Revista Científica Mundo de la Investigación y el Conocimiento

Galo Alfredo Flores Lagla [a]; José Augusto Cadena Moreano [b]; Edwin Edison Quinatoa Arequipa [c]; Manuel William Villa Quishpe [d]

Minería de datos como herramienta estratégica

Data mining as a strategic tool

Revista Científica Mundo de la Investigación y el Conocimiento. Vol. 3 núm.1, enero, ISSN: 2588-073X, 2019, pp. 955-970

DOI: 10.26820/recimundo/3.(1).enero.2019.955-970

URL: http://www.recimundo.com/index.php/es/article/view/400

Editorial Saberes del Conocimiento

Recibido: 20/11/2018 Aceptado: 05/01/2019 Publicado: 31/01/2019

Correspondencia: galo.flores@utc.edu.ec

a. Magister en Sistemas Informáticos Educativos; Diploma Superior en Didáctica de la Educación Superior; Ingeniero en Informática y Sistemas Computacionales; Universidad Técnica de Cotopaxi; galo.flores@utc.edu.ec
b. Magister en Ciencias de la Educación mención Planeamiento y Administración Educativa; Licenciado en Ciencias de la Educación en la Especialidad de Física y Matemáticas; Ingeniero en Informática y Sistemas Computacionales; Profesor de Educación Media en la Especialidad de Fisico-Matematicas; Universidad Técnica de Cotopaxi; jose.cadena@utc.edu.ec
c. Magister en Ingeniería de Software; Ingeniero en Informática y Sistemas Computacionales; Universidad Técnica de Cotopaxi; edwin.quinatoa@utc.edu.ec
d. Magister en Interconectividad de Redes; Diploma Superior en Comercio Exterior; Ingeniero en Sistemas e Informática; Licenciado en Sistemas Computacionales; Técnico Ejecutivo Analista de Sistemas; Tecnólogo en Computación e Informática; Universidad Técnica de Cotopaxi; manuel.villa@utc.edu.ec

RESUMEN

El desarrollo de la tecnología de la información ha generado una gran cantidad de áreas de investigación siendo la minería de datos una de ellas. La investigación en bases de datos y tecnología de la información ha dado lugar a un enfoque para almacenar y manipular estos datos precisos para una mayor toma de decisiones. La minería de datos es un proceso de extracción de información útil, deriva patrones y tendencias de enormes cantidades de datos. A estos patrones y tendencias se les conoce como modelo de minería de datos y puede ser aplicado en las empresas. A la minería de datos también se denomina proceso de descubrimiento de conocimiento, extracción de conocimiento a partir de datos, extracción de conocimiento o análisis de datos / patrones.

Este artículo presenta un primer paso hacia la unificación del marco para el descubrimiento del conocimiento en bases de datos. Se describen los elementos entre datos milfing, descubrimiento de conocimiento, y otros campos relacionados. Se definen los Procesos del descubrimiento del conocimiento en base de datos y algoritmos básicos de extracción de datos, se discuten problemas de aplicación y concluye con un enfoque de la aplicación de minería de datos en las empresas.

Palabras Claves: Minería de datos; Conocimiento; Procesos.

ABSTRACT

The development of information technology has generated a large number of investigation areas, with data mining being one of them. Investigation in databases and information technology has led to an approach to store and manipulate these precise data for greater decision making. Data mining is a process of extracting useful information, deriving patterns and trends from huge amounts of data. These patterns and trends are known as a data mining model and can be applied in the Business. Data mining is also called knowledge discovery process, knowledge extraction from data, knowledge extraction or data analysis / patterns.

This article presents a first step towards the unification of the framework for the discovery of knowledge in databases. The elements between milfing data, knowledge discovery, and other related fields are described. Processes of knowledge discovery are defined in database and basic data extraction algorithms, application problems are discussed and it concludes with a focus on the application of data mining in the business.

Keywords: Data mining; Knowledge; Processes.

Introducción.

A través de una amplia variedad de campos, se están recopilando datos y acumulado a un ritmo lento. Hay una urgente necesidad de una nueva generación de técnicas computacionales y herramientas para ayudar a los empresarios en la extracción de información útil (conocimiento) del rápido crecimiento en volúmenes de datos, estas técnicas y herramientas son las del campo emergente del descubrimiento del conocimiento en las bases de datos (KDD por sus siglas en Ingles). Este documento es un paso inicial hacia un marco común que esperamos permita entender la variedad de actividades en este multidisciplinario campo y cómo estos encajan entre sí. El proceso de descubrimiento del conocimiento es un conjunto de diversas actividades para dar sentido a los datos. En el centro de este proceso está la aplicación de métodos de minería de datos como patrón de descubrimiento. Se examina cómo la minería de datos utiliza y perfila algunos de sus métodos.

