Bibliominería, datos y el proceso de toma de decisiones*

Resumen

Este artículo trata sobre bibliominería de datos, que es minería de datos aplicada a grandes volúmenes de datos disponibles en las bibliotecas, como resultado de la operación de los principales sistemas transaccionales, tales como préstamos, referencia, adquisiciones, entre otros. Así, la bibliominería de datos es el proceso que tiene como propósito descubrir, extraer y almacenar información relevante de grandes bases de datos existentes en las bibliotecas, mediante la utilización de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores que pueden extraerse mediante distintas técnicas de minería de datos. Es importante señalar que para el análisis bibliométrico se requiere de la participación de equipos interdisciplinarios formados por ingenieros de sistemas, estadísticos y bibliotecólogos.

Palabras clave: minería de datos, bibliominería, toma de decisiones, patrones de datos, tendencia de datos, *datawarehouse*, big data.

Mynor Fernandez Morales
Máster en Administración de Empresas, Universidad de Costa Rica. Licenciado en Ciencias de la Computación e Informática. Profesor asociado investigador, Universidad de Costa Rica.
San José – Costa Rica
mynor.fernández@ycr.ac.cr
orcid.org/0000-0003-1162-994X

Roger Bonilla Carrión
Magíster Scientiae en Estadística, Universidad de Costa Rica. Licenciado en Estadística. Profesor de la Escuela de Bibliotecología y Ciencias de la Información, Universidad de Costa Rica.
San José – Costa Rica
roger.bonilla@ucr.ac.cr
orcid.org/0000-0002-8789-4494

Cómo citar este artículo: Fernández, Mynor; Bonilla, Roger (2020). Bibliominería, datos y el proceso de toma de decisiones. *Revista Interamericana de Bibliotecología*, *43*(2), eI8. https://doi.org/10.17533/udea.rib.v43n2eI8

Recibido: 2019-10-24 / **Aceptado**: 2020-03-17

* El presente texto es un avance del proyecto "Bibliominería de datos en el campo de las ciencias bibliotecológicas y de la información", de la línea Bibliotecas Universitarias de la Escuela de Bibliotecología y Ciencias de la Información de la Universidad de Costa Rica, que cuenta con el apoyo económico de la Vicerrectoría de Investigación - VI de la Universidad de Costa Rica. UCR, San Pedro 2060. San José – Costa Rica.

Bibliomining, Data, and the Decision Making Process

Abstract

This article deals with data libraries, where bibliomining is applied data mining on large volumes of data available in libraries, as a result of the operation of major transactional systems, such as loans, referrals, acquisitions, among others. Thus, the data library is the process that aims to discover, extract and store relevant information from large databases in libraries, through the use of search programs and identification of patterns and global relationships, trends, deviations and other Indicators that can be extracted through different techniques of data mining. It is important to note that bibliometric analysis requires the participation of interdisciplinary teams formed by system engineers, statisticians and librarians.

Keywords: Datamining, bibliomining, decision making, data patterns, trend data, datawarehouse, big data.

1. Antecedentes

Gracias al empeño de muchos especialistas de diferentes campos, han surgido nuevas aplicaciones tanto en software libre como en software privativo, que permiten la automatización integral de todos los servicios y procesos realizados en las unidades de información y la generación de servicios agregados al usuario y desarrollo de bibliotecas virtuales, además de múltiples aplicaciones en software privativo.

Es importante señalar que estas nuevas aplicaciones se fundamentan en bases de datos relacionales que vienen a facilitar el desarrollo de depósitos de datos (DD), que según Poe (1998) y como lo señala Chinchilla (2011), constituye "una base de datos de solo lectura, donde la información extraída de los sistemas operacionales corrientes de la empresa es transformada, integrada y resumida para luego ser usada con efectividad en el soporte de decisiones" (p. 56).

Con el desarrollo de los DD viene el surgimiento de los mercados de datos (conocidos en inglés como *datamarts*), que abarcan una parte funcional de la organización, como referencia, circulación, adquisiciones, o de los *datawarehouse* (en adelante DW), que abarcan toda la organización. Ambos serán el insumo principal de la bibliominería, donde, según Chinchilla (2011), los mercados de datos (MD) se pueden definir de la siguiente forma:

> La exposición se centra, particularmente, en que el modelo del MD está definido por la forma en que el usuario necesita ver la información y cómo quiere que se le presente; en este sentido, el MD posee las mismas características de un DD pero a un nivel más específico, ya que contiene diferentes combinaciones y selecciones de los datos que se encuentran en el DD y ofrece una mayor personalización de los datos del departamento, permitiendo un manejo más eficiente de la información histórica, ejecución de procesamiento independiente del resto de los departamentos y un costo de almacenamiento y procesamiento inferior. (p. 56)

De esta forma, los datos que resumen las actividades de la unidad de información, provenientes de los diferentes sistemas transaccionales soportados sobre los DD que utilizan bases de datos relacionales, deben ser consolidados en una base de datos que denominaremos DW, que será el soporte para los procesos de minería de datos que facilitan la toma de decisiones en la unidad de información.

