Instituto Tecnológico de Orizaba

MINERÍA DE DATOS (DATAMINING)

Autor:
Ixhel Escamilla Illescas
Website: www.ixhel.mx

FUNDAMENTOS DE INGENIERÍA ADMINISTRATIVA
Dr. Fernando Aguirre y Hernandez

ARTÍCULO DE INVESTIGACIÓN

TABLA DE CONTENIDO

RESUMEN

En éste artículo se revisa el significado de la Minería de datos (Datamining), su objetivo y los tipos de datamining que existen, así como el lugar que ocupa en en el proceso de generación del conocimiento y las técnicas que se emplean para ello. Se muestran también las áreas de aplicación para la misma y se discute que tan anónimos deben ser los datos o la falta de anonimato en ellos, debido a los factores de confidencialidad y confiabilidad, que también puede ser un área de exploración en los datos anómalos.

Agradezco al Instituto Tecnológico por los temas que componen la selección de base cultural que refuerzan mis modelos mentales, así como la libertad de pensamiento que me otorgan al escribir éste articulo, materia prima de mis relaciones sinápticas aparentemente inconexas, gracias por nutrir mis propuestas idealistas como un fomento racional a la sensación intuitiva de la necesidad de un cambio en el modelo de vida actual, por otorgarme las herramientas para progresar estas ideas de materialización de un futuro fuera del esquema que supone el éxito como una lucha por el poder y las terribles consecuencias de éstos patrones.

PALABRAS CLAVE

Datamining

Minerías de datos

Análisis de información

Procesamiento de datos

Bases de datos

INTRODUCCIÓN

La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativa relaciones, patrones y tendencias al examinar grandes cantidades de datos. Es el proceso de analizar los datos desde diferentes perspectivas, información que puede ser utilizada para aumentar los ingresos, reducir los costes, o ambos. El software de minería de datos permite a los usuarios analizar datos de muchas dimensiones o ángulos diferentes, categorizarlo, y resumir las relaciones identificadas. Las organizaciones utilizan potentes servidores para segmentar la información a través de volúmenes de datos y analizar los informes de investigación de mercado durante años. Sin embargo, las continuas innovaciones en la potencia de computación, almacenamiento en disco, y el software de estadística está aumentando drásticamente la exactitud del análisis al tiempo que reduce el costo. El objetivo principal de la minería de datos es el de extraer la información de un conjunto de datos, de seleccionarla y refinarla para poder transformarla en una estructura que sea comprensible para usarla posteriormente. Las organizaciones que emplean la minería de datos pueden ver rápidamente el retorno de su inversión puesto que dejan de dar pasos equivocados.

OBJETIVO DEL DATAMINING

Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnología de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos.

BASES TEÓRICAS

Los datos son los hechos, números o texto que pueden ser procesados por un recurso informático. La Web como ecosistema contiene y genera un universo de datos, tanto provenientes del propio contenido de sus páginas y la estructura de sus enlaces como de su uso por parte de las personas. Estos datos tienen una importancia crucial para el mejoramiento de la misma desde un punto de vista social y también comercial. Por esta razón la minería de datos ha crecido rápidamente y es una herramienta vital para entenderla y dar valor económico a los datos que obtenemos de ella.

TIPOS DE DATOS

Desde el punto de vista conceptual, los datos de una organización pueden ser **operacionales** o **transaccionales**, tales como ventas, costos, inventarios, nómina y contabilidad; datos **no operacionales**, tales como ventas de la industria, los datos de pronóstico y datos macro económicos y finalmente los **datos meta**, que son datos acerca de los datos en sí, como el diseño de base de datos lógicos o definiciones del diccionario de datos.

TIPOS DE DATAMINING

En una era globalizada, de los datos que viajan a través de internet se distinguen tres tipos de minería:

- Minería de contenido: texto, imágenes, etiquetas (tags), metadatos, etc.;
- Minería de estructura: enlaces y sus relaciones; y
- Minería de uso: interacción de las personas con la Web.

Los dos primeros tipos de datos se obtienen recolectando todo el contenido de los sitios web usando software especial llamado recolector (crawler). Un recolector comienza con un conjunto de sitios iniciales y sigue todos los enlaces que encuentra en las páginas según un cierto conjunto de reglas predeterminadas (por ejemplo, qué dominios o tipos de ficheros recorrer). Los datos de uso provienen de los registros (logs) de los servidores web y de aplicaciones específicas, como las de búsqueda.

