Revista Educación en Ingeniería
ISSN 1900-8260
Julio a Diciembre de 2013, Vol. 8, N°. 16, pp. 94-104 • © 2013 ACOFI • http://www.educacioneningenieria.org
Recibido: 15/01/2013 • Aprobado: 25/11/2013

CLASIFICACIÓN DIFUSA PARA DESCUBRIR PERFILES DE USUARIOS EN LA WEB

FUZZY CLASSIFICATION TO DISCOVER USER PROFILES ON THE WEB

Oscar Fernando Bedoya Leiva
Universidad del Valle, Cali (Colombia)

Resumen

El crecimiento explosivo de la web, tanto en el tamaño como en el uso, la ha convertido en una rica fuente para procesos de minería de datos. En este artículo se propone una nueva estrategia que combina técnicas de *clustering* y clasificación difusas, para predecir el comportamiento que en términos de navegación tendría un usuario. Además, se hace un análisis comparativo de la estrategia difusa propuesta con una técnica exacta existente. Los resultados de las pruebas evidencian las ventajas en cuanto a la exactitud en la tarea de clasificación y deja ver potenciales aplicaciones. El uso de técnicas difusas permite identificar la similitud y diferencia en un nuevo registro con respecto a cada *cluster*. Esta información es útil, entre otros, para mejorar el diseño de sitios web, realizar una personalización avanzada, definir perfiles de usuario y establecer filtros automáticos de información basados en el contenido y sistemas de recomendación.

Palabras claves: clasificación difusa, minería de datos en la web, perfiles de usuario.

Abstract

Explosive growth in both size and usage of the web has generated the need to use it as a rich source of data for mining processes. In this paper, a new strategy that combines techniques of fuzzy *clustering* and fuzzy classification, to predict user navigation behaviour is proposed. A comparative analysis between the fuzzy strategy proposed and an existing crisp technique is made. The results of several tests showed high accuracy related to the classification ability of the strategy proposed and introduced some potential uses. The use of fuzzy techniques would make it possible to identify the similarity and dissimilarity among new clickstreams and each *cluster*. This information is useful to improve the design of web sites, advanced customizing, user profile definition, automatic filters of information according to the content, and recommender systems, among others.

Keywords: fuzzy classification, user profiles, web mining.

Introducción

El crecimiento explosivo tanto en el tamaño como en el uso de la web ha generado la necesidad de usarla como una rica fuente para procesos de minería de datos. El término *web mining* hace referencia al descubrimiento de información y conocimiento a partir de su contenido, de su estructura y del comportamiento de los usuarios cuando lo acceden para navegar (Cooley et al.,1997; Han et al., 2000; Chikhi et al., 2007; Shinde & Kulkarni, 2008; Han et al., 2008; Kumar & Devi, 2010; Singh et al., 2010; Malik & Rizvi, 2011). A partir de la información registrada en el log de un servidor web, es posible analizar los patrones de navegación de los usuarios. Una entrada de él generalmente incluye, entre otros, la dirección IP, el tiempo de acceso, el método de petición utilizado, el URL de la página a la cual se accede, el protocolo de transmisión y el número de bytes transmitidos.

El comportamiento de los usuarios en la web se puede analizar utilizando diferentes enfoques y técnicas de minería de datos con el propósito de calcular perfiles de usuario (Anandampilai et al., 2007; Escobar-Jeria et al., 2006; Herder y Weinreich, 2005; Song y Shepperd, 2006). La aplicación de estas técnicas sobre el registro del URL generado por cada usuario cuando navega en la web ha sido comúnmente llamada *web usage mining* (Cooley et al., 1997). Uno de sus propósitos es adaptar los sitios web a los usuarios. En particular, las técnicas de *clustering* se han utilizado en *web usage mining* en la etapa de preprocesamiento de los datos. Una vez los datos del archivo log se han preprocesado, se aplican algoritmos de clasificación o asociación, entre otros, como tareas para extraer conocimiento sobre el comportamiento de los usuarios. El algoritmo de *clustering* que se aplica sobre los URL que se almacenan en el archivo log del servidor permite identificar grupos de usuarios en la web.

Normalmente, un algoritmo de *clustering* intenta agrupar datos con características similares de tal forma que se logre dividir un conjunto de N datos en K grupos, en el que los elementos que pertenecen a cada uno de ellos tienen características similares y al mismo tiempo los datos de diversos grupos, poca similitud. Dado que el criterio de decisión de un algoritmo de *clustering* en *web usage mining* es la similitud que existe entre las páginas visitadas por cada uno de los usuarios, es decir, entre las cadenas de clics (*clickstream*), se convierte en un inconveniente calcular de manera exacta dicha similitud. El siguiente ejemplo ilustra la situación: dados dos usuarios que navegan en la web y acceden a las siguientes páginas de un servidor:

http://www.nytimes.com/financialreport/wallstreet/realestate/

http://www.nytimes.com/financialreport/wallstreet/markets/

estos URL reflejan cierto grado de similitud en el registro de navegación de los usuarios, si se tiene en cuenta la ruta previa a la página final de consulta. Un algoritmo de *clustering* que utilice una función de similitud exacta (*crisp*), clasificaría estos URL en grupos distintos. Por otro lado, si usa una función de similitud difusa, podría tener en cuenta la relación que existe entre los URL y clasificarlos en el mismo grupo.

