DEPARTAMENTO DE LENGUAJES, SISTEMAS E INGENIERÍA DEL SOFTWARE

Facultad de Informática

Universidad Politécnica de Madrid

RESUMEN DEL TRABAJO DE INVESTIGACIÓN

(Programa de doctorado)

Cómputo de los logros de un sitio web mediante el análisis de las sesiones de sus usuarios

Autor: Esther Hochsztain

Tutora: Ernestina Menasalvas Ruiz

Fecha: Septiembre, 2002

Indice

Indice....2
Resumen del trabajo de investigación....2
1. Algoritmo de evaluación de los logros de un sitio web mediante el cómputo del valor de las sesiones de usuarios....2
2. Metodología para la estimación de la utilidad de una página web....9
Bibliografía analizada....14
Publicaciones que el trabajo ha generado....15

Resumen del trabajo de investigación

Los trabajos realizados se refieren a Web Mining. Se han desarrollado en dos áreas temáticas:

1- Algoritmo para la determinación del valor de una sesión en un sitio web, presentado en los artículos [1], [3] [4] [5]

2- Metodología para la estimación de la utilidad de una página web, presentado en el artículo [2]

A continuación se presentan los conceptos fundamentales de los trabajos realizados.

1. Algoritmo de evaluación de los logros de un sitio web mediante el cómputo del valor de las sesiones de usuarios

1.1) Resumen

La exitosa aplicación de técnicas de minería de datos en la Web requiere que éstas se adapten a los cambios continuos en los objetivos de los sitios web. Una de las razones por las cuales ha fallado la aplicación de técnicas de descubrimiento de conocimiento en datos extraídos de la web es que, en la mayoría de los casos, el análisis se ha concentrado exclusivamente en análisis de páginas y caminos más visitados sin tener en cuenta los objetivos del sitio web. Sin embargo, si se quieren extraer patrones útiles e interesantes, los datos de la web se deberían enriquecer con información relacionada con el negocio.

Se propone un algoritmo para determinar el valor de una sesión de un usuario en la web. Dicho algoritmo, tiene en cuenta las metas del sitio web, el comportamiento y perfil del usuario y los cambios en las políticas y objetivos marcados por los administradores del sitio web. La solución que se propone es innovadora en el sentido en que permite, tener en cuenta puntos de vista de distintos usuarios, e integrar la información del sitio web con las metas del negocio.

La entrada del algoritmo es una matriz de valores en la que cada casilla representa el valor que tiene avanzar desde una determinada página a otra. El artículo presenta también resultados experimentales basados en 2400 sesiones analizadas atendiendo a cuatro diferentes matrices.

Palabras Claves: Minería de datos en la web, análisis basado en grafos, valoración de sesiones de usuario

1.2) Introducción

El continuo crecimiento del World Wide Web, unido al entorno competitivo en el cual se mueven las organizaciones modernas, ha hecho necesario diseñar los sitios web teniendo en cuenta, como aspecto fundamental, el conocimiento que se puede extraer de las navegaciones de los usuarios que lo utilizan. Una de las formas de conocimiento más frecuentemente utilizadas consiste en descubrir cuáles son los caminos de usuario más frecuentes. Sin embargo, esto no es suficiente, haciéndose necesario integrar, por ejemplo, minería de datos con los objetivos del sitio web, con el propósito de conseguir que cada sitio web sea el más atractivo y como consecuencia el más competitivo.

La mayoría de las organizaciones que exploran el comportamiento de sus usuarios en la web utilizan, exclusivamente, datos de las secuencias visitadas (“clickstream”).

Hasta el momento, uno de los principales problemas en la aplicación de técnicas de data minign en datos de la web tiene que ver con la etapa de preprocesamiento de datos.

Los servidores web registran, comúnmente, una entrada por cada acceso en el archivo *log*. Entre los datos que se recogen se incluye la dirección IP, el tiempo de acceso, el método pedido, el URL de la página solicitada el protocolo de transmisión, un código de retorno y el número de *bytes* transmitido. El servidor log contiene, no obstante, muchas entradas que son irrelevantes o redundantes para la tarea de minería y que se requiere limpiar antes del preprocesamiento. Después de la limpieza, es necesario identificar y agrupar los datos en sesiones significativas [12]

Las técnicas inteligentes de web mining (intelligent web mining) pueden aprovechar los datos del clickstream una vez preprocesados para extraer conocimiento relacionado con la interacción de los usuarios con la Web [1][2], que se puede utilizar para tomar decisiones críticas de negocio.

Sin embargo, estos datos se deben enriquecer con información relativa al negocio si lo que se espera es ofrecer a las organizaciones conocimiento interesante y útil sobre el mismo y sobre sus clientes de forma que les permita competir. De acuerdo con [3] hoy en día, a menos que se pueda obtener y demostrar ganancia, no se podrá sobrevivir.

En este sentido, en este artículo se propone un algoritmo que a la vez que tiene en cuenta la información registrada en el servidor log mejora el análisis tradicional, puesto que integra información del negocio. El enfoque propuesto tiene en cuenta, para el cálculo de los valores de un enlace, los datos almacenados en el archivo log del servidor, los objetivos del negocio y el conocimiento disponible sobre el área o contexto del negocio.

El algoritmo permite calcular los valores acumulados, durante una sesión, teniendo en cuenta, tanto el análisis del comportamiento de los usuarios como las metas cambiantes del negocio.

La idea básica subyacente al algoritmo es muy similar al proceso de corrección de una prueba de evaluación de los estudiantes. En el caso de los exámenes, dependiendo de sus respuestas los alumnos suman o restan puntos a su calificación. Haciendo una analogía, las páginas visitadas por un visitante lo pueden alejar o acercar a la meta propuesta por la organización. Cuando éste se acerca a la meta, mientras visita las páginas, se añaden puntos; cuando se aleja se restan.

La solución que se propone en este artículo es innovadora porque considera diferentes caminos de evaluación a partir del punto de vista de diferentes usuarios integrando la información proveniente de la web con los objetivos del negocio. De esta manera, se ofrece un marco conceptual para analizar la evolución de las sesiones asignándoles un valor. El enfoque propuesto facilita también la detección de patrones de evolución a partir de sesiones de diferente valor.

El enfoque de representación utilizado se basa en un grafo dirigido como el propuesto en [4] y [5] y en las páginas web adaptativas propuestas en [6] [7] y [8]. El valor del cambio en la conducta de los usuarios es útil para tomar decisiones sobre la necesidad de adaptar las páginas web y sobre cómo hacerlo. Por otra parte, ésta propuesta se relaciona también con el descubrimiento de secuencias propuesto en [9] [10].

El algoritmo, requiere de una fase de preparación exhaustiva de los datos para identificar sesiones y usuarios tal y como se propone en [11].

El resto de la presente propuesta está organizada de la siguiente forma. En la sección 1.3 se presentan los conceptos básicos relacionados con el enfoque propuesto. En la sección 1.4, se describe el algoritmo para calcular la evolución del valor de las sesiones. En la sección 1.5 se presentan algunos criterios para analizar el valor de sesiones junto con un ejemplo de aplicación. La sección 1.6 presenta los resultados experimentales obtenidos al aplicar el algoritmo sobre un conjunto de 2400 diferentes sesiones. Finalmente, en la sección 1.7 se presentan las conclusiones y las futuras tareas de investigación que surgen del enfoque propuesto

1.3) Conceptos básicos

En esta sección se presentan algunos de los conceptos básicos en los cuales se apoya el algoritmo propuesto:

Sitio Web: Como en [13] un sitio web se define como un conjunto finito de páginas web.