Históricamente la noción de encontrar patrones útiles en los datos han recibido una variedad de nombres, incluido minería de datos, extracción de conocimiento, extracción de información, recolección de información, arqueología de datos y procesamiento de patrones de datos. El término minería de datos ha sido utilizado principalmente por estadísticos, analistas de datos y a lo largo de este documento se utiliza el término "patrón" para designar patrón o modelo extraído de los datos. La gestión de sistemas de información también ha ganado popularidad en el campo de la base de datos. El término KDD fue designado en el primer taller de KDD en 1989 (Piatetsky-Shapiro, 1991) para entender que el "conocimiento" es el producto final de un descubrimiento basado en datos.

KDD se refiere al proceso global de descubrimiento del conocimiento útil de los datos mientras que la minería de datos se refiere a un paso particular en este proceso. Minería de datos es la aplicación de algoritmos específicos para la extracción de patrones a partir de datos. La distinción entre el proceso KDD y el paso de minería de datos (dentro del proceso) es un punto central de este trabajo. Los pasos adicionales en el proceso de KDD, como la preparación, selección y limpieza de datos, incorporación de conocimiento previo y la correcta interpretación de los resultados de minería, son esenciales para asegurar que el conocimiento útil es derivado de los datos. La aplicación ciega de métodos en minería de datos (correctamente criticados como "dragado de datos" en la literatura estadística) puede ser una actividad peligrosa, que fácilmente conduce al descubrimiento de patrones inválidos. KDD ha evolucionado, y continúa evolucionando, desde la intersección de campos de investigación como el aprendizaje automático, reconocimiento de patrones, bases de datos, estadísticas, inteligencia artificial, adquisición de conocimiento para sistemas expertos, visualización de datos y computación de alto rendimiento. El objetivo unificador es extraer conocimientos de alto nivel a partir de datos de bajo nivel en el contexto de grandes conjuntos de datos. La minería de datos se ha convertido en una herramienta estratégica muy importante a la hora de extraer y analizar los datos en una institución o empresa.

Metodología.

Esta investigación está enfocada en el estudio y aplicación del descubrimiento del conocimiento en la minería de datos como factor unificador incluyendo los distintos factores que esto conlleva.

La revisión se ha centrado en textos, documentos y artículos científicos publicados disponibles en la web, considerando que aquella herencia de la globalización nos permite acceder a mayor y mejor información a través de las herramientas tecnológicas. El motor de búsqueda ha sido herramientas académicas de la web que direccionan específicamente a archivos con validez y reconocimiento científico, descartando toda información no confirmada o sin las respectivas referencias bibliográficas.

Resultados.

(Fayyad, Piatetsky-Shapiro, & Smyth, 1996) Definen KDD (**Descubrimiento del Conocimiento en Bases de Datos)** como el proceso no trivial de identificar patrones válidos, novedosos, útiles y comprensibles en datos.

Aquí los datos son un conjunto de hechos (por ejemplo, casos en una base de datos) y el patrón es una expresión en algún lenguaje que describe un subconjunto de los datos o un modelo aplicable a ese subconjunto. También designa ajustar un modelo a los datos, encontrando la estructura a partir de los datos, o en general cualquier descripción de alto nivel de un conjunto de datos. El término proceso implica que KDD consta de muchos pasos, que involucran preparación de datos, búsqueda de patrones, evaluación del conocimiento, y refinamiento, todo repetido en múltiples iteraciones. Por no trivial se quiere decir que se trata de una búsqueda o inferencia, es decir, no es un cálculo directo de cantidades predefinidas como calcular el valor promedio de un conjunto de números. Los patrones descubiertos deben ser válido en nuevos datos con cierto grado de certeza. También queremos que los patrones sean novedosos (al menos para el sistema, y preferiblemente para el usuario) y potencialmente útil, es decir, conducir a algún beneficio para

el usuario / tarea. Finalmente, los patrones deben ser comprensibles, si no inmediatamente, después de algún post-procesamiento.

