Búsqueda y Recomendación de contenido educativo en entornos virtuales de aprendizaje

Beatriz Fernández-Reuter[1,2], Elena Durán[1], Analía Amandi[2,3]

[1]Instituto de Investigaciones en Informáticas y Sistemas de Información (IIISI) - Facultad de Ciencias Exactas y Tecnologías (FCEyT) -
Universidad Nacional de Santiago del Estero (UNSE), Santiago del Estero
[2] CONICET, Comisión Nacional de Investigaciones Científicas y Técnicas
[3] ISISTAN, Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), Tandil
e-mail: bfreuter@unse.edu.ar; eduran@unse.edu.ar; amandi@exa.unicen.edu.ar

Resumen. Los entornos de aprendizaje virtual utilizados tanto en el desarrollo de cursos a distancia, como en el soporte al dictado de asignaturas presenciales, poseen una gran cantidad de información útil para el desarrollo de las diferentes actividades desempeñadas por los estudiantes en las distintas etapas del proceso de aprendizaje. Sin embargo, carecen de medios de búsqueda adecuados dentro del entorno, por lo que se dificulta encontrar la información para todos aquellos estudiantes que poseen una duda puntual. Si a estos entornos se les incorpora un mecanismo de búsqueda y recomendación que combine técnicas de inteligencia artificial, se pueden convertir en una herramienta muy potente que ayude a los estudiantes a evacuar sus dudas. Es por esto, que el presente trabajo propone un Enfoque para la Búsqueda y Recomendación que combina técnicas de Minería de Contenido Web y Recuperación de Información, para asistir a los estudiantes en la búsqueda de información adecuada, a partir del ingreso de una consulta.

Palabras Claves: Sistemas de Recomendación, Minería de Contenido Web, Recuperación de la Información, Entornos virtuales de aprendizaje.

1 Introducción

Los entornos virtuales de aprendizaje o entornos de e-learning son espacios donde tanto los docentes como los alumnos comparten material útil para la adquisición de conocimiento y el desarrollo de competencias dentro de una asignatura particular. Estos entornos proporcionan también un medio de consulta frecuente a través de los foros de discusión, y un espacio para el desarrollo de actividades pedagógicas y la producción individual y colaborativa de conocimiento. Sin embargo, a medida que se incrementa el material educativo disponible dentro del entorno virtual, se dificulta encontrar información valiosa para todo aquel estudiante que tenga una duda puntual, debido a que estos no proporcionan medios de búsqueda apropiados. Generalmente

estos entornos solo cuentan con buscadores que les permiten recuperar la información revisando los títulos del material pero no su contenido.

Para dar solución a este problema, se pueden utilizar técnicas de Recuperación de la Información, que permiten la búsqueda de documentos en base a la información o metadatos contenidos en los mismos [1]. Sin embargo, generalmente, solo una fracción reducida de esos documentos tiene información relevante para un usuario dado, y además puede haber contenido adicional que esté relacionado a lo que desea buscar, pero que no coincida específicamente con la consulta ingresada.

Ante esto, se podría combinar con las técnicas que utilizan los Sistemas de Recomendación, herramientas que ayudan al usuario a identificar el ítem mas interesante o relevante de un inmenso conjunto de elementos [2]. Estos sistemas, para realizar sus recomendaciones en plataformas web, se valen de diversas técnicas de Inteligencia Artificial, como por ejemplo, la Minería de Contenido Web que permite descubrir información útil desde los contenidos de la web, sean textos, imágenes, audio y video [3]. La combinación de estas técnicas permite que el estudiante al ingresar una consulta, no solo recupere aquella información que coincida con dicha consulta, sino que además, pueda obtener todo el material relacionado en cierta medida con la misma.

En este trabajo se presenta un enfoque que combina técnicas de Minería de Contenido Web y de Recuperación de la Información para la búsqueda y recomendación de todo el material disponible en entornos virtuales de aprendizaje, independientemente del formato en que se encuentre, y teniendo en cuenta el contenido de dicho material.

En las siguientes secciones se presentan algunos antecedentes de trabajos similares, se describe en detalle el enfoque propuesto y la validación del enfoque, aplicando el mismo a la recomendación de material educativo, a alumnos universitarios, en aulas virtuales de una asignatura particular.

