Buscador automático de material educativo en aulas virtuales

Beatriz Fernández Reuter y Elena Durán
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
e-mail: bfreuter@unse.edu.ar; eduran@unse.edu.ar

Resumen

Las Aulas Virtuales conforman una herramienta fundamental tanto de soporte para el dictado de asignaturas presenciales como a distancia. Esto es en parte gracias a todo el material disponible, en formatos de texto, video, imágenes, etc., que tanto docentes como alumnos publican en la misma. Sin embargo, carecen de un medio de búsqueda eficiente y que permita al estudiante obtener toda la información necesaria cuando posee una duda puntual. Es por esto, que en el presente trabajo se propone un Buscador automático de material educativo en aulas virtuales que además recomienda contenido adicional que tenga relación a una consulta ingresada por un estudiante.

Palabras clave: Análisis y Recuperación de la Información, Aula Virtual, Sistemas de Recomendación, Minería de Contenido Web

1 - Introducción

Hace algunos años se ha generalizado el uso de aulas virtuales tanto para el dictado de cursos a distancia, como de soporte y ayuda al dictado de asignaturas presenciales. Dentro de estas aulas, tanto los docentes como los alumnos comparten material útil para la asignatura, en diferentes formatos, así como también proporcionan un medio de consulta frecuente a través de los foros de discusión, en los cuales un estudiante puede realizar una pregunta, a la que el docente, o los mismos estudiantes, responden. Sin embargo, a medida que se incrementa el material educativo disponible dentro del aula virtual, se dificulta encontrar información valiosa para todo aquel estudiante que tenga una duda puntual. Esta dificultad, es en parte debido a que la mayoría de estas Aulas, no cuentan con un mecanismo de búsqueda, o bien, no brinda una buena solución dado que sólo realiza la búsqueda en base a los títulos del material presente en el aula y no en su contenido.

Para dar solución a este problema, se pueden aplicar técnicas de Recuperación de la Información. Las mismas consisten en el descubrimiento automático de documentos relevantes de acuerdo a un cierto criterio de búsqueda [3]. Es decir, que a partir del ingreso de una consulta, se recuperan todos aquellos documentos que tengan en su contenido alguna coincidencia con dicha consulta.

Con frecuencia un estudiante busca el material que le interesa en un lugar puntual dentro del aula virtual, por ejemplo entre los archivos asociados a un tema dentro de un aula virtual soportada en la plataforma MOODLE. Sin embargo, puede ocurrir que haya aportes importantes realizados por el docente o por alguno de sus compañeros en un foro, en una wiki, o en algún otro espacio del aula virtual. En estos casos, cobra importancia el uso de Sistemas de Recomendación como complemento de la búsqueda de información. Estos sistemas son aplicaciones inteligentes que asisten al usuario en el proceso de toma de decisiones, cuando debe escoger un ítem entre un inmenso conjunto de potenciales productos o servicios [13]. Para realizar sus recomendaciones se valen de diversas técnicas de Inteligencia Artificial como la Minería de Datos, que se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [4]. Dado que el contenido de las aulas virtuales se encuentran en la Web, el uso de técnicas de minería de datos para el

descubrimiento de información útil desde los contenidos textuales y gráficos de los documentos Web, es conocido como Minería de Contenido Web [4].

El objetivo del presente trabajo es presentar un sistema para buscar y recomendar al estudiante material de estudio disponible en el aula virtual de una asignatura determinada, utilizando técnicas de Análisis y Recuperación de Información y de Minería de Contenido Web.

En las siguientes secciones se presentan algunos antecedentes de trabajos similares, se describe en detalle el sistema propuesto, indicando los módulos que lo componen, la funcionalidad de cada módulo, así como las técnicas y algoritmos de recuperación de información aplicados. Se describe además el proceso de validación del sistema, aplicando el mismo a la recomendación, a estudiantes universitarios, de material educativo en el campo de la Simulación por Computadora.

