Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD

Karina Eckert, Roberto Suénaga
Universidad Gastón Dachary
Salta 1968, Posadas, Misiones-Argentina. Tel: +54 (0376) - 4438677
karinaeck@gmail.com, rsuenaga@ugd.edu.ar

Resumen

En el ámbito educativo es evidente la necesidad de disponer de sistemas de gestión que permitan tomar decisiones académicas y elaborar estrategias a partir del conocimiento oportuno, ya que esto no solo incide directamente sobre la funcionalidad de los departamentos académicos, u otras cuestiones internas, sino que también podrían incidir sobre actividades como las evaluaciones y acreditaciones de instituciones y carreras. Entre los problemas más complejos que enfrentan las instituciones de educación podemos mencionar: mejorar la calidad académica, disminuir la deserción y la reprobación, evitar el atraso estudiantil y los bajos índices de eficiencia relacionado con las tasas de graduación. Esto requiere gestionar estrategias y tomar medidas frente a estos acontecimientos; para ello es posible recurrir al proceso denominado Minería de Datos Educacional (MDE), es decir, la aplicación del proceso de Descubrimiento o Extracción de Conocimiento en Bases de Datos (KDD) en ámbito educativo.

En el presente trabajo se describe y expone la aplicación del proceso KDD (por su siglas en inglés), conocido como Minería de Datos (MD) en un entorno educativo, más precisamente a la información académica de la Universidad Gastón Dachary (UGD). El proceso consiste en una serie de etapas que parten de la selección y captura de los datos, pasando por una serie de actividades relacionadas a la integración, recopilación y el filtrado de los mismos (pre-procesamiento), para luego ser procesados, analizados y evaluados hasta obtener conocimiento adicional. Para ello, es necesario llevar a cabo un proceso iterativo que incluye numerosas consultas de selección a la base de datos, depuración de los datos, utilización de diferentes criterios de representación; también se aplican diferentes técnicas y algoritmos de MD, tanto descriptivas como predictivas.

Palabras clave: Descubrimiento de Conocimiento en Bases de Datos (KDD), Minería de Datos Educativos (EDM), Rendimiento Académico, Herramientas de Minería de Datos.

Contexto

El presente proyecto de investigación se encuentra dentro de una línea de investigación iniciada con la Tesis de Grado titulada *"Explotación de Datos Académicos a través de la Aplicación de Técnicas de Minería de Datos en Weka",* de la carrera Ingeniería en Informática con Orientación a Sistemas de Información de la Universidad Gastón Dachary (UGD). La tesis derivó en una

postulación y adjudicación de beca de investigación financiado por el Comité Ejecutivo de Desarrollo e Innovación Tecnológica (CEDIT), bajo la denominación: "*Exploración de datos académicos para la determinación de causas de deserción universitaria a través de la aplicación de Técnicas de Minería de Datos"*.

Introducción

Actualmente la sociedad se encuentra en la denominada era de la información, donde día tras día se incrementa significativamente la cantidad de datos almacenados en diferentes fuentes, estructuras y formatos. Empero, contrariamente a lo se espera, esta expansión de datos no siempre supone un aumento de conocimiento, puesto que procesarlos con los métodos clásicos resulta ser en muchos casos imposible o sumamente tedioso y con resultados superficiales e insatisfactorios. De modo que nos enfrentamos actualmente a la paradoja de que, cuantos más datos están disponibles, menos información se tiene, y algo peor que no tener información disponible es tener mucha y no saber qué hacer realmente con ella. La clave evidentemente está en tener la información adecuada, en el lugar y momento oportuno, y así incrementar la efectividad de toda organización. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista, por lo que hay que "torturarlos hasta que ellos confiesen" [1], que es una explicación informal de la actividad que se realiza mediante la, denominada Minería de Datos (MD).