Lo anterior implica que podemos definir medidas cuantitativas para evaluar patrones extraídos. En algunos casos, es posible definir medidas de certeza (por ejemplo, precisión de predicción estimada sobre nuevos datos) o utilidad (por ejemplo, ganancia, quizás en dólares ahorrados debido a mejores predicciones o aceleración en el tiempo de respuesta de un sistema). Nociones tales como la novedad y la comprensibilidad son mucho más subjetivo. En ciertos contextos comprensibles se puede estimar por simplicidad (por ejemplo, el número de bits para describir un patrón). Una noción importante, llamada curiosidad, generalmente se toma como una medida general del valor del patrón, combinando validez, novedad, utilidad y sencillez.

La minería de datos es un paso en el proceso KDD que consiste en la aplicación de análisis de datos y algoritmos de descubrimiento que, bajo limitaciones de eficiencia computacional aceptables, produce una enumeración particular de patrones sobre los datos. Tenga en cuenta que el espacio de los patrones es a menudo infinito, y la enumeración de patrones implica alguna forma de búsqueda en este espacio. Existen bases de datos en redes sociales, bancos, tiendas, hospitales y más. Para las empresas u organizaciones los datos son materia prima para poder encontrar patrones que favorezcan a interpretar fenómenos o sucesos, por ejemplo un usuario desea saber si puede acceder a un préstamo en el banco, qué producto se vende más según temporadas o cuáles son las causas de una enfermedad.

El proceso KDD es el proceso de usar la base de datos junto con cualquier selección requerida, pre procesamiento, sub muestreo, y transformaciones de la misma; aplicar métodos de

minería de datos (algoritmos) para enumerar patrones; y evaluar los productos de minado de datos para identificar el subconjunto de los enumerados patrones considerados "conocimiento". El componente de minería de datos del proceso KDD es resuelto por los medios algorítmicos por los cuales los patrones se extraen y se enumeran a partir de datos.

El proceso general de KDD incluye la evaluación y posible interpretación de los patrones "minados" para determinar qué patrones pueden ser considerados nuevos conocimientos. La noción de un proceso global impulsado por el usuario no es exclusivo de KDD, propuestas análogas se han presentado en estadísticas (Hand, 1994) y en aprendizaje automático (Brodley & Smyth, 1996).

El proceso de KDD es interactivo e iterativo, involucrando numerosos pasos con muchas decisiones tomadas por el usuario. Brachman & Anand (1996) proponen una vista del proceso de KDD que enfatiza la naturaleza interactiva y aqui se esboza ampliamente algunas de sus pasos básicos:

1. Desarrollar un entendimiento del dominio de la aplicación y el conocimiento previo relevante, y la identificación del objetivo del proceso KDD desde el punto de vista del cliente.

2. Crear un conjunto de datos de destino: selecciona un conjunto de datos, o centrándose en un subconjunto de variables o muestras de datos, en qué descubrimiento hay que realizar.

3. Limpieza y pre procesamiento de datos: operaciones básicas tales como la eliminación de ruido si es apropiado, recogiendo la información necesaria para modelar o dar cuenta para el

ruido, decidiendo estrategias para el manejo de campos faltantes de datos, teniendo en cuenta la información de secuencia de tiempo y cambios conocidos.

4. Reducción de datos y proyección: encontrar características de utilidad para representar los datos de mosaico en función del objetivo usando reducción de dimensionalidad o transformación de métodos para reducir el número de fichas efectivas de las variables en consideración o para encontrar representaciones invariantes para los datos.

5. Hacer coincidir los objetivos del proceso KDD (paso 1) para método particular de minería de datos: por ejemplo, resumen, clasificación, regresión, agrupamiento, etc

6. Eligiendo el (los) algoritmo (s) de minería de datos: seleccionando Método (s) que se utilizará para buscar patrones en los datos. Esto incluye decidir qué modelos y que parámetros pueden ser apropiados (por ejemplo, modelos para categorizar los datos son diferentes a los modelos en vectores sobre los reales) y la coincidencia de una minería de datos en particular con los criterios generales del proceso KDD. (por ejemplo, el usuario final puede estar más interesado entendiendo el modelo que sus capacidades predictivas).

7. Minería de datos; búsqueda de patrones de interés en un forma de representación particular o un conjunto de tales representaciones: reglas de clasificación o árboles, regresión, agrupación, y así sucesivamente. El usuario puede significativamente ayudar al método de minería de datos realizando correctamente los pasos anteriores.

8. Interpretando patrones minados, posiblemente volviendo a cualquiera de los pasos 1-7 para una nueva iteración. Este paso también puede implicar la visualización de los patrones / modelos extraídos, o visualización de los datos dados por los modelos extraídos.

