Perspectivas en recuperación y explotación de información electrónica: el "Data Mining"

Jesús Tramullas Saz
Universidad de Zaragoza
Departamento de Ciencias de la Documentación
e Historia de las Ciencias

0.1. Resumen

Este trabajo presenta una panorámica del concepto y de las técnicas identificadas por el término "data mining". Explica los principios y fases a desarrollar en un proceso de este tipo, y expone los principales tipos de herramientas disponibles. (Autor)

Palabras clave: Data Mining (Minería de Datos). Recuperación de Información. Bases de Datos.

0.2. Abstract

Presents the concept and the techniques that "data mining" expression identifies. Explain the principles and phases to develop a data mining process, and expose the main types of data mining tools. (Author)

Keywords: Data Mining. Data Warehousing. Information Retrieval. Data Bases.

1. Introducción: ¿recuperación de datos o recuperación de información?

La perspectiva tradicional de las Ciencias de la Información y la Documentación, ha favorecido la división entre "data retrieval" e "information retrieval", tal y como establecieron autores como Van Rijsbergen (1979) y Blair (1990). La primera de ellas se centraba en entornos en los que la información se encontraba altamente estructurada, como en las bases de datos relacionales, mientras que la segunda lidiaba con entornos muy poco estructurados, como por ejemplo, grandes cantidades de información textual. Teóricamente, las tareas de recuperación, en la primera de ellas, limitaban su dificultad al correcto uso de

perfeccionados, y altamente formalizados, lenguajes de recuperación, mientras que en la segunda influían numerosos factores intrínsecos y extrínsecos, como el lenguaje natural, los mecanismos de indización, etc. Las dificultades que entrañaba la recuperación de información ha generado gran cantidad de investigadores y publicaciones sobre herramientas y enfoques para explotar de forma consistente el conocimiento almacenado de forma poco estructurada, de los que los "agentes inteligentes" son una de las últimas, y más conocidas, manifestaciones.

Sin embargo, la evolución de los sistemas informáticos, más bien de las necesidades de las organizaciones que deben hacer frente a gran cantidad de datos y a un entorno altamente competitivo, han favorecido la aparición de nuevos enfoques en la explotación de los datos altamente estructurados. La gran cantidad que han ido acumulando en periodos de tiempo relativamente cortos, principalmente con finalidad contable o para estudios de mercadotecnia, pueden ser utilizados para predecir futuros comportamientos, identificar nuevas áreas de negocio, o detectar irregularidades de muy diverso tipo. Sin embargo, las técnicas tradicionales de recuperación de datos no permiten inducir nuevo conocimiento, sólo extraer el existente. Por lo tanto, es necesario desarrollar, probar y aplicar mecanismos que identifiquen esos patrones, de forma rápida y eficaz, entre los millones de registros de datos de todo tipo almacenados por la organización. De eso se encarga el "data mining", utilizando técnicas originadas por la inteligencia artificial en la década de 1980, y otras técnicas cuantitativas anteriores, muchas de las cuales tuvieron su campo de pruebas precisamente en la recuperación de información. Como primera aproximación, el lector interesado debería visitar los servidores web del Data Warehouse Resource center (URL : http://www.cio.com/CIO/rc_dw.html), y de The Data Mining Institute (URL : http://www.datamining.org).

2. Concepto de *data mining*

El término "data mining" ha sido ampliamente utilizado por las empresas informáticas para identificar a productos y aplicaciones que, de forma genérica, analizan grandes cantidades de datos, con la finalidad de encontrar patrones o principios entre ellos. Se trataría de un amplio conjunto de técnicas utilizadas mediante una aproximación informática, cuya finalidad sería explorar y descubrir relaciones complejas en grandes conjuntos de datos (Moxon, 1996). Estos conjuntos de datos se estructuran de forma tabular, es decir, han sido objeto de una organización y estructuración previa, adoptando normalmente la forma de estructuras relacionales, utilizando para su implementación sistemas de gestión de bases de datos relaciones, y, como mecanismos de recuperación, lenguajes estructurados de forma rígida, como QBE (Query By Example) o SQL (Structured Query Language), principalmente este último.