Debido a que se consideran datos históricos generados por los sistemas transaccionales, estos DW son grandes volúmenes de datos que almacenan información valiosa de la organización, que debe ser procesada con técnicas de minería de datos que se explicarán más adelante. Se debe tener claro que *bibliominería* es un término bastante nuevo, según de la Puente (2010):

> La minería de datos aplicada a las bibliotecas se denomina Bibliominería, término que deriva del inglés, bibliomining, como una derivación de los términos bibliometría (bibliometrics) y minería de datos (data mining). Se define como la combinación de minería de datos, Bibliometría, Estadística y herramientas de elaboración de informes y extracción de patrones de comportamiento, que se presentan en los sistemas bibliotecarios. (p. 3)

Es una realidad que el usuario es el norte principal de toda la actividad bibliotecaria, por lo que el amplio conocimiento sobre él es esencial para ofrecerle los servicios que requiere, como un objetivo estratégico que debe cumplir la unidad de información.

Rev. Interam. Bibliot. Medellín (Colombia) Vol. 43, número 2/mayo-agosto 2020 eI8 ISSN 0120-0976 / ISSN (en línea) 2538-9866
https://doi.org/10.17533/udea.rib.v43n2eI8

2. Introducción

Aunque es una realidad que las unidades de información son las proveedoras del componente estratégico y materia prima para el desarrollo de la sociedad del conocimiento, contradictoriamente se encuentran en el último lugar en las prioridades de inversión en las organizaciones (Arriola y Butrón, 2008).

Por tanto, las unidades de información generalmente carecen de presupuesto para realizar proyectos de automatización, de ahí la importancia estratégica de que los proyectos de automatización en que estas se involucren produzcan un mayor valor agregado en los servicios ofrecidos a los usuarios. Objetivo que se cumple a cabalidad con los proyectos de bibliominería de datos que facilitan el entendimiento de los usuarios para la satisfacción de sus necesidades.

A través del proceso de minería de datos, se utilizan técnicas estadísticas y técnicas de reconocimiento de patrones e identificación de tendencias, en la información almacenada. De esta forma, la bibliominería facilita una forma para conocer a nuestros usuarios, ofreciendo una gran contribución al momento de estudiarlos para ofrecerles los servicios de calidad que en realidad ellos requieren.

Los beneficios de la bibliominería para las bibliotecas están orientados a conocer mejor al usuario y, así, mejorar los servicios bibliotecarios. De acuerdo con Pal (2011), la bibliominería:

> Siempre revela un patrón de actividad dentro de la biblioteca, y que estos patrones pueden tener beneficios potenciales en tres diferentes niveles:
> 1. Para los usuarios a través de la mejora de los servicios bibliotecarios.
> 2. Para la gestión bibliotecaria, al proveer información para una mejor toma de decisiones.
> 3. Para la institución que alberga la biblioteca, a través de informes de los patrones relevantes que muestran el comportamiento de los usuarios. (p. 12)

Por tanto, se puede resumir que la minería de datos es el hallazgo de conocimiento a partir de datos, que se realiza a través de la exploración y análisis de grandes volúmenes de datos almacenados en las bases de datos de la unidad de información, con la finalidad de descubrir correlaciones significativas, nuevos patrones y tendencias entre los datos explorados y analizados, que faciliten la toma de decisiones para mejorar los servicios que esta presta.

De acuerdo con de la Puente (2010), el proceso de la bibliominería, según los diversos especialistas del área, se compone de seis fases:

> 1. Determinación de los campos temáticos de interés.
> 2. Identificación de fuentes de información internas y externas.
> 3. Recolección, depuración y proceso de ocultamiento de la identidad de usuarios en el almacén de datos del sistema o DW.
> 4. Selección de las herramientas de análisis.
> 5. Descubrimiento de patrones, tendencias y elaboración de informes.
> 6. Análisis e implementación de los resultados. (p. 4)

Para definir los campos temáticos de interés, se debe hacer un estudio en el área del usuario para determinar cuáles son sus requerimientos.

Luego de establecidos los requerimientos, se deben identificar las fuentes de información externas o internas que permitirán satisfacerlos de forma apropiada. Hasta aquí es un trabajo de análisis en el área del usuario que permitirá clarificar sus necesidades reales de información, en palabras resumidas, se está definiendo el "¿qué hay qué hacer?".

Posteriormente, lo que sigue es un trabajo de recolección de datos a partir de las fuentes identificadas, tanto internas como externas, para posteriormente realizar un proceso de depuración de datos, y finalmente establecer algoritmos que oculten la identidad de los usuarios que originan la información, ya que esta es de carácter privado y propiedad exclusiva del usuario.

Luego, se podrán hacer análisis bibliométricos para diagnosticar el funcionamiento de otras áreas de interés en la biblioteca, que requieran ser optimizadas, para satisfacer en mejor medida las necesidades de los usuarios.

Por último, aplicando estas técnicas de análisis de bibliominería, se facilitará el descubrimiento de patrones y tendencias ocultos en los datos que permitan la generación de informes en cuanto a dirección, para finalmente ejecutarlos a través de la toma de decisio-

nes en la organización, fundamentadas en los análisis realizados. Por tanto, el objetivo principal de la bibliominería de datos es proporcionar la información requerida para gestionar de mejor forma la organización, respondiendo a preguntas claves que le faciliten a esta el cumplimiento de sus objetivos estratégicos.

