GESTOR WEB SUSTENTABLE PARA EL MANEJO DE DOCUMENTOS ADMINISTRATIVOS EN INSTITUCIONES EDUCATIVAS

SUSTAINABLE WEB MANAGER FOR ADMINISTRATIVE DOCUMENT IN EDUCATIONAL INSTITUTIONS

Laura Camacho González[1], Selene Hernández Rodríguez[1] y Adolfo Aguilar Rico.[1]
lcamglez@gmail.com / hdezrod@gmail.com / adolforico2@gmail.com

Recibido: febrero 2, 2014 / Aceptado: junio 13, 2014 / Publicado: junio 19, 2014

RESUMEN. En la actualidad, si bien es cierto que muchas bondades de la tecnología moderna nos permiten vivir más cómodamente, comunicarnos más fácilmente, transportarnos rápidamente y realizar tareas que antes no habría sido posible realizar, también es cierto que algunos aspectos no han sido tan positivos y han impactado nuestro entorno. Tal es el caso del manejo en exceso del papel, lo que nos ha llevado a la tala innecesaria y ha impactado nuestra ecología. Por esta razón, en este trabajo se presenta el desarrollo de un gestor web para el manejo de documentos administrativos de una institución educativa, el cual funciona de manera computacional sin involucrar la tarea de imprimir cartas en ningún momento, a menos que el usuario necesite fuertemente una impresión de la versión final de la carta. Con este gestor web se tiene como objetivo reducir el manejo de papel en dos etapas por las que pasa la creación de un documento administrativo en una institución educativa, una de las cuales es la creación de un documento que involucra la impresión de varias versiones para corregir errores y la segunda es el almacenamiento. Además, se espera también que con este gestor se manejen los documentos administrativos de una institución, de manera electrónica durante todas las etapas por las cuales tiene que pasar, desde su creación hasta su revisión, autorización, firma electrónica, entrega y almacenamiento del documento.

PALABRAS CLAVE: Gestor web sustentable, manejo de documentos electrónicos, tecnologías web de búsqueda y recuperación de documentos.

ABSTRACT. Nowadays, although many benefits of modern technology enable us to live more comfortably, communicate more easily, transport us quickly and perform tasks that previously would not have been possible, it is also true that some aspects have not been positive and have impacted our environment. Such is the case of the excessive paper management, which has led to the unnecessary wood cutting and impacted our ecology. For this reason, in this paper the development of a web administrative documents manager for educational institutions is presented. This system works without involving the task of printing any document, unless the user strongly needs an impression of the final version of it. The development of this web manager aims to reduce paper handling in two stages of the creation of an administrative document in an educational institution, one of which is the printing of multiple versions in order to correct mistakes in the document, and the second is the storage of it. In addition, it is also expected that with the development of this manager, administrative documents of an institution are handled electronically during all of their process stages: creation, review, approval, electronic signatures, delivery and storage document.

KEYWORDS: Sustainable web manager, management of administrative document, documents search and retrieval web technologies.

[1] Instituto Tecnológico de Puebla. Avenida Tecnológico #420, Colonia Maravillas, C.P. 72220, Puebla, Puebla, México. www.itpuebla.edu.mx/

1. Introducción

En la actualidad, si bien es cierto que muchas bondades de la tecnología moderna nos permiten vivir más cómodamente, comunicarnos más fácilmente, transportarnos rápidamente y realizar tareas que antes no habría sido posible realizar, también es cierto que algunos aspectos no han sido tan positivos y han impactado nuestro entorno. Tal es el caso del manejo en exceso del papel, lo que nos ha llevado a la tala innecesaria y ha impactado nuestra ecología. Por esta razón, en este trabajo se presenta el desarrollo de un gestor web para el manejo de documentos administrativos de una institución educativa, el cual funciona de manera computacional sin involucrar la tarea de imprimir cartas en ningún momento, a menos que el usuario necesite fuertemente una impresión de la versión final de la carta, lo cual puede ocurrir ya que el papel es un elemento que ha perdurado históricamente. Sin embargo, con este gestor se reduce el número de impresiones necesarias durante la creación y modificaciones del documento.

Con este gestor web se tiene como objetivo reducir el manejo de papel en dos etapas por las que pasa la creación de un documento administrativo en una institución educativa, una de las cuales es la creación de un documento que involucra la impresión de varias versiones para corregir errores, y la segunda es el almacenamiento. Además, se espera también que con este gestor se manejen los documentos administrativos de una institución de manera electrónica durante todas las etapas por las cuales tiene que pasar, desde su creación y revisión, hasta la autorización, firma electrónica, entrega y almacenamiento del documento. Este manejo digital de documentos administrativos se ha automatizado mucho gracias a la capacidad de almacenamiento y procesamiento de las computadoras modernas, el crecimiento actual de Internet y el desarrollo de nuevas tecnologías de la información (TIC), con lo cual es posible desarrollar aplicaciones web cada vez más robustas.

Con el desarrollo de este gestor, se logra reducir el consumo de papel y se disminuye el uso de productos químicos que son incluidos en consumibles para impresoras de cualquier tipo. Esto es importante debido a que satisface la necesidad que tienen empresas, industrias y otras instituciones, de hacer sus procesos cada vez más sustentables [1-4]. Además, cabe destacar, como un gran beneficio de esta clase de sistemas, el ahorro de tiempo que significa respecto del proceso de gestión de documentos tradicional.

El gestor de documentos presentado está formado por los siguientes módulos:

- Una base de datos para almacenar la información referente a las cuentas de usuarios y los documentos almacenados en el servidor.
- Manejo de sesiones para restringir el acceso a usuarios permitidos, así como permitirles el acceso sólo a los documentos permitidos.
- Módulo para crear documentos, los cuales pueden ser firmados electrónicamente por miembros de la institución.
- Módulo para subir documentos administrativos existentes de la institución al servidor (disponibles en web para el uso remoto de diferentes miembros de la institución), asignándoles ciertos permisos de privacidad para poder ser vistos por los usuarios deseados.
- Módulo de búsqueda y recuperación de documentos de cierto interés por medio de una palabra o frase.

Cabe mencionar que el módulo de búsqueda y recuperación de documentos es un módulo importante para el

desarrollo del gestor de documentos propuesto. Para realizar este módulo se analizaron y compararon las características de algunas tecnologías web actuales. A partir de este análisis, se seleccionaron las tecnologías JiFile y Oracle Text, que de acuerdo al análisis realizado son las que mostraron las mejores características, ya que integran algoritmos robustos de búsqueda y recuperación de información sobre documentos de texto de manera privada (es decir, que no son públicos en Internet).

Para efectuar las comparaciones antes mencionadas entre las tecnologías JiFile y Oracle Text, se utilizó un conjunto de documentos formado por documentos administrativos reales de la División de Estudios de Posgrado e Investigación (DEPI) del Instituto Tecnológico de Puebla (ITP). Este conjunto de documentos está formado por 1065 documentos de texto (DOC), los cuales corresponden a los siguientes temas: alumnos, profesores y horarios, entre otros. De estos documentos se seleccionaron algunos temas para realizar algunas búsquedas, de los cuales se sabe el número de documentos relacionados con estos temas.

A partir de los resultados obtenidos de evaluar estas dos tecnologías para la búsqueda de documentos, se puede concluir que ambas tecnologías son muy competitivas al realizar la tarea de la búsqueda de texto completo.