El significado más aproximado de "data mining" sería el de "minado o extracción de datos". Por lo tanto, nos encontramos ante una metáfora, en el sentido de considerar a las grandes bases de datos como minas, dentro de las cuales es necesario localizar los filones que contienen los materiales preciosos (de hecho, en data mining se habla a menudo de *nuggets*, o pepitas de oro u otro metal precioso). Durante esa actividad de *mining* se localizan las *nuggets*, pero también los patrones que permiten localizar esas valiosas pepitas. Sin embargo, esta profusión del término ha obligado a establecer ciertas características intrínsecas que permiten diferenciar una aplicación para "data mining", de aplicaciones OLAP (*on-line analytical processing*, también llamadas de análisis multidimensional), y de aplicaciones DSS (*decision support system*). La principal diferencia entre ellas se encuentra en la aproximación al problema: mientras que las aplicaciones OLAP y DSS se centran en comprobar la validez de hipótesis establecidas previamente por el usuario, ya que son herramientas analíticas, las aplicaciones para data mining se dedican a explorar para descubrir, utilizando diferentes tipos de algoritmos para determinar las relaciones claves existentes entre los datos, que se encuentran ocultas al usuario, y las reglas que las rigen: "Data mining tools find patterns in the data and infer rules from them." (Edelstein, 1996). The Gartner Group ha definido *data mining* de esta forma:

> "Data mining is the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques." (Krivda, 1996)

La idea clave que preside las técnicas de *data mining* es la de descubrimiento (Kloesgen y Zytkow, sin fecha). Este descubrimiento es un proceso que busca aumentar el conocimiento sobre un campo o dominio dado, usando un método con unas tareas definidas, principalmente relacionadas con proceso de búsqueda. El conocimiento se adquiere mediante la obtención de patronos (obtención que puede conseguirse con diferentes técnicas, como métodos de extracción de reglas, de dependencias funcionales o estadísticos…), y mediante la evaluación de la validez de los patrones obtenidos. Por lo tanto, *data mining* es un proceso de descubrimiento cuantitativo y cualitativo de patrones ocultos en grandes cantidades de datos formalizados y organizados, para poder establecer reglas formales tomando como punto de partida los patrones identificados.

3. Herramientas y técnicas

Según Moxon (1996), las aplicaciones de *data mining* pueden describirse como una arquitectura de tres niveles, correspondientes a aplicaciones, aproximaciones, y algoritmos y modelos, arquitectura que se sitúa por encima de los repositorios de datos.

3.1. Aplicaciones

Las aplicaciones pueden clasificarse según el problema al que hagan frente, y que pueden tener las mismas características, sin perjuicio de que sean aplicadas en diferentes dominios. La adecuación de cada aplicación a un dominio dado se realizará mediante una parametrización específica, que refleje las características del dominio y del modelo.

3.2. Aproximaciones

Cada clase de aplicación se caracteriza por la utilización de un conjunto dado de algoritmos, que utiliza para extraer las relaciones relevantes existentes en los datos. Los cuatro grupos principales son los algoritmos de asociación, el análisis basado en secuencias, los análisis cluster, la clasificación y la estimación (Simoudis, 1996). Los algoritmos de asociación intentan identificar patrones de relación entre los datos, de forma que se puedan definir conjuntos. El análisis basado en secuencias se basa en localizar e identificar patrones a lo largo de una secuencia temporal, con la finalidad de poder establecer comportamientos en un tracto espacio-temporal. Los análisis *cluster* se emplean para identificar grupos de datos similares, que puedan utilizarse como punto de partida para otros análisis posteriores. Las técnicas de clasificación es una de las aproximaciones más aplicadas, y utiliza patrones preconstruidos para clasificar los registros analizados dentro de uno u otro grupo. Por último, la estimación añade a la clasificación la posibilidad de utilizar diferentes escalas o dimensiones dentro de los grupos.

3.3. Algoritmos y modelos

Los algoritmos aplicados dan como resultado la construcción de modelos, que pueden obtenerse, de la misma forma, por combinación de varios algoritmos. Cuatro son las herramientas principales utilizadas: redes neuronales, árboles de decisión, inducción de reglas y visualización de datos (Edelstein, 1996).

