Metodologías de Análisis de los Big Data en las Plataformas Educativas

M. R. Martinez-Torres, D. G. Reina, S. L. Toral, F. Barrero
Universidad de Sevilla, España

***Abstract*—La proliferación de nuevas plataformas educativas por Internet y el avance de la educación *online* ha abierto nuevas posibilidades de análisis debido al gran volumen de datos generados y almacenados en los servidores. Los usuarios dejan trazas de su actividad, y esta actividad posibilita nuevos análisis del comportamiento de estudiantes y de los contenidos compartidos, difícilmente realizables en la educación cara a cara tradicional. Este trabajo aporta un resumen de las diversas metodologías aplicables a los grandes volúmenes de datos generados por las plataformas educativas, clasificables dentro de los Big Data, así como los diversos campos en los que podrían aplicarse y las mejoras que podrían introducir en el desarrollo de las propias herramientas.**

***Index Terms*—Big Data, metodologías de análisis, participación, contenido compartido.**

I. Resumen

El análisis de los grandes volúmenes de datos disponibles *online* se ha convertido en una ciencia emergente que suele aglutinarse bajo la denominación de análisis de los Big Data [1]. La captura de datos no estructurados por la Web y su análisis se ha convertido en una herramienta clave y estratégica para la toma de decisiones y la asignación de recursos en muchas áreas, como la gestión empresarial, la inteligencia, los servicios de defensa, o las decisiones de autoridades públicas [2]. Muchas compañías como Microsoft, IBM, Google, o Amazon realizan fuertes inversiones con el fin de generar valor a partir del análisis de los Big Data. No obstante, en el terreno educativo todavía no han proliferado demasiado este tipo de análisis, a pesar de ser un ámbito donde las herramientas *online* y las plataformas de enseñanza han tenido un rápido y temprano auge, mucho antes que en otras áreas mencionadas.

El análisis de grandes volúmenes de datos en el ámbito educativo suele denominarse EDM (Educational Data Mining) o LA (Learning Analytics), y su objetivo es promover nuevos descubrimientos y avances en el terreno educativo mediante el uso de la información almacenada *online*. Contempla múltiples dimensiones, i.e., social, cognitiva, emocional, meta cognitiva, etc. y se centra en individuos, grupos de individuos o instituciones. En este contexto, el objetivo principal del trabajo es presentar las principales metodologías aplicables en el terreno de los EDM y LA para transformar información en datos estadísticamente analizables. Aunque gran parte de la información relevante puede obtenerse directamente de los archivos de log de los servidores educativos o de las bases de datos de usuarios registrados, existe otra información de gran interés en las plataformas educativas como es la interacción entre los usuarios, sus patrones de participación o el conocimiento compartido. Esta información no puede extraerse directamente, sino que es necesario aplicar metodologías como el análisis de redes sociales [3] o el procesamiento del lenguaje natural para materializar dicha información en datos procesables [4].

El artículo se estructura de la siguiente forma: en primer lugar, se parte de un análisis de la información relevante en las plataformas educativas. Se trata de determinar qué información es relevante desde el punto de vista educativo, y cuáles son los principales retos en la extracción de esta información. La siguiente sección presenta las diferentes metodologías que permiten transformar la información relevante en datos estructurados, así como las principales metodologías estadísticas relacionadas con el EDM. En lo que respecta a la participación de los usuarios, se muestra como modelar las interacciones de los usuarios como una red social, las principales características topológicas locales y globales que pueden extraerse de dicha red y su aproximación a modelos de redes complejas. En la parte de conocimiento compartido, se detallan las principales técnicas de análisis semántico para analizar el conocimiento creado y compartido. Finalmente, la sección IV detalla los principales retos futuros y la sección V las conclusiones.

II. Información disponible en las plataformas educativas

La información disponible por las instituciones de educación superior a través de sus bases de datos y, fundamentalmente, a través de sus plataformas educativas y campus virtuales permite la realización de nuevos análisis más complejos, más allá de la información recopilada sobre la base de cuestionarios. Esta información se caracteriza por ser abundante, accesible, pero en muchos casos no estructurada, lo que hace necesario un análisis previo que transforme la información en datos. A continuación se distinguen cuatro tipos de fuentes de información sobre las que llevar a cabo estos análisis.