En este artículo se propone una estrategia que combina técnicas de *clustering* y clasificación difusas para predecir, con base en los *clusters* identificados, el comportamiento que en términos de navegación tendría un usuario. El uso de técnicas difusas permite identificar la similitud y diferencia entre un nuevo registro con respecto a los datos almacenados en cada *cluster*. Esta información es útil en el mejoramiento del diseño de sitios y páginas web, personalización avanzada, definición de perfiles de usuario, filtros automáticos de información de acuerdo con el contenido y sistemas de recomendación, entre otros. En lo subsiguiente, este artículo está organizado en secciones. En la sección 2 se presenta el método propuesto para clasificación de URL y en la sección 3, los resultados de las pruebas realizadas; en la sección 4 se discuten los resultados que se obtienen cuando se aplican técnicas difusas y técnicas exactas; finalmente, en la sección 5 se exponen algunas conclusiones y lineamientos para un trabajo futuro.

Metodología

El modelo de clasificación difuso propuesto se basa en los algoritmos y medidas presentadas en (Joshi y Krishnapuram, 1998) y, particularmente, en el algoritmo *Robust Fuzzy c-Medoids* (RCMdd) que

implementa el proceso de *clustering* difuso con el objetivo de identificar grupos de URL similares. Así mismo, se plantea una alternativa para la identificación de grupos de acuerdo con las sesiones de los usuarios en la web.

El proceso de descubrimiento de conocimiento que se propone y se presenta en la figura 1 se divide en las siguientes etapas:

1. Preprocesamiento del archivo log.
2. Identificación de sesiones de usuario a partir del archivo log.
3. *Clustering* difuso para la identificación de los grupos de usuarios.
4. Clasificación difusa de un nuevo usuario a cada grupo identificado

El proceso de descubrimiento propuesto permite que un nuevo usuario sea clasificado utilizando la pertenencia difusa a *clusters* que se han identificado previamente. Como parte del análisis realizado se quiere evaluar qué tan apropiadas resultan las estrategias difusas utilizadas comparadas con técnicas exactas (*crisp*). Utilizando el proceso propuesto, se puede predecir el perfil de un nuevo usuario al conocer los grados de pertenencia difusa de su fragmento de archivo log con respecto a los grupos existentes.

