Minería de datos espaciales en búsqueda de la verdadera información*

Searching for True Information with Spatial Data Mining**

Mineração de dados espaciais na busca da verdadeira informação***

*María Ximena Dueñas-Reyes*****

* Fecha de recepción: 29 de julio de 2008. Fecha de aceptación para publicación: 23 de enero de 2009. Este artículo está basado en el proyecto de investigación denominado *Diseño e implementación de un algoritmo de Spatial Data Mining para la estimación de pérdidas no técnicas en el sector energético.* Caso de estudio: Empresa de Energía de Boyacá. Este proyecto fue financiado por la Universidad Distrital Francisco José de Caldas y cuenta con el número de registro interno 126 de 2008.

** Submitted on July 29, 2008. Accepted on January 23, 2009. This article is based on a research project called *Design and Implementation of a Spatial Data Mining Algorithm to Estimate Non- technical Losses in the Energy Sector.* Case Study: Empresa de Energía de Boyacá, financed by the Universidad Distrital Francisco José de Caldas registration N° 126/2008.

*** Data de recepção: 29 de julho de 2008. Data de aceitação para publicação: 23 de janeiro de 2009. Este artículo baseia-se no projeto de pesquisa denominado *Desenho e implementação de um algoritmo de Spatial Data Mining para a estimativa de perdas não técnicas no setor energético.* Caso de estudo: Empresa de Energía de Boyacá.

**** Ingeniera de Sistemas, Fundación Universitaria Juan de Castellanos, Tunja, Colombia. Estudiante de la Maestría en Ciencias de la Información y las Comunicaciones, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia. Correo electrónico: ximenadue@gmail.com.

Resumen

La información se ha convertido en un elemento clave en los procesos organizacionales. En los últimos años, la tecnología ha tenido un crecimiento acelerado como herramienta útil y necesaria para facilitar dichos procesos y mejorar la productividad. La inteligencia de negocio se ha definido como la transformación de datos en conocimiento, a fin de sustentar la toma de decisiones desde el punto de vista estratégico y táctico en el momento y lugar oportuno y de generar una ventaja competitiva y de efectividad. Entre las herramientas utilizadas para la exploración de datos está el procesamiento analítico en línea (OLAP), el cual, aunque permite obtener datos relevantes entre cantidades de información, es deficiente para analizar datos geográficos, por lo que ha surgido SOLAP, que brinda métodos de tratamiento especial para datos espaciales. La minería de datos se ha venido adaptando dentro de las empresas con el fin de realizar exploración y análisis de datos enfocados en el descubrimiento del conocimiento. Dada la importancia que la información espacial está tomando, surge la minería de datos espacial. Este proceso permite descubrir patrones útiles e inesperados dentro de los datos. Las técnicas de minería de datos espacial se aplican para extraer conocimiento a partir de grandes volúmenes de datos, los cuales pueden ser de tipo espacial y no espacial. Entre ellas se encuentran generalización, agrupación, asociación espacial, entre otras.

Palabras clave

Minería de datos, inteligencia de negocio, tecnología OLAP.

Abstract

Information is a key element of organizational processes. In recent years, the technology has had an accelerated growth, in line with the increase of what are regarded as useful and necessary tools to facilitate and expedite those processes giving an added value to productivity. Business intelligence has been defined as the transforming of data into knowledge, providing decision-making support at the strategic and tactical level where and when appropriate, providing a competitive advantage and increasing the effectiveness. One of the tools that has become useful for the exploration of data is *On-line Analytical Processing* (OLAP), which allows to obtain outstanding data among quantities of information, but it is faulty for the analysis of geographical data, for which SOLAP has arisen, which offers methods of special treatment for space data. Data mining has been adapted within companies, with the purpose of carrying out exploration and analysis of data focused on the discovery of knowledge. Because of the important place that space information is occupying nowadays, the spatial data mining has arisen. This process allows us to discover useful and unexpected patterns inside the data. The techniques of spatial data mining are applied to extract knowledge, starting from large volumes of data, which can be of space and non-space types. Among them are generalization, grouping, and space association.

Key words

Data mining, business intelligence, OLAP technology.

Resumo

A informação se converteu em um elemento chave nos processos organizacionais. Nos últimos anos, a tecnologia tem tido um crescimento acelerado como ferramenta útil e necessária para facilitar tais processos e melhorar a produtividade. A inteligência de negócio se define como a transformação de dados em conhecimento, a fim de sustentar a tomada de decisões desde o ponto de vista estratégico e tático no momento e lugar oportuno e de gerar una vantagem competitiva e de efetividade. Entre as ferramentas utilizadas para a exploração de dados está o processamento analítico online (OLAP), o qual, ainda que permita obter dados relevantes entre quantidades de informação, é deficiente para analisar dados geográficos, por isso apareceu o SOLAP, que oferece métodos de tratamento especial para dados espaciais. A mineração de dados tem se adaptado dentro das empresas com o objetivo de realizar exploração e analise de dados enfocados no descobrimento do conhecimento. Dada a importância que a informação espacial está tomando, surge a mineração de dados espacial. Este processo permite descobrir padrões úteis e inesperados dentro dos dados. As técnicas de mineração de dados espacial aplicam-se para extrair conhecimento a partir de grandes volumes de dados, os quais podem ser de tipo espacial e não espacial. Entre elas encontram-se generalização, agrupação, associação espacial, entre outras.

Palavras chave

Mineração de dados, inteligência de negócio, tecnologia OLAP.

Introducción

En un mercado cada vez más competitivo, el conocimiento cobra mayor relevancia. En este sentido, la inteligencia de negocio ofrece la posibilidad de obtener ventajas empresariales, mediante su aplicación en los procesos de toma de decisiones. Entre las técnicas que se utilizan, la minería de datos ha ocupado un lugar primordial en el mundo empresarial, ya que permite encontrar modelos ocultos dentro de las grandes cantidades de datos y generar conocimiento. Debido a los diversos datos que se manejan dentro de una corporación, los datos de tipo geográfico han tomado mayor relevancia, pues dan un valor agregado al análisis de los datos. A raíz de esto, surge la minería de datos espacial.

En este artículo se pretende abordar la importancia de la inteligencia de negocios y las diversas tecnologías que utiliza para lograr una toma de decisiones más acertada en los diversos ámbitos organizacionales. Entre ellas se encuentra la minería de datos espacial, la cual logra integrar los datos de tipo geográfico dentro del análisis, y así obtener una información precisa y efectiva para la corporación. Lo anterior bajo el supuesto de que todo ocurre en algún lugar en el espacio y en un momento de tiempo dado; por ende, caracterizar las dimensiones espacio y tiempo permitirá realizar análisis y descubrimientos de conocimiento más acertados.