A. Perfiles de estudiantes

Las instituciones de educación superior poseen en sus bases de datos una amplia información sobre los perfiles de los estudiantes, así como de toda su vida académica, desde que realizan su ingreso. Esta información se usa habitualmente a efectos de clasificación y categorización. Por ejemplo, para clasificar los perfiles de los estudiantes matriculados, o determinar las tasas de finalización o abandono. No obstante, más raramente suelen utilizarse estos datos para estimación o predicción. Por ejemplo, para predecir la tasa de abandono de alumnos, estimar el número de matriculados [5]. Una aplicación particularmente interesante consiste en determinar la probabilidad que un determinado perfil de estudiante ingrese en una u otra carrera, lo que serviría apara determinar el grado de aceptación de una titulación [6].

B. *Información sobre el aprendizaje*

En la última década, la práctica totalidad de instituciones de educación superior han incorporado plataformas de gestión del aprendizaje (LMS, Lerning Management Systems) como parte de sus campus virtuales. Si bien el objetivo de estas plataformas es mejorar el rendimiento y la efectividad en el proceso de enseñanza y servir como apoyo al desarrollo de asignaturas, también pueden proporcionar una gran cantidad de información debido a la traza que los estudiantes dejan al utilizar sus recursos. Los archivos log de los servidores web permiten almacenar datos diversos sobre la actividad de los usuarios. Por ejemplo, el tiempo que el estudiante pasa realizando una determinada tarea, el aprovechamiento de ese tiempo según las interacciones que el estudiante haya realizado, el número de veces que el estudiante se conecta a la plataforma para la realización de las tareas asignadas, etc. Estos datos permiten modelar el comportamiento de los estudiantes durante su aprendizaje. A diferencia de los métodos tradicionales, que requieren de observación directa o de una grabación en video, los datos anteriores pueden obtenerse fácilmente de los registros de los servidores y no son intrusivos. En esta línea se han identificado aplicaciones de la minería de datos en la exploración y análisis de los datos alojados en el LMS [7], identificación de patrones [8], y evaluación de la actividad del estudiante dentro de la plataforma para la identificación de estilos de aprendizaje [9], entre otras. También permiten la visualización del aprendizaje en forma de curvas de aprendizaje, que representan las mejoras en el aprendizaje de una habilidad o aptitud del estudiante frente al número de oportunidades de entrenar esa habilidad o aptitud [10].

C. *Interacciones*

Una de las principales ventajas de las plataformas de gestión del aprendizaje es que facilita de manera virtual las interacciones entre los alumnos y de éstos con el profesor. Estas interacciones también constituyen una fuente de información. En primer lugar, proporcionan información sobre el grado de participación del alumno, que puede usarse a efectos de evaluación. Pero también permite descubrir patrones de comportamiento de los alumnos y su organización informal modelando las interacciones como una red social [11]. Mediante técnicas de análisis de redes sociales, pueden descubrirse características locales de los usuarios y globales de la red. Por ejemplo, qué alumnos ocupan las posiciones más centrales, en qué medida las interacciones con otros alumnos facilitan el aprendizaje, o cuál es el grado de cohesión dentro de un curso.

D. *Contenido*

Las interacciones de los alumnos normalmente tienen lugar en forma de texto escrito, que queda almacenado en los servidores para su posterior análisis. Este texto se puede analizar algorítmicamente mediante técnicas de procesamiento del lenguaje natural, integrando los resultados dentro de las propias plataformas o entornos educativos. Por ejemplo, existen numerosos trabajos enfocados a hacer herramientas de e-learning personalizadas mediante la incorporación de recomendaciones, acciones correctoras o resolución automática de preguntas analizando los contenidos generados por el alumno [12]. También se han llevado a cabo agentes para la realización de acciones tutoriales automáticas, identificando de forma inteligente las preguntas del alumno [13]. Estas mismas técnicas pueden también utilizarse para evaluar a los alumnos, más allá de respuestas tipo test. En general, es posible realizar estas evaluaciones cuando se trata de respuestas cortas con construcciones gramaticales sencillas [14].

Otra aplicación interesante del procesamiento del lenguaje natural se refiere a la clasificación de textos en general, y el análisis de sentimiento en particular. El análisis de sentimiento es la detección y clasificación de textos atendiendo a un tipo de actitud, en su forma más simple positiva o negativa, hacia el objeto de dicho texto [15]. Esto puede ser interesante para evaluar no ya las respuestas de conocimiento de los alumnos, sino los comentarios compartidos en foros de discusión. En particular, analizar las actitudes positivas o negativas hacia la materia o parte de ella, o hacia determinadas herramientas o pruebas.

