Aplicación de un algoritmo de extracción de reglas difusas para minería de uso web

Cristóbal José Carmona del Jesus **Pedro González García**
Víctor Manuel Rivas Santos **María José del Jesus Díaz**

Depto. Informática. Universidad de Jaén, {ccarmona, pglez, vrivas, mjjesus}@ujaen.es

Resumen

El gran impulso de las nuevas tecnologías hace que la World Wide Web sea actualmente una de las fuentes de información más importantes en el momento. Dentro del campo de la minería de uso web, que analiza la información de la WWW, se aplican diferentes tipos de técnicas. En nuestra aproximación se procesa la base de datos MSNBC mediante una técnica evolutiva capaz de extraer conocimiento de interés sobre el uso de la web y expresarlo en forma de reglas difusas de descripción de subgrupos. Los resultados obtenidos en esta primera aproximación se caracterizan por la obtención de reglas que cubren la mayoría de los ejemplos con alto nivel de relevancia y confianza.

Palabras Clave: Minería de Uso Web, Patrones Frecuentes, MSNBC, Descubrimiento de Subgrupos, Sistemas Evolutivos Difusos.

1 Introducción

El uso de las nuevas tecnologías y la explosión de información que ha surgido a raíz de la World Wide Web, provoca que sea una de las fuentes de datos más importantes que existen en la actualidad. Por este motivo se ha convertido en una de las áreas de investigación más importantes dentro del descubrimiento de información en grandes bases de datos.

En minería web se distinguen tres dominios de extracción de conocimiento de acuerdo con la naturaleza de los datos [6][31]: *minería de contenido web* que se centra en la extracción de conocimiento del contenido de los documentos o de sus descripciones, *minería de estructura web* que intenta descubrir conocimiento a partir de la forma en que se enlazan (o se incluyen) unos documentos con otros y la *minería de uso web* que obtiene el conocimiento de los ficheros de registro de los servidores.

En este trabajo nos centraremos en el dominio de la *minería de uso web*, aplicando una técnica de minería de datos descriptiva como el descubrimiento de subgrupos. Nuestra implementación utiliza técnicas de *computación flexible*, como los *algoritmos evolutivos* [16][18] y la *lógica difusa* [43][44], que facilitan la interpretación por los expertos del conocimiento extraído y mejoran los resultados obtenidos por los algoritmos clásicos en resolución de bases de datos con una alta dimensionalidad. Aplicaremos esta técnica sobre la base de datos del *MSNBC Anonymous Web Data*, que es una base de datos que se caracteriza por su alta dimensionalidad y que describe las páginas visitadas por los usuarios que accedieron a la web del MSNBC[1] el 28 de septiembre del 1999.

El artículo se organiza de la siguiente forma: En la sección 2 se definen los aspectos más importantes de la minería de uso web. En la sección 3 se introduce el problema que pretendemos resolver, en la sección 4 se define el descubrimiento de subgrupos, a continuación en la sección 5 se describe la aproximación del algoritmo SDIGA. En la sección 6 se encuentra la experimentación llevada a cabo sobre la base de datos *MSNBC Anonymous Web Data*. Para concluir, en la sección 7 se detallan las conclusiones y posibles trabajos futuros a realizar.

2 Minería de uso web

La minería de uso web extrae conocimiento de datos de utilización de la web o *web logs*. Un *web log* es una lista de datos de referencias de páginas. En estos *logs* se pueden examinar, desde una perspectiva del cliente o desde una perspectiva de servidor (la más destacada

[1] http://www.msnbc.msn.com

de ellas es la del lado del servidor), las páginas alojadas en cada uno de los sitios, en la que el proceso de minería de uso web descubre información sobre el uso y el acceso que los usuarios hacen del sitio.

La minería de uso web es una de las áreas de investigación que más desarrollo está teniendo en la actualidad para descubrir conocimiento oculto dentro de la Web [38], y consta de tres fases distintas y separadas que identificamos de forma secuencial: Preparación de los datos, Minería de datos y Evaluación e Interpretación.

