CONSULTORA DE CIENCIAS DE LA INFORMACIÓN
BUENOS AIRES
ARGENTINA

Serie

DOCUMENTOS DE TRABAJO

Área: Tecnología

Gestión del conocimiento y Minería de datos

Marcelo de la Puente

Agosto 2010
N°019

ISSN 1852 - 6411

Editor: Patricia Allendez Sullivan. Asistente Editorial: Mariana Sabugueiro

Puente, Marcelo de la

Gestión del conocimiento y Minería de datos. Buenos Aires: Consultora de Ciencias de la Información, 2010.

ISSN 1852 - 6411

1. Gestión del Conocimiento 2. Minería de Datos. 3. Conocimiento. 4. Información.
I. Título

Resumen:

La Gestión del Conocimiento se refiere al conjunto de procesos desarrollados en una organización para crear, organizar, almacenar y transferir el conocimiento. La Minería de Datos es la disciplina que tiene por objetivo la extracción de conocimiento implícito en grandes bases de datos. La Minería de Datos tiene un papel fundamental en el proceso de convertir en explicito al conocimiento implícito y en las distintas etapas del proceso de Gestión del Conocimiento en las organizaciones.

Introducción

La **Gestión del Conocimiento** es un concepto actualmente altamente aplicado en las organizaciones[1] que procura transferir el conocimiento y la experiencia de sus recursos humanos, de forma tal que quede disponible y pueda ser utilizado por otros miembros de dicha organización.

Aunque las aportaciones teóricas que han configurado este campo pueden rastrearse hasta unos cincuenta años atrás, su implementación en el ámbito empresarial ha tenido lugar en el período que abarca desde el año 1985 hasta el momento actual. Usualmente el proceso implica una variedad de distintas técnicas para capturar, organizar y almacenar el conocimiento del personal de la organización para transformarlo en un activo intelectual que brinde beneficios y se pueda compartir. En la actualidad las tecnologías de la información permiten contar con herramientas que apoyan este proceso, es decir, la recolección, la transferencia y la administración sistemática de la información, junto con los sistemas diseñados para dar un mejor uso a dicho conocimiento.

Todo esto implica que la información suministrada por este conjunto de herramientas les sirva a los individuos clave de la organización para apoyar la

[1] Entendemos por organización a las Bibliotecas, Archivos y Centros de Documentación, así como a toda otra forma de denominar los centros de almacenamiento y recuperación de información.

toma de decisiones y reducir el riesgo vinculado a las mismas. En el ámbito de las nuevas tecnologías de la información y de actividades de consultoría relacionadas al tema de la inteligencia competitiva de las organizaciones, la Gestión del conocimiento cobra una importancia vital. La administración del conocimiento se ha convertido en un asunto primordial en las empresas ya que se han percatado de que una gran parte de su valor como entidades que brindan un servicio de valor agregado al usuario, depende de la capacidad de las mismas para crear y administrar el conocimiento. Existen estudios que han determinado que una parte importante del valor de una organización, se relaciona con sus activos intangibles, de los cuales el conocimiento es un activo fundamental

La gestión del conocimiento, entonces, se refiere al conjunto de procesos de negocios desarrollados en una organización para crear, almacenar, transferir y aplicar el conocimiento; incluye una variedad de técnicas en sus distintas fases entre las que se encuentra la **Minería de Datos o Data Mining**.
En este trabajo exploraremos la relación entre la Gestión del Conocimiento y la Minería de Datos y como esta última contribuye al proceso total de la dicha gestión.

La Minería de Datos y las etapas de la Gestión del Conocimiento

La **Minería de datos** también denominada extracción de datos es la práctica por medios automáticos o semiautomáticos de la búsqueda y la exploración en grandes almacenes de datos de relaciones no visualizadas previamente, dando por resultado el descubrimiento de patrones significativos entre los mismos y reglas. Para lograr este propósito la Minería de Datos emplea técnicas estadísticas, de automatización del conocimiento y de reconocimiento de patrones (observar datos de una sola fuente, recursos de información, etc.)

Para ver la contribución de la Minería de Datos al ámbito de la Gestión del Conocimiento debe analizarse previamente, el ciclo completo de la Gestión del Conocimiento en las organizaciones.