3. Conceptos teóricos sobre bibliominería

Teniendo presente que las técnicas de minería de datos aplicables a bibliotecas se conocen con el nombre de bibliominería, es importante señalar que, según Nicholson (2003), "la bibliominería es la aplicación de herramientas estadísticas y de reconocimiento de patrones a una gran cantidad de datos relacionados con los sistemas bibliotecarios, con la finalidad de ayudar en la toma de decisiones o para justificar los servicios" (p. 146).

La bibliominería es una disciplina muy ligada al uso de técnicas estadísticas que nos permiten extraer patrones de comportamiento de grandes volúmenes de datos. Según Aluja (2001, p. 484), las técnicas de minería de datos más utilizadas son las siguientes.

3. 1. Análisis factoriales descriptivos

Permiten hacer visualizaciones de realidades multivariantes complejas y, por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.

3.2. Market Basket Analysis

También llamad análisis de la cesta de la compra. Permite detectar qué productos se adquieren conjuntamente, permite incorporar variables técnicas que ayudan en la interpretación, como el día de la semana, localización, forma de pago. También puede aplicarse en contextos diferentes del de las grandes superficies, en particular el e-comercio, e incorporar el factor temporal.

3.3. Técnicas de *clustering*

Son técnicas que parten de una medida de proximidad entre individuos y, a partir de ahí, buscan los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

3.4. Series temporales

A partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad, y así poder hacer predicciones para el futuro, tales como cifra de ventas, previsión de consumo de un producto o servicio, etc.

3.5. Redes bayesianas

Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.

3.6. Modelos lineales generalizados

Son modelos que permiten tratar diferentes tipos de variables de respuesta, por ejemplo, la preferencia entre productos concurrentes en el mercado. Al mismo tiempo, los modelos estadísticos se enriquecen cada vez más y se hacen más flexibles y adaptativos, permitiendo abordar problemas cada vez más complejos (GAM, Projection Pursuit, PLS, MARS, etc.).

3.7. Previsión local

La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una cierta variable de respuesta. La técnica consiste en situar los individuos en un espacio euclídeo y hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos.

3.8. Redes neuronales

Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son "una caja negra".

Rev. Interam. Bibliot. Medellín (Colombia) Vol. 43, número 2/mayo-agosto 2020 eI8 ISSN 0120-0976 / ISSN (en línea) 2538-9866
https://doi.org/10.17533/udea.rib.v43n2eI8

3.9. Árboles de decisión

Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.

3.10. Algoritmos genéticos

También aquí se simula el modelo biológico de la evolución de las especies, solo que a una velocidad infinitamente mayor. Es una técnica muy prometedora. En principio, cualquier problema que se plantee —como la optimización de una combinación entre distintos componentes, estando estos sujetos a restricciones— puede resolverse mediante algoritmos genéticos.

Según Rao, citado también por Aluja (2001), la estadística es una metodología para la extracción de la información y manejar la cantidad de incertidumbre en las decisiones que tomamos (p. 479).

Así, este conjunto de herramientas estadísticas nos servirán para identificar tendencias y patrones de comportamiento almacenados en forma implícita en las bases de datos de la unidad de información, lo que nos facilitará la toma de decisiones sobre aspectos que afectan a los usuarios en el uso de los servicios. Es importante destacar que el proceso de bibliominería es un trabajo interdisciplinario que requiere del aporte de bibliotecólogos, informáticos y estadísticos.

El objetivo primordial de la bibliominería es entonces extraer patrones de uso de la colección a partir de la información contenida en grandes volúmenes de datos disponibles en la biblioteca, con la finalidad de poder predecir cuáles serán los futuros comportamientos de los usuarios en el uso de los servicios de la biblioteca, para facilitar la toma de decisiones en una unidad de información, con el fin de realizar una mejor gestión de estas.

4. Bibliominería y aprovechamiento en la biblioteca para mejor toma de decisiones

Antes de entrar en el tema central de este artículo, que es la bibliominería aplicada en bibliotecas, es importante analizar lo que dice Fernández (2012a) sobre la propuesta de valor de una organización:

> Una unidad de información, luego de hacer un planeamiento estratégico, define cuál será la propuesta de valor para sus usuarios, que al final se traduce en el conjunto de servicios que la unidad ofrece al usuario, bajo una dimensión de tiempo y espacio que satisfaga las necesidades del usuario. Esta propuesta de valor permitirá enrolar como clientes a todos aquellos usuarios que la propuesta de valor establecida satisfaga sus expectativas de servicio. (p. 5)

Entonces, una biblioteca que cumpla su propuesta de valor tendrá unos usuarios satisfechos, mientras otra que no la cumpla, probablemente, enfrentará la problemática de que sus usuarios migrarán a otras bibliotecas.

De esta manera, a través de la bibliominería es posible determinar si se está cumpliendo o no la propuesta de valor, así como determinar cuáles son las necesidades del usuario a través del estudio del uso que hacen ellos de la colección. Entre los aspectos que podríamos obtener con el uso de la bibliominería están 1) cuáles son los principales temas de interés de los usuarios, 2) cuáles son sus principales búsquedas, 3) cuáles búsquedas fueron exitosas y cuáles infructuosas y 4) identificación de las palabras claves que ellos utilizan para búsqueda de material. Todo con la finalidad de poder dimensionar las necesidades reales de los usuarios y así poder mejorar el servicio prestado.