Este documento está estructurado de la siguiente manera: en la primera sección se da una visión general sobre la definición y la importancia de la inteligencia de negocio. En la segunda sección se describe la importancia y las características de la inteligencia de negocio espacial y de *Spatial OLAP*. En la tercera sección se define la minería de datos, sus características, modelos y algoritmos, así como la importancia de la minería de datos espacial, y se enuncian algunas técnicas y algoritmos que se utilizan en la minería de datos espacial. Por último, se concluye.

1. Inteligencia de negocio

Hoy en día, la información se ha convertido en un recurso vital para el desarrollo y la evolución de cualquier empresa, donde la competitividad hace necesario obtener información de una manera rápida y eficiente frente a su diario crecimiento. A raíz de esto, surge la *inteligencia de negocio* como un nuevo concepto orientado al tratamiento inmediato de los datos, la información y, en definitiva, el conocimiento, con el fin de mejorar los procesos de cualquier organización frente a las exigencias del mercado (Reinschmidt y Francoise, 2000). El filósofo Albin Tofel manifiesta "El dueño de la información es el dueño del poder", y a esta afirmación hay que adicionarle: en el lugar y momento oportuno.

Este tipo de inteligencia se define como la transformación de los datos en conocimiento, para sustentar la toma de decisiones en los ámbitos estratégico y táctico, en el momento y lugar oportuno, y obtener una ventaja competitiva a través de la gestión de conocimiento, orientado al apoyo a la toma de decisiones, con el propósito de incrementar la efectividad de la empresa. La gestión, el manejo y el control de la información también hacen parte de este nuevo concepto, con base en herramientas de análisis que permiten mejorar el rendimiento de las organizaciones (Moss y Atre, 2003; Marcano, Yelitza y Talavera, 2006).

2. Inteligencia de negocio espacial

Las herramientas de inteligencia de negocio se han centrado en el tratamiento de los datos con el fin de brindar una información óptima, que permita a los analizadores generar una buena toma de decisiones con el fin de mejorar los procesos y, por ende, los niveles competitivos de la organización frente al mercado. Con el aumento de la información dentro de las organizaciones, también han aumentado los diversos tipos de datos, y uno de los que está tomando relevancia en la actualidad es el tipo de dato geográfico (Reinschmidt y Francoise, 2000).

En la actualidad los reportes generados por las herramientas de inteligencia de negocio presentan datos muy simples en forma gráfica, a través de histogramas, tortas, barras, entre otros, pero por esta misma simplicidad carecen de cierta información (Gonzales, 2004). Esta limitación lleva a que los datos generados no sean completamente efectivos en la toma de decisiones. "Valiosas ideas se ocultan en la gran cantidad de datos operativos y de negocios generada por las organizaciones", dice Steve Trammell de ESRI: "Hacia el futuro las organizaciones, tienen que darse cuenta de que la adición de la dimensión de análisis geográfico en las aplicaciones sofisticadas de inteligencia empresarial, presentan resultados con un mayor conocimiento de las decisiones" (CBR, 2005, s. p.).

Los datos de tipo geográfico que se manejan hoy en día dentro de las organizaciones se presentan a través de direcciones de clientes, proveedores, sucursales, entre otros. Al aplicar la inteligencia de negocios en un caso concreto como la discriminación de sectores con niveles de pérdida, la visualización de información permitirá, de una forma más fácil, presentar los datos generados por la inteligencia de negocio y tener una mejor comprensión y análisis de estos (Roddick y Spiliopoulou, 2002; SAS, 2004).

"Gran parte de la información almacenada en bases de datos de la empresa de hoy tiene algún tipo de contenido geográfico en la misma, ya sea que se relaciona con una dirección o de una región o un código postal", dice Renaud Besnard, director de *marketing* del producto solución para SQL Server 2005 en Microsoft. "Desafortunadamente, esta información normalmente sólo se ve a través de una hoja de cálculo, una tabla o un gráfico. Pocas veces se analizan en su formato más lógico: en el mapa" (CBR, 2005, s. p.).

A partir de lo anterior, surge la inteligencia de negocio espacial, la cual agrega la variable espacio y logra obtener información más precisa y efectiva para la toma de decisiones acertadas en todos los ámbitos de la organización y una mejor generación de conocimiento.

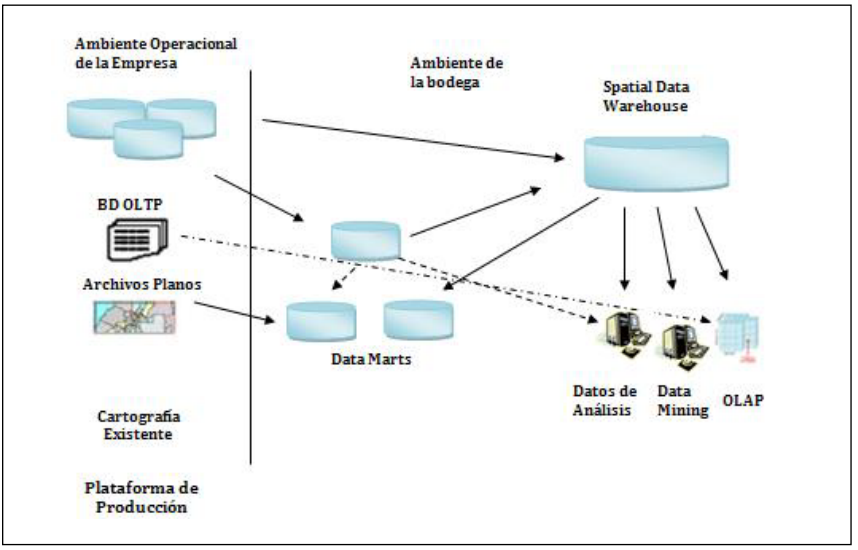
2.1 Spatial Data Warehouse

Las bodegas de datos surgieron como soluciones orientadas al análisis de la información de una organización con el fin de mejorar la eficiencia y eficacia de esta ante el mercado. "El *Data Warehouse* es una colección de datos orientados a temas, integrados, no volátiles, historiados, organizados para el apoyo de un proceso de ayuda a la decisión" (Inmon, 2005, p. 29). Este tipo de repositorios permite un eficiente procesamiento analítico de datos y, consecuentemente, una óptima toma de decisiones gerenciales.

