Data Warehouse y Data Mining Aplicados al Estudio del Rendimiento Académico

D. L. LA RED MARTINEZ, J. C. ACOSTA, L. A. CUTRO

V. E. URIBE, A. R. RAMBO

Dpto. Informática. FACENA. Univ. Nac. del Nordeste
(3400) Corrientes. Argentina
E-mail: lrmdavid@exa.unne.edu.ar

31 de octubre de 2009

ABSTRACT

En la Sociedad de la Información y el Conocimiento (SIC) es imprescindible la aplicación de técnicas de Almacenes de Datos (Data Warehouses: DW) y de Minería de Datos (Data Mining: DM) basadas en clustering, entre otras, para la búsqueda de perfiles de los alumnos de la Asignatura Sistemas Operativos (SO) de la FACENA de la UNNE según su rendimiento académico, situación demográfica y socio económica, con el propósito de determinar a priori situaciones potenciales de éxito o de fracaso académico, lo cual permitiría encarar las medidas tendientes a minimizar los fracasos. El presente trabajo tiene por objetivo brindar una breve descripción de aspectos relacionados con el almacén de datos construido y algunos procesos de minería de datos desarrollados sobre el mismo.

Palabras Claves: Base de Datos, Almacén de Datos, Minería de Datos, Clustering, Cluster Demográfico.

1 INTRODUCCIÓN

En el contexto de la SIC (Joyanes Aguilar, 1997), (Bolaños Calvo, 2001), (Taquini, 2001), (Peiró, 2001) y a los efectos de la determinación de los perfiles característicos de los alumnos de SO de la FACENA de la UNNE, se ha construido un DW con información personal, académica, demográfica y socio económica de los alumnos y de su núcleo familiar, el cual se ha comenzado a explorar con técnicas de DM, presentándose en este trabajo algunos de los resultados obtenidos (aún preliminares y parciales). Este artículo se ha estructurado de la siguiente manera: primeramente se planteará el objetivo principal del mismo, luego se hará una muy breve revisión de los principales conceptos involucrados en cuanto a DW y DM, seguidamente se indicará brevemente el software utilizado, para continuar con la metodología seguida y la presentación de algunos resultados obtenidos, para finalizar con las conclusiones y líneas futuras de acción, los reconocimientos y las referencias.

2 OBJETIVO PRINCIPAL

El objetivo principal de este trabajo es encontrar perfiles de alumnos por medio de la aplicación de técnicas de DM a un DW con datos académicos, socio económico y demográfico correspondientes a alumnos de SO de la Licenciatura en Sistemas de Información (LSI) de la FACENA de la UNNE.

3 REVISIÓN CONCEPTUAL

Un DW es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales (Kubski, 2005), (Cutro, 2008). Es también un conjunto de datos integrados orientados a una materia, que varían con el tiempo, y que no son transitorios, los cuales soportan el proceso de toma de decisiones de una administración (Inmon, 1992), (Inmon, 1996), (Simon, 1997), (Trujillo, Palomar & Gómez, 2000).

La DM es la etapa de descubrimiento en el proceso de KDD (Knowledge Discovery from Databases), es el paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados (Fayyad, Grinstein & Wierse, 2001), (Fayyad, Piatetskiy-Shapiro, Smith, & Ramasasmy, 1996), (Han & Kamber, 2001), (Hand, Mannila & Smyth, 2000).

Es también un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos. Está muy ligada a los DW ya que los mismos proporcionan la información histórica con la cual los algoritmos de minería obtienen la información

necesaria para la toma de decisiones (Gutiérrez, 2001), (IBM Software Group, 2003).

La DM es un conjunto de técnica de análisis de datos que permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y extraer patrones y tendencias para predecir comportamientos futuros (Simon, 1997), (Berson & Smith, 1997), (Frawley, Piatetsky-Shapiro & Matheus, 1992), (White, 2001).