2. Tecnologías Web existentes para Búsqueda y Recuperación de Documentos

Existen muchas clases de buscadores. Entre ellos, se encuentran buscadores por palabras clave, buscadores por categorías, buscadores geográficos, metabuscadores y buscadores de texto completo [5-8]. Este trabajo se enfoca en las búsquedas de texto completo, las cuales son adecuadas para la búsqueda y recuperación de documentos administrativos. La búsqueda de texto completo es una técnica de búsqueda sobre documentos electrónicos o una colección de documentos en una base de datos. Este motor de búsqueda examina todas las palabras en cada documento almacenado para posteriormente encontrar coincidencias en el criterio de búsqueda. Para realizar búsquedas de texto completo sobre documentos, existen diferentes tecnologías de información, las cuales van desde productos comerciales, productos de código abierto, bibliotecas o incluso bases de datos orientadas a documentos. Las búsquedas principalmente pueden realizarse en documentos con formato estándar, estos incluyen XML (eXtensible Markup Language), YAML (Yet Another Markup Language), JSON (JavaScript Object Notation) y BSON (Binary JSON), así como PDF (Portable Document Format) y documentos de Microsoft Office. Además, la mayoría de las tecnologías comparadas para realizar la búsqueda y recuperación de documentos son de código libre y con características similares. Sin embargo, algunas de estas tecnologías tienen objetivos diferentes al planteado. Por ejemplo, DataparkSearch se enfoca en imágenes y nosotros deseamos desarrollar una aplicación más enfocada en texto. Lemur Toolkit, por su parte, realiza búsquedas en documentos XML [9].

Actualmente, varias tecnologías web recientes se enfocan en el manejo de archivos de tipo XML (eXtensible Markup Language), los cuales son un tipo de archivos muy utilizados debido a su fácil manejo y portabilidad [10]. Sin embargo, debido a que muchas empresas e instituciones aún tienen sus documentos en otro tipo de formatos con texto plano no marcado, se propone en este trabajo el desarrollo de un gestor de documentos el cual sea aplicable a las empresas e instituciones que manejan documentos de diversos formatos, como DOC y PDF, y para las cuales no sea tan sencilla una migración a XML. En este caso se descarta Google Search para este trabajo, ya que maneja los documentos de manera pública, por lo que la privacidad de los documentos no está garantizada y ésta es una característica importante para el desarrollo de un gestor de documentos administrativos de una institución o empresa, la cual busca la confidencialidad de su información.

Finalmente, las tecnologías de JiFile y Oracle Text presentan características que las hacen ver robustas y atractivas en cuanto a la búsqueda y recuperación de documentos, aun cuando entre estas dos tecnologías es importante mencionar que JiFile es gratuito mientras que Oracle Text no lo es. Por lo expuesto hasta el momento se han seleccionado las tecnologías de JiFile y Oracle Text para realizar pruebas e incluir alguna de estas tecnologías en el gestor de documentos propuesto, con el objetivo de lograr un funcionamiento exitoso.

Estas dos tecnologías tienen un funcionamiento similar, ya que ambas consisten de dos etapas: la primera es una etapa de pre-procesamiento sobre el conjunto de documentos y la segunda etapa corresponde a la tarea de búsqueda. En la primera etapa se suben al servidor en el que está hospedado nuestro gestor web, todos los documentos sobre los que se realizarán las búsquedas posteriormente. A partir de estos documentos, ambas tecnologías aplican una serie de algoritmos para pre-procesar los documentos y la información contenida en ellos, de tal manera que se generan estructuras y datos que serán de ayuda durante la etapa de búsqueda y recuperación de documentos.

Durante la etapa de búsqueda, ambas tecnologías utilizan operadores para realizar la búsqueda de documentos, con los cuales se define el algoritmo a utilizar para la búsqueda (con sus respectivos parámetros, si es el caso) de acuerdo al tipo de búsqueda que se desea realizar. Los diferentes operadores con los que se puede realizar una búsqueda se muestran en la Tabla 1. En la primera columna de esta tabla se muestra la descripción del operador (DESCRIPCIÓN) y se muestra también entre paréntesis el símbolo utilizado durante la consulta, en la segunda columna (OPERADORES) se muestra el nombre del operador, en la tercera columna (JiFile) se muestra con una palomita cuando el operador está disponible para JiFile y finalmente en la cuarta columna se muestra cuando el operador está disponible para la tecnología de búsqueda Oracle Text.

Tabla 1. Lista de operadores de JiFile y Oracle Text.

DESCRIPCIÓN	*OPERADORES*	*JIFILE*	*ORACLE TEXT*
Busca coincidencia exacta de un término en los documentos.	Término exacto	√	√
Busca coincidencia exacta de múltiples palabras utilizando el operador AND. (&)	AND	√	√
Busca coincidencia de alguna o algunas de las palabras de una búsqueda utilizando el operador OR. (|)	OR	√	√
Búsqueda por palabras que están cercanas una de otra con el operador NEAR. (;)	NEAR	√	√
Suma puntajes de búsqueda individuales y compara el puntaje acumulado con el valor límite. (,)	ACCUM	√	√
Resta el puntaje de la búsqueda del segundo término al puntaje de la búsqueda del primer término. (-)	MINUS	√	√
Elimina registros basados en la búsqueda de términos. (~)	NOT	√	√
Especifica una sustitución aceptable para una palabra en una búsqueda. (=)	EQUIV	√	√
Busca cadena de texto que tenga un patrón.	%	**X**	√
Limita la expansión de la cadena de texto a exactamente a tres caracteres.	___	**X**	√
Recupera coincidencias de expresiones con calificación de términos que son más grandes que el límite. (>)	THRESHOLD	**X**	√

Busca términos que comparten la misma raíz lingüística.	STEMM	√	√
Expande búsquedas para incluir palabras que se escriben de forma similar al término especificado. Búsqueda de términos mal escritos. (?)	FUZZY	√	√
Expande la búsqueda a palabras que tienen sonidos similares. (!)	SOUNDEX	**X**	√
Búsqueda por concepto y temas.	ABOUT	√	√

3. Gestor web de documentos administrativo propuesto

En este trabajo se expone el desarrollo de un Gestor Web de Documentos Administrativos para el manejo de documentos administrativos, útil tanto para el sector empresarial (empresas privadas y públicas) como para cualquier otro sector (investigación, educativo o bibliotecas electrónicas), el cual permita realizar la creación, almacenamiento, manejo y búsqueda de documentos. Se propone que el gestor de documentos tenga las siguientes características (ver Figura 1):

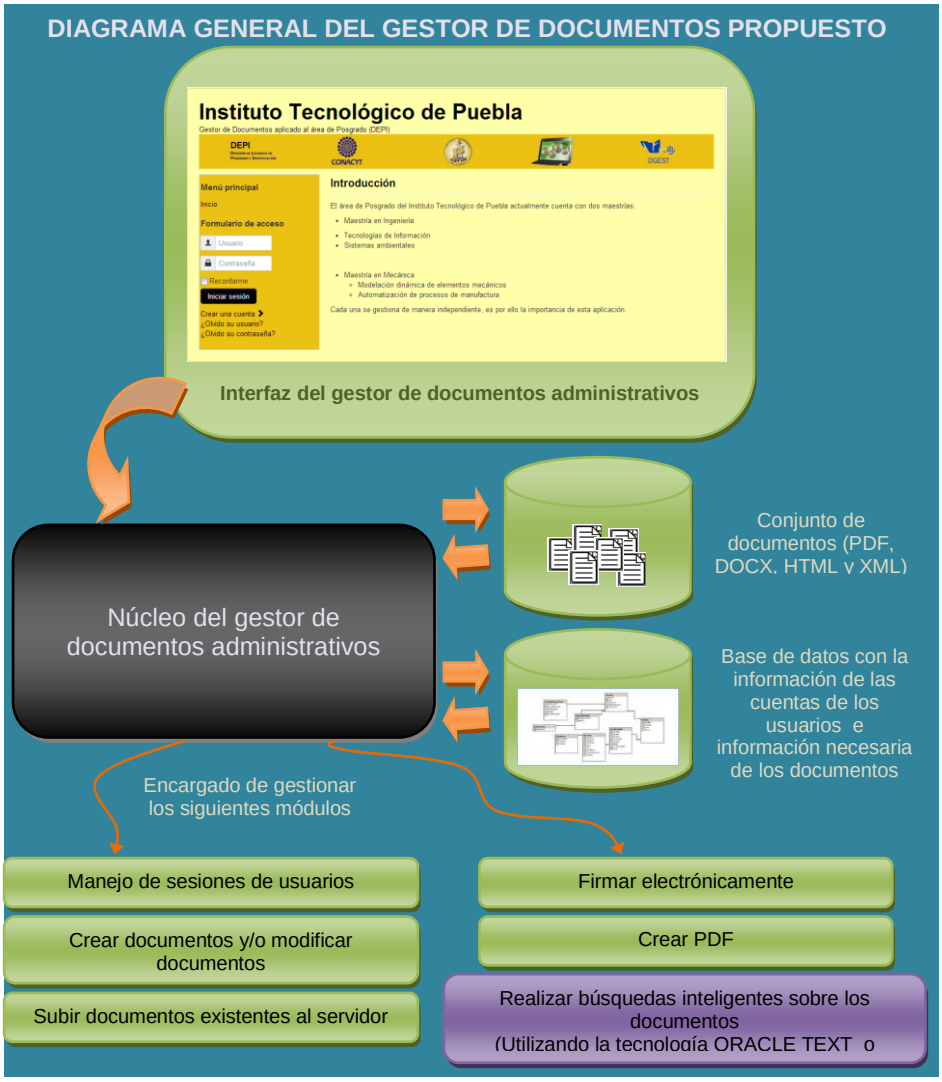

Figura 1. Diagrama general del gestor de documentos propuesto.