La actividad más habitual llevada a cabo dentro de la minería de uso web es la mejora del diseño del sitio, aunque según Facca y Lanzi [11], los ficheros *web log* se pueden aplicar para mejorar en distintas áreas de aplicación de la minería de uso web. Para la resolución de estos problemas se pueden aplicar distintos tipos de técnicas de minería de datos. Estas áreas y técnicas son las siguientes:

- Personalización, en la que analizando la secuencia de páginas a las que accede un usuario se pueden desarrollar un perfil sobre el mismo. Dentro de este área se pueden encontrar entre otras, una técnica de *fuzzy clustering* que obtiene perfiles de usuario y los usa para generar una web de interés para el usuario automáticamente [21], una técnica de reglas de asociación, donde el soporte su ajusta automáticamente [29], o también, una técnica basada en la agregación de perfiles de usuario, con *clustering* y reglas de asociación [34]. Entre otras técnicas relacionadas con la lógica difusa también se pueden encontrar distintos trabajos como por ejemplo el de Delgado[9], que mediante reglas de asociación difusas intenta obtener el comportamiento de los usuarios en la web, o trabajos como [32] que trata los perfiles de usuario en la Web. En [33] se puede encontrar una revisión de las técnicas utilizadas dentro de este campo.
- *Caching*, donde se estudian técnicas de clasificación para saber qué páginas visitará el usuario y cargarlas con anterioridad. Dentro de este área se puede encontrar un modelo de predicción para mejorar los sistemas de Internet basados en técnicas de reglas de asociación y clasificadores [26], además de predicción de caching basada en *n-gram* que es un algoritmo de reglas de asociación basado en la filosofía GSDF [42], entre otras.
- Diseño, donde se estudia la calidad y efecitividad global de las páginas del sitio, Podemos encontrar una técnica para descubrir patrones de secuencias y clasificadores para mejorar el diseño de un sitio web [5], técnicas mediante clasificadores para obtener un sistema web adaptativo para facilitar los accesos de los usuarios [12], técnicas que miden la calidad de navegación, del servicio o de la página mediante técnicas de patrones de secuencia [39], o, técnicas que se basan en el *backtracking* de los usuarios para descubrir comportamiento para el diseño, buscando patrones de secuencia en las páginas [40].
- e-comercio, donde podemos encontrar entre otras, una técnica de clustering para medir las relaciones secuenciales en las páginas web, basándose en lógica difusa y algoritmos genéticos [17].

Considerando el problema que pretendemos resolver, donde se procesa la base de datos del MSNBC, que describe las páginas visitadas por los usuarios que accedieron a la web el 28 de septiembre del 1999, se aplicará la técnica de descubrimiento de subgrupos para obtener reglas que sean capaces de mostrar a los expertos los patrones de navegación para este sitio más relevantes dentro de la base de datos. El descubrimiento de subgrupos es una técnica de inducción descriptiva, cuya aplicación dentro del campo de la minería de uso web es novedosa y hasta el momento desconocida dentro de este área de aplicación.

3 Problema a resolver

El problema que se pretende resolver consiste en la extracción de patrones frecuentes en la base de datos *MSNBC Anonymous Web Data*, disponible en el repositorio UCI[2] [2]. Esta base de datos, obtenida de los ficheros *log* de los servidores del portal MSNBC y preprocesada y almacenada como un conjunto de sesiones de usuario, describe las páginas visitadas por los usuarios que accedieron a la web del MSNBC el 28 de septiembre del 1999. Las características principales de esta base de datos se pueden observar en la tabla 1.

Tabla 1: Características principales de MSNBC

Información Relevante	
Número de usuarios	989818
Número medio de visitas por usuario	5,7
Longitud mínima de peticiones por sesión	1
Longitud máxima de peticiones por sesión	500
Número de URLs por categoría	10 a 5000

La base de datos MSNBC, consiste en un conjunto de datos de longitud variable con información sobre 989818 usuarios, donde para cada usuario se almacena la información de una *sesión*, es decir, las páginas visitadas por el usuario en un periodo de 24 horas, por orden de petición. En esta base de datos, cada una

[2] http://www.ics.uci.edu/∼mlearn/MLRepository.html

de las páginas se asocia a una categoría (cat) a la que se asigna un código (cód); así la categoría frontpage se codifica con el número 1, la categoría news con el número 2 y así sucesivamente hasta las 17 categorías existentes, tal y como muestra la tabla 2. Dentro de cada categoría, de las 17 que hay, se pueden encontrar asociadas un número de páginas web que pueden oscilar entre 10 y 5000, y el número de peticiones realizadas por un usuario en una sesión, pueden ir desde 1 como mínimo hasta 500 como máximo, es decir, una sesión puede tener un tamaño que oscila entre 1 y 500.