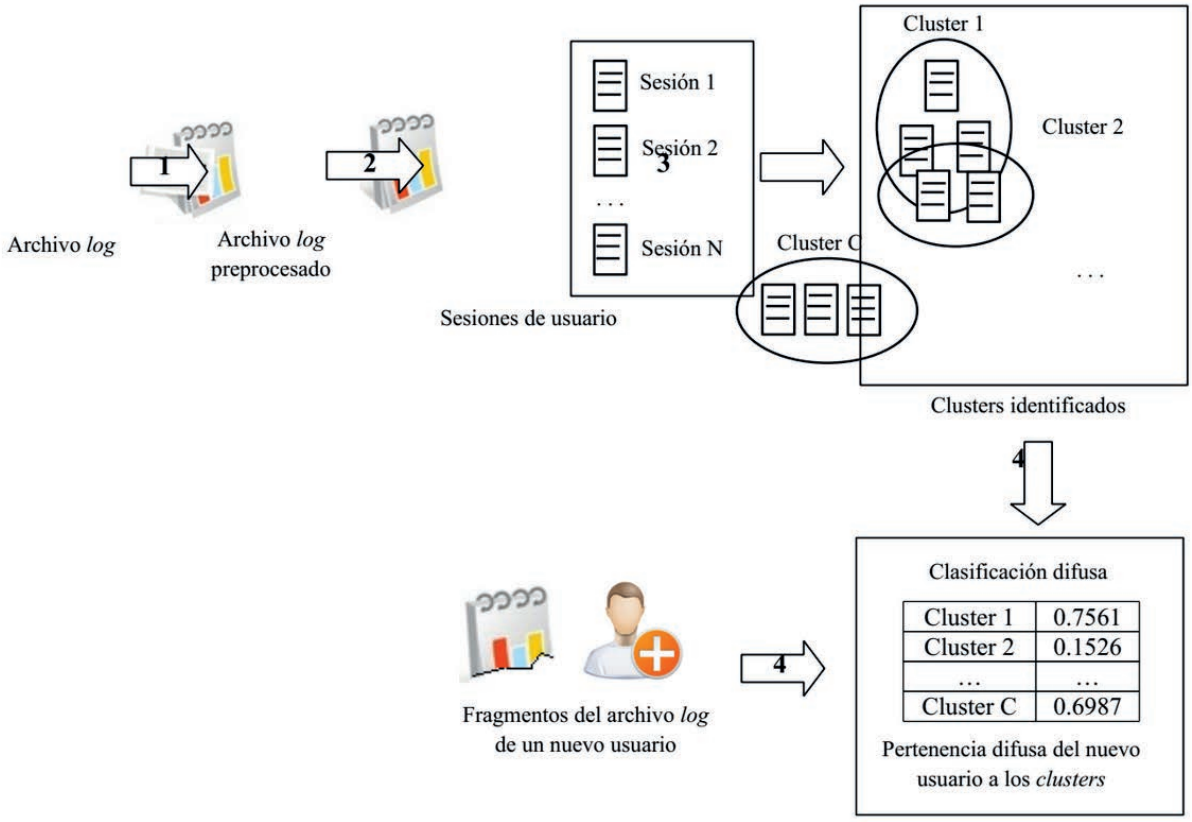


Figura 1. Proceso de descubrimiento propuesto

Preprocesamiento del archivo log

En el log de un servidor web se registran los siguientes datos: dirección IP del usuario, tiempo de acceso, método solicitado (Get, Post), URL de la página accedida, protocolo (http/1.0), código retornado, número de bytes transmitidos.

El preprocesamiento que se lleva a cabo sobre el archivo log consiste en eliminar los registros en los cuales se accede a archivos de imágenes (por ejemplo: gif, jpeg, png), aquellos en los que el acceso a una página web se realiza bajo un método de petición diferente a *Get*, o aquellos cuyo código retornado indica que el URL solicitado no se pudo acceder, es decir, se presenta un código de error.

Identificación de sesiones de usuario a partir del archivo log

Las entradas del archivo log se agrupan en sesiones de usuario. Una sesión de usuario se define como una secuencia de accesos temporalmente compactos. Ya que los servidores web no utilizan nombres de usuarios en los archivos log, es decir que son anónimos, una sesión se define como los accesos desde el mismo IP de modo que el tiempo entre dos accesos consecutivos esté dentro de un umbral previamente especificado. Para los casos de prueba analizados en este artículo se consideró un tiempo de sesión de una hora.

Clustering difuso para la identificación de los grupos de usuarios

Cada URL tiene asignado un número único $j \in \{1,2,\ldots,N_U\}$, donde es el número total de URL válidos, es decir, aquellos que conforman el archivo log después del preprocesamiento. Por lo tanto, la i-ésima sesión de usuario se codifica como un atributo binario N_U-dimensional (vector $s^{(i)}$, donde es 1 si el usuario accedió al j-ésimo URL durante la sesión i, y 0 en caso contrario). La unión de las N_s sesiones extraídas del archivo log se denota con S.

Las medidas de similitud entre dos sesiones $s^{(k)}$ y $s^{(l)}$ intentan incorporar tanto la estructura del sitio web como los URL involucrados. En Joshi y Krishnapuram (2000) se considera primero el caso simple, en el que todos los URL accedidos en las sesiones son distintos. Por lo tanto, se puede usar el coseno del ángulo entre $s^{(k)}$ y $s^{(l)}$ como una medida ($M_{1,kl}$) de similitud. Esto simplemente mide el número de URL comunes accedidos durante las dos sesiones.

Una posible estrategia para estimar la similitud de URL es analizar su contenido. Esto lleva a definir una medida de similitud en la estructura del URL. El sitio web se modela como un árbol con nodos que representan diferentes URL. La similitud entre dos URL se determina midiendo la sobreposición de caminos desde la raíz del árbol hasta los nodos correspondientes. Por lo tanto, se define la similitud difusa entre el i-ésimo y el j-ésimo URL como:

$$S_u(i,j) = min\left(1, \frac{|p_i \cap p_j|}{\max(1, \max(|p_i|, |p_j|) - 1)}\right) \quad (6)$$

donde p denota el camino recorrido desde la raíz hasta el nodo correspondiente al i-ésimo URL y $|p_i|$ indica la longitud de este camino. Así, la similitud difusa entre las sesiones se define por medio de la similitud entre los URL visitados en dos sesiones, con respecto al número total de URL visitados:

$$M_{2,kl} = \frac{\sum_{i=1}^{N_U} \sum_{j=1}^{N_U} s_i^{(k)} s_j^{(l)} S_U(i,j)}{\sum_{i=1}^{N_U} \sum_{j=1}^{N_U} s_j^{(l)}} \quad (7)$$