Sea W un sitio web y sea Ω un conjunto finito representando las páginas contenidas en W. Cada página tiene asignado un identificador único, de manera que un sitio web consistiendo de m páginas se representa como $\Omega = \{\alpha_1, \ldots, \alpha_m\}$. $\Omega(i)$ representa el *i-éseimo* elemento o página.

Dos páginas especiales, que se denotan como α_0 y α_∞ , corresponden a la página desde la cual el usuario entra al sitio web y la que visita antes de salir de la sesión, respectivamente [14]

Representacion de un sitio web: un sitio web se considera un grafo dirigido, definido como (N,E), donde N es un conjunto de nodos y E es un conjunto de arcos. Un nodo se corresponde con una página web y un arco con un enlace.

Páginas objetivo: Las páginas objetivo son los nodos que se desean alcanzar. La forma de determinarlas forma parte del algoritmo. Estas se definen de acuerdo con las metas de negocio, el perfil del navegador y su historia o comportamiento pasado. De esta manera, es posible, que una página sea página objetivo en una visita de un usuario al sitio web y no sea parte del conjunto de páginas objetivo en una posterior visita del usuario al mismo sitio web.

Enlace: Un enlace es un arco con origen en la página α_i y destino en la página α_j. Los enlaces se representan por medio del par (α_i, α_j).

Valor de un enlace: La principal acción del usuario es seleccionar un enlace para obtener la siguiente página (o terminar la sesión). Esta acción toma diferentes valores dependiendo de la distancia o cercanía a la página o conjunto de páginas objetivo.

El valor del enlace (α_i , α_j) se representa por medio de un número real v_{ij} ($v_{ij} \in \Re$ for $0 \le i,j \le n$):

- Si $v_{ij} > 0$, consideramos que el navegante, yendo del nodo i al nodo j, está más cerca de las páginas objetivo.
- (Si $v_{ij} > 0$, $v_{il} > 0$, $v_{ij} > v_{il}$ entonces, se considera que es mejor ir de la página α_i a la α_j que ir de la página α_i a la α_l)
- Si vij < 0 se considera que el navegante, que va de la página αi a la αj, se está alejando de las páginas objetivo. (Si $v_{ij} < 0$, $v_{il} < 0$, $v_{ij} < v_{il}$ entonces es peor ir de la página α_i a la página α_i que ir de la α_i a la α_k)
- Si I$v_{ij} = 0$ consideramos que el enlace no representa ni una ventaja ni una desventaja en la búsqueda del objetivo.

Sesión: es una secuencia de páginas visitadas por un usuario. El archivo de registro de accesos al sitio web contiene información relacionada con la identificación del usuario (dirección IP), URL de la página solicitada y fecha y tiempo de la solicitud. Con esta información se puede reconstruir la sesión representada como un vector de páginas recorridas: *S[1],S[2],...., S[n]*.

Las sesiones se denotan por S siendo |*S*| su longitud (número de páginas visitadas). Las sesiones se representan como vectores de manera que *S[i]* representa la i-ésima página visitada S[$i \in \Omega$ 1≤i≤ n., con, |S| = n.

Es importante destacar que las páginas del sitio web visitadas durante una sesión se pueden repetir. Por ejemplo, si la primera y sexta páginas visitadas son la página 3, $S[1]=S[6]= \alpha_3$. Sin embargo, las páginas contenidas en el sitio web $\alpha_1, \ldots, \alpha_m$ no se repiten dado que conforman un conjunto.

Secuencia inicial de longitud k: (k páginas iniciales *S[1],S[2],...., S[k*]): las primeras k páginas recorridas durante una sesión representan una secuencia de los k-1 enlaces iniciales de la sesión.

Valor de una secuencia inicial de longitud k: *S[1],S[2],...., S[k)]*: este valor se calcula como la suma de cada uno de los valores de las páginas *S[k]* a las cuales llega el usuario recorriendo los enlaces*(S[1], S[2])*, *(S[2], S[3]), ..., (S[k-1], S[k])* y se denota por *AV(k).*

$$AV(k)= v_{S[1],S[2]} + v_{S[2],S[3]} +...+ v_{S[k-1],S[k]} \quad 2 \le k \le n$$

El valor acumulado de una secuencia inicial se puede definir como:

$$AV(k)= AV(k-1)+ v_{S[k-1],S[k]}$$

Valor de Sesión: Se calcula como la suma de los valores de los enlaces recorridos durante una sesión completa (páginas visitadas) y se denota por *VA(n),* donde

$$AV(n)=v_{S[1],S[2]} + v_{S[2],S[3]} +...+ v_{S[n-1],S[n].} \quad n \ge 2$$

Valor promedio de una sesión: éste representa el valor promedio de cada enlace recorrido durante una sesión. Denotado por *AAV(n),* se define como el valor total de la sesión dividido por el número total de enlaces recorridos. El número de enlaces recorridos es n-1, al final de una sesión en la cual se han recorrido n páginas.

$$AAV(n) = AV(n) / (n-1)$$

Interpretación del Valor promedio acumulado de una secuencia inicial de longitud k (AAV(k)): Este valor ofrece al administrador del sitio web el valor promedio generado para cada uno de los enlaces recorridos hasta alcanzar la página k-ésima . Si ejecutáramos el algoritmo durante una sesión (en tiempo real) obtendríamos una medida útil que es independiente del número de enlaces recorridos. Por ejemplo, si tenemos páginas web adaptativas (tienen en cuenta diferentes parámetros) podríamos ofertar productos y/o páginas más atractivas a aquellos usuarios con un bajo Valor Acumulado en una secuencia inicial de longitud k. De esta manera, se podría incrementar el valor promedio acumulado de cada usuario en una sesión.

Si el número de páginas recorridas (*k*) se incrementa, el Valor Acumulado de una secuencia inicial promedio en las k páginas iniciales (*AAV(k)*), se puede:

- Incrementar, lo que significa que la sesión se acerca la meta.
- Decrementar: cuando la sesión se aleja de la meta
- Permanecer constante: cuando la sesión ni se acerca ni aleja de la meta

1.4) Algoritmo para el cálculo del valor de una sesión

El algoritmo tiene por objetivo conocer lo cercano que está el comportamiento de un usuario del sitio web de los objetivos de la organización. Medimos la distancia de los objetivos utilizando el valor de los

enlaces recorridos. El algoritmo de cálculo de evolución del valor de una sesión se basa en el recorrido de un grafo.

La entrada es una matriz de valores V[m,m] que contiene el valor de los enlaces en un sitio web, que se determinan en base a los procesos de negocios de la organización y los objetivos del sitio web. El análisis de los procesos de negocios brinda un marco conceptual para determinar el valor de los enlaces, en función de cuanto acercan (o alejan) al usuario de las páginas establecidas como objetivo del sitio web.