Tabla 2: Códigos para las categorías del MSNBC

cat	cód	cat	cód	cat	cód
frontpage	1	misc	7	summary	13
news	2	weather	8	bbs	14
tech	3	health	9	travel	15
local	4	living	10	msn-news	16
opinion	5	business	11	msn-sports	17
on-air	6	sports	12		

Una de las ventajas de usar la base de datos MSNBC, es la gran difusión que tiene dentro de la literatura actual con la aplicación de distintas técnicas para la resolución de problemas de minería de uso web, como por ejemplo:

- Búsqueda de secuencias de patrones mediante el algoritmo FS-Miner [10] que aplica técnicas de reglas de asociación para la aplicación de *caching* y *pre-fetching* sobre los *web logs*.
- Descubrimiento de órdenes parciales en las secuencias de las sesiones de usuario mediante estructuras [15].
- Búsqueda de patrones frecuentes [19][20] que presenta tres algoritmos capaces de obtener patrones, secuencias y representaciones gráficas, respectivamente.
- Técnicas de *rough clustering hierarchy* de datos secuenciales [25].
- Hiperenlaces adaptativos mediante las secuencias de acceso [30] que es una aproximación mediante teoría de grafos para mejorar la navegación.
- Técnicas de *clustering* basadas en medidas de probabilidad para la obtención de resultados [35].
- Descubrir patrones para describir el comportamiento de los usuarios mediante técnicas de *clustering* [37].

Como el objetivo es la obtención de patrones frecuentes mediante reglas interpretables tras la aplicación de descubrimiento de subgrupos, ha sido necesario realizar una preparación de los datos, para convertir la base de datos de longitud variable a una base de datos con tamaño fijo. De esta forma obtenemos un conjunto de datos de 17 variables (categorías) y 989818 usuarios (sesiones), donde cada categoría almacena el número de peticiones realizadas por el usuario para cada sesión.

El descubrimiento de subgrupos tiene la finalidad de obtener subgrupos de la población, donde dando el conjunto de datos y una propiedad de esos datos en la que el usuario esté interesado se buscan subgrupos que sean interesantes para el usuario en el sentido de que tengan una distribución estadística inusual respecto a la propiedad resaltada por el usuario. Esta filosofía se puede asemejar en cierto modo a la extracción de patrones frecuentes, tal y como se puede observar en los trabajos de Iváncsy [19][20]. En estos trabajos, se obtienen patrones frecuentes partiendo de unos determinados niveles de umbral y confianza, consiguiendo así los mejores patrones dentro de la base de datos. Por contra, en nuestra aproximación se obtienen reglas basándose exclusivamente en un umbral de soporte, el cual detallaremos más adelante.

Como se detalla en la sección 5, la aproximación se basa en técnicas de *soft-computing* para la resolución del problema. Dentro de este campo, SDIGA utiliza las técnicas de lógica difusa y algoritmos genéticos. Mediante estas técnicas se obtiene un procesamiento más flexible para la obtención de resultados, ya que permiten tratar los problemas asemejándose más a los problemas que existen en la realidad. Tal y como se puede observar en [36], las técnicas de *soft-computing* aplicadas a los problemas de minería web para el reconocimiento de patrones son muy relevantes. No solo la lógica difusa y los algoritmos genéticos, sino también técnicas como las redes neuronales o los conjuntos rugosos [1].

4 Descubrimiento de Subgrupos

Las técnicas de descubrimiento de subgrupos generan modelos basados en reglas cuya finalidad es descriptiva, empleando una perspectiva predictiva para obtenerlos [28]. Las reglas utilizadas en la tarea de descubrimiento de subgrupos tienen la forma $Cond \rightarrow Clase$, donde la propiedad de interés para el descubrimiento de subgrupos es el valor de la Clase que aparece en el consecuente de la regla [13][27] y el antecedente es una conjunción de variables (parejas atributo-valor) seleccionadas entre las variables del conjunto de datos.

