Sistema para la administración de asignaturas basado en clustering de documentos

Marylin Giugni O., Mariel Sarmiento, Yusdene Borrero y Luis León G.
mgiugni@uc.edu.ve, marielsarmiento@yahoo.com, yusdene@gmail.com, lleon@uc.edu.ve
Universidad de Carabobo. Facultad de Ciencias y Tecnología (FACYT).
Departamento de Computación

Resumen

Uno de los principales problemas en la actualidad, es el aumento exponencial de la información, lo cual no solo ocurre en el ámbito industrial, sino también en el sector educativo, donde los profesores deben mantenerse en constante investigación y en la búsqueda de todas aquellas innovaciones relacionadas con su área de enseñanza. En consecuencia, cuando un educador encuentra datos relevantes siente la necesidad de almacenarlos en los medios disponibles para tal fin, e incluso podría compartirlos o redistribuirlos entre los demás miembros de la organización educativa, lo que conduce a la duplicación de los datos y por ende al uso inadecuado de los recursos disponibles en la institución. Por otro lado, el proceso de organizar esta información no es tarea fácil, ya que es necesario invertir tiempo para su lectura y clasificación, de acuerdo a ciertos parámetros establecidos por la propia asignatura. Aunado a esto, una de las responsabilidades de los docentes, es el control de las actividades relacionadas con la administración de las asignaturas, así como también el seguimiento del desempeño de sus estudiantes. En este sentido, la Facultad de Ciencias y Tecnología de la Universidad de Carabobo (FACYT-UC) consciente de la necesidad de disminuir el tiempo que los profesores dedican a la docencia administrativa, ha generado aplicaciones que le sirven de apoyo durante esta actividad. Es por ello, que este artículo tiene por objetivo presentar el desarrollo de un sistema para la administración de asignaturas, el cual esta basado en la técnica de agrupación (*clustering*) de documentos, y se fundamentó en la metodología INSITE. Fue desarrollado bajo la plataforma J2EE (*Java Enterprise Edition)* utilizando el manejador de Base de Datos MySQL.

Palabras Claves: Educación, *clustering*, información, docencia administrativa, INSITE.

Introducción

En los últimos años, las Tecnologías de Información y Comunicación, han pasado a formar parte de toda organización. En el ámbito educativo, han transformado las formas de generar, gestionar, difundir la información y el conocimiento (Roa y otros, 2005).

En este sentido, cuando un investigador descubre datos valiosos, los almacena de modo que pueda reutilizarlos las veces que sea necesario, esta situación se puede repetir una y otra vez, incrementándose así, tanto el volumen de la información almacenada, como la complejidad de la búsqueda o recuperación de un dato en particular, aun cuando la misma se encuentre organizada, esto debido a que el desbordamiento de información se hace difícil de manejar por el hombre (Bartolomé, 1996).

Ahora bien, los resultados obtenidos a través un estudio realizado en el año 2005, por la Universidad de Berkeley, demostraron que hay un cambio en el patrón de crecimiento de la información creada por la humanidad, de un comportamiento basado en un crecimiento casi lineal se esta pasando a un crecimiento exponencial (Rojas, 2005). Esto evidencia la importancia de que la información se encuentre disponible de forma organizada, con el fin de optimizar la búsqueda y consulta de ésta.

Asimismo, en el ámbito educativo durante el proceso de enseñanza-aprendizaje se genera un alto volumen de información, en el cual los profesores responsables de dictar el contenido de cada una de las asignaturas realizan investigaciones y búsquedas bibliográficas para dictar sus clases. Cada uno de ellos realiza esta exploración con criterios y preferencias distintas, almacenando la información en diversos dispositivos, entre los cuales se tienen: documentos impresos, correo electrónico, dispositivos de almacenamiento portátil, entre otros. Así, "para que nuestros sistemas educativos operen con más eficiencia, siendo dinámicos y flexibles, habrá que insistir en que el capital estructural sirve integralmente sólo si existe un capital humano que lo sepa utilizar, implicando una eficiente gestión del conocimiento organizacional" (Barojas y Jiménez, 2006).

