Ediciones pensando el futuro
Editions thinking the future
Informática 2013
XV CONVENCIÓN Y FERIA INTERNACIONAL
15TH INTERNATIONAL CONVENTION AND FAIR
www.informaticahabana.com
www.informaticahabana.cu
ISBN: 978-959-7213-02-4

# SLD201 APLICACIÓN DE TÉCNICAS DE LA MINERÍA DE DATOS PARA USO EFICIENTE DE INTERNET EN EL CNGM

## SLD201 APLICATION OF TECHINICAL DATA MINING FOR EFFICIENT USE OF INTERNET IN CNGM

*Ing. Lissette Nuñez Maturel[1], Ing. Yaimara Alvarez Zaldivar [2], Ing. María de los A González Torres [3]*

1 Centro Nacional de Genética Médica, Cuba, lissette@cngen.sld.cu
2 Centro Nacional de Genética Médica, Cuba, yaimara@cngen.sld.cu
3 Centro Nacional de Genética Médica, Cuba, magonzalez@cngen.sld.cu

**RESUMEN:** *Una de las causas de que Internet sea un servicio limitado en las universidades, empresas y centros de investigaciones en Cuba son las restricciones impuestas por el bloqueo de los EUA* y el uso no adecuado por los investigadores, en muchos casos por desconocimiento de métodos de búsquedas más eficientes*. A partir de esta problemática el presente trabajo tiene como objetivo* trazar estrategias con vistas a obtener el máximo aprovechamiento de internet en un centro de investigación. *Se realizó un estudio observacional descriptivo tomándose una muestra de 71 usuarios de Internet y las 25 Urls más visitadas. Haciendo uso de la herramienta Weka se aplicaron* los algoritmos de conglomerados EM y SimpleKMean. Se obtuvo grupos de usuarios y similitudes entre ellos que apoyan el trabajo de búsqueda de información para descargarla y sugerirla a los usuarios.

***Palabras Clave:*** Internet, ancho de banda, algoritmos de conglomerados, EM, SimpleKMean.

**ABSTRACT:** *One reason that the Internet is a limited service in universities, companies and research centers in Cuba are the restrictions imposed by the U.S. blockade and the improper use by researchers, often for lack of search methods more efficient. Since this problem this paper aims to develop strategies in order to get the maximum use of the internet in a research center. We conducted a descriptive study of 71 taking a sample of Internet users and the 25 most visited Urls. Using the tool Weka applied EM clustering algorithms and SimpleKMean. Obtained user groups and similarities among them who support the work of searching for information to download and suggest it to users.*

***KeyWords:*** Internet, bandwidth, informatic security, Sawmill

## 1. INTRODUCCIÓN

La conexión de nuestro país a Internet se ha visto entorpecida entre otros factores por el bloqueo económico, comercial y financiero del gobierno de los EUA y por el uso no adecuado de los investigadores, en muchos casos por desconocimiento de métodos de búsquedas más eficientes. Tal es así que la autorización para la conexión se logró en el año 1996 a pesar de que los orígenes de la llamada red de redes se remontan a 1969, cuando se estableció la primera conexión de computadoras, conocida como ARPANET. Debido a las leyes que impone el embargo, Cuba no puede conectarse a la fibra óptica internacional que rodean sus costas lo que obliga a realizar la conexión vía satelital, opción que resulta más cara y limitada. [1].

A pesar de estas restricciones en los últimos años han aumentando considerablemente en el país el número de personas con acceso a Internet y los usuarios con correo electrónico gracias a las universidades, centro de investigaciones, escuelas y joven club de computación y electrónica. A través

de la red de la salud, INFOMED, acceden cerca de 30 mil profesionales, médicos e investigadores. [2]

Internet es una amplia red de ordenadores que nos puede proporcionar canales de comunicación, información y formación sobre cualquier tema, en cualquier momento y en cualquier lugar. Además permite que todos podamos producir y distribuir conocimientos, y nos proporciona un nuevo entorno de interrelación social, donde se pueden desarrollar todo tipo de actividades: entretenimiento, trabajo, comercio, arte, expresión de emociones y sentimientos. [3].Sin embargo el mal uso de este recurso puede provocar que el ancho de banda disponible sea insuficiente.

En ocasiones los usuarios no tienen el conocimiento y la destreza para utilizar este servicio. Para ello se implementan en diversas instituciones del país cursos y entrenamientos de alfabetización informacional, medidas de divulgación de las políticas de seguridad informática unidas a otras para mantener este servicio como son el uso de servidores proxy, identificadores de usuario y contraseñas, sistemas de filtrado de contenido y análisis detallado de las trazas de navegación de Internet (log).

