Aplicación de Técnicas de Data Mining en Gestión de Docentes de Educación Superior

Lucía Rosario Malbernat; María Patricia Clemens; Analia Elena Varela; Ezequiel Matías Urrizaga

Grupo DM-ES, Departamento de Sistemas, Universidad CAECE, Subsede Mar del Plata
Gascón 2464, Mar del Plata, Buenos Aires, República Argentina
+54 233 499-3400
{lmalbernat, mpclemens, avarela, eurrizaga}@ucaecemdp.edu.ar

Resumen

En estudios previos se analizó la factibilidad de incorporar actividades virtuales según las competencias docentes aplicando técnicas de Data Mining y se concluyó que, "en relación con la Preparación y la Actitud para la modalidad virtual, los docentes pueden clasificarse como Innovadores, Indiferentes y Refractarios", de acuerdo con lo propuesto por la hipótesis de investigación.

En la línea de investigación que se reporta en este trabajo se han tomado los modelos y datos recabados en dichos estudios, se recabaron nuevos datos y están en etapa de exploración distintas técnicas de segmentación, variando sus métodos y parámetros para valorar su aplicación a través de distintas herramientas informáticas para Data Mining.

De este modo, se evaluará el impacto que tienen las técnicas, métodos, parámetros y herramientas en los resultados y en el propio procesamiento de los datos, para desarrollar un modelo de segmentación de docentes según su preparación y actitud para incorporar tecnología en educación superior basado en la comparación de resultados que permita desarrollar una herramienta informática para la aplicación del modelo.

Los datos recogidos, que han sido revisados, depurados, ampliados y actualizados, provienen de distintas unidades académicas de universidades de gestión pública y privada.

Palabras clave: Data Mining, segmentación, educación superior, TIC, Preparación docente, actitud.

Contexto

La investigación que se reporta en este trabajo, titulada "Aplicación de técnicas de Data Mining en gestión de docentes de educación superior (DM-ES)", continúa estudios previos vinculados con la innovación en educación universitaria, está radicada en el Departamento de Sistemas de Universidad CAECE y aprobada por R.R. 549/13 para el período 2014-2015.

Se relaciona con dos proyectos, aprobados también para el período 2014-2015, uno radicado en el Centro de Investigación en Procesos Básicos, Metodología y Educación (CIMEPB) de la Facultad de Psicología de la Universidad Nacional de Mar del Plata y el otro, en la Facultad de Ciencias Económicas y Sociales de la misma Universidad.

El primer proyecto mencionado, denominado, "Competencias para la Inno-

vación Docente en Enseñanza Superior: Preparación y Actitud para el Uso de las TIC", ha aportado la depuración de las variables de segmentación con las que se venía trabajando, Preparación y Actitud, de modo de obtener datos confiables y validados para el procesamiento mediante las técnicas de Data Mining. El segundo proyecto, "Estudio de la identidad, la cultura y el clima organizacional en la Universidad y su influencia en el desarrollo de la Profesión Académica", cuenta, entre sus objetivos específicos, con uno estrechamente relacionado con la investigación que se reporta pues será el nexo para definir un modelo de gestión de la innovación docente a partir de los resultados a los que se arribe al procesar los datos.

Dicho objetivo plantea "Analizar el impacto del uso de TIC para fines académicos en la cultura organizacional a fin de aportar información útil para la toma de decisiones vinculada con la gestión e implementación de estos proyectos diseñando estrategias de motivación a los docentes, según su perfil innovador".

Introducción

Data Mining está definida en el IT *Glossary* de *Gartner Group* [GG14] como proceso de descubrimiento de correlaciones significativas, patrones y tendencias que se obtienen examinando grandes volúmenes de datos almacenados en repositorios, empleado tanto tecnologías de reconocimiento de patrones como técnicas estadísticas y matemáticas. También ha sido definida como la extracción no trivial de datos implícitos, previamente desconocidos y potencialmente útiles a partir de datos [FPM92].