En la figura 1 se muestra la "arquitectura" de un DW.

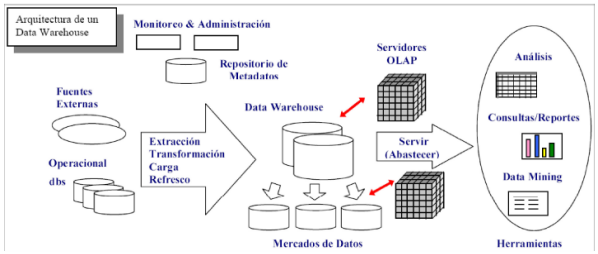

Figura 1: Arquitectura de un Data Warehouse.

4 SOFTWARE UTILIZADO

Se ha utilizado el IBM Data Warehouse Edition (DWE) V.9.5, que incluye al DB2 Enterprise Server Edition (DB2 ESE), al Design Studio (DS) y al Intelligent Miner (IM).

En la figura 2 se muestra el Centro de Control del DB2 ESE.

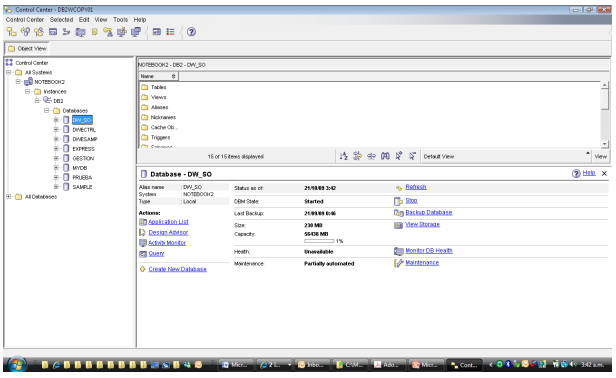

Figura 2: Centro de Control del DB2 ESE.

En la figura 3 se presenta una pantalla del Design Studio.

5 METODOLOGÍA SEGUIDA

El presente estudio se realizó sobre datos obtenidos mediante encuestas realizadas al alumnado de SO, considerando además los resultados de las distintas instancias de evaluación previstas durante el cursado de dicha

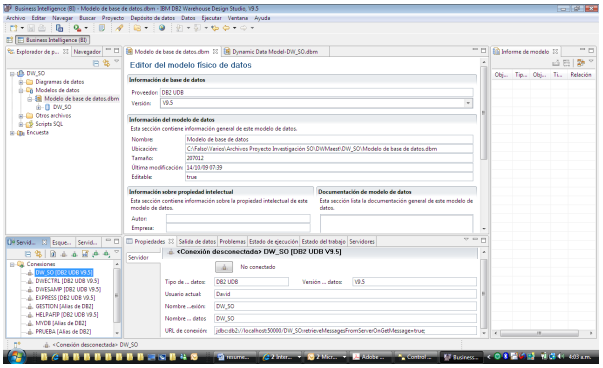

Figura 3: Pantalla del Design Studio.

asignatura. Se utilizó un entorno integrado de gestión de bases de datos y data warehouse (DB2 versión 9.5), obtenido de la empresa IBM mediante los Acuerdos firmados entre dicha empresa y la UNNE; dicho entorno permite la extracción de conocimiento en bases de datos y DW mediante técnicas de DM como ser clustering (o agrupamiento de datos) que consiste en la partición de un conjunto de individuos en subconjuntos lo más homogéneos posible, el objetivo es maximizar la similitud de los individuos del cluster y maximizar la diferencia entre clusters. El cluster demográfico es un algoritmo desarrollado por IBM e implementado en el IM, componente del DWE, entorno antes mencionado, que resuelve automáticamente los problemas de definición de métricas de distancia / similitud, proporcionando criterios para definir una segmentación óptima (Grabmeier, & Rudolph, 1998), (Baragoin, Chan, Gottschalk, Meyer, Pereira & Verhees, 2002), (Ballard, Rollins, Ramos, Perkins, Hale, Dorneich, Cas Milner & Chodagam, 2007), Ballard, Beaton, Chiou, Chodagam, Lowry, Perkins, Phillips & Rollins, 2006).