Con el objetivo de optimizar el uso del canal de Internet teniendo en cuenta las limitaciones que tiene el país y de garantizar la calidad de los servicios que ofrece, Infomed implementó un sistema de cuotas asignando un cupo diario de tráfico a cada institución. El consumo de la cuota diaria depende de la cantidad de usuarios que estén laborando en el centro y utilizando el servicio.

En ocasiones la cuota asignada se acaba antes de que termine el horario laboral. De ahí la necesidad del análisis y procesamiento de los datos para trazar estrategias con vistas a obtener el máximo aprovechamiento de este servicio.

## 2. CONTENIDO

### 2.1 Materiales y métodos

Infomed es el proveedor de Internet del Centro Nacional de Genética Médica (CNGM) asignándole diariamente una cuota. En el centro existen 87 usuarios. Por lo que es importante usar los servicios telemáticos con fines de trabajo e investigación.

Se realizó un estudio observacional descriptivo tomándose una muestra de 71 usuarios de Internet activos y las 25 Urls más visitadas. De este modo se obtuvo un vector de Urls Vur(u) {Ur1, Ur2,…, Urn}, donde Ur es un número que identifica únicamente a una dirección Web dentro del log. Un ejemplo del vector para un usuario se observa a continuación:

Vur(Maria) {1,2,14, 20,23,28,30,32}

De este modo se tuvo un total de 87 vectores lo cual constituyó el juego de datos con el que se realizó el análisis.

Para determinar las 25 páginas mas visitadas se analizaron los logs de internet utilizando el software Sawmill 7.28.

Para agrupar a los usuarios en dependencia de sus preferencias investigativas en Internet (se tomó el vector de Urls), se empleó el análisis de conglomerados. Con el uso de la herramienta Weka se aplicaron varios algoritmos de conglomerados: EM y SimpleKMean.

El método Expectation Maximization (EM), es un método de agrupación en clústeres blando. Esto significa que un punto de datos siempre pertenece a varios clústeres, y que se calcula una probabilidad para cada combinación de punto de datos y clúster. (4)

SimpleKMeans: Es un algoritmo clasificado como Método de Particionado y Recolocación. Este método representa cada uno de los clusters por la media (o media ponderada) de sus puntos, es decir, por su controide. (5)

EM está basado en métodos probabilísticos y mediante validación cruzada permite determinar la cantidad apropiada de clústeres que se pueden formar mientras que SimpleKMean requiere por su parte de que se le indique el número de clústeres a formar.

Primeramente, fue necesario transformar los datos que se tenían a un formato en el que Weka trabaje: Formato de Archivo Atributo-Relación (ARFF, por sus siglas en inglés). Para un buen desempeño de los algoritmos de conglomerados es preciso trabajar con valores numéricos, por lo que la mejor manera de representar la información fue mediante una matriz donde las filas constituyeron los usuarios y las columnas cada Url. Los valores de las celdas fueron 0 o 1 indicando si el usuario visitó o no la Url.

De esta manera se obtuvo un archivo ARFF con los datos de los usuarios. Un fragmento de este archivo se muestra a continuación:

```
@relation URL
@attribute Usuario numeric
@attribute URL1 numeric
@attribute URL2 numeric
@attribute URL3 numeric
(…)
@data
```

```
1,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0
2,1,1,1,0,1,0,1,1,0,0,0,0,0,1,1,0,0,0,
0,0,0,0,0,0,0,0
3,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,
0,0,0,0,0,0,0,0
(…)
```

## 2.2 Resultados y discusión

Usando el algoritmo EM se asignó a cada instancia (usuario) una probabilidad de pertenencia a cada clúster y mediante validación cruzada decidió cuantos grupos crear. Se formaron 3 grupos con 26, 38 y 7 usuarios. (fig.1)

```
Clustered Instances

0       36 ( 51%)
1       28 ( 39%)
2        7 ( 10%)
```

**Fig 1 Usuarios asignados a cada clúster según EM**

Los grupos formados indican 3 principales tendencias de navegación en internet que tienen lugar en el centro. En un primer grupo aparecen 36 usuarios en su mayoría investigadores que trabajan temas relacionados con la genética clínica y la biología molecular; en un segundo grupo aparecen 28 usuarios todos investigadores que trabajan fundamentalmente temas relacionados con epidemiología, la Genética Bioquímica y el Estrés Oxidativo; y por último se identifica un tercer grupo de 7 usuarios que investigan en diversos temas como las tecnologías de la información, el acontecer internacional y temas generales de la genética y la inmunología.