La Minería de Datos (MD), entre otras técnicas, utiliza Inteligencia Artificial para encontrar patrones y relaciones entre los datos, permitiendo la creación de modelos y representaciones abstractas de la realidad. La MD es una etapa dentro del proceso mayor llamado Descubrimiento de Conocimiento en Base de Datos (KDD), aunque en la mayoría de los entornos, ambos términos se usan de manera indistinta. El proceso completo de KDD incluye entre otras cosas, la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a los patrones identificados por las técnicas y algoritmos de MD. Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean [2]. KDD es el "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos", relevantes y nuevos sobre un fenómeno o actividad mediante algoritmos eficientes, dadas las crecientes órdenes de magnitud en los datos [3]. Las metas son: procesar automáticamente grandes cantidades de datos crudos, identificar los patrones más significativos y relevantes, presentarlos como conocimiento apropiado para satisfacer las demandas del usuario. El conocimiento se obtiene para llevar a cabo acciones, ya sea incorporándolo dentro de un sistema de desempeño o para almacenarlo y reportarlo a los interesados, en este sentido, implica un proceso interactivo e iterativo.

Lo que en verdad hace la MD es reunir las ventajas de varias disciplinas como la estadística, la inteligencia artificial, la computación gráfica, los sistemas de bases de datos y el procesamiento masivo, principalmente usando como materia prima las bases de datos de las organizaciones. La MD como etapa en el proceso de KDD es el "paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre-procesados" [3]. Se coleccionan los datos y se espera que de ellos emerjan hipótesis. Se busca que los datos describan o indiquen por qué son como son. Luego entonces, se valida esa hipótesis inspirada por los datos en

los datos mismos, será numéricamente significativa, pero experimentalmente inválida. De ahí que la MD debe presentar un enfoque exploratorio, y no confirmador. Usar la MD para confirmar las hipótesis formuladas puede ser peligroso, pues se está haciendo una inferencia poco válida [4] [5].

La deserción, el rezago estudiantil y los bajos índices de eficiencia terminal se encuentran entre los problemas más complejos y frecuentes que enfrentan las Instituciones de Educación Superior.

La Minería de Datos Educacional (EDM) es una disciplina emergente, preocupada por el desarrollo de métodos para explorar las características singulares de los datos que provienen de entornos educativos, y utilizar esos métodos para comprender mejor el desempeño de los estudiantes, y las condiciones en las cuales ellos aprenden [6][7][8]. Es el proceso de transformar los datos en bruto, recopilados por los sistemas de enseñanza, en información útil que pueda utilizarse para tomar decisiones y responder preguntas de investigación [8][9][10][11]. La aplicación de la MD en el ámbito de la enseñanza, tiene como objetivo obtener una mejor comprensión del proceso de aprendizaje de los estudiantes y de su participación global en el proceso, orientado a la mejora de la calidad y la eficiencia del sistema educativo [10][12]. A partir de toda la información disponible, las diferentes técnicas de MD pueden ser aplicadas a fin de descubrir conocimiento útil que ayude a mejorar el proceso educativo, siendo este conocimiento muy diverso. Está dirigida a los alumnos, profesores o autoridades académicas, quienes a partir de ella pueden identificar tres tipos de objetivos: pedagógicos (ayuda en el diseño de contenidos didácticos, mejoras en el rendimiento académico de los alumnos), de gestión (optimizar la organización y mantenimiento de infraestructuras educativas, áreas de interés, cursos más solicitados) y comerciales (permite realizar segmentación del mercado y facilita la captura de alumnos) [10].

Líneas de investigación y desarrollo

El trabajo de investigación sigue la estructura del Procedo KDD, que costa de cinco fases (Figura 1).