- *Funcionamiento en Internet:* Actualmente, tener en Internet nuestra información nos permite la flexibilidad de tener acceso a ella en cualquier lugar y en cualquier momento. En el caso de un gestor web de documentos sería posible manejar, modificar, leer o buscar documentos desde cualquier lugar geográfico.
- *Manejo de sesiones:* Para restringir el acceso a usuarios permitidos, así como permitirles el acceso sólo a los documentos permitidos.
- *Creación y almacenamiento de documentos:* Con esta bondad del sistema, sería posible que los diferentes usuarios tuvieran una interfaz amigable de la aplicación web con la cual pudieran crear sus documentos, firmarlos electrónicamente y compartirlos con otros usuarios miembros de la institución. También es importante permitir el almacenamiento de documentos ya existentes en la institución, puesto que ocurre que, al momento de adoptar una herramienta nueva para el manejo de documentos, ya se cuenta con una gran cantidad de documentos administrativos existentes en formatos PDF o DOC, a menos que la institución en cuestión fuera de nueva creación.
- *Manejo de documentos:* Con esta característica es posible manejar diferentes usuarios dentro del sistema, a los cuales se les pueden asignar diferentes permisos para tener acceso a los documentos. También, cuando un usuario agregue o cree un nuevo documento, puede compartirlo con otros usuarios para ver, modificar, firmar o realizar búsquedas sobre los documentos.
- *Búsqueda y recuperación de documentos:* Este punto es importante, ya que realizar una búsqueda exhaustiva de algún tema en particular, sobre una gran cantidad de documentos digitales, resulta una tarea difícil o imposible. Por esto, se propone un módulo de búsqueda y recuperación de documentos, a través de una palabra o frase, para encontrar los documentos relacionados.

Para desarrollar el gestor propuesto, se diseñó el diagrama de clases que se muestra en la Figura 2, con el cual es posible manejar los elementos y aspectos de este gestor de documentos. A partir de la Figura 2 se puede observar que se proponen dos clases principales, las cuales son "*Usuario*" y "*Documento*". La clase *Usuario* maneja algunos atributos de los usuarios, como son: el tipo de privilegios, cargo, datos personales, la firma electrónica de cada usuario, etc. De esta manera, es posible manejar los distintos tipos de usuarios existentes con sus respectivos rangos dentro de la empresa, los cuales se ligarán directamente con los permisos a los documentos. Por ejemplo, si se trata de un coordinador de área, tiene permiso para ver todos los documentos, mientras que otros usuarios de menor rango como los profesores, sólo pueden ver sus propios documentos y los documentos de otros usuarios compartidos con ellos. De esta manera, quedan protegidos los documentos con información importante como salarios, ya que no todos los usuarios tienen permiso de ver todos los documentos.

En la clase *Documento* se manejan todas las características ligadas a un documento, como son: tipo de documento, estado actual del documento (si ya se firmó, si se encuentra en modificaciones, etc.), usuarios relacionados con el documento, entre otros. También, en esta clase se manejan los siguientes métodos: crear, modificar, eliminar y realizar búsquedas de documentos, así como crear archivos con extensión PDF.

Además, se desarrolló una base de datos para almacenar la información necesaria para el manejo de usuarios y la información adicional sobre los documentos (por ejemplo, los usuarios relacionados con cada documento, los tipos de privilegios de usuarios, entre otros), la cual se muestra en la Figura 3.

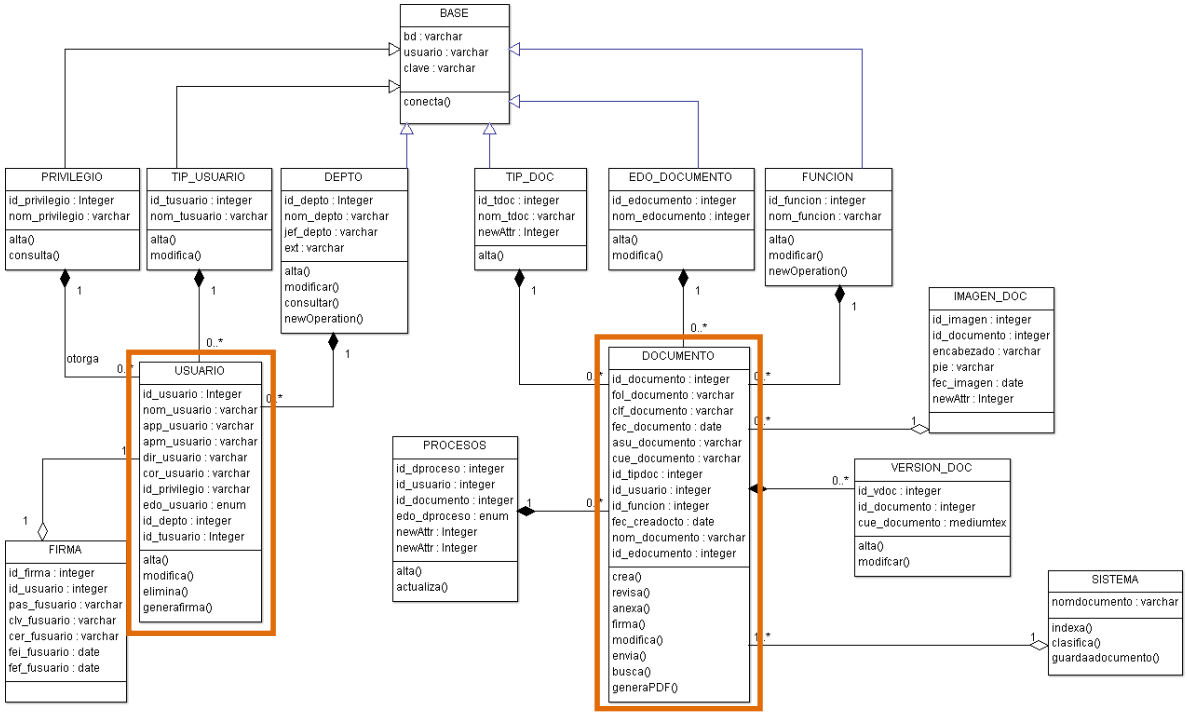

Figura 2. Diagrama de clases del gestor de documentos propuesto.

