minería de Contenido.

En otra de las lecturas, definen la Minería de uso como el descubrimiento automático de patrones de acceso de usuario desde los servidores Web. Las organizaciones coleccionan grandes cantidades de datos en sus operaciones diarias, generadas automáticamente por sus servidores web y colleciones de logs de acceso a servidor.

Analizan qué datos pueden ayudar a las organizaciones a determinar el tiempo de vida de sus compradores, cruzar estrategias de marketing en sus productos y realizar campañas promocionales, entre otras cosas.

Existen muchas herramientas de análisis web que tienen mecanismos para recoger informes sobre la actividad de los usuarios en los servidores y varios filtros de formularios de datos. Usando estas herramientas es posible determinar el número de accesos al servidor y a ficheros individuales, tiempo de las visitas y nombres de dominio a los que se accede. Sin embargo, estas herramientas son diseñadas para moderar el tráfico en los servidores y en muchas ocasiones no realiza análisis de las relaciones de datos o ficheros accedidos en el servidor.

2. Etapas de procesamiento

2.1. Preprocesamiento

El preprocesamiento consiste en convertir la información de uso, contenido y estructura obtenida de varias fuentes de datos disponibles en las abstracciones de datos necesarias para el descubrimiento de patrones.

Se divide 3 preprocesamientos:

2.1.1. Preprocesamiento de uso

El preprocesamiento de uso podría decirse que es la tarea más compleja en el proceso de Minería de Uso Web debido a que los datos disponibles suelen estar incompletos. A menos que se utilice un mecanismo de seguimiento de cliente, sólo la dirección IP, agente y seguimiento de servidor están disponibles para identificar usuarios y sesiones de servidor. Muchas de los problemas encontrados son:

- **Única dirección IP / Múltiples sesiones de servidor**: Por lo general los Internet Service Providres (ISPs) tienen un grupo de servidores proxy mediante el cual pueden acceder los usuarios. Un servidor proxy puede tener muchos usuarios accediendo a una Web en el mismo periodo.
- **Múltiples direcciones IP / Una sesión única de servidor**: Algunos ISPs o herramientas de privacidad aleatoria, asignan cada solicitud de un usuario a un grupo de direcciones IP. En este caso una única sesión de servidor puede tener múltiples direcciones IP.
- **Múltiples direcciones IP / Usuario único:** Un usuario que accede a la web desde diferentes máquinas puede tener diferentes IPs para cada sesión. Esto hace que el seguimiento de visitas cuente a ese usuario único como más de una visita.
- **Múltiples Agentes / Usuario único:** También, un usuario que usa más de un navegador, incluso en la misma máquina, puede aparecer como múltiples usuarios.

Asumiendo que cada usuario ha sido identificado (a través de cookies, logins o análisis de agentes o IP), el *click-stream* para cada usuario debe ser dividido en sesiones. Algunas solicitudes de páginas desde otros servidores no están disponibles a veces, es difícil saber exactamente cuándo un usuario ha abandonado la web. A menudo se suele evaluar teniendo en cuenta que más de 30 minutos de espera, significan que el usuario ya ha salido de la web. Cuando se inserta una ID de sesión en cada URL, la definición de una sesión es establecida por el servidor de contenido.

Mientras está disponible a menudo el contenido exacto servido como resultado de cada acción realizada por cada usuario, conocido gracias al log del servidor, esto se convierte muchas veces en algo necesario para tener acceso a la información de los servidores de contenido. Los servidores de contenido

pueden mantener variables de estado para cada sesión activa, la información necesaria para determinar qué contenido ha sido visto por el usuario, no está siempre disponible en la URL.

2.1.2. Preprocesamiento de contenido

El preprocesamiento de contenido consiste en convertir el texto, las imágenes, los scripts y otros ficheros multimedia en formas que son útiles para el proceso de Minería de Uso de la Web. A menudo, estos consiste en realizar minería de contenido como una clasificación o clustering. Mientras aplicamos minería de datos a los contenidos de la web, esto es un area de investigación interesante, en el contexto de la mineria de Uso de la Web el contenido de un sitio puede ser usado para filtrar la entrada o salida de los algoritmos de descubrimiento de patrones. Por ejemplo, los resultados de un algoritmo de clasificación pueden utilizarse para limitar los patrones descubiertos a estos contenidos de páginas vistas sobre un cierto tema o clase de productos. Además para clasificar o hacer cluster de páginas vistas basados en temas o tópicos, las páginas vistas pueden también clasificarse acorde con nuestras intenciones de uso. Las páginas vistas pueden ser destinadas a transmitir información, recopilar información del usuario, permitir la navegación, o algunas combinaciones de estos usos.

