Retos de la Minería de Datos en las Ciencias e Ingeniería

Patricia Galván Morales
Instituto Tecnológico de Celaya
patricia.galvan@itcelaya.edu.mx

Franco Fabio García González
Instituto Tecnológico de Celaya
franco.garcia@itcelaya.edu.mx

José Emigdio Godoy Zárate
Instituto Tecnológico de Celaya
emigdio.galvan@itcelaya.edu.mx

José Benigno Molina Castro
Instituto Tecnológico de Celaya
Benigno.molina@itcelaya.edu.mx

Viridiana Omaña Silva
Instituto Tecnológico de Celaya
viridiana.omana@itcelaya.edu.mx

Resumen

El presente ensayo propone un conjunto de ideas que ofrezcan al lector una visión global de la minería de datos que sirva como sugerencias de aplicación para las ciencias y la ingeniería.

Palabra(s) Clave(s): minería de datos.

1. Introducción

Desde la popularización de la computadora personal en la década de los ochenta del siglo XX, el crecimiento exponencial de la información por año ha sido una constante. Las áreas de la ingeniería y, específicamente, las ciencias computacionales mediante la inteligencia artificial se dieron a la tarea de investigar y crear soluciones para el aprovechamiento de los datos, transformando muchas disciplinas desde la "pobreza de datos a la riqueza de los datos y el llamado para el desarrollo de métodos de uso intensivo de datos que conlleven a la investigación de las ciencias y la ingeniería" (Han & Gao, 2009).

El interés no debe ser puesto solamente en la idea de manejar los datos y descubrir la utilidad de su información, sino también en el tratamiento en línea, que sean accesibles en las redes de datos y desarrollar herramientas de minería de datos que sean aprovechadas para analizar dichos datos.

La minería de datos surgió como el conjunto de herramientas que permite descubrir en la información esos elementos o patrones que apoyen la toma de decisión de los negocios, las ciencias, la tecnología y la sociedad.

En muchos casos, la metodología tradicional para transformar los datos en conocimiento consiste en hacer un análisis e interpretación de dichos datos, de ahí que, se analizarán los datos que reflejen las tendencias de los mismos. Cabe señalar que cuando la cantidad de datos es abundante, se sobrepasa la capacidad humana para entenderlos sin la ayuda de herramientas computacionales potentes.

La minería de datos es el proceso automatizado del descubrimiento de información útil en grandes repositorios de datos (Tan, Steinbach, & Kumar, 2006). No toda la información extraída de los repositorios de datos se debe considerar minería de datos, existen consultas de registros individuales o conjuntos bien identificados que pueden resolverse mediante álgebra relacional y otras herramientas de lenguaje estructurado de consultas.

Los avances de la computación y las comunicaciones en las redes de datos han creado múltiples sistemas distribuidos y omnipresentes, algunos ejemplos son los usos del Internet, intranets, redes de área local y redes inalámbricas (Kargupta & Sivakumar,

2004); el almacenamiento de esa información hace un gran repositorio de acceso multicultural y universal.

La minería de datos es pues, un conjunto de herramientas basadas en la inteligencia artificial y la estadística que nos permite descubrir comportamientos de información que no son fáciles de determinar mediantes lenguaje estructurado de consultas o algebra relacional, se aplica tanto para repositorios estáticos de datos propiedad de una empresa o institución o, pueden ser repositorios públicos que se encuentren en la red (por sí mismo en un repositorio o distribuidos en varios repositorios).

1.1. Tipos de minería de datos

Tamilselvi y Kalaiselvi (2013) exponen dos tipos principales de minería de datos: predictiva y descriptiva.

1.1.1. Minería de datos predictiva

Se genera un modelo del sistema descrito por los datos dados. Utiliza algunas variables o campos en el conjunto de datos para predecir los valores futuros desconocidos o de otras variables de interés.

1.1.2. Minería de datos descriptiva

Se centra en encontrar patrones que describe los datos que puede ser interpretado por los seres humanos.

1.2. Proceso de descubrimiento de conocimiento

La minería de datos es una parte integral del descubrimiento de conocimiento en bases de datos (“Knowledge Discovery in Databases” – KDD), el cual es un proceso de convertir datos en bruto en información útil (Tan, Steinbach, & Kumar, 2006).

En 1996, Fayyad y Piatesky-Shapiro publicaron un conjunto de tareas que permiten encontrar patrones útiles y que no son triviales en el KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), la figura 1 presenta el diagrama a bloques del proceso de KDD en el que, a partir del conjunto de datos en bruto, se debe hacer una preparación de los datos para determinar los parámetros, estructuras de datos y niveles de información que se

persiguen en el proyecto, y obteniendo un formato adecuado para que posteriormente se apliquen los algoritmos propios de la minería de datos tales como asociación, clasificación y agrupación entre otros, mismos que generan el conjunto de patrones de información que serán útiles para el descubrimiento del conocimiento, dichos patrones son analizados en el post-procesamiento con el fin de producir la interpretación final de la información y el reporte final.