Las matrices de valoraciones son, en consecuencia, determinadas por los directivos de negocios de la organización. Es de destacar que pueden ser calculadas para cada perfil de usuario y por tanto hacen posible adaptar los objetivos empresariales de acuerdo al comportamiento de los usuarios.

Como consecuencia, las matrices de valoraciones V son adaptables en dos aspectos:

- Los objetivos empresariales no son fijos ni únicos, pueden modificarse o ser considerados desde diferentes puntos de vista. Por ejemplo marketing, ventas, auditoria y relaciones públicas pueden analizar una misma sesión desde sus puntos de vista. Para reflejar diferentes puntos de vista el algoritmo deberá ejecutarse con diferentes matrices de valoraciones como input.
- Por otro lado, la entrada del algoritmo puede incorporar (además de la matriz de valoraciones M) la identificación del usuario. La matriz de valores de los enlaces puede adaptarse en función del perfil del usuario, definido por ejemplo en un proceso previo de segmentación de los usuarios.

Las salidas del algoritmo son la evolución del valor acumulado y del valor acumulado promedio durante la sesión.

Pseudocódigo del algoritmo

```
Input: Value links matrix V[m,m]

Initialización

AV=0    //Added Value=0
AAV=0     //Average added value=0
k=1          //number of nodes=1
read S[k] //read the first traversed page in the Web site S[1]

Pseudocode:
While new pages are traversed
  k = k +1 //compute the traversed
            page sequential number
  read S[k] // read the next
               traversed page
  /*  the selected link is
     (S[k-1],S[k])
      1≤S[k-1]≤m-1  1≤S[k] ≤m
      2≤k≤n */

      AV = AV + V(S[k-11],S[k])
      // Add link traversed value to
         accumulated value
      AAV =  AV / k-1
      // Compute average link
         traversed value
      Plot values
      // needs the previous value to be stored

Output:  Final Accumulated Value and Final Average Accumulated Value
/* or Accumulated Value and Average Accumulated Value evolution  if all values
are stored. */
```

1.5) Valor de una Sesión

La principal ventaja del algoritmo propuesto es el cálculo de la evolución del valor de una sesión., que constituye un elemento relevante en la adopción de decisiones de diseño relativas a sitios y páginas web.

Una empresa puede beneficiarse con estos resultados al detectar la necesidad de incorporar nuevas páginas, realizar ofertas online o efectuar ventas cruzadas. Frecuentemente, los ejecutivos no comprenden cómo usar la tecnología y qué tipo de análisis efectúa [3]. Nuestra propuesta sugiere adaptar la tecnología a las métricas de los ejecutivos. Su principal ventaja es que no requiere de grandes esfuerzos para entenderla y por tanto el esfuerzo requerido para utilizarla se minimiza.

A continuación presentamos ejemplos de análisis del valor de una sesión seguidos de un ejemplo que ilustra el comportamiento del algoritmo.

1.6) Análisis de la evolución del valor de una sesión

Para analizar la evolución del valor de una sesión se presenta una gráfica en la cual las abcisas representan la cantidad de enlaces recorridos y las ordenadas el valor acumulado hasta el último enlace recorrido.

En la Figura 1.1 se muestra la evolución del valor de una sesión, que disminuye al comienzo y que luego aumenta. Durante esta sesión el usuario se fue alejando del objetivo y posteriormente volvió a él.

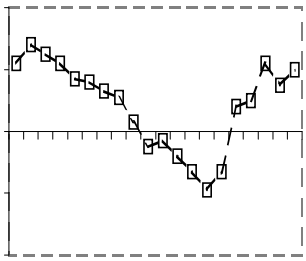

Figura 1.1 Evolución del valor acumulado de una sesión

Las Figuras 1.2 y 1.3 describen tres sesiones que tienen similar evolución de valor. De esta manera es posible encontrar dos patrones en las seis sesiones**.**

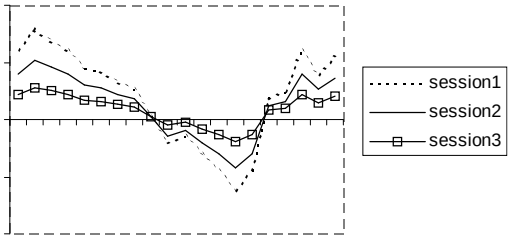

Figura 1.2- Evolución del valor acumulado (sesiones 1 a 3).

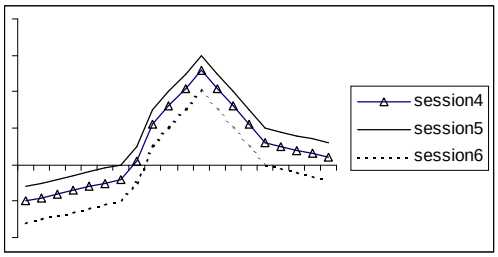

Figura 1.3-Evolución del valor acumulado (sesiones 4 a 6)

A partir de las sesiones 1 a 6 que aparecen en la Figura 1.4 se pueden obtener dos patrones interesantes.

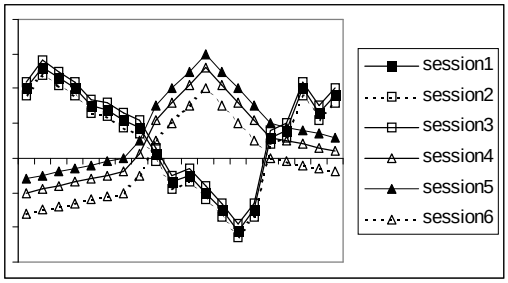

Figura 1.4 - Evolución del valor acumulado (sesiones 1 a 6)

Combinando el análisis de la evolución del valor con las características de las páginas del sitio web es posible identificar las páginas que llevan a secuencias ascendentes o descendentes con respecto al valor de la sesión. En particular, estas páginas serían las previas o iniciales de ciclos de valores descendentes. Esto puede permitir detectar aspectos a modificar en las mismas.

1.7) Cálculo del valor de una sesión

En esta sección se presenta un ejemplo de cálculo del valor de una sesión. El ejemplo permite observar la baja complejidad del algoritmo propuesto. En lugar de constituir una limitación, su simplicidad puede ser considerada un elemento positivo, porque facilita la comprensión de sus resultados por parte de los administradores de un sitio web.

La entrada del algoritmo es la siguiente matriz de valores de los enlaces V[4,4]

	Origen (α_i)			
Destino (α_j)	α_1	α_2	α_3	α_4
α_1	3	2	3	6
α_2	4	1	2	2
α_3	-5	-1	-1	-1
α_4	-6	-2	-3	-1

En la Figura 1.5 se muestra el grafo con la asignación de valores a los arcos asociada a la matriz anterior. Es fácil observar que en el grafo existen dos nodos objetivo (α1 y α2), porque los arcos de entrada a estos nodos son positivos y los arcos que salen hacia el resto de los nodos (α3 y α4) toman valores negativos.

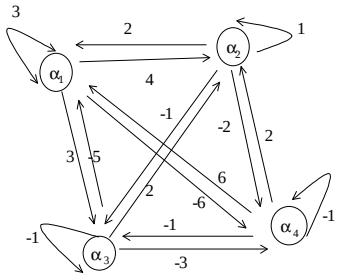

Figura 1.5- Grafo con los valores de los *enlaces*

Los enlaces que coinciden en origen y destino, que se observan en la diagonal de la matriz de valoraciones y en el grafo correspondiente, representan la acción de *reload* (volver a cargar) una página.

