Vol. III, No. 3, septiembre - diciembre, 2012, pp. 57 - 64

Minería de uso Web aplicada a registros de navegación por Internet

Darian Horacio Grass Boada
Correo electrónico:dgrass@uci.cu
Universidad de las Ciencias Informáticas, La Habana, Cuba

Alejandro Rosete Suárez
Correo electrónico:rosete@ceis.cujae.edu.cu
Instituto Superior Politécnico José Antonio Echeverría, Cujae, La Habana, Cuba

Jesús Eladio Sánchez García
Correo electrónico:grupoest@icmf.inf.cu

Valia Guerra Ones
Correo electrónico:vguerra@icimaf.cu
Instituto de Cibernética Matemática y Física, La Habana, Cuba

Artículo Original

Resumen
En el trabajo se describe un proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD en sus siglas en inglés) realizado en el entorno de los registros de navegación por Internet en la Universidad de las Ciencias Informáticas. En este contexto, se detalla un proceso de Minería de Uso Web utilizando como fuentes de datos los registros de navegación por Internet archivados en el servidor proxy, así como información descriptiva de los usuarios del servicio de navegación alojada en los sistemas de gestión del personal de la institución. Se combinan técnicas estadísticas, numéricas y de agrupamiento con vistas a identificar grupos similares de usuarios en el uso de las cuotas de navegación por Internet y de esta forma apoyar en la toma de decisiones a la Dirección de Redes y Seguridad Informática u otras direcciones de la universidad. Se describen los métodos y técnicas utilizadas, así como el procedimiento definido para llevar a cabo la tarea descriptiva de agrupamiento. En el mismo se propone un nuevo uso de la descomposición matricial CUR para identificar el número posible de grupos a identificar por el algoritmo de agrupamiento k-medoides. Por último, se muestran los experimentos realizados, la evaluación de los grupos obtenidos, además de ejemplos de algunos de los patrones obtenidos, y de esta forma apoyar actividades orientadas a la gestión y seguridad del servicio de navegación por Internet

Palabras clave: agrupamiento, registros de navegación por Internet, técnicas numéricas y estadísticas

Recibido: 14 de junio del 2012 Aprobado: 2 de agosto del 2012

INTRODUCCIÓN

El avance de las tecnologías de la información y las comunicaciones (TICs) permitió un incremento exponencial en el volumen de información almacenada en los sistemas de cómputo, con lo cual las organizaciones han podido satisfacer sus necesidades cotidianas, pero ha superado la capacidad de estas de analizar y transformar la información en conocimiento útil que ayude al mejor funcionamiento. En estas condiciones surge la *minería de datos* como alternativa para la obtención de patrones ocultos en un conjunto de datos.

La misma se define como "… el proceso de extracción no trivial a partir de las bases de datos de información desconocida y potencialmente útil". [1]

Uno de los dominios con alto crecimiento de información lo constituye la World Wide Web, en donde los datos brutos de la Web se han convertido en una vasta fuente de información. Por consiguiente, el uso de las técnicas de minería de datos se ha hecho necesario para descubrir patrones ocultos, [2] desarrollándose diferentes técnicas englobadas bajo la denominación de Minería Web. [3]

Dentro de esta clasificación se encuentra la Minería de Uso Web, la cual consiste en "… aplicar las técnicas de minería de datos para descubrir patrones de uso de los datos de la Web a fin de comprender y servir mejor a las necesidades de los usuarios que naveguen por ella". [4] De esta forma, se buscan patrones de acceso general para entender el comportamiento y las tendencias de los usuarios, con el fin de reestructurar contenidos de los sitios y publicarlos de forma más accesible o para dirigir a los usuarios a lugares concretos durante la navegación. También se pueden realizar búsquedas de uso tipificado donde se analizan las tendencias individuales de cada visitante para adaptar dinámicamente la información a partir de un perfil de usuario.

El primer paso en el proceso de Minería de Uso Web consiste en reunir los datos pertinentes. Existen dos fuentes principales de datos las que corresponden a los dos sistemas de software que interactúan durante una sesión Web: los datos en el servidor Web y los datos en el cliente. Además, cuando existen "intermediarios" en la comunicación cliente-servidor, también pueden convertirse en fuentes de datos de uso, por ejemplo: servidores proxy (ISP en sus siglas en inglés, Internet Service Provider). [5]

Las empresas comerciales, así como los investigadores académicos han desarrollado una amplia gama de herramientas que realizan varios algoritmos de minería de datos en los archivos de registro procedentes de servidores Web, con el fin de identificar el comportamiento del usuario en un sitio Web en particular. Sin embargo, un área que ha recibido mucha menos atención es la investigación de comportamiento del usuario en los servidores proxy. [6]

Estos son sistemas de software que suelen ser empleados por una empresa relacionada con Internet y actúan como intermediario entre un *host* interno y la Internet para que la empresa puede garantizar la seguridad, un control administrativo y servicios de almacenamiento en caché. Pueden ser una valiosa fuente de información para estudiar el comportamiento en el uso de las cuotas de navegación por Internet de un conjunto de usuarios de una organización.