III. METODOLOGÍAS DE ANÁLISIS

Las principales técnicas relacionadas con la minería de datos educativos y análisis del aprendizaje se describen a continuación siguiendo la clasificación propuesta en [16].

A. *Técnicas estadísticas predictivas*

Las técnicas estadísticas predictivas tienen como objetivo predecir el comportamiento de un aspecto de los datos (variable dependiente) como combinación del resto de características (variables independientes). Las técnicas predictivas se basan esencialmente en regresiones y clasificaciones supervisadas. Normalmente estas técnicas se aplican para predecir el rendimiento académico de los alumnos. Por ejemplo, establecer modelos predictores sobre el índice de aprobados de un curso, sobre su nota media, o predecir el tiempo que un estudiante tardará en completar una tarea. Relacionado con lo anterior son los clasificadores y árboles de decisión, que también pueden utilizarse para clasificar grupos de alumnos o predecir su nota final [17], [18]. No obstante, las técnicas predictivas también pueden utilizarse para predecir si el alumno posee una determinada aptitud o competencia. Esto es lo que se conoce como estimación de conocimiento latente, llamado así porque el conocimiento no es una variable directamente observable. En este caso, algoritmos como el conocido Bayesian Knowledge Tracing tratan de determinar en qué medida un estudiante conoce una determinada aptitud o habilidad a partir de su rendimiento pasado con esa habilidad [19]. Esta información resulta de gran utilidad para determinar en qué medida una plataforma educativa cumple con su objetivo, para informar a los profesores o incluso para realizar acciones correctoras pedagógicas de manera automática.

B. *Descubrimiento de estructuras*

El objetivo en este caso es el descubrimiento de estructuras y patrones en los datos capturados. Abarca los diversos algoritmos de clustering y el análisis factorial. En el caso de los entornos educativos, los algoritmos de clustering permiten clasificar a los alumnos según una determinada característica y ver la evolución de su aprendizaje en el tiempo [20]. A diferencia de los algoritmos de clustering, que tratan de encontrar agrupaciones dentro de una nube de puntos, el análisis factorial se basa en realizar agrupaciones de variables para

determinar un conjunto reducido de variables latentes, obtenidas como combinación lineal de las variables originales. En general, se utiliza para reducir la dimensionalidad del problema. Por ejemplo, es posible identificar muchas variables que caractericen el comportamiento de los estudiantes, o tener una taxonomía muy elevada de características que deben poseer las herramientas de e-learning. En lugar de trabajar con un elevado número de características, estas pueden agruparse en unos factores latentes, reduciendo la dimensionalidad del problema final [21].

C. *Minería de relaciones*

La minería de relaciones trata sobre el descubrimiento de las relaciones entre las variables dentro de un conjunto extenso de datos. La forma más simple de minería de relaciones son las correlaciones. Un paso más allá son las relaciones de causalidad. La minería causal trata de generar métodos eficientes para descubrir las relaciones causales en bases de datos observacionales. Las relaciones de causalidad son diferentes a las de predicción ya descritas. El hecho de que dos eventos covaríen no significa que exista una relación causal entre ellos. Las relaciones de causalidad intentan dar respuesta a preguntas como qué comportamiento de los estudiantes causa el aprendizaje, o qué ocurre sin una acción tutorial suministra algunas pistas tras un error [22]. La minería de reglas de asociación busca encontrar automáticamente reglas del tipo "if-then" dentro de grandes volúmenes de datos. Las reglas de asociación implican causalidad, pero ésta puede ir en ambos sentidos. En entornos educativos, puede servir para descubrir por ejemplo asociaciones entre cursos que cursan los estudiantes, o asociaciones entre los pasos que siguen para completar un problema. Como caso particular de lo anterior, la minería de patrones secuenciales es la minería de patrones que ocurren frecuentemente relacionados al tiempo u otras secuencias. Por ejemplo, descubrir una secuencia temporal de acciones de los usuarios cuando interactúan con una herramienta de e-learning [23].