Según Piatetsky-Shapiro [P05], en los últimos años, la minería de datos ha atravesado una gran transición pasando por las distintas fases del ciclo de sobre-expectación de la tecnología y, mientras el mundo siga produciendo datos de todo tipo a un ritmo cada vez mayor, la demanda de la minería de datos seguirá creciendo. Para Britos [B08], la explotación de información es una subdisciplina de la Informática que aporta a la Inteligencia de Negocio las herramientas para la transformación de información en conocimiento que contribuirá a la toma de decisiones de gestión y generación de planes estratégicos en las organizaciones.

Las técnicas de Data Mining están ampliamente difundidas en diversas disciplinas y rubros pero resultan escasos los reportes relacionados con la gestión educativa del cuerpo docentes que describan el uso de estas técnicas en Iberoamérica.

Para la *International Educational Data Mining Society*, la minería de datos educacional es una disciplina emergente, preocupada por el desarrollo de métodos para la exploración de datos que provienen de centros educativos que puedan ser usados para comprender mejor a los estudiante, y realizar ajustes en sus aprendizajes [JED15]. Basta ver que los artículos publicados en su JEDM - Journal of Educational Data Mining efectivamente se enfocan hacia las métricas del aprendizaje y no de la gestión administrativa de las instituciones educativas, tal como ocurre con la *Society for Learning Analytics Research* (SoLAR), una red interdisciplinaria integrada por investigadores internacionales, que explora el rol e impacto de las analíticas en la enseñanza, el aprendizaje, la formación y el desarrollo. Sus producciones se reflejan en la Journal of Learning Analytics.

También da cuenta del contexto que se describe el libro Tendencias de la Minería de datos [GRA04] que recopila el trabajo de los grupos que integran la Red de Minería de Datos y Aprendizaje, subvencionada por el Ministerio de Ciencia y Tecnología de España, con el fin de difundir

las principales líneas de investigación españolas, destacándose investigaciones en áreas disímiles a la educativa.

Otro exponente de este son las Jornadas Argentinas de Data Mining, llevadas en cabo anualmente en la UBA desde 2006. Las temáticas tratadas en las presentaciones han abarcado procesos de diversos tipos, incluida la gestión empresarial o productiva, pero sólo se hace mención al tratamiento de datos sobre educación en 2014, en una presentación sobre el Sistema de Información Universitaria (SIU).

Por otra parte, durante muchos años el interés de la comunidad de investigadores se ha centrado más en los algoritmos que en los procesos pero la creciente incorporación de la visión de ingeniería en proyectos de software ha resultado en la necesidad de que ese tipo de visión se aplique en proyectos de minería [GBP11].

Se detecta así, una carencia en gestión universitaria relacionada con la aplicación sistemática de técnicas de minería u otros tratamientos de datos vinculados a los docentes y en ese marco se ha propuesto la investigación que aquí se describe.

Líneas de Investigación, Desarrollo e Innovación

La investigación que se reporta indaga variaciones en la técnica de segmentación aplicada, método elegido, parámetros utilizados y herramienta informática empleada para procesar datos vinculados con la preparación y actitud de los docentes para incorporar TIC, a fin de evaluar el impacto que tienen en los resultados (conformación de la agrupación) y en el propio procesamiento de los datos, midiendo precisión y eficiencia.

La noción de lo que constituye un buen agrupamiento y la calidad de los segmentos tiene aspectos intrínsecos y extrínsecos que están siendo evaluados.

Hasta el momento se han aplicado dos tipos de métodos: uno probabilístico (EM, *Expectation Maximization*) y otro basado en k-particiones (k-means), [MCV14].
Las herramientas informáticas utilizadas para el análisis de los datos han sido Weka (*Waikato Environment for Knowledge Analysis*), PSPP (*Perfect Statistics Professionally Presented*) y RapidMiner versión *Starter*.