1.8) Resultados Experimentales

El análisis y las pruebas se hicieron con datos de 2400 sesiones de usuario provenientes de un sitio web de comercio electrónico. Los datos fueron procesados teniendo en cuenta 4 matrices diferentes. Se descartaron las sesiones que recorrían 10 páginas o menos, dado que la propuesta no resulta interesante para analizar sesiones cortas.

Los valores de las cuatro matrices utilizadas premian los siguientes aspectos:

1. área de noticias ,
2. página principal
3. registro de usuarios
4. realización de compras .

En la figura 1.6 se muestra que al analizar los valores de los acumulados por sesión con las cuatro matrices se observan patrones claramente definidos. Se utiliza una matriz de diagramas de dispersión, que muestra todas las combinaciones posibles de los valores acumulados por sesión obtenidos con las cuatro matrices consideradas. La primera fila y la primera columna presentan cálculos obtenidos con la matriz 1, la segunda fila y la segunda columna los obtenidos con la matriz 2, y así sucesivamente.

En el cruce de la primera fila y la segunda columna se presenta el gráfico que vincula valores de sesión obtenidos con las matrices 1 y 2. En el cruce de la segunda fila y la primera columna se presenta (con los

ejes invertidos) el análisis de las mismas matrices. No se cruzan los valores obtenidos con una matriz, consigo misma, dado que siempre el gráfico estará formado por puntos ubicados en la diagonal principal.

Al analizar el gráfico, se observa por ejemplo como los valores obtenidos con las matrices 1 y 2 muestran son totalmente opuestas.

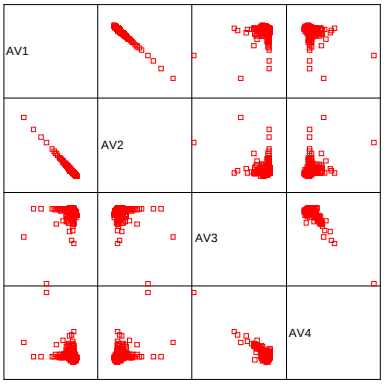

Figura 1.6 - Valores acumulados con las 4 matrices consideradas.

En la Figura 1.7 se muestran los valores promedio obtenidos con las cuatro matrices consideradas. Se observa el mismo patrón que para los valores acumulados.

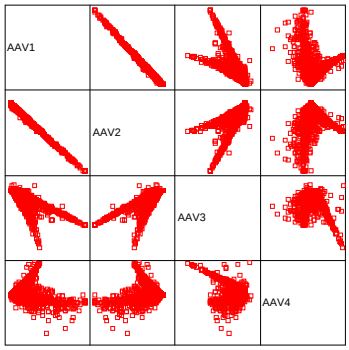

Figura 7 - Valores promedio con las 4 matrices consideradas

Con el objetivo de resumir los valores promedio de cada enlace recorrido en una sesión obtenidos con las cuatro matrices, calculamos media y desviación estándar, que se presentan en la Tabla 1.1.

El valor positivo +4,8048 muestra que las sesiones son favorables al ser analizadas con el criterio subyacente a la matriz 2. Sin embargo, el valor promedio --3.3382 muestra una evaluación desfavorable si se analizan las sesiones con el criterio que brinda la matriz1. Al observar los valores de resumen para el valor promedio al final de la sesión, se observa que los caminos, en promedio, se adecuan a los objetivos reflejados en la matriz 2 y no se adecuan a los planteados en los objetivos reflejados en la matriz 1

Matriz	Media	Desviación Estandar
1	-3,3382	2,1435
2	4,8048	2,8445
3	-,8720	1,2446
4	,1187	,3767

Tabla 1.1 - Media y desviación estándar del valor acumulado promedio

Por último, un ejemplo de los patrones que se obtienen con las diferentes matrices se presentan en la Figura 1.8

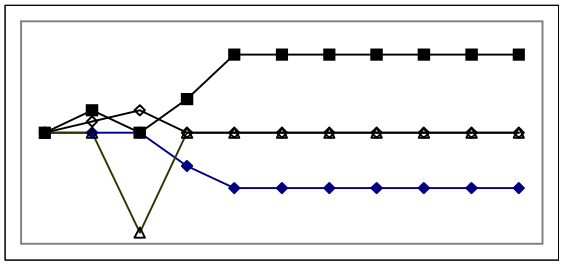

Figura 1.8 Valores obtenidos con datos experimentales.

1.8)1. Conclusiones y trabajo futuro

Se ha presentado un enfoque para calcular el valor de una sesión de usuario en un sitio web teniendo en cuenta tanto su comportamiento como las metas establecidas por el administrador del sitio web. El enfoque propuesto ha sido probado con datos reales y demuestra ser útil para seguir el comportamiento de los usuarios mientras navegan en un sitio. Por otra parte, del análisis a posteriori de las sesiones, es posible identificar sesiones similares incluso cuando no se accede a las mismas páginas. Asimismo, la posibilidad de definir distintos tipos de matrices, de acuerdo a distintos criterios posibilita el poder realizar análisis bajo diferentes puntos de vista.El mayor problema del algoritmo radica en el hecho de que se requieren las matrices de valoración de entrada y de momento éstas se calculan manualmente. Estamos trabajando en la actualidad en un prototipo para el cálculo de estas matrices aplicando técnicas de data mining. Otro problema del algoritmo propuesto es que no tiene en cuenta los tiempos de permanencia del usuario web en una página. En la siguiente versión del algoritmo está prevista una modificación para tener en cuenta estos tiempos.

1.9) Bibliografía

[1]B. Mobasher, N. Jain, E. Han, and J. Srivastava. (1997) "Web mining: Pattern discovery from WWW transaction". *In Int Conference on Tools with Artificial Intelligence*, pages 558-567, New port.
[2] J.Han, M."*Kamber .Data Mining: Concepts and Techniques"* . Academic Press USA 2001
[3] G. Piatetsky-Shapiro "*Interview with Jesus Mena, CEO of WebMiner, author of Data Mining your Website"* Date: June 24, 2001 http://www.kdnuggets.com/news/2001/n13/13i.html
[4] J. Borges and M. Levene. "Mining navigation patterns with hypertext probabilistic grammars" *Research Note RN/99/08*, Department of Computer Science - University College London, 1999.
[5] J. Borges and M. Levene. "Data mining of user navigation patterns". *Web Usage Mining, in Lecture Notes in Artificial Intelligence (LNAI 1836)* B. Masand and M. Spliliopoulou, editors,. Springer-Verlag, Berlin, 2000.
[6] M. Perkowitz and O. Etzioni "Adaptive Web Sites: Automatically Synthesizing Web Pages". In *Proceedings of AAAI98*.
[7] M. Perkowitz and O Etzioni. "Adaptive web sites: Conceptual cluster mining". *In Sixteenth International Joint Conference on Articial Intelligence*, Stockholm, Sweden, 1999.
[8] M. Perkowitz and O. Etzioni. "Towards adaptive Web sites: Conceptual framework and case study". In *Artificial Intelligences* 118, 2000.
[9] M. Spiliopoulou, L. Faulstich, and K. Wilkler. "A data miner analyzing the navigational behaviour of web users". In *Proc. Of the Workshop on Machine Learning in User Modelling of the ACAI99*, Greece.
[10] M. Spiliopoulou, C. Pohle, and L. Faulstich. "Improving the effectiveness of a web site with web usage mining". In *Proceedings WEBKDD99*.
[11] R. Cooley, B. Mobasher, and J. Srivastava. "Data preparation for mining world wide web browsing patterns". *Knowledge and Information Systems*, 1(1), February 1999.
[12] E. Menasalvas, S. Millán, J.M. Peña, M. Hadjimichael, O. Marbán " Subssesions: A granular approach to click path analysis" In *Proceedings WCCI'2002*
[13] C. Shalabi, F. Banaei-Kashaani, J. Faruque and A. Faisal. (2001) "Feature matrices: A model for e-Ecient and anonymous web usage mining". *In Proceedings of EC-Web 2001*, Germany, September
[14] E. Menasalvas, O. Marbán , S. Millán , and J. M. Peña "Intelligent Web Mining" in *Intelligent Exploration of the Web series Studies in Fuzziness and Soft Computing*,. Springer-Verlag 2002 P.S.Szczepaniak, J. Segovia, J. Kacprzyk, L.A. Zadeh – editors