Debido a la creciente información que actualmente manejan las empresas, fue necesario crear nuevas herramientas que ofrezcan un manejo ágil, eficaz y eficiente de esta. A raíz de esto, surgieron los *Spatial Data Warehouse*, los cuales combinan las bases de datos espaciales y las tecnologías del *data warehouse* y logran un mayor apoyo para el manejo de datos de tipo espacial (Malinowski y Zimánky, 2008). Una bodega de datos espacial se caracteriza por que está orientada a temas, no es volátil, está integrada y es de tiempo variante y de geografía del dato (Bohórquez, 2000; Inmon, 2005). Los *Spatial Data Warehouse* manejan un esquema mutidimensional igual a las bodegas de datos tradicionales, pero agregan una extensión de tipo espacial que permite el manejo de

elementos espaciales y no espaciales (Malinowski y Zimánky, 2008), así como lo señala la Figura 1.

Figura 1. Esquema global de *Spatial Data Warehouse*



Fuente: (Bohórquez, 2000).

Dentro de los principales componentes de una arquitectura *Spatial Data Warehouse* se encuentran el *Procesamiento Analítico en Línea* (OLAP, por su sigla en inglés) y el *Data Mining* (minería de datos). El primero provee un análisis intrusivo de alto desempeño sobre la bodega de datos, mientras el segundo se encarga de explotar, escarbar y analizar de forma exhaustiva la bodega.

2.2 SOLAP (Spatial OLAP)

Con el OLAP es posible obtener datos relevantes entre la gran cantidad de información que se maneja a diario dentro de una empresa. Los conceptos multidimensionales que se manejan en el OLAP incluyen: dimensiones, atributos de medidas, miembros, hechos y cubos de datos, los cuales están compuestos por un conjunto de medidas agregadas de acuerdo con el conjunto de dimensiones. El uso de cubos OLAP permite realizar un análisis de tipo multidimensional de modo ágil y eficiente (Wrembel y Koncilia, 2007; Baltzer, 2006).

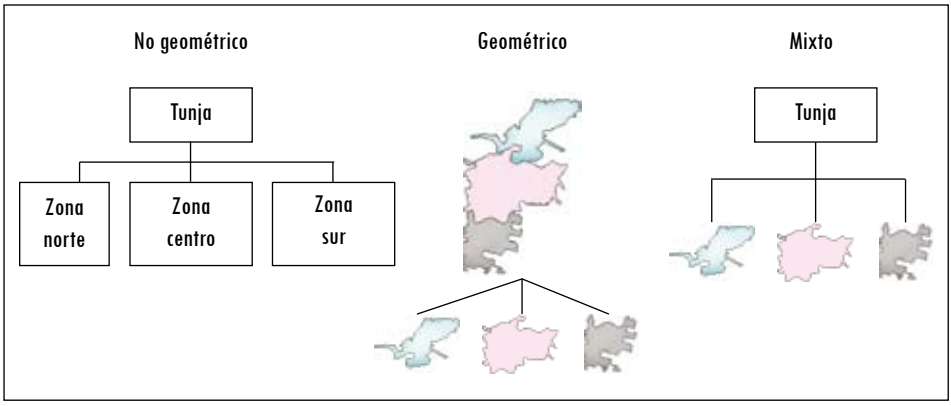
Las herramientas de análisis usadas en la actualidad, como el OLAP, toman el dato de tipo geográfico como un atributo, sin darle la importancia que requiere, y limitan así la información generada. A partir de esta necesidad se

han generado herramientas como: SOLAP y bodegas de datos espaciales, que permiten integrar funcionalidades de sistemas de información geográficos (SIG) para la gestión de este tipo de datos, incluidas operaciones de consolidación y formas de navegación entre los diversos elementos espaciales (Matias y Moura-Pires, 2005; Abril y Pérez, 2007).

El *Spatial OLAP* se ha definido como "un tipo de software que permite una rápida y fácil navegación dentro de bases de datos espaciales y que ofrece muchos niveles de granularidad de la información, muchos temas, muchas épocas y muchos modos de visualización sincronizada" (Bédard, Proulx y Rivest, 2005, p. 5). Este tipo de herramientas permiten utilizar el potencial de las herramientas OLAP y agregar métodos para la explotación de información de tipo geográfico (Rivest *et al*., 2005).

El SOLAP soporta tres tipos de dimensiones: espaciales no geométricas, las cuales abarcan elementos de referencia espacial nominal; espaciales geométricas, que permiten representar sus elementos en todos los detalles de tipo geométrico, y espaciales mixtas, que comprenden figuras de tipo geométrico para varios niveles de detalle (Rivest *et al*., 2003) (Figura 2).

Figura 2. Tipos de dimensiones soportados por SOLAP



Fuente: presentación propia de la autora.

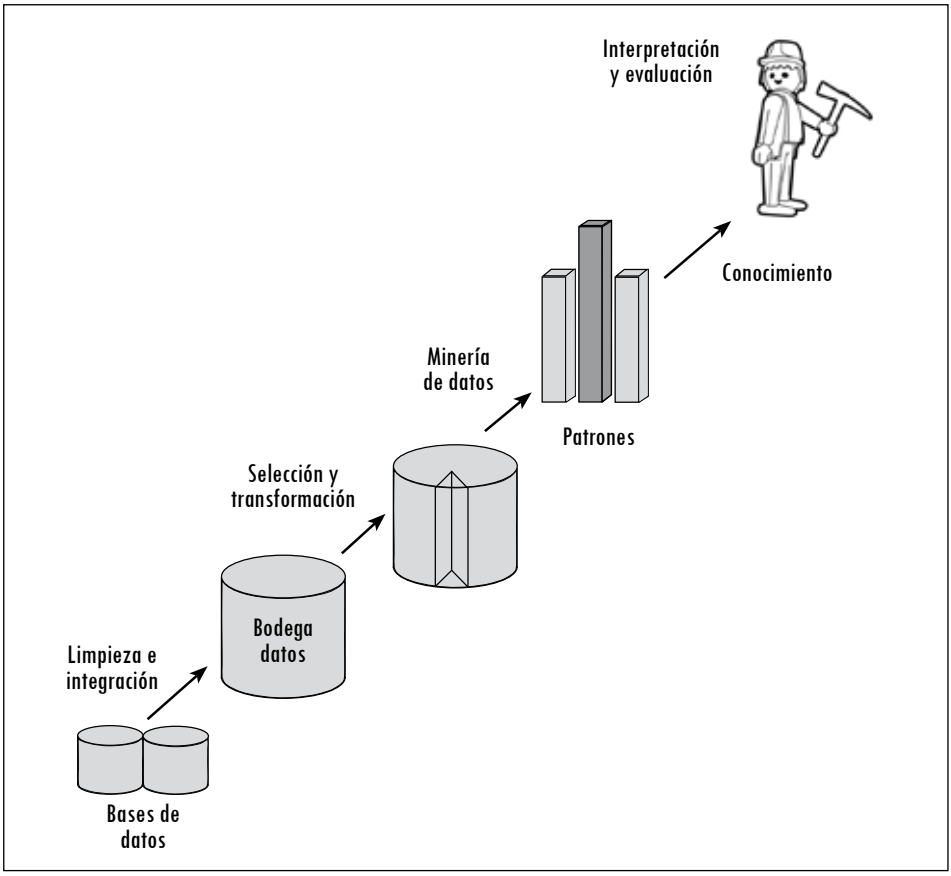
3. Minería de datos espaciales (*Spatial Data Mining*) y técnicas aplicadas