Ahora bien, con la finalidad de identificar cuáles eran los medios de recuperación de información utilizados por los profesores del Departamento de Computación de la FACYT-UC, para la preparación de sus clases, prácticas y exámenes, se aplicó una encuesta a 18 de ellos, lo cual representa el 72% de la población docente de este departamento. Así se obtuvieron los siguientes datos: El 100% de los encuestados realiza la búsqueda de información en Internet; el 62% busca en su propio computador; el 31% lo hace en el computador de otro profesor de la cátedra; el 15% realiza sus búsquedas en la biblioteca; mientras que el 8% busca información en un servidor central destinado a colocar documentación de todas las asignaturas que se dictan en este Departamento.

Como se puede apreciar, el medio utilizado por excelencia para la preparación de clases es la Internet, posteriormente el docente almacena dicha información en diversos dispositivos, entre los cuales se encuentra: el correo, un computador en su hogar, en el servidor de la facultad, en un computador de la oficina, en discos portátiles, entre otros.

En este sentido, considerando el alto volumen de información asociada a una asignatura,al cual se le suma el hecho de que la información se encuentra ubicada en diferentes dispositivos y por lo tanto de forma descentralizada, esto tiende a incrementar la dificultad en la recuperación de los datos y generar retardo en los tiempos de consulta de la misma.

Considerando que ésta situación se mantiene en el proceso de enseñanza de la FACYT_UC denota la importancia de disponer de un sistema que sirva de soporte a la administración del conocimiento, a través del cual se pueda almacenar, acceder y transferir la información y el conocimiento explícito existente dentro de la organización. Esto genera una memoria organizacional que permita explotar la experiencia ganada por la facultad a través de los años, así como garantizar el flujo de conocimiento e información, brindando un medio para registrarla y

compartirla con todas las entidades que la necesiten, ofreciendo así mayor apoyo a los profesores.

En este sentido, FACYT-UC consciente de la necesidad de disminuir los tiempos de docencia administrativa, se planteó a través de este trabajo el objetivo de desarrollar un sistema que le permitiera al profesor clasificar, etiquetar y actualizar el flujo de información con el propósito de disminuir los tiempos de respuesta en las consultas y mejorar la administración del conocimiento explícito de una asignatura.

De esta forma, se plantearon los siguientes objetivos específicos:

- Investigar la existencia de sistemas, aplicaciones y tecnologías que puedan ser empleadas para resolver la problemática con el fin de obtener el estado del arte.
- Identificar los requerimientos de los usuarios del sistema mediante técnicas de recopilación de información adecuadas para establecer el alcance y las restricciones del *software*.
- Estudiar distintas metodologías orientadas al desarrollo de aplicaciones Web, con la finalidad de utilizar la que mejor se adapte a la solución del problema.
- Estudiar los algoritmos de *clustering* de datos a fin de seleccionar aquel que permita localizar directamente la información necesitada.

Metodología

Esta investigación se inició con una revisión bibliográfica de: procesos de almacenamiento, recuperación y reutilización de información asociada al proceso de administración de asignaturas; sistemas automatizados para gestionar el conocimiento generado en las universidades, procesos para la clasificación adecuada de documentos en formato digital, entre otros.

Posteriormente se efectuó un diagnóstico del dominio de estudio, para ello se llevaron a cabo observaciones directas sobre los procesos de gestión de las asignaturas de la FACYT, específicamente en la escuela de computación. Además, se aplicaron encuestas a 18 profesores de esta institución, tal como se indicó previamente, lo cual permitió tener una visión global del proceso en estudio e identificar las necesidades específicas de los docentes.

Cabe destacar, que se ha tomado como caso de estudio la asignatura Algoritmos y Programación II, ya que ésta es considerada columna vertebral dentro del currículo de la carrera de Licenciatura en Computación, además, de ella se tiene un historial de datos desde el año 1997 hasta el 2006.