- Redes neuronales: colección de nodos conectados, con flujos de entrada y de salida, y procesos en cada nodo. Estas redes son capaces de aprender, y de aplicar lo aprendido a otros conjuntos de datos. Los resultados son difíciles de interpretar, por los que algunos sistemas añaden nuevos algoritmos que traducen los resultados a reglas.
- Árboles de decisión: dividen los datos en grupos jerárquicos, tomando como criterio el valor de determinadas variables. El valor de sus resultados depende de los tipos de datos.
- Inducción de reglas: crea grupos no jerárquicos, que pueden superponerse.
- Visualización de datos: en realidad, no es un modelo, ya que sólo muestra imágenes que resumen los resultados obtenidos. Sin embargo, la capacidad de estas visualizaciones para mostrar de forma sencilla gran cantidad de

información, pro ejemplo con más de cuatro variables.

4. El proceso de data mining

La aplicación de un proceso de data mining sobre un conjunto de datos es una tarea compleja, que debe estar relacionada con la organización en la que se aplique. En primer lugar, se trata de un trabajo que tiene como objetivo identificar y utilizar información oculta. Para asegurar el éxito del proceso es necesario cumplir tres condiciones principales (Simoudis, 1996). En primer lugar, el conjunto de datos debe integrarse en una visión completa de la organización; en segundo lugar, los datos deben ser "minados", y, por último, la información obtenida debe ser organizada y presentada de forma que facilite la toma de decisiones complejas.

Para ofrecer soporte a la toma de decisiones, es necesario transformar los datos en información elaborada. El proceso de data mining posee cuatro fases:

- Selección: consiste en seleccionar los tipos de datos que van a ser utilizados. No todos los datos existentes en un almacén pueden ser pertinentes para una actuación de este tipo. Por lo tanto, la selección de los más adecuados agiliza el proceso y puede servir como mecanismo previo de filtrado, evitando errores de desviación.
- Transformación: una vez seleccionados, suele ser necesario transformar los datos elegidos, cambiando su forma o presentación, o derivando nuevas categorías de la existentes.
- Mining o minado: es el proceso de minado en sí mismo, durante el cual se aplican diferentes técnicas al conjunto de datos. Puede ser necesario realizar algún ajuste adicional de los datos durante esta fase.
- Interpretación: la información obtenida se analiza en términos del objetivo fijado. El propósito de este análisis no se reduce a establecer una visualización o presentación adecuada: también se trata de filtrar y validar la información obtenida.

El punto 3, correspondiente específicamente al proceso de data mining, consiste, a su vez, en cuatro operaciones, que las aplicaciones de data mining suelen utilizar de forma combinada (Gerber, 1996; Brethenoux, 1996):

- Modelado predictivo: se trata de analizar el conjunto de datos seleccionados para generar automáticamente un modelo que pueda explicar un comportamiento. Para esta tarea se utilizan técnicas de razonamiento inductivo, como por ejemplo redes neuronales.

- Segmentación de la base de datos: el objetivo es identificar y separar grupos de registros que posean un comportamiento o características similares, de forma que se pueda resumir el contenido de la base de datos.
- Análisis de enlaces: se trata de establecer la existencia de relaciones o conexiones relevantes entre registros dentro de la propia base de datos.
- Detección de desviaciones: es la operación opuesta a la segmentación, e identifica a los registro que no pertenecen a ningún grupo, para establecer una explicación.
- Software: las aplicaciones informáticas.

Las aplicaciones y programas informáticos utilizados en los procesos de minado de datos reciben un nombre especial que los identifica: se les llama genéricamente "siftware". Bajo este paraguas se pueden encontrar a menudo sistemas DSS y OLAP, por lo que es necesaria una revisión de los métodos y algoritmos que utilizan para identificar claramente a las verdaderas aplicaciones de *data mining.*

El creciente interés por parte de las organizaciones en la explotación del *data mining* se ve reflejado en la existencia en el mercado de gran cantidad de aplicaciones comerciales que ofrecen cobertura a estas tareas. Sólo en el mercado norteamericano pueden encontrarse más de treinta empresas que comercializan aplicaciones de este tipo, y más de 50 aplicaciones diferentes (Hall, 1996). A ellas hay que añadir la gran cantidad de aplicaciones experimentales desarrolladas a través de proyectos de investigación por universidades, laboratorios u otras entidades. Por lo general, estas herramientas se ejecutan en entornos con sistemas operativos UNIX, aunque cada vez es mayor el número que ofrece también versiones para sistemas Windows, en cualquiera de sus versiones. El usuario puede

