IMPLEMENTACIÓN DE UNA APLICACIÓN WEB DE BÚSQUEDA EMPRESARIAL Y SU INTEGRACIÓN EN EL SISTEMA GESTOR FIDUCIA FONDOS JEE DE LA EMPRESA GESTORINC. S.A.

Jessica Balarezo Galarza[1], Henry Coral Coral[2], Santiago Salvador[3]

1 Universidad de las Fuerzas Armadas - ESPE, Ecuador, karo793@gmail.com

2 Universidad de las Fuerzas Armadas - ESPE, Ecuador, hrcoral@espe.edu.ec

3 Universidad de las Fuerzas Armadas - ESPE, Ecuador, mssalvador@espe.edu.ec

RESUMEN

Con el tiempo el almacenamiento de documentos en repositorios relacionales (bases de datos) de las empresas fue creciendo constantemente; sin embargo al momento de buscar información dentro de cada uno de ellos se volvió una tarea compleja que muchas veces era imposible de realizarse debido a las limitaciones tecnológicas. Por esta razón muchas empresas se vieron en la necesidad de adaptar sus sistemas empresariales transaccionales con sistemas de búsqueda documental los cuales tienen la capacidad de acceder a estos repositorios y extraer la información de cada documento para ejecutar los procesos de búsqueda. Actualmente existen diferentes herramientas en el mercado de licenciamiento privado o público para ser integradas en diferentes plataformas de software de las empresas. El presente artículo detalla la implementación de un sistema de Búsqueda Documental genérico el cual puede ejecutar de forma independiente búsquedas de texto en documentos de formato propietario como Microsoft Office, PDF, entre otros y que se encuentran almacenados en una base de datos relacional. El producto final desarrollado y que es la parte central de este artículo se llama “Gestor Search”; una aplicación web desarrollada bajo los estándares de la plataforma JEE y que integra el uso de las librerías Apache Tika y Apache Lucene, para cumplir con este propósito, siendo una aplicación de búsqueda empresarial que cuenta con las funcionalidades de indexación, búsqueda y acceso a los documentos.

Palabras Clave: Búsqueda, Tika, Lucene.

ABSTRACT

Eventually storing documents in relational repositories (databases) of companies grew steadily; however when seeking information within each a complex task that often became impossible to perform was due to technological limitations. For this reason many companies saw the need to adapt their systems with transactional enterprise information retrieval systems which have the ability to access these repositories and extract information from each document to run the search processes. Currently exist several tools on the market for private or public licensing that can be integrated into the various software platforms companies. This paper details the implementation of a generic search Documentary system which can run independently text searches on documents proprietary format as

Microsoft Office, PDF, among others, that are stored in a relational database. The final product developed that is the central part of this article is called "Gestor Search"; a web application developed under the standards of the JEE platform and integrates the use of libraries and Apache Tika Apache Lucene, to accomplish this purpose. Gestor Search is an application enterprise search functionality has indexing, search and access to documents.

KeyWords: Search, Tika, Lucene.

1. INTRODUCCIÓN

Con el paso de los años las empresas han optado por el uso de sistemas transaccionales, algunos de estos sistemas permiten el almacenamiento de documentos en bases de datos relacionales utilizando para esto campos binarios. Si bien este mecanismo provee la funcionalidad de gestión de los documentos no permite la búsqueda de textos dentro de los mismos. Lo cual hace difícil el acceso a los mismos. Debido a esta necesidad se han creado algunas herramientas que son compatibles con los sistemas existentes para poder integrarlas y así crear una sección específica para la búsqueda de la información y facilitar el trabajo diario.

Una de estas herramientas que permite la extracción del texto de los documentos es Oracle Secure Enterperise Search distribuida por Oracle Corporation, sin embargo su costo de licenciamiento, la complejidad de su implantación sin una previa capacitación y la adquisición de una infraestructura de hardware adecuada para su uso hace que la misma no sea muy utilizada en la actualidad; así mismo, también existen otras herramientas de menos costo pero de difícil integración que se encuentran a disposición del público hoy en día.

En base a estos antecedentes la empresa GestorInc. S.A. auspicio este trabajo de investigación para el desarrollo de una herramienta de búsqueda empresarial que pueda ser ofertada a sus clientes sin costo adicional, sino como un servicio adicional, generando así una ventaja competitiva para la empresa.