Los destinos de uso de una página vista pueden también filtrar las sesiones antes o despues del descubrimiento de patrones. En orden de comenzar los algoritmos de minería de contenido en páginas vistas, la información debe ser primeramente convertida en un formato cuantificable. Algunas versiones de modelos de espacios vectoriales se utilizan para complementar esto. Los ficheros de texto se pueden romper en vectores de palabras. Palabras clave o descripciones de texto pueden sustituirse por gráficos o multimedia. El contenido de las páginas vistas estáticas puede preprocesarse fácilmente por el análisis de HTML y reformateado de información o ejecutando algoritmos tradicionales.

Las páginas vistas dinámicas presentan más de un desafío. Los servidores de contenido que emplean técnicas de personalización y / o dibujan a bases de datos para construir las páginas vistas deben ser capaces de formar más páginas vistas que puedan preprocesarse. Un conjunto de sesiones de servidor dado puede acceder solamente a una fración de páginas vistas para un sitio dinámico. También el contenido puede ser revisado en bases regulares. El contenido de cada página vista para ser procesada debe estar reunido, o por una solicitud HTTP de un rastreador, o por una combinación de plantillas, script y acceso a bases de datos. Si solo una parte de una página ve que son accedidas son preprocesadas, la salida de cualquier algoritmo de clustering o clasificación puede ser sesgada.

2.1.3. Preprocesamiento de Estructura

La estructura de un sitio es creada por los enlaces de hipertexto entre páginas vistas. La estructura puede ser obtenida y preprocesada de la misma manera que el contenido de un sitio. Otra vez, el contenido dinámico (además de los enlaces) plantean más problemas que las páginas estáticas. Para cada sesión de usuario se tiene que crear una estructura de sitio diferente.

2.2. Inferencia de patrones (pattern discovery)

El descubrimiento de patrones se basa en métodos y algoritmos desarrollados desde muchos campos como estadísticas, minería de datos, aprendizaje automático y reconocimiento de patrones. Los métodos desarollados por otros campos deben estar en consideración de diferentes tipos de abstracciones de datos y bases de conocimiento disponibles para la minería web. Por ejemplo, en el descubrimiento de reglas de asociación, la noción de una transacción para el análisis de mercado no debe estar en consideración del orden en que los objetos son seleccionados. Sin embargo, en la minería de uso de la web, una sesión

de servidor es una secuencia ordenada de solicitudes de páginas por un usuario. Además, debido a la dificulta de identificar sesiones únicas, se requiere la base de datos adicional.

2.2.1. Análisis estadístico

Las técnicas estadísticas es el método más común para extraer base del conocimiento sobre los visitantes de una Web. Analizando los ficheros de sesiones, uno puede realizar diferentes tipos de análisis estadísticos descriptivos (frecuencia, significado, medio...) en variables como páginas vistas, viendo tiempo y longitud de un directorio de navegación. Muchas herramientas de análisis de tráfico web producen un informe periódico que contiene información estadistica como las páginas más visitadas frecuentemente, media de tiempo de visita sobre una página o media de longitud de un directorio de un sitio. Estos informes pueden incluir análisis de errores de bajo nivel limitados detectando puntos de entrada no autorizados o encontrando URIs no válidas. Además en la profundida de estos análisis, este tipo de bases de conocimiento pueden ser potencialmente útiles para mejorar la ejecución de los sistemas, aumentar la seguridad de los sistemas, facilitar las tareas de modificación del sitio y proveyendo soporte para decisiones de marketing.

2.2.2. Reglas de asociación

La generación de reglas de asociación puede ser utilizada para relacionar páginas que son más referenciadas juntas a menudo en una sesión única de servidor. En el contexto de la minería de Uso de la Web, las reglas de asocación hacen referencia a conjuntos de páginas que son accedidas juntas con un valor de apoyo superior a un umbral especificado. Estas páginas pueden no estar conectadas directamente a otras mediante hiperenlaces. Por ejemplo, el descubrimiendo de reglas de asociación usando el algoritmo Apriori puede revelar una correlación entre usuarios que visitan una página que contiene productos electrónicos y quienes acceden a una página sobre equipamiento deportivo. Aparte de ser aplicable para empresas y aplicaciones de marketing, la presencia o ausencia de estas reglas puede ayudar a diseñadores Web a reestructurar su sitio Web. Las reglas de asociación pueden también servir como métodos heurísticos para documentos en calidad de reducir la latencia por usuario cuando cargan una página desde un sitio remoto.

2.2.3. Clustering

Clustering es una técnica que agrupa juntos un conjunto de objetos que tienen características similares. En el dominio del Uso Web, hay dos tipos de clusters interesantes para ser descubiertos: clusters de uso y cluster de páginas.

Clustering de usuarios tiende a establecer grupos de usuarios exhibiendo patrones de navegación similares. Como la base del conocimiento es especialmente útil para inferir la demografía del usuario en orden de ejecutar segmentaciones de mercado en las aplicaciones de comercio electrónico o proveer contenido web personalizado para los usuarios.

