USO DE LA METODOLOGÍA CRISP-DM PARA ANALIZAR EL SISTEMA DE NAVEGACIÓN DE LA UNIVERSIDAD DE MATANZAS.

Ing. Rita Milena Hernández Díaz[1], Ing. José Enrique Díaz Ramos[1]

1. Universidad de Matanzas, rita.hernandez@umcc.cu

Resumen

Con la amplia disponibilidad de datos producidos en el mundo de la informática, las técnicas tradicionales estadísticas para su procesamiento no aprovechan la información de valor cognitivo implícita en ellos. Por tanto, se necesitan implementar nuevas técnicas y herramientas que permitan solucionar esta problemática, siendo la Minería de Datos una alternativa para la obtención de patrones ocultos en un conjunto de datos.

Se presenta una propuesta de metodología CRISP-DM para analizar el sistema de navegación de la Universidad de Matanzas; con el objetivo de obtener patrones, tendencias y relaciones dentro de los datos, que describan el uso por los usuarios a las cuotas de navegación por Internet. Se desarrolla un proceso de descubrimiento de conocimiento, aplicando el algoritmo seleccionado en un caso de estudio; a partir de una muestra de datos generados por el servidor proxy. Con los resultados obtenidos se demuestra la utilidad de la presente investigación.

***Palabras claves**: Minería de datos; metodología; CRISP-DM; Internet.*

Introducción

Monografías 2020

ISBN: 978-959-16-4472-5

En la Universidad de Matanzas, Internet constituye una de las principales fuentes de consulta de conocimiento para la docencia, la investigación y la producción. Como mecanismo de control de acceso se utiliza un servidor proxy, el cual se emplea como intermediario en la comunicación entre un *host* interno e Internet. Para la navegación los usuarios cuentan con un sistema de cuotas, implementado mediante un análisis que se desarrolla en la Dirección de Redes y Seguridad Informática de la institución a partir de los datos almacenados en los registros generados por el servidor proxy.

Para la aplicación de las reglas implementadas en el sistema de cuotas, la Dirección de Redes analiza un elevado volumen de información de forma automatizada utilizando diferentes herramientas que permiten la generación de reportes estadísticos. A partir de estos reportes, los administradores infieren determinados comportamientos de los usuarios lo que permite definir las políticas de navegación. Estos reportes no tienen en cuenta diferentes variables y observaciones que posibiliten asociaciones con un elevado nivel de exactitud. Por otra parte, no se realiza análisis de datos de forma inteligente que describa el comportamiento de los usuarios, constituyendo esto una limitación en términos de competitividad y comercialización.

Una forma muy valiosa de análisis de la información es la Minería de Datos (MD). La MD es una de las etapas de lo que se conoce como el Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases: KDD). Este proceso consta de varias fases e incorpora diferentes técnicas de Aprendizaje Automático, la Estadística, las Bases de Datos, los Sistemas de Toma de Decisiones, las técnicas de Visualización y otras áreas de la informática y de la gestión de información. La MD puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. (Arias, 2011)

Como se ha descrito, en la Dirección de Redes de la Universidad de Matanzas no se realiza análisis inteligente, por lo cual se hace necesaria una herramienta que a partir de los registros de navegación, permita obtener patrones, tendencias y relaciones dentro del conjunto de datos para conocer mejor el uso que le dan los usuarios a las cuotas de navegación, aportando al sistema de toma de decisiones de la universidad una visión más amplia.

Por todo lo anterior, se hace necesaria la búsqueda de nuevos modelos, algoritmos y técnicas computacionales que permitan, a través de procesos inteligentes, obtener resultados que sean de interés para los especialistas y directivos del centro y que brinde nuevos conocimientos sobre la información generada.

Desarrollo

Se define como Minería de Datos al proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. En esencia, la tarea fundamental de la Minería de Datos es encontrar modelos inteligentes. Por sus características es imprescindible que sea realizado como un proceso automático o semiautomático. (Rojas, 2015)

Existen diversas metodologías de desarrollo para proyectos de MD tales como: CRISP-DM (*CRoss Industry Standard Process for Data Mining*), SEMMA (*Sample, Explore, Modify, Model, Asses*), Metodología de las cinco A"s (*Asses, Access, Analyze, Act, Automate*), Modelo de proceso de Minería de Datos de *Two Crows*, CRITIKAL (*Client-Server Rule Induction Technology for Industrial Knowledge Acquisition from Large Databases*) y Metodología SQL Server- 2005.

La metodología CRISP-DM es suficientemente amplia y flexible, y la más usada gracias a su fácil adaptación a proyectos de MD.