2 - Antecedentes

Respecto a la búsqueda de contenido educativos, Jun et. al [6] proponen un framework para la búsqueda de documentos almacenados en diferentes formatos, en el Centro de Recursos Educativos de Shanghai, una base de recursos de alta calidad que ayuda a las personas a obtener la información educativa que requiere. Otro trabajo similar es el de Shao et al [11], quienes proponen un motor de búsqueda para todo tipo de recurso educativo almacenado en el repositorio de una Universidad. Puustjärvi y Pöyry [10] proponen un sistema de recuperación de información en lo que ellos denominan Universidad Virtual, un portal único donde los estudiantes pueden encontrar los objetos de aprendizaje provenientes de diferentes aulas virtuales.

Con relación a la combinación de los buscadores con los sistemas de recomendación, se puede mencionar la propuesta de Souali et al. [12] quienes se valen de sistemas de filtrado basados en contenidos para recomendar material de estudio y lecciones, a partir de una solicitud del estudiante.

De la revisión de antecedentes realizada se puedo concluir que si bien existen numerosos trabajos que utilizan la recuperación de la información sobre materiales educativos, sin embargo, ninguno de estos aplica este enfoque dentro de un Aula Virtual, donde el estudiante posee toda la información necesaria y específica de una asignatura. Además, sólo uno de estos combina la búsqueda con los sistemas de recomendación, pero lo hace solo aplicando técnicas de filtrado basado en contenido.

Es por esto, que el presente trabajo propone un Buscador de todo el material educativo incluido en el Aula virtual de una asignatura, recomendando además, mediante el uso de técnicas de Minería de contenido Web, otro material complementario que esté relacionado a la consulta ingresada por el estudiante.

3 – Descripción del sistema propuesto

El sistema de recomendación automática propuesto en este trabajo, tiene el objetivo de utilizar el material educativo disponible en el aula virtual de una asignatura, para guiar al estudiante en la búsqueda de información. Para esto, el estudiante ingresa una o varias palabras clave vinculadas al tema sobre el que tiene dudas y que desea consultar. El sistema ejecuta la búsqueda y devuelve todo el material y las preguntas del Foro de Discusión, donde se encuentren las palabras clave ingresadas por el estudiante. Además, identifica a qué temática corresponden dichas palabras, con el fin de recomendar otro material que esté relacionado y que pueda ser de interés para el estudiante.

Como se puede observar en la **Figura 1**, el sistema toma como entrada todo el material disponible en la Aula virtual de una asignatura (trabajos prácticos, actividades, dispositivas de clase, wikis, videos, imágenes, etc.) en cualquier formato, como pdf, doc, jpg, avi, mp3, etc., y las interacciones realizadas en los foros de discusión. Esta información, pasa por una etapa de Preprocesamiento, que la prepara para las etapas siguientes de Indexación-Búsqueda y Clasificación-Recomendación.

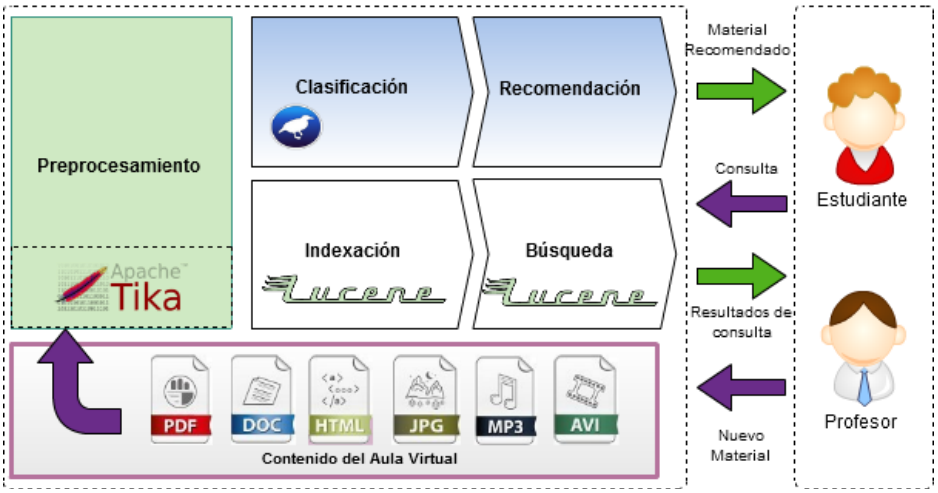

Figura 1: Arquitectura del buscador automático de material educativo en aulas virtuales

A continuación se describe en detalle cada uno de los módulos que componen al sistema.