2 Antecedentes

En el área de la educación, existen numerosos sistemas de recomendación que sirven de soporte en las diferentes etapas del aprendizaje. Algunos ejemplos que se pueden mencionar son, Pinter et al. [4] quienes desarrollaron un sistema de recomendación para guiar al estudiante en el proceso de aprendizaje, proponiendo los caminos de aprendizaje más efectivos basándose en técnicas de filtrado colaborativo, solicitando opiniones acerca de las actividades que realizaron y del material que utilizaron. Otro caso es el de Rule et al. [5] quienes utilizando ontologías y reglas de filtrado, sugieren a los estudiantes algunos temas que necesitan aprender basándose en su nivel de conocimientos, perfil del estudiante y algunas evaluaciones que realizan a los mismos. En la propuesta de Sun y Xie [6] demuestran como la minería de uso web es útil en la recomendación de links a visitar dentro de un ambiente de e-leraning. Otra aplicación es la de Souali et al. [2] quienes se valen de sistemas de filtrado basados en contenidos para recomendar material de estudio y lecciones, a partir de una solicitud del estudiante.

Si bien se puede observar que existen trabajos que utilizan la minería web para la recomendación o bien, que aplicando otras técnicas recomiendan a partir del ingreso

de una consulta por parte de los estudiantes, ninguno de estos combina las técnicas aquí propuestas. Se espera que combinando ambas técnicas, los estudiantes que posean una duda puntual, obtengan mejores resultados a la hora de buscar información y que además cuenten con la posibilidad de acceder a material adicional que esté relacionado a su consulta.

3 El enfoque propuesto

El enfoque que se propone en este trabajo combina técnicas de Análisis y Recuperación de Información y de Minería de Contenido Web para recomendar al estudiante material de estudio disponible en un entorno virtual de aprendizaje, a partir del ingreso de una consulta.

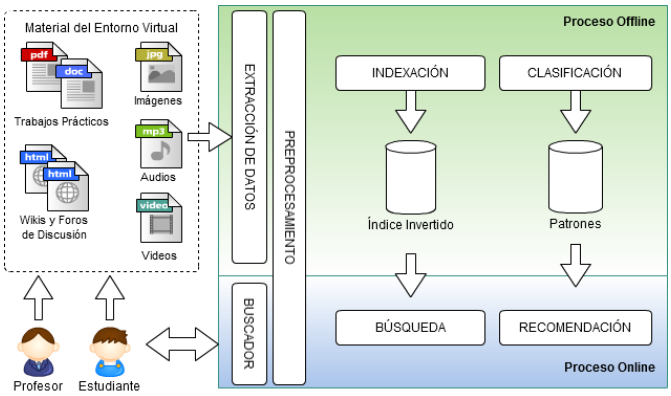

Fig. 1. Búsqueda y Recomendación de contenido para entornos virtuales de Aprendizaje

Como se puede observar en la **Fig. 1**, se presentan dos procesos fundamentales, uno offline de Indexación y Clasificación y otro online de Búsqueda y Recomendación. Para esto, se toma como entrada el material disponible, por ejemplo dentro de aula virtual (archivos de contenido, wikis, foros de discusión, cuestionarios, glosarios, etc) de una asignatura, en cualquier formato: textos, imágenes, audios y/o videos. Esta información, pasa por una etapa de Extracción de Datos y/o de Preprocesamiento, que la prepara para la ejecución de los dos procesos arriba mencionados. A continuación se describe en detalle cada uno de los procesos propuestos en el enfoque.

3.1 Proceso Offline

Este proceso consiste en preparar toda la información que se utilizará para la ejecución del Proceso Online, es decir, para la Búsqueda y Recomendación. Para esto, genera como salida el índice en donde se realizará la búsqueda del material y el modelo de patrones que permitirá realizar la recomendación de material adicional.

Debe ser ejecutado periódicamente o bien, o cada vez que los docentes o alumnos publiquen nuevo material educativo en el entorno virtual de aprendizaje, a fin de ac-

tualizar el índice y, permitir así que mejore la precisión en el reconocimiento de patrones en el modelo de clasificación.

Este proceso consta de una serie de actividades que se describen a continuación.

Extracción de Datos: Dado que el contenido de un entorno virtual de aprendizaje puede ser de diferentes tipos, textuales, imágenes, audios, videos, metadatos e hipervínculos, y posee datos no estructurados (doc, .pdf, etc.), muy poco estructurados (HTML), semi-estructurados (XML) y estructurados (bases de datos); es preciso obtener todos los datos textuales del contenido presente en el entorno virtual, para luego poder procesarlo. Para ello, se utilizó Apache Tika [7], un software que posee un conjunto de herramientas que permitan detectar y extraer metadatos y contenido de texto estructurado desde varios tipos de documentos, usando librerías de parsing existentes.