La bibliominería trata de determinar los patrones de uso de la colección por parte de los usuarios, la extracción de patrones de comportamientos de los usuarios en el uso de los servicios bibliotecarios, con utilidad para la toma de decisiones y la selección de recursos, la organización de la colección y la planificación de los servicios por parte de los directores de unidades de información.

También es posible determinar en qué áreas de la colección los usuarios tienen más problemas con los servicios que recibe, cuáles son los patrones de uso que los usuarios aplican para el uso de la colección, así como las áreas y temas de interés que ellos tienen, con el fin de fortalecer aquellas áreas donde los usuarios presentan mayor demanda de servicios.

Además, se pueden determinar patrones de uso en el área de circulación a través del análisis de los registros transaccionales del área de circulación, como préstamo, devolución, renovación, para determinar el tiempo pro-

medio de respuesta a los usuarios. Con esta información se puede asignar más personal a determinados turnos, o brindar capacitación a funcionarios, para mejorar sus deficiencias y para incrementar el nivel de servicio.

En el área de adquisiciones es importante conocer qué materiales se utilizan con mayor frecuencia, qué materiales tienen un número de ejemplares no acorde con la demanda y qué materiales tiene poco uso, lo que es importante para futuras asignaciones presupuestarias para realizar posibles compras de más ejemplares, o por el contrario desechar cualquier intento de adquisición de material con poco uso; es decir, buscar correlaciones entre material poco utilizado, usuario y proveedores, así como la racionalidad del costo con que se adquieren estos materiales. En esta área se pueden hacer análisis sobre documentos de poco uso, buscar correlaciones de estos materiales de poco uso con los proveedores respectivos, así se pueden revisar los niveles de precios que estos proveedores tienen para esos materiales y, por tanto, poder realizar mejores negociaciones de precio en la adquisición de material.

En el área de referencia se pueden evaluar las preguntas que hacen los usuarios, cuáles son las respuestas del personal, el tiempo de respuesta, la ruta seguida para atender las consultas, con la finalidad de buscar formas más sistemáticas y exitosas en la atención de las consultas de los usuarios.

También es importante evaluar la frecuencia de uso y el grado de satisfacción que presentan los usuarios por cada uno de los servicios que ofrece la biblioteca, desde que el usuario solicita el servicio hasta que se le ofrece, si este fue exitoso o infructuoso; además, conocer la frecuencia con que se demanda el servicio por la comunidad de usuarios.

5. Necesidad de un *datawarehouse* para aplicar la bibliominería en una biblioteca

Un DW es un deposito histórico de la información producida por los sistemas transaccionales de la organización, donde las operaciones que se aplican sobre él son de lectura de la información almacenada para alimentar los sistemas de inteligencia de negocios.

Así, el DW es el componente central de la inteligencia de negocios y se entiende que la bibliominería de datos es un campo especializado de la inteligencia de negocios, cuyo objetivo general es proporcionar la información necesaria para gestionar el negocio, brindando la información necesaria para la toma de decisiones. El objetivo principal es responder a preguntas importantes que faciliten a la organización el alcance de sus objetivos estratégicos.

Lo primero que se debe tener como norte cuando se analiza una organización para proponer una solución de DW es la compresión del usuario hacia la propuesta de solución que se le presenta, la cual debe estar en un lenguaje del negocio, que sea intuitiva y fácil de utilizar, descartando lenguaje técnico propio del departamento de tecnologías de la información, ya que con el uso de lenguaje técnico se propicia la obstrucción de la comunicación con los usuarios finales.

El DW es una base de datos que soporta las necesidades de información del usuario para la toma de decisiones, y se mantiene por separado de las bases de datos operativas de la organización, que son producto de los sistemas transaccionales que soportan las operaciones básicas de la unidad de información.

Es una forma de ganar ventajas competitivas en el mercado a través del conocimiento. Entonces aquella organización que logre obtener ventajas competitivas basadas en el conocimiento almacenado en un DW obtiene una posición más sólida, por lo que para sus competidores será más difícil alcanzarla o superarla. Esta posición estratégica de los DW justifica aún más la importancia y necesidad de su tenencia para una organización.

Una vez que una organización posea un DW que se utilice como base para soportar la toma de decisiones estratégicas, estas van a ser más acertadas, ya que con el DW se puede conocer más a fondo el negocio.

Es lógico que se deberá realizar un análisis de costos/beneficios que justifique la creación de un DW, en donde se tiene que el retorno de la inversión es muy alto para su desarrollo, de ahí que se justifique la creciente demanda de las organizaciones para el desarrollo de un DW para soportar la toma de decisiones.

Rev. Interam. Bibliot. Medellín (Colombia) Vol. 43, número 2/mayo-agosto 2020 eI8 ISSN 0120-0976 / ISSN (en línea) 2538-9866
https://doi.org/10.17533/udea.rib.v43n2eI8

Por tanto, con la tenencia de un DW se podrán aplicar técnicas de bibliominería de datos que faciliten la gestión de la unidad de información a través de una mejor toma de decisiones, justificada por el análisis de los datos ocultos en los grandes volúmenes de datos almacenados en él, que permitan mejorar el servicio brindado a los usuarios.