Los pasos realizados durante el presente trabajo han sido los siguientes:

- Recolección de los datos.
- Tratamiento y depuración de los datos.
- Preparación de la base de datos y del DW correspondiente sobre la plataforma de trabajo seleccionada.
- Selección de la técnica de minería de datos para la realización del estudio (predominantemente clustering).
- Generación de diferentes gráficos para el estudio de los resultados.
- Estudio de los resultados obtenidos.
- Obtención de las conclusiones.

En esta etapa se trabajó con una porción (Data Mart: DM) del DW, cuya estructura se muestra en la figura 4.

6 RESULTADOS OBTENIDOS

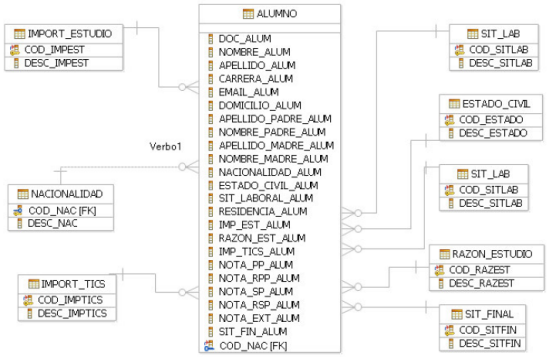

Figura 4: Estructura del DM utilizado, parte del DW.

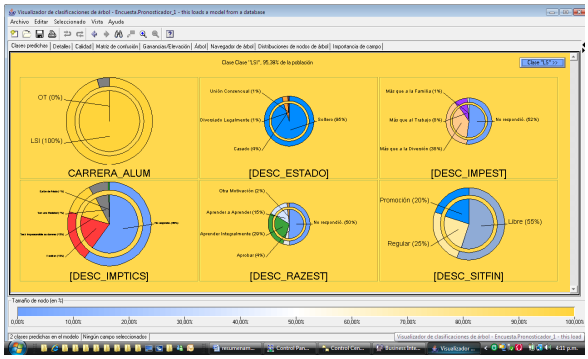

Figura 5: Licenciatura en Sistemas de Información.

Se obtuvieron diferentes clasificaciones mediante la utilización (preferentemente) de técnicas de clustering, según diferentes criterios de agrupación de los datos.

Se utilizó la siguiente equivalencia de nombres y significados de variables:

- SIT_LABORAL_ALUM: Situación laboral del alumno.
- IMP_EST_ALUM: Importancia dada al estudio por el alumno.
- RAZON_EST_ALUM: Razón para estudiar según el alumno.
- IMP_TICS_ALUM: Importancia dada a las TICs por el alumno.
- NOTA_PP_ALUM: Nota primer parcial.
- NOTA_RPP_ALUM: Nota recuperatorio primer parcial.
- NOTA_SP_ALUM: Nota segundo parcial.
- NOTA_RSP_ALUM: Nota recuperatorio segundo parcial.
- NOTA_EXT_ALUM: Nota recuperatorio extraordinario.
- SIT_FIN_ALUM: Situación final del alumno luego del cursado.

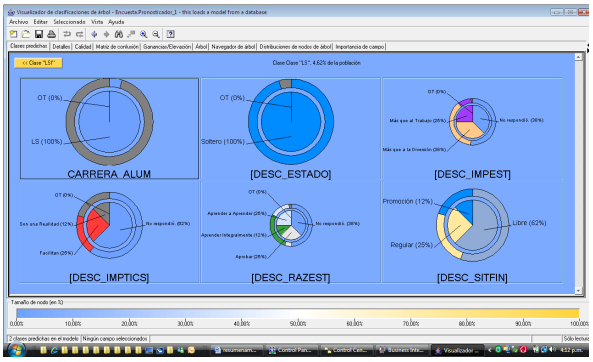

Figura 6: Licenciatura en Sistemas.