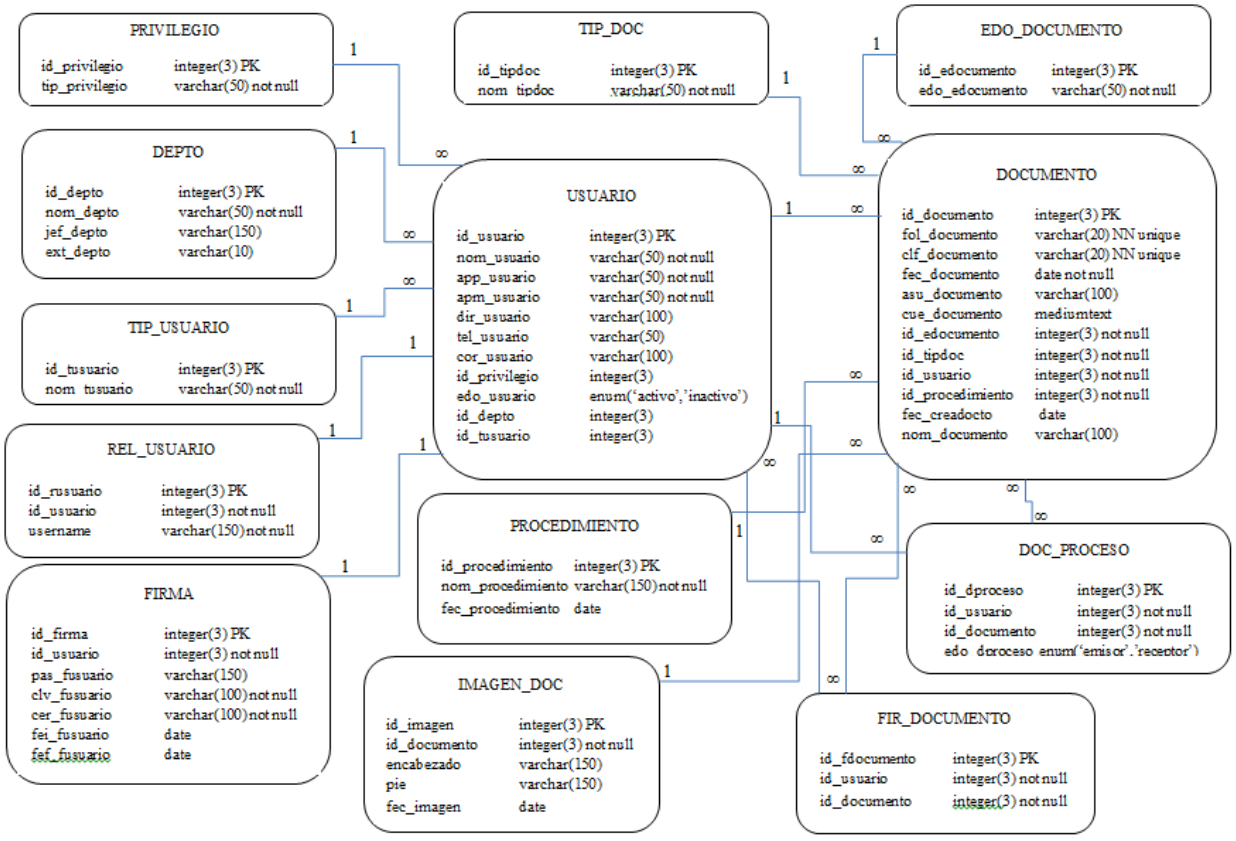

Figura 3. Base de datos propuesta para el manejo de documentos.

Así, con la implementación realizada de los diseños mostrados (diagrama de clases y base de datos), es posible realizar diferentes tareas en el gestor de documentos. Una vez dado de alta un usuario, se puede iniciar sesión y realizar alguna de las siguientes actividades (ver diagrama de Actividades en la Figura 4):

- *Subir documento al gestor:* Es posible subir al gestor documentos en formato DOC o PDF existentes en la empresa o institución. Es posible, también, compartir estos documentos con otros usuarios. Si no se comparte, el documento sólo lo pueden ver el usuario que subió el documento y usuarios con jerarquía más alta.
- *Crear nuevo documento:* Se puede utilizar la interfaz del gestor para crear nuevos documentos y compartirlos con otros usuarios, así como agregar usuarios para permitir modificaciones al documento y/o para firmar electrónicamente. Un documento nuevo es autorizado (o está listo) cuando todos los usuarios relacionados con este documento ya lo han firmado electrónicamente.
- *Revisar bandeja de documentos:* Por medio de ésta, es posible visualizar todos los documentos a los que el usuario tiene acceso, ya sea porque son de él o porque pertenece al grupo de usuarios a los cuales está compartido el documento. Además, por medio de esta bandeja es posible indicar el estado de cada documento. Por ejemplo, si falta que revise modificaciones o si faltan firmas electrónicas, entre otras características.
- *Consultar documentos:* Por medio de las tecnologías de búsqueda evaluadas en este trabajo, se realiza la búsqueda de texto completo sobre los documentos a los que tiene acceso el usuario.

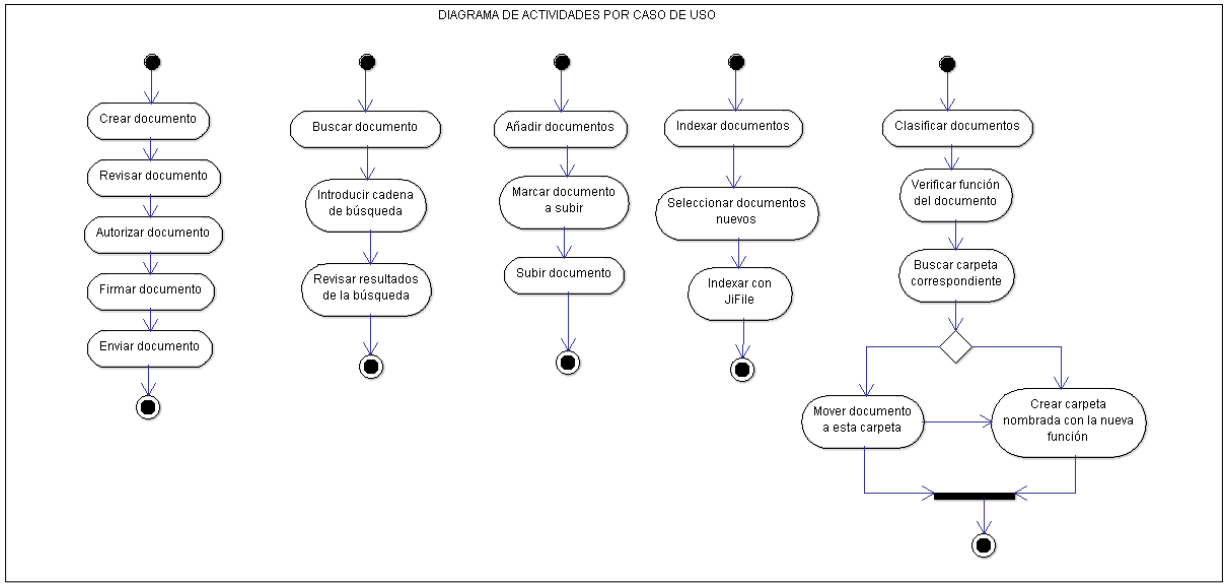

Figura 4. Diagrama de Actividades del Gestor de Documentos propuesto.

4. Resultados experimentales

En esta sección se muestran como resultado del desarrollo del gestor de documentos propuesto, el diagrama de clases del gestor y algunas interfaces, así como algunas pruebas realizadas con las tecnologías Oracle Text y JiFile, las cuales se realizaron con el objetivo de seleccionar una de estas dos tecnologías para implementarla en el módulo de búsqueda y recuperación de documentos de este gestor.

El gestor propuesto se desarrolló sobre el sistema operativo de Windows 7 Home Basic y se

utilizaron diferentes gestores de bases de datos para poder utilizar las dos tecnologías de búsqueda de documentos evaluadas. Para el caso de JiFile, se utilizó el gestor MySQL 5.5.20, mientras que para el manejo de Oracle Text se utilizó Oracle 11.0. Para el caso de JiFile se utilizó el gestor de contenidos de Joomla versión 3.0, el cual permite incorporar de manera fácil extensiones de diferentes empresas, tal es el caso de JiFile que permite realizar búsquedas de texto completo sobre documentos. Para realizar el desarrollo de los módulos restantes del gestor propuesto, se utilizó el lenguaje de programación PHP 5. Para el caso de Oracle Text se utilizó la herramienta Oracle Forms, la cual nos permite desarrollar aplicaciones web. Sin embargo, es importante mencionar que la versión final y completa de este gestor se desarrolló en Joomla utilizando la tecnología de búsqueda de documentos JiFile, Apache, PHP y MySQL.