6. Método

6.1. Criterios técnicos para el diseño del *datawarehouse*

Es importante destacar que cuando se está diseñando el DW no se debe pensar en productos o marcas, ya que esto sería un completo error, por cuanto las soluciones de inteligencia de negocios son para el negocio y no para el departamento de tecnologías de información, en donde frecuentemente sí se piensa en productos y marcas. Por tanto, se debe obviar el entorno tecnológico y centrarse en obtener datos sobre el negocio y cómo opera, cuáles sus necesidades, cuáles son sus expectativas, a mediano y largo plazo, para poder entender hacia dónde se dirige el negocio. Por tal razón, el diseñador del DW debe estar estrechamente ligado con el negocio y no con el departamento de tecnologías de la información, pues, cuando se diseña un DW, todo debe ser expuesto en lenguaje del negocio y nunca en lenguaje técnico.

El objetivo de la bibliominería de datos es poder responder preguntas importantes que faciliten a la organización el cumplimiento de un objetivo estratégico y que faciliten también el cumplimiento de la propuesta de valor a sus usuarios. El DW debe estar orientado hacia un tema estratégico e integrado para el soporte de la toma de decisiones, en el que primero e realice un diseño en papel, con un lenguaje de negocios que responda a las interrogantes estratégicas que permitan facilitar la toma de decisiones.

Para el diseño de un DW, el modelo estrella (véase Figura 1) es el más utilizado. De acuerdo con Chinchilla (2011), el modelo estrella se puede explicar de la siguiente manera:

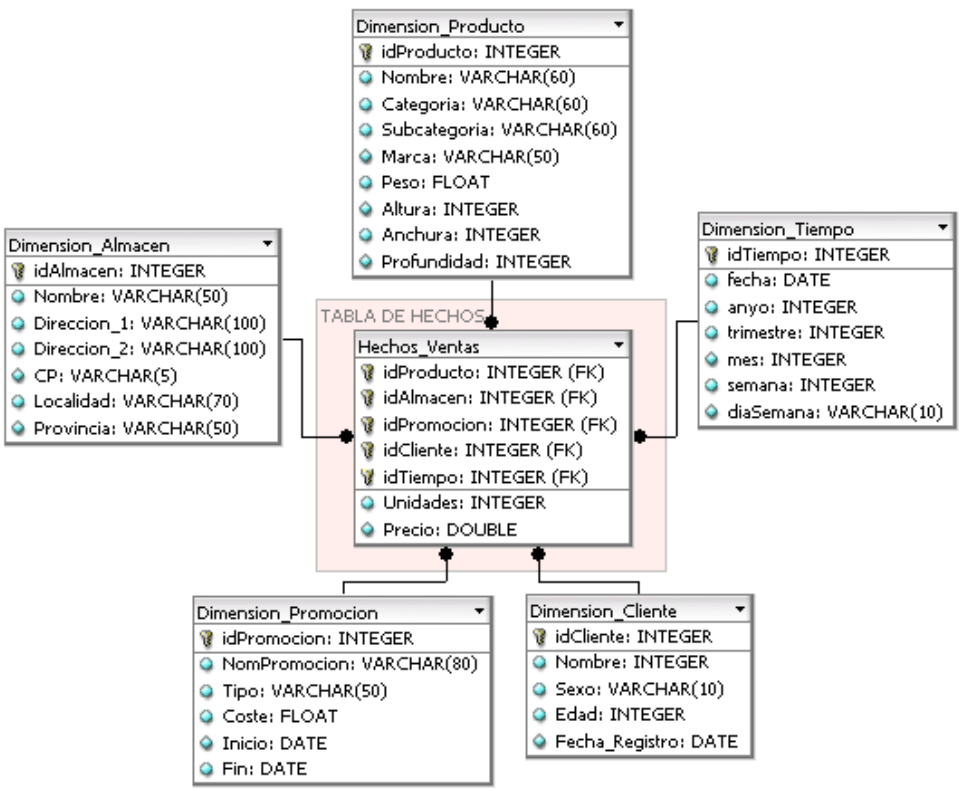

Figura 1. Ejemplo de modelo estrella.
Fuente: figura con carácter ilustrativo, tomado de https://es.wikipedia.org/wiki/Esquema_en_estrella

> Consiste en una tabla central construida con llaves foráneas provenientes de un conjunto de tablas llamadas dimensiones. Las llaves foráneas en conjunto corresponden a la llave primaria de dicha tabla central, que, conforme a la terminología de base de datos multidimensional también es llamada tabla de hechos. Según Cortéz (1999), cada una de las tablas de dimensión tiene una sola llave primaria que corresponde exactamente a uno de los componentes de la llave primaria compuesta en la tabla de hechos.
>
> La tabla de hechos contiene, además de las llaves foráneas, una o más medidas (hechos numéricos) que ocurren como resultado de combinar las dimensiones que definen cada registro. Las tablas de dimensiones contienen información descriptiva. (p. 60)

Además, consta de una tabla central, llamada la tabla de hechos, en la que se almacenan valores cuantitativos que indican la magnitud del hecho que se está midiendo; alrededor de esta tabla de hechos se encuentran las tablas de dimensiones, que contienen los aspectos que se quieren medir en la tabla de hechos. El modelo estrella es un modelo multidimensional, donde cada dimensión es una perspectiva diferente de lo que se quiere analizar a través de su asociación con la tabla de hechos que alberga los valores cuantitativos para cada una de la dimensiones establecidas.