Por otra parte, considerando que la investigación tiene como objetivo principal el desarrollar un sistema basado en *clustering* de documentos, el cual facilite la administración de asignaturas, y mejore la productividad del docente, se hizo necesario utilizar una metodología de desarrollo de software que permitiera guía el desarrollo y puesta en marcha del sistema, considerando aspectos de calidad del sistema, tales como: usabilidad, confiabilidad, seguridad y mantenibilidad, entre otros. Para ello se realizó un estudio exhaustivo de diversos enfoques metodológicos, tales como: UWE (UML-Based Web Engineering) (Koch, 2001), XP (*eXtreme Programming)* (Kent, 2001), INSITE (Wairua, 2002), entre otros. A partir de dicho estudio, se seleccionó INSITE como

metodología para el desarrollo del sistema, por mantener la integridad del diseño, ayudando a minimizar los defectos de diseño y errores humanos durante la especificación.

Durante la fase de implementación se codificaron las tareas programadas utilizando software de licencia GPL (GNU Public License), es decir, software libre. Se utilizó la plataforma de desarrollo J2EE con todos sus componentes, el manejador de Base de Datos MySQL versión 5.0 y el servidor Web Tomcat 5.5.

Desarrollo del sistema

Como se indicó previamente el desarrollo del sistema esta guiado por la metodología INSITE, a través de la cual se llevaron a cabo las fases de análisis, diseño, codificación y pruebas del sistema. En esta sección se describirán los aspectos resaltantes de este desarrollo, destacando la audiencia y funcionalidades del sistema, así como los algoritmos de clasificación y agrupación de documentos, ejes del trabajo.

Es importante señalar que el sistema es conocido como CORAL, y sus siglas estas relacionadas con la **L**ocalización, **A**dministración, **R**ecuperación y **O**rganización de las **C**lases, asociadas a una asignatura.

La audiencia que corresponde a éste trabajo está constituida por los profesores adscritos al departamento de computación de la FACYT y por los administradores de la plataforma tecnológica de la misma facultad. De manera que los primeros tendrán acceso sólo a las funcionalidades propias de la gestión de una asignatura y los segundos realizarán únicamente las acciones correspondientes a la administración de las cuentas de los docentes.

Es necesario señalar que tanto los administradores como los profesores, deben haber ingresado a su respectiva cuenta de usuario para realizar cualquiera de las acciones que correspondan a su rol. Los escenarios que se describen a continuación representan algunas de las acciones que realizan los usuarios dentro del sistema CORAL.

Escenario 1: El profesor a través de la herramienta podrá añadir documentos al repositorio de la asignatura, los cuales serán clasificados de acuerdo a sus principales características, lo que facilitará su posterior recuperación. Para esto deberá elegir la opción <<agregar al repositorio>>, suministrar algunos datos y pulsar la opción <<agregar>> para culminar exitosamente la acción e iniciar su clasificación.

Escenario 2: El profesor podrá buscar y abrir cualquier documento que se encuentre almacenado en el repositorio. En este sentido, puede realizar dos tipos de búsqueda: rápida y detallada; para la primera deberá colocar las palabras claves en la sección <<búsqueda rápida>>, pulsar la opción <<buscar>>. Por otro lado, si el docente selecciona la <<búsqueda detallada>> deberá proporcionar ciertos datos específicos, además de las palabras claves, para realizar una búsqueda con mayor precisión. En ambos casos aparecerá una lista de los documentos relacionados a las palabras claves o características dadas por el profesor, entonces, de esa lista podrá abrir los documentos que

desee. Para eliminar un documento tendrá que pulsar la opción <<eliminar>> que aparecerá a un lado de éste y después de confirmar tal solicitud, el documento quedará eliminado del repositorio.

Escenario 3: El administrador de la asignatura (profesor) podrá generar la planificación correspondiente a ésta, seleccionando la opción <<administrar asignatura>> y luego <<planificación>>, desde allí creará las clases, las secciones de clases con sus respectivos alumnos asignando a cada una de éstas el profesor asociado. Además el administrador de la asignatura tendrá derecho a actualizar o eliminar cualquiera de los aspectos relacionados con la planificación. También desde las opciones <<clases>> y <<secciones>> podrá observar los respectivos datos y modificarlos.

Escenario 4: El administrador del sistema tiene el permiso para crear cuentas de usuario y conceder cualquiera de los tres tipos de roles: profesor, administrador de asignatura o administrador del sistema, desde la sección <<administrar usuarios>>, e ingresando a <<crear usuario>>. Por otro lado para modificar o eliminar una cuenta de usuario, debe acceder a los datos del usuario desde <<buscar usuario>> o <<listar usuarios>>. Además con la opción <<administrar sistema>> puede agregar nuevas asignaturas e ingresar los datos necesarios para formar los grupos en los que serán clasificados los documentos correspondientes a la asignatura.

