Sistema de Recuperación de Información

Motor de Búsqueda: Innuendo

Epifanio Tula, Luis Gerónimo

Medeot, Matías Daniel

Universidad Tecnológica Nacional, Facultad Regional Córdoba

Abstract

El presente trabajo tiene por objetivo implementar un sistema de recuperación de información documental, comúnmente denominado "motor de búsqueda", teniendo por objetivo la investigación de las técnicas de recuperación de información para luego volcar los conocimientos adquiridos en un sistema propio. Por lo tanto, logramos desarrollar un sistema flexible, escalable y que corresponde a parámetros de calidad similares a otros productos de la comunidad open-source donde se conjugan tanto las técnicas de optimización en back-end como últimas tecnologías en front-end con tiempos de respuestas aceptables.

Palabras Clave

Recuperación de información documental, motor de búsqueda, indexación, posteo, skiplist, AJAX, Java 6, portable, multiplataforma, multiformato, html, pdf, xml, doc. xls, ppt, txt, zip, índices invertidos

Introducción

Hoy en día, debido a la integración de nuestra vida cotidiana a los sistemas informáticos, la necesidad de almacenamiento y recuperación de información es algo tan vital e indispensable para poder desarrollar nuestras actividades. Para responder a las necesidades de información de los usuarios surgen los sistemas de recuperación de información o, comúnmente denominados, motores de búsquedas.

Estos motivos nos han conducido a la construcción de un pequeño pero robusto motor de búsqueda, denominado "Innuendo", que pudiera servir para diversos fines. Los objetivos básicos a la hora de diseñarlo han sido:

a) Investigación del campo de recuperación de información

b) Disponer de un programa que implementara algunos de los modelos teóricos más difundidos en recuperación de la información

c) Disponer de una implementación que sirva de referencia para estudio de sistema de recuperación de información y al mismo tiempo refleje las condiciones de uso en un ambiente cuasi-realista

La recuperación de la información es la disciplina encargada de la representación, almacenamiento, y la organización de la información, y su posterior acceso y recuperación de información que responda a las necesidades de un usuario.

Antes de profundizar en el tema, primero tenemos que diferenciarla de la recuperación de datos.

En la recuperación de datos:

- Los datos se pueden estructurar en tablas, árboles, listas, etc. para recuperar posteriormente exactamente lo que se quiere.
- Se conoce exactamente lo que se está buscando.
- Se evalúa principalmente la eficiencia (velocidad y espacio).

En la recuperación de información:

- El texto no tiene estructura clara y no es fácil de analizar.
- En RI no existe "la respuesta correcta". Cada documento puede ser más o menos relevante, y esto varía según el usuario y la situación.
- En RI, la velocidad y espacio importa, pero más la calidad de la búsqueda.
- La RI busca una aproximación a la respuesta sobre lo que el usuario busca.

Por lo tanto, el problema al que se enfrenta la RI se puede definir como: "Dada una necesidad de información (consulta) y un conjunto de documentos; ordenar los documentos de más a menos relevancia para esta necesidad, y presentar un subconjunto de los más relevantes.

Este problema se puede dividir en dos subproblemas:

- Elegir un modelo para calcular la relevancia de un documento frente a una consulta.
- Diseñar algoritmos y estructuras de datos que lo implementen eficientemente (índices).

Para resolver estos 2 problemas, debemos entender cómo es la arquitectura de un sistema recuperación de información. Básicamente el proceso de RI se divide en dos etapas: la indexación y la búsqueda.

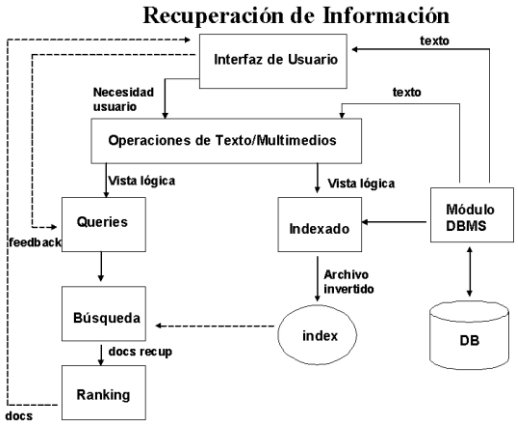

Figura 1: Arquitectura de un sistema de información

La etapa indexación se divide en las siguientes subetapas:

a) Crawling: búsqueda los diferentes documentos a indexar ya sea a nivel de sistemas archivos local, intranet o internet.

b) Parsing de documentos: modelización de los documentos a ser indexados.

c) Análisis: La forma más común de representar un documento de texto es por un sistema de términos indexados o palabras claves. Estos términos son extraídos con el siguiente proceso:

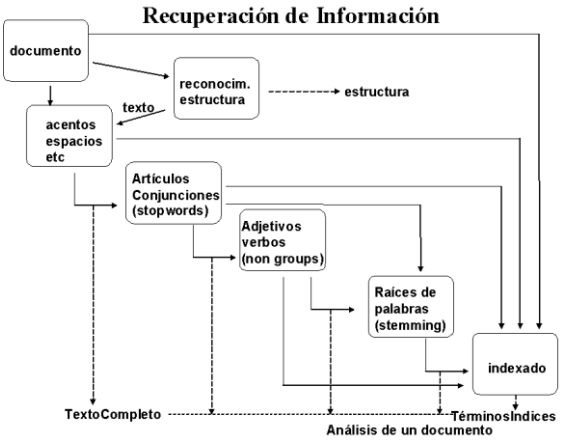

Figura 2: Proceso de Análisis

En esta etapa se eliminan y normalizan los términos del documento a ser indexado, se descartan las palabras que constituyen "stopwords" (ej. la, las, y, los, en, etc.), se eliminan acentos, espacios, se reducen palabras a su raíz gramatical (stemming), etc.
Todo esto es necesario para evitar redundancia en el índice y logar optimización del mismo.

d) Indexado: construye un fichero invertido (o índice invertido) de

palabras, con punteros a los documentos en que dicha palabra tiene una o más ocurrencias.

En la etapa de búsqueda se llevan a cabo los siguientes pasos genéricos:

a) Parsing de consulta: transforma la consulta para mejorar la búsqueda.
b) Búsqueda: recupera los documentos que contienen una determinada palabra o palabras claves del índice.
c) Ordenación o Ranking: ordena todos los documentos recuperados de acuerdo a una métrica de relevancia (modelo de búsqueda)
d) Presentación de los resultados a los usuarios.

Otros componentes a tener en cuenta en el desarrollo de un motor de búsqueda, son:

Interfaz de usuario: gestiona toda la interacción con el usuario.

- Entrada de las consultas
- Salida de documentos.
- Retroalimentación de relevancia.
- Visualización de los resultados.

Elementos y metodología de trabajo

Introducidos en esta problemática, intentaremos entender los modelos de búsqueda que intentan dar una respuesta a la recuperación de información y particularmente utilizamos como elemento y metodología de trabajo.

Un modelo de RI es la especificación sobre como representar documentos y consultas y cómo comparar unos y otros. El objetivo de todo modelo es obtener un orden (ranking) de los documentos recuperados que refleje la relevancia de estos con la consulta del usuario.

La clasificación de los modelos se muestra a continuación en la figura 2, nosotros nos centramos en el modelo más utilizado en esta disciplina, el modelo vectorial generalizado.

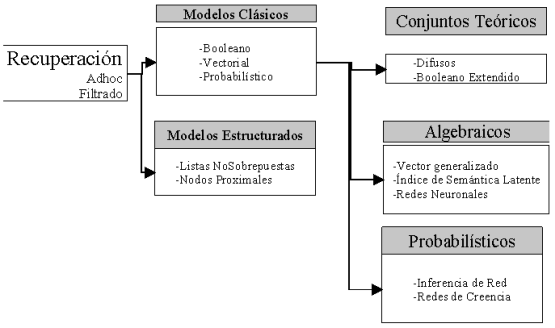

Figura 3: Clasificación de los modelos teóricos de búsqueda

Características del modelo Vectorial

Brevemente, según este modelo, cada documento es representado mediante un vector de n elementos, siendo n igual al número de términos indizables que existen en la colección documental. Hay, pues, 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. Si la palabra no está presente en el documento, ese valor es igual a 0. En caso contrario, ese valor es calculado teniendo en cuenta diversos factores, dado que una palabra dada puede ser más o menos significativa (tanto en general como, sobre todo, en ese documento en concreto); este valor se conoce con el nombre de peso del término en el documento.

Siempre según el modelo del espacio vectorial, las consultas son representadas también mediante un vector de las mismas características que las de los documentos (variando los valores numéricos de cada elemento en función de las palabras que forman parte de la consulta, claro está). Esto permite calcular fácilmente una función de similaridad dada entre el vector de una consulta y los de cada uno de los

documentos. El resultado de dicho cálculo mide la semejanza entre la consulta y cada uno de los documentos, de manera que, aquéllos que, en teoría, se ajustan más a la consulta formulada, producen un índice más alto de similaridad. Naturalmente, se asume que la consulta se formula en lenguaje natural (podría ser, incluso un documento de muestra, para recuperar los que fuesen parecidos a él), y, de lo dicho se deduce que el resultado de la consulta consiste en una lista de documentos ordenada en orden decreciente en función de su similaridad con la consulta. Diversos factores son modificables dentro de este modelo de recuperación. Así, entre otros, el sistema de cálculo de pesos. Los esquemas habituales se basan, de una u otra forma, en las frecuencias de aparición de la palabra cuyo peso se quiere calcular. Así, muchos sistemas utilizan algún mecanismo basado en la multiplicación del IDF del término por su frecuencia en el documento en cuestión. El propio IDF (Inverse Document Frequency) tiene varias versiones, por lo general basadas en una función inversa del número de documentos en que aparezca el término en cuestión.
En la figura podemos observar la representación gráfica de la similaridad de dos documentos.