Resulta útil para el grupo de redes y servicios técnicos informáticos del centro y para los especialistas en bibliotecología y ciencias de la información los resultados que arroja este algoritmo de conglomerados. De esta manera se pueden detectar afinidades entre los usuarios y posibilidades de recomendar los materiales que se descarguen o elaboren en la institución.

El algoritmo Simple K-Mean es (uno de los algoritmos más empleados en aplicaciones científicas e industriales que se basa en establecer puntos llamados centroides que representan la media de un conjunto de datos. Los centroides son empleados para formar los grupos, determinando los puntos con menor distancia a un centroide.

En este algoritmo se debe especificar la cantidad de grupos a formar y tomando en cuenta el resultado obtenido al aplicar el algoritmo EM indicamos 3. Cuando se ejecutó el algoritmo se obtuvieron resultados menos concluyentes que los obtenidos por el EM, y teniendo en cuenta que el tercer grupo formado por el algoritmo EM incluía usuarios con un comportamiento variable en internet, se decide aumentar el número de grupos a formar a 5 y aumentar el número de semillas a 19, para mejorar la efectividad del agrupamiento.

Se obtuvo un agrupamiento diferente de usuarios manteniendo en un mismo clúster a aquellos que tienen necesidades de búsquedas más afines y se formaron otros grupos con nuevas relaciones. (Fig.2)

Se obtuvo un grupo que consulta información relacionada con mutaciones del ADN, un grupo que aborda la epidemiología, un grupo las tecnologías de la información, un grupo la genética clínica y otro la genética bioquímica y el estrés oxidativo. Muchos usuarios coinciden en los mismos grupos que resultaron del algoritmo anterior, pero con este surgen afinidades que no se vieron anteriormente.

```
Clustered Instances

0       15 ( 21%)
1       21 ( 30%)
2        7 ( 10%)
3       16 ( 23%)
4       12 ( 17%)
```

**Fig. 2: Usuarios asignados a cada clúster según SimpleKMean**

De esta manera se pueden emplear ambos algoritmos para obtener grupos de usuarios y similitudes entre ellos que apoyen el trabajo de búsqueda de información. Ambos resultados son útiles para descargar información actualizada afín a cada grupo identificado, de modo que se les pueden sugerir a los usuarios artículos, eventos, etc.

## 3. CONCLUSIONES

El estudio realizado permitió identificar las principales fuentes de información que emplean los usuarios del centro. Esta información es de gran valor ya que permite mejorar la calidad del servicio de Internet y facilitar su uso racional.

Las técnicas de minería de datos empleadas para analizar los datos de los usuarios resultaron de gran utilidad para tener una visión acertada de las preferencias de información de los usuarios. Su aplicación apoyará la optimización del uso del ancho de

banda, al descargar y publicar documentación de interés en un lugar visible, lo que reduce el tiempo de búsqueda.

## 4. REFERENCIAS BIBLIOGRÁFICAS

1. A. E. d. Valle, "Estados Unidos bloquea Internet en Cuba (I)," Juventud Rebelde, 2 de Noviembre 2006.

2. M. d. R. e. d. Cuba, "Informatización en Cuba."

3. "IMPORTANCIA DEL BUEN USO DE INTERNET." vol. 2012, 2011.

4. msdn.microsof.com. [En línea] 2012. [Citado el: 28 de Noviembre de 2012.] http://msdn.microsoft.com/es-es/library/cc280445.aspx.

5. **S, Santiago Fernando Suárez.** slideshare. [En línea] 2010. [Citado el: 26 de Noviembre de 2012.] http://www.slideshare.net/nando85/clasificacin-automtica-de-documentos.

## 5. SÍNTESIS CURRICULARES DE LOS AUTORES

**Lissette Nuñez Maturel**, nace el 1ro de enero de 1986 en La Habana. Graduado del nivel superior en la Universidad de Ciencias Informáticas como Ingeniero en Ciencias Informáticas, en el año 2008. Se desempeña como responsable de seguridad informática en el Centro Nacional de Genética Médica, ubicado en Ave.31 esq.146 #3102.Cubanacán, Playa, La Habana. Profesor Instructor y aspirante a investigador. Publicó un artículo en la Revista Cubana de Informática Médica (RCIM) y participó los eventos UCIENCIA 2012 y en la I Jornada Virtual de Informática y Software Libre (InfoSoft2012).