Para el caso especial en el cual todos los URL accedidos durante la sesión $S^{(k)}$ tienen similitud cero con respecto a los URL accedidos durante la sesión, $S(l), SU(i,j) = 0$ si i≠j. Por lo tanto, se reduce a:

$$M_{2,kl} = \frac{\sum_{i=1}^{N_U} s_i^{(k)} s_i^{(l)}}{\sum_{i=1}^{N_U} s_i^{(k)} \sum_{j=1}^{N_U} s_j^{(l)}} \quad (8)$$

y cuando las dos sesiones son idénticas, este valor se simplifica así:

$$M_{2,kk} = \frac{1}{\sum_{i=1}^{N_U} s_i^{(k)}} \quad (9)$$

lo cual puede ser considerablemente pequeño dependiendo del número de URL accedidos. Según Joshi y Krishnapuram (2000), esta medida no es intuitiva pues idealmente la similitud debería ser máxima para dos sesiones idénticas. Además, la similitud será subestimada en sesiones que comparten algunos URL idénticos. En general, para sesiones en las que la similitud entre los URL es baja, proporciona una medida alta y exacta de similitud. Por otro lado, cuando la similitud entre URL es alta, logra una medida aun mejor. Por esto se usa el máximo entre M_1 y M_2 como medida de similitud. Para realizar el *clustering*, la similitud se trasforma en una medida de disimilitud.

$$d_s^2(k,l) = (1 - M_{kl})^2 \quad (10)$$

Con esta medida de disimilitud es posible que para dos sesiones distintas se tenga disimilitud cero. Así ocurre cuando:

$$\sum_{i=1}^{N_U}\sum_{j=1}^{N_U} s_i^{(k)} s_j^{(l)} S_U(i,j) = \sum_{i=1}^{N_U} s_i^{(k)} \sum_{j=1}^{N_U} s_j^{(l)} \quad (11)$$

que equivale a:

$$\sum_{j=1}^{N_U} s_i^{(k)} s_j^{(l)} S_U(i,j) = s_i^{(k)} \sum_{j=1}^{N_U} s_j^{(l)} \text{ donde i:1,2,…,} N_U \quad (12)$$

Esto es particularmente verdadero si la similitud difusa obtenida a partir de la ecuación (6) es 1, para todos los URL accedidos en las dos sesiones. Al realizar el proceso de *clustering* de acuerdo con las sesiones, éste se trabaja a un nivel de granularidad más alto que al hacerlo de acuerdo con los URL, ya que se intenta agrupar las sesiones más similares, y no los URL como ocurre en RCMdd. Es decir, cada URL contribuye con un determinado peso para hacer que la sesión a la cual pertenece se acerque o se aleje de cada una de las otras sesiones.

En el algoritmo propuesto, las sesiones de usuarios se asignan al *cluster* más cercano, lo cual crea *clusters* de la forma:

$$X_i = \left\{ s^{(k)} \in S \mid d_{ik} < d_{jk}, \forall_{i \neq j} \right\} \text{ donde } 1 < i < C \quad (13)$$

Las sesiones en el *cluster* se consolidan en una sesión que servirá de perfil P. Los componentes de son los pesos de los URL que representan el número de accesos de un URL durante las sesiones de X_i.

$$P_{ij} = p\left(s_j^{(k)} = 1 \mid s^{(k)} \in X_i\right) = \frac{|X_{ij}|}{|X_i|} \quad (14)$$

donde los pesos de los URL P_{ij} miden la importancia de un URL dado en el i-ésimo perfil.

Clasificación difusa de un nuevo usuario a cada grupo identificado

Al clasificar un nuevo ejemplo, es decir, una nueva sesión de usuario preprocesada, se utiliza la medida de disimilitud entre sesiones para calcular las distancias entre cada uno de los *clusters* y el nuevo ejemplo. En este caso, un *cluster* se trata como una sesión, ya que es el consolidado de la información de todos los URL que le pertenecen. Este proceso se puede describir en el algoritmo de la figura 2.

```
for j=1 to C do
    calcule d[j][ejemplo] usando Ecuación (10)
end-for
for j=1 to C do
    despliegue el valor de pertenencia difusa del
    ejemplo al cluster j
end-for
calcule el menor d[j][ejemplo]
asigne ejemplo al cluster j
```

Figura 2. Algoritmo de clasificación

Resultados

En esta sección se presentan los resultados obtenidos al aplicar los algoritmos de clasificación difuso y exacto sobre diferentes conjuntos de datos. El objetivo de las pruebas fue comparar el comportamiento de los algoritmos exacto y difuso al clasificar usuarios nuevos con base en fragmentos de sus sesiones. De acuerdo con las cuatro actividades que formaron parte del proceso de descubrimiento propuesto en la figura 1, se va a evaluar solamente la etapa 4. Para efectos del objetivo de las pruebas, se supone que las etapas 1, 2 y 3 que permiten conseguir los *clusters* se han alcanzado exitosamente.