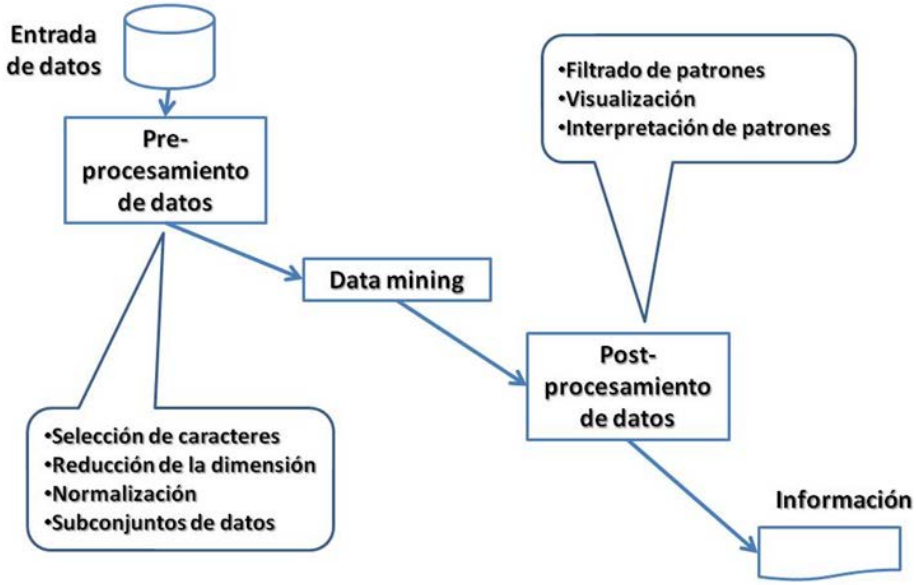

Figura 1. Proceso General de KDD. (Elaboración propia basada en (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) y (Tan, Steinbach, & Kumar, 2006))

2. Retos de investigación

Existen varios retos de investigación tanto de la minería de datos como de su aplicación en las ciencias y la ingeniería, a continuación se mencionan algunos que los autores consideran importantes:

- en las redes de datos,
- en el descubrimiento, comprensión y utilización de patrones,
- en los flujos de datos de las redes,
- en los repositorios de textos estáticos, en la web y otros datos no estructurados,
- por dominios específicos tales como los biológicos e ingeniería de software.

Con la utilización de los *buscadores*[1] o *motores de búsqueda* en las redes ha sido necesario crear un conjunto de técnicas basadas en la minería que permitan la optimización de la búsqueda, mejora el motor de búsqueda y genere una nueva frontera de investigación con amplias aplicaciones tanto en las redes sociales como en la web 2.0 (O'reilli, 2005).

El reto es ir más allá de las redes homogéneas y ahondar profundamente en las redes heterogéneas y multidimensionales, donde existe una gran cantidad de información natural, técnica, social y redes de aplicación de las ciencias y la ingeniería (tales como las redes de investigación genética, biológica, de proteínas, etc.). Dichas redes son altamente evolutivas, dinámicas e interdependientes (Han & Gao, 2009).

Para el descubrimiento de patrones, en grandes almacenes de datos se han utilizado diversas técnicas, principalmente han sido utilizados para la clasificación y asociación, los algoritmos que permiten identificar los patrones están evolucionando de acuerdo a las exigencias tanto del tamaño de la información a analizar como de los recursos computacionales que se tienen.

En los flujos de datos de las redes se puede encontrar un gran volumen de información en tiempo real, con cambios dinámicos y posiblemente infinitos, además de poseer un conjunto multidimensional de características. El uso de las técnicas de síntesis que permiten integrar elementos de voz, video y datos que sean útiles al minero. Se pueden reutilizar algoritmos y técnicas de ventana deslizante para determinar los tiempos de investigación y que los resultados sean patrones en un periodo de tiempo estimado (Datar, Gionis, Indyk, & Motwani, 2002). Uno de los principales algoritmos utilizados en estas investigaciones es el de "árbol de decisión muy rápida" (*VFDT – Very Fast Decision Tree)* (Witten, Frank, & Hall, 2011).