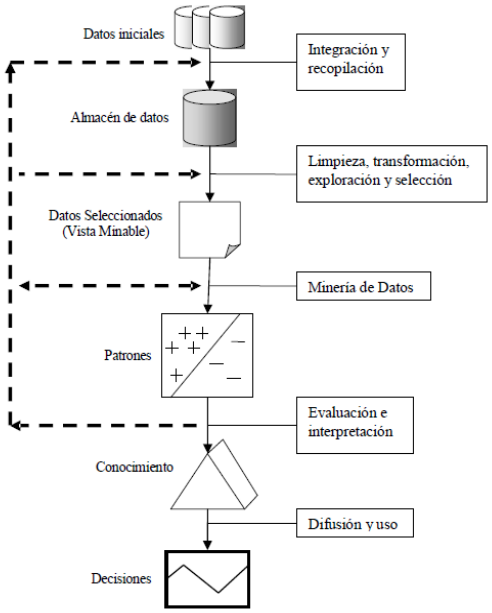

Figura 1: Fases del Proceso de KDD [13]

Durante el desarrollo del proceso de KDD, suele ser necesario interrumpir en algún punto de las fases del proceso y volver a comenzar en alguno de los pasos anteriores, siendo así un proceso *iterativo* e *interactivo* necesario para lograr una alta calidad del conocimiento a descubrir [3].

La fuente de datos proviene de la información académica de la UGD, datos proporcionados al ingreso (personales y antecedente de estudio a la institución) y durante el lapso de sus estudios en la institución; con la debida protección de datos personales, sin identificación de individuos, descartando cualquier información que pueda identificar directa o indirectamente a los alumnos (DNI, Apellidos, Matriculas, etc.), creando una

vista minable con las características (atributos) de las titulaciones seleccionadas, una colección de individuos, en este caso alumnos de la UGD, sobre los cuales se realizó el estudio, a la que se le aplica la fase de Minería de Datos para poder extraer conocimiento útil en lo que se refiere al rendimiento académico del alumnado. Dicha vista minable contiene información de las carreras tanto del departamento de administración como el de informática, correspondiente un periodo de 10 años, que va desde el año 1999 al 2009 respectivamente. La precisión de los datos a utilizar depende en gran medida de la completitud de los mismos.

Objetivos y Resultados

Objetivo General

Identificar características y patrones de comportamiento relacionadas con el desempeño académico de los alumnos, a través del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) y herramientas de Minería de Datos. Los modelos identificados se propondrán como contribución a la toma de decisiones en el ámbito de la gestión académica.

Objetivos Específicos

Utilizar herramientas de Minería de Datos para detectar patrones y relaciones entre los datos de la trayectoria académica de los estudiantes.

Identificar cuáles son las variables y áreas que inciden sobre el desempeño, permanencia y graduación de los alumnos universitarios.

Elaborar recomendaciones sobre los posibles usos de los resultados obtenidos y las características de las fuentes de información que sirvan como estrategias innovadoras y apropiadas para la gestión académica.

Actividades realizadas para el alcance de los objetivos:

Búsqueda y análisis de información referente a los aspectos teóricos relacionados con el proceso de descubrimiento de conocimiento en base de datos (KDD), técnicas de minería de datos (MD), como ser: pre-procesado de los datos, clasificación, segmentación, y asociación; así como los algoritmos disponibles en ellas. Estas actividades estaban precisamente destinadas a identificar y describir el proceso, técnicas y algoritmos de MD dentro del ámbito educativo/académico.

Obtención de datos pertinentes a la investigación proporcionados por la UGD para su análisis. Integración y recopilación de datos: para esta etapa se utilizaron técnicas como la ejecución de instrucción en SQL en la base de datos utilizando los criterios correspondientes a los fines de la investigación, con el fin de generar el almacén de datos con el cual se trabajará en las siguientes etapas.

Conversión de los datos seleccionados: para poder procesar los datos en algunas de las herramientas de minería de datos, se requiere la conversión al formato CSV (archivos separados por comas) en cada una de las variantes y conjuntos de datos que se generen para tal fin.

Ejecución de pruebas con la herramienta de MD Weka: analizando los resultados preliminares obtenidos en el período anterior, se procedió a llevar a cabo una serie de pruebas con los datos seleccionados y acondicionados para observar el comportamiento de las técnicas y los algoritmos de MD de acuerdo a las nuevas series de datos seleccionadas, aplicando distintas variantes y ajustes a los parámetros de ejecución en los distintos algoritmos y adaptaciones de los datos con el fin de obtener variantes en los resultados.