4.1 Interfaces del gestor de documentos propuesto

En esta sección se muestran algunas interfaces desarrolladas para este gestor de documentos, las cuales fueron evaluadas y probadas por algunos usuarios finales, y resultaron ser satisfactorias y de fácil manejo. Para realizar la evaluación y prueba del gestor se realizó una fase de pruebas Alfa y una fase de pruebas Beta.

Durante la fase de pruebas Alfa, se realizó un trabajo de manera conjunta entre los desarrolladores del gestor y un grupo de 7 usuarios finales (un coordinador, una secretaria, dos maestros y tres alumnos) para definir y desarrollar las interfaces necesarias para toda la gestión de documentos propuesta y que, además, resultaran de fácil manejo para los usuarios finales.

Durante la fase de pruebas Beta, se evaluó el funcionamiento de un prototipo completo del gestor de documentos propuesto por parte de los usuarios finales, con el objetivo de detectar errores o posibles mejoras al sistema. También se tomó en cuenta la opinión de los usuarios finales, por medio de un cuestionario, para evaluar algunas métricas: exactitud (es decir, si el gestor tenía errores), tiempo (es decir, si el usuario tiene la impresión de que puede llevar a cabo de manera rápida cualquier tarea), facilidad y respuesta emocional (es decir, si al utilizar el gestor el usuario se sintió satisfecho, tenso, etcétera).

Así, en la Figura 5 se observa la pantalla principal del gestor propuesto, una vez que el usuario ha entrado al gestor de documentos mediante un usuario y contraseña (si no se cuenta con ello se puede registrar y esperar autorización del sistema). Una vez que el usuario se ha registrado al sistema, observará las opciones disponibles dependiendo de sus privilegios. Un ejemplo de un usuario registrado en el sistema se puede observar en la Figura 5.

En la Figura 6 se muestran documentos añadidos al gestor de documentos. También se muestra la interfaz para crear un nuevo documento (Figura 7). Antes de crear el cuerpo del documento, debe ser llenada y guardada la primera parte del formulario presentado. Después, para crear el cuerpo del documento se debe utilizar el editor presentado y posteriormente se debe guardar. En la Figura 8 se muestra la interfaz de búsqueda de documentos implementada en el gestor propuesto.

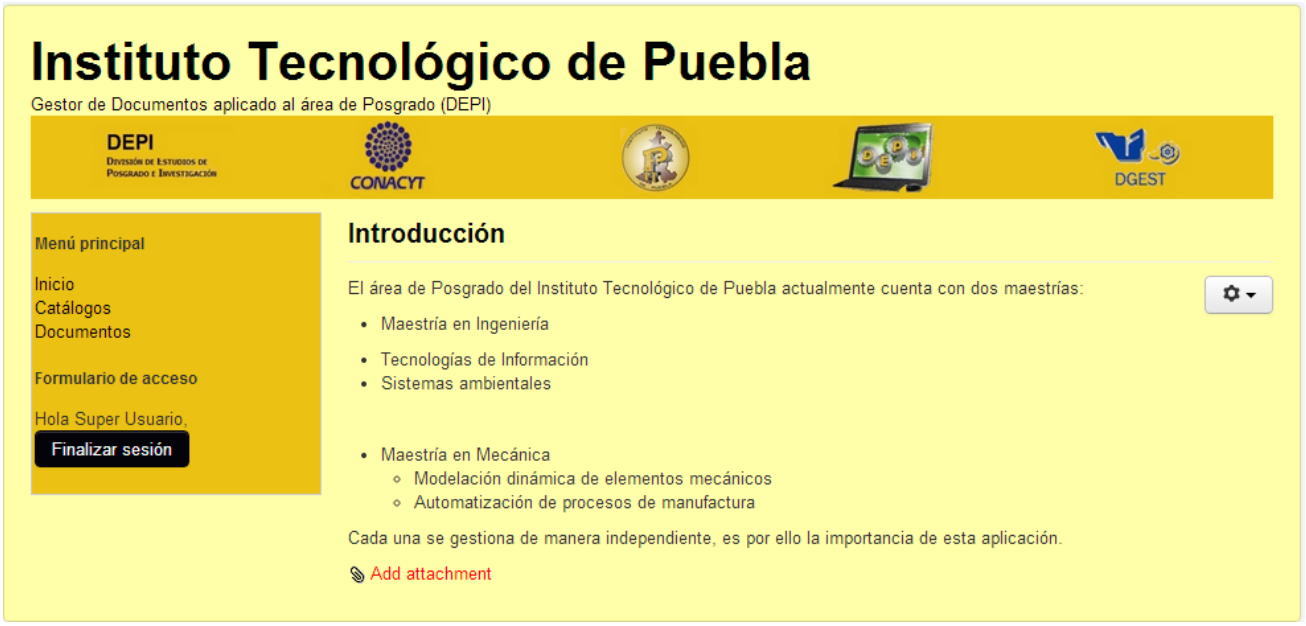

Figura 5. La interfaz muestra las opciones disponibles al usuario conectado.

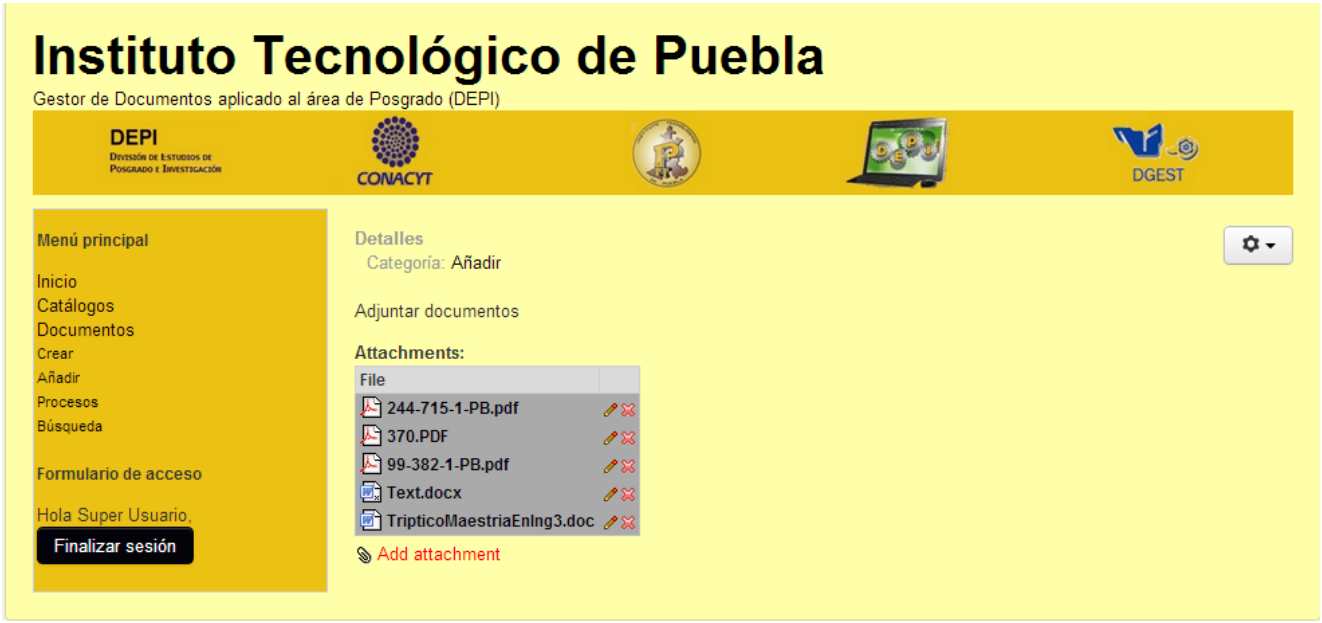

Figura 6. Interfaz para anexar documentos.