Seguidamente se muestran algunos de los resultados obtenidos.

Minería de Clasificación según Carrera: figuras 5, 6.

Miniería de Clasificación según Situación Final del Alumno: figuras 7, 8.

Se considera Libre al alumno que no ha cumplimentado la aprobación de los trabajos prácticos y de laboratorio, Regular a quien ha cumplimentado dichas exigencias pero con un promedio inferior a 7 en la escala 0-10,

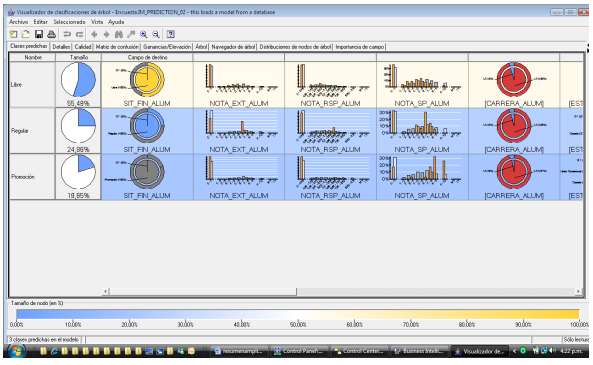

Figura 7: Situación final del alumno1.

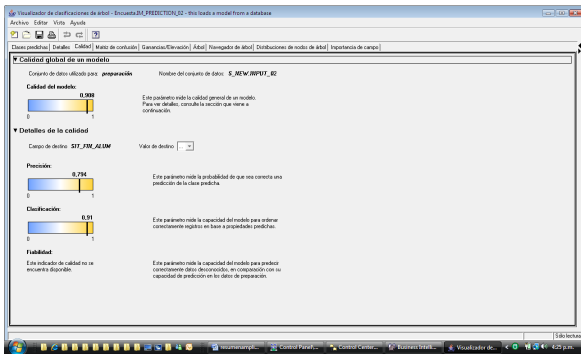

Figura 8: Situación final del alumno - calidad del modelo.

finalmente se considera Promoción a quien ha cumplimentado las exigencias con un promedio igual o superior a 7.

Minería de Clasificación según Importancia Dada al Estudio: figuras 9, 10.

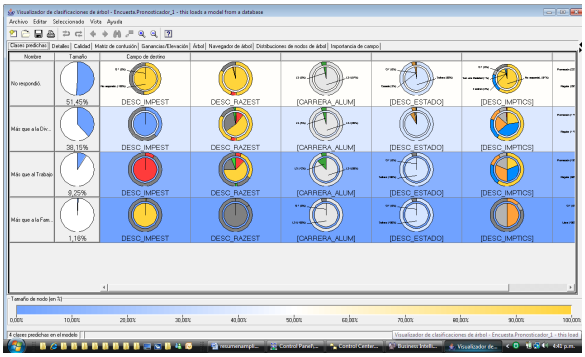

Figura 9: Importancia dada al estudio - 1.

Minería de Clustering Demográfico según Situación Final del Alumno como variable principal: figuras 11, 12, 13.

Minería de Clustering de Kohonen según Situación Final del Alumno como variable principal: figuras 14, 15, 16.

7 CONCLUSIONES Y LÍNEAS FUTURAS

Se obtuvieron diversos modelos de minería de datos referidos a diversos aspectos de los alumnos de la asignatura mencionada, lo que permitió descubrir el perfil de dichos estudiantes, destacándose especialmente lo siguiente:

- Los libres son en su mayoría solteros, y en menor proporción divorciados, los libres indican en mayor porcentaje darle más importancia al estudio que a la diversión que los regulares y los promocionados.