Una de las técnicas de inteligencia de negocio que se ha venido utilizando dentro de las empresas es la minería de datos, la cual a partir de la exploración y el análisis se enfoca en descubrir conocimiento. Según Fayad, la minería de datos es "un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos"

(Fayad, 1996, p. 5). La minería reúne ventajas de diversos campos como lo son: la estadística, la inteligencia artificial, la computación gráfica, las redes neuronales, entre otros" (Bramer, 2007, s. p.). La minería de datos brinda un acceso y navegación retrospectiva de los datos y de esta manera genera información precisa y oportuna a partir del apoyo de tres tecnologías (Bramer, 2007): (1) recolección de datos, (2) multiprocesador y (3) algoritmos de minería de datos.

A través de la minería de datos se logra descubrir información en forma de patrones, cambios, asociaciones y estructuras significativas de grandes cantidades de datos almacenados en *Data Warehouse*. Para poder llevar a cabo el descubrimiento de conocimiento se deben seguir una serie de pasos iterativos: limpieza, integración, selección, transformación, minería de datos, evaluación de patrones y representación de conocimiento (Roddick y Lees, 2001) (Figura 3).

Figura 3. Pasos para la minería de datos



Fuente: (Han y Kamber, 2006).

La aplicación de algoritmos de minería ha permitido detectar patrones en los datos y, por ende, crear modelos que sustenten la toma de decisiones, y así contribuir al mejoramiento de los índices de competitividad o del problema en particular (Roddick y Lees, 2001). La minería de datos consiste en (Bramer, 2007):

- *Análisis de dependencias*. La dependencia puede ser probabilística, es decir, a partir de un valor se puede predecir el de otro elemento. La dependencia permite determinar el valor de otras dependencias o ser funcional.
- *Identificación de clases*. Esta reconoce los grupos que describen datos y que pueden ser utilizados para la entrada de otros sistemas.
- *Descripción de conceptos*. Este tipo de descripción permite identificar registros comunes entre categorías denominadas *descripción de características*, donde difieren de sus grupos.
- *Detección de desviaciones, casos extremos o anomalías.* A través de ello se identifican cambios significativos en los datos respecto a sus valores y se logran filtrar datos irrelevantes para las futuras tomas de decisiones.

La arquitectura de la minería de datos consiste en (Roddick y Lees, 2001) (Figura 4):

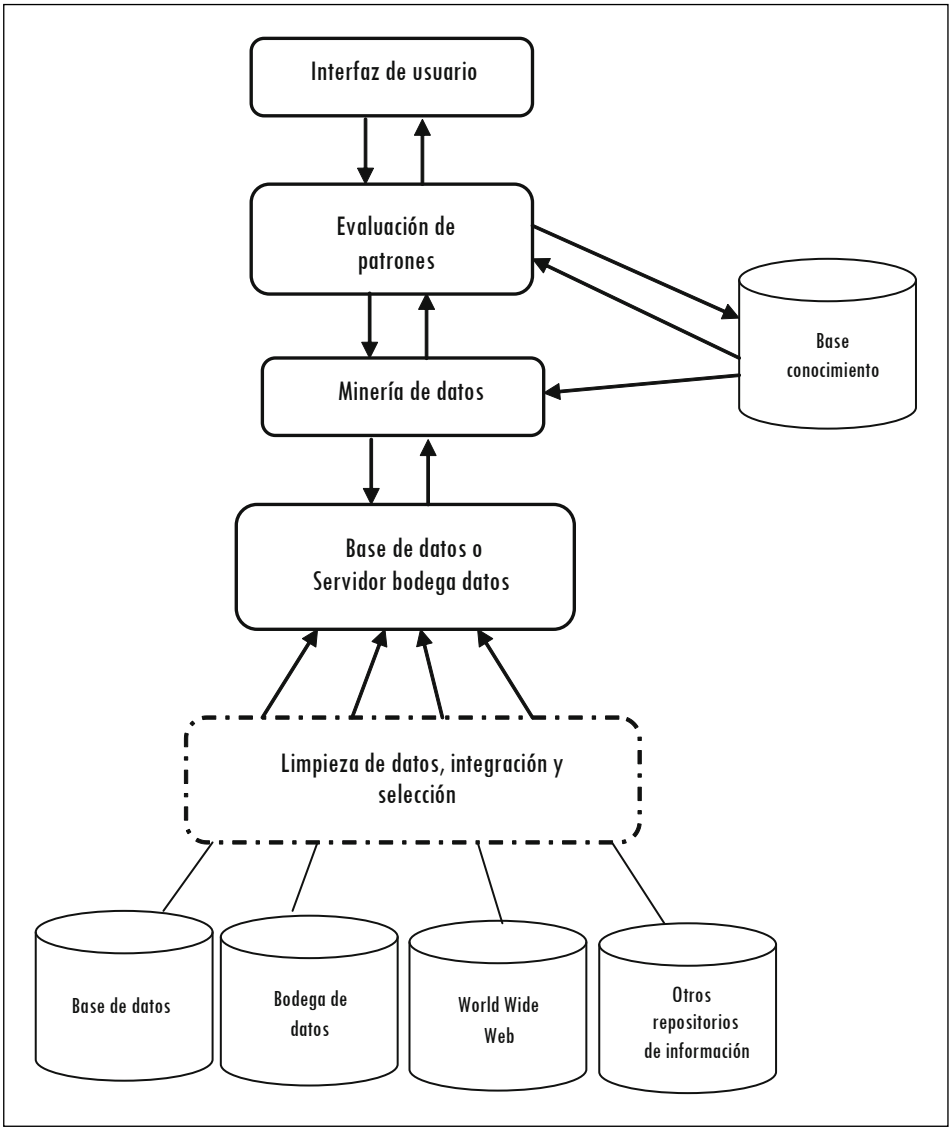
- *Bases de conocimiento*. Con el fin de dirigir búsquedas y evaluar patrones se tiene un completo conocimiento del dominio.
- *Algoritmo de minería de datos*. A través de estos se pueden realizar distintos tipos de análisis de los datos en búsqueda del conocimiento.
- *Evaluación de patrones*. En conjunto con los algoritmos de minería, se busca evaluar los diversos patrones con el fin de encontrar el más interesante.