3.1 - Preprocesamiento

El preprocesamiento comienza con la detección y extracción de los metadatos y el contenido del texto de todo el material del aula virtual de la asignatura, utilizando el software Apache Tika [8]. Este software posee un conjunto de herramientas que permiten detectar y extraer metadatos y contenido de texto estructurado desde varios tipos de documentos.

Sobre el texto extraído, se corrigieron, de forma manual, errores ortográficos y de tipeo y se reemplazaron abreviaturas por palabras completas. Luego, de forma automática se convirtieron a minúsculas, se eliminaron acentos y caracteres especiales. Para mejorar el rendimiento del sistema se eliminaron las palabras irrelevantes tales como los artículos y se sustituyeron las palabras por su raíz (*stem*) [3], lo que permite ampliar la consulta con las variantes morfológicas de los términos usados.

Para estas dos últimas actividades mencionadas, se utilizaron las herramientas de Lucene[9] y Weka[1] 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. En la **Figura 2** se puede ver parte del listado de términos obtenido.

No.	Name
1	_Clase_
2	abiert
3	abstraccion
4	abstract
5	accion
6	actitud
7	activ
8	actual
9	acuerd
10	adecu
11	administracion
12	admit
13	advanc
14	agreg
15	agregacion

Figura 2: Listado de términos obtenidos en el Preprocesamiento

3.2 - Indexación

A partir del listado de términos obtenido en el módulo de Preprocesamiento, se generó un índice inverso de términos utilizando la librería de Lucene. El índice generado contiene una lista de términos presente en cada uno de los documentos, un enlace a los documentos en donde se encuentra el mismo y el peso del término, dado por la frecuencia de aparición. En la **Figura 3** se puede ver parte del índice generado.

Cabe aclarar que este módulo debe ser ejecutado cada vez que se agrega nuevo material al Aula Virtual.

Rank	Freq	Field	Text
1	65	contenido	simulacion
2	51	contenido	sistem
3	34	contenido	model
4	29	contenido	variabl
5	25	contenido	metod
6	22	contenido	diferenci
7	21	contenido	pregunt
8	21	contenido	diagram
9	19	contenido	aleatori
10	18	contenido	dinamic
11	17	contenido	objetiv
12	16	contenido	simular
13	15	contenido	numer
14	15	contenido	demor
15	14	contenido	discret
16	13	contenido	aplicar

Figura 3: Parte del índice generado por Lucene

3.3 - Clasificación

Este módulo consiste en analizar el contenido del Aula Virtual a fin de generar y entrenar un modelo capaz de reconocer patrones en los documentos. Este modelo será utilizado en el módulo de Recomendación para clasificar las consultas ingresadas por los estudiantes, determinando la unidad temática de la asignatura a la que pertenecen y así poder recomendar material adicional.

Al igual que en el modulo de Indexación, el modelo resultante de la clasificación se deberá reconstruir cada vez que un docente o alumno agregue material al Aula virtual a fin de que adquiera mayor precisión en el reconocimiento de patrones. 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[2, 5].

3.4 - Búsqueda

Este módulo consta de un motor de búsqueda cuyo objetivo es proveer al estudiante una interfaz mediante la cual pueda ingresar una consulta y recupere todo el material relacionado.

Este componente fue desarrollado con la librería Lucene, lo que permite traducir las palabras ingresadas por el estudiante, a una forma que sea interpretada por la librería y pueda buscar en el índice todos los documentos que satisfagan dicha consulta. Para esto, la consulta debe pasar primero por un preprocesamiento como el descripto en el punto 3.1.