2. Metodología para la estimación de la utilidad de una página web

2.1) Resumen

Los sitios web necesitan ser muy atractivos para los visitantes, debido a que se ubican en un entorno sumamente competitivo. Proponemos en un enfoque para analizar y determinar el nivel de agrado de los usuarios de un sitio web que tienda a asegurar la satisfacción de sus usuarios, en base a su tipo de páginas y el tipo de usuarios.

Proponemos un enfoque granular basado en la idea de que una página puede ser considerada como un conjunto de características o factores y cada uno de ellos puede percibirse en diferentes niveles de granularidad. El enfoque propuesto permite estimar una medida de la utilidad que brinda a un usuario cada nivel de cada factor en particular. En una página en particular, cada factor toma un cierto. La medida global de utilidad por una cierta página se determina considerando conjuntamente los niveles que presenta dicha página en cada factor de diseño.

Palabras clave: minería de datos, minería de la web, enfoque granular, factores de diseño, utilidad de una página

2.2) Introducción

Para diseñar páginas web atractivas uno de los principales desafíos es encontrar los factores que deben tenerse en cuenta. Cuando se resuelve este problema el segundo desafío es encontrar un modelo para cuantificar su relevancia. El problema requiere un modelo que considere los atributos y su cuantificación y como tenga en cuenta diferentes perfiles de usuarios y tipos de páginas. Considerando que en la web pueden distinguirse diferentes dominios (i.e. educacionales, empresariales, administrativos, gubernamentales), la metodología propuesta incluye aspectos comunes a todos los sitios web.

El diseñador de sitios web debe actuar en función del dominio de trabajo. Por tanto, hay decisiones de diseño que no pueden definirse de una forma genérica. En este sentido, debemos distinguir tipos de páginas (i.e. comercio electrónico, información) y perfiles de usuarios y características (i.e. habilidad en el uso de computadoras, sensibilidad y formación artística). El principal objetivo de este trabajo es brindar a los administradores de un sitio una metodología para determinar el grado de afabilidad de un sitio web, que en el largo plazo, pueda ayudarlos a diseñar páginas y sitios atractivos. Una respuesta a la pregunta "¿qué paginas atraen más la atención de los usuarios?" facilitará el diseño de sitios web adaptativos y a predecir el comportamiento de los visitantes del sitio de acuerdo a sus perfiles y a las páginas que han visitado en la sesión. Hará también posible vincular el diseño de las páginas con los objetivos empresariales del sitio web.

Para identificar características que aumentan el agrado por las páginas necesariamente deben incluirse factores usualmente tomados en cuenta en el diseño de sitios web. Pero también serán incluidos aspectos nunca usados o considerados antes irrelevantes. Esto ultimo seguramente ayudará a mejorar la calidad de los sitios considerando que diferentes usuarios tienen diferentes gustos, preferencias y desagrados, y la utilidad de una página puede asociarse a los perfiles de usuarios (i.e. educación, género, edad, pasatiempos, religión). Por tanto la metodología propuesta requiere considerar usuarios objetivo para adaptar el hallazgo de nuevos atributos de diseño a los perfiles de los usuarios. El análisis de patrones de uso de la web puede ser relevante para identificar atributos que atraen a los usuarios. Suponemos que los usuarios vistan con mayor frecuencia aquellas páginas que los atraen más porque por su diseño o por la información que contienen. Nuestra tarea es descubrir el valor que el usuario asigna a un sitio web y/o a una página.

Proponemos un enfoque granular para descubrir el valor que un usuario asigna a una página, cuantificando cada una de los factores de diseño utilizando un enfoque que descompone su valor o utilidad. La principal idea que subyace la metodología propuesta es la siguiente: un visitante evalúa el valor de una página combinando los valores individuales que asigna inconscientemente a los atributos de dicha página. La utilidad de una página web del punto de vista del usuario es un juicio subjetivo que. represente una preferencia global por la página web. Esta preferencia del usuario es un marco conceptual para cuantificar el valor de una página web.

En nuestra propuesta, suponemos que el valor (utilidad) de una página se basa en los valores individuales asignados a cada factor de diseño. Y agregando los valores individuales de cada factor obtendremos la utilidad conjunta de la página. Las páginas con mayor utilidad serán consideradas más atractivas y supondremos que tienen mayor probabilidad de ser elegidas.

La propuesta se estructura de la siguiente forma: En la Sección 2.3 se presenta la metodología propuesta para calcular la utilidad de páginas web. En la Sección 2.4 se muestra un ejemplo de utilización de la metodología propuesta.

2.3) Metodología para calcular la utilidad de páginas web

Proponemos un enfoque metodológico basado en la estimación la utilidad que posee una página para un cierto usuario. Nuestra propuesta utiliza el análisis conjunto multivariado [HA+98] [GC+89] para la construcción del modelo y para estimar sus parámetros. Deben efectuarse varios ajustes para adaptar la metodología existente al diseño no experimental usado..

En el modelo para explicar el comportamiento de los usuarios la variable dependiente (y) es el tiempo que un usuario permanece en una página. Asumimos que existe una relación directa entre el tiempo de permanencia y el agrado por la página, de modo que cuanto más permanece el usuario en una página más le agrada. Asumimos que el tiempo de permanencia depende tanto de las alternativas de diseño como de los contenidos de la página. Queremos descubrir como diseñar páginas que hagan que el tiempo de permanencia aumente.

Consideramos factores de diseño a aquellos elementos que pueden modificarse al diseñar la página. Cada factor de diseño puede ser implementado en diferentes niveles. Entre los factores que podrían ser tenidos en cuenta consideramos:

- El tipo de imágenes que contiene la página: estática (nivel 1), dinámica (nivel 2)

- El color de fondo: suave (nivel 1), fuerte (nivel 2)
- El tipo de lenguaje utilizado: técnico (nivel 1), coloquial (nivel 2)
- El tamaño de letra: grande (nivel 1), pequeño (nivel 2)

Si bien en lo anterior sólo se sugieren dos niveles para cada factor, podrían considerarse más. Se requiere que se asigne un único nivel a cada factor en cada página utilizada en el experimento. Construimos un modelo que explica el tiempo de permanencia en función de los factores de diseño.