Los algoritmos de minería de datos son procedimientos bien definidos, es decir, procesos codificados como un conjunto finito de reglas que toman los datos como entradas y sus salidas como modelos o patrones (Kurgan y Musilek, 2006). De acuerdo con los métodos de aprendizaje, se pueden clasificar los algoritmos en (Svetlozar, 2003):

- *Algoritmos de tipo predictivo*. Predicen el valor de un atributo del conjunto de datos a partir de otros atributos conocidos. Cuando la etiqueta de los datos se conoce, se induce una relación entre esta y otra serie de atributos.
- *Algoritmo de descubrimiento de conocimiento.* Descubren patrones y tendencias en los datos actuales y permite obtener beneficio de estos.

Las técnicas de minería de datos pueden clasificarse de acuerdo con los dos grandes grupos de minería de datos, como se observa en la Tabla 1 (Moreno *et al*., 2001).

Figura 4. Arquitectura de minería de datos



Fuente: (Han y Kamber, 2006, p. 9).

Tabla 1. Técnicas de minería de datos

Supervisados	No supervisados
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento
Series temporales	Reglas de asociación
	Patrones secuenciales

Fuente: (Moreno *et al.*, 2001).

Entre las técnicas de minería de datos que existen actualmente las más utilizadas son:

- *Redes neuronales*. Esta técnica de inteligencia artificial se ha convertido en una herramienta de uso frecuente para descubrir categorías comunes en los datos, ya que son capaces de detectar y aprender patrones complejos y sus características. Una de las características principales de las redes neuronales es su capacidad de trabajar con datos incompletos, incluso paradójicos (Hernández *et al*., 2007).
- *Árboles de decisión*. En este tipo de representación cada nodo es una decisión que, a su vez, genera reglas para la clasificación de un conjunto de datos. Los árboles de decisión son de fácil interpretación y admiten atributos de tipo discreto y continuo (Sánchez, Miranda y Cerda, 2004).
- *Predicción*. El análisis de predicción está relacionado con las técnicas de regresión. La idea de este tipo de análisis es descubrir la relación entre variables ya sean independientes o dependientes. Por ejemplo, si las ventas son una variable independiente, entonces el beneficio puede ser una variable dependiente. Mediante el uso de datos históricos de ambas ventas y beneficios, las técnicas lineales o no lineales de regresión pueden producir una curva que permita la predicción de beneficios en el futuro (Sánchez, Miranda y Cerda, 2004; Olson y Denle, 2008).
- *Patrones secuenciales*. Estos realizan un análisis que permite encontrar patrones similares en los datos de transacciones durante un período de negocio. Los analistas pueden usar estos patrones para identificar relaciones entre los datos. Los modelos matemáticos son patrones secuenciales detrás de la lógica normativa, la lógica difusa u otras. En la fase de minería de datos, es posible estudiar varias secuencias similares para identificar las tendencias futuras en el desarrollo de transacciones. Este enfoque es útil en el tratamiento de bases de datos con características de series de tiempo (Olson y Denle, 2008).

- *Naïve Bayes*. Se trata de una técnica que combina la clasificación y predicción, con el fin de construir modelos para predecir posibles resultados a partir de asociaciones en los datos históricos (Domingo y Lowd, 2005).
- *Reglas de asociación*. A partir de estas se pueden descubrir relaciones entre elementos de una base de datos o bodega, lo cual permite analizar la información. Agrawal define una regla de asociación como una implicación de la forma $X \Rightarrow Y$, donde $X, Y \subset I$ y $X \cap Y = 0$ (Sánchez, Miranda y Cerda, 2004).
- *Series de tiempo*. Las series de tiempo en la minería de datos permiten buscar patrones a partir de grandes cantidades de datos. Algunas de sus variables están en función del tiempo. Esta técnica se utiliza a partir del comportamiento histórico de los datos, que permite modelar los componentes básicos de la serie, y así se logra hacer predicciones (Martínez de Pisón *et al*., 2005).

La minería de datos permite descubrir el conocimiento no sólo en el ámbito organizacional, sino sustentar investigaciones en la rama biológica. Así, en la reunión *Chips to Hits* '99 se planteó que actualmente uno de los cuellos de botella de los ensayos con tecnologías basadas en *biochips* se encuentra en la carencia de herramientas bioinformáticas adecuadas para el análisis y gestión de los datos, debido a los enormes volúmenes de datos que ellos generan. Así mismo, se resaltó la necesidad de emplear las técnicas de la minería de datos como la mejor forma de obtener conocimientos a partir de los resultados experimentales (Martín, López y Maojo, 1999).

Actualmente la información tiene un papel importante en la competitividad y la productividad de cualquier organización. Entre los diversos tipos de datos que se manejan, la información de tipo geográfico está tomando gran relevancia en la toma de decisiones organizacionales (Bramer, 2007).

Los datos de tipo espacial reflejan la primera ley de la geografía enunciada por Tobler (1979), quien afirma que en el análisis geográfico todo está relacionado con todo, pero las cosas cercanas están más relacionadas entre sí que las cosas lejanas (Martínez de Pisón *et al*., 2005; Martín, López y Maojo, 1999). A partir de ello han surgido herramientas que han logrado extraer conocimiento de este tipo de datos: la minería de datos espacial. Este proceso permite descubrir patrones útiles y relaciones entre datos que se han desconocido en las bases de datos espaciales (Santos y Amaral, 2000).

Aunque a primera vista se podría pensar que la minería de datos espacial es similar a la minería de datos tradicional, existen múltiples diferencias entre ellas; una de ellas radica en los objetos. Los objetos de tipo espacial manejan un

componente descriptivo —por ejemplo, nombres, datos de población, cantidades de ventas— y un componente espacial —donde se incluye su geometría, la cual se representa por puntos, líneas y polígonos—. Los puntos representan espacialmente una geometría de cero dimensiones, que denota una localización en el espacio; las líneas representan un objeto de una dimensión y denotan un conjunto conectado de puntos; mientras que los polígonos son formados por composiciones de líneas ordenadas (Malinowsky y Zimányi, 2004).

Las relaciones entre los objetos de tipo tradicional son aquellas explícitas en la entrada de los datos como las relaciones aritméticas, de orden, de subclases y entre miembros, a diferencia de las relaciones en objetos con características espaciales, las cuales están implícitas en la entrada de los datos y muestran las relaciones entre ellos, como se aprecia en la Tabla 2 (Martín, Kriegel y Sander, 2001).