La Universidad de las Ciencias Informáticas (UCI) cuenta con un servicio de navegación por Internet donde se genera un considerable volumen de información que registran los servidores proxy. A la hora de la toma de decisiones en la gestión y seguridad del servicio de navegación por Internet, la Dirección de Redes y Seguridad Informática (DRSI) no aprovecha el conocimiento implícito en los registros de navegación que describa el uso de las cuotas de navegación por Internet.

Del problema anterior se deriva el siguiente objetivo general: investigar, experimentar y aplicar diferentes técnicas descriptivas de agrupamiento de minería de datos para obtener patrones de navegación en términos de grupos similares en el uso de las cuotas de navegación por Internet que apoyen la toma de decisiones en la DRSI. De esta forma, se propone la combinación de técnicas numéricas, estadísticas y de algoritmos de agrupamiento. Se evalúan los grupos encontrados y se muestran algunos de los patrones obtenidos, y de esta forma apoyar las actividades orientadas a la gestión y seguridad del servicio de navegación por Internet.

MATERIALES Y MÉTODOS

A continuación se describe la tarea de minería de datos realizada, metodología y herramienta de análisis de datos usadas, así como las técnicas numéricas, estadísticas y algoritmo de agrupamiento utilizado. Se finaliza con una propuesta para realizar la tarea descriptiva: Agrupamiento.

Agrupamiento (clustering o segmentación)

La tarea de minería de datos realizada fue el *agrupamiento*, la cual consiste en agrupar un conjunto de objetos físicos o abstractos en clases de objetos similares. Un grupo es una colección de objetos que son similares entre sí y diferentes a los objetos de otros grupos. [7] Por tanto, el objetivo de esta tarea es obtener grupos entre los elementos de un conjunto de datos, de tal manera, que los elementos asignados al mismo grupo sean bien similares. [8]

Estos grupos de objetos similares pueden considerarse como una forma de compresión de los mismos, [7] lo que permite aplicar normas y técnicas más adecuadas debido a la reducción del tamaño del conjunto original. Las dos clases más representativas del agrupamiento son: el *agrupamiento particional* y el *agrupamiento jerárquico*. En el presente trabajo se realiza un agrupamiento particional, [9] el cual requiere como entrada el número de grupos a buscar.

Metodología y herramienta de análisis de datos

Se decidió emplear la metodología CRISP-DM [10] para el desarrollo de esta investigación, ya que tiene a su favor las fortalezas de concebir el proyecto de KDD de forma global y estrechamente relacionado con el negocio en cuestión, además de ser de distribución libre, y por tanto, estar en continuo perfeccionamiento.

Para la realización de las fases del proceso KDD: recopilación y preparación de los datos se llevó a cabo a través de una herramienta diseñada e implementada para tal propósito. En la fase de minería de datos, se utilizó como método para encontrar los grupos el algoritmo k-medoides [7] desarrollado en la herramienta de análisis de datos RapidMiner. [11] El mismo se adecua al tipo de variables que describen el problema abordado (variables cualitativas), además de proporcionar facilidades de experimentar con diferentes puntos de inicialización y funciones de similitud. Necesita como requisito el número de grupos, aspecto tratado mediante el uso de técnicas numéricas y estadísticas descritas en las próximas secciones.

Métodos numéricos: Descomposición matricial CUR como herramienta en el análisis exploratorio de datos

En esta sección se describe la descomposición matricial CUR y cómo esta puede ser de ayuda cuando se trata de

determinar el número de clases en las que se agrupa un conjunto grande de datos.

Dada una matriz real A, las descomposiciones matriciales aleatorias conocidas por CUR se basan en la determinación de tres matrices C, U y R tales que el producto CUR es una aproximación de la matriz A, C y R están formadas por algunas columnas y filas de A, respectivamente. Se conocen varias descomposiciones CUR que se diferencian en las cotas de error obtenidas y en el criterio para elegir las columnas y filas que forman las matrices C y R. [12-14] En particular, en este trabajo se utiliza la descomposición CUR propuesta en [15] que consiste en construir C y R a partir de la determinación de un factor de importancia para cada columna de la matriz de datos.

Las columnas y filas de la matriz son seleccionadas aleatoriamente según la distribución de probabilidad establecida por los factores de importancia, los que se interpretan naturalmente como sensores de la influencia de cada columna en la mejor aproximación de menor rango de la matriz de datos. El factor de importancia de la columna *j* denotado como Π_j se define como:

$$\pi_j = \frac{1}{k}\sum_{p=1}^{k}\left(v_j^{\,p}\right)^2 \qquad (1)$$

donde:

v_p^j: *j* -ésima componente del *p* -ésimo vector singular derecho.

k : número aleatorio de filas o columnas a utilizar.