El concepto de descubrimiento de subgrupos fue formulado inicialmente por Klösgen [23] y Wrobel [41], como "dado un conjunto de datos y una propiedad de

esos datos que sea de interés para el usuario, buscar subgrupos que sean de mayor interés para el usuario". En este sentido, se dice que un subgrupo es interesante cuando tiene una distribución estadística inusual respecto a la propiedad en la que estamos interesados. El objetivo es descubrir propiedades características de subgrupos construyendo reglas individuales sencillas (con una estructura comprensible y con pocas variables), altamente significativas y con soporte alto (que cubran muchas instancias de la clase objetivo).

Entre los distintos modelos más interesantes que obtienen descripciones de subgrupos representados de diferentes formas, tenemos: *AprioriSD* [22], el cual adopta un algoritmo de aprendizaje de reglas de asociación al descubrimiento de subgrupos; *CN2-SD* [28], cuyo propósito es inducir subgrupos en forma de reglas utilizando como medida de calidad la relación entre el ratio de verdaderos positivos y el ratio de falsos positivos; y *SD-MAP* [3], que es un algoritmo exhaustivo de extracción de reglas de descubrimiento de subgrupos basado en el algoritmo *FP-Growth* de reglas de asociación.

Para llevar a cabo una tarea de descubrimiento de subgrupos, es necesario definir cuatro elementos fundamentales [4]:

- El tipo de variable objetivo. Puede ser binaria, nominal o numérica. Dependiendo del número de variable objetivo son posibles diferentes cuestiones analíticas.
- El lenguaje de descripción que especifique los individuos de la población de referencia que pertenecen al subgrupo.
- La función de calidad que mide el interés de los subgrupos. Se han propuesto gran variedad de funciones de calidad [14][24][23]. Las funciones de calidad aplicables en cada caso vienen determinadas por el tipo de variable ojetivo y el problema analítico.
- La estrategia de búsqueda. Este aspecto es muy importante, puesto que el espacio de búsqueda crece exponencialmente con el número de posibles expresiones que pueden formar parte de una descripción de un subgrupo.

Uno de los aspectos más importantes de cualquier enfoque de inducción de reglas que describan subgrupos es la elección de las medidas de calidad a utilizar, tanto para seleccionar las reglas, como para valorar los resultados obtenidos en el proceso.

5 SDIGA: Subgroup Discovery Iterative Genetic Algorithm

El algoritmo SDIGA es un modelo evolutivo para la extracción de reglas difusas para la tarea de descubrimiento de subgrupos. En [8] puede encontrarse una descripción detallada del mismo.

Este algoritmo puede trabajar con reglas difusas o nítidas y con distintas estructuras de regla dependiendo del problema que deseemos resolver (DNF o canónicas). En este trabajo utilizaremos reglas difusas canónicas en las que el antecedente está compuesto por una conjunción de pares variable-valor.

El modelo evolutivo propuesto sigue el enfoque IRL, en el que el que cada cromosoma representa una regla, la solución del AG es el mejor individuo, y la solución global está formada por los mejores individuos de una serie de ejecuciones sucesivas [7]. Así, el núcleo de SDIGA es un AG que utiliza una etapa de post-procesamiento basada en una búsqueda local que incrementa la generalidad de la regla obtenida. Este AG híbrido se incluye en un proceso iterativo para la extracción de un conjunto de reglas difusas de descripción de subgrupos soportado por zonas distintas del espacio de búsqueda, que extrae reglas mientras se cubran nuevos ejemplos y las reglas generadas superen el umbral de confianza. Este proceso iterativo permite la extracción de reglas diferentes en sucesivas ejecuciones del algoritmo.