Este modelo representa los datos del proceso de negocio en hechos y dimensiones. La tabla de hechos estará rodeada de las tablas de dimensiones, que contienen los metadatos sobre los que se analiza un área determinada del negocio sobre los hechos que se encuentran almacenados en la tabla central.

De esta forma, lo más importante es saber qué quiere analizar la gerencia, y con base en esto se construye el modelo para responder a esas necesidades. Se parte de lo que se quiere saber, luego se sigue la determinación de las fuentes, donde el problema es determinar de dónde se obtiene esa información. Posteriormente, cuando se entienda que la información se encuentra distribuida en varios sistemas transaccionales de la unidad de información —como contabilidad, adquisiciones, sistema de préstamos, el CRM, y diferentes bases de datos que pueden ser internas o externas—, toda esta información proveniente de los sistemas operativos se deberá consolidar en el DW.

En resumen, en el modelo de datos, la tabla de hechos se relaciona con las tablas de dimensiones de la siguiente forma: las tablas de dimensiones contienen una clave primaria, y en la tabla de hechos la clave principal estará compuesta por las claves primarias de las tablas de dimensiones. Entonces las tablas de dimensiones y de hechos se asocian a través de identificadores existentes en ambas tablas. Por tanto, cada dimensión tiene una clave primaria única que se enlaza a través de integridad referencial a la tabla de hechos. Se debe asegurar que en la tabla de hechos se puedan realizar los cruces deseados de las dimensiones, y que estos cruces produzcan un valor correcto y válido. Esta actividad se realiza en el análisis de la información que se debe realizar antes de cargar el DW, que se verá en el próximo apartado de este artículo.

Cuando diseñamos un DW, a diferencia de una base de datos relacional, aplicamos el principio de desnormalización, contrario al de normalización comúnmente utilizado en el diseño de bases de datos relacionales, ya que lo más importante en el DW es la velocidad en el tiempo de respuesta a las diferentes consultas que se realicen, por tanto, la introducción de redundancia controlada para acelerar el tiempo de respuesta en la tabla de hechos de un DW es una situación completamente permitida y deseada en su diseño. Chinchilla (2011) se refiere a este tópico con la siguiente afirmación:

> La utilización de modelos estrella, el manejo de redundancia y la falta de normalización muchas veces chocan con los esquemas de desarrollo de bases de datos y requieren de tiempo para ser asimilados. Pero para el manejo de bases de datos con información masiva para la toma de decisiones no es posible con los sistemas convencionales de administración transaccional. (p. 65)

Para realizar el diseño de un DW para una organización, se recomienda estudiar la publicación del Chinchilla (2011), donde explica claramente cómo realizar el diseño de un mercado de datos *(datamart)* en una organización, actividad que hace uso de la misma metodología que se requiere para diseñar un DW, con la única diferencia de que un mercado de datos se utiliza para una parte de la organización, que puede ser un departamento o una sección, mientras que un DW se construye para toda la organización en forma completa.

Rev. Interam. Bibliot. Medellín (Colombia) Vol. 43, número 2/mayo-agosto 2020 eI8 ISSN 0120-0976 / ISSN (en línea) 2538-9866
https://doi.org/10.17533/udea.rib.v43n2eI8

7. Resultados

7.1. Implantación del *datawarehouse*

Para el proceso de implantación del DW, se debe planear y diseñar cuidadosamente la extracción de datos de las diferentes fuentes, así como el proceso de limpieza, transformación y carga en el DW. Los datos se extraen de las bases de datos operacionales y se enriquecen con las bases de datos externas. Para esto se diseña el proceso de extraer, transformar y cargar (ETL, por sus siglas en inglés), que es el proceso que organiza el flujo de datos entre los diferentes sistemas en una organización.

De esta forma, este proceso llamado ETL se puede resumir en las siguientes actividades.

- Se deben extraer los datos desde los sistemas origen, que frecuentemente son los sistemas transaccionales que se utilizan en la unidad de información para automatizar las labores operativas de la unidad.
- Se hace un análisis de los datos extraídos para verificar que estos cumplen con las normas establecidas en la unidad.
- Se transforman los datos, aplicando una serie de reglas del negocio o funciones sobre los datos extraídos para convertirlos a un formato determinado y deseado por la unidad.
- Finalmente los datos serán cargados al DW.

Por esta razón, la información en el DW debe estar consolidada, limpia y verificada, con relaciones validadas dentro del DW, en donde debe diseñarse cuidadosamente la adquisición y consolidación de datos de las diferentes fuentes, así como la ejecución de los procesos de limpieza y validación requeridos.