9. Consolidando el conocimiento descubierto: incorporando este conocimiento en otro sistema para la acción adicional, o simplemente documentándolo y reportándolo a partes interesadas.

El componente de minería de datos del proceso KDD a menudo implica repetidas iteraciones de métodos particulares de minería de datos. Los objetivos de descubrimiento de conocimiento están definidos por el uso previsto del sistema en las que se puede tomar en cuenta ciertos aspectos como son: **verificación**, donde el sistema es limitado para verificar la hipótesis del usuario, **descubrimiento**, donde el sistema encuentra de forma autónoma nuevos patrones. Además se subdivide el objetivo de descubrimiento en **Predicción**, donde el sistema encuentra patrones para el propósito de predecir el comportamiento futuro de algunas entidades; y **Descripción**, donde el sistema encuentra patrones con el fin de presentarlos a un usuario en una forma humana comprensible. La mayoría de los métodos de extracción de datos se basan en técnicas probadas de aprendizaje automático, reconocimiento de patones y estadísticas, clasificación, agrupamiento, regresión, así sucesivamente.

Cabe destacar que de los muchos métodos de minería de datos anunciados en la literatura, en realidad solo son unas pocas técnicas fundamentales. La actual representación del modelo subyacente está siendo utilizado por un método particular (es decir, la forma funcional de f en el mapeo x f (x)) generalmente proviene de una composición de un pequeño número de opciones conocidas: polinomios, splines, funciones de núcleo y base, umbral / funciones booleanas, etc.

Así, los algoritmos tienden a diferir principalmente en el criterio de bondad de ajuste utilizado para evaluar el modelo ajuste, o en el método de búsqueda utilizado para encontrar un buen ajuste.

Aunque los términos entre predicción y descripción no son parecidos (algunos de los modelos predictivos puede ser descriptivo, en la medida en que sean comprensibles, y viceversa), la distinción es útil para el entendimiento del objetivo general de descubrimiento. La relativa importancia de la predicción y descripción para particulares aplicaciones de minería de datos pueden variar considerablemente, sin embargo, en el contexto de KDD, la descripción tiende a ser más importante que la predicción. Esto está en contraste a muchas aplicaciones de aprendizaje automático y reconocimiento de patrones donde la predicción es a menudo el objetivo principal. (Parker, 2004)

Se alcanzan los objetivos de predicción y descripción a través de los siguientes métodos primarios de minería de datos.

- **Clasificación**: aprender una función que mapea (clasifica) un elemento de datos en uno de varias clases predefinidas.
- **Regresión**: aprender una función que mapea un dato elemento a una variable de predicción de valor real y el descubrimiento de relaciones funcionales entre variables.
- **Agrupamiento**: identificación de un conjunto finito de categorías o grupos para describir los datos. Muy relacionado con el clustering, es el método de estimación de densidad de probabilidad que consiste en técnicas para estimar a partir de datos, la

función de densidad de probabilidad multivariada conjunta de todas las variables / campos en la base de datos.

- **Resumen:** encontrar una descripción compacta para un subconjunto de datos, por ejemplo, la derivación de resumen o reglas de asociación y el uso de visualización multivariable.
- **Técnicas de modelado de dependencia**: encontrar un modelo que describa dependencias significativas entre variables. (por ejemplo, aprendiendo de redes de creencias).
- **Detección de cambios y desviaciones**: descubriendo los cambios más significativos en los datos anteriormente medidos o normativos.

Habiendo descrito los métodos generales de minería de datos, el siguiente paso es construir algoritmos específicos para implementar estos métodos, se puede identificar tres principales componentes en cualquier algoritmo de minería de datos y estos son:

1. **Modelo de Representación**: el lenguaje utilizado para describir patrones descubribles. Si la representación es demasiada limitada, entonces sin la cantidad de tiempo de entrenamiento necesario producirá un modelo preciso para los datos. Es importante que un analista de datos comprenda las supuestas representaciones que pueden ser inherentes en un método particular. Es igual de importante que un diseñador de algoritmos establezca claramente que asunciones representativas están siendo hechas por un algoritmo en particular.

2. **Criterios de evaluación del modelo:** declaraciones cuantitativas (o "funciones de ajuste") de qué tan bien cumple los objetivos del proceso KDD un determinado patrón (un modelo y sus parámetros). Por ejemplo, los modelos predictivos a menudo son juzgados por la precisión de predicción empírica en algún conjunto de prueba. Los modelos descriptivos pueden ser evaluados a lo largo de las dimensiones de precisión predictiva, novedad, utilidad y comprensibilidad del modelo ajustado.