Son diversas las técnicas que se emplean para analizar estos datos, pero sin duda la más popular es lo que se llama aprendizaje automático. Consiste en aprender como predecir variables en función de otras variables a través de subconjuntos de datos completos y luego evaluar cuán buena es la predicción en otro subconjunto de datos. El algoritmo resultante se usa en los datos reales con la suposición de que su desempeño será similar. Este proceso se repite en el tiempo para ir mejorando la herramienta con casos difíciles. Para esto se pueden utilizar árboles de decisión, máquinas de soporte vectorial o redes neuronales, entre otros.

La minería de datos es sólo una etapa del proceso de extracción del conocimiento a partir de datos **(KDD)**. Éste proceso consta de varias fases como la preparación de datos (selección, limpieza y transformación), su exploración y auditoría, minería de datos propiamente dicha (desarrollo de modelos y análisis de datos), evaluación, difusión y utilización de modelos (output). Además, el proceso de extracción del conocimiento incorpora muy diferentes técnicas (árboles de decisión, regresión lineal, redes neuronales artificiales, técnicas bayesianas, máquinas de soporte vectorial, etc.) de campos diversos (aprendizaje automático e inteligencia artificial), estadística, bases de datos, etc., y aborda una tipología variada de problemas (clasificación, categorización, estimación/regresión, agrupamiento, etc.)

KDD
(KNOWLEDGE DISCOVERY IN DATABASES)

1 El KDD comienza con la ***recopilación e integración*** de la información a partir de unos datos iniciales de que se dispone (fase de selección de datos). Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Generalmente, la información que se quiere investigar sobre cierto tema de interés para la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas. Muchas de esas fuentes son las que se utilizan para el trabajo transaccional. El análisis posterior será mucho más sencillo si la fuente es unificada, accesible y desconectada del trabajo transaccional. Aparte de información interna de la organización, los almacenes de datos (Data Warehouse) pueden recoger información externa, entonces la disponibilidad de grandes volúmenes de información en esta fase nos lleva a la necesidad de usar técnicas de muestreo para la selección de datos.

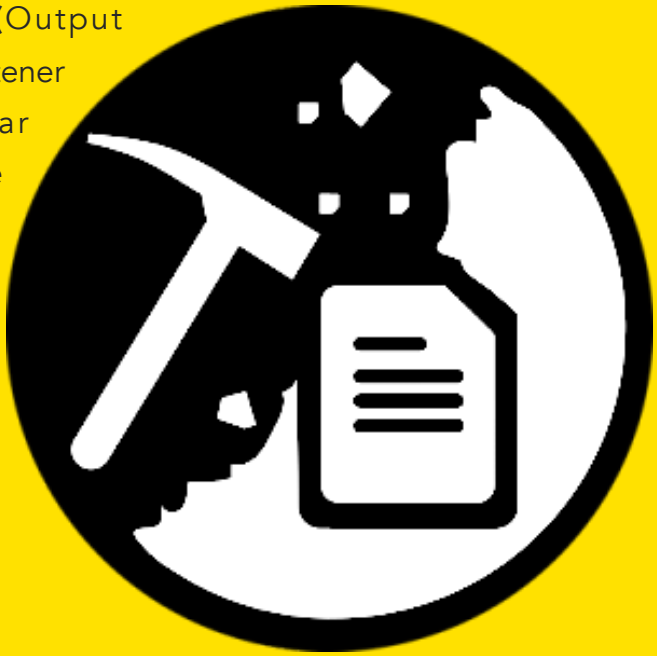
2 La siguiente fase del KDD integra la ***exploración, limpieza de datos y transformación*** (Data Cleaning). Se deben eliminar el mayor número posible de datos erróneos o inconsistente e irrelevantes.

En esta fase se utilizan herramientas de consulta (Query Tools) y herramientas estadísticas (Statistics Tools) casi exclusivamente. En la exploración se usan técnicas de análisis exploratorio de datos como los histogramas y los diagramas de caja, que ayudan a detectar datos anómalos o atípicos (Outliers). La presencia de datos atípicos y valores desaparecidos puede llevarnos a usar algoritmos robustos para filtrar la información y reemplazar valores mediante técnicas de imputación y a transformar datos continuos en discretos mediante técnicas de discretización. Entre las técnicas avanzadas de transformación tenemos las de reducción y aumento de la dimensión.