Como producto de esta investigación se realizó el desarrollo de la aplicación Gestor Search la cuál utiliza las herramientas Apache Lucene y Apache Tika distribuidas bajo la licencia Apache, que en conjunto permiten la extracción de texto desde formatos propietarios (binarios) de documentos y la posterior indexación del mismo para la ejecución de búsquedas empresariales.

El presente artículo se definió de la siguiente manera: En la sección 2 se detallan las características técnicas de las herramientas Apache Lucene y Apache Tika. En la sección 3 se encuentra el diseño e implementación de la aplicación Gestor Search. La sección 4 se presenta las conclusiones que se obtuvieron al final del proyecto y los trabajos futuros en base a los resultados obtenidos del mismo.

2. MATERIALES Y MÉTODOS

La búsqueda de documentos en la empresa es una parte fundamental ya que se realizan varias cargas al día, por eso la gestión documental es una parte fundamental para la indexación y búsqueda de los mismos.

2.1 Gestión Documental

La práctica de la gestión documental está dada por un conjunto definido de normas técnicas y algunas prácticas generales que son usadas para: especificar en qué tiempo deben almacenarse los documentos, la administración del flujo de documentos de todo tipo en una organización, la recuperación de la información, la eliminación de los documentos que ya no sirven, la conservación de todos los documentos que se consideren valiosos indefinidamente.

Teniendo en cuenta el papel muy importante que tiene la gestión documental dentro de una organización está también agiliza la búsqueda de los documentos almacenados en los repositorios siempre y cuando se administre bien la gestión documental. (Nayar, 2010)

2.2 Apache Lucene

Apache Lucene posee todas las funciones de un motor de búsqueda de texto escrito completamente en Java que tiene un alto rendimiento y tiene versiones en otros lenguajes como Perl, Python o .NET. Se trata de una tecnología desarrollada que puede integrarse a cualquier aplicación que requiera la búsqueda de texto, especialmente multiplataforma. Lucene ofrece dos servicios principales: la indexación de texto y búsqueda de texto. Estas dos operaciones son relativamente independientes entre sí. (Erik Hatcher, 2004) (Carpenter, 2011)

Las principales características son: (Foundation, Introducción para Apache Lucene, 2010)

- Posee la capacidad de búsqueda en varios idiomas.
- Indexa cualquier archivo que pueda convertirse en texto.
- Realiza la clasificación de la búsqueda para que aparezca primero los mejores resultados.
- Las búsquedas pueden realizarse por campos ya sean por contenidos, autores, títulos, etc. (Foundation, Apache Lucene Core, 2011-2012)

2.3 Apache Tika

Apache Tika es una librería desarrollada en Java, que detecta y extrae los metadatos y el contenido de texto estructurado de documentos en diferentes formatos (Microsoft Word, Excel, Power Point, PDF, txt, etc) mediante bibliotecas analizadoras existentes que no son creados por Apache Tika pero son basadas en otros proyectos existentes de código abierto.

La gran ventaja de Tika para la extracción de datos es que no se necesita indicarle que tipos de documentos se analizará, si no que lo realiza automáticamente. Tika es un proyecto de la Fundación Apache, y anteriormente fue un sub-proyecto de Apache Lucene. (Chris A. Mattmann, 2012)

3 DISEÑO E IMPLEMENTACIÓN

3.1 Arquitectura de la Aplicación Gestor Search

Gestor Search es una aplicación empresarial construida bajo los estándares de la plataforma JEE, que utiliza principalmente las librerías de los productos Apache Lucene y Apache Tika; para implementar el concepto de “Búsqueda Empresarial” en los diferentes documentos que se encuentran almacenados en la Base de Datos del sistema Gestor Fiducia Fondos.

La aplicación desarrollada está constituida por dos funcionalidades principales: Funcionalidad de Indexación Incremental de Documentos y Funcionalidad de Búsqueda de Documentos, cada una de ellas cuenta con algunas funcionalidades internas.

La arquitectura lógica de la aplicación desarrollada se la describe en la Figura 3.1.