Las ventajas encontradas en esta metodología son:

- El proyecto de MD es visto de forma global y estrechamente relacionado al negocio en cuestión.
- Fue diseñada de forma neutra a la herramienta que se utilice para el desarrollo del proyecto, brindando la facilidad de uso con cualquiera de ellas.
- Es una metodología de distribución libre.
- Muchas de las metodologías que se pueden encontrar en la actualidad se basan en este estándar.
- Es la que cuenta con mayor aceptación por parte de los desarrolladores de procesos de extracción de conocimientos a partir de datos.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de MD en seis fases, cuya sucesión no es rígida, y se puede mover entre ellas siempre que se requiera: (Vercellis, 2009)

- Análisis del problema: fase inicial enfocada a entender los objetivos y requerimientos desde una perspectiva de negocio; para luego definirlos en términos de un problema de Minería de Datos y diseñar un plan para satisfacerlos.

- Comprensión de los datos: se hace una recolección y exploración inicial de los datos para familiarizarse con ellos e identificar problemas de calidad. Además, se trata de descubrir o estimar las relaciones más evidentes para formular las primeras hipótesis sobre información oculta en ellos.

- Preparación de los datos: esta fase cubre todas las actividades necesarias para construir la colección de datos que finalmente será minada a partir del grupo inicial. Incluye la colección, exploración, limpieza, transformación y construcción de datos.

- Modelado: durante esta fase se aplican varias técnicas de modelado. Comúnmente existen varias técnicas para resolver un problema de Minería de Datos del mismo tipo. Incluye la evaluación desde el punto de vista de precisión de los modelos.

- Evaluación: al llegar a esta fase se tendrán los modelos de mayor calidad desde la perspectiva de la precisión. Se impone una evaluación de los modelos y de los pasos que se siguieron para su construcción, a fin de determinar si responden apropiadamente a los objetivos de negocio que se determinaron en la primera fase. Es de vital importancia analizar si alguna regla del negocio no fue tomada en cuenta con el suficiente peso.

- Despliegue: en dependencia de los requerimientos y objetivos, la fase de despliegue puede ser tan simple como generar un reporte, o tan compleja como emprender un proceso de KDD de mayor envergadura. En ocasiones, son los clientes y no los desarrolladores quienes implementan esta fase; deben comprender cómo desarrollarla. Se hace imprescindible documentar y presentar los resultados de manera que todos los puedan entender.

Las fases del proyecto de minería de acuerdo a lo establecido por la metodología CRISP-DM interactúan entre ellas de forma iterativa durante el desarrollo del proyecto, formando una secuencia cíclica.

Resultados de la investigción a través de las fases del Proyecto de minería utilizando la metodología CRISP-DM:

- Análisis del problema: para los administradores de redes resulta importante conocer el comportamiento sobre el uso de las cuotas de acceso a Internet y en función del conocimiento obtenido, implementar políticas de navegación que aporten a la producción de la universidad. En la estructura propuesta para la base de datos se encuentran recogidos datos como usuario, facultad, área desde donde se realizó la conexión, clasificación de la página accedida, entre otros. Sobre estos datos no se ha hecho ningún estudio que permita conocer información interesante sobre el uso de la cuota de acceso a Internet.

- Comprensión de los datos: Los datos de interés para realizar el proceso de minería de datos se tomaron a partir de los registros de navegación por Internet archivados en el servidor proxy y otros datos propuestos. Los administradores de red de la universidad tienen asignados segmentos de IP para cada facultad y área desde donde se realiza la conexión, esta característica se utilizó para obtener estas variables a partir del campo IP. Otro dato que se obtiene es la clasificación o

categoría de la página (ocio, irrelevante, de interés y nacional) y el dominio al que pertenece.

A partir del estudio realizado a una muestra de 30 usuarios se llegó a la conclusion que:

- El 30% de los accesos a internet se realizan a sitios de ocio.
- El 30% de los accesos a internet se realizan a sitios nacionales.
- El 16% de los accesos a internet se realizan a sitios irrelevantes.
- El 24% de los accesos a internet se realizan a sitios de interés.
- El 50% de los accesos a internet se realizan en el horario de la mañana.

Los dominios utilizados presentan el siguiente comportamiento:

	Dominio	**Por ciento**
1	youtube.com	14 %
2	facebook.com	9 %
3	twitter.com	4 %
4	taringa.net	3 %
5	wikipedia.com	12 %
6	patriagrande.com.ve	9 %
7	juventudrebelde.cu	3 %
8	ecured.cu	9 %
9	cubadebate.cu	3 %
10	ain.cu	7 %
11	prensalatina.cu	11 %
12	proverbia.net	7 %
13	mascotas.com	9 %

- Preparación de los datos: En el análisis de minería de datos es conveniente realizar transformaciones sobre el conjunto de datos con el fin de mejorar la precisión de los modelos de aprendizaje que se desarrollarán posteriormente. Con la realización de esta tarea se analizan los datos y se combinan en el caso de que se encuentren en fuentes diversas. Luego se prosigue a la integración de la información que se

extrae de tablas diferentes y se crea una o más tablas que contienen información útil sobre los mismos objetos. Además puede que se generen nuevos registros o columnas que generalicen la información de múltiples tablas.