Por otra parte, el clustering de páginas descubrirá grupos de páginas teniendo contenido relacionado. Esta información es útil para los buscadores de Internet y los proveedores de asistencia Web. En ambas aplicaciones, páginas de HTML dinámicas o estáticas pueden ser creadas para sugerir hiperenlaces relacionados para el usuario acorde con las consultas del usuario o el historial pasado de información necesitada.

2.2.4. Clasificación

La clasificación es la tarea de mapear un objeto de datos en una de las clases predefinidas. En el dominio Web, uno esta interesado en el desarrollo de un perfil de usuarios a lo largo de una clase o categoría particular. esto requiere extracción y selección de características que describan de la mejor manera las propiedades de una clase o categoría dada. La clasificación puede ser hecha usando algoritmos de aprendizaje inductivos supervisados como clasificación de árboles de decisión, clasificadores de ingenuos Bayesian, clasificadores de vecinos cercanos, Máquinas de Vectores de Soporte, etc. Por ejemplo, la clasificación en logs de servidores puede llevar a descubrir reglas importantes como: 30 % de los usuarios que estan online en el directorio /Product/Music tienen entre 18 y 25 añosy viven en la Costa Oeste.

2.2.5. Patrones secuenciales

La técnicas de descubrimiento de patrones secuenciales intenta encontrar patrones de inter-sesiones en los que la presencia de un conjunto de objetos es seguida por otro objeto en conjunto de sesiones o episodios en un tiempo ordenado. Usando este enfoque, los mercados Web muestran anuncios apuntando a ciertos grupos de usuario. Otros tipos de análisis temporales que pueden ser ejecutados en patrones de secuencia incluyen análisis de tendencias, detecciones de puntos de cambio o análisis similiares.

2.2.6. Dependencia de modelado

Modelar dependencia es otra tarea de descubrimiento de patrones en la Minería Web. El objetivo es desarrollar un modelo capaz de representar dependencias significantes entre las variables varias en el dominio Web. Como ejemplo, uno puede estar interesado en construir un modelo representando los estados diferentes se someten mientras compramos en una tienda online basada en acciones elegidas. Hay muchas técnicas de aprendizaje que pueden ser empleadas para modelaar la navegación de los usuarios. Como técnicas incluye Modelos Markov Hidden y Redes Belief Bayesian. Modelando los patrones de uso web no solamente proveeran un framework teórico para analizar el comportamiento de los usuarios pero es potencialmente útil para predecir el consumo de recursos web en un futuro. La informaión puede ayudar a desarrollar estrategias para incrementar las ventas de productos ofertados por el sitio Web o mejorar la navegación a conveniencia de los usuarios.

2.3. Análisis de patrones

El análisis de patrones es el último paso en el proceso de minería de uso de la web. La motivación detrás del análisis de patrones es filtrar reglas que no son interesantes o patrones de un conjunto encontrado en la fase de descubrimiento de patrones. La metodología de análisis exacto está usualmente dirigida por la aplicación por la cual la mineria web es realizada. La forma más común del análisis de patrones consiste en un mecanismo de consultas a una base de conocimiento como por ejemplo SQL. Otro método es cargar los datos de uso en un cubo de datos en orden de ejecutar operaciones OLAP.

2.4. Algunas herramientas existentes

3. Técnicas de aprendizaje aplicadas a minería de uso

En el texto *WebWatcher: A Tour Guide for the World Wide Web* de *Thorsten Joachims, Dayne Freitag y Tom Mitchell* se explican distintas técnicas de aprendizaje aplicadas a la minería de uso. ¿ Cuál es la forma que debe tener la base de conocimiento para el Web Watcher ? En general, esta tarea es sugerir un enlace apropiado dando un interés y una página Web. En otras palabras, Web Watcher necesita una base de conocimiento de la siguiente función de objetivo:

Calidad de Enlace : Página x Interés x Enlace -> [0,1]

El valor de la *Calidad de Enlace* es interpretado como la probabilidad de que un usuario seleccione *Enlaces* dada la *Página* actual y el *Interés*. Existen principalmente tres enfoques de aprendizaje. El primer enfoque usa guías previas como fuente de información para aumentar la representación interna de cada hiperenlace seleccionado. El segundo enfoque se basa en el aprendizaje por refuerzo. La idea es encontrar guías a través de la Web así como la cantidad de información relevante encontrada sobre la trayectoria que es maximizada. El tercer enfoque es el método que combina ambos enfoques anteriores.

3.1. Aprendizaje desde Giras Anteriores

En el primer enfoque, el aprendizaje está logrado por la anotación de cada hiperenlace con el interés de los usuarios quienes tocan estos hiperenlaces en visitas anteriores. Así, cada vez que un usuario siga un hiperenlace la descripción de el hiperenlace es aumentada añadiendo las palabras clave que los usuarios han escrito en el inicio de su visita. La descripción inicial de un hiperenlace es el texto subrayado. Para sugerir hiperenlaces durante una visita el WebWatcher compara el interes de los usuarios actuales con las descripciones de todos los hiperenlaces en la página actual. El WebWatcher sugiere estos hiperenlaces que tienen una descripción similar al interés de los usuarios.