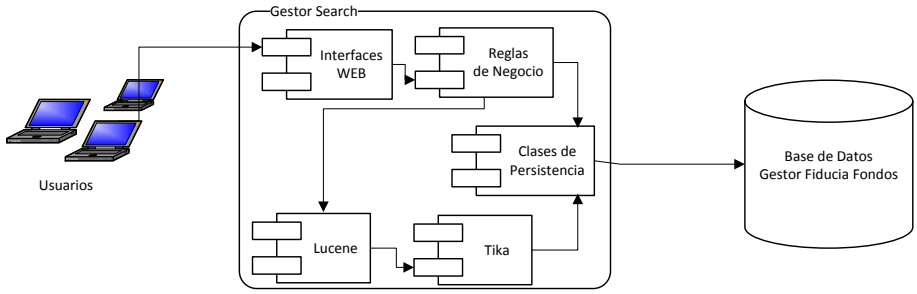

Figura 3. 1 Arquitectura Lógica de Gestor Search

3.1.1 Funcionalidad de Indexación

Esta funcionalidad fue la más compleja de implementar en la aplicación y se considera la más importante, ya que de esta depende la ejecución de la búsqueda de texto en documentos. Apache Lucene es una librería que realiza la indexación en base a archivos de texto, sin embargo, los archivos de texto plano no son un estándar de uso y almacenamiento en las empresas, por lo que fue necesario en primer lugar extraer el texto de los documentos que se encuentran en formatos propietarios como (PDF, archivos de Microsoft Office, archivos de OpenOffice, etc.); para que este texto sea utilizado por Lucene y pueda ser indexado utilizando para esto las librerías de Apache Tika.

El proceso de indexación es el tratamiento de los datos originales en una muy eficiente búsqueda de referencia cruzada con el fin de facilitar las búsquedas rápidas almacenadas en un índice para su rápido acceso y eliminar el proceso de exploración secuencial lento. Un índice de Lucene se almacena en un único directorio del sistema de archivos en un disco duro, está compuesto de segmentos, documentos, campos y términos. Cada segmento contiene uno o más documentos. Cada documento tiene uno o más campos, y cada campo contiene una o más términos. Cada término es un par de cadenas que representan un nombre de campo y un valor.

El esquema general de la funcionalidad de indexación se lo presenta en la Figura 3.2 que representa al diagrama de las clases desarrolladas en la implementación de la funcionalidad.

