Mejora de la Recuperación de Información en entorno Oracle: Aplicación práctica a Recursos Cartográficos

Víctor Dart [1], Juan Carlos Martínez [1], José V. Ballester [2], Francisco Rangel[1]

[1] Corex Soluciones Informáticas, Valencia, España
{victor.dart,juancarlos.martinez, francisco.rangel}@corex.es
[2] Dpto. Sistemas Informáticos y Computación, Universitat Politècnica de València, España
jvballester@dsic.upv.es

Abstract. Actualmente es bien conocido que uno de los elementos principales de un portal es el buscador, el cual debe permitir a los visitantes acceder a la información lo más rápido y con la mayor precisión posible. El objetivo de este trabajo es incrementar la eficiencia y precisión de la recuperación de información en una Web que contiene recursos en diferentes idiomas y que están almacenados en una BD Oracle. Se describe la creación e integración de un motor de búsqueda *multilingüe* basado en índices de terminologías. Para ello se ha utilizado la tecnología Lucene para la recuperación de información, usando sus métodos de indización y búsqueda de texto, refinados con la integración de *stopwords* y *stemming*, la creación de analizadores propios para cada idioma, y se ha incluido la posibilidad de realizar búsquedas simples y avanzadas, tanto textuales como numéricas.

Keywords. Recuperación de Información, Clasificación, Oracle, Lucene, Stemming

1 Introducción y motivación

En la actualidad, hay Webs que tienen un buscador de información muy básico, basado en realizar una comprobación simple del texto completo solicitado con la información almacenada, habitualmente en Bases de Datos (BD), obteniendo resultados con poca precisión y con unos tiempos de respuesta elevados. Este proceso se hace aún más complejo cuando la información está almacenada en distintos idiomas.

Nuestro proyecto presenta la creación de dos sistemas de indexación y búsqueda sobre una base de datos Oracle, utilizando dos tecnologías distintas de manera que se pueda evaluar cual de ellas será la más adecuada para el sistema Web de recursos cartográficos. La solución propuesta, además, pretende ser altamente extensible de modo que en un futuro sea posible añadir cualquier funcionalidad adicional.

Tomaremos como punto de partida una Web que tiene información de distintos recursos, en varios idiomas y que tiene la información almacenada en tablas de texto de Oracle.

Adicionalmente, a la solución óptima le impondremos una serie de restricciones técnicas (compatible con Java 1.6 o superior y Oracle 10g), económicas (menor coste de implantación) y de servicio (el tiempo de parada del buscador sea el menor).

2 Elección de los sistemas

Según los requisitos expuestos, se propusieron los siguientes sistemas como posibles soluciones a implantar: Oracle Text y Lucene.

Oracle Text, anteriormente conocido como *ConTextOracle*, esta basado en los estándares de SQL para realizar las funciones de indexación, búsqueda y análisis de datos de texto almacenados en una base de datos Oracle, pudiendo realizar dichas funciones incluyendo estrategias como operadores booleanos, coincidencia de patrones, mezclado de consultas temáticas, palabras claves...entre otros.

La segunda opción planteada es la librería de búsqueda de texto completo de referencia en la categoría de software libre, Lucene, una *API* de código abierto para la recuperación de información, enfocada directamente en la indexación y búsqueda de texto completo. Este sistema fue originalmente implementada en *JAVA* , aunque cuenta con versiones para otros lenguajes como *Delphi, Perl, C#, C++, Python, Ruby y PHP*.

Dado que la base de datos se encuentra en Oracle, la opción más adecuada parece ser Oracle Text, pero este sistema presenta algunas desventajas frente a Lucene como veremos en la sección de experimentación.

3 Experimentación de las propuestas

A continuación describiremos los detalles de cada una de las implementaciones, divididas en tres secciones comparativas: características generales, indexación y búsqueda.

3.1 Características generales

Una de los puntos interesantes a medir, es la flexibilidad, la capacidad de migrar la implementación en diferentes plataformas. Lucene, al ser una capa independiente de la base de datos y estar enfocada a nivel de aplicación, es más adaptable a cambios, por tanto, migrar la base de datos a otro servidor constituiría una tarea sencilla.

En cuanto a la adaptabilidad, Oracle Text esta sujeto a la política comercial del fabricante, por el contrario, Lucene esta respaldada por una comunidad de software libre muy activa evolucionando e incrementando las funcionalidades continuamente.

La solución de Oracle tiene un coste en licencias, el cual es inexistente en Lucene, por lo que este parámetro está claramente decantado hacia la segunda alternativa.

3.2 Indexación

En este apartado veremos un estudio sobre la indexación de información entre los dos sistemas. A modo de comparación, en la Fig. 1 se presentan dos gráficas, cada una correspondiente a un sistema, donde el eje de las *y* muestra el tamaño de los ficheros en *MB*, mientras que el eje de las *x* hace referencia al tiempo de ejecución en minutos.