La métrica usada para computar similitudes entre un estado de interés del usuario y una descripción de hiperenlaces está basada en una técnica desde recuperaciones de información. Descripciones de intereses e hiperenlaces son representadas por muchos vectores de característcias altamente-dimensionales, donde cada dimensión representa una palabra particular en el idioma inglés. El elemento (llamado peso de palabra) de un vector es calculado usando TFIDF. Basándonos en esta representación de similitudes vectoriales son calculadas como el coseno entre vectores.

El algoritmo WebWatcher usa sugerir hiperenlaces considerando todos los hiperenlaces de la página actual. Para cada hiperenlace, la lista de palabras asociadas se utiliza para calcular la similitud del interés del usuario actual. El valor de la Calidad del Enlace para cada hiperenlace es estimada siendo la media de similitud, k (normalmente 5), la mejor clasificada. Un hiperenlace es sugerido si este valor de calidad de enlace está por encima del umbral. El máximo número de hiperenlaces sugeridos por página es tres.

3.2. Aprendizaje desde Estructuras de Hipertexto

En el punto anterior describí un método de aprendizaje que auemntaba un hiperenlace dado con los estados de interés que los usuarios habían seleccionado anteriormente. En este punto, describiré un segundo método de aprendizaje que aumenta un hiperenlace dado usando palabras encontradas en páginas corrientes. Este enfoque está basado en el aprendizaje por refuerzo. El objetivo es encontrar directorios a través de la Web que maximicen la cantidad de información relevante encontrada.

Aprendizaje con refuerzo

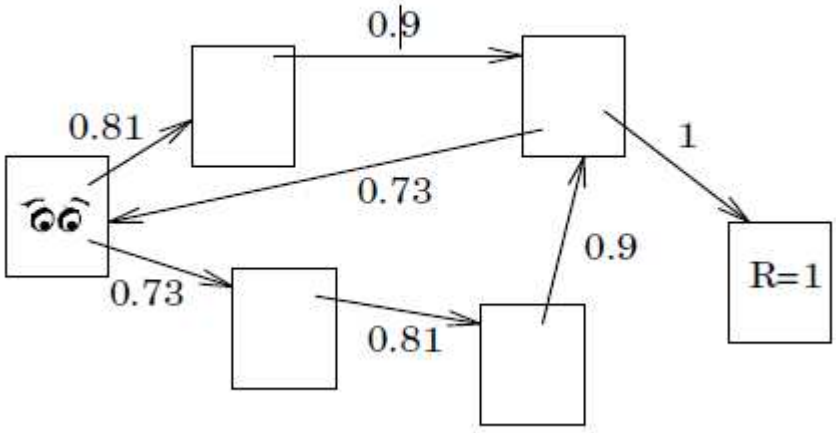
El aprendizaje con refuerzo permite a los agentes aprender estrategias de control que seleccionan acciones óptimas en ciertas configuraciones. Considermos a un agente navegando desde un estado a otro estado por acciones realizadas. En cada estado s el agente recibe una recompensa *R(s)*. Lo mejor de una acción a puede ser expresada en términos de una función de evaluación *Q(s,a)* , definidos para todos los posibles pares de estado-acción. Si el agente puede aprender esta función, entonces esto sabrá cómo actuar en cada estado. Más precisamente

$Q(s_t, a) = \sum_{i=0}^{\infty} \gamma^i \cdot R(s_t + 1 + i)$

donde s_t es el estado, el agente está en el tiempo t, y donde γ es el factor de descuento $0 \leq \gamma < 1$ que determina cómo descontar el valor de las recompensas recividas en un futuro. Bajo ciertas condiciones, la función Q puede ser iterativamente aproximado actualizando la estimación para Q(s,a) repetidamente como se muestra en la siguiente fórmula:

$Q_{n+1}(s, a) = R(s') + \gamma_{a' \in actions-in-s'}[Q_n(s', a')]$

s' es el estado resultante de la realización de la acción a en estado s. Un Q(s,a) es conocido, la estrategia de control óptimo para el agente es repetidamente realizar la acción a que maximice *Q(s,a)* para el estado actual s.



En la figura anterior, las cajas representas posibles estados del agente. Los ejes representan acciones que conducen al agente desde un estado a otro. Los ejes son anotados con valores de la función Q(s,a). El estado que está más a la derecha tiene una recompensa de 1. la recompensa es 0 en todos los otros estados. Si el agente siempre sigue la acción con el valor más alto de Q, este cogera el estado de recompensa en el número más bajo de pasos y esto maximizará el descuento recibido.