3.5 - Recomendación

El módulo de Recomendación, tiene la finalidad de proponer material educativo relacionado a la consulta que realizó el estudiante y que pueda ser de interés. Para esto, al momento que un estudiante realiza una consulta, el sistema preprocesa la misma de acuerdo a lo descripto en el punto 3.1, y al resultado de esto lo clasifica haciendo uso del modelo generado en el punto 3.3, a fin de determinar a qué categoría o clase pertenece. Luego, busca y recupera todo el material educativo de dicha categoría que no haya sido incluido por el Buscador y lo propone como material adicional.

4 - Evaluación y Resultados

La evaluación del sistema se llevó a cabo en dos etapas. En la primera se evaluó el Módulo de Clasificación y en la segunda el Módulo de Indexación y Búsqueda.

Actualmente, se está trabajando en la validación del Módulo de Recomendación con los estudiantes que cursan las asignaturas de Simulación en la Universidad Nacional de Santiago del Estero, Universidad Católica de Santiago del Estero y Universidad Nacional del Chaco Austral.

4.1- Evaluación del Módulo de Clasificación

Para la construcción de un modelo de clasificación, es necesario contar con un volumen importante de información, ya que esto aporta más precisión en el reconocimiento de patrones. Por este motivo, para contar con mayor número de documentos, se trabajó con el material extraído de las aulas virtuales de los años 2011, 2012 y 2013, correspondientes a asignaturas vinculadas a la temática Simulación, perteneciente a las tres universidades antes mencionadas, totalizando 204 documentos en formato PDF, DOC, TXT, HTML, JPG y AVI.

Con los datos obtenidos se ejecutaron las etapas de preprocesamiento, y clasificación, según se describió en la sección 3.1 y 3.3. 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 del sistema de recomendación. Se trabajó con 7 categorías o clases en total, una por cada uno de los seis temas centrales de estas asignaturas, según lo indicaron los docentes responsables de los espacios curriculares. Se agregó una séptima categoría para material complementario que no está específicamente relacionado a los temas centrales. Las categorías resultaron de la siguiente manera:

- ***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 prueba, se utilizó el mismo conjunto de material educativo para entrenar, y para probar el modelo.
Para evaluar la efectividad en el reconocimiento de las clases o categorías a las que pertenecen los documentos del aula virtual, se analizaron los resultados obtenidos entre las instancias clasificadas correctamente y aquellas instancias que fueron clasificados de forma incorrecta, tal y como se puede ver en la **Figura 4**. Además, se compararon la cantidad de documentos que se clasificaron en un determinado Tema con los que efectivamente pertenecían al mismo. Esto último se puede ver en la **Figura 5**.

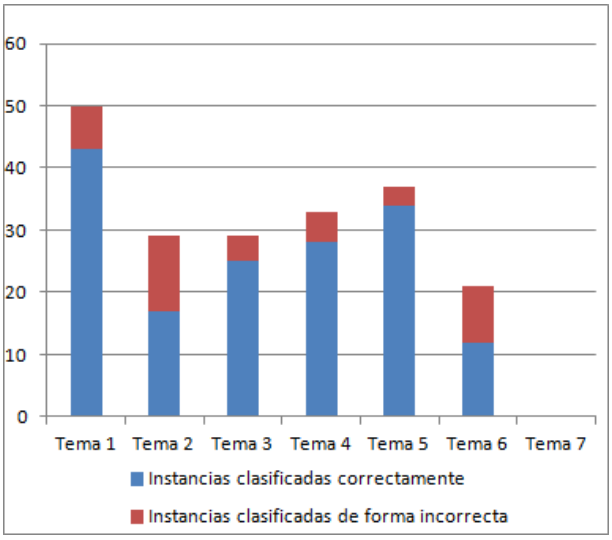

Figura 4: Total de instancias clasificadas correctamente frente a las clasificadas incorrectamente