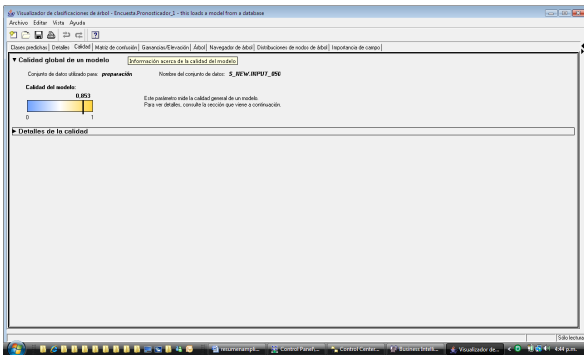

Figura 10: Importancia dada al estudio: calidad del modelo.

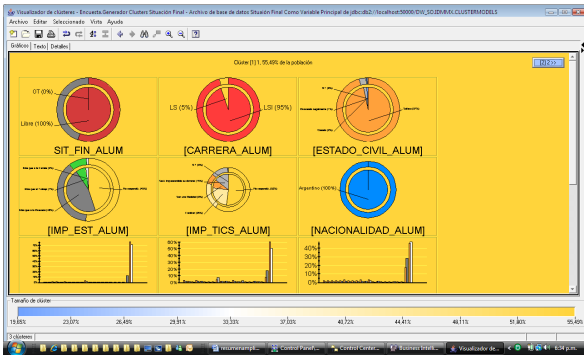

Figura 11: Situación final del alumno: Libre.

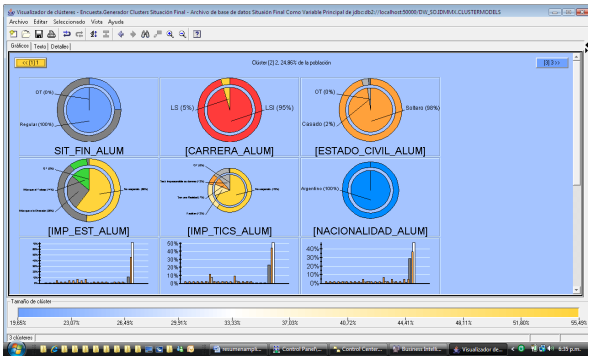

Figura 12: Situación final del alumno: Regular.

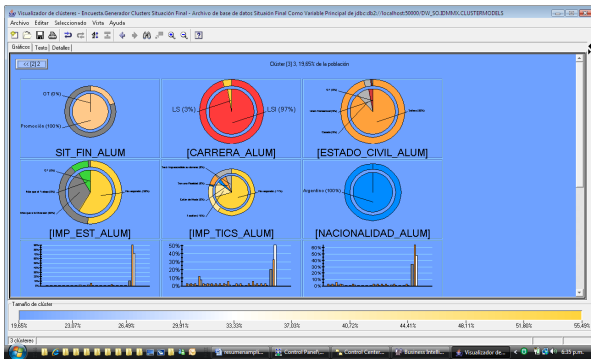

Figura 13: Situación final del alumno: Promoción.

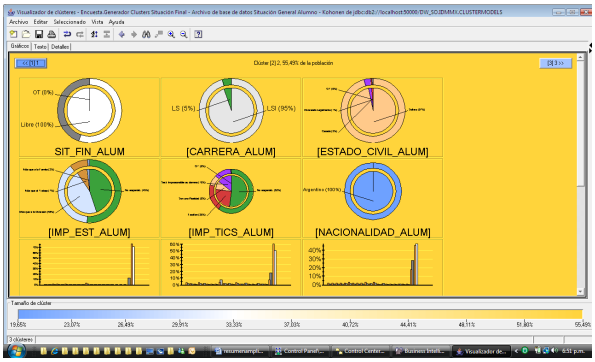

Figura 14: Situación final del alumno: Libre.