Preprocesamiento: Para que la recuperación de información sea efectiva, es necesario preparar los datos recuperados para su posterior indexación y clasificación.

Sobre el texto extraído en la etapa anterior, se corrigieron, errores ortográficos y de tipeo, se reemplazaron abreviaturas por palabras completas, se convirtieron a minúsculas, se eliminaron acentos y caracteres especiales.

Para mejorar la performance tanto de la clasificación como del tiempo de búsqueda se eliminaron los stopwords [1] y se aplicó Stemming [1]. Esto último permite ampliar la consulta con las variantes morfológicas de los términos usados. Para estas dos últimas actividades mencionadas, se utilizaron las herramientas de Lucene [1] y Weka [8] y se redujo notablemente el espacio de términos.

Como resultado de este módulo se obtuvo un listado de términos o tokens, conformado por los stem de las palabras contenidas en cada uno de los documentos.

Esta actividad es común para los Proceso Offline y Online.

Indexación: A partir del listado de tokens obtenido en el Preprocesamiento, se generó un índice inverso de términos utilizando la librería de Lucene. Este índice contiene una lista de términos presente en cada uno de los documentos, un enlace a los documentos en donde se encuentra la misma y el peso del término, dado por la frecuencia de aparición.

Clasificación: Consiste en analizar el contenido recuperado del entorno virtual de aprendizaje, a fin de generar y entrenar un modelo capaz de reconocer patrones en los documentos. Este modelo será utilizado en el Proceso Online para clasificar las consultas ingresadas por los estudiantes, determinando el tema al que pertenecen y así poder recomendar material adicional.

Para la construcción del modelo de clasificación se empleó el método de Máquina de Vectores Soporte (SVM, del inglés Support Vector Machine), ampliamente utilizado en problemas de categorización de texto por ser rápido y efectivo [9], [10].

Toma como entrada los tokens obtenidos durante el Preprocesamiento, conformando un vocabulario y les asigna un peso que indica su importancia o contribución en la regla de clasificación. Este método fue implementado en el software Weka ejecutando el algoritmo SMO (Sequential Minimal Optimization) [11]. Este software al aplicar el dicho algoritmo a problemas multi-clases utiliza la clasificación por pareja (1-vs-1). Es decir que construye ***n*** clasificadores binarios por las ***n*** combinaciones posibles de las clases que se utilizaron para la clasificación. Se utilizó un Kernel polinomial del tipo $K(x,y) = <x, y>^p$ con el exponente igual a 1, conformando así un Kernel lineal

K(x,y) = <x,y>. Además se utilizaron márgenes duros, es decir, un parámetro C = 1 [3].

3.2 Proceso Online

Este proceso trabaja de forma online interactuando directamente con los estudiantes. Tiene la finalidad de brindar todo el material educativo del entorno virtual de aprendizaje que coincida con cierto criterio de búsqueda ingresado por el alumno, y de recomendar material adicional que esté relacionado.

El proceso se puede describir básicamente en cinco pasos:

1. El alumno ingresa una serie de palabras claves del tema que desea buscar.
2. Se realiza un preprocesamiento a estas palabras según lo explicado anteriormente.
3. Un motor de búsqueda desarrollado con la librería Lucene, traduce las palabras ingresadas por el estudiante a una forma que sea interpretado por la librería. Busca en el índice inverso generado durante la ejecución del Proceso Offline y devuelve todo el material educativo que satisfaga dicha consulta.
4. Utilizando el modelo generado en la Clasificación, se determina a que temas (dentro de un conjunto de temas predefinidos) pertenecen las palabras claves ingresadas por el estudiante para la búsqueda.
5. Se recupera todo aquel material del entorno virtual de aprendizaje que pertenezca al tema determinado en el paso 4 y que no haya sido incluida en el conjunto devuelto por el buscador Lucene en el paso 2.

4 Evaluación y Resultados

Para evaluar el enfoque propuesto se trabajó con el material extraído de las aulas virtuales implementadas durante los años 2011, 2012 y 2013, correspondientes a asignaturas vinculadas a la temática Simulación, perteneciente a tres universidades argentinas (Universidad Nacional de Santiago del Estero, Universidad Católica de Santiago del Estero y Universidad Nacional del Chaco Austral), totalizando 204 documentos en diferentes formatos. Los datos obtenidos fueron sometidos a las etapas de extracción de datos y de preprocesamiento, según se describió en la sección anterior.