Entre las diferentes fases de la Gestión del Conocimiento, encontramos:

Descubrimiento, captura y creación del conocimiento: en esta fase el conocimiento tácito o implícito de la organización se captura a través de una variedad de técnicas entre ellas la **Minería de Datos**, también por medio de la colaboración, directorios expertos o sistemas inteligentes que usan patrones o estrategias de búsquedas particulares, etc.

Clasificación y representación: En esta fase se utilizan herramientas para organizar la información, como los metadatos, taxonomías, ontologías, mapas temáticos para la visualización gráfica de la información, *clustering*, una técnica de agrupamiento de la información en *clusters* o categorías por medio de un algoritmo que clasifica los términos en función de la distancia, de su cercanía tomando vectores de entrada con características comunes, comúnmente usada por la **Minería de datos** ; indización automática, etc.

Recuperación de la información: en esta fase intervienen motores de búsqueda, agentes inteligentes, la competencia del *browsing* o navegación general versus búsqueda específica, arquitectura del conocimiento y de la información y **Minería de datos**

Diseminación de la información: comunicación, publicación, incluyendo Internet versus Intranet, etc.

Puede observarse claramente, que la Minería de datos interviene tanto en la parte de entrada o *input*, como en el eslabón intermedio del ciclo, la organización de la información y finalmente en la salida o en el *output* del mismo, la recuperación de la información.

A continuación se analizará más detalladamente, la función de la Minería de datos en cada una de las etapas de la Gestión del Conocimiento.

Descubrimiento y captura del conocimiento

La organización adquiere conocimiento de una gran variedad de formas, según el tipo de conocimiento que se esté buscando. Los primeros sistemas de gestión del conocimiento buscaban construir almacenes de conocimientos, informes, presentaciones y mejores prácticas. Estos primeros esfuerzos se han ampliado para incluir documentos no estructurados, como el correo electrónico. En otros casos, las organizaciones adquieren más conocimientos al desarrollar redes de expertos en línea con el propósito de que en la empresa los empleados puedan encontrar al experto que posee el saber específico que buscan.

El primer paso que hay que tomar para descubrir el conocimiento, es conocer cuáles pueden ser las fuentes potenciales del mismo, que pueden diferir de una organización a otra, en general, se puede enunciar la siguiente lista: E-Mail, Internet, conocimiento público, clientes, competidores, suscripciones, cadena de suministros, adquisición de contenidos, noticias, conocimiento interno, bases de datos, agencias de información, etc.

En la Gestión del conocimiento existe una jerarquía de la información, que comienza con los **datos**, que consisten en un conjunto discreto de elementos objetivos acerca de distintos eventos, que pueden derivar de hechos, información, estadísticas o similares, tanto históricos como derivados del cálculo o de la experimentación. En el contexto de las empresas los datos pueden ser descriptos como registros estructurados de transacciones u operaciones comerciales propias de la organización. Los datos en sí mismos, carecen de sentido, ya que solo describen lo que sucede de manera parcial y no proporcionan juicio o interpretación, ni favorecen la toma de decisiones.

En el siguiente nivel, la **información,** consiste en datos dotados de relevancia y de propósito. Consiste en conocimiento comunicado o recibido concerniente a

hechos o circunstancias particulares, es cualquier conocimiento adquirido mediante comunicación, aprendizaje o instrucción. La información se construyen a través de datos a los que en cierto sentido, se le ha aplicado un control de calidad previo, se los ha contextual izado. Para la Teoría de la Comunicación de Shannon que describe la información como *mensaje*, normalmente en forma de un documento en un soporte dado, impreso, audiovisual, etc., es el receptor de la comunicación el que le da la categoría de información al mensaje. Las mediciones cuantitativas de la gestión de la información tienden a incluir medidas como conectividad y tipo de transacciones, por ejemplo, cuantos tipos de mensaje se intercambian por un canal particular, por ejemplo, el E-mail, pero todas estas mediciones carecen de sentido a menos que aporten nuevas perspectivas o aporten a la solución de un problema.

Los datos se convierten en **información** cuando se les añade un sentido mediante la contextualización de los mismos, es decir, conocer para qué propósito han sido recolectados, su categorización, es decir, se conocen las unidades de análisis o los componentes clave de los datos, para puedan ser analizados matemática o estadísticamente. Se debe efectuar un proceso de corrección de los mismos para eliminar los datos erróneos y finalmente, se debe efectuar una condensación de los datos, para que la información contenida en los mismos sea resumida, es decir, que sean concisos.