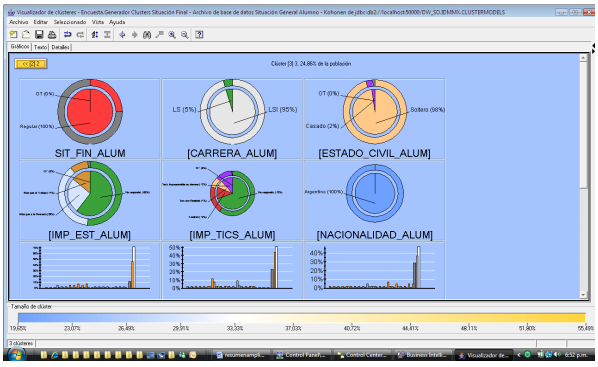

Figura 15: Situación final del alumno: Regular.

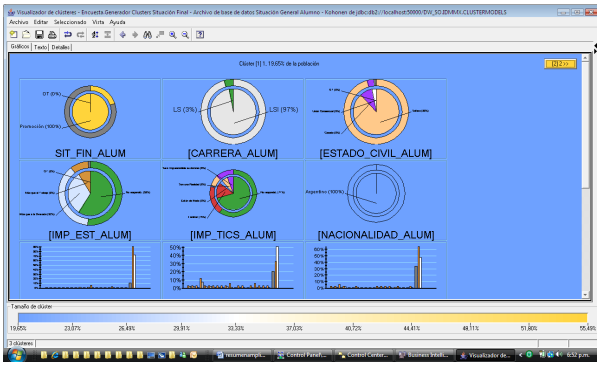

Figura 16: Situación final del alumno: Promoción.

Indican en un igual porcentaje que las TICs facilitan el estudio y que es importante su dominio.

- Los regulares son solteros y en unión consensual, le dan más importancia al estudio que a la diversión y en mayor porcentaje que los libres incluso que al trabajo.
- Entre los promocionados figuran algunos casados y solteros y aparecen quienes consideran que las TICs simplemente están de moda.
- Los tres grupos indican como razón para el estudio con mayor porcentaje la de aprender integralmente. Entre los libres figura con un mayor porcentaje que en los otros dos grupos la razón de estudiar para aprobar.
- La mayoría en sendos grupos no trabaja, pero el porcentaje de quienes trabajan entre los promocionados es mayor que entre los libres y los regulares.

Se ha podido comprobar las grandes ventajas de la utilización de tecnologías y software de última generación que soportan sistemas multiplataforma.

Se hace notar que los resultados logrados son sólo la etapa preliminar de los diversos estudios que se tiene previsto realizar, incorporando las demás variables del DW.

Se tiene previsto desarrollar las siguientes líneas futuras de acción:

- Avanzar en la investigación con la utilización de minería de datos como ser redes neuronales, redes bayesianas, arboles de decisión, etc., aplicadas al almacén de datos utilizado hasta ahora generalmente con las técnicas de clustering.
- Aplicar las técnicas de minería de datos utilizadas, pero sobre otras bases de datos de alumnos de otras asignaturas y carreras para comparar los resultados obtenidos.

9 RECONOCIMIENTOS

El presente trabajo se encuadra en el Proyecto de Investigación "El Desigual Aprovechamiento de las TICs en el Proceso de Enseñanza – Aprendizaje de los Sistemas Operativos en la FACENA de la UNNE", acreditado por la Secretaría de Ciencia y Técnica de la UNNE como PI-120-07 (Res. 369/08 CS).

El software utilizado, Data Warehouse Edition V.9.5, que incluye al DB2 Enterprise Server Edition, al Design Studio y al Intelligent Miner, se han obtenido de la empresa IBM Argentina S.A., en el marco de la Iniciativa Académica de dicha empresa y de los Acuerdos realizados entre la misma y la FACENA de la UNNE (Acuerdo del 18/06/04 D, Res. 1417/04 D, Res. 858/06 CD).