3 La fase siguiente es la propia ***minería de datos*** que se llevará acabo a partir del desarrollo de modelos predictivos y descriptivos (Model Development) y mediante un análisis de datos (Data Analysis). Una vez recogido los datos de interés, un explorador puedes decidir qué tipo de patrón quiere descubrir. El tipo de conocimiento que se desea extraer va a marcar claramente la técnica de minería de datos a utilizar.

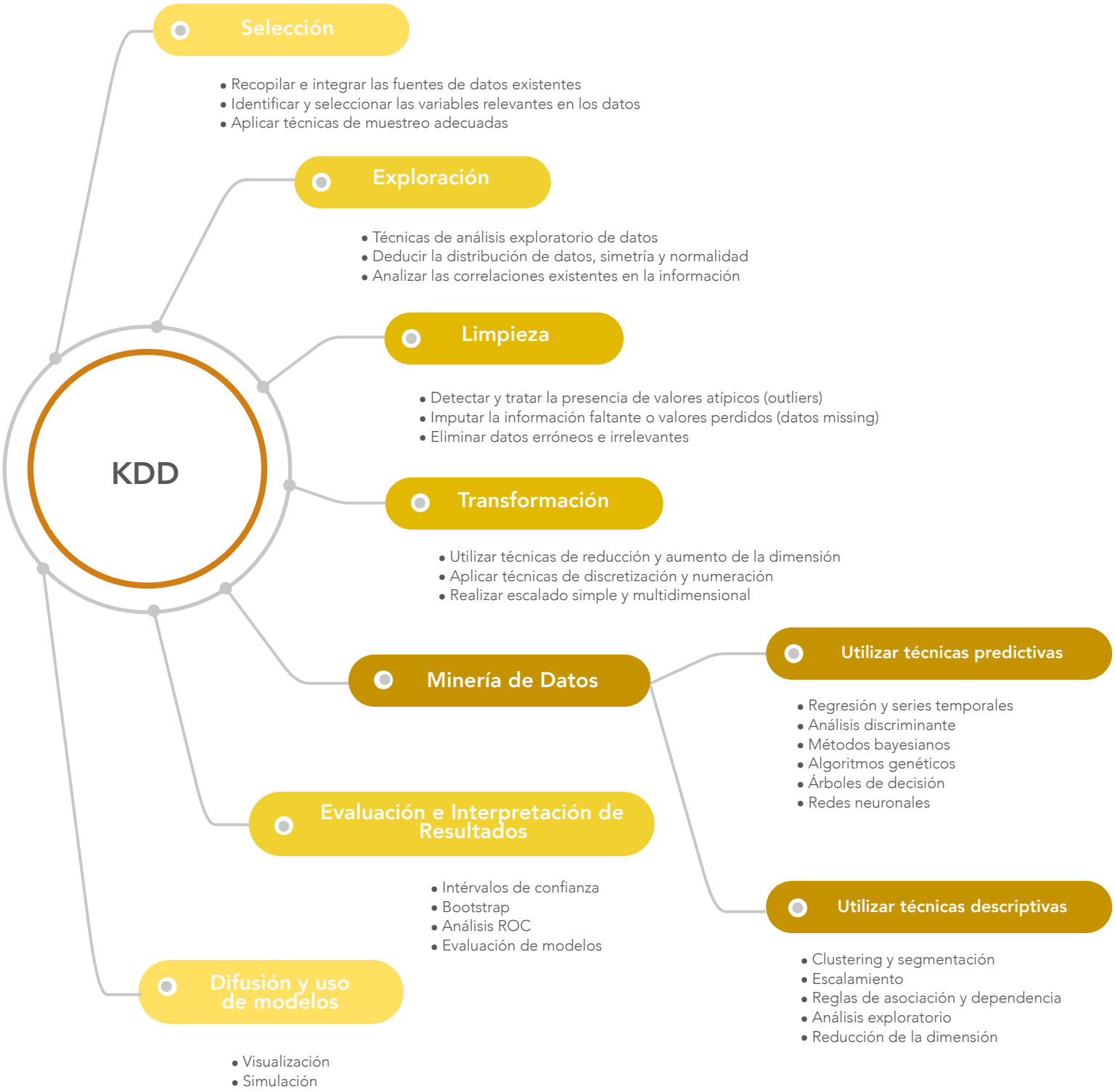
Para seleccionar y validar los modelos anteriores es necesaria una nueva fase consistente en el uso de criterios de evaluación de hipótesis. Despliegue del modelo a veces es trivial pero otras veces requiere un proceso de implementación o interpretación. En esta fase se utilizan adicionalmente herramientas estadísticas y de visualización (Visualization Tools).