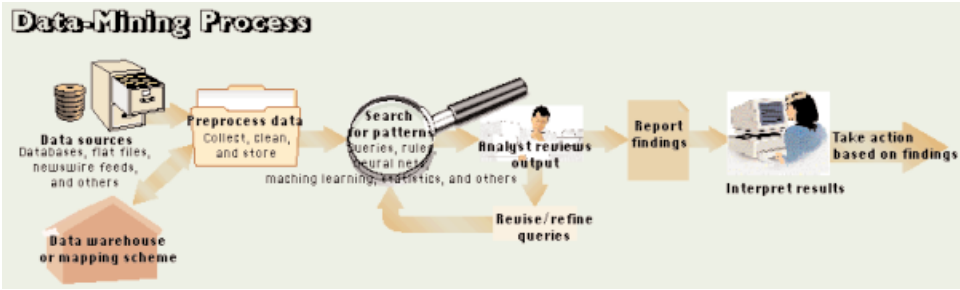

Fig.21 El proceso de minería de datos (tomado de Byte, 1996).

encontrar completos listados y referencias actualizadas de los productos siftware existentes en el mercado en Data Warehousing Information Center (URL : http://pwp.starnetinc.com/larryg/datamine.html), y en KD Mine: Data Mining and Knowledge Discovery Resources Index (URL : http://info.gte.com/~kdd/index.html).

El conjunto de aplicaciones puede englobarse en ocho grupos principales, según las técnicas de descubrimiento que utilizan para analizar los datos y/o para construir los modelos predictivos, sin perjuicio de que una aplicación puede incorporar varias técnicas de descubrimiento. Los ocho grandes grupos corresponden a (Hall, 1996, p.2-3):

- Reglas y árboles de decisión.
- Redes neuronales.
- Estadística convencional.
- Técnicas avanzadas de visualización.
- Técnicas difusas.
- Sistemas basados en el conocimiento.
- Múltiples.

Las herramientas más comunes y extendidas se engloban en el primer grupo. Sin embargo, las basadas en redes neuronales están ganando terreno, por su capacidad de manejar mejor datos incompletos o inconsistentes. Las basadas en estadística convencional son de uso común, y las de visualización de datos han ido incorporando los desarrollos de métodos específicos derivados de la ingeniería y similares. Más escasas son las basadas en lógica difusa y en SBC, por la necesidad de incorporar personal humano altamente especializado. Por último, los sistemas múltiples son los más conocidos, debiendo citarse como referencia obligada Intelligent Miner, de IBM, Darwin, de Thinking Machines, MinetSet, de SGI, y Clementine, de ISL.

4. Limitaciones y perspectivas

La compleja teoría en la que se asientan los métodos de data mining propugnan la utilización de un fuerte componente algoítmico, en detrimento, en numerosas ocasiones, de otros aspectos. Brevemente, se puede argumentar que la importancia concedida al análisis de los conjuntos de datos ha dado como resultado un "descuido" en los aspectos relacionados con la preparación de los mismos, y con la presentación de resultados. Por esta razón se hace necesario un significativo trabajo de preprocesado y de postprocesado de los datos seleccionando cuidadosamente los conjuntos de datos a explorar y los mecanismos de representación de las respuestas, para obviar esta dificultad. Las cuestiones claves resul-

tan ser la susceptibilidad a datos sucios, ya que estas herramientas no incorporan modelos de datos de alto nivel; la incapacidad de explicar los resultados de forma inteligible, ya que la respuesta suele adoptar la forma de indicadores numéricos o gráficos excesivamente crípticos; y la laguna en la representación de los datos, ya que gran parte de las herramientas no son capaces de manipular estructuras relacionales, que deben convertir a ficheros planos. Estas cuestiones ponen de manifiesto la importancia de los sistemas de "data warehousing", cuya revisión desborda totalmente los objetivos de este trabajo, como requisito previo a la utilización de data mining.

Otras limitaciones surgen del estado actual de la tecnología. Las actuales versiones de UNIX son de 32 bits, a la espera de la generalización de sistemas de 64 bits. Esto hace que en muchos de ellos el límite del fichero de tamaño a manipular sea de 2 Gb., por lo que se impone una forzosa selección de los datos a utilizar. Por otra parte, el tipo de procesamiento del fichero hace que no todas las herramientas puedan aprovecharse de las posibilidades de la computación en paralelo, lo que ralentiza notoriamente los cálculos. Frente a estas cuestiones, otros análisis indican que la tendencia será hacia la descentralización de las herramientas de *data mining*, actualmente privilegio de grandes computadoras, hacia los PC de sobremesa de la organización, que utilizan los conjuntos de datos disponibles mediante *data warehousing*, a través de las redes de área local, e incluso de Internet o sus sucesoras (Gerber, 1996).