Aprendizaje reforzado e Hipertexto

Imagina un agente web buscando por páginas en las que se encuentra la palabra "inteligente". Para este agente, los estados corresponden a páginas web y las acciones corresponden a hiperenlaces. En este caso, nosotros definimos recompensa $R_{inteligente}(s)$ para una página particular s para ser el valor TFIDF de *inteligente* para s. El agente entonces aprenderá una función $Q_{inteligente}(s, a)$ cuyo valor

para páginas s e hiperenlaces a es la suma del valor descontado TFIDF de *inteligente* sobre la gira comenzada con a. Esto puede usar esta función Q para escoger el mejor enlace a cada paso en la gira.

WebWatcher usa una función separada de recompensa $R_w(s)$ y aprende una función de distinción $Q_w(s, a)$ para todas las palabras w. En ejecución el sistema recomienda hiperenlaces por los cuales la suma de $Q_w(s, a)$ para las palabras en las descripciones del interés de los usuarios es alta. Una recompensa adicional es dar si el interés encuentra el texto subrayado del hiperenlace.

Debido a que WebWatcher no puede esperar que los usuarios siempre se adhieran a páginas, una cuestión de núcleo en la implementación de este enfoque es cómo aprenden una aproximación general para cada una de las funciones-Q $Q_w(s, a)$ que aplican incluso a estados no vistos (páginas) y acciones (hiperenlaces). Cada hiperenlace a es descrito por el vector representación TFIDF del texto de anclaje subrayado, cada página s análogamente por su título. Describimos la similitud entre el hiperenlace $a_1$en la página s_1y el hiperenlace $a_2$en la página s_2 para ser la distancia entre a_1 y a_2, más (heurísticamente) veces la distancia entre s_1 y s_2. La distancia entre dos vectores está definida por ser el coseno del ángulo entre los vectores, de acuerdo con las medidas estándares de similitud usados en la recuperación de información.

4. Sitios web adaptativos

4.1. Definición y objetivos

En primer lugar, cabe destacar que la World Wide Web se está convirtiendo en el medio clave de distribución de información, entretenimiento y comunicación. Cada sitio web se crea con una serie de objetivos diferentes dependiendo de los usuarios a los que vayan dirigidos. El problema de los buenos diseños web se agrava por factores más allá de las distintas características que tienen los diferentes visitantes y sus respectivos objetivos. Primero, el mismo visitante debe poder buscar información diferente en varias ocasiones. En segundo lugar, muchos sitios crecen más allá de su diseño original, acumulando enlaces y páginas en lugares inadecuados. En tercer lugar, un sitio debe ser diseñado para un tipo de uso particular, pero ser usado de muchas formas distintas en la práctica; las espectativas de diseño iniciales pueden verse truncadas. También a menudo un diseño de un sitio web es bastante antiguo en HTML, mientras la navegación web es dinámica, dependiente del tiempo e ideosincráctica. Para dar solución a este problema, se crearon sitios web adaptativos: los sitios web que mejora automáticamente su organización y presentación aprendiendo de los patrones de acceso de los usuarios.

El diseño web es un problema en el diseño de la interface de usuario. Sin embargo, en contraste con vendedores de software empaquetado, pocos diseñadores de sitios web pueden darse el lujo en sus sitios web para formalizar la usabilidad probando con laboratorios especiales. Afortunadamente, los usuarios web interactúan directamente con un servidor mantenido por los inventores del servicio o autores del contenido que está siendo servido. Como resultado, los datos de su comportamiento son almacenados en los logs de los servidores. Estos datos en bruto son aplastantes para un webmaster con exceso de trabajo para procesar regularmente, los logs de servidores son utilizados para los análisis automáticos.

¿ Cómo podemos construir un sitio web que se mejore a lo largo del tiempo en respuesta a las interacciones de los usuarios con este sitio ? Este desafío posee un número difícil, pero no imposible de cuestiones:

- **¿ Qué tipos de generalizaciones podemos dibujar desde los patronces de acceso de usuarios y qué tipo de cambios podemos realizar ?** Si por ejemplo nosotros tenemos una tienda de material de oficina, y tenemos unos parámetros de visión de páginas por fabricante,

y observamos que las personas buscan muchos bolígrafos, pero dándoles igual la marca a la que pertenezcan, tendremos que hacer un nuevo enlace que permita ver únicamente los bolígrafos, dando igual el fabricante al que pertenezcan.

- **¿ Cómo podemos diseñar un sitio por adaptabilidad ?** Podemos ofrecer un mapa web en el cual el usuario pueda ver rápidamente todas las secciones que tiene la web. Otra idea sería crear un tour en la web o presentar la web como una base de datos.
- **¿ Cómo podemos colaborar con eficacia con un webmaster humano para sugerir y justificar posibles adaptaciones ?** Nuestro sistema puede acumular observaciones y sugerencias y presentárselas al webmaster.
- **¿ Cómo podemos llevar más allá de una sola vez algoritmos de aprendizaje a los sitios web que continuamente mejoran con la experiencia ?** Nuestros sitios web adaptativos con el tiempo acumularán un gran número de datos sobre nuestros usuarios y deberán usar su historial para evolucionar y mejorar.