Para evaluar el comportamiento del algoritmo de clasificación se utilizaron archivos log. Los registros de éstos fueron la entrada inicial del proceso de *clustering* difuso descrito en la sección 2. En las tablas 1, 2 y 3 se presentan los URL que pertenecen a sesiones similares que se agruparon en 3 *clusters* a partir de uno de los archivos log.

Tabla 1. Datos agrupados en el primer *cluster*

Cluster 1	
1	/cnn.com/ALLPOLITICS/index.html
2	/cnn.com/2003/ALLPOLITICS/04/29/mideast.roadmap/mideast0.html
3	/cnn.com/2003/ALLPOLITICS/04/29/mideast.roadmap/mideast2.html
4	/cnn.com/2003/ALLPOLITICS/04/29/mideast.roadmap/mideast1.html
5	/cnn.com/HEALTH/index.html
6	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins.html
7	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins1.html
8	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins2.html
9	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins3.html
10	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins4.html
11	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins5.html
12	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins6.html
13	/cnn.com/HEALTH/04/29/guatemalan.twins.ap/twins7.html
14	/bookstore.gpo.gov/index.html
15	/bookstore.gpo.gov/science/book1.html
16	/bookstore.gpo.gov/tecnical/book2.html
17	/bookstore.gpo.gov/tecnical/book3.html

Tabla 2. Datos agrupados en el segundo *cluster*

Cluster 2	
1	/cnn.com/ALLPOLITICS/index.html
2	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber.html
3	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber1.html
4	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber2.html
5	/cnn.com/TRAVEL/ski.report/ski1.html
6	/cnn.com/TRAVEL/ski.report/xcountry.html
7	/ebussiness.com/index.html
8	/ebussiness.com/local/ny/printers/cannon/692.html
9	/ebussiness.com/local/ny/printers/cannon/692a.html
10	/ebussiness.com/local/ny/printers/cannon/692b.html
11	/ebussiness.com/local/printers/cannon/720.html
12	/ebussiness.com/local/printers/cannon/720.html
13	/bookstore.gpo.gov/tecnical/book1.html
14	/nytimes.com/WEATHER/index.html
15	/nytimes.com/WEATHER/Europe/frct.html
16	/nytimes.com/WEATHER/Europe/temp.html
17	/nytimes.com/WEATHER/NAmerica/temp.html

Tabla 3. Datos agrupados en el tercer *cluster*

Cluster 3	
1	/cnn.com/ALLPOLITICS/index.html
2	/cnn.com/2003/ALLPOLITICS/04/29/ONU01.html
3	/cnn.com/2003/ALLPOLITICS/04/29/ONU01.html
4	/cnn.com/2003/ALLPOLITICS/04/29/ONU01.html
5	/electronicsusa.com/index.html
6	/electronicsusa.com/cdrw/HP/mo1.html
7	/electronicsusa.com/cdrw/HP/mo2.html
8	/electronicsusa.com/cdrw/HP/mo3.html
9	/electronicsusa.com/cdrw/HP/mo4.html
10	/wpost.com/SPORTS/index.html
11	/wpost.com/SPORTS/nba/playoffs/minlalG1.html
12	/wpost.com/SPORTS/nba/playoffs/minlalG2.html
13	/wpost.com/SPORTS/index.html
14	/wpost.com/SPORTS/nba/playoffs/minlalG1.html
15	/wpost.com/SPORTS/nba/playoffs/minlalG2.html
16	/bookstore.gpo.gov/tecnical/book3.html
17	/bookstore.gpo.gov/tecnical/book4.html

Para ilustrar el proceso llevado a cabo, en la tabla 4 se muestra una sesión de usuario, cuyos fragmentos son la entrada de los algoritmos de clasificación difuso y exacto.

Tabla 4. Sesión de un nuevo usuario

Sesión de usuario por clasificar	
1	/cnn.com/ALLPOLITICS/index.html
2	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber.html
3	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber1.html
4	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber2.html
5	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber3.html
6	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber4.html
7	/cnn.com/2003/ALLPOLITICS/04/29/bush.amber.ap/bushamber5.html
8	/cnn.com/HEALTH/04/29/topic1.html
9	/cnn.com/HEALTH/04/29/topic2.html
10	/cnn.com/HEALTH/04/29/topic3.html
11	/cnn.com/HEALTH/04/29/topic4.html
12	/cnn.com/HEALTH/04/29/topic5.html
13	/cnn.com/HEALTH/04/29/topic6.html
14	/cnn.com/HEALTH/04/29/topic7.html
15	/bookstore.gpo.gov/tecnical/book.html
16	/bookstore.gpo.gov/tecnical/book1.html
17	/bookstore.gpo.gov/tecnical/book23.html
18	/bookstore.gpo.gov/tecnical/book31.html
19	/bookstore.gpo.gov/tecnical/book45.html
20	/bookstore.gpo.gov/tecnical/book59.html