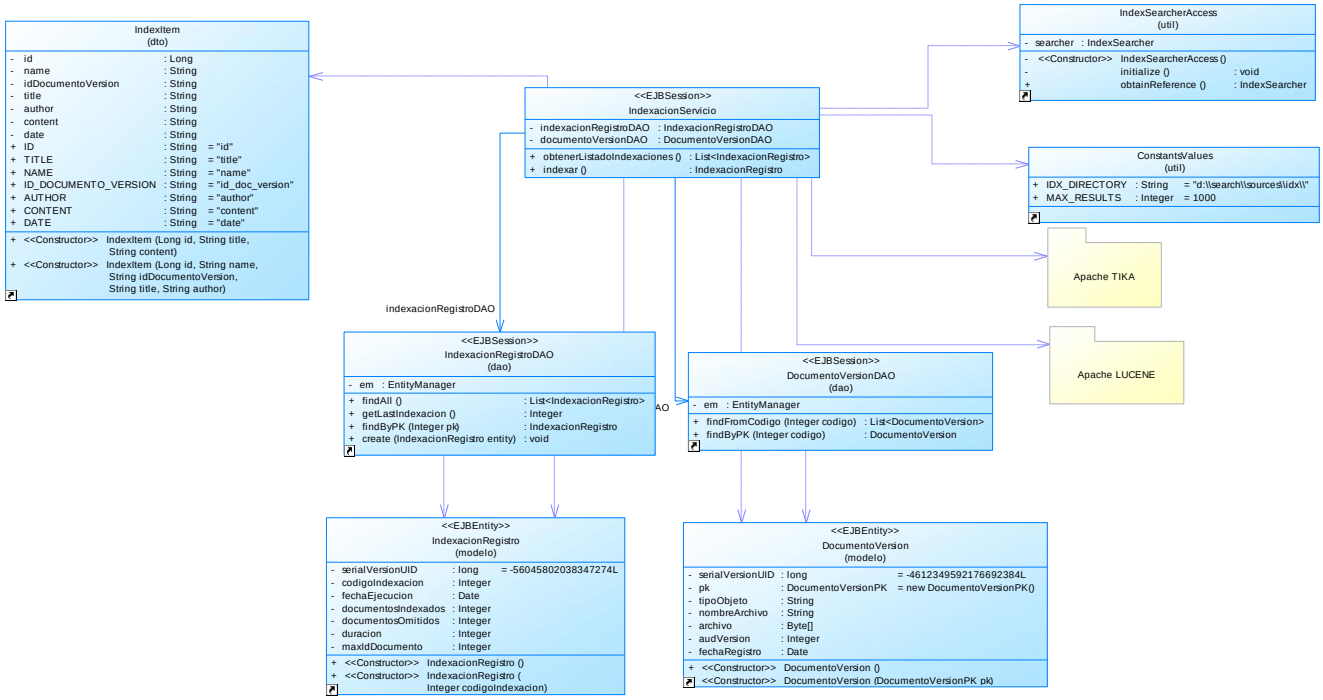

Figura 3. 2 Diagrama de clases de funcionalidad de Indexación

La clase principal de esta funcionalidad es ***IndexacionServicio***; esta clase define el método ***indexar***(); el cual se encarga de implementar todo el proceso de indexación explicado anteriormente.

El proceso de indexación implementado es incremental, esto significa que cada vez que se ejecuta la indexación de archivos primero se busca en la base de datos información relacionada al último proceso de indexación para obtener el ID del último documento indexado, con esta información el proceso indexa solamente los nuevos archivos, es decir, los archivos cuyo ID sea mayor al obtenido; de esta forma, el proceso de indexación no tarda en su ejecución y tampoco se adueña de recursos que pueden afectar al rendimiento de todo el sistema transaccional.

3.1.2 Funcionalidad de Búsqueda

Esta funcionalidad se encarga de la búsqueda en los archivos de los índices generados con la cadena de texto ingresado. La librería Apache Lucene implementa esta funcionalidad, adicionalmente se desarrolló una lógica propia parte para obtener los fragmentos de texto coincidentes de cada documento encontrado y resaltar las palabras que coinciden con la cadena de búsqueda enviada; después de obtener el texto que coincidió se presenta un link para la descarga de cada uno de los archivos que forman parte del resultado en la interfaz de usuario.

El esquema general de la funcionalidad de búsqueda se lo presenta en la Figura 3.3 que representa al diagrama de las clases desarrolladas en la implementación de la funcionalidad.