D. *Análisis de redes sociales*

El análisis de redes sociales consiste en modelar una comunidad como un grafo, donde los nodos representan a los usuarios identificados por un email o un alias y los arcos las interacciones entre los usuarios. Mediante el análisis de redes sociales se puede modelar la participación de los alumnos en una plataforma de aprendizaje, o los trabajos colaborativos entre ellos. Cada nodo como miembro de la red social posee una serie de características topológicas, como el grado (número de arcos que inciden o salen del nodo), su centralidad (de intermediación, cercanía, de autovalor) o su coeficiente de clustering [24]. La propia red considerada en su conjunto también posee características medibles, como su tamaño (número de nodos), el diámetro de la red o su ASP (Average Shortest Path). Asimismo, las características locales de los nodos se pueden promediar para toda la red dando lugar a un grado, centralidad o coeficiente de clustering global de la red. Todas estas características pueden utilizarse para detectar determinado perfiles de usuarios o identificar grupos o subcomunidades dentro de la red global [25], [26]. Por ejemplo, qué tipo de participación desempeña cada alumno en trabajos colaborativos, cuáles son los que ocupan una posición clave para el buen desarrollo del trabajo, qué papel desempeña el profesor dentro de la red o cuál es el grado de cohesión de los alumnos matriculados en un curso.

Las redes sociales también se estudian desde la perspectiva de los modelos de redes complejas. Por lo general, los foros en Internet se comportan como redes de escala libre, donde las contribuciones de los usuarios siguen una distribución en ley de potencias [27]. Esto significa que la mayor parte de las contribuciones las realizan un reducido número de usuarios, que es lo que se conoce como desigualdad participativa [28].

E. *Procesamiento del lenguaje natural*

El procesamiento del lenguaje natural es un conjunto de técnicas algorítmicas para analizar el lenguaje humano. La aproximación más sencilla consiste en partir de un conjunto de términos o palabras clave que representen una taxonomía de un campo de conocimiento, y a partir de él generar una matriz de incidencia términos documentos, donde cada celda de la matriz contenga el número de veces que cada término aparece en cada documento. El modelo de espacio de vectores [29] considera un espacio de tantas dimensiones como términos o documentos, y las similitudes de términos o documentos se calculan mediante la proximidad de los vectores fila o columna de matriz anterior (por ejemplo, mediante el coseno del ángulo entre dos vectores). El principal problema de esta técnica es que cuando se trabaja con muchos documentos, la dimensionalidad es muy elevada, y la mayoría de los elementos de la matriz son ceros. Una alternativa consiste en reducir esta alta dimensionalidad proyectando los términos o los documentos en un subespacio en el que la estructura semántica resulta más clara. En este subespacio se pueden aplicar las mismas medidas de similitud de términos o documentos con resultados más fácilmente interpretables [30]. Uno de los algoritmos de reducción de dimensionalidad más populares es el conocido como análisis semántico latente (LSA, Latent Semantic Analysis), que descompone la matriz términos documentos mediante una descomposición en valores singulares, quedándose con las dimensiones correspondientes a los autovalores más elevados [31].

La clasificación de textos es otra de las aplicaciones más habituales en el procesamiento del lenguaje natural. La clasificación consiste en asignar documentos a un conjunto predefinido de clases, normalmente, mediante algoritmos de aprendizaje supervisado. Un caso particular de la clasificación de texto es el análisis de sentimiento, que consiste en detectar actitudes de los textos hacia objetos o personas, dentro de un conjunto pre definido de clases. En su forma más simple, el análisis de sentimiento identifica la actitud positiva o negativa de los textos.

En el ámbito educativo, el procesamiento del lenguaje natural tiene interés en el análisis de las discusiones *online* a través de foros [32] así como el análisis de los sentimientos y emociones de los usuarios [33].

IV. FUTUROS RETOS

La proliferación y constante desarrollo de las plataformas educativas en Internet esta cambiando no sólo el papel desarrollado por los alumnos o usuarios, sino también el de los docentes. Los usuarios de estos sistemas han pasado de ser simples espectadores a usuarios activos con capacidad de decidir sobre su propio aprendizaje. Al

mismo tiempo, estas interacciones de los usuarios con las plataformas educativas generan un elevado volumen de información, que queda almacenado y disponible para su posterior análisis. La aplicación de técnicas de data mining y de análisis sobre estos datos posibilita la realización de nuevos descubrimientos sobre patrones de comportamiento de los usuarios. Mediante ellos, los docentes también pueden detectar y rastrear nuevos problemas y aplicar las posibles correcciones que podrían mejorar el proceso de enseñanza/aprendizaje.