En el artículo mencionado, [15] los autores proponen la utilización de la descomposición CUR para mejorar el análisis exploratorio de datos pues consiguen expresarlos en términos de un número pequeño de columnas y/o filas de la matriz de datos. Con respecto al método de las componentes principales (PCA en sus siglas en inglés) en el que los datos son expresados a partir de los mayores vectores singulares, este análisis facilita la interpretación pues los vectores singulares pierden significado en términos del problema del que provienen los datos.

En este trabajo se propone una aplicación diferente de la descomposición matricial CUR en el análisis de datos que está motivada por la necesidad de determinar un número adecuado de grupos para los datos que se tienen. La estrategia propuesta consiste en estudiar el comportamiento numérico de los factores de importancia determinados por la descomposición CUR calculados a partir de (1) para detectar grupos que presentan patrones similares en cuanto al índice determinado.

Aunque la estrategia no es de uso universal, en ocasiones (en particular en los datos provenientes del problema de este trabajo) el vector de probabilidad conformado por todos los Π_j arroja una información que permite el establecimiento del número de las clases o grupos buscados. En la sección de Resultados se muestran los factores de importancia calculados por la descomposición CUR aplicada a la matriz transpuesta de los datos.

Métodos estadísticos: Análisis factorial de las correspondencias múltiples como herramienta en el análisis exploratorio de datos y análisis canónico como método de evaluación

El análisis factorial de las correspondencias múltiples (AFCM) es una técnica factorial creada por Benzécri en 1973. [16] Constituye una generalización del análisis de las correspondencias simple (ACS), también del mismo autor. El ACS analiza una tabla de contingencia de dos entradas, mientras que el AFCM puede considerarse como el análisis de una tabla del mismo tipo, pero con múltiples entradas. En cierto sentido, se asocia al análisis de componentes principales (ACP) y muchos lo consideran como el ACP para datos cualitativos.

En el AFCM se logra una descomposición de la inercia total (una medida de la dispersión total análoga a la matriz de covarianzas del ACP) contenida en la nube de puntos y a través de esta se obtienen factores para las filas, así como para las columnas de la matriz de observaciones transformadas. Esta propiedad permite que sea posible representar en un mismo gráfico tanto los individuos como las variables.

En el presente trabajo se utiliza en forma exploratoria con vistas a conocer la posible existencia de agrupamientos entre los individuos, al mismo tiempo que saber cuáles son las variables que provocan estas uniones. De aquí se puede deducir un tamaño aproximado del número de grupos que debe darse como valor de entrada en el método *k*-medoides.

Como método de evaluación de los grupos se tomó el análisis canónico (AC), [17] conocido también como análisis factorial discriminante (AFD) en la literatura estadística francesa. El tratamiento algebraico coincide en líneas generales con la formulación del análisis de varianza multivariado (MANOVA) [17] y parte de la descomposición de la matriz de varianzas total en las componentes intra e inter, donde esta última es la matriz de varianzas de los puntos medios de los grupos supuestos. En este trabajo se utiliza como técnica comprobatoria de la división en grupos propuesta por el método k-medoides, así como en los grupos definidos a partir del factor de importancia de la descomposición matricial CUR.

Procedimiento para el agrupamiento

A partir del agrupamiento particional a encontrar y estudiar, se define un procedimiento para realizar el mismo, en donde se combinan en una primera fase exploratoria de los datos, el método numérico CUR y el método estadístico del AFCM para identificar posibles números de grupos. Posteriormente, los resultados identificados en esta fase se evalúan mediante el análisis canónico.

Si la evaluación de los grupos identificados no resulta consistente se regresa a definir nuevos posibles grupos; en caso contrario, se pasa a una próxima fase en busca de los

patrones mediante el algoritmo *k*-medoides para lo cual se utiliza como parámetro de número de grupos el obtenido en el paso previo. Se realiza diferentes ejecuciones del algoritmo *k*-medoides para lo cual se varían aleatoriamente los puntos iniciales y se experimenta con diferentes funciones de similitud: coeficiente Roger-Tanimoto, coeficiente Russel-Rao y el coeficiente de coincidencias simples. [9,18]

Todos estos experimentos son realizados mediante la utilización de la herramienta de análisis de datos RapidMiner. Por último, el mejor agrupamiento encontrado se evalúa con el análisis canónico el cual comprueba la validez de los grupos obtenidos.