- Modelado: Sobre un mismo conjunto se pueden emplear diversas técnicas, pero se debe tener en cuenta que algunas plantean requerimientos específicos sobre la forma de los datos. Por lo tanto frecuentemente es necesario regresar a la fase de preparación de datos. Para realizar la fase de modelado se realizan las pruebas con el algoritmo Apriori. Con este se hacen las pruebas correspondientes sobre el conjunto de datos definido. Los algoritmos de asociación permiten la búsqueda automática de reglas que relacionan conjuntos de atributos entre sí.

Resultados obtenidos para conocer la cantidad de usuarios de la muestra que tienden a conectarse a sitios clasificados en diferentes horarios:

Horario	**Usuarios**	**Categoría**
mañana	1	nacional
mañana	1	interés
mañana	1	ocio
almuerzo	2	nacional
almuerzo	2	ocio
tarde	1	interés
tarde	1	ocio, nacional e interés
comida	2	ocio
comida	4	nacional
comida	1	irrelevante
comida	2	interés

noche	2	ocio
noche	3	irrelevante
noche	1	nacional, interés
madrugada	4	ocio
madrugada	1	interés
madrugada	3	nacional
madrugada	1	irrelevante

Resultados obtenidos para conocer del departamento de Informática, los tipos de sitios a los que más se acceden según la muestra.

Departamento	**Categoría**
Informática	ocio
Informática	nacional

- Evaluación y despliegue: Para la evaluación del modelo escogido se tiene en cuenta el cumplimiento de los objetivos propuestos. En este paso se revisa el proceso para determinar si debe repetirse alguna fase en caso de existir algún error. Si el modelo generado es válido se procede al despliegue del proyecto. Se determina que las Reglas de Asociación con el uso del algoritmo Apriori es la técnica con la que se obtiene los mejores resultados a partir de cada funcionalidad propuesta.
  Reglas obtenidas a partir de los datos seleccionados para la minería:

  - ✓ El dominio facebook.com es más visitado en el horario de la madrugada.
  - ✓ Los usuarios del departamento de Informática tienden a conectarse más a los sitios nacionales y de ocio.

Las reglas obtenidas responden a los objetivos deseados, por lo que se concluye que fueron cumplidos los propósitos trazados correspondientes al descubrimiento de patrones ocultos en los datos; que permiten describir las relaciones existentes en los atributos de los datos analizados.

En el proyecto no se propone repetir ningún paso, ya que después del análisis no se han encontrado fallas, ni se ha omitido ninguna variable que pudiera limitar el éxito de los resultados.
Finalmente se determina que el proyecto puede ser desplegado, se toman los resultados de la evaluación y se concluye una estrategia para el despliegue. Las acciones propuestas son: resumen de los resultados obtenidos, decidir para cada resultado el conocimiento o la información proporcionada a sus usuarios y cómo estos podrán ser utilizados dentro de los sistemas de la organización.

Conclusiones

La realización de este trabajo permitió identificar la metodología CRISP-DM como una guía valiosa para el desarrollo de proyectos de Minería de datos. Con la utilización de las reglas de asociación se obtinen relaciones dentro de los datos que describen el comportamiento de los usuarios en el uso de las cuotas de navegación.

Cuando se llevaron a cabo los diferentes experimentos sobre los datos, el Apriori se concluye que resulta eficiente para la obtención del conocimiento implícito a partir de los registros y se obtienen los resultados adecuados en correspondencia con las funcionalidades propuestas.

ISBN:

Referencias bibliográficas

2005. Vol. VIII, No. 26 . *Licencia de software libre orientada a proteger la libre distribución, modificación y uso del software. Ingenierías.* 2005.
Arias, Rigoberto Mora, Pino, Omar Vidal y Pérez, Lisandra Guerrero. 2011 . *2. Arias, Rigoberto Mora, Pino, Omar VidAplicación de técnicas de minería de datos con Weka Acknowledge Explorer. .* 2011 .
J, Marín R. y Palma. 2008. *Inteligencia artificial, técnicas, métodos y aplicaciones.* Madrid : s.n., 2008.
Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Ian H. Witten y Peter Reutemann. 1. Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Ian H. Witten y Peter Reutemann. The WEK2009. *The WEKA Data Mining Software: An Update.* 1. Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Ian H. Witten y Peter Reutemann. The WEK2009.
Pierrakos, Dimitrios. 2003. *Web Usage Mining as a Tool for Personalization: A Survey. .* 2003.
Rojas, Roberth Paúl Bravo Castro y María Esther Ruilova. 2015. *Árboles de clasificación (inteligencia artificial avanzada).* Ecuador : s.n., 2015.
Vercellis, C. 2009. *C. Business Intelligence: Data Mining and Optimization for Decision Making. .* 2009.