En los próximos años, el campo de análisis de los Big Data en entornos educativos irá progresivamente alcanzando su madurez, lo que a su vez plantea nuevos retos. En primer lugar, los Big data educativos serán cada vez mayores, lo que supondrá modificar el almacenamiento y procesamiento computacional de los datos. A diferencia de lo que ocurría con métodos tradicionales de análisis mediante encuestas, los datos pasarán de ser un recurso escaso a un recurso abundante, o incluso sobre abundante. Esto también significa que un mayor número de investigadores comenzará a trabajar en estos temas, al existir menos barreras de entrada. Desde un punto de vista metodológico, las técnicas de predicción e inferencia estadística serán cada vez más utilizadas, abordando no sólo cuestiones técnicas, sino también implicaciones sociales. Por último, la interactividad cada vez mayor de las herramientas educativas y la popularidad de las redes sociales provocarán que las técnicas de análisis de redes sociales y de procesamiento del lenguaje natural sean cada vez más populares. Por una parte, estas técnicas permitirán el análisis cuantitativo de elementos intangibles como la participación y el conocimiento compartido. Pero por otra parte, la utilización de datos e información volcada por los usuarios también plantea nuevos problemas éticos y de confidencialidad.

V. CONCLUSIONES

El objetivo del artículo es revisar las metodologías de análisis de los Big Data en entornos educativos, partiendo de las fuentes de información para a continuación abordar las técnicas de procesamiento más relevantes. El objetivo común es poner en valor una información que ya existe almacenada en los servidores, y con la que se puede mejorar apreciablemente la efectividad de los procesos de aprendizaje.

ACKNOWLEDGEMENTS

Este trabajo ha sido financiado por la Consejería de Economía, Innovación, Ciencia y Empleo, Junta de Andalucía (Proyecto de Excelencia referencia P12-SEJ-328).

REFERENCES

[1] S. Fiore, A. D'Anca, C. Palazzo, I. Foster, D.N. Williams, G. Aloisio, "Ophidia: Toward Big Data Analytics for eScience", Procedia Computer Science, Vol. 18, pp. 2376-2385, 2013.

[2] H. Chen, R. H. Chiang, V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact", MIS Quarterly, Vol. 36, no. 4, pp. 1165-1188, 2012.

[3] S. L. Toral, M. R. Martínez-Torres, F. Barrero, "Analysis of Virtual Communities supporting OSS Projects using Social Network Analysis", Information and Software Technology, Vol. 52, Iss. 3, pp. 296-303, 2010.

[4] M.R. Martínez-Torres, S. L. Toral, F. Barrero, D. Gregor, "A text categorisation tool for open source communities based on semantic analysis", Behaviour & Information Technology, Vol. 32, Iss. 6, pp. 532-544, 2013.

[5] S. Kotsiantis, K. Patriarcheas, M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education", Knowledge-Based Systems, Vol. 23, Iss. 6, pp. 529–535, 2010.

[6] Z. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data", Proceedings of Informing Science & IT Education Conference (InSITE'2010), pp. 647-665, 2010.

[7] R. Mazza, C. Milani, Exploring usage analysis in learning systems: Gaining insights from visualisations. Workshop on Usage analysis in learning systems at 12th International Conference on Artificial Intelligence in Education, pp. 1-6. Nueva York, 2005.

[8] L. Talavera, E. Gaudioso, Mining student data to characterize similar behavior groups in unstructured collaboration spaces. Proc. 16th European Conf. Artificial Intelligence (ECAI), 2004.

[9] E. Mor, J. Minguillón, E-learning personalization based on itineraries and long-term navigational behavior. Proceedings of the 13th international world wide web conference, 2004.

[10] S. Ritter, J. R. Anderson, K. R. Koedinger & A. Corbett, "Cognitive Tutor: Applied research in mathematics education". Psychological Bulletin & Review, Vol. 14, Iss. 2, pp. 249-255, 2007.

[11] C. Reffay, T. Chanier, Social Network Analysis Used for Modelling Collaboration in Distance Learning Groups, Intelligent Tutoring Systems, Lecture Notes in Computer Science, Vol. 2363, pp 31-40, 2002.

[12] O. C. Santos, J. G. Boticario, D. Pérez-Marín, Extending web-based educational systems with personalised support through User Centred Designed recommendations along the e-learning life cycle, Science of Computer Programming, doi: 10.1016/j.scico.2013.12.004, 2014.