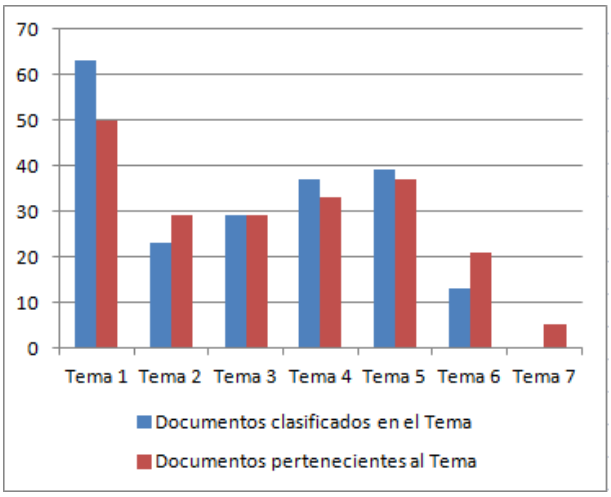

Figura 5: Total de documentos clasificados en cada Tema frente a los efectivamente pertenecientes al mismo

Como se puede observar, en el *Tema 1* se clasificaron un total de 63 instancias de las cuáles, solo 43 fueron correctamente clasificadas. Esto no se puede considerar un error en sí, debido a que se trata de una categoría que contiene la introducción a los principales conceptos de la asignatura y que luego son tratados en mayor detalle en Temas posteriores. Con respecto al *Tema 2*, solo se lograron clasificar 17 de las 29 instancias de forma correcta y 6 instancias fueron clasificadas incorrectamente como pertenecientes a esta categoría. Esto, al igual

que en el *Tema 1*, se debe a que los conceptos tratados en dichos temas, 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.
El *Tema 6* presenta una diferencia significativa entre lo clasificado correctamente y lo que verdaderamente pertenece a la categoría. Esto 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 documentos pertenecientes al *Tema 6*, por lo que se diferencia notablemente 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, es decir, en el *Tema 3*, 25 instancias clasificadas correctamente de 29; en el *Tema 4*, 28 de 33 y en el *Tema 5*, 34 de 37. No existen instancias clasificadas en el *Tema 7*, lo que se supone se debe a la baja cantidad de documentos que correspondían a este Tema.
Continuando con la evaluación de la efectividad del clasificador, se calcularon además, las métricas de Precisión, Recall y Valor-F. La precisión mide la proporción de los valores que fueron clasificados correctamente sobre el total de los clasificados en una determinada clase. El recall mide la proporción de los ejemplos clasificados correctamente en una clase y los que efectivamente pertenecen a la misma. El Valor-F es simplemente una métrica que combina tanto la precisión como el recall [1].
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.

4.2 - Evaluación del Módulo de Búsqueda
Para la generación del índice inverso descripto en el Módulo de Indexación, se tomaron de forma individual las aulas virtuales de las diferentes universidades descriptas anteriormente. Esto, a diferencia del módulo de clasificación, no requiere de cierto volumen de información, por el contrario, debe estar basado en el contenido de cada aula, ya que guarda un enlace al documento, para poder acceder en caso que sea el material buscado por el estudiante. Esto es que, no se puede indexar documentos pertenecientes a un aula virtual y ser accedido desde otra aula y mucho menos si pertenece a otra universidad.
Luego de realizar el preprocesamiento y generar el índice, como se describieron en los puntos 3.1 y 3.2, se ejecutaron diversas consultas, utilizando el Módulo de Búsqueda. Sobre los resultados obtenidos por el buscador, se calcularon las métricas de precisión y recall obteniendo muy buenos resultados tal como lo muestra la **Figura 6**. En este caso, la precisión hace referencia a la fracción del total de documentos relevantes recuperados sobre el total de documentos recuperados, y el recall determina el total de documentos relevantes recuperados sobre el total de documentos relevantes[7].
En muchos casos, cuando la búsqueda estaba enfocada en un tema particular de la asignatura, tanto la Precisión como el Recall eran igual a 1.
Sin embargo, al hacer búsquedas de varias palabras claves de forma conjunta sin indicar la conexión lógica entre ellas, sobre todo si alguna de estas palabras no era un tema puntual de la asignatura, sino mas bien, un concepto general y transversal de la misma, se obtenía un Recall alto, pero la Precisión se reducía notablemente. Esto se debe a que el conector lógico por defecto utilizado en el buscador, es el OR, lo que produce que el buscador recupere todos los documentos que contengan al menos una de las palabras claves

ingresadas. Al realizar consultas especificando claramente la conexión lógica entre las palabras, se obtenían nuevamente valores de Precisión y Recall altos.