Análisis y elaboración de resultados preliminares sobre las pruebas realizadas. Comprobación de resultados mediante confrontación de diferentes técnicas de MD.

Búsqueda y análisis de información referente a las prestaciones de una serie de herramientas de minería de datos, en busca de las más adecuadas para la investigación.

Formación de Recursos Humanos

El trabajo corresponde a una tesis de grado de Ingeniería en Informática, que durante su desarrollo recibió una beca de iniciación a la investigación científica, otorgada por el Comité Ejecutivo de Desarrollo e Innovación Tecnológica de la Provincia de Misiones (CEDIT).

Referencias

[1] MOLINA, FÉLIX LUIS CARLOS. “Data mining: torturando a los datos hasta que confiesen”. [En línea] 2002. http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html#bibliografia.

[2] “Data Mining y el Descubrimiento del Conocimiento”. VALCÁRCEL ASENCIOS, VIOLETA. 2, s.l. : Revista de la Facultad de Ingeniería Industrial, 2004, Vol. (7), págs. 83-86. ISSN: 1560-9146 (impreso) / ISSN: 1810-9993 (electrónico).

[3] FAYYAD, U. M., PIATETSKY SHAPIRO, G., SMYTH, P., UHTURUDSAMY, R., “Advances in Knowledge Discovery & Data Mining”,1ª Edition, Editorial MIT Press, Cumberland, Rhode Island, EE.UU., 1996. ISBN: 9780262560979.

[4] VERÓNICA S. BOGADO Y MARIANA C. ARRUZAZABALA, “Sistemas Operativos – Descubrimiento de Conocimiento en Base de Datos”, Corrientes Argentina: Universidad Nacional del Nordeste - Facultad de Ciencias Exactas, Naturales y Agrimensura (UNNE), Trabajo Monográfico, 2003.

[5] CARLOS MARIN, “Minería de Datos”, http://mineriadatos.blogspot.com/.

[6] RAMASWAMI, M. y BHASKARAN, **R.** “A Study on Feature Selection Techniques in Educational Data Mining”. Journal of Computing, 2009, Vol. Vol.1 , Issue 1.

[7] INTERNATIONAL WORKING GROUP ON EDUCATIONAL DATA MINING. [En línea] http://www.educationaldatamining.org/.

[8] “Educational Data Mining 2008 - The 1st International Conference on Educational Data Mining”. JOAZEIRO DE BAKER, RYAN S., BARNES, TIFFANY y BECK, JOSEPH. Montréal Québec Canada : s.n., 2008. URL: http://www.educationaldatamining.org/EDM2008/uploads/proc/full%20proceedings.pdf.

[9] “Proceedings of Educational Data Mining workshop, held in conjunction with the 8th International Conference on Intelligent Tutoring Systems”. HEINER, CECILY, BAKER, RYAN y YACEF, KALINA. Jhongli Taiwan : s.n., 2006.

[10] INFORMATION, SCIENCE TODAY. “Educational data mining”. [En línea] http://www.infosciencetoday.org/document-management/educational-data-mining.html.

[11] “Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education”. HEINER, CECILY, HEFFERNAN, NEIL y BARNES, TIFFANY. Marina del Rey CA. USA : s.n., 2007.

[12] VENTURA SOTO, SEBASTIÁN. “Minería de Datos en Sistemas Educativos”. [En línea] http://sci2s.ugr.es/docencia/doctoM6/EducationalDataMining.pdf.

[13] HERNÁNDEZ ORALLO, J., RAMÍREZ QUINTANA, M.J. y FERRI RAMÍREZ, C. “Introducción a la Minería de Datos”. 1a Edición. Madrid España: Editorial Pearson, 2004. ISBN: 84-205-4091-9.