Una importante cuestión a considerar es la forma en la que el diseño de la base de datos que contiene los datos objeto de estudio puede afectar los resultados obtenidos mediante las técnicas de *data mining*. Por ejemplo, Stiwell (1995) ha revisado las características del sistema RECON, de Lockheed Corporation, y las ha aplicado a un caso teórico, concluyendo que una cuestión crucial es el papel desempeñado por las técnicas de optimización .

Resulta de sumo interés la revisión de un reciente artículo de R.D. Small (1997) en el que se detallan los mitos que rodean al *data mining*. Situando el *data mining* en su justo término, concluye que no es la panacea en el descubrimiento de información, ni el último paso en la explotación de grandes bases de datos, sino una técnica actual, todavía experimental en gran parte de los casos, que será integrada progresivamente en las aplicaciones de bases de datos de cualquier campo.

7. Algunas aplicaciones a las Ciencias de la Información y la Documentación

El estudio detallado de las posibilidades abiertas con el desarrollo de las técnicas de *data mining* debe servir para recuperar una discusión clásica en el mun-

do de las ciencias de la información y la documentación. Se trata del eterno problema de la "recuperación de información" (*information retrieval*, IR), en entornos informáticos, y de la adecuación de los resultados a las necesidades del usuario. Información que sabemos mediatizada por multitud de factores, tanto externos al recurso de información electrónica, como inherentes a éste. En primer lugar, la aplicación de técnicas de *data mining* podría realizarse sobre el contenido de la ecuación de búsqueda tradicional, derivando de ésta modelos de contenido, más que la actual comprobación de existencia/condición. La utilización de estos modelos de contenido facilitaría la recuperación de aquellos documentos que, sin cumplir estrictamente la lógica representada en la ecuación, podrían, en cambio responder a la necesidad planteada. En segundo lugar, también podría aplicarse al estudio de la utilización de los descriptores en la construcción de bases de datos documentales, así como a los procesos de indización de las mismas, pudiendo derivar de ello mecanismos de corrección para la recuperación, ya que resulta evidente que la utilización de estos descriptores no es la misma a lo largo de todo el ciclo de vida del recurso de información electrónica. En tercer lugar, sería interesante confrontar estas técnicas con los resúmenes textuales incluidos en muchas bases de datos documentales, ya que se obtendrían nuevos conocimientos sobre el proceso de construcción de los mismos, así como de las estructuras informativas subyacentes, que a su vez podrían relacionarse con la investigación señalada en primer lugar.

En cualquier caso, las técnicas de *data mining* son una "nueva" forma de recuperar información, más que simples datos, cuyos principios son plenamente coincidentes con las áreas de interes del especialista en información. Los futuros productos informáticos que vayan saliendo al mercado deberían integrar un nivel cada vez mayor de "inteligencia" para facilitar el tratamiento y recuperación, y dentro de esa inteligencia deberían encontrarse, por lo menos, alguna de las técnicas de data mining.

8. Conclusiones

El *data mining* es la obtención de nueva información, que puede ser transformada en conocimiento en el contexto de una organización dada, tomando como punto de partida patrones y reglas ocultos en grandes cantidades de datos que reflejan las actividades de la organización que se trate. Esta adquisición se realiza mediante la aplicación a un conjunto seleccionado de datos de diferentes técnicas, como la clasificación, los árboles de decisión, las dependencias, etc.

El objeto básico de trabajo es el conjunto de datos que se reúne en las bases de datos de la organización. Por lo tanto, cobran especial importancia las cuestiones relacionadas con el diseño y mantenimiento de las mismas, lo que relaciona directamente al *data mining* con los sistemas de *data warehousing*. La expan-

sión cada vez mayor de éstos, los almacenes de datos, facilitará el desarrollo de mejores, más simples y más económicas aplicaciones para *data mining*. Las técnicas de descubrimiento y de aprendizaje inductivo de estos sistemas se incorporan a otras aplicaciones informáticas de diverso tipo, con lo que se pueden augurar nuevas generaciones de herramientas informáticas capaces de aprender de sus usuarios.