Tabla 2. Relaciones entre objetos con y sin características espaciales

Relaciones no espaciales (explícito)	Relaciones espaciales (frecuentemente implícito)
Aritméticas	Orientado a conjunto: unión, intersección…
Ordenamiento	Topológicas: solapamiento, inclusión…
Subclases de	Métricas
Partes de	Dinámicas

Fuente: (Shekhar, Zhang, Huang Yan y Vatsavai, 2008).

El objetivo de la minería de datos espaciales es encontrar relaciones entre objetos de tipo espacial y no espacial a través de relaciones, como las topológicas, las de orientación espacial y las de distancia de información (Hsu, Li y Wang, 2008). La extracción de patrones en conjuntos de datos espaciales es más complicado que en conjuntos de datos tradicionales (datos numéricos), debido a la complejidad de los datos, relaciones y autocorrelación espacial (Martín, Kriegel y Sander, 2001).

Para poder entender la importancia de la minería de datos espacial se puede tomar el caso de un analista de una empresa de distribución de energía eléctrica en la cual se han detectado pérdidas de tipo no técnico dentro de la red de la ciudad; encontrar la relación entre el estrato socioeconómico del usuario y la pérdida genera patrones no esperados en la minería de datos tradicional.

Aunque las técnicas y algoritmos de la minería de datos tradicional y espacial son similares, hay que recalcar que los últimos deben manejar características especiales debido a la complejidad de los datos (Martín, Kriegel y Sander, 2001; Assaf, Ran y Dan, 2003; Eun-Jeong *et al*., 1998):

- La cantidad de datos de tipo espacial ha venido creciendo vertiginosamente dentro de las empresas y se ha convertido en pieza clave para el descubrimiento de patrones; por lo tanto, los algoritmos deben estar preparados para grandes cantidades de información.
- Para el tratamiento de datos espaciales se debe manejar el razonamiento espacial y técnicas de optimización de búsquedas de tipo espacial.
- Este tipo de algoritmos debe tener presente a los vecinos de los objetos, ya que estos pueden tener una influencia significativa en el objeto mismo.

Las técnicas de minería de datos espaciales se aplican para extraer conocimiento a partir de grandes volúmenes de datos, los cuales pueden ser de tipo espacial y no espacial. En particular, son utilizadas para encontrar relaciones entre datos espaciales y no espaciales, entender los datos de tipo espacial, entre otras razones. Estas técnicas son parecidas a las de minería de datos tradicional, pero con el factor espacial como valor agregado. Entre ellas se encuentran la generalización, la agrupación, la exploración de asociación espacial, entre otros (Harms, Deogun y Goddard, 2003; Al-Hamami, Mohammad y Hassan, 2006).

3.1 Técnica de generalización

La técnica de generalización para la minería de datos espaciales requiere el uso de jerarquías de conceptos; en este caso se tienen dos tipos de jerarquía: temática o espacial. Para entender mejor este concepto, la jerarquía temática puede ejemplificarse si, en el caso del sector eléctrico, se asimila medidor y transformador a infraestructura, mientras que la jerarquía espacial se puede asociar con varios sectores en un municipio (Santos y Amaral, 2000).

Entre los algoritmos de esta técnica se encuentran la generalización de datos espaciales y no espaciales. En la primera, a partir de la jerarquía de datos espaciales, la generalización se puede realizar inicialmente en los datos espaciales, mientras que en la segunda se realiza una inducción orientada sobre atributos no espaciales, que logran que este tipo tenga un mayor grado de concepto a partir de la generalización (Santos y Amaral, 2000; Vyas, Kumar y Tiwary, 2007).

En los algoritmos anteriormente mencionados se supone que la jerarquía está previamente dada; sin embargo, se pueden encontrar situaciones donde estas no

sean a priori o, al tratar con la jerarquía de componentes espaciales, las regiones *merging* se encuentran en un nivel más bajo que los de la jerarquía normal de una región. A partir de lo anterior surgen algoritmos que no requieren este tipo de jerarquía y se encuentran dentro de las técnicas de agrupación.

3.2 Técnica de agrupación

En este tipo de minería, una de las técnicas más relevantes es la clasificación de los objetos de acuerdo con características similares. Es posible afirmar que la ubicación geográfica cumple con un papel importante en la determinación de un fenómeno específico. En la mayoría de casos, la dependencia espacial puede expresarse no sólo como de primer orden, sino también de segundo orden, es decir, para cada zona son importantes sus vecinos inmediatos y los vecinos de sus vecinos (Giraldo, 2007; Man y Nikos, 2004).

Los algoritmos que hacen parte de la técnica de agrupación pueden catalogarse en algoritmos de agrupación particional, de agrupación jerárquica y de agrupación basada en localización. En el agrupamiento particional, los algoritmos agrupan los objetos de acuerdo con su grado de similitud; en estos se encuentran los métodos *k-means* y *k-medoid*. En la agrupación jerárquica se desarrollan operaciones de agrupamiento, como *bottom up* y *top down*. Por último, en el de localización las agrupaciones se realizan de acuerdo con sus relaciones locales mediante algoritmos basados en densidad o distribución aleatoria (Assaf, Ran y Dan, 2003; Han, Lamber y Tung, 2001).

Entre los algoritmos que se utilizan en esta técnica para el tratamiento de datos espaciales se encuentra el CLARANS, el cual consiste en búsquedas aleatorias en grupos limitados de datos, que combinan características de los algoritmos PAM y CLARA, y forma una muestra de datos en las distintas fases de búsqueda (Raymond y Han, 1994; Eun-Jeong *et al*., 1998). A partir de la importancia de los datos espaciales de CLARANS se deriva el SD CLARANS (aproximación dominante espacial), el cual busca descubrir características no espaciales en grupos espaciales, y el NSD CLARANS (aproximación dominante no espacial), el cual busca descubrir clústeres espaciales en grupos de datos no espaciales (Raymond y Han, 1994).

3.3 Métodos de exploración de asociación espacial

El descubrimiento de reglas de asociación espacial permite establecer, como su nombre lo indica, reglas que asocian objetos espaciales con uno o más objetos espaciales. Una regla de asociación se define como $X \rightarrow Y$, donde X y Y son conjuntos

de predicados espaciales o no espaciales (Vyas, Kumar y Tiwary, 2007). Se debe recordar que los predicados espaciales permiten calcular relaciones entre objetos y devuelven un valor booleano, entre los que se encuentran valores como *disjoint, touches*, *overlaps*, *contains*, *adjacent_to*, *near_by*, *inside*, *close_to,* entre otros.