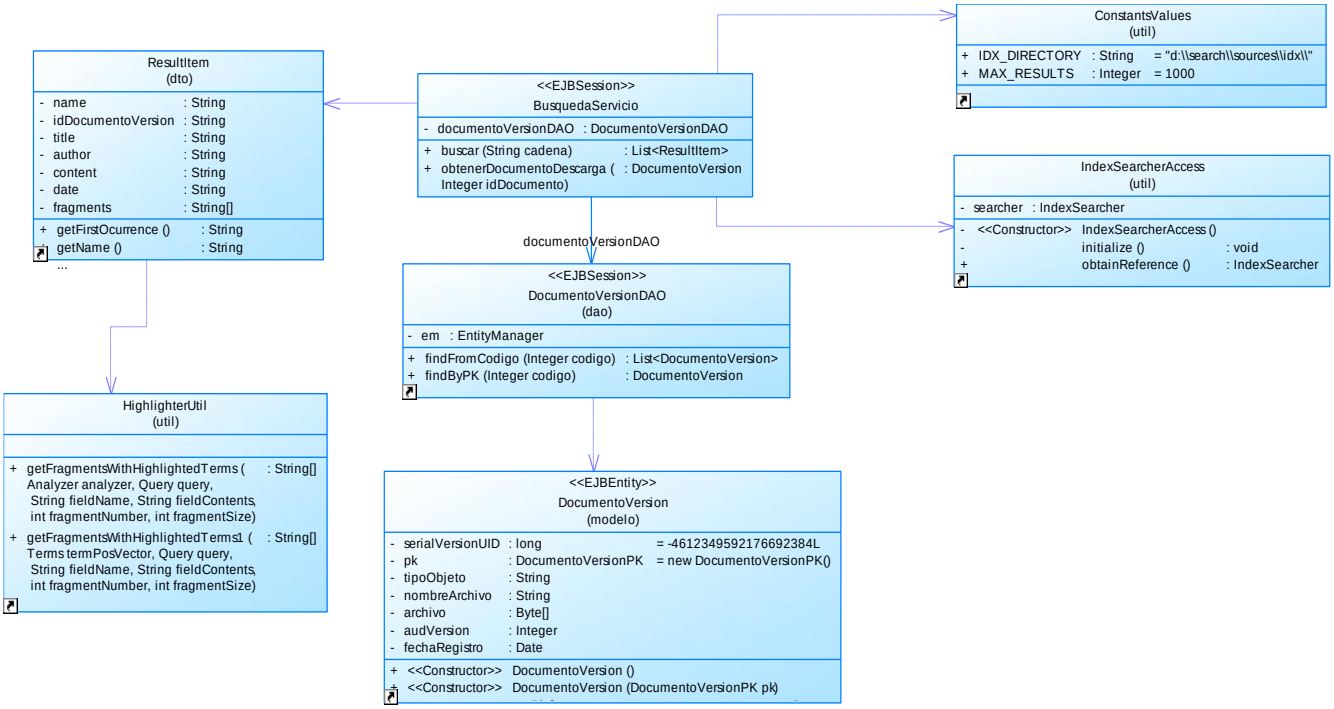

Figura 3. 3 Diagrama de clases de la funcionalidad de Búsqueda

La clase principal de esta funcionalidad es ***BusquedaServicio***. El proceso de búsqueda recibe desde la interfaz de usuario, la cadena de búsqueda, la cual contiene el texto que el usuario desea buscar en los documentos indexados. La clase ***BusquedaServicio*** tiene el método ***buscar(String cadena)*** que internamente implementa el proceso de búsqueda, para esto, primero utiliza la librería Lucene, quien es la encargada de buscar la cadena de texto en los archivos de índices anteriormente generados, y retorna los archivos de resultado en orden de importancia; para esto utiliza una serie de algoritmos estándares.

3.1.3 Integración de Gestor Search con el sistema Gestor Fiducia Fondos

La integración de la aplicación Gestor Search con el sistema Gestor Fiducia Fondos JEE, se la ha realizado por medio de la inclusión de dos funcionalidades, estas tienen internamente elementos de tipo “iframe” que realizan el llamado a la funcionalidad de indexación y a la funcionalidad de búsqueda, de esta forma se tiene una independencia del sistema Gestor Fiducia Fondos JEE con la aplicación Gestor Search, permitiendo a esta última su propia evolución, disminuyendo al máximo la cohesión entre las dos aplicaciones.

En la figura 3.3 se muestra la interfaz de indexación con el resultado de cuantos documentos se indexaron, la fecha, duración. En la figura 3.4 se muestra la pantalla de búsqueda, en la cual se ingresa cualquier palabra o frase que se desee buscar y finalmente en la figura 3.5 se puede ver los resultados de búsqueda con los documentos ordenados desde el más importante que contiene la palabra o frase ingresada anteriormente, la interfaz de usuario de despliegue de resultados es similar a la del buscador Google para que el usuario sienta familiaridad al momento de utilizar esta funcionalidad.