Resultados y Objetivos

En investigaciones anteriores se diseñó un algoritmo de segmentación basado en el método k-means [M13], aplicado utilizando SQL en un motor de base de datos, que agrupaba a los docentes según su actitud innovadora en Refractarios, Indiferentes e Innovadores, considerando la preparación y actitud para la virtualizar sus materias de grado [M12]. Dado que cada sujeto (docente) estaba cuantitativamente representado por dos indicadores, Preparación (P) y Actitud (Q), el dataset utilizado para procesar la muestra caracterizaba a cada instancia con un par (p,q).

Durante la investigación que se reporta, se aplicó el algoritmo K-means a la muestra original, reiteradamente, utilizando Weka, PSPP y RapidMiner.

Inicialmente, se generaron luego de 6 ó 7 iteraciones, utilizando esas 3 herramientas, 3 segmentos, los cuales no presentaban cambios sustantivos con el procesamiento previo en lo que hacía a la conformación de los clústeres (22% de la muestra correspondió a los Refractarios, 62% a los Indiferentes y 16% a los Innovadores).

A los fines de mejorar la conformación de los segmentos, y sin estudiar aun las cualidades de los clúster a través de índices internos de validación de los agrupa-

mientos, se observaron los resultados de utilizar K-means con K = 4, ya que se detectó que el clúster de Indiferentes tenía instancias que tendían a acercarse a los Refractarios e instancias fronterizas con los Innovadores.

Surgió de esta nueva corrida en Weka, que, luego de 13 iteraciones, los 2 clúster centrales continuaron siendo mayoritarios (68% del total), absorbiendo algunos pocos refractarios (menos del 20% inicial) y casi un 40% de los Innovadores.

Así, lo que originalmente se consideró como el grupo de Indiferentes, fue dividido en sujetos que hemos dado en llamar Reticentes por estar más cercanos a los Refractarios que a los Innovadores y en Flemáticos por tener una actitud y preparación superior a la media sin llegar a estar entre los indiscutiblemente innovadores.

Al procesarse la muestra de datos tanto con PSPP como con RapidMiner, utilizando el algoritmo K-means con K=4, luego de 10 iteraciones se arribó a idénticos resultados tanto en lo que hace a la asignación de grupos para cada docente como en la ubicación de los centroides.

Luego, se analizaron los resultados de procesar los datos con K-means en Weka seleccionando el valor 5 para k, resultado al que se arribó luego de ser 10 iteraciones. En este caso, los clústeres de los extremos (Refractarios e Innovadores) no sufrieron ninguna modificación.

Así, 25% de los sujetos analizados continuaron estables en el grupo de Refractarios y 14% en el grupo de Innovadores, pero generándose un nuevo grupo dentro de los Indiferentes.

Este nuevo grupo fue denominado Desorientados, interpretándose que poseía un alto nivel de actitud con una baja preparación, por debajo de la media.

La generación de 5 grupos obtuvo una división más equilibrada de la muestra, con una varianza no mayor a 5%. No obstante el grupo correspondiente a sujetos con mayor preparación y actitud (Innovadores), se mantuvo conformado por el 14% de los docentes.

A diferencia de lo que venía pasando hasta el momento, al procesar los datos con RapidMiner utilizando k-means para generar 5 grupos, varió notoriamente la composición de los clústeres de modo que, el clúster de Reticentes fue conformado por el 14% de la muestra, el 23% correspondió al grupo con baja preparación y alta actitud (Desorientados), los Reticentes conformaron el grupo más denso con el 51% de los docentes y un 8% de los docentes quedó en el grupo Flemático, que, en este caso, presentaban alta preparación y actitud, ya que menos del 5% había quedado en el grupo de Innovadores.

Así, se pasó de la composición más equilibrada en cuanto a la conformación de los grupos obtenida en Weka, a una propuesta de agrupamiento que continuaba teniendo un clúster (el de los Reticentes) con una probabilidad de pertenencia de un sujeto mayor al 50% contra una probabilidad de pertenencia al grupo de innovadores de menos del 5%.

Por otra parte, al aplicar el algoritmo EM, dado que es probabilístico, se obtuvieron resultados distintos a k-means. Con Weka, para 3 grupos, utilizando una distribución normal, los grupos quedaron conformados con un 22,5% de sujetos Refractarios, 55% de Indiferentes y 22,5% de Innovadores.