En el tercer nivel, el **conocimiento**, consiste en una aprensión de hechos, verdades o principios como resultado del estudio, investigación o de la erudición en general. Implica familiaridad con un tema particular o rama del saber. El conocimiento deriva de la información, como ésta deriva de los datos, para transformarse en conocimiento, la información debe pasar por un proceso comparativo, en el que se analiza cómo se ajusta la información a una situación dada, comparándola con otras situaciones ya conocidas, se deben tomar en cuenta, además, las consecuencias o implicaciones que tiene la información para la toma de decisiones en la organización. Se debe considerar cómo se relaciona ese fragmento de conocimiento con otros y finalmente se tiene que tomar conciencia de la opinión de otras personas de la organización sobre este

fragmento de conocimiento particular, ya que el conocimiento se obtiene de los individuos y se transmite por medios estructurados de comunicación, documentos, libros, redes digitales, o en contactos persona a persona, mediante conversaciones y aprendizaje. Las tecnologías de la información deben actuar como facilitadores para la comunicación del conocimiento a lo largo de una organización, ya que sea esta esté centralizada en un solo lugar geográfico o distribuida en muchas localidades.

En un último nivel de la pirámide se encuentra la **sabiduría o inteligencia**, saber definido o entendido como la capacidad de comprender los principios, como contraposición al conocimiento que solo comprende patrones y de la información que analiza relaciones y cuya acumulación puede dar lugar al capital intelectual de una organización.

En la gestión del conocimiento, generalmente se consideran dos categorías básicas de conocimiento, el **conocimiento tácito** y el **conocimiento explícito**

El **conocimiento tácito** es el conocimiento personal o implícito, almacenado en los cerebros del personal de la organización, difícil de formalizar, registrar y articular. Se desarrolla por medio de un proceso de prueba y error y va conformando el conocimiento de un individuo sobre diferentes temas.

El **conocimiento explícito** es el conocimiento almacenado en distintos soportes o medios físicos. La interacción entre el conocimiento tácito y el explícito es lo que da lugar a procesos de generación de nuevo conocimiento

La transmisión del conocimiento tácito no resulta fácil y para que pueda ser rentabilizado es necesario sustraerlo del contexto de origen y formalizarlo, con lo que se genera un "ciclo de conversión" que se describe en tres procesos:

1) **De tácito a tácito**: El paso de conocimiento de tácito a tácito se produce a través de procesos de socialización, es decir, a través de la adquisición de conocimientos e información mediante la interacción directa con el mundo exterior: con otras personas, con otras culturas, etc.

2) **De tácito a explícito**: Se produce a través de la externalización, que podríamos definir como el proceso de expresar algo, el diálogo. Externalizar es convertir imágenes y/o palabras a través del diálogo.

3) **De explícito a explícito**: Este paso se denomina combinación. Como su propio nombre indica, se combinan diferentes formas de conocimiento explícito mediante documentos o bases de datos.

A la hora de utilizar técnicas de minería de datos en un determinado proyecto, el proceso que se está llevando a cabo es una "extracción no trivial de información implícita, previamente desconocida a partir de los datos", a nivel del conocimiento explicito, con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles, que se debe interiorizar, para posteriormente externalizarlo en la toma de decisiones.

La Minería de Datos utiliza una gran variedad de técnicas, pero se puede formular un esquema general que siguen todas:

1. Selección y procesamiento de los datos: generalmente los datos disponibles en una base de datos no se encuentran en la forma más adecuada para el tratamiento de los mismos por medio de algoritmos que utiliza la Minería de Datos, por lo que es necesario realizar una

operación de filtrado de valores incorrectos o efectuar un muestreo, es decir extraer una muestra de la población total que se considere representativa de la misma, para poder trabajar con un número más reducido de datos. El filtrado que se realiza sobre los datos puede efectuarse de diferentes maneras, puede ser aleatorio, puede aplicarse la citada técnica de *clustering*, para formar *clusters* o grupos de datos con características homogéneas o similares en base a documentos clasificados anteriormente y a la inferencia de reglas de agrupamiento.