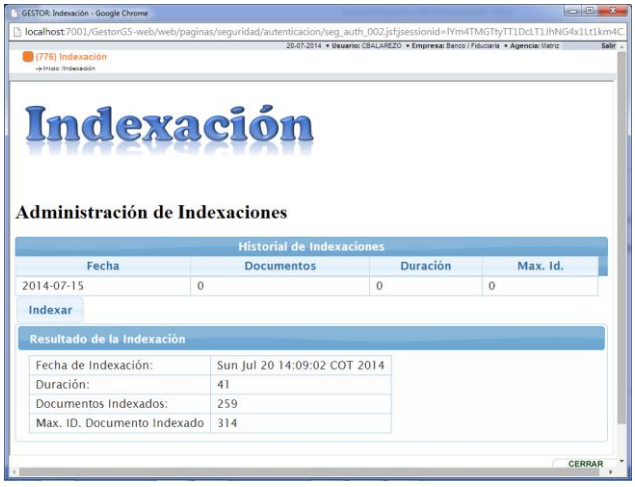

Figura 3. 4 Administración de Indexaciones

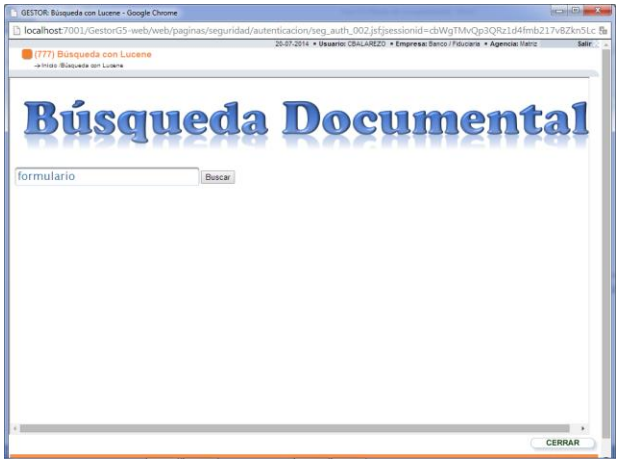

Figura 3. 5 Página de búsqueda documental con Lucene

Figura 3. 6 Resultados de la búsqueda

4 CONCLUSIONES Y TRABAJOS FUTUROS

Terminado el desarrollo de la aplicación Gestor Search utilizando las librerías Apache Lucene y Tika y su integración con el sistema Gestor se ejecutaron pruebas en las que se obtuvo buenos resultados en tiempos de búsqueda, menores a 2 segundos utilizando un conjunto de 500 archivos los cuales aproximadamente representan 10GB de espacio en disco.

Para la empresa GestorInc. S.A. fue una gran ayuda la implementación de esta aplicación ya que se consiguió facilitar el proceso de búsqueda que los clientes realizan sobre los documentos almacenados en el repositorio del sistema Gestor G5 Trust.

La ventaja de la implementación de la aplicación Gestor Search es su compatibilidad con el sistema Gestor tanto en requisitos de hardware y software, por esta razón los clientes no se ven en la obligación de incurrir en costos relacionados a la adquisición de una infraestructura.

Concluida la implementación de la aplicación de búsqueda empresarial Gestor Search y una vez que se han realizado de forma satisfactorias las pruebas funcionales de la misma, la empresa Gestor prevé iniciar un proyecto formal para la comercialización de la misma a sus principales clientes.

5 AGRADECIMIENTOS

A la Empresa Gestor, a la Unidad Estratégica de Desarrollo y a la Unidad Corporativa de Sistemas por brindar toda la ayuda requerida en la infraestructura tecnológica para el desarrollo de este proyecto.

A los Ingenieros Henry Coral y Santiago Salvador por ayudarme con sus conocimientos y guía en todo el proceso de desarrollo de este proyecto y a mi familia por su apoyo incondicional para cumplir una meta más en mi vida profesional.

6 REFERENCIAS BIBLIOGRÁFICAS

- Carpenter, B. (2011). *Lucene version 3.0 tutorial.* LingPipe, Inc.
- Chris A. Mattmann, J. L. (2012). *Tika in Action.* Shelter Island, NY 11964: Manning Publications Co.
- Erik Hatcher, O. G. (2004). *Lucene in Action.* Canadá: Manning Publications.
- Foundation, T. A. (2010). *Introducción para Apache Lucene*. Obtenido de http://www.javacodegeeks.com/2010/05/introduction-to-apache-lucene-for-full.html
- Foundation, T. A. (2011-2012). *Apache Lucene Core*. Obtenido de http://lucene.apache.org/core/
- Nayar, L. (2010). *La gestión documental. Conceptos básicos.* Buenos Aires: Mariana Sabuguerio.
- Sonawane, A. (2009). *Using Apache Lucene to search text*. Obtenido de http://www.ibm.com/developerworks/library/os-apache-lucenesearch/