En los repositorios de textos estáticos tales como blogs y redes sociales, la minería de datos se ha convertido en una herramienta de gran valor para determinar patrones de palabras y otros elementos visuales que puedan ser identificados para su clasificación y determinen tendencias a favor o en contra, para la censura u otros. También ha sido muy importante para los grandes repositorios de investigaciones médicas y biomédicas

[1] Herramientas informáticas para la búsqueda de información mediante palabras clave sobre un tema específico.

de donde, con las constantes actualizaciones se hacen los grandes almacenes que permiten identificar patrones útiles. Las tecnologías en el procesamiento de la minería de datos en texto o también denominado *text mining* incluyen estracción de información, seguimiento de temas, sumarización, categorización, agrupación y enlaces de conceptos (Han & Gao, 2009).

Los dominios de datos donde más se ha investigado y donde hay más oportunidad de seguir investigando es en la biología y la ingeniería de software.

- La investigación en la rama de la biología y biomédica ha permitido la evolución y crecimiento de la bio-informática, sin embargo, es también uno de los grandes retos de investigación, el tamaño de la información de los genes y proteínas ha requerido de funciones matemáticas que permitan la optimización y uso de las técnicas de minería de datos para la descripción y análisis de patrones.
- Por otro lado, la ejecución de programas por si mismos genera una gran cantidad de datos dignos de ser analizados mediante técnicas de minería de datos para la optimización de los recursos computacionales donde se aplican dichos sistemas así como de la mejora de los resultados que deben esperarse, se puede identificar el tipo de análisis estático y análisis dinámico, basado en los casos donde el sistema debe coleccionar conjuntos de datos mediante el trazado de los programas para un análisis pre y post ejecución y, finalmente en su ejecución en tiempo real.

3. La minería de datos y la ética

El uso de los datos y la minería de datos han tenido importantes implicaciones éticas principalmente para datos acerca de la gente.

Se deben utilizar con mucho profesionalismo al encontrar patrones de discriminación (racial, sexual, religioso y otros) en cualquier área del conocimiento, excepto cuando su propósito haya sido ese precisamente en las ramas de la medicina y las ciencias biológicas.

En la era del Internet existe información no solo almacenada en repositorios bien conocidos, sino que existe información individualizada y personal a lo largo del mundo,

en diferentes idiomas y en diferentes formatos, sin embargo, las técnicas de minería de datos permiten que se puedan encontrar patrones en fuentes públicas y privadas no seguras.

Existen diferentes algoritmos para garantizar la privacidad de los datos, podemos asumir que tenemos algún e conjunto de datos en el cual la privacidad es relevante (o se basa en información potencialmente sensible sobre individuos), queremos que un analista obtener respuestas a preguntas acerca de los datos en su conjunto, pero sin dejar que el analista de deducir información privada sobre cualquier individuo. La idea está en cómo convertir esta meta intuitiva en una definición precisa que mantenga los datos privados como privados.

4. Conclusiones

La ciencia y la ingeniería son terreno fértil para la investigación mediante las técnicas de minería de datos. Sigue siendo un reto el desarrollo de minería de datos invisible para los sistemas, tal que apoyen al sistema pero que no lo ralenticen en tiempo real. En el presente artículo se presentan solamente algunas de las áreas en las que se han desarrollado proyectos, herramientas de software y hardware, investigaciones que han permitido mejoras no solo a las áreas de investigación pura sino que también, a las áreas de investigación aplicada para el beneficio de los negocios, las ciencias, la tecnología y la sociedad.

Bibliografía

[1] FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. From Data Mining toKnowledge Discovery in Databases. American Association for Artificial Intelligence. All rights reserved. 0738-4602-1996, 37-54.

[2] HAN, J., & GAO, J. Research Challenges for Data Mining in Science and Engineering. En H. Kargupta, J. Han, P. Yu, R. Motwani, & V. Kumar, Next Generation of Data Mining (págs. 3-27). Chapman & Hall / CRC. 2009.

[3] KARGUPTA, H., & SIVAKUMAR, K.. Existential pleasures of distributed data mining. En H. Kargupta, A. Joshi, K. Sivakumar, & Y. Yesha, Data mining next generation challenges and future directions (págs. 3-25). MIT Press. 2004.

[4] O'REILLI, T. O´reilly.com. Recuperado el 14 de 01 de 2014, de What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software: http://oreilly.com/ 30 de 09 de 2005.

[5] TAMISELVI, R., & KALAISELVI, S. An Overview of Data Mining Techniques and Applications. International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, 506-509. 2013.

[6] TAN, P.-N., STEINBACH, M., & KUMAR, V. Introduction to data mining. Pearson Education, Inc. 2006.

[7] WITTEN, I., FRANK, E., & HALL, M. Data mining: practical Machine Learning Tools and Techniques. Burlington, MA.: Morgan Kaufmann Publishers. 2011.