Para evaluar el poder de predicción de la técnica de clasificación se consideró como entrada para los algoritmos exacto y difuso un conjunto de fragmentos de sesiones de usuario de diferente longitud. La longitud de los fragmentos correspondió al 25%, 50% 75% y 100% de la sesión original. El objetivo de realizar las pruebas de esta forma es evaluar la habilidad de las técnicas para clasificar un nuevo usuario web sin necesidad de conocer el recorrido completo de su navegación. En la tabla 5 aparecen valores que representan disimilitud entre diferentes porcentajes de la sesión del nuevo usuario y cada uno de los *clusters*.

Tabla 5. Disimilitud con respecto a cada *cluster*

Porcentaje del registro original	*Cluster 1*	*Cluster 2*	*Cluster 3*
25%	0,7403362	0,7214033	0,7857142
50%	0,9054894	0,9104237	0,9152789
75%	0,8946779	0,9134454	0,9285714
100%	0,9350840	0,9464285	0,9510084

De acuerdo con los valores de la tabla anterior, se presenta para cada porcentaje de registro el cluster más y el menor similar se presenta en la tabla 6.

Tabla 6. Similitud y diferencia con *clustering* difuso

Porcentaje del registro original	***Cluster* más similar**	***Cluster* menos similar**
25%	2	3
50%	1	3
75%	1	3
100%	1	3

La tabla 7 muestra los resultados obtenidos por el clasificador exacto. Independientemente del porcentaje de la sesión del nuevo usuario, la decisión fue siempre asignar al *cluster* 2.

Tabla 7. Similitud y diferencia con *clustering* exacto

Porcentaje del registro original	*Cluster 1*	*Cluster 2*	*Cluster 3*
25%	0	1	0
50%	0	1	0
75%	0	1	0
100%	0	1	0

El clasificador difuso necesita, al menos, un 50% de los URL de la sesión del nuevo usuario para clasificarlo de manera correcta. Por otra parte, el clasificador exacto no puede detectar ésta como la mejor opción. Para todos los porcentajes de la sesión consideró al *cluster* 2 por tener más URL similares entre este grupo y el nuevo usuario. Esto indica que el clasificador difuso tiene en cuenta aquellos URL que no son idénticos pero sí similares. Esto permite que la distancia total entre cada grupo y el nuevo usuario se aproxime más al valor real.

En la figura 3, se presentan los resultados de la clasificación sobre fragmentos de múltiples sesiones de usuario. Se tomaron 10 casos de prueba para cada uno de los porcentajes por considerar. La clasificación es exitosa a partir del 75% de la sesión completa del

nuevo usuario. Con el 50% se consiguió la mejor clasificación en el 80% de los casos de prueba. Por otro lado, el clasificador exacto tan sólo logró la mejor asignación del nuevo usuario en el 40% de los casos, en los que se consideró el 100% de la sesión, y en el 20% de los casos, cuando consideró el 25%.

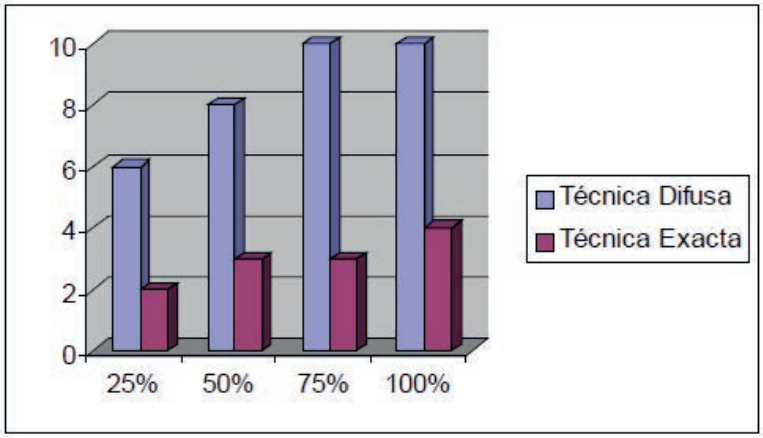


Figura 3. Cuadro comparativo *Clustering* difuso versus exacto

Se modelaron otros diez casos de prueba para cada uno de los porcentajes, con diferentes *clusters* y nuevos usuarios por clasificar, cuyas sesiones estaban compuestas por 40 URL. Los resultados obtenidos se pueden observar en la figura 4.