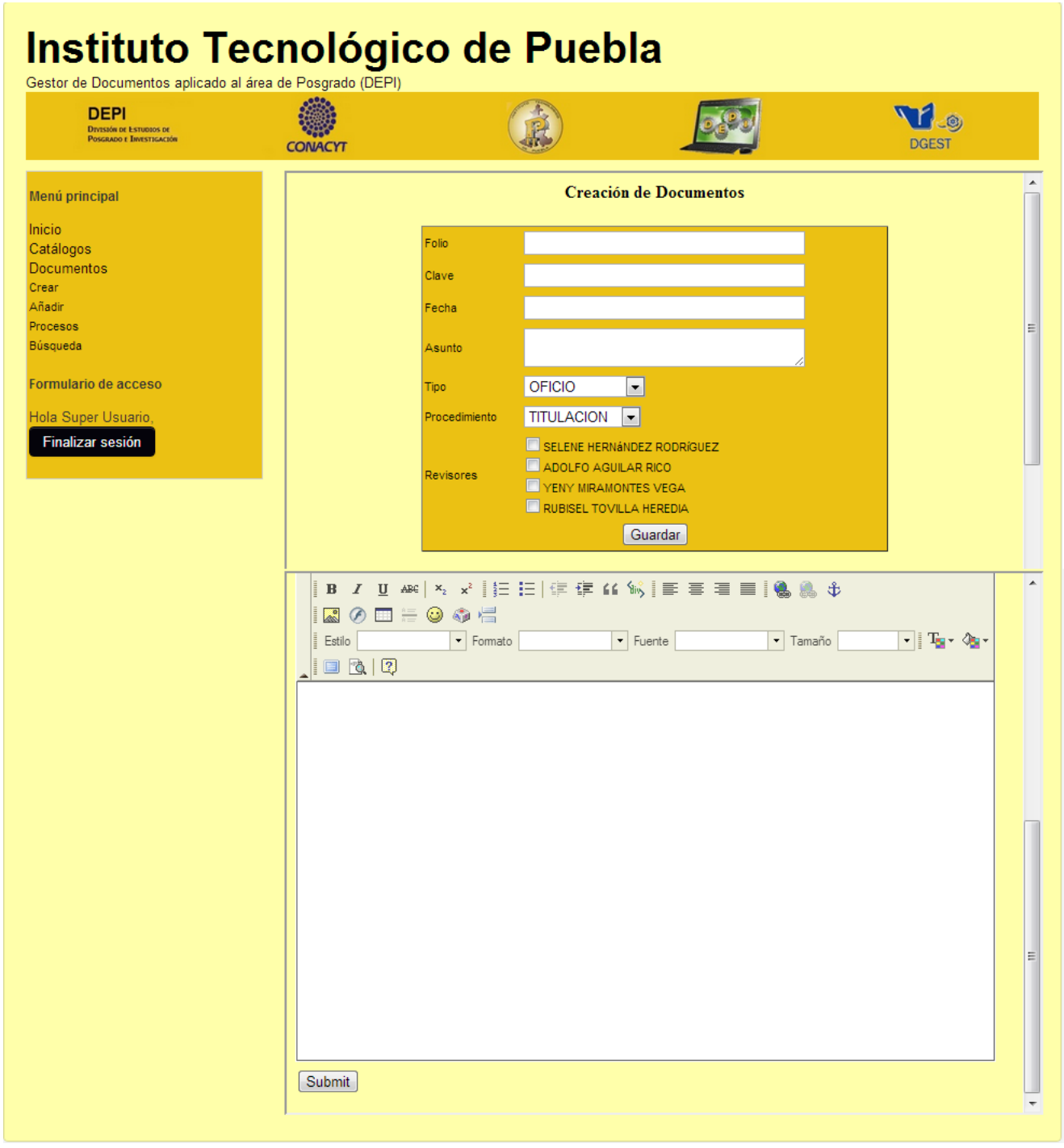

Figura 7. Interfaz de creación de un documento nuevo.

Figura 8. Interfaz de búsqueda de documentos.

4.2 Evaluación del funcionamiento de las tecnologías Oracle Text y JiFile

Con el objetivo de evaluar el funcionamiento del módulo de búsqueda y recuperación de documentos del gestor propuesto, se analizaron las tecnologías de Oracle Text y JiFile, las cuales fueron seleccionadas a partir del análisis mencionado en la sección 2. Estas dos tecnologías se probaron con los operadores "TÉRMINO EXACTO" y "ABOUT", descritos en la Tabla 1. El operador "TÉRMINO EXACTO" realiza una búsqueda exhaustiva sobre todos los documentos, buscando exactamente el término o frase deseada. El operador "ABOUT" realiza una búsqueda más robusta, ya que considera el caso de sinónimos y raíces de palabras para considerar conjugaciones de verbos o el singular y plural, entre otras cosas. Este operador también utiliza un índice de palabras, de tal manera que, dada una búsqueda, no es necesario analizar exhaustivamente todos los documentos sino sólo esta estructura, siendo así más rápida y robusta la búsqueda.

Las pruebas antes mencionadas se realizaron en una Laptop con las siguientes características: 4 gigabytes en RAM, disco duro de 500 gigabytes y procesador Intel Core i3-2310M CPU a 2.10 gigahertz. Para realizar la comparación entre JiFile y Oracle Text, se utilizó un conjunto de documentos administrativos reales de la División de Estudios de Posgrado e Investigación (DEPI) del Instituto Tecnológico de Puebla (ITP).

4.2.1 Criterios de comparación para evaluar el funcionamiento de Oracle Text y JiFile

Para la evaluación del funcionamiento de las tecnologías de Oracle Text y JiFile se utilizó el porcentaje de documentos recuperados correctamente de acuerdo a la fórmula (1), donde *DocsTotales* corresponde al número total de documentos relacionados con cierto tema y *DocsRecCorrectamente* corresponde al número de documentos que se recuperaron correctamente mediante alguna tecnología.

$$PorcDocs = \frac{DocsRecCorrectamente * 100}{DocsTotales} \qquad (1)$$

También, en los sistemas de recuperación de información se utilizan para evaluar las siguientes medidas: precisión, especificación, exactitud y rendimiento, las cuales dividen el resultado arrojado por los buscadores en resultados relevantes y no relevantes, como se muestra en la Figura 9:

		HUMANO		
		Relevante	*Irrelevante*	
TIC	*Recuperados*	a	b	k = a + b *Todos los documentos recuperados.*
	No recuperados	c	d	m = c + d *Todos los documentos omitidos*
		r = a + c *Todos los documentos relevantes de la colección*	s = b + d *Todos los documentos irrelevantes de la colección*	n = a + b + c + d *Todos los documentos de la colección*

Figura 9. Cálculo de métricas sobre documentos

Precisión: La precisión es la fracción de documentos recuperados que son realmente relevantes para los criterios definidos en la búsqueda, entre el total de documentos recuperados por el sistema (a + b). Se define como la proporción entre el número de documentos relevantes recuperados y el total de documentos relevantes en la colección y se define mediante la siguiente ecuación:

$$Precisión = \frac{a}{k}$$

Especificidad: La especificidad (recall), es la fracción del conjunto de documentos relevantes que han sido correctamente recuperados por el sistema, y mide la habilidad de recuperar información relevante de toda la colección de documentos. El cálculo de la especificidad requiere el conocimiento del número de exacto de documentos relevantes que hay en la colección (a + c). También se realiza una estimación del número de documentos relevantes no recuperados (c) como se especifica en la siguiente ecuación:

$$Especificidad = \frac{a}{r}$$

Exactitud: La exactitud se entiende como la capacidad del sistema de identificar los elementos, tanto relevantes como no relevantes, de acuerdo a la clasificación realizada por un experto del área correspondiente. Su fórmula es la siguiente:

$$Exactitud = \frac{a+d}{n}$$

Rendimiento: Para obtener el rendimiento general del sistema, se seleccionó la medida-F propuesta por Manning y Schütze [11], la cual combina, dentro de una misma métrica, la precisión y la especificidad sin ser afectada por el tamaño de la colección. El rendimiento se calcula tomando como base la precisión y la especificidad, y se puede ver el rendimiento como el mayor número de aciertos con el menor número de fallas. Se define como sigue:

$$Rendimiento = \frac{2pr}{p+r}$$

Estas cuatro funciones están acotadas en el intervalo [0, 1]. Cuando el valor es cercano a 1 significa que se está obteniendo un buen funcionamiento, es decir, la herramienta es precisa (para el caso de la función *Precisión*), específica (para el caso de la función *Especificidad*), exacta (para el caso de la función *Exactitud*) o de buen rendimiento (para el caso de la función *Rendimiento*), mientras que si el valor es cercano a cero se tiene el caso opuesto.