La función de adaptación (*fitness*) se calcula de acuerdo a la ecuación 1, en la que se utilizan el soporte (Sop_c) como medida de generalidad, la confianza ($Conf$) como medida de precisión y el interés (Int) de la regla como medida de novedad. Las variables ω_1, ω_2 y ω_3 son los pesos asignados a cada una de las medidas.

$$fitness(c) = \frac{\omega_1 \times Sop_c(c) + \omega_2 \times Conf(c) + \omega_3 \times Int(c)}{\omega_1 + \omega_2 + \omega_3} \quad (1)$$

Tal y como se puede ver en [8], estas medidas se definen de la siguiente forma:

- **Soporte:** Es una medida del grado de cobertura que la regla ofrece a los ejemplos de la clase. Para promover la obtención de diferentes reglas en sucesivas ejecuciones del algoritmo, en su cálculo solo se consideran los ejemplos no cubiertos por las reglas previamente generadas.
- **Confianza:** La confianza de una regla es una medida estándar que determina la frecuencia relativa de los ejemplos que satisfacen tanto el antecedente

como el consecuente de una regla entre aquellos que satisfacen sólo el antecedente.

- **Interés:** Para el cálculo del interés, en la práctica es adecuado utilizar criterios objetivos como medidas que permitan seleccionar reglas potencialmente interesantes y criterios subjetivos para que el usuario final determine reglas realmente interesantes.

La lógica difusa mejora la interpretabilidad de las reglas extraídas debido a la utilización de una representación del conocimiento más cercana al experto, permitiendo además la utilización de variables numéricas sin necesidad de realizar una discretización previa. Los conjuntos difusos que definen las etiquetas lingüísticas se definen mediante las correspondientes funciones de pertenencia especificadas por el usuario o definidas mediante una partición uniforme si no está disponible el conocimiento del experto. Como se puede observar en la figura 1, se utilizan particiones uniformes con funciones de pertenencia triangulares (en este caso en concreto con 5 etiquetas).

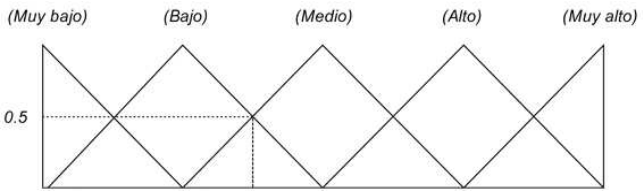

Figura 1: Conjunto difuso utilizado por SDIGA para 5 etiquetas

6 Experimentación realizada

La experimentación se ha llevado a cabo sobre la base de datos MSNBC. Considerando que en los algoritmos de descubrimiento de subgrupos hay que indicar cuál es el atributo objetivo, se necesitaría realizar un estudio para cada una de las categorías de la base de datos (como la variable objetivo tiene 17 valores posibles, habría que realizar una experimentación para cada uno de éstos, obteniendo una base de datos distinta para cada categoría). Sin embargo, en la literatura actual podemos encontrar distintos estudios realizados sobre MSNBC [19], que indican como categorías relevantes para la extracción de patrones frecuentes las categorías *frontpage*, *news* y *on-air*. Por esta razón, se ha realizado la experimentación sólo sobre estas tres categorías.

El objetivo principal que se pretende para esta primera aproximación de aplicación de descubrimiento de subgrupos en el área de la minería de uso web es obtener un conjunto de patrones de MSNBC para cada una de las categorías seleccionadas como variable objetivo y descubrir conocimiento sobre ellas.

Debido al gran tamaño de la base de datos (989818 sesiones descritas mediante 17 variables), es necesario realizar una reducción de instancias para reducir el tiempo de procesamiento y almacenamiento. En la literatura actual se pueden encontrar distintas técnicas de reducción de instancias aplicadas a la base de datos MSNBC [10][20][25]. En esta experimentación, hemos aplicado una selección de instancias aleatoria del 10% sobre la base de datos, obteniendo una nueva base de datos con 98982 usuarios.

Se ha realizado una *5-Validación Cruzada*, generándose cinco pares distintos de entrenamiento y prueba con un 20% de patrones para el entrenamiento y un 80% para el conjunto de prueba. Las particiones se han realizado en esta proporción ya que debido al gran tamaño de la base de datos, el porcentaje de entrenamiento es suficiente para que el modelo generado sea capaz de describir información a partir del conjunto de prueba.