El objetivo de crear sitios webs de auto-mejora es una tarea similar: una en cuya realización se requieren avances en diferentes áreas de la IA. Existen 2 enfoques para crear sitios web adaptativos.

4.2. Aproximaciones

Los sitios deben ser adaptativos en dos caminos principales. Primero, el sitio debe enfocar a la personalización: modificar páginas web en tiempo real para satisfacer las necesidades de los usuarios individuales. En segundo lugar, los sitios deben enfocar a la optimización: alterando los sitios para hacer la navegación más fácil para todos. En este trabajo explicaré dos aproximaciones o enfoques con ejemplos sobre la investigación de la IA. Si nosotros modificamos nuestras páginas web online o offline, nosotros debemos usar información sobre nuestros patrones de acceso de usuarios y la estructura de nuestro sitio. Mucha de esta información está disponible en los logs de acceso y en los sitios HTML, pero esto muchas veces no es suficiente: comentaré en este mismo trabajo cómo ayudar a la adaptabilidad con meta-información - información sobre los contenidos de las páginas.

4.2.1. Personalización

La personalización está ajustando las presentaciones de los sitios para los usuarios individualmente. La personalización permite la mejora de la "finura", desde el interface se puede completar a medida cada usuario individual. Un camino para el sitio para responder a visitantes particulares está el permitir personalizaciones manuales: permitiendo a los usuarios especificar las opciones que se quieren mostrar que han recordado durante su estancia en el sitio desde la primera visita hasta la siguiente. La red de Microsoft (como http://www.msn.com), por ejemplo, permite a los usuarios crear páginas iniciales con noticias e información personalizadas. Siempre, los visitantes pueden ver su página personalizada entren cuando entren, se guardan sus personalizaciones.

La predicción de ruta, por otra parte, personaliza automáticamente intentando adivinar dónde quieren ir los usuarios. Un sistema de predicción de ruta debe responder por lo menos a las siguientes cuestiones:

- **¿Qué estamos prediciendo?** Debemos intentar predecir los siguientes pasos de usuario. Por ejemplo, si nosotros podemos predecir qué enlace en una página va a seguir un usuario concreto,

nosotros debemos resaltar ese enlace o llevarlo a la parte más alta de la web para que sea visto. De forma alternativa, nosotros debemos intentar predecir los objetivos de los usuarios de forma eventual; si nosotros podemos determinar qué página en el sitio es buscada por unvisitante, podemos presentarla de manera inmediata.

- **¿ En qué nos basamos para hacer predicciones ?** Debemos usar acciones particulares para predecir lo que quiere cada usuario. Por otra parte, podemos generalizar para múltiples perfiles de usuario y así agilizar y dar la información de forma más rápida.
- **¿ Qué tipos de modificaciones podemos hacer en nuestras bases de nuestras predicciones ?** Podemos hacer tan poco como destacar los enlaces seleccionados o tanto como la síntesis de una página nueva que creemos que el usuario quiere ver.

Por ejemplo, el WebWatcher aprende a predecir qué enlaces van a seguir los usuarios en una página particularmente con una función que tiene en cuenta sus intereses específicos.

Un enlace que WebWatcher cree que te va a gustar, se suele resaltar de alguna manera gráficamente y se coloca en la parte superior de la web. Los visitantes de un sitio son encuestados para saber exactamente qué es lo que buscan. WebWatcher utiliza rutas de personas que indican sucesos como ejemplos de buena navegación.

En lugar de predecir las siguientes acciones, intentamos predecir los objetivos que tienen esos usuarios. Para ello, el reconocimiento de objetivos es uno de los problemas difíciles de identificar por una serie de acciones. Lesh y Etzioni dieron una posible solución a este problema basada en un framework de dominio-independiente. Ellos modelaron las acciones de usuario como un planning de operaciones. En el dominio web, nosotros observamos la navegación de un usuario a través de nuestro sitio e intentamos determinar qué páginas busca realmente.

¿ Es posible formalizar la navegación de usuario como un planning de procesos que son dóciles para el reconocimiento de objetivos ? ¿ Actúan los usuarios en la web con suficiente evidencia en sus propuestas ?

El proyecto AVANTI esta enfocado en la personalización dinámica basada en las necesidades de los usuarios. Funciona de forma similar al WebWatcher, teniendo en cuenta las necesidades y búsquedas de los usuarios, en función de eso, provee una serie de páginas que cree que son interesantes para el usuario en particular.