3. **Método de búsqueda**: consta de dos componentes: parámetro de búsqueda y búsqueda de modelos. Una vez que el modelo de representación y los criterios de evaluación del modelo son fijos, entonces el problema minería de datos se ha reducido a una tarea de optimización: encontrando los parámetros / modelos seleccionados que optimizan los criterios de evaluación. En la búsqueda de parámetros el algoritmo debe buscar para los parámetros que optimizan la evaluación del modelo.

Minería de datos en las empresas.

La minería de datos o (Big Data) es sin duda la tecnología dominante. Empresas y herramientas están adoptando esta tendencia para analizar y visualizar las cantidades masivas de datos. Históricamente, Excel ha sido una herramienta común para analizar y visualizar una colección de datos. Hoy, sin embargo, las empresas necesitan más que el conjunto estándar de funciones que Excel ha ofrecido, para obtener una visión útil de las grandes bases de datos.

Las empresas cuentan con grandes bases de datos, cuyos datos derivan de diferentes fuentes de almacenamiento en diferentes formatos y diferentes contenedores. Las últimas versiones de Excel se combinan con Power BI y se puede utilizar para acceder a una gran

cantidad de datos de la empresa. Las empresas podrán acceder a los datos almacenados en Azure, Hadoop, Active Directory, y más y nuevas capacidades como PowerChart que permitirá visualizar los datos de forma más fácil y útil.

Tratamiento de los datos para la toma de decisiones en las empresas.

La Minería de Datos es el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información no estructurada (interna y externa a la compañía) en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento y así dar soporte a la toma de decisiones en la empresa. (Marcano Aular, 2007)

Apoyo a la toma de decisiones

Las herramientas de apoyo a la toma de decisiones podrían utilizarse también como herramientas para eliminar los resultados innecesarios e irrelevantes obtenidos en el proceso de minería de datos. Igualmente pueden ser consideradas de este tipo, herramientas tales como las hojas de cálculo (Excel), sistemas expertos, sistemas de hipertexto, sistemas de gestión de información de Web y cualquier otro sistema que ayude a analistas y gestores a manejar eficazmente grandes cantidades de datos e información. Recientemente ha aparecido un área nueva llamada gestión del conocimiento, la cual trata de manejar eficazmente los datos, la información y el conocimiento de una organización. (Marcano Aular, 2007)

Conclusiones.

Se ha presentado algunas definiciones de nociones básicas en el campo KDD. Un objetivo primordial es aclarar la relación entre descubrimiento de conocimiento y minería de datos. Se proporciona una visión general del proceso de KDD y Métodos básicos de minería de datos. Dado el amplio espectro de métodos y algoritmos de minería de datos, la visión general es inevitablemente limitada en su alcance: hay muchas técnicas de minería de datos, particularmente métodos especializados para tipos particulares de datos y dominios. A pesar de que varios algoritmos y aplicaciones pueden aparecer bastante diferente en la superficie, no es raro encontrar que comparten muchos componentes comunes. El entendimiento en la minería de datos e inducción de modelos en este componente aclara la tarea de cualquier algoritmo de minería de datos y hace que sea más fácil para el usuario entender su estado general y su contribución y aplicabilidad al proceso KDD.

Este documento representa un paso hacia un marco común que en última instancia proporcionará una visión unificada de los objetivos y métodos generales comunes utilizados en KDD. Es de esperar que esto conduzca eventualmente a una mejor comprensión de la variedad de enfoques en este campo multidisciplinario.

Minería de datos se ha convertido en una herramienta estratégica especialmente en las empresas porque por medio de ella se podrá realizar una extracción de la información, un análisis de datos y a su vez estos resultados permitirán tomar mejores decisiones sobre la empresa.

Bibliografía.

Brodley, C., & Smyth, P. (1996). *Applying classification algorithms in practice.*

Cheeseman, P. (1990). *Encontrar el modelo más probable en Modelos Computacionales de Descubrimiento Científico y Formación Teórica.*

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery.*

Hand, D. (1994). *Deconstructing statistical questions.* .

Marcano Aular, Y. J. (2007). Minería de Datos como soporte a la toma de decisiones empresariales. *Scielo*, 104-118.

Parker, G. (2004). *Data Mining: Modules in emerging fields.*

Piatetsky-Shapiro, G. (1991). *Knowledge Discovery in Real Databases*