2. Selección de características a estudiar: una vez que ya está determinada la población de estudio, si se considera que el conjunto de datos con el que se trabaja es muy amplio, se realiza una selección de las características a observar, las variables del problema, con el propósito de simplificar el procesamiento de los datos y de agilizarlo. Se puede utilizar una gran variedad de técnicas estadísticas o métodos gráficos que permitan visualizar las relaciones existentes entre los datos. La visualización de datos es una técnica adecuada para localizar patrones en un conjunto de datos. Puede ser utilizada al comienzo del proceso de Minería de Datos para tener obtener una referencia de la calidad del conjunto de los datos
3. Uso de algoritmo de extracción de conocimiento: se utiliza la técnica de Minería de datos elegida previamente, con el objetivo de elaborar un modelo de conocimiento con los patrones de comportamiento y de asociación entre las distintas variables de estudio, algoritmos que encuentran reglas simples de clasificación de un conjunto de instancias o ejemplos, de valores perdidos y atributos numéricos, modelos estadísticos, árboles de decisión, etc. Los árboles de decisión, por ejemplo, son estructuras con forma arborescente que representan conjuntos de decisiones. Estas decisiones, a su vez, generan reglas para la clasificación del conjunto de datos, lo cual puede llevar a predicciones basadas en esta clasificación efectuada. En el árbol se comparan los atributos en cada nodo o punto del mismo, para decidir si un ítem pertenece a una clase o a otra. A cada nodo del árbol se encuentra asociada la cantidad de información necesaria para decidir si

una instancia debe ser clasificada con un sí o con no, es decir, si satisface o no la condición especificada en cada nodo.

4. El análisis predictivo utiliza técnicas de Minería de Datos, datos históricos y suposiciones sobre condiciones futuras, para predecir el comportamiento de eventos determinados, por ejemplo, el comportamiento de clientes en base al lanzamiento de un nuevo servicio o producto.
5. Interpretación y evaluación de los resultados: se verifica que los resultados obtenidos sean consistentes y se compara con los obtenidos mediante los métodos estadísticos y gráficos. Se determina finalmente si aportan alguna novedad en las relaciones ocultas entre los datos que permitan guiar toma de decisiones. Si existe una discrepancia se deben tomar en cuenta los que se consideren más confiables o reiniciar todo el proceso.

Una de las principales funciones que la Minería de Datos le brinda a la Gestión del Conocimiento es precisamente, la capacidad de hacer explícito este conocimiento que permanece como tácito o implícito en los recursos humanos de las organizaciones al codificarlo en bases de datos, lo cual resulta de vital importancia cuando el conocimiento codificado o explícito es insuficiente para enfrentar una situación determinada.

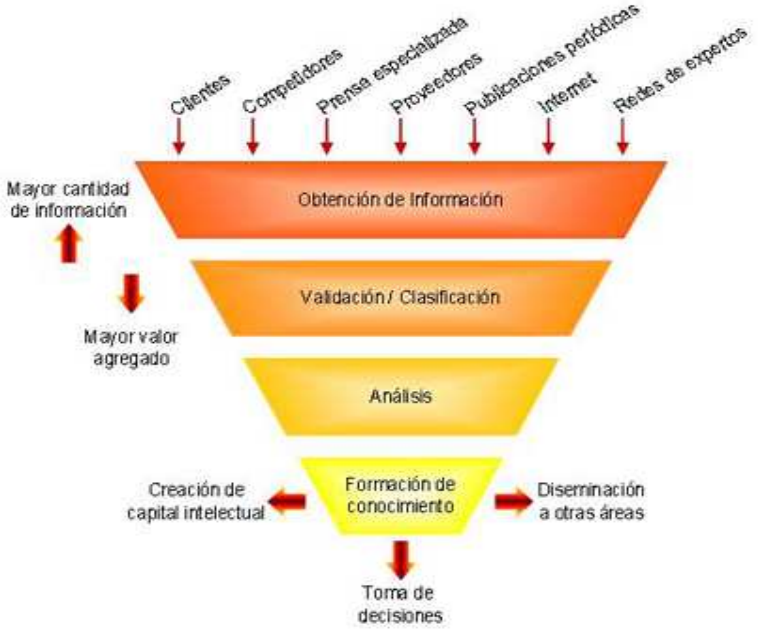
Descubrimiento

La fase de descubrimiento y captura del conocimiento tienen límites difíciles de precisar y comparten elementos comunes. Lo primero que debe hacerse para descubrir nuevo conocimiento es saber cuáles son las fuentes de información potenciales con las que cuenta la organización, estás difieren en distintos casos, pero en general pueden dividirse en:

- Competidores

- Clientes
- Internet
- Proveedores
- Socios
- Bases de datos
- Agencias de información
- Publicaciones periódicas

A estas fuentes externas se les debe sumar el conocimiento interno de la organización. Toda esta variedad de fuentes brindan una gran variedad y diversidad en la cantidad y tipos de datos que pueden recolectarse, lo cual convierte al exceso de información en un problema clave para la adquisición de conocimiento. La filtración de datos relevantes resulta, entonces, de vital importancia. Los primeros sistemas de gestión del conocimiento buscaban construir archivos o repositorios de documentos, informes, presentaciones y mejores prácticas. Con el tiempo este primer repertorio se amplió hasta incluir los documentos no estructurados, generalmente en formato digital, como el E-mail. En otros casos, las organizaciones adquieren conocimiento al desarrollar redes de expertos en línea, con el propósito de que los recursos humanos de la organización encuentren al experto en el tema de interés.



Las organizaciones requieren lograr la creación de conocimiento descubriendo nuevos patrones en los datos almacenados en el sistema. En este punto cobra importancia el **OLAP o Procesamiento analítico en línea**[2], este tipo de procesamiento soporta el análisis de datos mulidimensionales, el cual permite a los usuarios ver los datos de diferentes formas utilizando múltiples dimensiones. Cada atributo o aspecto de la información representa una dimensión diferente, tales como producto, precio, costo, etc. Por ejemplo, se puede conocer el uso de un servicio determinado de la organización. OLAP permite a los usuarios del sistema obtener respuestas en línea a preguntas particulares en un período de tiempo sumamente rápido, aún cuando los datos estén almacenados en bases de datos de gran tamaño. Una organización puede utilizar una base de datos multidimensional especializada o una herramienta que genere una visualización de los datos multidimensionales en base de datos relacionales, es decir, bases de datos que representan los datos como tablas bidimensionales o relaciones, en donde las diferentes tablas se consideran archivos y cada una contiene datos sobre una entidad y sus atributos. Se efectúa, entonces, una extracción selectiva de los datos multidimensionales para su análisis.

La Minería de Datos va más allá del OLAP ya que proporciona conocimiento acerca de los datos de la organización que no se podrían obtener con el primero, al encontrar patrones y relaciones ocultas en grandes bases de datos e infiriendo reglas acerca de ellas para predecir el comportamiento futuro. Los patrones y reglas hallados se usan para orientar la toma de decisiones y predecir el efecto de tales decisiones. Los tipos de información que se pueden obtener con esta técnica son *las asociaciones, secuencias, clasificaciones y pronósticos*

Las *asociaciones* son ocurrencias enlazadas a un único evento, en las *secuencias*, los eventos se asocian con el paso del tiempo, la *clasificación* reconoce patrones que describen al grupo al cual pertenece un elemento dado

[2] Para profundizar este tema se puede leer el Documento de Trabajo Nª 14 , perteneciente a esta misma colección.

por medio del análisis de los elementos existentes que se han clasificado. El *agrupamiento* o *clustering* funciona de forma similar a la clasificación cuando aún no se han definido grupos, se realiza como se cito anteriormente por medio de una clasificación automática y la *generación de pronósticos* usa las predicciones de otra forma, emplea una serie de valores existentes para pronosticar que otros valores asociados se hallarán.

El **Text Mining** y el **Web Mining** son variaciones de la Minería de Datos. En el primer caso, tenemos las técnicas de Minería de datos, pero aplicadas a descubrir patrones ocultos en textos, es el proceso de extraer información del análisis de textos, que es útil para determinados propósitos; en el segundo estas técnicas están aplicadas a la red. Estas técnicas se emplean en analizar los datos que se hallan en documentos y páginas Web. Todos los que visitan un sitio en Internet dejan huellas digitales (direcciones de IP, navegador, etc.) que los servidores automáticamente almacenan en una bitácora de accesos (Log). Las herramientas de Web mining analizan y procesan estos logs para producir información significativa. Debido a que los contenidos de Internet consisten en varios tipos de datos, como texto, imagen, vídeo, metadatos o hipervínculos investigaciones recientes usan el término multimedia data mining (minería de datos multimedia) como una instancia del Web mining), para tratar ese tipo de datos asociados al crecimiento de Internet se han desarrollado técnicas que permiten descubrir y analizar la información existente en Internet. El Web Mining ayuda a descubrir información, localizar documentos relacionados, mostrar distintos temas, así como a evaluar el nivel de satisfacción de los recursos web, por parte de los internautas, etc..