Es importante recordar que un DW es para toda la organización, mientras que un datamart es para un departamento o sección de la organización y tener siempre presente que tanto el DW como el *datamart* se diseñan conforme a las necesidades del negocio, ya sea para mejorar la rentabilidad, reducir costos, mejorar la satisfacción del cliente o cualquier tema estratégico que desee analizar la gerencia.

El DW es una base de datos que soporta las decisiones y se mantiene por separado de las bases de datos operativas de la organización, las cuales permiten la lectura y actualización de la información. Mientras que el *datamart* es una base de datos que permite sola la lectura de la información, de los diferentes procesos de consulta orientados a contestar diferentes inquietudes estratégicas.

Por tanto, una vez implantado el DW en una unidad de información, es posible aplicar técnicas de bibliominería sobre él para enriquecer nuestras posibilidades de obtener información para hacer más efectiva la gestión de la unidad de información a través de una mejor toma de decisiones.

7.2. Usos actuales de la bibliominería

- Descubrimiento de patrones y tendencias de los datos.
- Árboles de decisiones.
- Conglomeración k-medias y conglomeración jerárquica.
- Detección de valores extremos.
- Análisis de series de tiempo y predicción.
- Conglomeración de series de tiempo y clasificación.
- Reglas de asociación.
- Minería de texto.
- Análisis de redes sociales.
- Escalamiento multidimensional.

7.3. Herramientas para el uso de la bibliominería

Como se mencionó al principio de este artículo, las bibliotecas generalmente tienen poco presupuesto para soportar proyectos de automatización, por tanto, serán descritas cinco herramientas de software libre que pueden ser utilizadas para estos proyectos de minería de datos en las unidades de información.

Algunos de los paquetes de software que se podrían utilizar son Orange, Weka, JHepWork, Knime y RapidMiner.

A continuación se explican las principales características de cada uno de ellos, con la aclaración de que cualquiera que escoja el usuario para soportar sus procesos de minería de datos deberá ser objeto de una investigación más detallada.

7.3.1. Orange

Aplicación de software libre para minería de datos y análisis predictivo, bajo licencia GPL. Es una herramienta poderosa, pero a la vez es amigable e intuitiva y permite una programación visual, rápida y versátil para un análisis exploratorio de datos. Desarrollado en C++. La aplicación Orange es multiplataforma, ya que es soportada por Windows, Linux y Mac. Esta permite el modelado predictivo a través de árboles de clasificación, regresión, logística, clasificación de Bayes y reglas de asociación. Además de métodos de descripción de datos, mapas autoorganizados y *clustering* (agrupamiento de datos).

Por último, tiene técnicas de validación de modelos y validaciones cruzadas. Además dispone de un componente de programación visual fácil y potente, rápido y versátil, para el análisis exploratorio de datos y su visualización.

7.3.2. Weka

Es una aplicación de software libre que también soporta minería de datos y análisis predictivo, asi como preprocesamiento, *clustering*, clasificación, regresión, visualización y características de selección. Es una de las herramientas para aplicación de tareas de *data mining* más reconocidas. Se encuentra bajo licencia GNU-GPL. Esta escrita en lenguaje Java, lo que la hace muy portable. Weka permite procesos previos, *clustering*, clasificación, regresiones, visualización y selección de propiedades.

Debido a que su interfaz es el *browser* Explorer, la aplicación Weka es de fácil uso. Esta es una aplicación multiplataforma que corre en Windows, Linux o Mac. Además proporciona acceso a bases de datos vía SQL.

7.3.3. JHepWork

También es una aplicación de software libre para análisis de datos y visualización. Contiene librerías científicas en Java para funciones matemáticas, y algoritmos de minería de datos. Se encuentra bajo licencia GPL. Es una aplicación que permite el análisis de grandes volúmenes de datos y análisis estadístico. Es un software multiplataforma que corre en Windows, Mac y Linux. JHepWork es de fácil uso, ya que utiliza una interfaz de usuario comprensible y amigable.

Esta herramienta es un poco más avanzada y se requiere más alto conocimiento, el lenguaje usado es Jython (Java + Pythom), aunque también funciona a la perfección en Java. También puede ser usada para llamar librerías JHepWork numéricas y gráficas.

7.3.4. Knime (Konstanz Information Miner)

Aplicación de software libre de fácil uso y compresión para integración de datos, procesamiento, análisis y exploración de datos. Se distribuye bajo licencia GPL. Ofrece a los usuarios la capacidad de crear de forma visual flujos o tuberías de datos, ejecutar selectivamente algunos o todos los pasos de análisis, y luego estudiar los resultados, modelos y vistas interactivas. Es un software de integración de datos amigable, intuitivo y fácil de usar, que permite el procesamiento, análisis y exploración de datos, desde su plataforma.

Es una aplicación multiplataforma que corre en ambiente Windows, Linux o Mac. Es una herramienta de fácil uso y compresión. Knime está escrita en Java, está basada en Eclipse y hace uso de sus métodos de extensión para soportar diferentes *plugins*, proporcionando así una funcionalidad adicional.

7.3.5. RapidMiner

Es una aplicación de software libre que se utiliza para minería de datos y análisis predictivo. Produce sus resultados en archivos XML y cuentan con la interfaz gráfica del mismo programa. Se distribuye bajo licencia AGPL. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico.