En la figura 1, se observan algunas de las funcionalidades del sistema, las cuales han sido descritas en los párrafos previos.

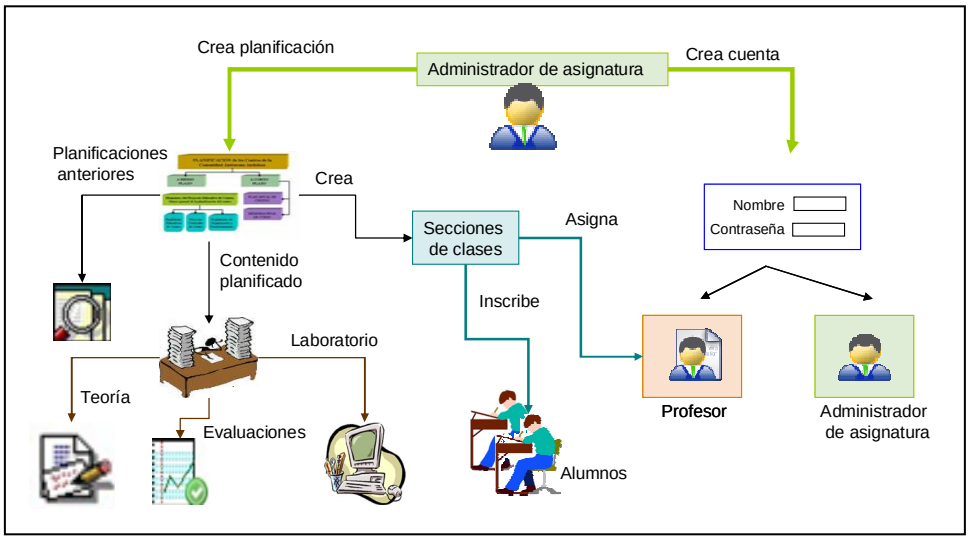

Figura 1. Algunas funcionalidades de CORAL

Ahora bien, con el propósito de optimizar el proceso de búsqueda y recuperación de los documentos almacenados en el sistema, se realizó una investigación de los algoritmos diseñados para tal fin; de allí surge la necesidad de utilizar estrategias para caracterizar previamente estos documentos, con la finalidad de establecer una representación numérica que permita ejecutar el algoritmo de *clustering* adecuadamente y lograr la recuperación eficaz de los mismos. Cabe resaltar que esta técnica también llamada agrupamiento, es la que permite la identificación de

tipologías o grupos, donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros (La Serna y otros, 2004).

En este sentido, se utilizó el modelo del espacio vectorial de Salton (La Serna y otros, 2004) en el cual cada documento es representado mediante un vector de n elementos, siendo n igual al número de términos indizables que existen en la colección documental. Por lo tanto, hay un vector para cada documento y en cada vector, un elemento para cada término o palabra susceptible de aparecer en el documento. Cada uno de esos elementos es cubierto u ocupado con un valor numérico conocido como el peso del término en el documento. Este peso es calculado mediante el esquema TF-IDF (Figuerola, Zazo y Alonso, 2002), mediante el cual se determinó la frecuencia que tiene una palabra dentro de un documento y el número de repeticiones de la misma en todo el conjunto de documentos, esto con el propósito de filtrar aquellas palabras comunes que no representan alguna información importante.

Por otra parte, durante este proceso de búsqueda se encontró que los algoritmos de *clustering* de datos (COBWEB, K-Vecinos, *Espectation-Maximization* y K-Medias), mejoran el rendimiento de los motores de búsqueda mediante la categorización previa de los documentos, debido a que se sostiene que los documentos fuertemente asociados tienden a ser relevantes para la misma consulta, por lo que el mismo identifica tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros.