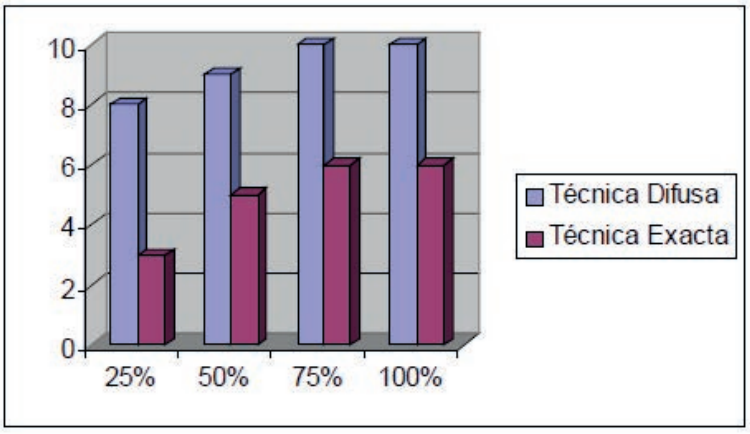


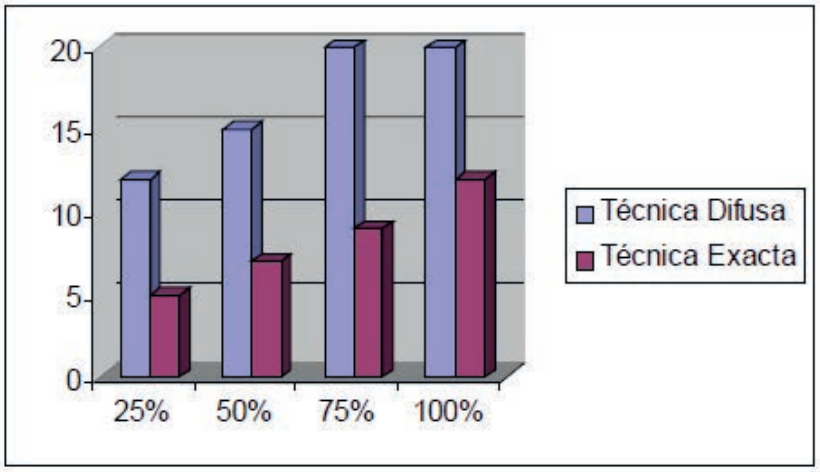
Figura 4. *Clustering* difuso versus exacto

Discusión

La técnica difusa tiene un mejor rendimiento al momento de clasificar con respecto a la técnica exacta. Por ejemplo, en nueve de los diez casos en los que se consideró el 50% de la sesión, y en ocho de los diez casos en los se tuvo presente tan sólo el 25%, la clasificación difusa fue exitosa.

El algoritmo de clasificación se ejecutó sobre otro conjunto de prueba consistente en un archivo log de 250 URL. Las pruebas de clasificación se hicieron con 20 casos por cada porcentaje, los cuales corresponden a fragmentos de sesiones de 50 URL. Para este nuevo conjunto, los resultados obtenidos se describen en la figura 5.

Figura 5. *Clustering* difuso versus exacto



El algoritmo de *clustering* difuso se aplicó sobre los 11079 URL contenidos en otro archivo log de prueba. La salida del algoritmo consistió en 172 *clusters*. Con base en estos *clusters* identificados se ejecutó el algoritmo de clasificación difusa con diferentes porcentajes de la sesión de un nuevo usuario.

Los resultados obtenidos con el algoritmo de clasificación difuso se presentan en la tabla 8.

Tabla 8. Clasificación difusa con diferentes fragmentos de sesiones

Porcentaje de la sesión total a clasificar (%)	*Cluster* elegido por el clasificador	Disimilitud difusa entra la sesión del nuevo usuario y el *cluster* elegido
25	76	0,4355622
50	24	0,3914759
75	79	0,2514783
100	79	0,2498748

Los resultados indican que la decisión del clasificador, también en este caso, no varía cuando la entrada corresponde al 75% y 100% de la sesión total del nuevo usuario. Sin embargo, para el 25% y el 50% de la sesión se obtuvieron resultados diferentes. Para analizar este caso particular, se calculó además la disimilitud difusa entre los clusters 24 y 79, a los cuales se asignan las sesiones correspondientes al 50% y 75%, respectivamente. La disimilitud difusa entre los clusters 24 y 79 es de 0.1230457, lo que indica que son clusters muy similares; esto explica la asignación hecha por el algoritmo

Otra prueba se hizo sobre un archivo log de 17330 URL's, dividido en 326 clusters de usuarios por el algoritmo de *clustering* y con una nueva sesión de usuario de longitud de 100 URL's. Los resultados obtenidos se muestran en la tabla 9.