Se ha desarrollado una experimentación sobre estas bases de ejemplos con los algoritmos clásicos de descubrimiento de subgrupos, Apriori-SD [22] y CN2-SD [28], pero no se obtienen resultados para poder realizar una comparación con SDIGA, ya que la necesidad de aplicar una discretización sobre la base de datos para el procesamiento de éstos, produce una gran pérdida de información que provoca la obtención de reglas vacías. Este problema no se presenta con SDIGA, puesto que no necesita discretización.

Dado el carácter no determinista de SDIGA, se han realizado tres ejecuciones independientes para cada uno de estos pares de conjuntos entrenamiento-prueba y los resultados mostrados en la tabla 3 son la media de los resultados. Como parámetros de la experimentación de SDIGA, se ha utilizado un umbral de confianza del 80%, un tamaño de la población del AG de 100, un máximo de 10000 evaluaciones de individuos para el AG, y una distribución de pesos de 0.4, 0.3 y 0.1 para el soporte, confianza e interés, respectivamente.

Para realizar la experimentación, se ha realizado la comparación de resultados considerando distinto número de etiquetas para los conjuntos difusos. De esta forma tenemos los algoritmos *SDIGA-3*, *SDIGA-5* y *SDIGA-7*, que utilizan 3,5 y 7 etiquetas respectivamente para la descripción de las variables continuas.

Los resultados se pueden observar en la tabla 3, donde *Objetivo* es la categoría utilizada como variable objetivo, *Alg* es el nombre del Algoritmo utilizado, *NReg* es el número de reglas obtenidas por el algoritmo,

Tabla 3: Resultados obtenidos de los experimentos realizados.

Objetivo	*Alg*	*NReg*	*Var*	$Sop_c(\%)$	$Conf(\%)$	*Int*	*Rel*
frontpage	*SDIGA-3*	**2**	**2**	92.90	**50.59**	**1.02**	**153.80**
	SDIGA-5	3	**2**	**95.09**	**50.59**	1.00	128.90
	SDIGA-7	**2**	**2**	94.01	50.29	1.00	135.23
news	*SDIGA-3*	**2**	**2**	99.13	64.52	**1.04**	**582.93**
	SDIGA-5	3.13	2.01	**99.17**	**64.74**	1.03	402.85
	SDIGA-7	**2**	**2**	98.70	64.17	1.03	326.48
on-air	*SDIGA-3*	**2**	**2**	**90.18**	50.17	**1.02**	120.21
	SDIGA-5	3	**2**	86.26	50.09	**1.02**	183.06
	SDIGA-7	**2**	**2**	87.75	**50.53**	1.01	**226.37**

Var es el número de variables que contiene cada regla, Sop_c, *Conf* e *Int* son el Soporte, la Confianza y el Interés, y *Rel* es la relevancia, medida muy utilizada en tareas de descubrimiento de subgrupos que se calcula en términos de su razón de verosimilitud (*likelihood ratio*) normalizada con la razón de verosimilitud del umbral de relevancia (99%).

Analizando los resultados obtenidos de la experimentación, se puede observar que el número de etiquetas con el que se obtiene mejores resultados depende de la categoría a estudiar. Así, para las categorías *frontpage* y *news* son mejores los conjuntos de 5 etiquetas, mientras que para la categoría *on-air* se obtienen mejores resultados con el conjunto de 7 etiquetas.

Se obtienen conjuntos con pocas reglas, con reglas sencillas interpretables, por el uso de reglas difusas lingüísticas, y en todos los casos con valores de soporte y confianza muy similares. Son reglas que abarcan casi la totalidad del conjunto de ejemplos pero con una confianza ligeramente baja. Además, en el caso de las clases o grupos on-air y frontpage son más difíciles de representar de forma precisa con el conjunto de instancias seleccionadas y el método de selección de instancias aleatorio utilizado.

7 Conclusiones y Trabajos futuros

Se ha presentado una primera aproximación a la aplicación de un algoritmo de descubrimiento de subgrupos con reglas difusas para un problema de minería de uso web. Hasta donde conocemos, no se han aplicado algoritmos de descubrimiento de subgrupos con lógica difusa a este problema. Las aproximaciones más cercanas son [17][21], que se centran en técnicas de agrupamiento difuso. Los resultados obtenidos muestran que es posible extraer conocimiento descriptivo en este tipo de problemas pero que es necesario hacer un análisis detallado del preprocesamiento, es decir de la selección de instancias.