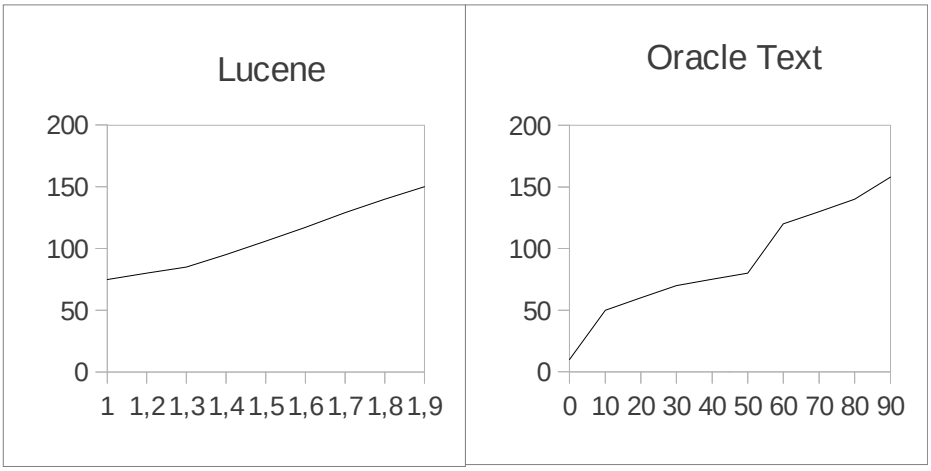

Fig. 1. Comparación de indexación de datos

A simple vista podemos observar varios puntos. El tiempo para la indexación de un fichero es siempre superior con Oracle que con Lucene, tanto es así que, un fichero casi tarda cuatro veces más con un sistema que con el otro. También se puede ver, que mientras Lucene sigue una evolución constante en el tiempo, Oracle presenta una irregularidad notable durante el proceso.

3.3 Búsqueda

Se realizaron diez experimentos sobre las búsquedas, evaluando los tiempos de ejecución y la cantidad de datos obtenidos, así como la precisión en estos.

Para ellos se ejecutaron las siguientes búsquedas en los dos sistemas: *límits administratius, hidrografía, xarxa de carreteres, incendios forestales, contaminación atmosférica, boscos públics, fishing zone, erosió litoral* y *lake.*

En la siguiente gráfica se muestra tres barras, las dos primeras hacen referencia a la cantidad de datos obtenidos por cada uno de los sistemas, y la última a la precisión de los resultados obtenidos.

Como se observa en la gráfica de la Fig. 2, Lucene es capaz de devolver más resultados que Oracle en la mayoría de las búsquedas realizadas, y con una precisión bastante buena.

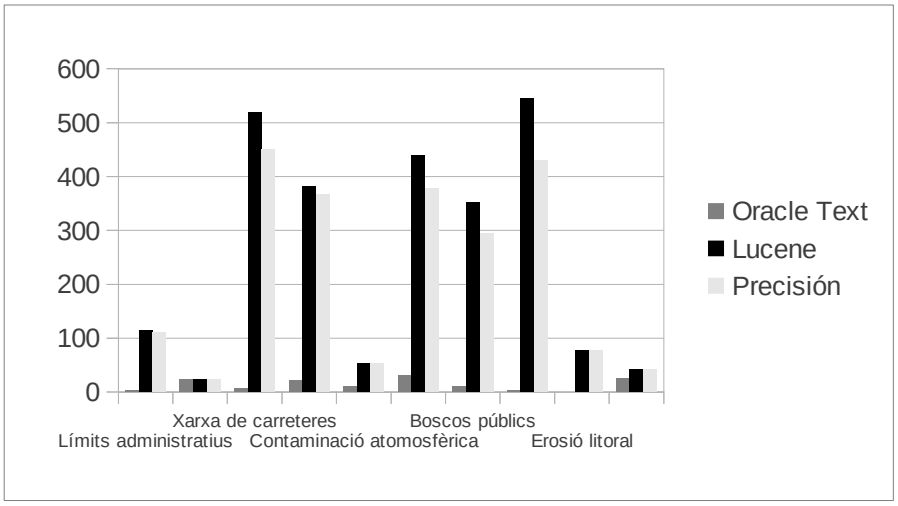

Fig. 2. Comparación de búsqueda de datos

En la Fig. 3. se muestra el gráfico de tiempos de ejecución de las búsquedas.