Cabe destacar que fue seleccionado el algoritmo de K-medias para realizar la categorización de los documentos, debido a la sencillez y facilidad de implementación que éste presenta, además del alto rendimiento que ostenta en la clasificación de datos, ya que es un procedimiento iterativo que parte de una distribución inicial y en cada iteración puede cambiar la asignación de un individuo (documento) a los grupos para conseguir una mejor partición, hasta dar con una satisfactoria. Por esta razón es uno de los algoritmos de clasificación más populares, tanto a nivel científico como industrial.

Conclusiones

- Al investigar la existencia de sistemas y tecnologías que contribuyeran a la solución del problema, se encontraron algunos trabajos que ofrecieron grandes aportes, sin embargo no se encontró sistema alguno que cumpliera con las características necesarias para la administración de asignaturas.
- El empleo de la propuesta metodológica constituyó una guía para el proceso de desarrollo de manera sencilla, flexible y robusta, evitando errores en la especificación, permitiendo además incluir pruebas de usabilidad para lograr la aceptación del usuario, así como la obtención de resultados exitosos en la implementación del software.
- La implementación del *clustering* de datos como técnica de clasificación de documentos permite optimizar los motores de búsqueda para localizar los documentos de gran relevancia para un profesor.
- La implementación tanto del repositorio de documentos, como de la técnica de clasificación para manejar el alto volumen y el flujo de información generado dentro del proceso de formación de una asignatura, establece las bases para la gestión del conocimiento dentro de la

FACYT-UC, considerando la importancia que tiene para una organización, sobre todo si es educativa, el valor agregado que ofrece tal gestión para aumentar la productividad y el logro de objetivos en la misma.

- El uso de la tecnología J2EE como plataforma de desarrollo aseguró la portabilidad del sistema. Así mismo, la implementación del *software* se realizó de manera simplificada gracias a las características de modularidad y reusabilidad provistas por esta plataforma.

Bibliografía

Barojas J. y Jiménez E. (2003). *Gestión del Conocimiento Organizacional en Educación.* XIX Simposium Internacional de Computación en la Educación 2003. [Artículo en línea]. Consultado el 20 de noviembre de 2006 en: http://bibliotecadigital.conevyt.org.mx/colecciones/documentos/somece/29.pdf

Bartolomé A. (1996). *Preparando para un nuevo modelo de conocer*. Revista electrónica de tecnología educativa. [Revista en línea] Num. 4. Consultado el 17 de enero de 2006. Disponible en: http://www.uib.es/depart/gte/revelec4.html

Figuerola C.; Zazo A. y Alonso J. (2002). *Caracterización automática de documentos en español y la normalización de términos.* DoIS (Documents in Information Science) [Artículo en línea]. Consultado el día 18 de enero de 2006 en: http://imhotep.unizar.es/jbidi/jbidi2000/14_2000.pdf

Kent, B. y Martin, F. (2001). *Planning Extreme Programming*. Addison- Wesley.

Koch, N. (2001). *Software Engineering for Adaptative Hypermedia Applications*. Ph. Thesis, FAST Reihe Softwaretechnik Vol(12), Uni-Druck Publishing Company, Munich. Germany

La Serna N.; Román U.; Osorio N. y Benito O. (2004). *Estudio y evaluación de los sistemas de Recuperación de información*. Revista de investigación de Sistemas e Informática. [Revista en línea] Vol. 1, N° 1. Consultada el 20 de febrero de 2006 en: http://sisbib.unmsm.edu.pe/BibVirtual/Publicaciones/risi/portada.htm

Rojas, O. (2006). *El boom de las soluciones de almacenamiento*. The GBM Journal. Business Transformation. [Revista en línea] Año 8 Edición 31. Consultada el 20 de febrero de 2006 en: http://www.gbm.net/bt/BT31/08_bluetech.html

Roa J.; Gramajo S. ; Virgil R. ; Ramírez R. ; Karanik M. y Perez J. (2005). *Mejora de la Plataforma de e-learning Moodle utilizando Redes Neuronales*. Primeras Jornadas de Educación en Informática y TICs. Buenos Aires. Argentina. [Pagina Web en línea]. Disponible en: http://cs.uns.edu.ar/jeitics2005/Trabajos/pdf/33.pdf

Wairua Consulting (2002) *The INSITE™ Methodology*. INSITE Guidebook. [Página Web en línea] Disponible en: http://www.wairua.co.nz/docs/insite.pdf