En la técnica de asociación se introducen dos conceptos: mínimo soporte y mínima confianza. En las grandes bases de datos pueden encontrarse múltiples asociaciones entre los objetos, pero estas asociaciones deben poder aplicarse a pequeños grupos; por esto se deben filtrar las asociaciones utilizando mínimo soporte y mínima confianza (Agrawal, Imielinskib y Swami, 1993; Mennis y Liu, 2005).

La entrada de este tipo de reglas especifica una relación de *n-tuplas* con atributos espaciales, relación espacial de vecinos, concepto de jerarquía para los atributos, selección de tipos de objetos relevantes, mínimo soporte y mínima confianza (Martin, Kriegel y Sander, 2001; Mennis y Liu, 2005).

Para garantizar que se generen ciertas reglas de asociación espacial se deben tener en cuenta las limitaciones sintácticas y las de apoyo. Las primeras indican las restricciones que se tienen sobre un tema específico que pueden aparecer en la regla, mientras que las segundas revisan las operaciones que se consideran necesarias para la unión de predicados tanto en el antecedente como el consecuente de la regla (Cheungi *et al*., 1996; Ping-Yu, Yen-Liang y Chun-Ching, 2004).

Uno de los algoritmos más conocidos en la asociación espacial se denomina a priori, desarrollado por Agrawal, en 1993. Este algoritmo trabaja básicamente en dos pasos: en el primero, los grandes ítems son determinados de acuerdo con la frecuencia de los elementos dentro del grupo, mientras que en el segundo paso se detectan las reglas de asociación (Lutu, 2002; Harms, Deogun y Goddard, 2003).

4. Conclusiones

Hoy en día, para que las organizaciones puedan alcanzar un mayor grado de competitividad frente a las necesidades del mercado, deben utilizar la inteligencia de negocio, proceso que se basa en la recolección, el procesamiento y el almacenamiento de los datos generados por la empresa, con el fin de crear conocimiento. La minería de datos espaciales, a diferencia de la minería tradicional, rescata la importancia de los datos geográficos que a diario pasan inadvertidos; sin embargo, en ellos se oculta un conocimiento valioso como patrones de comportamiento, que son sinónimo de mejoramiento en índices de productividad y competitividad empresarial.

Las técnicas de minería de datos espacial pueden catalogarse como conjuntos de estrategias que permiten solucionar problemas específicos en los ámbitos del negocio donde los datos de tipo espacial tienen relevancia. Estas técnicas apoyan la identificación de patrones y descubren asociaciones entre datos espaciales y no espaciales, agrupamientos, entre otros. Los algoritmos en la minería de datos espacial deben manejar grandes cantidades de datos y razonamientos y optimizaciones de búsqueda de tipo espacial. Este tipo de minería está tomando gran relevancia en diversos campos donde los datos de tipo geográfico se manejan a diario, como son los sistemas de información geográfica, *geo-marketing*, control de tráfico, detección de fraudes, entre otros.

Referencias

ABRIL, D. y PÉREZ, J. Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a bases de datos espaciales. *Revista Ingeniería de Investigación*. 2007, vol. 27, núm.1, pp. 58-67. ISSN 0120-5609.

AGRAWAL, R.; IMIELINSKI, T. y SWAMI, A. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. 1993, pp. 1-4.

AL-HAMAMI, A.; MOHAMMAD, A. y HASSAN, S. Applying data mining techniques in intrusion detection system on web and analysis of web usage. *Information Technology Journal*. 2006, vol. 5, núm. 1, pp. 1-4.

ASSAF, S.; RAN, W. y DAN, T. A high-performance distributed algorithm for mining association rules. In *The Third IEEE International Conference on Data Mining (ICDM'03)*. Melbourne, 2003.

BALTZER, O. *Spatial OLAP and data mining*. 2006 [web en línea]. <http://www.cs.dal.ca/news/def-1808.shtml>. [Consulta: 15-04-2008].

BÉDARD, Y.; PROULX, M. J. y RIVEST, S. Enrichissement du OLAP pour l'analyse géographique: exemples de réalisation et différentes possibilités technologiques. En BENTAYEB, F.; BOUSSAID, O.; DARMONT, J. y RABASEDA, S. (Eds.). *Entrepôts de Données et Analyse en ligne, RNTI B_1*. Paris: Cépaduès, 2005, pp. 1-20.

BOHÓRQUEZ, J. E. Aproximación metodológica de un Spatial Data Warehouse. 2000 [documento en línea]. < http://proceedings.esri.com/library/userconf/latinproc00/colombia/spatial_data.pdf>. [Consulta 20-04-2009].

BRAMER, M. Date for data mining. En *Principles of data mining*. London: Springer, 2007, pp. 11-20.

CBR. *Spatial business intelligence*. 2005 [web en línea]. <http://www.cbronline.com/article_cbr.asp?guid=8D70BDDB-616B-4D66-8D8F-5FFCCD9AF431>. [Consulta: 15-04-2008].

CHEUNGI, D. *et al*. Maintenance of discovered association rules in large databases: an incremental updating technique. *Proceedings of ICDE*. 1996, pp. 1-3.

DOMINGO, P. y LOWD, D. Naive bayes models for probability estimation. *ACM International Conference Proceeding Series*. 2005, vol. 119, pp. 529-536.

EUN-JEONG, S. *et al*. A spatial data mining method by clustering analysis. *Proceedings of the 6th International Symposium on Advances in Geographic Information Systems*. Washington, 6-7 de noviembre de 1998.

GIRALDO, R. Análisis exploratorio de variables regionalizadas con métodos funcionales. *Revista Colombiana de Estadística*. 2007, vol. 30, núm. 1, pp.115-127.

GONZALES, M. *Spatial business intelligence. The spatial & visual components for effective BI* [Documento en línea]. 2004. <http://www.brio.nl/files/SpatialBI-v4-1-Generic_01.pdf> [Consulta: 10-05-2008].

HAN, J.; KAMBER, M. y TUNG, A. Spatial clustering methods in data mining: a survey. En MILLER, H. y HAN, J. *Geographic data mining and knowledge discovery*. London: Taylor and Francis, 2001.

HAN, J. y KAMBER, M. *Data mining: Concepts and techniques*. 7th. ed. Morgan Kaufmann, 2006.

HARMS, S. K.; DEOGUN, J. y GODDARD, S. Building knowledge discovery into a geo-spatial decision support system. *Proceedings of the 2003 ACM symposium on Applied Computing*, 2003, pp. 445-449.

HERNÁNDEZ, J. *et al*. Parte 3: Técnica de minería de datos. En *Introducción a la minería de datos*. New York: Pearson Prentice Hall, 2007, pp. 281-351.

HSU, W.; LI, M. y WANG, J. Parte 1: Spatial data mining introduction. *Temporal and spatio-temporal data mining*. Hershey: Igi Publishing, 2008, pp. 1-10.