4 La fase posterior del KDD es la relativa a la ***difusión y el uso del conocimiento*** derivado de las técnicas de minería de datos a través de los modelos correspondientes que habitualmente desembocan en la generación de resultados (Output Generation). El modelo puede tener muchos usuarios y necesitar difusión, con lo que puede requerir ser expresado de una manera comprensible para ser distribuido en la organización. Etapa se utilizan herramientas de visualización, presentación y transformación de datos.



KDD y MINERIA DE DATOS

De acuerdo a López, C. P. (2007), la clasificación de las fases del proceso de extracción del conocimiento para la minería de datos, podría reducirse en el siguiente esquema:



TÉCNICAS DE DATAMINING

La clasificación inicial de las técnicas de minería de datos distingue entre ***técnicas predictivas***, en las que las variables pueden identificarse inicialmente en dependientes e independientes y ***técnicas descriptivas,*** en la que todas las variables tienen inicialmente el mismo estatus así como ***técnicas auxiliares***.

Las **Técnicas Predictivas** especifican el modelo para los datos en base a un conocimiento teórico previo. Ésa su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa.

- Algoritmos genéticos: técnicas de optimización que utilizan procesos en un diseño basado en los conceptos de la evolución natural.
- Las redes neuronales artificiales: modelos predictivos no lineales que se asemejan a las redes neuronales biológicas en la estructura.
- Los árboles de decisión: CART y CHAID son técnicas de árbol de decisión utilizadas para la clasificación de un conjunto de datos. Proporcionan un conjunto de reglas que se pueden aplicar a un nuevo conjunto de datos (sin clasificar) para predecir qué registros tendrán un resultado dado.

En las **Técnicas Descriptivas** no se le asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes e independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente partiendo del reconocimiento de patrones. Están enfocadas al descubrimiento del conocimiento embebido en los datos.

- Clustering y segmentación: Técnica de clasificación descriptiva que clasifica individuos observaciones dentro de grupos previamente definidos.
- Reducción de la dimensión: se refiere a la atención del factoriales, los componentes principales las correspondencias entre ellos etc.

Las **Técnicas Auxiliares** son herramientas de apoyo más superficiales limitadas. Se trata de nuevos métodos basados en técnicas estadísticas descriptivas, consultas e informes y enfocados en general hacia la verificación.

APLICACIÓN DEL DATAMINING

En las organizaciones, los modelos de Datamining, se pueden aplicar en los siguientes escenarios:

- Previsión: Permite calcular las ventas y predecir cargas o tiempos de inactividad de los servidores.
- Riesgo y probabilidad: Ayuda a elegir a los mejores clientes para la correcta distribución de correo y asigna probabilidades de diagnóstico o algunos otros resultados.
- Recomendaciones: Sirve para determinar productos que se pueden vender juntos y generar algunas recomendaciones.
- Buscar secuencias: Analiza artículos que clientes han introducido en un carrito de compra y así predecir posibles eventos.
- Agrupación: Separa clientes o eventos en clústeres determinados y así analizar o predecir afinidades.

La minería de datos es principalmente usada por las empresas con un fuerte enfoque del consumidor (minorista, financiero, comunicación y organizaciones de marketing). Permite a estas empresas determinar las relaciones entre los factores internos como el precio, el posicionamiento del producto, o las habilidades del personal, y los factores externos, tales como los indicadores económicos, la competencia y demografía de los clientes. Y, que les permite determinar el impacto en las ventas, satisfacción del cliente y las ganancias corporativas. Por último, les permite "profundizar" en la información de resumen para ver los datos transaccionales detallados.

La minería de datos es también utilizada en sectores gubernamentales, agencias anti terrorismo, universidades, investigaciones especiales, deportes y medicina.