Además, todo el material fue clasificado manualmente por los docentes de las cátedras, a fin de obtener una clasificación testigo contra la cual comparar la efectividad de la clasificación y recomendación. Se trabajó con 7 categorías o clases en total, descriptas a continuación:

- Tema 1 - Introducción a la Simulación: 50 documentos
- Tema 2 - Metodología de Simulación: 29 documentos
- Tema 3 - Generación de variables aleatorias: 29 documentos
- Tema 4 - Simulación de Eventos Discretos: 33 documentos
- Tema 5 - Simulación Continua con Dinámica de Sistemas: 37 documentos
- Tema 6 - Nuevas Tendencias de la Simulación: 21 documentos
- Tema 7 - Otros temas no incluidos en los anteriores: 5 documentos

Al no contar con un conjunto de documentos de pruebas, se utilizó el mismo conjunto de material educativo para entrenar, y para probar el modelo mediante una validación cruzada de 10 Folds [8].

Para evaluar la efectividad en el reconocimiento de las clases o categorías a las que pertenecen los documentos de la plataforma de e-learning, se analizaron los resultados obtenidos en la matriz de confusión [8], obtenida luego de ejecutar el algoritmo SMO en Weka, tal y como se muestra en la **Tabla 1**. Se adicionaron a la matriz la suma de Totales para facilitar la lectura de la misma. En la parte inferior totaliza las instancias clasificadas automáticamente en cada clase, y a la derecha, la cantidad total de documentos por Tema que fueron clasificadas manualmente.

	Tema 1	Tema 2	Tema 3	Tema 4	Tema 5	Tema 6	Tema 7	Total Documentos por clase
	43	2	1	2	2			50
	9	17	1	1		1		29
		1	25	3				29
	3		2	28				33
	2			1	34			37
	4	3		1	1	12		21
	2			1	2			5
Total de instancias clasificadas por clase	**63**	**23**	**29**	**37**	**39**	**13**	**0**	**204**

Table 1. Matriz de Confusión obtenida al ejecutar el algoritmo SMO en Weka

Uno de los puntos a remarcar en base al análisis de la matriz es la diferencia existente entre los clasificado como Tema 1 y Tema 2 respecto a los que efectivamente pertenece a esos temas. Esta notable diferencia se debe a que en ambos temas se tratan conceptos que tienen una granularidad mayor que los demás, es decir, no poseen contenidos específicos, sino mas bien, conceptos que son abarcados durante todo el dictado de la asignatura.

Lo ocurrido con el Tema 6 y en mayor medida, con el Tema 7, puede ser producto de la diferencia de documentos utilizados para el entrenamiento, respecto a las otras categorías. Es decir que, para entrenar el modelo, sólo se utilizaron 21 y 5 documentos respectivamente, a diferencia de los demás, que contaban con más de 29 documentos. Estas diferencias en el conjunto de entrenamiento, suelen perjudicar el reconocimiento de patrones, debido a que el algoritmo tiene pocos ejemplos que le permitan reconocer a una consulta con mayor precisión como perteneciente a una clase o a otra.

Los demás Temas no presentaron una variación importante entre sus clasificaciones, ya que muestran una diferencia que no supera las 5 instancias, entre la cantidad total que corresponden a cada una y lo efectivamente clasificado.

Continuando con la evaluación de la efectividad del clasificador, se calcularon además, las métricas de Precisión, Recall y Valor-F [8]. El algoritmo logró clasificar el 77.94% de las instancias correctamente, con una Precisión = 0.77, un Recall = 0.779 y un Valor-F = 0.766.

Con respecto al motor de búsqueda de Lucene, se ejecutaron una serie de consultas, determinando para cada una de ellas la cantidad de documentos relevantes recupe-

rados en relación al total de documentos recuperados y al total de documentos relevantes para la consulta. Con estos datos, se calcularon las métricas de precisión y recall [1] de cada ejecución. Los resultados obtenidos se muestran en la **Fig. 2**.