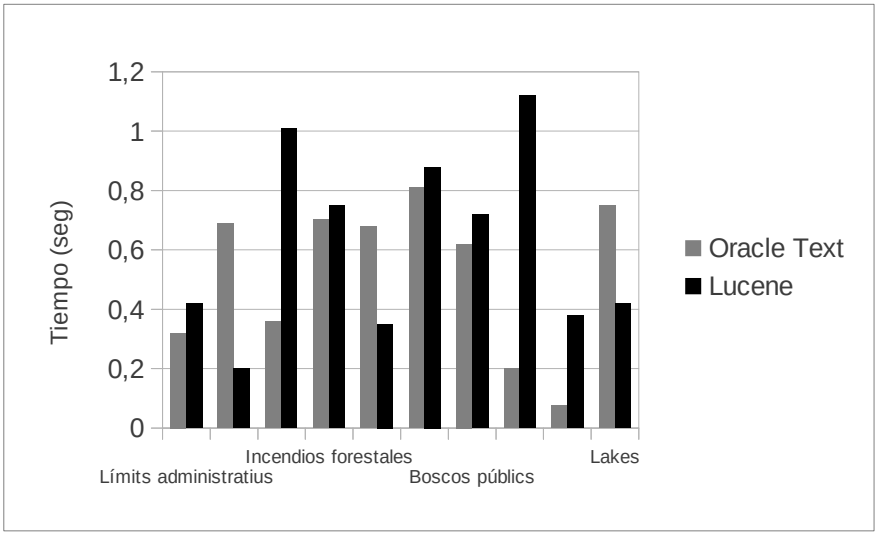

Fig. 3. Comparación de tiempos de ejecución en la búsqueda de datos

Como podemos ver, Oracle Text suele ser más rápido que Lucene en la obtención de documentos, pero si comparamos el tiempo y la cantidad de documentos extraídos, Oracle Text queda drásticamente penalizado.

4 Implantación

Una vez seleccionada la mejor alternativa, se estudiaron las tareas a realizar para conseguir el objetivo de mejorar la eficiencia y precisión del buscador.

Como conclusión a los estudios y trabajos realizados, determinamos que el objetivo se conseguía aplicando la siguiente secuencia de actividades:

(a) Tratar independientemente cada idioma en que está la información, por tanto, se realiza una indexación con Lucene de cada idioma por separado.

(b) La información asociada a cada uno de los recursos se indexará según la siguiente estructura:

  (1) Con el dato original.

  (2) Aplicando *stemming* al dato, utilizando los *stemmers Snowball.*

(c) Se desdobla el proceso de búsqueda en varias actividades para incrementar el número de recursos localizados y que están asociados a los intereses del usuario:

  (1) Para la búsqueda simple o libre, el usuario introduce las palabras a buscar, y el buscador del sistema valorará más el orden de aparición de estas palabras, así como la distancia entre ellas en los documentos recuperados.

  (2) Para la búsqueda avanzada se han incorporado las funcionalidades de operadores lógicos y restrictivos, como son *“”,OR, AND, +, -, NOT, ()*. Este modo de búsqueda es activado automáticamente siempre que se utilice alguno de estos operadores.

  (3) Para todas las búsquedas se realizar un filtrado de palabras vacías o *stopwords*, para eliminar información no relevante. Dicho filtro, es aplicado también en la fase de indexación.

(d) Para evitar la “parada” del buscador en la Web de recursos, el proceso de indexación se realiza *off-line*, y una vez finalizada la tarea se reemplaza el fichero de índices activo por el nuevo, proceso casi inmediato y sin repercusión para los posibles usuarios de la Web en el instante de la actualización.

5 Caso experimental

Los trabajos y estudios descritos en las secciones previas se han trasladado a un caso experimental para obtener resultados objetivos del incremento logrado. En concreto se ha implantado en la Web que contiene el Inventario de la Infraestructura de Datos Espaciales de Cataluña (IDEC) del Instituto Cartográfico de Cataluña.

En el estudio de búsquedas realizadas sobre diferentes términos se ha evidenciado claramente el incremento conseguido, estableciendo como medida final que se ha mejorado 42 veces el ratio de descubrimiento y recuperación de recursos que cumplen satisfactoriamente las condiciones establecidas por los términos buscados, lo que ha permitido pasar de 60 recursos recuperados a un total de 2.506 recursos con las acciones propuestas.

6 Trabajos futuros

Se pueden incrementar las funcionalidades descritas con anterioridad incluyendo diferentes tratamientos adicionales, entre los que se encuentran:

- Tratamiento de sinónimos, antónimo, homónimos, hiperónimos, hipónimos.
- Soporte *translingüe* de topónimos, para que el idioma en el que se escriben los topónimos sea independiente del idioma de búsqueda seleccionado.
- *Syntax Highlighting*.

7 Conclusión

Tras el análisis de los dos sistemas planteados, deja a la luz que el motor de búsqueda Lucene es mucho más apropiado para realizar búsquedas de texto completo a un conjunto de datos, ya que este está especializado en bases de datos textuales y por tanto ofrece mejores posibilidades para el tratamiento de estos.

Utilizando Lucene y siguiendo los pasos descritos en el apartado anterior, se ha logrado una mejora sustancial de la recuperación de información *multi-idioma* en un entorno Oracle, incrementando tanto la eficiencia como la precisión.

8 Agradecimientos

La realización del presente trabajo ha sido financiado por la empresa COREX Soluciones Informáticas de Valencia. (www.corex.es)

Referencias

1. Eric Hatcher y Otis Gospodnetic, "Lucene in Action", Manning Publications co., 2004, ISBN: 1932394281
2. Lucene wiki: http://wiki.apache.org/lucene-java/PoweredBy/
3. Introduction to Oracle Text: http://docs.oracle.com/
4. Infraestructura de Dades Especials de Catalunya: http://www.geoportalidec.cat/geoportal/cas/