RESULTADOS Y DISCUSIÓN

Fase de preparación de los datos

Los datos de interés para realizar el proceso de KDD se consultaron de diferentes fuentes. En el caso de los registros de navegación por Internet archivados en el servidor proxy (logs proxy), se tomaron los datos pertenecientes a un mes (noviembre 2010), cuyo volumen en bruto (sin ningún procesamiento llevado a cabo) se encuentra cercano a los 30 Gigabytes. Por otro lado, se consultó la información de interés para el estudio de los usuarios que realizaron navegación en el mes mencionado. Estos datos se encuentran registrados en los sistemas de gestión del personal de la institución.

Se definieron las variables del estudio mediante entrevistas a los especialistas de la DRSI. A partir de estas, para una mejor comprensión y análisis de la información recopilada, se realizó una exploración inicial de las variables mediante gráficas e índices estadísticos. De esta forma, se desarrollaron transformaciones: discretización, jerarquía de conceptos, creación de nuevos atributos, filtrado de los elementos más representativos, etcétera. La información recogida del servidor proxy se agrupó mediantes sesiones de navegación definidas para el problema en cuestión, con la finalidad de su mejor comprensión. Para más detalle de la preparación de los datos se debe consultar [19].

Luego del análisis exploratorio, las transformaciones realizadas a las variables, así como la creación de nuevas características se obtuvo el conjunto de datos a estudiar en la fase del modelado. Se tiene una vista minable o matriz de datos para cada tipo de usuario, en donde todas las variables son cualitativas nominales. Las tablas 1 - 3 muestran algunas de las variables del estudio que describen las vistas minables conformadas.

Tabla 1
Datos del usuario estudiante

Atributo	Descripción	Tipo	Valores
Índice académico	Índice académico	Nominal	3
Procedencia académica	Procedencia académica	Nominal	3
Sexo	Sexo	Nominal binaria	2

Tabla 2
Datos del usuario trabajador

Atributo	Descripción	Tipo	Valores
Cargo	Cargo ocupacional	Nominal	5
Área	Área de trabajo	Nominal	29
Región de procedencia	Región del país de donde procede	Nominal	3

Tabla 3
Datos de las sesiones de navegación

Atributo	Descripción	Tipo	Valores
Rango IP	Segmento de red de la petición	Nominal	13
Categoría URLs	Clasificación de la página web solicitada	Nominal	21
Horario	Horario de la petición	Nominal	6
Consumo	Tamaño (Kbyte) de los recursos solicitados	Nominal Binaria	2
Peticiones	Número de peticiones por sesión de navegación	Nominal Binaria	2
Día	Período del mes de la petición	Nominal	9

Las matrices de datos o visas minables analizadas, así como el número de instancias analizadas en cada una de estas son: Trabajadores externos (430 instancias), Trabajador recién graduado en adiestramiento (2 406), Trabajadores no docentes (1 533), Trabajadores docentes (1 5128), Estudiantes de 3er. año (1 122), Estudiantes de 4to. año (1668), Estudiantes de 5to. año (1 367).

Fase de minería de datos: Tarea descriptiva agrupamiento

Como se describió en la sección anterior, se definió un procedimiento que realiza un análisis exploratorio combinando el método numérico CUR y el método estadístico del AFCM para encontrar posibles números de grupos; esto último necesario en el método *k*-medoides. En el presente trabajo se utiliza este algoritmo por su sencillez y adecuación al problema abordado. El mismo constituye una modificación del algoritmo *k*-medias para trabajar con datos cualitativos. A continuación se describe un caso de estudio del procedimiento definido y posteriormente se muestran los resultados alcanzados en las restantes matrices de datos (vistas minables).

Caso de estudio: Trabajadores externos

La figura 1 muestra los factores de importancia dados por el método CUR a cada individuo de la matriz de datos, ordenados ascendentemente. A partir de estos valores, luego

de varias iteraciones se definen tres intervalos de importancia: factor importancia *Bajo* (0,0001-0,0006), factor de importancia *Medio* (0,0007-0,0008) y factor importancia *Alto* (el resto). A partir de estos intervalos construidos se realizó un AFCM (figuras 2,3 y 4) para comprobar la consistencia de los mismos.

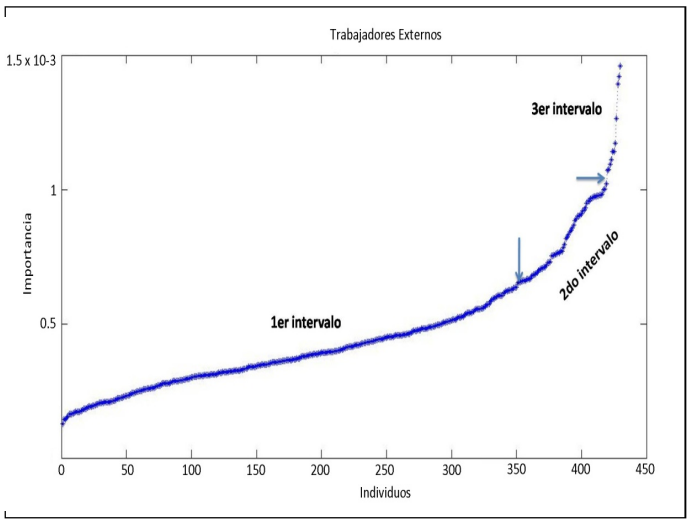

Fig. 1. Método CUR. Trabajadores externos.