Web mining: Según el objetivo que se persigue, se puede efectuar una clasificación en tres grandes grupos de dominios de extracción de la información en función de la naturaleza de los datos:

1. Minería de contenidos web: Busca la regularidad y dinámica de los contenidos en la Web. Los documentos Web pueden ser datos sin estructurar, archivos HTML parcialmente estructurados, o

información procedente de bases de datos generadas en páginas con formato HTML. Estos documentos hipertexto incluyen texto y también a imágenes, audio, vídeo, metadatos e hiperenlaces. La metodología utilizada en este apartado, va desde las tradicionales relaciones entre términos hasta la tecnología que se utiliza en la minería de textos.

2. Minería de estructura web o Web Mining de estructura: intenta descubrir la organización de los enlaces del conjunto de hiperenlaces dentro del documento para generar un informe estructural sobre la página y el sitio web. Obtenemos información acerca de si los usuarios encuentran la información, si la estructura de sitio es demasiado ancha o demasiado profunda, si los elementos están colocados en los lugares adecuados dentro de la página, si la navegación se entiende, cuáles son las secciones menos visitadas y su relación con el lugar que ocupan en la página central.
3. Minería web de uso: es la aplicación de las técnicas de data mining para descubrir pautas de conducta a la hora de utilizar la web por parte de los usuarios. Esta extracción se refiere a patrones de navegación que podemos descubrir en nuestros usuarios y nos pueden servir para mejorar la misma.

Text Mining: dado que el ochenta por ciento de la información de una compañía está almacenada en forma de documentos, las técnicas como la categorización de texto, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, entre otras, apoyan al text mining o Minería de Textos. En ocasiones se confunde la Minería de Textos con la recuperación de la información o Information Retrieval . Esta última consiste en la recuperación automática de documentos relevantes mediante indización de textos, clasificación, categorización, etc. Generalmente se utilizan palabras clave para encontrar una página relevante. En cambio, la Minería de Textos se refiere a examinar una colección de documentos y al descubrimiento de nueva información no contenida en ningún documento individual de la colección. El texto es información más amorfa y no estructurada, al igual que la información digital que se encuentra en Internet,

por lo que es más difícil manejar el procesamiento automático de la información.

Una forma particular de Text Mining o Minería de Textos es la *Extracción de Metadatos*. Los metadatos son datos sobre datos, se refieren generalmente a aspectos, como autor, título, esquema de clasificación, descriptores, etc. Los metadatos pueden considerarse como un registro bibliográfico enriquecido y actualmente estructurado del documento, tal que su objetivo es expandir y acompañar con frases al objeto tomando la noción de extracción de entidades. Pueden incluirse una gran variedad de atributos de estos recursos de información, como e-mail, direcciones, resúmenes, tablas de contenido, URL, ISBN, estructuras químicas, ecuaciones matemáticas, etc.

El procesamiento documental se puede acelerar y mejorar significativamente con el uso adecuado de los metadatos, se pueden realizar búsquedas, enlaces y referencias cruzadas entre los documentos representados por los metadatos. Para identificar este tipo de objetos digitales se usan atributos que se completan con piezas individuales de información estructuradas, denominadas *información de extracción*. Se realiza el análisis sintáctico automático de palabras del texto para determinar relaciones entre los distintos elementos por medio del procesamiento del lenguaje natural.

El problema fundamental para el procesamiento automático de información contenida en textos es su falta de estructura. Luego de que el sistema analiza las oraciones del texto desde los puntos de vista, morfológico y sintáctico, sigue el análisis semántico en el que se determina el significado. En el análisis morfológico se detectan las unidades mínimas que pueden formar una palabra, como el reconocimiento de prefijos y sufijos. Este análisis guarda estrecha relación con el léxico. El léxico es un conjunto de información de cada palabra que el sistema utiliza para el procesamiento. Las palabras del diccionario están representadas por una entrada léxica y en el caso de que tenga más de un significado o diferentes categorías gramaticales, tendrá asignadas diferentes entradas.