INMON, W. H. Parte 2. The data warehouse environment. En *Building data warehouse*. 4th ed. Indianapolis: Wiley, 2005, pp. 9-46.

KURGAN, L. y MUSILEK, P. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*. 2006, vol. 21, núm. 1, pp. 1-24.

LUTU, P. An integrated approach for scaling up classification and prediction algorithms for data mining. *ACM International Conference Proceeding Series*, 2002, vol. 30, pp. 110-117.

MALINOWSKI, E. y ZIMÁNKY, E. Parte 2. Introduction to databases and data warehouse. En *Advanced data warehouse design: from conventional to spatial and temporal applications*. Berlin: Springer, 2008, pp. 16-51.

—. Representing spatiality in a conceptual multidimensional model. *Proceedings of the 12th ACM Int. Symp. on Advances in Geographic Information Systems, ACM GIS*, 2004, pp. 12-21.

MAN, L. Y. y NIKOS, M. Clustering objects on a spatial network. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, Paris, 2004.

MARCANO, A.; YELITZA, J. y TALAVERA, R. Minería de datos como soporte a la toma de decisiones empresariales. *Revista de Ciencias Humanas y Sociales*, 2006, núm. 52, pp. 104-118.

MARTÍN SÁNCHEZ, F.; LÓPEZ CAMPOS, G. y MAOJO GARCÍA, V. Bioinformática y salud: impactos de la aplicación de las nuevas tecnologías para el tratamiento de la información genética en la investigación biomédica y la práctica clínica. *Informática y Salud*. 1999 [web en línea]. <http://www.seis.es/i_s/i_s19/i_s19l.htm>. [Consulta: 22-03-2008].

MARTIN, E.; KRIEGEL, H.-P. y SANDER, J. Algorithms and applications for spatial data mining. In MILLER, H. y HAN, J. *Geographic data mining and knowledge discovery*. London: Taylor & Francis, 2001. pp. 1-10.

MARTÍNEZ DE PISÓN, F. *et al*. Minería de datos en series temporales para la búsqueda de conocimiento oculto en históricos de procesos industriales. *I Congreso Español de Informática,* Granada, 13-16 de septiembre de 2005.

MATIAS, R. y MOURA-PIRES, J. *Spatial On-Line Analytical Processing (SOLAP): A tool the to analyze the emission of pollutants in industrial installations*. 2005 [documento en línea]. <http://centria.fct.unl.pt/~jmp/page11/page14/files/EPIA05-RM-JMP.pdf>. [Consulta: 30-03-2008].

MENNIS, J. y LIU, J. W. Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. *Transactions in GIS*, 2005, pp. 1-17.

MORENO, M. *et al*. *Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software*. 2001 [documento en línea]. <http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf>. [Consulta: 25-03-2008]

MOSS, L. y ATRE, S. Guide to the development steps. En *Business intelligence roadmap: the complete project lifecycle for decision-support applications*. New York: Addison Wesley, 2003. 0-201-78420-3.

OLMO, J. y MARTÍNEZ, T. Aplicación de la teoría de variables regionalizadas en la investigación de marketing. *Revista Europea de Dirección y Economía de la Empresa*. 1992, vol. 1, núm. 1, pp. 125-132.

OLSON, D. y DENLE, D. Parte 2. Data mining methods and tools. *Técnicas avanzadas de minería de datos*. Berlin: Springer, 2008. pp. 39-144.

PING-YU, H.; YEN-LIANG, C. y CHUN-CHING, L. Algorithms for mining association rules in bag databases. *Information Sciences*. 2004, vol. 166, pp. 31-35.

RAYMOND T. N. y HAN, J. CLARANS: a method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*. 2002, vol. 14, núm. 5, pp. 1004-1006.

—. Efficient and effective clustering methods for spatial data mining. *Proceedings of the 20th International Conference on Very Large Data Bases*: September 12-15, 1994.

REINSCHMIDT, J. y FRANCOISE, A. *Business intelligence certification guide* [Libro en línea]. IBM Redbooks, 2000. <http://www.redbooks.ibm.com/abstracts/sg245747.html>. [Consulta: 25-03-2008]. ISBN e-book 0738415111.

RIVEST, S. *et al*. *SOLAP technology: merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data*. 2005 [Documento en línea]. <http://sirs.scg.ulaval.ca/Yvanbedard/article_nonprotege/400.pdf>. [Consulta: 02-04-08].

—. *SOLAP: a new type of user interface to support spatio-temporal multidimensional data exploration and analysis*. 2003 [documento en línea]. <http://sirs.scg.ulaval.ca/Yvanbedard/article_nonprotege/344.pdf>. [Consulta: 07-04-08].

RODDICK, J. y LEES, B. Paradigms for spatial and spatio-temporal data mining. In MILLER, H. y HAN, J. *Geographic data mining and knowledge discovery*. London: Taylor & Francis, 2001.

RODDICK, J. F. y SPILIOPOULOU, M. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and data engineering*. 2002, vol.14, núm. 4, pp. 750-767.

SÁNCHEZ, D.; MIRANDA, M. y CERDA, L. Reglas de asociación aplicadas a la detección de fraude con tarjetas de crédito. En *Actas del XII Congreso Español sobre Tecnologías y Lógica Fuzzy*, Jaén, 15-17 de septiembre de 2004.

SANTOS, M. y AMARAL, L. *Knowledge discovery in spatial databases: the qualitative approach*. 2000 [documento en línea]. <http://www.car.busmgt.ulst.ac.uk/papers/santos.pdf>. [Consulta: 10-04-2008].

SAS. *ESRI offer solutions for GIS and business intelligence needs*. 2004 [web en línea]. <http://www.sas.com/news/preleases/102004/news1.html>. [Consulta: 05-05-2008].

SHEKHAR, S.; ZHANG, P.; HUANG YAN, R. y VATSAVAI R. R. Trends in spatial data mining. En KARGUPTA, H. y JOSHI, A. (Eds.). *Data mining: next generation challenges and future directions*. AAAI/MIT Press, 2008, pp. 357-380.

SVETLOZAR, N. Mining qualified association rules in distributed databases. En *Workshop on data mining and exploration middleware for distributed and grid computing*. Minneapolis: University of Minnesota, 2003.

VYAS, R.; KUMAR, L. y TIWARY, U. Exploring spatial ARM (Spatial Association Rule Mining) for geo-decision support system. *Journal of Computer Science*. 2007, vol. 3, núm. 11, pp. 1-3.

WREMBEL, R. y KONCILIA, C. Parte 1. Modeling and designing. En *Data warehouse and OLAP: concepts, architectures and solutions*. New York: IRM Press, 2007, pp. 1-58.