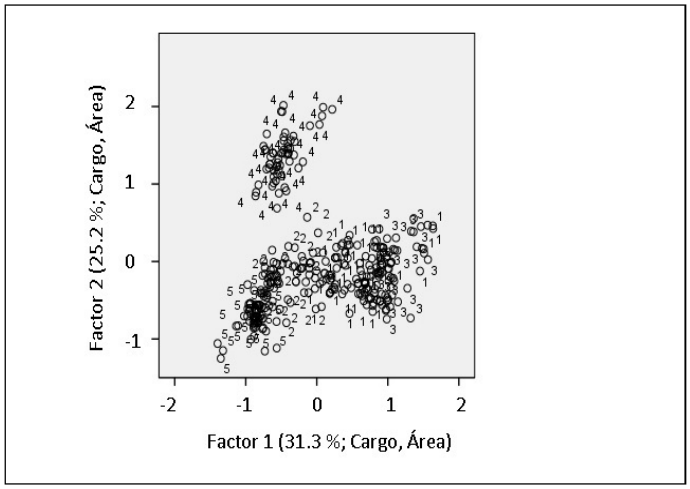

Fig. 2. Análisis factorial de las correspondencias múltiples. Trabajadores externos. Primer intervalo CUR.

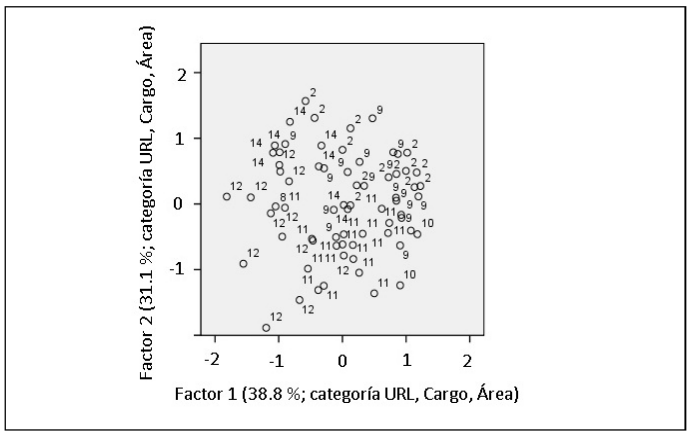

Fig. 3. Análisis factorial de las correspondencias múltiples. Trabajadores externos. Segundo intervalo CUR.

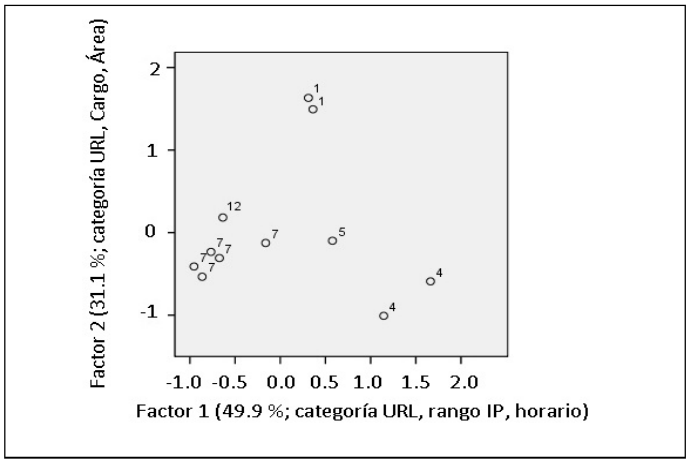

Fig 4. Análisis factorial de las correspondencias múltiples. Trabajadores externos. Tercer intervalo CUR.

Observar como el primero y segundo intervalos definidos a partir de CUR tienen una fuerte cohesión, disminuyendo en el tercer intervalo. Posteriormente se realizó el análisis canónico a partir de estos intervalos definidos obteniéndose como resultado la figura 5 y la tabla 4.

Resulta importante resaltar que los intervalos definidos si bien siguieron un criterio visual, se hicieron varias iteraciones definiendo los puntos de cruce, para luego evaluar las particiones obtenidas mediante el análisis canónico como se muestra en la tabla 5.