El análisis sintáctico analiza cómo cada una de las palabras de la oración que se combinan para formar construcciones correctas desde el punto de vista sintáctico. El resultado consiste en generar las diferentes categorías sintácticas formadas por las distintas unidades léxicas de una oración.

El análisis de contenido del texto adquiere el conocimiento contenido en el mismo, transformando el contenido literal y explícito en una representación más fluida que captura la estructura referencial. Dicha representación debe tener en cuenta no sólo la estructura superficial del texto, sino también operaciones referenciales tales como analogías y generalizaciones.

Entre las diferentes técnicas utilizadas para el análisis textual están:

-Modelización de textos: se busca generar un entorno para el modelado del texto, incluyendo facilidades para la investigación y simulación de procesos cognitivos

-Independencia del dominio temático: las operaciones deben descansar en **ontologías** o formas de representación del dominio temático, que proporcionan una serie de conceptos pertenecientes a dicho dominio y a sus relaciones recíprocas, en ella se indican conceptos *sinónimos*, *hipónimos* (términos más específicos de una clase genérica), *hiperónimos* (términos genéricos que pueden usarse para referirse a conceptos más específicos), *merónimos* (términos cuyo significado constituye una parte del significado global de otro término) y *antónimos* (términos con significados opuestos)

-Representación por medio del *parsing*, el cual representa significados usando las relaciones gramaticales a través de la división de las oraciones en sus componentes gramaticales y estableciendo relaciones cruzadas entre dichos componentes.

En conclusión en la Minería de Textos se darían en general, las siguientes etapas:

1. Recuperación de información, es decir, seleccionar los textos pertinentes.
2. Extracción de la información incluida en esos textos: hechos, acontecimientos, datos clave, relaciones entre ellos, etc.
3. Por ultimo se realizaría lo que antes definíamos como **Minería de Datos** para encontrar asociaciones entre esos datos claves previamente extraídos de entre los textos.

Captura, clasificación y almacenamiento

Una vez que se han descubierto los documentos, patrones y las reglas expertas deben almacenarse con la finalidad de los recursos humanos de la organización puedan recuperar y utilizar todo este conocimiento.

Generalmente el almacenamiento del conocimiento requiere la creación de una gran base de datos en que la que la colección de documentos relevantes digitalizados sean indizados y etiquetados de acuerdo a reglas consistentes, en este nivel encontramos los distintos sistemas de gestión documental, las bases de datos de conocimientos y los sistemas expertos, que ayudan a las organizaciones a preservar el conocimiento adquirido incorporándolo en los procesos y en la cultura de la organización.

Generalmente un **almacén de datos** o *data warehousing* es una base de datos que almacena datos actuales e históricos de potencial interés para los encargados de la toma de decisiones de la organización. El almacén de datos consolida y normaliza una gran variedad de datos provenientes de distintas fuentes, internas (datos operativos, de usuarios, históricos, etc.) y externas, de la organización. En el almacén todos estos datos se combinan y se reorganizan en una gran base de datos central diseñada para elaborar informes y análisis para el apoyo de la toma de decisiones por parte de los directivos de las organizaciones. Es a partir de este almacén que se pueden realizar análisis

mucho más profundos a través de la Minería de datos y puede llegarse al descubrimiento de nuevos patrones, relaciones y conocimientos más profundos para orientar esta toma de decisión.

En esta etapa, la Minería de datos puede aportar distintas técnicas de *clasificación*, como obtener un modelo que permita asignar una clase concreta de un sistema de clases, como en el caso de los árboles de clasificación o CART, árboles de clasificación: Una técnica de árbol de decisión usada para la clasificación de un conjunto de datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando dos divisiones cuyos resultados pueden expresarse en lenguaje SQL o en el modelo Bayesiano[3] o de probabilidad, en la *clasificación Bayesiana*, uno de los algoritmos más usados en la clasificación de textos, se construyen las distintas clases de documentos basándose en una estimación de las probabilidades de que un elemento pertenezca a cada clase, dado un nuevo documento

En el caso del método de *clustering*, se hace corresponder cada caso a una clase determinada, con un conjunto de clases construidas con la datos de entrada del sistema, utilizando medidas de similaridad o proximidad, se agrupan los datos según una gran variedad de criterios. Generalmente, los vectores de un mismo grupo (o *clústers*) comparten propiedades comunes. El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo. Las técnicas más usadas son la distancia mínima y los mapas autoorganizados de Kohonen[4] mediante el uso de redes neuronales. Las redes neuronales son un modelo de aprendizaje y procesamiento automático de la información inspirado en la forma en que

[3] Es un modelo en el que se considera constante la información que se posee.