8 REFERENCIAS

Ballard, Ch.; Beaton, A.; Chiou, D.; Chodagam, J.; Lowry, M.; Perkins, A.; Phillips, R. & Rollins, J. (2006). Leveraging DB2 Data Warehouse Edition for Business Intelligence. IBM International Technical Support Organization. IBM Press. USA.

Ballard, Ch.; Rollins, J.; Ramos, J.; Perkins, A.; Hale, R.; Dorneich, A.; Cas Milner, E. & Chodagam, J. (2007). Dynamic Warehousing: Data Mining Made Easy. IBM International Technical Support Organization. IBM Press. USA.

Baragoin, C.; Chan, R.; Gottschalk, H.; Meyer, G.; Pereira, P. & Verhees, J. (2002). IBM International Technical Support Organization Enhance Your Business Applications. Simple Integration of Advanced Data Mining Functions. IBM Press.

Berson, A. & Smith, S. J. (1997). Data Warehouse, Data Mining & OLAP. Mc Graw Hill. USA.

Bolaños Calvo, B. (2001). Las Nuevas Tecnologías y los Desafíos Teórico – Prácticos en los Sistemas de Educación a Distancia: Caso UNED de Costa Rica. Temática: Universidades Virtuales y Centros de Educación a Distancia. UNED. Costa Rica.

Cutro, A. (2008). Minería de Datos Aplicada a la Encuesta Permanente de Hogares. Trabajo Final de Aplicación de la Licenciatura en Sistemas de Información dirigido por el Prof. David Luis la Red Martínez. Corrientes. Argentina.

Fayyad, U.M.; Grinstein, G. & Wierse, A. (2001). Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann. Harcourt Intl.

Fayyad, U.M.; Piatetskiy-Shapiro, G.; Smith, P.; Ramasasmy, U. (1996). Advances in Knowledge Discovery and Data Mining. AAAI Press / MIT Press. USA.

Frawley, W. J.; Piatetsky-Shapiro, G & Matheus, Ch. J. (1992). Knowledge Discovery in Database An Overview. Al Magazine.

Grabmeier, J. & Rudolph, A. (1998). Techniques of Cluster Algorithms in Data Mining version 2.0. IBM Deutschland Informationssysteme GmbH. GBIS (Global Business Intelligence Solutions). Germany.

Gutiérrez, J. M. (2001). Data Mining, Extracción de Conocimiento en Grandes Bases de Datos. España.

Han, J. & Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hand, D.J.; Mannila, H. & Smyth, P. (2000). Principles of Data Mining. The MIT Press. USA.

IBM Software Group. (2003). Enterprise Data Warehousing whit DB2: The 10 Terabyte TPC-H Benchmark. IBM Press. USA.

Inmon, W. H. (1992). Data Warehouse Performance. John Wiley & Sons. USA.

Inmon, W. H. (1996). Building the Data Warehouse. John Wiley & Sons. USA.

Joyanes Aguilar, L. (1997). Cibersociedad. Mc Graw Hill. España.

Kubski, M. (2005). Aplicación Orientada al Descubrimiento de Conocimiento en Bases de Datos. Trabajo Final de Aplicación de la Licenciatura en Sistemas de Información dirigido por el Prof. David Luis la Red Martínez. Corrientes. Argentina.

Peiró, J. M. (2001). Las competencias en la sociedad de la información: nuevos modelos formativos. Centro Virtual Cervantes. España.

Simon, A. (1997). Data Warehouse, Data Mining and OLAP. John Wiley & Sons. USA.

Taquini, A. C. (h). (2001). Educación Superior y Ciberespacio.

Trujillo, J. C., Palomar M. & Gómez, J. (2000). Applying Object-Oriented Conceptual Modeling Techniques To The Design of Multidimensional Databases and OLAP Applications. First International Conference On Web-Age Information Management (WAIM'00). Lecture Notes in Computer Science 1846:83-94.

White, C. J. (2001). IBM Enterprise Analytics for the Intelligent e-Business. IBM Press. USA.