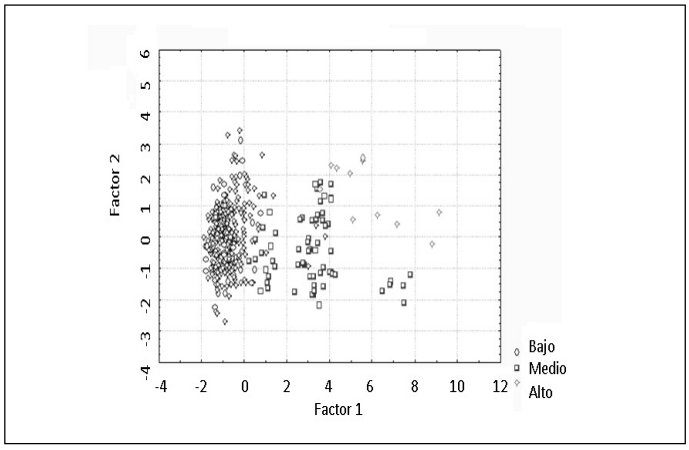

Fig. 5. Análisis canónico. Trabajadores externos. Agrupamiento según intervalos CUR.

Tabla 4
Análisis canónico.a los intervalos CUR

Intervalo	Correcto %	11(%)	I2 (%)	I3 (%)
11	99,1	347	3	0
I2	63,8	19	44	6
I3	90,9	0	1	10
Total	93.3	366	48	16

Tabla 5
Análisis canónico a los grupos k-medoides

Grupo	Correcto (%)	G1(%)	G2 (%)	G3 (%)
G1	80.8	210	13	37
G2	84.2	3	16	0
G3	81.5	25	3	123
Total	81.2	238	32	160

Una vez identificado (mediante el procedimiento iterativo descrito anteriormente) el número de grupos se pasó a realizar 33 experimentos en total mediante el método *k*-medoides. Para ello se utilizaron las tres funciones de similitud descritas en secciones anteriores, así como varios puntos iniciales como semillas. El número de grupos definidos fue de tres. Los grupos obtenidos en los diferentes experimentos se evaluaron mediante un operador de evaluación de la herramienta RapidMiner, por último, el mejor agrupamiento se evaluó mediante el análisis canónico. En la figura 6 se muestran los resultados del agrupamiento del mediante *k*-medoides.

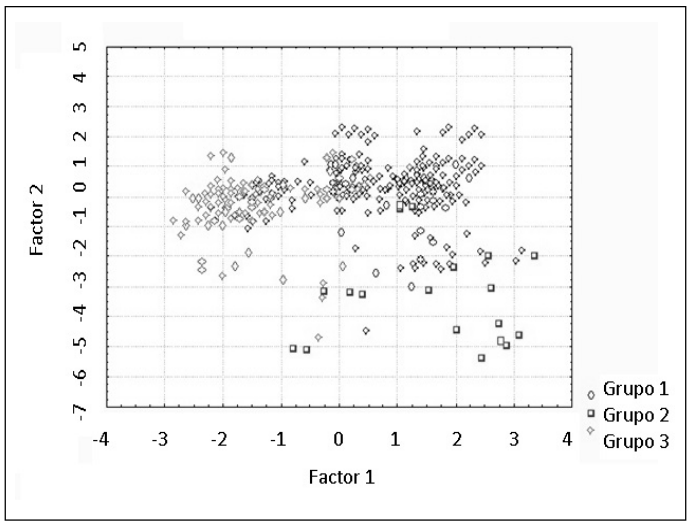

Fig. 6. Análisis canónico. Trabajadores externos. Agrupamiento según *k*-medoides.

Los grupos obtenidos se analizaron con los especialistas de la Dirección de Redes y Seguridad Informática (DRSI) para su interpretación y validación. Los patrones encontrados en la muestra de los trabajadores externos se muestran en la tabla 6.

El grupo 1 es el de mayor número de objetos representando a los que realizan navegación por sitios de noticias en la 3era. semana del mes. El de menor porcentaje (grupo 3) está caracterizado por visitas a sitios de seguridad informática. Se debe destacar que las URLs pertenecientes a este grupo conciernen en su mayoría a sitios de *hosting* o compartición de archivos, así como a servidores de anuncios (ads, de sus siglas en inglés).

Resultados generales

El procedimiento descrito anteriormente se realizó con las restantes matrices de datos (vistas minables) definidas. Las tablas 7 y 8 muestran los análisis canónicos realizados a los intervalos definidos a partir de CUR y al agrupamiento obtenido según corresponda.

Se puede observar que los intervalos definidos a partir de CUR tienen una buena consistencia, arrojando evaluaciones mediante el análisis canónico por encima del 85 %. Se debe resaltar que los intervalos 1 y 3 son los de mejores valores, teniéndose en el intervalo 2 los resultados más discretos, lo que muestra la posibilidad de definir en ocasiones 2 intervalos en lugar de 3. Asímismo, la evaluación de los agrupamientos obtenidos a partir del método *k*-medoides resultaron en su mayoría por encima del 75 %.