Tabla 9. Clasificación difusa con diferentes fragmentos de sesiones

Porcentaje de la sesión total a clasificar	*Cluster* elegido por el clasificador	Disimilitud difusa entre la sesión del nuevo usuario y el *cluster* elegido
25%	110	0,5234121
50%	304	0,5914759
75%	45	0,3514787
100%	45	0,2498743

La decisión del clasificador difuso fue la misma cuando se consideró el 75% y el 100%. En este sentido, el algoritmo es útil para identificar grupos similares a una entrada dada sin tener que contar con el 100% de la sesión.

Conclusiones

En este artículo se presentó un modelo que combina las técnicas *clustering* difuso y clasificación difusa para hacer minería de datos en la web. En el enfoque propuesto es posible calcular la similitud y diferencia que se presenta entre los registros almacenados en el archivo log de un servidor web, información que puede ser útil en el diseño e implementación de aplicaciones personalizadas.

El modelo propuesto fue evaluado con múltiples conjuntos de prueba, diferentes en el número de *clusters*, en el tamaño del archivo log y en la longitud de las sesiones de usuario. Todos los resultados fueron correctos a partir del 75% del tamaño del registro; incluso en la mayoría de las pruebas a partir del 50% del tamaño del registro se obtuvo el resultado correcto. De los resultados obtenidos es posible inferir que a mayor distancia intercluster, mayor calidad en el proceso de clasificación. El modelo propuesto podría ser útil en el proceso de personalización de sitios web en la medida en que permite identificar el perfil del usuario durante el proceso de navegación.

Se está trabajando en implantar el algoritmo de clasificación difuso en tiempo real en un sitio web. Se realizará otra extensión de este trabajo para extraer textos ligados a imágenes y proveer mecanismos de anotación automática en sistemas de etiquetado de imágenes con el fin de interpretar y representar mediante símbolos el contenido semántico de las mismas.

Referencias

Anandampilai, B. Shunmuganathan, K.L. Vasudevan, V. (2007). A Mutiagent System for Web Mining Using Adjustable User Profile and Vibrant Confederacy. In Proceeding of International Conference on Computational Intelligence and Multimedia Applications, Vol 1, ACM Press, pp. 139-144.

Chikhi Nacim, Rothenburger Bernard, Aussenac-Gilles Nathalie. (2007). A Comparison of Dimensionality Reduction Techniques for Web Structure Mining, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp.116-119.

Cooley R., Mobasher B., and Srivastava J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. In Proc. of the 9th IEEE Int. Conf. on Tools with Artificial Intelligence, ICTAI'97.

Escobar-Jeria V., Martin-Bautista M., Sánchez D. y Vila M. (2006). Web Mining Application with Fuzzy Logic. Primera Conferencia Internacional sobre Ciencias y Tecnologías Multidisciplinarias de la Información (InScit). pp 77-81.

Han J., Pei J., Mortazavi-Asl B., Chen O., Dayal U., and Hsu M. (2000). FreeSpan: frequent pattern-projected sequential pattern mining. KDD 2000. pp. 355-359.

Han Qingtian; Gao XiaoYan; Wu Wenguo. (2008). Study on Web Mining Algorithm based on Usage Mining. Computer-Aided Industrial Design and Conceptual Design, 2008. CAID/CD 2008. 9th International Conference, pp.1121-1124.

Herder E. Weinreich GH. (2005). Interactive Web Usage Mining with the Navigation Visualizer. In Proc. of the CHI 2005 Conference on Human Factors in Computing Systems, ACM Press.

Joshi A. and Krishnapuram R. (1998). Robust Fuzzy Clustering Methods to Support Web. SIGMOD Workshop on Data Mining and Knowledge Discovery.

Joshi A. and Krishnapuram R. (2000). On Mining Web Access Logs. In Proc. of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 63-69.

Kumar Sravan and Devi Naveena. (2010). Learner's Centric Approach for Web Mining et al. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1(2).

Malik, S.K.; Rizvi, S.A.M. (2011). Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation. Computational Intelligence and Communication Networks (CICN), 2011 International Conference, 7-9 Oct. 2011. pp. 465-469.

Shinde Subhash, Kulkarni U.V. (2008). A New Approach For On Line Recommender System in Web Usage Mining, Proceedings of the 2008 International

Conference on Advanced Computer Theory and Engineering pp. pp. 973-977.

Singh, B.; Singh, H.K. (2010). Web Data Mining research: A survey. Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference, pp.1-10.

Song, Q., Shepperd, M.,(2006). Mining web browsing patterns for e-commerce. Comput. Ind. 57, pp. 622–630.

Sobre el autor

Oscar Fernando Bedoya Leiva
Ingeniero de Sistemas. Magíster en Ingeniería con énfasis en Ingeniería de sistemas y computación. Profesor de la Escuela de Ingeniería de Sistemas y Computación de la Universidad del Valle, Cali (Colombia). Integrante del grupo de investigación Bioinformática.
oscar.bedoya@correounivalle.edu.co