El valor promedio (μ) del tiempo de permanencia puede aumentar o disminuir en función de los niveles considerados para cada factor (βab). Por tanto, los parámetros β tienen dos subíndices: el primero identifica el factor y el segundo el nivel de dicho factor.. También se considera un término de error (ε). Si consideramos tres factores de diseño el modelo resultante es : $yijk = \mu+\beta 1i+\beta 2j+\beta 3k+\varepsilon ijk$

Para determinar el tiempo de permanencia se estiman los parámetros del modelo μ , $\beta 1i$, $\beta 2j$, $\beta 3k$ a través de los estimadores $\hat{\mu}$, $\hat{\beta}_{1i}$, $\hat{\beta}_{2j}$, $\hat{\beta}_{3k}$ respectivamente, donde. i, j, k varían entre 1 y la cantidad de niveles de los factores 1,2,3 respectivamente. Los valores de $\hat{\beta}_{1i}$, $\hat{\beta}_{2j}$, $\hat{\beta}_{3k}$ para cada nivel se utilizan para estimar si el tiempo de permanencia aumenta o disminuye en función de las alternativas de diseño utilizadas en cada página.

El procedimiento de estimación puede ser métrico o no métrico en función de si el método utilizado para transformar la variable dependiente es lineal o monótono. La estimación de los parámetros puede requerir iteraciones, dependiendo del modelo especificado.

La utilidad de una página se determina en función del nivel de cada uno de los factores que influye en su diseño (niveles de los atributos). Se propone una función que determina la utilidad de una página en función de diferentes combinaciones de atributos. Como consecuencia, las páginas con mayor utilidad son más atractivas y por tanto tendrán mayor probabilidad de ser elegidas.

Obtención de los datos

Analizar el comportamiento de los visitantes de un sitio, y en particular las decisiones que adoptan al visitar una página, permite obtener información relativa a la relevancia de cada factor de diseño de dicha página. Por tanto, cada página puede evaluarse tomando en cuenta diferentes atributos (factores) de diseño y sus respectivos niveles (valores). La metodología tradicional del análisis conjunto [HA+98] se basa en un diseño experimental. Se presentan a una persona diferentes opciones de diseño que son combinaciones de atributos (factores) con diferentes niveles. El usuario manifiesta su preferencia global para cada una de las opciones presentadas. Nuestra propuesta se basa en esta metodología pero, en nuestro caso, en lugar de diseñar un experimento para consultar al usuario, se analizan los logs del servidor web.

Debido a que en nuestra propuesta no se usan variables independientes ni existe una persona controlando a los usuarios, estamos en presencia de un diseño no experimental. Constituye un diseño "expost facto" (luego que los hechos ocurrieron) debido principalmente a que se observa al usuario y luego se determina el presunto factor causal.

Identificación de atributos

Para determinar los atributos relevantes de una página web pueden utilizarse los siguientes métodos:

1. Juicio de expertos
2. Métodos cualitativos, generalmente en base a un pequeño número de personas entrevistadas. Pueden basarse en grupos motivacionales o entrevistas en profundidad.
3. Identificación experimental. Utilizamos este último. Primero utilizamos la técnica propuesta considerando todos los factores de diseño posibles de una página web con el objetivo de identificar los más relevantes. En segundo lugar, solamente los atributos considerados relevantes en la etapa anterior se tendrán en cuenta. Para aplicar este método se requiere que el sitio web considerado contenga páginas con diferentes criterios de diseño. Este procedimiento propuesto se detalla a continuación:

Determinación de las variables independientes:

1. Identificar de factores de diseño.
2. Describir los niveles considerados de cada factor de diseño.
3. Describir las páginas web en función de los factores y niveles identificados anteriormente. En resumen, cada página se caracterizará como una lista de vares de valores de la forma:

$$(\text{factor1-nivel}_{1x}, \text{factor2-nivel}_{2y}, \ldots, , \text{factork-nivel}_{ky})$$

Identificación de la o las variable Dependiente

La o las variables dependientes consisten en las medidas que nos interesa considerar respuesta a las alternativas de diseño de una página. Por ejemplo, el tiempo de permanencia en la página, el número de clicks, etc. pueden ser consideradas variables dependientes.

Proceso de estimación

Consiste en estimar la utilidad de todos los niveles en todos los factores para el usuario. Estas estimaciones parciales de los niveles individuales de los factores se usan para determinar la estimación global de la utilidad de una página. Los conceptos preliminares considerados en el experimento son:

- Población objetivo: los usuarios de las páginas consideradas
- Unidad experimental: una visita de un usuario a una página.
- Parámetros: atributos de la página(tipo de página, objetivo principal, habilidades requeridas).
- Variables de respuesta (variables dependientes): la utilidad de una página.
- Factores (variables independientes): características que afectan a las variables dependientes. Un factor es un atributo de diseño (i.e. tipo de imágenes en la página, tamaño de letra). Estamos interesados en identificar el impacto de estos factores, definidos como la utilidad de los atributos de diseño. Las opciones para identificar los factores se han mencionado previamente.
- Niveles: diferentes valores que puede tomar una variable independiente (i.e. las imágenes en una página pueden ser estáticas o dinámicas, el tamaño de letra puede ser grande o chica).

2.4) Ejemplo

A continuación se presenta un ejemplo de estimación de parámetros y de análisis de sus resultados. Los factores de diseño de una página (variables independientes del modelo) y sus correspondientes niveles se presentan en la tabla 2.1.

FACTOR		Nivel 1	Nivel 2
β_1	Tipo de imágenes	β_{11} = estáticas	β_{12} = dinámicas
β_2	Tamaño de letra	β_{21} = grande	β_{22} = pequeña
β_3	Color de fondo	β_{31} = suave	β_{32} = fuerte

Tabla 2.1. - Ejemplo de factores y sus niveles.

Presentamos los siguientes datos relacionados con la utilidad (tiempo de permanencia) que un usuario asigna a diferentes combinaciones de los tres factores de diseño considerados.

Tipo de imágenes	Tamaño de letra	Color de fondo	Tiempo
estáticas	grande	Suave	15
estáticas	grande	Fuerte	12
estáticas	pequeña	Suave	12
estáticas	pequeña	Fuerte	8
dinámicas	grande	Suave	18
dinámicas	grande	Fuerte	16
dinámicas	pequeña	Suave	18
dinámicas	pequeña	Fuerte	14

Tabla 2.2 - Datos del ejemplo

Los parámetros de yijk = $\mu+\beta 1i+\beta 2j+\beta 3k+\varepsilon ijk$ se estiman considerando las siguientes restricciones $\beta 11+\beta 12= \beta 21+\beta 22=\beta 31+\beta 32= 0$. que indican que la suma utilidades de los niveles de cada atributo debe ser nula. El término de error es εijk. El análisis conjunto métrico utilizado crea una matriz de diseño de efecto principal para las variables especificadas.