Tabla 7
Análisis canónico a los intervalos CUR

Matriz datos	11 (%)	12 (%)	13 (%)	Total (%)
Externos	99,1	63,8	90,9	93,3
Graduados adiestramiento	98,5	21,6	92,1	87,8
No docentes	96,9	46,9	84,9	87,6
Docentes	98,3	40,8	91,7	90,1
Estudiantes 3er. año	95,1	39,9	88,6	86,9
Estudiantes 4to. año	97,0	29,7	100	86,6
Estudiantes 5to. año	95,7	40,8	91,7	86,5

Tabla 6
Patrones obtenidos en vista minable trabajadores externos

Rango IP	Categoría URLs	Semana	Horario	Peticiones	Área	Categoría	%
Residencia	Noticias	3	Mañana	Medio	Producción	Tercerizados	73,7
Docente 2	Publicación	2	Tarde	Alto	Facultad-4	Adjunto	13,3
Residencia	Seguridad informática	2	Medio día	Medio	Producción	Tercerizados	13,0

Tabla 8
Análisis canónico grupos k-medoides

Matriz datos	G1 (%)	G2 (%)	G3 (%)	Total (%)
Externos	80,8	84,2	81,5	81,2
Graduados adiestramiento	90,0	68,0	63,4	80,2
No Docentes	87,0	65,8	64,8	77,8
Docentes	93,0	75,1	71,9	88,2
Estudiantes 3er. año	69,5	90,6	69,5	80,7
Estudiantes 4to. año	58,2	82,3	70,7	74,3
Estudiantes 5to. año	84,1	76,6	34,1	76,5

CONCLUSIONES

Se propone un procedimiento para realizar la tarea descriptiva de extracción de conocimiento: Agrupamiento en el contexto de la Minería de Uso Web. En el mismo se combinan la descomposición matricial CUR y el Análisis Factorial de las Correspondencias Múltiples (AFCM) como un análisis exploratorio inicial para identificar el número de grupos (parámetro necesario en el método *k*-medoides). La novedad en la utilización del método CUR estuvo dirigida a identificar un grado de *importancia* en los individuos a agrupar y no en las variables que los caracterizan; esto se debió al tipo de problema abordado: Gran número de individuos a agrupar; poca dimensión en los datos, lo cual permitiera identificar un número posible de grupos a encontrar. El método AFCM permitió la verificación de lo anterior. Posteriormente se utiliza el método *k*-medoides para encontrar los grupos, evaluados estos mediante el análisis canónico, alcanzandose resultados por encima del 75 %. Se cumplieron los objetivos propuestos, y se obtuvieron grupos (patrones) que describen el uso de las cuotas de navegación por Internet de diferentes usuarios de la universidad, con vistas a apoyar actividades de gestión y seguridad del servicio de navegación por Internet.