4.2.2 Pruebas realizadas con el conjunto de documentos de la DEPI del ITP

Se contó con un conjunto de documentos extraídos de la DEPI del ITP formado por 1065 documentos de texto (DOC), los cuales corresponden a los siguientes temas: alumnos, profesores, horarios, entre otros. De estos documentos se seleccionaron los temas que se muestran en la Tabla 2 para realizar algunas búsquedas,

de los cuales se sabe el número de documentos relacionados con estos temas.

Tabla 2. Conjunto de documentos de la DEPI

TEMAS	*TOTAL*
Tecnologías de la información	42
TIC	10
Horario irregular	39
Evaluación de un lodo residual como moderador en la fertilidad del suelo	10
Pago sinodales	47
Ambiental	212
Alejandro Perez Cortes	20
Propedéutico	36
María Evelinda Santiago Jiménez	120
PNPC	11

Se realizaron pruebas de recuperación, con diez temas diferentes, a través de búsquedas simples, búsquedas con dos palabras o búsquedas con un tema de tesis. Los resultados de la recuperación de información se muestran en la Tabla 3. En la primera columna (TEMAS) de esta tabla se muestra el término a buscar (por ejemplo: "Tecnologías de la Información"), y en la segunda columna (TOTAL) se muestra el total de documentos que están clasificados con este término. Por ejemplo, existen 42 documentos (de los 1065 documentos totales) clasificados como documentos relacionados con "Tecnologías de la Información".

Tabla 3. Búsqueda y recuperación de documentos con la información de la DEPI

		BÚSQUEDA POR TÉRMINO EXACTO				*BÚSQUEDA CON EL OPERADOR ABOUT*			
TEMAS	*TOTAL*	*JIFILE*	*Porc de archivos correctos*	*ORACLE*	*Porc de archivos correctos*	*JIFILE ABOUT*	*Porc de archivos correctos*	*ORACLE ABOUT*	*Porc de archivos correctos*
Tecnologias de la información	42	1059-36	85.71	103-36	85.71	1059-36	85.71	103-36	85.71
TIC	10	9-8	80.00	9-9	90.00	9-8	80.00	9-9	90.00
Horario irregular	39	65-18	46.15	66-18	46.15	65-18	46.15	66-18	46.15
Evaluación de un lodo residual como moredador en la fertilidad del suelo	10	1061-6	60.00	1040-6	60.00	1061-6	60.00	1040-6	60.00
Pago sinodales	47	52-39	82.98	51-39	82.98	52-39	82.98	51-39	82.98
Ambiental	212	227-50	23.58	231-50	23.58	227-50	23.58	231-50	23.58
Alejandro Perez Cortes	20	102-15	75.00	109-15	75.00	102-15	75.00	109-15	75.00
Propedeutico	36	14-13	36.11	14-14	38.89	14-13	36.11	14-14	38.89
María Evelinda Santiago Jiménez	120	311-104	86.67	411-104	86.67	311-104	86.67	411-104	86.67
PNPC	11	8-8	72.73	8-8	72.73	8-8	72.73	8-8	72.73
% Promedio de documentos recuperados			64.89		66.17		64.89		66.17

El resultado del primer tipo de búsqueda (por término exacto) se observa en la columna 3 de la Tabla 3 para JiFile, y en la columna 5 para Oracle Text. En la columna 3 se observan dos números separados por

un guión, el primer número representa los documentos recuperados por la tecnología de búsqueda y el segundo es el número de documentos recuperados correctamente, es decir, que realmente pertenecen al tema para la tecnología JiFile (y en la columna 5 para Oracle Text). Por ejemplo, en el primer renglón de la tercera columna, se puede ver el número 1059-36, lo que quiere decir que de los 1059 documentos que reporta como resultado JiFile, 36 pertenecen a este tema. Los 36 documentos que regresa correctamente JiFile son el 85.71% de los 42 documentos relacionados con este tema (PorcDocs, descrito en la ecuación 1). Este porcentaje se muestra en la columna 4 para JiFile y en la columna 6 para Oracle Text. Este mismo análisis se repite para el operador ABOUT, cuyos resultados se muestran en las columnas 7-10. A partir de esta tabla (Tabla 3) se puede observar que se obtienen exactamente los mismos resultados utilizando ambas tecnologías.

Tabla 4. Tiempos en la recuperación de información de la DEPI

TIEMPOS DE RECUPERACIÓN (segundos)				
BÚSQUEDAS	*TÉRMINO*		*ABOUT*	
	Oracle Text	*JiFile*	*Oracle Text*	*JiFile*
Tecnologias de la informacion	0.05	0.61	0.05	0.86
TIC	0.01	0.02	0.01	0.02
Horario irregular	0.03	0.02	0.03	0.04
Evaluación de un lodo residual como moredador en la fertilidad del suelo	0.45	0.61	0.45	0.61
Pago sinodales	0.03	0.04	0.03	0.04
Ambiental	0.11	0.17	0.11	0.14
Alejandro Perez Cortes	0.05	0.03	0.05	0.03
Propedeutico	0.01	0.01	0.01	0.02
María Evelinda Santiago Jiménez	0.18	0.20	0.18	0.20
PNPC	0.01	0.02	0.01	0.02
PROMEDIO	0.08	0.14	0.08	0.16

En la Tabla 4 se muestran los tiempos de la recuperación de documentos. A partir de esta tabla, se puede observar que el tiempo requerido por cada tecnología es muy similar, siendo Oracle Text el que obtuvo menor tiempo de recuperación tanto por búsquedas por término simple como en la búsqueda por el operador ABOUT. Esto se puede observar mejor en la Figura 10.

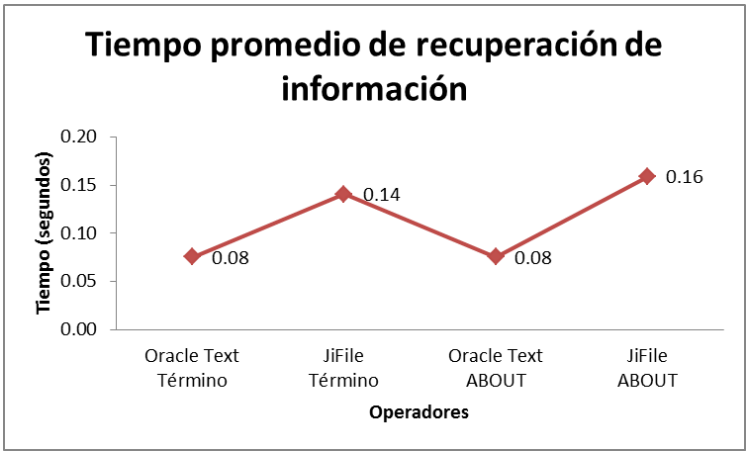

Figura 10. Tiempo promedio de recuperación de información de la DEPI

Con este conjunto de documentos, también se evaluaron las siguientes métricas: precisión, recuperación, especificidad y rendimiento, obteniendo los resultados que se muestran en la Tabla 5. A partir de esta tabla se puede observar que se obtienen mejores resultados con la tecnología Oracle Text. Sin embargo, cabe resaltar que la diferencia es muy pequeña. Estos resultados se muestran gráficamente en la Figura 11.