En este ejemplo, . R2 = 0.94436 y R2Corregido = 0.9026. La Tabla 2.3 presenta la relevancia estimada de cada factor., se puede apreciar que el tipo de imagen predominante es el factor más importante, seguido del color de fondo , siendo el tamaño de letra el factor con menor relevancia estimada. Debe tenerse en cuenta que la tabla ANOVA brinda, en este caso, sólo una aproximación al ajuste del modelo conjunto, debido a que los supuestos de normalidad e independencia no se cumplen.

FACTOR		IMPORTANCIA
β_1	Tipo de imágenes	46.342%
β_2	Tamaño de letra	21.951 %
β_3	Color de fondo	31.707%

Tabla 2.3- Estimaciones de la importancia de los factores

Las estimaciones de utilidad presentadas en la Tabla 4 permiten identificar los niveles preferidos de cada atributo. Los niveles con utilidad positiva se prefieren a aquellos con utilidad negativa. Las estimaciones de la utilidad de cada uno de los niveles de todos los factores considerados se presenta en la Tabla 2.4.

FACTOR		Nivel 1	Utilidad Estimada	Nivel 2	Utilidad Estimada
β_1	Tipo de	β_{11} = estáticas	$\hat{\beta}_{11}$ = -2.375	β_{12} = dinámicas	$\hat{\beta}_{12}$ = +2.375

FACTOR		**Nivel 1**	**Utilidad Estimada**	**Nivel 2**	**Utilidad Estimada**
	imágenes				
β_{2}	Tamaño de letra	β_{21} = grande	$\hat{\beta}_{21} = +1.125$	β_{22} = pequeña	$\hat{\beta}_{22} = -1.125$
$\beta_{3.}$	Color de fondo	β_{31} = suave	$\hat{\beta}_{31} = +1.625$	β_{32} = fuerte	$\hat{\beta}_{11} = -1.625$

Tabla 2.4- Utilidad estimada de los niveles de factores de diseño

El valor positivo $\hat{\beta}_{12}$ =+2.375 del nivel dinámicas del factor tipo de imágenes muestra preferencia por este tipo de imágenes en contraste con el valor negativo $\hat{\beta}_{11}$ = -2.375 de las imágenes estáticas. Procediendo de la misma forma con los restantes atributos podemos decir que los niveles preferidos de los tres factores considerados son imágenes dinámicas, letra grande y color de fondo suave.

La estimación del la media general $\hat{\mu}$ es 14.250. Para la combinación preferida de Tipo de imágenes, Tamaño de letra y Color de fondo la utilidad estimada es $\hat{y}$ =14,125 + 2,375 + 1,125 + 1,625 = 19,25, siendo el valor observado de la preferencia de dicha combinación y =18. Para la combinación menos preferida la utilidad estimada y el valor observado de la preferencia son respectivamente 14,125+ -2,375+ -1,125 + -1,625 = 9 25 = $\hat{y}$ e y =8.

La utilidad puede ser considerada como los valores que se predicen en un modelo de regresión. El coeficiente de determinación entre la utilidad de cada combinación y el tiempo observado es R2. Los factores que presentan mayor utilidad se consideran los más relevantes en la determinación de los tiempos de permanencia estimados.

2.5) Conclusiones

La metodología presentada permite estimar la utilidad de una página en función de su diseño. Se supone que los usuarios permanecerán más tiempo en aquellas páginas que les resultan más interesantes.

El principal resultado de nuestro enfoque es que las páginas web pueden ser parametrizadas en función de diferentes factores de diseño y podrían diseñarse dinámicamente con el objetivo de adaptarse a las preferencias de los usuarios (estimadas durante el transcurso de la sesión de dicho usuario). Por consiguiente, el enfoque propuesto permite que los diseñadores de páginas web tomen decisiones por al poseer información de la contribución relativas de cada factor de diseño de la página y sus respectivos niveles en el agrado (o utilidad) que la página genera. El diseñador puede estimar la mejor combinación de atributos (la que genera mayor utilidad) para cada página en particular.

El enfoque también considera información relativa a los perfiles de los usuarios en relación a su preferencia por cierto tipo de páginas. Esto permite el diseño de páginas para grupos predefinidos de usuarios, si los diseñadores saben con anticipación los segmentos de usuarios que serán potenciales visitantes de la página. Por consiguiente, la preferencia de los usuarios por una página puede ser tomada en cuenta en algoritmos de web mining adaptativo.

2.6) Referencias

[AGJ00] Andersen J., Giversen A., Jensen A. Larse R., Bach T., Skyt J. Analysing clickstreams using subsessions. Proc. DOLAP-OO, pp. 25-32, 2000

[BM00] Borges J., Levene M. A fine grained heuristic to capture web navigation patterns. SIGKDD Exploration, 2(1) pp 40-50, 2000.

[CY00] Chang Wei-Lun, Yuan Soe-Tsyr. A synthesized Learning Approach for Web-Based CRM. Working Notes of Workshop on Web Mining for E-commerce: Challenges and Opportunities. August 20, 2002 Boston USA pp. 43-59

[GS00]Gaul Wolfang, Schmidt-Thieme Lars. Mining web navigation path fragments. Workshop on Web Mining for E-Commerce - Challenges and Opportunities. Working notes pp.105-110. Kdd-2000, August 20,2000, Boston, MA.

[Ga01] John Gajan Rajakulendran. Personalised Electronic Customer Relationships: Improving The Quality of Data Within Web Clickstreams - Individual Project (MSc) - Newcastle University (UK) & Universidad Politecnica Madrid - Supervisor: E. Menasalvas (UPM)

[GA96] Ildefonso Grande, Elena Abascal - Fundamentos y Técnicas de Investigación Comercial - ESIC - España 1996

[GC+89] Paul E. Green, Frank J. Carmone, JR. Scott M. Smith - Multidimensional Scaling Concepts and applications - Allyn and Bacon- A Division of Simon & Schuster, USA 1989.

[HK01] Han J., Kamber M. Data Mining: Concepts nad Techniques. Acadc. Press, USA 2001

[HA+98] Joseph F. Hair, Jr, Rolph E. Anderson, Ronald L. Tathan, William C. Black - Multivariate Data Analysis - Prentice Hall USA 1988

[HM02] Hochztain E., Menasalvas E. Sessions value as measure of web site goal achievement. Technical Report. Universidad Politécnica de Madrid, 2002

[KNY00] Kato H., Nakayama T., Yamane Y. Navigation Analysis Tool based on the Correlation between Contents Distribution and Access Patterns. Workshop on Web Mining for E-Commerce - Challenges and Opportunities Kdd-2000, August 20,2000, Boston, MA

[LAR00] Lin Weiyang, Alvarez Sergio, Ruiz Carolina. Collaborative Recommendation via Adaptative Association Rule Mining. Working Notes of Workshop on Web Mining for E-commerce: Challenges and Opportunities. August 20, 2002 Boston USA pp. 35-41

[MB+97] Salvador Miquel, Enrique Bigné, Jean-Pierre Lévy, Antonio Carlos Cuenca, Mª José Miguel - Investigación de Mercados-Mc Graw - Hill /Interamericana de España - 1997

[MJHS97] Mobasher B., Jain N., Han, E-H., Srivastava J. Web Mining: Pattern Discovery from World Wide Web Transactions. In International Conference on Tools with Artificial Intelligence, pp. 558-567, New Port 1997

[MMP+02]] Menasalvas E., Millán S., Peña J., Hadjimichael M., Marbán O. Subsessions: a granular approach to click path analysis. In Proc. WICI'02

[PM01] From: Gregory Piatetsky-Shapiro 2001: Interview with Jesus Mena, (WebMiner)

[SFBF00] Shahabi C., Faisal A., Banaei F., Faruque J. INSITE:A tool for real-time knowledge Discovery from users web navigation. In Proc. VLDB-2000, 2000.