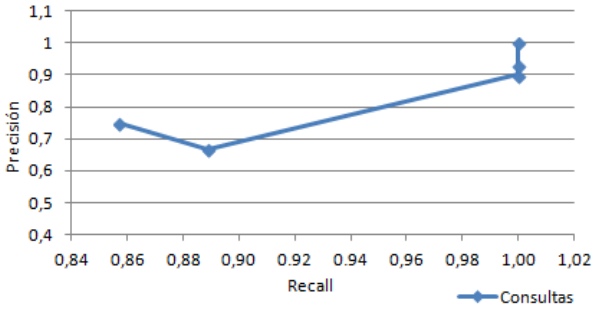

Figura 6: Gráfico de Precisión y Recall del Buscador Lucene

5 - Conclusión

A partir de los experimentos realizados, se demostró que es factible la construcción de un sistema para la búsqueda y recomendación para aulas virtuales, a través de la aplicación de las técnicas de Minería de Contenido Web y de Recuperación de la Información.

Los resultados obtenidos en la validación de los módulos Clasificación y de Búsqueda demostraron la efectividad de los mismos.

Actualmente se está trabajando en la evaluación del módulo de Recomendación con usuarios reales; lo que permitirá determinar la efectividad de la recomendación de acuerdo a la clasificación propuesta y corroborar que es posible asistir a los estudiantes en la búsqueda, dentro de un aula virtual, de material educativo relacionado a una duda puntual, independiente del formato en el que se encuentre; recomendándole la consulta de material adicional, disponible dentro del aula virtual, relacionado a su consulta

Como líneas acción futuras se piensa incorporar búsqueda semántica y aplicar técnicas de para mejorar las recomendaciones realizadas a los estudiantes, de acuerdo a características personales y de comportamiento del mismo.

Referencias

[1] Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D. 2013. *WEKA Manual for Version 3-6-9*. University of Waikato, Hamilton, New Zealand.

[2] Feldman, R. and Sanger, J. 2007. *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*.

[3] Han, J. and Kamber, M. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

[4] Hernandez Orallo, J., Ramirez Quintana, M.J. and Ferri Ramirez, C. 2004. *Introducción a la Minería de Datos*.

[5] Joachims, T. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. *10th European Conference on Machine Learning Chemnitz, Germany* (1998), 137–142.

[6] Jun, X., Gao, S. and Wu, G. 2006. Intelligent Resource Retrieval on Education Resource Base. *1 st International Symposium on Pervasive Computing and Applications* (2006), 461–465.

[7] Konchady, M. 2008. *Building Search Applications - Lucene, LingPipe and Gate*. Mustru Publishing.

[8] Mattmann, C.A. and Zitting, J.L. 2012. *Tika in action*. Manning Publications Co.

[9] McCandless, M., Hatcher, E. and Gospodnetic, O. 2010. *Lucene in action*. Manning Publications Co.

[10] Puustjärvi, J. and Pöyry, P. 2006. Information Retrieval in Virtual Universities. *International Journal Distance Education Technologies*. 4, 3 (2006), 36–47.

[11] Shao, L., Gou, X. and Li, J. 2011. Research and design of a vertical search engine for educational resources. *International Conference on Advanced*

Intelligence and Awareness Internet (AIAI 2011) (2011), 159–163.

[12] Souali, K., Afia, A. El, Faizi, R. and Chiheb, R. 2010. A New Recommender System For E-Learning Environments. *IEEE*. (2010), 1–4.

[13] Werthner, H., Hansen, H. and Ricci, F. 2007. Recommender Systems. *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. (2007), 167–167.