[4] Este concepto fue desarrollado en el Documento de Trabajo Nª 14 perteneciente a esta colección.

funciona el sistema nervioso. Se trata de un conjunto de neuronas interconectadas en red que colaboran para producir un elemento de salida.

Los mapas autoorganizados o **SOM** (Self-Organizing Map), también llamados **Redes de Kohonen** son un tipo de red neuronal no supervisada, competitiva, distribuida de forma regular en una rejilla de, normalmente, dos dimensiones, cuyo fin es descubrir la estructura subyacente de los datos introducidos en ella. A lo largo del entrenamiento de la red, los vectores de datos son introducidos en cada neurona y se comparan con el *vector de peso* característico de cada neurona. La neurona que presenta menor diferencia entre su vector de peso y el vector de datos es la neurona ganadora (o *BMU*) y ella y sus vecinas verán modificados sus vectores de pesos.

Se usan también técnicas de *resumen* en las que se utilizan diversas representaciones de datos, como visualización de datos, generación automática de informes, etc.

Conclusiones

Puede concluirse que la Minería de Datos constituye una parte muy importante en todo el ciclo de la creación, organización, recuperación y difusión de nuevo conocimiento en el Ciclo de la Gestión del Conocimiento en las organizaciones.

Esto le permite a una organización mejorar significativamente en su eficiencia y en su eficacia. La Minería de Datos interviene a lo largo de todo el proceso, en la Adquisición, Captura, Organización, Almacenamiento y en la Recuperación del Conocimiento de las organizaciones. No obstante, es importante tomar en consideración que tipo de conocimiento, tácito o explícito se ha descubierto, analizar la forma de generación, organización y recuperación del mismo para lograr una potenciación del papel de la Minería de Datos en el Ciclo de la Gestión del Conocimiento. La Minería de Datos se usa tanto en la búsqueda de la información, como en el análisis de la misma para descubrir relaciones ocultas entre los datos, brinda métodos de procesamiento de la información, de

clasificación automática y de almacenamiento. Sirve para fomentar *el planeamiento estratégico*, entendiéndose por tal, el proceso sistemático de identificación, adquisición, análisis, registro y presentación de las oportunidades y amenazas de una organización, para apoyar de esta manera, la toma de decisiones por parte de los directivos. La Gestión del Conocimiento aliada con la Minería de Datos debe permitirle a la organización la creación continua de capital intelectual para generar un valor agregado en los servicios que la organización brinda a sus clientes, aportando de esta forma, un plus de competitividad a la organización en comparación con otras organizaciones del mismo nicho.

Bibliografía

Advances in Knowlegde Discoveryt and Data Mining (1996) , edited by Usama Fayyad.. (et. al.), Cambridge (Ma.): MIT Press

Frappaolo, C. (2006). *Knowlegde Management*. John Willey

Herramientas de Gestión del Conocimiento. En: www.**gestiondelconocimiento**.com/documentos2/america/herramientas.pdf Acceso, 1,sep. 2010

Larose, D. (2006). *Data Mining: Methods and Models*. New Jersey: John Willey

Laudon, K. (2008) J. Laudon. *Sistemas de información gerencial: Administración de la información digital.* México: Pearson.

Leite Lopez, L (2009). *Minería de datos para inteligencia competitiva*. En: www.delfos.co.cu/boletines/bsa/PDF/13**mineria**%20de%20**datos**.pdf Acceso, 1,sep., 2010

Martínez Orol, J. *La Gestión del Conocimiento y su vinculación a los Data Warehouses*. En: www.gestiopolis.com/.../**gestion-del-conocimiento**-y-su-vinculacion-a-los-**data**-warehouses.htm, Acceso, 1, sep, 2010

Valhondo, D. (2003). *Gestión del conocimiento: Del mito a la realidad*, Madrid: Díaz de Santos.

Valcarcel Acensios, V. (2004) Data Mining y el descubrimiento de conocimiento. En: *Industrial Data*, 7(2), p. 83-86.

Witten, I. (2005). E. Frank. *Data Minino: Practical Machina Learning Tools and Techniques*. 2nd. ed. New York. Elsevier