[13] K. S. Song, X. Hu, A. Olney, A. C. Graesser, A framework of synthesizing tutoring conversation capability with web-based distance education courseware, Computers & Education, Vol. 42, Iss. 4, pp. 375–388, 2004.

[14] R. J. Mislevy, J. T. Behrens, K. E. Dicerbo, & R. Levy, Design and discovery in educational assessment: evidence-centered design, psychometrics, and educational data mining. Journal of Educational Data Mining, Vol. 4, Iss. 1, pp. 11–48, 2012.

[15] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums, ACM Transactions on Information Systems, Vol. 26, Iss. 33, no. 12, 2008.

[16] R.S. Baker, K. Yacef, The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, Vol. 1, no. 1, pp. 3-17, 2009.

[17] J. Mostow, J. González-Brenes, & B. H. Tan, Learning classifiers from a relational database of tutor logs. Proceedings of the 4th international conference on educational data mining, pp. 149–158, 2011.

[18] J. Mccuaig, J. & Baldwin, Identifying successful learners from interaction behaviour. In Proceedings of the 5th international conference on educational data mining, pp. 160–163, 2012.

[19] Z. A. Pardos & N. T. Heffernan, Navigating the parameter space of bayesian knowledge tracing models: visualizations of the convergence of the expectation maximization algorithm. In Proceedings of the 3rd international conference on educational data mining, pp. 161–170, 2010.

[20] R. Nugent, N. Dean, & E. Ayers, Skill set profile clustering: the empty kmeans algorithm with automatic specification of starting cluster centers. Proceedings of the 3rd international conference on educational data mining, pp. 151–160, 2010.

[21] C. Patarapichayatham, A. Kamata, & S. Kanjanawasee, Evaluation of model selection strategies for cross-level two-way differential item functioning analysis. Educational and Psychological Measurement, Vol. 72, Iss. 1, pp. 44–51, 2012.

[22] B. Shih, K. Koedinger, and R. Scheines, Optimizing Student Models for Causality. Proceedings of the 13th International Conference on Artificial Intelligence in Education, Los Angeles, California, USA, 2007.

[23] J. Sabourin, J. Rowe, B. Mott, J. Lester, When Off-Task in On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. Proceedings of the 15th International Conference on Artificial Intelligence in Education, 534-536, 2011.

[24] M. R. Martinez-Torres, A Genetic Search of Patterns of Behaviour in OSS Communities, Expert systems with applications, Vol. 39, no. 18, pp. 13182-13192, 2012.

[25] S.L. Toral, M.R. Martínez-Torres, F. Barrero, Analysis of Virtual Communities supporting OSS Projects using Social Network Analysis, Information and Software Technology, Vol. 52, Iss. 3, pp. 296-303, 2010.

[26] S.L. Toral, M.R. Martínez-Torres, F. Barrero, Virtual Communities as a resource for the development of OSS projects: the case of Linux ports to embedded processors, Behaviour and Information Technology, Vol. 28, Iss. 5, pp. 405-4119, 2009.

[27] S. Valverde, G. Theraulaz, J. Gautrais, V. Fourcassie, R.V. Sole, Self-organization patterns in wasp and open source communities, IEEE Intelligent Systems, Vol. 21, Iss. 2, pp. 36-40, 2006.

[28] S.L. Toral, M.R. Martínez-Torres, F. Barrero, Modelling mailing list behaviour in open source projects: the case of ARM embedded Linux, Journal of Universal Computer Science, Vol. 15, no. 3, pp. 648–664, 2009.

[29] G. Salto, and M. J. McGill, An Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[30] D. Cai, X. He, and J. Han, Document Clustering Using Locality Preserving Indexing, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, Iss. 12, pp. 1624-1637, 2005.

[31] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman,. Indexing by latent semantic analysis, Journal of the American Society of Information Science, Vol. 41, Iss. 6, pp. 391–407, 1990.

[32] G. Dyke, D. Adamson, I. Howley, & C. P. Rosé, Enhancing scientific reasoning and explanation skills with conversational agents. IEEE Transactions on Learning Technologies, Vol. 6, Iss. 3, pp. 240–247, 2013.

[33] S. K. D'Mello, S.D. Craig, A. W. Witherspoon, B. T. McDaniel, and A. C. Graesser, Automatic Detection of Learner's Affect from Conversational Cues. User Modeling and User-Adapted Interaction, Vol. 18, Iss. 1-2, pp. 45-80, 2008.