4.2.2. Optimización

Mientras que la personalización se enfoca a lo individual, la optimización intenta mejorar el sitio en su conjunto. En lugar de hacer cambios por cada usuario, los sitios aprenden de todos los usuarios para hacer el sitio más fácil de navegadr. Este enfoque permite incluso nuevos usuarios, sobre quién no sabe nada, el beneficio de las mejoras.

Debemos ver un diseño web como un punto particular en el espacio de posibles diseños. Mejorando el sitio, entonces, corresponde a buscar en el espacio de un "diseño mejor". Asumiendo que nosotros tneemos un camino de medidas mejores, debemos ver esto como un clásico problema de IA. Una cualidad métrica posible debería ser la medida de una cantidad de esfuerzo que el usuario debe realizar para encontrar lo que desea en nuestra página web. El esfuerzo es definido como la función de un número de enlaces y la dificulta de encontrar esos enlaces. Por ejemplo, un sitio cuya página popular resulta que está escondida despues de pasar por cinco enlaces, se puede mejorar haciéndola más accesible. Podemos navegar a través del espacio realizando transformaciones en el sitio.

¿ Qué longitud tiene el espacio de búsqueda y cuál es la estrategia de búsqueda apropiada ? ¿ Podemos reestructurar el espacio para evitar búsqueda en porciones largas ?

Primero se dibuja el diseño de un sistema con repertorios de transformaciones que apuntan a mejorar la organización del sitio; las transformaciones incluyen reordenado y remarcado de enlaces como sintetizando nuevas páginas. Estos sistemas aprenden de los patrones comunes en los logs de acceso de los usuarios y deciden cómo transformar el sitio para explotar esos patrones y hacer el sitio más fácil de navegar.

Una transformación ambiciosa es el clustering - sintetizando una marca de una nueva página web que contiene enlaces a un conjunto de objetos relacionados. Partiendo de los datos disponibles, el sistema debe inferir este conjunto de páginas en el sitio que son agrupados. Esta inferencia debe estar basada en contenido o en patrones de navegación de usuario. Como exámenes de aproximación final, los estudiantes tienden a buscar múltiples soluciones en cada visita. Incluso conociendo páginas de soluciones que no estan enlazadas de forma directa, los visitantes pueden ir de uno a otro sin problemas. El sistema utilizado en este experimento, podría crear una nueva página con un enlace a cada solución y hacer esta nueva página que este disponible para los visitantes del sitio.

4.2.3. Meta-Información

La capacidad de un sitio web para adaptarse puede verse obstaculizada por el escaso conocimiento sobre su contenido y la estructura proporcionada por el código HTML. Por ejemplo, suponemos que una página contiene una lista de enlaces. ¿ Es apropiado añadir un nuevo enlace en la parte más alta de la lista ? La respuesta depende de del contenido de la lista - un sitio adaptativo no debe añadir un enlace a la parte más ala de una lista de enlaces de un tema concreto ; además, si la lista esta ordenada alfabéticamente, entonces el nuevo objeto puede ser añadido en un punto apropiado. Claramente la capacidad de un sitio web para adaptarse puede ser producida con meta-información: información sobr su contenido, estructura y organización.

Un camino para proveer meta-información es representar los contenidos del sitio en un framework formal con precisión semántica como una base de datos o una red semántica. Este enfoque o aproximación fue pionero por el sistema de gestión de webs STRUDEL, que intentó separar la información disponible en un sitio web de su presentación gráfica. A lo largo de la manipulación de sitios web en el nivel de las páginas y enlaces, los sitios webs podían ser especificados usando el lenguaje de STRUDEL. Además, los sitios webs pueden ser creados y actualizados mediantes consultas STRUDEL. Por ejemplo, una empresa puede crear páginas iniciales para sus empleados mediante la fusión de datos de sus bases de datos de "gestión" y "empleados". Una página podría ser creada para todas las personas en cualquier otra base de datos. Además, cada página de gestión tendria enlaces a sus empleados y viceversa.

Esta aproximación podría facilitar la adaptabilidad porque STRUDEL activa un sitio para razonar sobre su descripción lógica y detecta casos donde las adaptaciones pueden ser truncadas. STRUDEL tiene mecanismos para actualizar el sitio de forma automática y correctamente. El coste de la construcción de envolturas para webs existentes y particularmente sitios no construidos, es bastante alto.

Las etiquetas meta-content son las más "ligeras de peso". Una descripción formal coexiste con documentos HTML. Podemos escoger cuántos sitios queremos editar y la complejidad de los mismos. Estas anotaciones de meta-etiquetas tienen conexiones entre distintas partes de la web.

Una aproximación de este tipo es el formato de Apple Meta-Content. MCF (Meta-Content-Format) es un intento de establecer un estandar para anotaciones de meta´contenido para la web. Cuando un usuario visita un sitio con MCF, con un navegador que tiene activo el MCF, puede elegir navegar por

el sitio en tres dimensiones. SHOE es un lenguaje para añadir ontologías simples a las páginas web. SHOE añade declaraciones básicas ontológicas al HTML; una página puede referirse a una ontología particular y declarar clasificaciones por sus relaciones.