DISCUSIÓN

Uno de los temas más importantes relacionados con la minería de datos es su privacidad, y para mantenerla muchas veces se *anonimizan* identificadores de usuario, direcciones IP o cualquier otro dato que pueda identificar a una persona. Sin embargo con el uso de IPs dinámicas, distintas personalidades, computadores compartidos, etcétera, es difícil poder identificar a una persona, más aún cuando esos datos están distribuidos entre el proveedor de internet y el sitio web. Una técnica muy usada para preservar la privacidad es que los datos sean k-anónimos, es decir que no se pueda distinguir los datos de una persona de al menos otras (k-1) personas. Por ejemplo, si k=10, habrá subconjuntos de datos de personas de al menos tamaño 10 que son iguales. En algunos tipos de datos garantizar que sean k-anónimos no es trivial, cosa que pasó cuando AOL publicó en la Web un registro de su buscador que incluía consultas con sesiones anónimas. Con estos datos un periodista identificó a una persona usando una sesión que contenía en las preguntas un código postal y un medicamento poco usual: cruzó esos datos con información pública de los hospitales correspondientes a esa zona. Por esta razón los buscadores han decidido no publicar este tipo de información y limitar el tiempo de almacenamiento de este tipo de datos (18 meses en el caso de Google y Microsoft Live y sólo 13 meses en Yahoo!) y guardarlos usando técnicas de anonimización más poderosas.

Otro tema de discutir, es la minería de datos anómalos, que representa un área de la minería de datos que aborda el problema de la detección de datos raros o comportamientos inusuales en los datos. Esta disciplina tiene una alta aplicación en disímiles escenarios, entre los que se destacan el aseguramiento de ingresos en las telecomunicaciones, la detección de fraudes financieros, la seguridad y la detección de fallas en la gestión de organizaciones orientadas al desarrollo de proyectos de software.

CONCLUSIÓNES

Las tendencias en minería de datos son las de la misma Web. Estamos viendo sólo su comienzo, y queda mucho por hacer. La minería de datos se presenta como una tecnología que está emergiendo con varias ventajas, como el punto de encuentro entre investigadores y personas de negocios, y el ahorro de grandes cantidades de dinero a la organización además de que permite abrir nuevas oportunidades de negocio. Trabajar con el Datamining implica cuidar tantos detalles que al final permite la toma de decisiones de manera precisa.

Existe una gran diversidad de datos, en conjuntos cada día más voluminosos, y que abarcan periodos de tiempo más largos. En cada uno de ellos hay innumerables preguntas a responder y casos extraños a encontrar. Estas preguntas pueden ser desde medidas específicas hasta modelos para el comportamiento de millones de personas. Podemos plantear la posibilidad de descubrir el futuro de las ciencias sociales y hacia donde está encaminada la humanidad en la correcta explotación de los datos que nos describen como sociedad, en conjunto con las tendencias exploratorias que muestran comportamientos recurrentes, no solamente para las organizaciones, sino también para el desarrollo humano.

Tema de tesis: Estudio de patrones utilizando técnicas de Datamining para la detección de actitudes sociales que refuerzan el comportamiento sustentable como herramienta motivacional en las organizaciones.

REFERENCIAS

López, C. P. (2007). Minería de datos: técnicas y herramientas. Editorial Paraninfo.

Baeza-Yates, R. (2009). Tendencias en minería de datos de la Web. El Profesional de la Información, 18(1), 5–10. https://doi.org/10.3145/epi.2009.ene.01

Maté Jiménez, C. (2014). Big data. Un nuevo paradigma de análisis de datos. Revista: Anales de Mecánica y Electricidad, Periodo: 1, Volumen: XCI, Número: VI, Página inicial: 10, Página final: 16. Recuperado de https://repositorio.comillas.edu/xmlui/handle/11531/4873

Castro Aguilar, G. F., Pérez Pupo, I., Piñero Pérez, P. Y., Martínez, N., & Crúz Castillo, Y. (2016). Aplicación de la minería de datos anómalos en organizaciones orientadas a proyectos. Revista Cubana de Ciencias Informáticas, 10, 195–209. Recuperado de http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S2227-18992016000500015&lng=es&nrm=iso&tlng=en

Chakrabarti, Soumen. (2002)Mining the Web: discovering knowledge from hypertext data. Morgan-Kaufmann Publishers.

Durán Mena, C. (6 de Agosto de 2014). Forbes México. Obtenido de https://www.forbes.com.mx/mineria-de-datos-informacion-precisa-y-relevante/