En este trabajo se ha utilizado un método de selección de instancias aleatorio con una reducción muy importante (90% por la dimensionalidad de la base de datos original). Es conveniente estudiar el efecto de otros métodos de selección de instancias y la influencia del sesgo del mismo en el método posterior. La integración de un método de selección de instancias dentro del propio método de minería de datos, podría dar buenos resultados.

Como trabajos futuros se pretende realizar un completo estudio estadístico sobre técnicas de preprocesamiento para esta base de datos. El objetivo es por un lado, demostrar que es posible realizar una selección de instancias sobre MSNBC perdiendo la mínima información posible, y por otro, realizar un análisis de la influencia del método de preprocesamiento utilizado en el conocimiento extraído tras aplicar un método de minería de datos descriptiva posterior.

Además, se pretende obtener mediante descubrimiento de subgrupos las secuencias que existen en la base de datos, intentando ir más allá de la mera extracción de patrones frecuentes e indicando las secuencias que los usuarios realizan para llegar a una categoría objetivo.

Referencias

[1] D. Arotaritei and S. Mitra. Web mining: a survey in the fuzzy framework. *Fuzzy Sets and Systems*, 148(1):5–19, 2004.

[2] A. Asuncion and D.J. Nweman. UCI Machine Learning Repository. *University of California*, 2007.

[3] M. Atzmüller and F. Puppe. Sd-map - a fast algorithm for exhaustive subgroup discovery. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *PKDD*, volume 4213 of *Lecture Notes in Computer Science*, pages 6–17. Springer, 2006.

[4] M. Atzmüller, F. Puppe, and Hans-Peter Buscher. Towards Knowledge-Intensive Subgroup Discovery. In Andreas Abecker, Steffen Bickel, Ulf Brefeld, Isabel Drost, Nicola Henze, Olaf Herden, Mirjam Minor, Tobias Scheffer, Ljiljana Stojanovic, and Stephan Weibelzahl, editors, *LWA*, pages 111–117. Humbold-Universität Berlin, 2004.

[5] B. Berendet. Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*, 6(1):37–59, 2002.

[6] R. Cooley, B. Mobasher, and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. *On Tools with Artificial Intelligence*, pages 558–567, 1997.

[7] O. Cordón, M.J. del Jesus, F. Herrera, and M. Lozano. MOGUL: A Methodology to obtain genetic fuzzy rule-based systems under the iterative rule learning approach. *Internation Journal of Intelligent Systems*, 14:1123–1153, 1999.

[8] M.J. del Jesus, P. González, F. Herrera, and M. Mesonero. Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A case study in marketing. *IEEE Transactions on Fuzzy Systems*, 12(3):296–308, 2007.

[9] M. Delgado, N. Marín, D. Sánchez, and M.A. Vila. Fuzzy Asociation Rules: General Model and Applications. *IEEE Transactions on Fuzzy Systems*, 11:214–225, 2003.

[10] M. El-Sayed, E. Rudensteiner, and C. Ruiz. FS-Miner: Efficient and Incremental Mining of Frequent Sequence Patterns. 2004.

[11] F.M. Facca and P.L. Lanzi. Mining interesting knowledge from weblogs: a survey. In *Data & Knowledge Engineering*, volume 53, pages 225–241. Elsevier, 2005.

[12] Y. Fu, M. Creado, and C. Ju. Reorganizing web sites based on user access patterns. In ACM Press, editor, *Tenth International Conference on Information and Knowledge Managemen*, pages 583–585, 2001.

[13] D. Gamberger and N. Lavrac. Expert-Guided Subgroup Discovery: Methodology and Application. *J. Artif. Intell. Res. (JAIR)*, 17:501–527, 2002.

[14] D. Gamberger and N. Lavrac. Active subgroup mining: a case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, 28(1):27–57, 2003.

[15] G. Garriga, P. Díaz-López, and J.L. Balcázar. A lattice-based method for structural analysis. 2005.

[16] D.E. Golberg. Genetic Algorithms in search, optimization and machine learning. *Addison-Wesley*, 1989.

[17] B. Hay, G. Wets, and K. Vanhoof. Clustering navigation patterns on a website using a sequence alignment method.