Mientras el meta-contenido es menos pesado que el STRUDEL, también tiene sus desventajas. La anotación por meta-contenidos ha de ser actualizada de forma manual cuando el contenido cambia. No puede automatizarse su adaptación debido a su dependencia del HTML. Cualquier adaptación implica modificar el HTML original.

5. Áreas de investigación relacionadas

La Minería Web se utiliza generalmente en tres caminos, **Minería Web de Contenido, Minería Web de Estructura** y **Minería Web de Uso** (que es de la que hemos hablado en este trabajo).

La **Minería Web de Contenido** es un proceso automático que va más allá de la extracción de palabras clave, ya que los datos se analizan para poder generar información de los documentos que se encuentran en la Web, ya sea, artículos, material audiovisual, documentos HTML, entre otros. La extensa Web puede revelar más información que la que se encuentra contenida en los documentos, por ejemplo, los enlaces apuntando hacia un documento indican la popularidad del documento, mientras algunos vínculos salen de un documento indican la riquieza o quizás la variedad de temas cubiertos en el documento. Esto puede ser utilizado para comparar las citaciones bibliográficas.

En esta área se encuentra la **Minería Web de Estructura**, el cuál consiste en estudiar las estructuras de los enlaces. Y por último, la **Minería Web de Uso,** que es un proceso de descubrimiento automático de patrones de accesos o uso de servicios de la Web, centrándose en el comportamiento de los usuarios cuando interactúan en la Web.

6. Conferencias internacionales

Algunas de las conferencias internacionales que abordan el tema de la minería de uso de la web, son las siguientes:

- International Conference on Databases Theory (ICDT)
- Internacional Conference on Very Large Data Base IBM Almaden Research Center
- International World Wide Web Conference
- Conference on Artificial Intelligence (AAAI198)
- International Conference on Machine Learning (ICML)
- International Conference on Distributed Computing Systems
- European Conference on Machine Learning (ECML-98)
- International Conference Machine Learning
- International Conference on Knowledge Discovery and Data Mining
- International Computer Software and Applications Conference on Prolonging Software Life

Referencias

[1] R. Cooley, B. Mobasher, and J. Srivastava. Web mining and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, ICTAI, 1997

[2] J. Srivastava, R. Cooley, M. Deshpande, P. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 2000.

[3] B. Mobasher. Web Usage Mining and Personalization. Chapter in Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press, 2004

[4] Mike Perkowitz and Oren Etzioni. Adaptive Web Sites: an AI Challenge, IJCAI, 1997

[5] Thorsten Joachims, Dayne Freitag and Tom M. Mitchell.Web Watcher: A Tour Guide for the World Wide Web, IJCAI, 1997

[6] Abiteboul S. (1997) Querying semi-structured data. *En Proceedings of the International Conference on Databases Theory (ICDT)*

[7] Backer E. (1995) Computer-assisted reasoning cluster analysis. *Prentice Hall Internacional (UK) Ltd* Hertfordshire UK.

[8] Broder A., Kumar S., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A. y Wiener J. (2000) Graph structure in the web.

[9] Chang G., Healy M., McHugh J. y Wang J. (2001) Minning the world wide web: An information search approach.

[10] Cernuzzi L. y Molas M. (Septiembre 2004) Integrando diferentes técnicas de data mining en procesos de web usage mining.

[11] Duran B. y Odell P. (1974) Cluster analysis: A survey.

[12] Escobar-Jeria V., Martin-Bautista M., Sanchez D. y Vila M.A (2006) Minería web: Aplicaciones con lógica difusa.

[13] Etzioni 0. (1996) The world wide web: Quagmire or gold mine.

[14] García F. y Gil A. (2002) Personalización y recomendación en aplicaciones de comercio electrónico.

[15] Kim K. y Cho S. (2001) Personalized mining of web documents using link

[16] Kandel A. Fuzzy techniques in pattern recognition.

[17] Molina L. (2002) Data mining: torturando a los datos hasta que confiesen.

[18] Sánchez D. (1999) Adquisición de relaciones entre atributos en base de datos relacionales.

[19] Tan A. (1999) Text mining: Promises and challenges.

[20] Zadeh L. (1975) The concept of linguistic variable and its application to approximate reasoning..

[21] Víctor Heughes Escobar Jeria (2007) Minería Web de Uso y Perfiles de Usuario: Aplicaciones con Lógica Difusa.

[22] http://es.wikipedia.org/wiki/Web_mining

[23] http://webusagemining.com/

[24] http://www.slideshare.net/kamui002/analizador-de-estructuras-de-navegacin-aplicando-minera-de-uso-web-y-mi

[25] http://www-2.cs.cmu.edu/afs/cs/project/theo-6/web-agent/www/project-home.html