[SFKFF01] Shahabi Cyrus, Farnoush Banaiei-Kashaani, Jaabed Faruque, Adil Faisal. Feature Matrices: A model for e-Ecient and anonymous web usage mining. Proc. of EC-Web 2001.

Bibliografía analizada

Andersen J., Giversen A., Jensen A. Larse R., Bach T., Skyt J. Analysing clickstreams using subsessions. Proc. DOLAP-OO, pp. 25-32, 2000

Borges J., Levene M. A fine grained heuristic to capture web navigation patterns. SIGKDD Exploration, 2(1) pp 40-50, 2000.

J. Borges and M. Levene. "Mining navigation patterns with hypertext probabilistic grammars" Research Note RN/99/08, Department of Computer Science - University College London, 1999.

J. Borges and M. Levene. "Data mining of user navigation patterns". Web Usage Mining, in Lecture Notes in Artificial Intelligence (LNAI 1836) B. Masand and M. Spliliopoulou, editors,. Springer-Verlag, Berlin, 2000.

Chang Wei-Lun, Yuan Soe-Tsyr. A synthesized Learning Approach for Web-Based CRM. Working Notes of Workshop on Web Mining for E-commerce: Challenges and Opportunities. August 20, 2002 Boston USA pp. 43-59

R. Cooley, B. Mobasher, and J. Srivastava. "Data preparation for mining world wide web browsing patterns". Knowledge and Information Systems, 1(1), February 1999.

Gaul Wolfang, Schmidt-Thieme Lars. Mining web navigation path fragments. Workshop on Web Mining for E-Commerce - Challenges and Opportunities. Working notes pp.105-110. Kdd-2000, August 20,2000, Boston, MA.

John Gajan Rajakulendran. Personalised Electronic Customer Relationships: Improving The Quality of Data Within Web Clickstreams - Individual Project (MSc) - Newcastle University (UK) & Universidad Politecnica Madrid - Supervisor: E. Menasalvas (UPM)

Ildefonso Grande, Elena Abascal - Fundamentos y Técnicas de Investigación Comercial - ESIC - España 1996

Paul E. Green, Frank J. Carmone, JR. Scott M. Smith - Multidimensional Scaling Concepts and applications - Allyn and Bacon- A Division of Simon & Schuster, USA 1989.

Han J., Kamber M. Data Mining: Concepts nad Techniques. Acadc. Press, USA 2001

Joseph F. Hair, Jr, Rolph E. Anderson, Ronald L. Tathan, William C. Black - Multivariate Data Analysis - Prentice Hall USA 1988

Hochztain E., Menasalvas E. Sessions value as measure of web site goal achievement. Technical Report. Universidad Politécnica de Madrid, 2002

Kato H., Nakayama T., Yamane Y. Navigation Analysis Tool based on the Correlation between Contents Distribution and Access Patterns. Workshop on Web Mining for E-Commerce - Challenges and Opportunities Kdd-2000, August 20,2000, Boston, MA

Lin Weiyang, Alvarez Sergio, Ruiz Carolina. Collaborative Recommendation via Adaptative Association Rule Mining. Working Notes of Workshop on Web Mining for E-commerce: Challenges and Opportunities. August 20, 2002 Boston USA pp. 35-41

Salvador Miquel, Enrique Bigné, Jean-Pierre Lévy, Antonio Carlos Cuenca, Mª José Miguel - Investigación de Mercados-Mc Graw - Hill /Interamericana de España - 1997

Mobasher B., Jain N., Han, E-H., Srivastava J. Web Mining: Pattern Discovery from World Wide Web Transactions. In International Conference on Tools with Artificial Intelligence, pp. 558-567, New Port 1997

Menasalvas E., Millán S., Peña J., Hadjimichael M., Marbán O. Subsessions: a granular approach to click path analysis. In Proc. WICI'02

E. Menasalvas, O. Marbán , S. Millán , and J. M. Peña "Intelligent Web Mining" in Intelligent Exploration of the Web series Studies in Fuzziness and Soft Computing,. Springer-Verlag 2002 P.S.Szczepaniak, J. Segovia, J. Kacprzyk, L.A. Zadeh – editors

From: Gregory Piatetsky-Shapiro 2001: Interview with Jesus Mena, (WebMiner)

M. Perkowitz and O. Etzioni "Adaptive Web Sites: Automatically Synthesizing Web Pages". In Proceedings of AAAI98.

M. Perkowitz and O Etzioni. "Adaptive web sites: Conceptual cluster mining". In Sixteenth International Joint Conference on Articial Intelligence, Stockholm, Sweden, 1999.
M. Perkowitz and O. Etzioni. "Towards adaptive Web sites: Conceptual framework and case study". In Artificial Intelligences 118, 2000.
Shahabi C., Faisal A., Banaei F., Faruque J. INSITE:A tool for real-time knowledge Discovery from users web navigation. In Proc. VLDB-2000, 2000.
Shahabi Cyrus, Farnoush Banaiei-Kashaani, Jaabed Faruque, Adil Faisal. Feature Matrices: A model for e-Ecient and anonymous web usage mining. Proc. of EC-Web 2001.
M. Spiliopoulou, L. Faulstich, and K. Wilkler. "A data miner analyzing the navigational behaviour of web users". In Proc. Of the Workshop on Machine Learning in User Modelling of the ACAI99, Greece.
M. Spiliopoulou, C. Pohle, and L. Faulstich. "Improving the effectiveness of a web site with web usage mining". In Proceedings WEBKDD99.

Publicaciones que el trabajo ha generado

Durante la etapa de suficiencia investigadora fueron presentados los siguientes artículos en carácter de coautora:

[1] **Sessions Value as measure of web site goal achievement -** 3rd ACIS International Conference on Software Engineering Artificial Intelligence, Networking and Parallel/Distributed Computing - Madrid - España- 2002 SPND (aceptado)

[2] **A granular approach for analyzing the degree of affability of a web site**- The Third International Conference on Rough Sets and Current Trends in Computing RSCTC'2002 - Pennsylvania - USA - Octubre 2002 (aceptado).

[3] **Web Site Goal Achievement Measured by a Sessions Value Algorithm** - Web Mining for Usage Patterns and User Profiles. Edmonton, Alberta, Canada WEBKDD'02 (no aceptado)

[4] **Algoritmo de evaluación de los logros de un sitio web mediante el cómputo del valor de las sesiones de usuarios.** VII Jornadas de Ingeniería del Software y Bases de Datos - El Escorial, Madrid- 2002 -JISBD2002

[5] **Algoritmo de Cómputo del Valor de las Sesiones de Usuarios para Evaluación de los Logros de un Sitio Web -**Conferencia Latinoamericana de Informática - Montevideo, Uruguay -CLEI2002