El *data mining*, sus aplicaciones y sus técnicas son una técnica de mercado prometedor, pero aún inmaduro. Moxon (1996) ha apuntado que los principios y técnicas del *data mining* podrán aplicarse a otras representaciones de datos, como datos espaciales, textuales, imágenes y multimedia. Además, su principal objeto de trabajo son las bases de datos relacionales, por lo que se plantea el problema de desarrollar herramientas que sean capaces de utilizar las nuevas bases de datos multimedia orientadas a objetos, o los sistemas hipertextuales. Desde este enfoque, métodos y técnicas como ésta, y otras futuras que puedan surgir, resultarán de suma importancia para los especialistas en información y documentación, ya que proveerán de un valor añadido a sus productos, en cuanto se supera la mera "recuperación de datos" o "recuperación de información", para entrar en una nueva fase caracterizada por el descubrimiento de conocimiento a través de cualquier tipo de recuperación.

9. Referencias

Adriaans, P. ; Zantiage, D. (1996). Data Mining. Addison-Wesley, 1996.

Brethenoux, E. (1996). Buried Treasure. Some dos and don'ts of data mining. // CIO Magazine, (Oct. 1996). URL : http://www.cio.com/CIO/100196_gartner.html

Chen, M.S.; et alii (1996). Data Mining: An Overview from a Database Perspective. // IEEE Transactions on Knowledge and Data Engineering, 8 : 6 (Dec. 1996).

Dhar, V.; Stein, R. (1996). Seven Methods for Transforming Corporate Data Into Business Intelligence. Prentice Hall, 1996.

Edelstein, H. (1996). Technology How To: Mining Data Warehouses. // Information Week (Jan. 1996). URL : http://techweb.cmp.com/iw/561/61oldat.htm

Fayyad, U.M.; (et al.) (1995). Advances in Knowledge Discovery and Data Mining. MIT Press, 1995.

Fayyad, U.M.; Piatetsky-Shapiro, G.; Padhraic, S. (1996). Data Mining and Knowledge Discovery in Databases: An overview. // Communications of the ACM. 39: 11.

Gardner, C. (1996). Data Mining Technology. (April 1996). URL : http://booksrv2.raleigh.ibm.com/cgi_bin/bookmgr.cmd/BOOKS/datamine/

Gerber, C. (1996). Excavate your Data. // Datamation (May 1996). URL : http://www.datamation.com/PlugIn/issues/1996/may1/05asoft3frame.html

Hall, C. (1996). Data Mining Tools. // Data Management Series. 1 (1996).

Hedberg, S. (1996). Searching for the mother lode: tales of the first data miners. // IEEE Expert Special issue on data mining. (October 1996). URL : http://www.computer.org/pubs/expert/1996/insight/x5004/x5004.htm

Hodel, A. (1995). Data Mining: A New Weapon for Competitive Advantage. // IBM Software Quaterly. 24 (1995). URL : http://www.software.ibm.com/sq/issues/vol24/data.htm

Imielinski, T.; Mannila, H. (1996). A Database Perspective on KDD. // Communications of the ACM. 39 : 11 (1996).

Kloesgen, W.; Zytkow, J. (sin fecha). Machine Discovery Terminology. URL : http://org-wis.gmd.de/projects/explora/terms.html

Krivda, C.D. (1996). Unearthing Underground Data. // LAN Magazin (1996). URL : http://www.lanmag.com/9605mine.htm

Moxon, B. (1996). Defining Data Mining. // DBMS Online DBMS Dara Warehouse Supplement (August 1996). URL : http://www.dbmsmag.com/9608d53.html

Simoudis, E. (1995). Data Mining: A Technology Comes of Age. // IBM Software Quaterly, 24 (1995). URL : http://www.software.ibm.com/sq/issues/vol24/data-tech.htm

Small, R.D. (1997). Debunking Data Mining Myths. // Information Week (January 1997). URL : http://techweb.cmp.com/iw/614/14oldat.htm

VV.AA.(1995). State of the Art Section on Data Mining. // BYTE (October 1995). URL : http://www.byte.com/art/9510/sec8/sec8.htm

Wu, X. (1995). Knowlege Acquisition from Databases. Norwood : Ablex, 1995.