Tabla 5. Evaluación de métricas de la recuperación de información sobre documentos de la DEPI

BÚSQUEDA	*PRECISIÓN*		*ESPECIFICIDAD*		*EXACTITUD*		*RENDIMIENTO*	
	JIFILE	*ORACLE TEXT*	*JIFILE*	*ORACLE TEXT*	*JIFILE*	*ORACLE TEXT*	*JIFILE*	*ORACLE TEXT*
TECNOLOGÍAS DE LA INFORMCIÓN	0.034	0.350	0.857	0.857	0.034	0.931	0.065	0.497
TIC	0.889	1.000	0.800	0.900	0.997	0.999	0.842	0.947
HORARIO IRREGULAR	0.277	0.273	0.462	0.462	0.936	0.935	0.346	0.343
EVALUACIÓN DE UN LODO RESIDUAL COMO MEJORADOR EN	0.006	0.006	0.600	0.600	0.006	0.025	0.011	0.011
PAGO SINODALES	0.750	0.765	0.830	0.830	0.980	0.981	0.788	0.796
AMBIENTAL	0.220	0.216	0.236	0.236	0.682	0.678	0.228	0.226
ALEJANDRO PÉREZ CORTÉS	0.147	0.138	0.750	0.750	0.914	0.907	0.246	0.233
PROPEDÉUTICO	0.929	1.000	0.361	0.389	0.977	0.979	0.520	0.560
MARÍA EVELINDA SANTIAGO JIMÉNEZ	0.334	0.253	0.867	0.867	0.791	0.697	0.483	0.392
PNPC	1.000	1.000	0.727	0.727	0.997	0.997	0.842	0.842
PROMEDIO	0.459	0.500	0.649	0.662	0.731	0.813	0.437	0.485

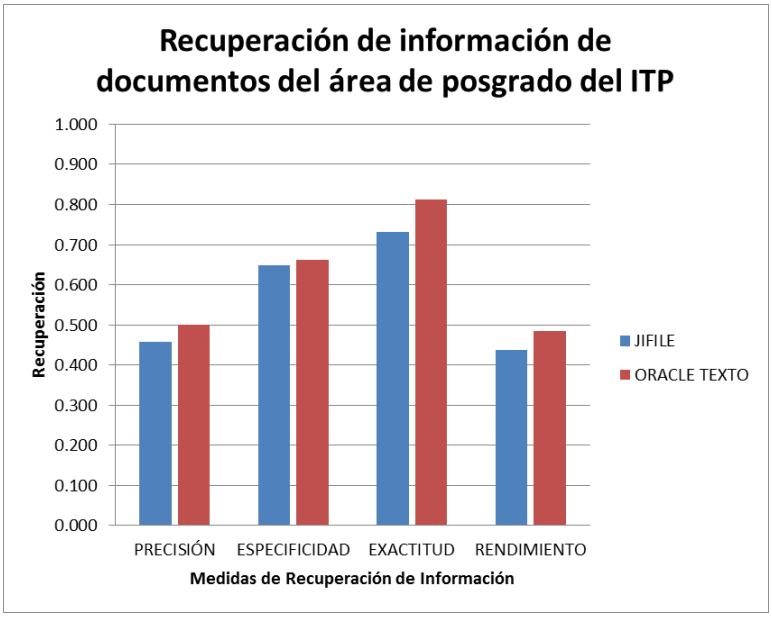

Figura 11. Comparación de las métricas de evaluación.

4.2.3 Discusión de resultados

De acuerdo a la medida de comparación *PorcDocs*, la cual representa el porcentaje de documentos recuperados correctamente, es posible concluir que los resultados obtenidos son muy similares para las dos tecnologías de búsqueda y recuperación de documentos, por lo que resulta difícil tomar una decisión.
Mediante las métricas de recuperación de información (precisión, especificidad, exactitud y rendimiento), se puede observar que utilizando Oracle Text se obtienen ligeramente mejores resultados. Sin embargo, dado que los resultados son muy similares para ambas tecnologías, decidir cuál de ellas elegir tal vez recaiga en el hecho de que una de ellas es comercial (Oracle Text) y la otra es libre (JiFile).

También es importante mencionar que la tecnología Oracle Text tiene una mayor especialización en el idioma inglés. Tomando en consideración lo expuesto hasta el momento, se seleccionó la tecnología JiFile como el motor de búsqueda para ser implementado en el gestor de documentos propuesto en este trabajo.

Conclusiones

En este trabajo se propone un sistema de gestión de documentos administrativos aplicable a instituciones como el Instituto Tecnológico de Puebla. Este sistema es necesario para llevar un mejor control de documentos.

Este sistema permitirá consultar documentos pendientes, los cuales se pueden verificar y autorizar desde cualquier computadora que cuente con Internet, así como también su resguardo eficiente, debido a que se archivan de manera electrónica, dejando a un lado la preocupación de que se llegue a perder algún documento o traspapelar en algún archivero.

Este proyecto centró su atención en los servicios existentes para búsqueda de documentos

electrónicos, así como los algoritmos que los conforman. Se utilizó el servicio que otorga la extensión JiFile para el gestor de contenidos Joomla, y esta extensión tiene su base en la biblioteca Lucene. Mediante el motor de búsqueda JiFile se resolvió el reto de búsqueda inteligente sobre los documentos electrónicos con que cuenta la DEPI.

Los resultados obtenidos con las pruebas realizadas en ambas tecnologías de búsqueda arrojaron resultados similares tanto en tiempo como en recuperación de documentos, por lo que se eligió la extensión JiFile por ser de código abierto y de bajo costo. Finalmente este gestor de documentos contribuye a reducir el consumo de papel y tóner porque todos los procesos se manejan de manera electrónica.

Referencias

1. Climente, Carlos (17 julio 2001) ¿Qué es la oficina sin papel? *Ideas y negocios en red*. Disponible en: http://winred.com/internet/relaciones-con-los-clientes-y-nuevas-tecnologias/gmx-niv113-con20.htm
2. InspirAction, 2009. "Reciclaje del tóner". Disponible en: http://www.inspiraction.org/
3. Encinas, Juan 2008. "El tóner de impresoras". Disponible en: http://basura-electronica.blogspot.com/
4. Mota, Gabriel 2009. "Riesgos provocados por el tóner de fotocopiadoras e impresoras láser" Disponible en: www.cta.org.ar/base/IMG/pdf/riesgos_impresoras.pdf
5. Lennon, M., Peirce, D.S., Tarry, B.D. and Willett, P. (1981), "An evaluation of some conflation algorithms for information retrieval", Journal of Information Science, Vol. 3 No.4, pp. 177-183.
6. Büttcher, Stefan. Charles L.A. Clarke y Gorden V. Cormack(2010) Information Retrieval, implementating and evaluating search engines. MIT Press, USA. pp. 1-29, 105-131
7. Grossman, David A. Frieder, Ophir (2004), Information Retrieval, Algorithms and Heuristics, Second Edition, USA, Springer, pp 182-184.
8. Manning, Christoper D. y SchützeHinrich (1999) Foundations of statistical natural language processing. MIT Pres, Cambridge, Massachusetts, USA.
9. Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (1998). Extensible markup language (XML). World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210.
10. Baeza-Yates, R., Boldi, P., & Goméz Hidalgo, J. (enero-febrero de 2007). Presentación: buscando en la Web del futuro. Novática(185), 3-4.
11. Manning, Christoper D. y Schütze Hinrich (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, Massachusetts, USA.

Bibliografía complementaria

- Deerwester S., Dumals S., Furnas T., Landauer G. and Harshman R. "Indexing by latent semantic analysis". J. Amer. Soc. Inform. Sci. 41, pp. 391-407, 1990.
- IiFile (2011) IFile-Introduzioneall'utilizzo – versione 1.1
- JiFile (2011) JIFILE – IntroduzioneAll'Utilizzo - versiones 1.0
- Loney, Kevin. Oracle Database The complete reference, Master the Powerful Features of the Lastest Database Release, McGrawHill, Oracle Press 2009. ISBN: 978-0-07-159876-7
- Lovins, J.B. (1968), "Development of a stemming algorithm", Mechanical Translation and Computational Linguistics, Vol. 11 Nos 1 and 2, pp. 22-31.
- McCandless, Michael. Hatcher, Erik. Gospodnetic, Otis. (2010), Lucene in Action, Second Edition, USA. Manning.
- Oracle (September 2012) Oracle Database, Installation Guide 11g Release 2 (11.2) for Linux E24321-07
- Oracle (May 2012) Oracle Database, Installation Guide 11g Release 2(11.2) for Microsoft Windows