REFERENCIAS

1. **CHEN CHEN, M. S.; HAN, J.** *et al***.** "Data Mining: An Overview from a Database Perspective". *IEEE Transactions on Knowledge and Data Engineering*, 1996, vol. 8, núm. 6, pp. 866-883.
2. **SÁNCHEZ, G. G.; ÁVILA, S. D.** *et al***.** "Preprocesamiento de bases de datos masivas y multidimensionales en minería de uso web para modelar usuarios: comparación de herramientas y técnicas con un caso de estudio". En *III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA*. Madrid, España, 2005, pp. 193-202. ISBN 84-609-6891-X.
3. **PAGOLA, J. E. M.** *Estado del Arte del Web Mining*. Centro de Aplicaciones de Tecnologías de Avanzada. Serie: GRIS. Reporte: 001. La Habana, 2007.
4. **IVÁNCSY, R.; VAJK, I.** "Frequent Pattern Mining in Web log Data". *Acta Polytechnica Hungarica*, 2006, vol. 3, núm. 1, pp. 77-90.
5. **PIERRAKOS, D.; PALIOURAS, G.,** *et al***.** Web Usage Mining as a Tool for Personalization: A Survey. User Modeling and User-Adapted Interaction, 2003, vol. 13, núm. 4, p. 311-372. ISSN 0924-1868.
6. **KERKHOFS, J.; VANHOOF, K.** *et al***.** "Web Usage Mining on Proxy Servers: A Case Study". En *Workshop on Data Mining For Marketing Applications*. September, 2001.
7. **HAN, J.; KAMBER, M.** *Data Mining: Concepts and Techniques*. Editado por: Gray, J. San Francisco, CA: Morgan Kaufmann Publishers, 2006. vol. xxviii, 743 pp. Data Management Systems. ISBN 1-55860-901-6.
8. **ORALLO, J. H.; QUINTANA, J. R.** *et al***.** Introducción a la Minería de Datos. Madrid: Pearson Educación S.A., 2004. vol. xviii, 658 pp. ISBN 84-205-4091-0.
9. **XU, R.; II, D. C. W.** *Clustering*. Editado por: Fogel, D. B. New Jersey: John Wiley & Sons, Inc., 2009. vol. x, 341 pp. IEEE Press Series on Computational Intelligence. ISBN 978-0-470-27680-8.
10. **CHAPMAN, P.; CLINTON, J.** *et al***.** CRISP-DM 1.0 Step-by-Step Data Mining Guide SPSS Inc., 2000 [Consultado el 10\09 de 2010]. Disponible en: http://www.crisp-dm.org/download.htm.
11. RAPID-I COMPANY. RapidMiner. Report the Future [página web]. Rapid-I, 2010 [Consultado el 19\02\ de 2011]. Disponible en: http://rapid-i.com/content/view/181/190/.
12. **GOREINOV, S. A.; TYRTYSHNIKOV, E. E.** *et al***.** Theory of Pseudo-Skeleton Matrix Approximation. Linear Algebra and its Applications. 1997, vol. 261, núm. 1, pp. 1-21.
13. **STEWART, G. W.** The Decompositional Approach to Matrix Computation. *Computing in Science & Engineering*, 2000, vol. 2, núm. 1, pp. 50-59. ISSN 1521-9615.
14. **FRIEZE, A.; KANNAN, R.** *et al***.** *Fast Monte Carlo Algorithms for Finding Low-Rank Approximation*. J. Assoc. Comput. Mach, 2004, vol. 51, pp. 1025-1041.
15. **MAHONEY, M. W.; DRINEAS, P.** "CUR Matrix Decompositions for Improved Data Analysis". *Proc. Natli. Acad. Sci*. USA, 2009, vol. 106, pp. 697-702.
16. **BENZÉCRI, J. P.** *L'Analyse des Données: Analyse des Correspondences*. 22 ed. París, Francia: Dunod, 1973. vol. 2, 619 pp. ISBN 204007225X.
17. **MARDIA, K. V.; KENT, J. T.** *et al***.** *Multivariate Analysis*. Londres: Academic Press. 1979, Probability and Mathematical Statistics. ISBN 0-12-471252-5.
18. **YINA, Y.; YASUDAB, K.** "Similarity Coefficient Methods Applied to the Cell Formation Problem: a Comparative

inVestigation". *Computers & Industrial Engineering*. 2005, vol. 48, pp. 471-489.

19. **BOADA, D. H. G.; SUÁREZ, A. R.** "Minería de datos aplicada a los registros de navegación por Internet: Preparación de datos". En *I Conferencia Internacional de Ciencias Computacionales e Informática*. La Habana, 2011, pp. 6. ISBN 978-959-7213-01-7.

AUTORES

Darian Horacio Grass Boada
Licenciado en Ciencias de la Computación, Máster en Ciencias de la Computación, Profesor Asistente, Departamento de Programación e Ingeniería de Software, Universidad de las Ciencias Informáticas (UCI), La Habana, Cuba

Alejandro Rosete Suárez
Ingeniero en Sistemas Automatizados de Dirección, Doctor en Ciencias Técnicas, Profesor Titular, Facultad de Ingeniería Informática, Instituto Superior Politécnico José Antonio Echeverría, Cujae, La Habana, Cuba

Jesús Eladio Sánchez García
Licenciado en Matemática, Doctor en Ciencias Matemáticas, Investigador Titular, Departamento de Matemática, Instituto de Cibernética Matemática y Física, La Habana, Cuba

Valia Guerra Ones
Licenciada en Matemática, Doctora en Ciencias Matemáticas, Investigadora Auxiliar, Departamento de Matemática, Instituto de Cibernética Matemática y Física, La Habana, Cuba

Web Usage Mining Applied to Records of Navigation by Internet

Abstract

This paper presents a Knowledge Discovery on Databases (KDD) process applied on the internet surfing logs at the University of Informatics Sciences. In this context, it describes a Web-Usage Mining process using as data sources; the internet surfing logs stored by the proxy server, and also descriptive information regarding the users of such surfing service, which was provided by the institution's personnel management systems. Statistical, numerical and clustering techniques were combined seeking to identify user groups with similar internet surfing account usage, in hopes of providing important information for decision making processes carried out by the Network Management and Security Office or other areas of the institution. This paper describes the methods and techniques used, and the procedure utilized for performing the descriptive clustering task. This procedure proposes the use of the CUR matricial decomposition to identify the possible number of groups to identify by the k-medoides clustering algorithm. Lastly, the experiments carried out and the evaluations of the groups obtained are described and examples of some of the patterns obtained are presented.

Key words: clustering, internet browsing records, numerical and statistical techniques