Rev. Interam. Bibliot. Medellín (Colombia) Vol. 43, número 2/mayo-agosto 2020 eI8 ISSN 0120-0976 / ISSN (en línea) 2538-9866
https://doi.org/10.17533/udea.rib.v43n2eI8

Facilita crear modelos predictivos. Es de fácil uso a través de su interfaz gráfica para el usuario. Es una aplicación multiplataforma que corre en ambiente Windows, Linux o Mac. Además permite utilizar los algoritmos incluidos en el software Weka.

Finalmente, es importante mencionar que debido al gran auge de aplicaciones nuevas cada día, en todas las áreas del saber humano, no se pretende ni es el objetivo de este artículo hacer una lista exhaustiva de todas las posibles aplicaciones en software libre disponibles en la web para minería de datos, de ahí la importancia que el lector interesado realice una investigación periódica de nuevas herramientas de software libre para el uso de bibliominería en su unidad de información.

7.3.6. R+/ RStudio

En R+ existe una interfaz gráfica para el manejo de bibliominería llamada Rattle.[1] Rattle es un GUI popular para el manejo de minería de datos a través de R+ y presenta estadística y visualmente resúmenes de datos, datos transformados y también se pueden ajustar algunos modelos.

8. Consideraciones finales

Aunque las bibliotecas en general tienen bajos presupuestos para realizar proyectos de automatización, los beneficios que producirá a los usuarios finales un proyecto de bibliominería en una unidad de información superan con creces los costos, por lo que el rápido retorno de la inversión de estos proyectos justifica ampliamente su desarrollo.

Para realizar un proyecto de bibliominería, se deberá disponer de los sistemas transaccionales origen, que automatizan las operaciones del día a día en la unidad de información, los cuales estarán soportados sobre bases de datos relacionales, que serán las fuentes primarias de información. Posteriormente, estas bases de datos relacionales de los sistemas origen serán sometidas a procesos ETL en los DW, los cuales serán la base para realizar los procesos de bibliominería. Aunque existe software privativo para la implementación de proyectos de biblominería, también existe una buena gama de aplicaciones de software libre para este propósito.

1 Véase https://rattle.togaware.com/

En este punto, debe quedar claro que el software libre no elimina el costo humano de la implementación del proyecto, pero sí tiene un alto impacto en la reducción de costos por no tener que pagar licencias onerosas.

Con el uso de computación en la nube, también se pueden reducir los costos de inversión en infraestructura tecnológica, a través del uso de servidores en la nube, los cuales tienen un costo dinámico representado por el uso de almacenamiento y la velocidad que se requiera, eliminando así el costo de grandes servidores que pasarán ociosos gran parte del tiempo en su vida útil; aunque conlleva un importante requisito como lo es la tenencia de un buen ancho de banda de acceso a internet, para el acceso a los servidores que darán soporte a los procesos de bibliominería en la unidad de información. Tal como lo señala (Fernández, 2012b).

> La computación en la nube está cambiando el modelo tecnológico actual. La obligación de mantener equipos de cómputo con configuraciones sofisticadas de gran capacidad y de alto valor da paso a la simple tenencia de estaciones de trabajo de bajo costo que acceden a Internet a través de banda ancha y a potentes servidores virtuales. Esto tendrá un impacto positivo en la masificación del uso de Internet y en la virtualización del almacenamiento y el poder de las computadoras. (p. 2)

Un proyecto de bibliominería es de carácter interdisciplinario y requiere el concurso de especialistas en el área de bibliotecología, en el área de cómputo y en el área de estadística. En realidad, es un proyecto complejo que producirá resultados de análisis de datos que serán una plataforma sólida para apoyo en la toma de decisiones.

9. Referencias

1. Aluja, Tomás (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Qüestiió*, *25*(3),479-498.
2. Arriola, Oscar; Butrón, Katya (2008). Sistemas integrales para la automatización de bibliotecas basados en software libre. *ACIMED*, *18*(6).
3. Chinchilla, Ricardo (2011). Mercado de datos: conceptos y metodologías de desarrollo. *Tecnología en Marcha*, *24*(3), 55-66.

4. De la Puente, Marcelo (2010). Bibliominería y minería de datos. *Consultora de Ciencias de la Información*, *15*. http://www.slideshare.net/pattsul/014

5. Fernández Mynor (2012a). Gestión estratégica y la automatización de las unidades de información. *Revista de la Escuela de Bibliotecología, Documentación e Información*, *2*(1).

6. Fernández, Mynor (2012b). Computación en la nube para automatizar unidades de Información. *Bibliotecas*, *30*(1).

7. Nicholson, Scott (2003). The bibliomining process: Data warehousing and data mining for library decision-making. *Information Technology and Libraries*, *22*(4).

8. Pal, Jiban (2011). Usefulness and applications of data mining in extracting information from different perspectives. *Annals of Library and Information Studies*, *58*(1), 7 -16.

9. Poe, Vidette (1998). *Building a data warehousefor decision support*. New Jersey: Prentice-Hall.

Rev. Interam. Bibliot. Medellín (Colombia) Vol. 43, número 2/mayo-agosto 2020 eI8 ISSN 0120-0976 / ISSN (en línea) 2538-9866
https://doi.org/10.17533/udea.rib.v43n2eI8