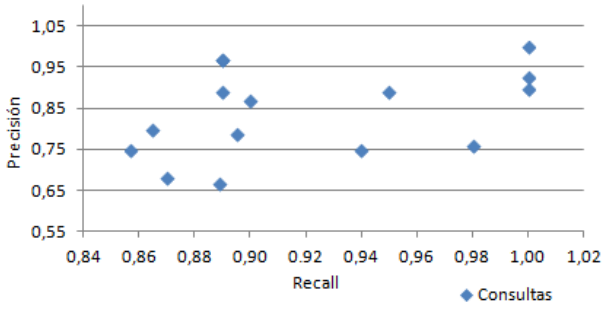

Fig. 2. Gráfico de Precisión y Recall del Buscador Lucene

Como se puede observar los valores de Recall siempre superan el 85%, es decir que todas las consultas devuelven la mayor parte de los documentos relevantes para la misma. Sin embargo, en algunas de estas consultas los valores de Precisión apenas superan el 65%, lo que se traduce a una enorme cantidad de documentos recuperados. Estos valores se obtuvieron al ejecutar búsquedas por varias palabras claves de forma conjunta sin indicar la conexión lógica entre ellas, sobre todo cuando alguna de estas palabras estaban relacionadas a un concepto general de la asignatura. Al modificar las consultas especificando claramente la conexión lógica entre las palabras, se obtenían nuevamente valores de Precisión y Recall altos.

Actualmente, se está trabajando en la validación de la recomendación con los alumnos que cursan las asignaturas de Simulación en las universidades antes mencionadas.

5 Conclusión

A partir de los experimentos realizados, se demostró que es factible la aplicación del enfoque para la búsqueda y recomendación de material educativo en entornos virtuales de aprendizaje, combinando técnicas de Minería de Contenido Web y de Recuperación de la Información.

Con el enfoque propuesto es posible asistir a los estudiantes en la búsqueda de material educativo relacionado a temáticas de la asignatura, independiente del formato en el que se encuentre, recomendado además, información sobre otros contenidos relacionados al tema en cuestión.

A corto plazo, se contará con los resultados de las pruebas de la función de recomendación, lo que permitirá determinar que tan efectiva resultaría la misma de acuerdo a la clasificación propuesta.

A mediano plazo, se prevé también, replantear las clases seleccionadas para la clasificación, de manera que todas tengan el mimo nivel de granularidad y que sus con-

tenidos se solapen lo menos posible. Se considera que esto ayudaría a mejorar aún más la efectividad en la clasificación, y por lo tanto en la recomendación. Además, teniendo en cuenta que las técnicas utilizadas para la clasificación están más orientadas a la aplicación sobre grandes volúmenes de datos, a diferencia de la utilizada en este trabajo que cuenta con poca cantidad de información, se pretende seguir incorporando datos históricos, correspondientes al material de aulas virtuales de años anteriores, así como también de objetos de aprendizaje vinculados a la temática.

Se está trabajando, también, para optimizar los resultados obtenidos en las búsquedas incorporando un enfoque semántico.

Además, es importante destacar que, sobre la base del enfoque presentado en este trabajo, se podrían incorporar nuevas características para mejorar y personalizar las recomendaciones a los estudiantes, de acuerdo a características personales y de comportamiento del mismo.

6 Referencias

1. M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in action*, Second Edi. Manning Publications Co, 2010, p. 528.
2. K. Souali, A. El Afia, R. Faizi, and R. Chiheb, "A New Recommender System For E-Learning Environments," in *International Conference on Multimedia Computing and Systems (ICMCS)*, 2011, pp. 1–4.
3. J. Hernandez Orallo, M. J. Ramirez Quintana, and C. Ferri Ramirez, *Introducción a la Minería de Datos*. Editorial Pearson, 2004, p. 680.
4. R. Pinter, T. Marušić, D. Radosav, and S. Maravić Čisar, "Recommender System in E-student web-based adaptive educational hypermedia system," in *MIPRO*, 2012, pp. 1270–1273.
5. S. Shishehchi, S. Y. Banihashem, and N. A. M. Zin, "A Proposed Semantic Recommendation System for E-learning - A Rule and Ontology Based E-learning Recommendation System," in *Information Technology (ITSim), 2010 International Symposium*, 2010, pp. 1–5.
6. J. Sun and Y. Xie, "A Recommender System Based on Web Data Mining for Personalized E-Learning," *2009 Int. Conf. Inf. Eng. Comput. Sci.*, pp. 1–4, Dec. 2009.
7. C. A. Mattmann and J. L. Zitting, *Tika in action*. Manning Publications Co., 2012, p. 257.
8. R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, *WEKA Manual for Version 3-6-9*. University of Waikato, Hamilton, New Zealand, 2013, p. 303.
9. T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *10th European Conference on Machine Learning Chemnitz, Germany*, 1998, vol. 1398, pp. 137–142.
10. R. Feldman and J. Sanger, *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*. 2007, p. 423.
11. J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burge, and A. Smola, Eds. 1998.