Al obtener 4 grupos con EM, los clústeres se modificaron una vez más para cambiar significativamente al grupo central, destacado por su tamaño, con más de la mitad de los docentes (52%), equivalente al anteriormente identificado como grupo de Flemáticos, sin modificar significativamente a los Innovadores (14%). Los Refractarios, pasaron a incluir al 17% de los sujetos, y los Reticentes, al 16%.

Se puedo advertir, luego, que, con la implementación de 5 grupos, se generó una división más homogénea en la muestra.

Resta ahora verificar si estos segmentos se mantienen estables en las nuevas muestras de datos tomadas, finalizar los estudios de calidad de los clústeres a partir de medirse su cohesión interna y separación inter-clústeres y seleccionar las mejores segmentaciones a partir de los criterios intrínsecos y extrínsecos fijados.

Formación de Recursos Humanos

El grupo de investigación en Data Mining aplicado a la educación superior, DM-ES, está integrado por Lucía Rosario Malbernat, Directora del Proyecto, y María Patricia Clemens, ambas docentes del Departamento de Sistemas; Analia Elena Varela y Ezequiel Matías Urrizaga, egresados de la Licenciatura en Sistemas y los Estudiantes avanzados de Ingeniería en Sistemas, Franco Biagioli, Ignacio Salvarrey y Gabriel Verdi, quienes se encuentran realizando en el marco de esta investigación sus respectivos trabajos finales de carrera.

Referencias

[B08] P. Britos, Procesos de explotación de información basados en sistemas inteligentes. Tesis de Maestría. La Plata, Buenos Aires, Argentina, 2008

[FPM92] W.J. Frawley, G. Piatetsky-Shapiro, C.J.. Matheus, Knowledge Discovery in Databases: An Overview. AAAI Fall 1992, California, 1992, pp. 57-69

[GBP11] R. García-Martínez, P. Britos, P, Pesado, R, Bertone, F. Pollo-Cattaneo, D., Rodríguez, P., Pytel, J., Vanrell Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching. Sello Editorial Universidad de Medellín ISBN 978-958-8692-32-6, 2011, pp. 83-99

[GG14] Gartner Group. IT Glossary. http://www.gartner.com/it-glossary/data-mining.

[GRA04] R. Giráldez J., Riquelme, & J. Aguilar-Ruiz, (Eds.) Tendencias de la Minería de Datos en España. Red Española de Minería de Datos. TIC2002-11124-E ISBN 84-688-8442-1. España, 2004

[JED15] JEDM - Journal of Educational Data Mining. Vol 7, No 1 (2015)

[KR02] R. Kimball, M.R. Ross, The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling 2a ed. USA: Wiley Publishing, Inc, 2002, pp. 243-254

[M12] L. R. Malbernat, Innovación en Educación universitaria: Factibilidad de incorporar actividades virtuales según las competencias docentes. Tesis de Maestría de la Universidad Nacional de Mar del Plata. Facultad de Ciencias Económicas y Sociales. Argentina.

[M13] L. R. Malbernat Incorporating virtual activities in Higher Education: a mathematical model for describing teachers according to their skills. XVIII Congreso Argentino de Ciencias de la Computación, CACIC 2013, RedUnci ISBN 978-987-23963-1-2, 2013, pp 609-619.

[MCV14] L. R. Malbernat, M. P. Clemens, A. E. Varela, E. Urrizaga. Aplicación de técnicas de Data Mining en Gestión de Docentes de Educación Superior. Impacto en el Desarrollo de la Profesión Académica. IV Congreso sobre Nuevas Tendencias en la Formación Permanente del Profesorado. UNTREF. Argentina. 2014 pp. 2278-2295.

[P05] G. Piatetsky-Shapiro, Data Mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from “university” to “business” and “analytics”. Data Min Know Disc DOI 10.1007/s 10618-006-0058-2 15, Springer Science + Business Media LCC, 2007, pp. 99–105