[18] J.H. Holland. Adaptation in natural and artificial systems. *University of Michiga Press*, 1975.

[19] R. Iváncsy and I. Vajk. Different Aspects of Web Log Mining. 2005.

[20] R. Iváncsy and I. Vajk. Frequent Pattern Mining in Web Log Data. *Acta Polytechnica Hungarian*, 3(1), 2006.

[21] T. Kamdar. *Creating adaptive web servers using incremental web log mining*. PhD thesis, Computer Science Department, University of Maryland, Baltimore Country, 2001.

[22] B. Kavsek and N. Lavrac. APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20:543–583, 2006.

[23] W. Klösgen. Explora: A Multipattern and Multistrategy Discovery Assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. 1996.

[24] W. Klösgen and J. Zytkow. *Handbook of Data Mining and Knowledge Discovery*. Oxford, 2002.

[25] P. Kumar, P.R. Krishna, R.S. Bapi, and S. Kumar De. Rough Clustering of Sequential Data. *Data Knowl. Eng.*, 63(2):183–199, 2007.

[26] B. Lan, S. Bressan, B.C. Ooi, and K.-L. Tan. Rule-assisted prefetching in web-server caching. In ACM Press, editor, *Ninth International Conference on Information and Knowledge Management (CIKM 2000)*, pages 504–511, 2000.

[27] N. Lavrac, B. Cestnik, D. Gamberger, and P.A. Flach. Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned. *Machine Learning*, 57(1-2):115–143, 2004.

[28] N. Lavrac, B. Kavsek, P.A. Flach, and L. Todorovski. Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[29] W. Lin, S.A. Alvarez, and C.Ruiz. Efficient adaptive support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6(1):83–105, 2002.

[30] P. Lingras and R. Lingras. Adaptative hyperlinks using page access sequences and minimum spanning trees. 2006.

[31] Z. Markov and D.T. Larose. *Data Mining The Web. Uncovering patterns in Web Content, Structure and Usage*. Wiley, 2007.

[32] M.J. Martín-Bautista, D.H. Kraft, M.A. Vila Miranda, J. Chen, and J. Cruz. User profiles and fuzzy logic for web retrieval issues. *Soft Comput.*, 6(5):365–372, 2002.

[33] B. Mobasher. *Web Usage Mining and Personalization*. CRC Press LLC, 2005.

[34] B. Mobasher, H. Dai, and M. Tao. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6:61–82, 2002.

[35] V. Nikulin. Universal Clustering with Family of Power Loss Functions in Probabilistic Space. In Marcus Gallagher, James M. Hogan, and Frédéric Maire, editors, *IDEAL*, volume 3578 of *Lecture Notes in Computer Science*, pages 311–318. Springer, 2005.

[36] S.K. Pal and S. Mitra. *Neuro-Fuzzy Pattern Recognition: Methods in Soft-Computing*. Wiley, New York, 1999.

[37] G. Pallis, F. Angelis, and A. Vakali. Model-bases cluster analysis for web user sessions. 3488:219–227, 2005.

[38] M. Spiliopoulou. The laborious way from data mining to web mining. *International Journal of Computer System*, 14:113–126, 1999.

[39] M. Spiliopoulou and C. Pohle. *Data Mining for Measuring and Improving the Success of Web Sites*, volume 5, pages 85–114. Kluwer Academic Publishers, 2001.

[40] R. Srikant and Y. Yang. Mining web logs to improve website organization. In *WWW*, pages 430–437, 2001.

[41] S. Wrobel. An Algorithm for Multi-relational Discovery of Subgroups. In Henryk Jan Komorowski and Jan M. Zytkow, editors, *PKDD*, volume 1263 of *Lecture Notes in Computer Science*, pages 78–87. Springer, 1997.

[42] Q. Yang and H.H. Zhang. Web-log mining for predictive web caching. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):1050–1054, 2003.

[43] L.A. Zadeh. Information Control. *Fuzzy sets*, 8:338–353, 1965.

[44] L.A. Zadeh. The concept of a linguistic variable and its applications to approximate reasoning, Parts I, II, III. *Information Science*, 8-9:199–249, 301–357, 43–80, 1975.