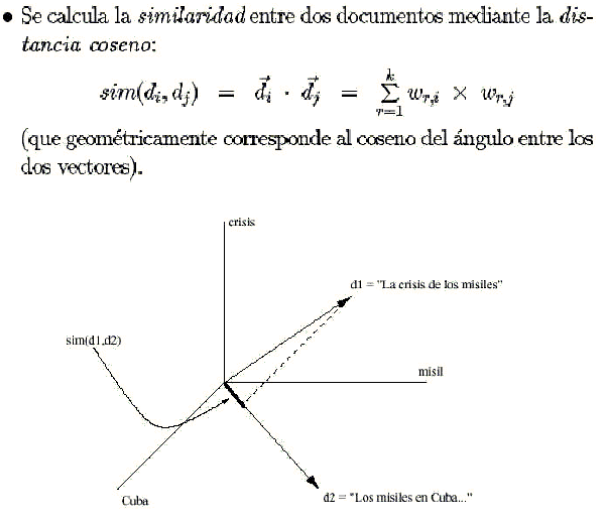

Figura 4: Representación gráfica de la similaridad de dos documentos

Resultados
Logramos desarrollar un motor de búsqueda con las siguientes características:

- Índice basado en un sistema de archivos con acceso aleatorio, en el mismo se utilizan técnicas para ahorrar espacio de almacenamiento, como compresión por sufijos, Int y Long variables, compresión por d-gaps, también se utilizan técnicas para optimizar el proceso de tiempo de búsqueda como skiplist en disco, etc.
- Indexación de los tipos de documentos más conocidos como xml, pdf, doc, xls, html, ppt, etc.
- Utilización del modelo vectorial con fórmula de similaridad mejorada
- Modulo Analizador: cuya función es normalizar los términos a ser indizados.
- Interfaz de usuario intuitiva y óptima debido a la disminución de consumo de ancho de banda por uso de tecnología AJAX.
- Basado en tecnología open-source.
- Consultas de búsquedas flexibles: capacidad de utilización de operadores booleanos (AND, NOT, etc.)
- Diseño flexible y modular permitiendo la extensibilidad del mismo
- Parámetros de medición (precisión, eficiencia y tiempo de búsqueda) acordes con lo de productos similares
- Su uso está orientado a la indexación de documentación de pequeñas y medianas empresas donde el tiempo de respuesta es aceptable para el cliente.
- Diseño totalmente orientado a objetos y desarrollado bajo estándares de UML

En cuanto a la evaluación del rendimiento de la recuperación, es necesario tener en cuenta los siguientes parámetros:

- Evaluar la calidad de los documentos recuperados en una búsqueda.

  En las sucesivas pruebas podemos observar que los documentos

recuperados responden a las necesidades de información requeridas por el usuario en términos de precisión de búsqueda y calidad de la misma.

- Evaluación del rendimiento del sistema (tiempo de respuesta y espacio de almacenamiento):
  En el proceso de indexación tiene una velocidad de 30 MB/min y el tamaño del mismo es de alrededor de 20-25% del tamaño del texto indexado, siendo bastante eficiente ya que utiliza una técnica de construcción de sub-índices. El proceso. El tiempo de respuesta a la búsqueda corresponde a parámetros análogos a los tiempos de respuesta de productos similares.

Discusión

Dado el carácter y objetivos planteados al diseñar "Innuendo", es evidente que no se trata de un proyecto cerrado. Antes bien, la sencillez y flexibilidad de que se le ha dotado, permite la adición de nuevas prestaciones. De algún modo podemos decir que se diseñó así precisamente para eso.

Entre las acciones futuras previstas se encuentra la implementación de un algoritmo de lematización específico para el idioma castellano. También, la evaluación y estudio de otros modelos de búsqueda como "índice de semántica latente" o la inclusión de redes neuronales para la mejora de la indización y el proceso de búsqueda resultante.

También nos planteamos la necesidad de su extensión a la búsqueda en el entorno Web que requiere la investigación de técnicas de búsqueda en redes de documentos (documentos inter-referenciados) como es el caso de los sitios Web, ej. clúster de documentos, algoritmos de pageRank, crawling, etc.

Conclusión

Logramos cumplir con los objetivos inicialmente planteados, aplicando los conocimientos adquiridos en el apasionante campo de la recuperación de la información. Se ha diseñado un sistema de recuperación documental que aplica el modelo del espacio vectorial, lo suficientemente abierto y flexible para ser utilizado en labores docentes, de investigación, y así mismo orientado y adaptable a pequeñas y medianas empresas. La sencillez de su arquitectura permite tanto la fácil observación de resultados y estructuras intermedias como la modificación y añadido de nuevos módulos y, por consiguiente, la experimentación.

Agradecimientos

Fundamentalmente a Valerio Fritelli, quien con su entusiasmo en la docencia motivo en nosotros el anhelo en la obtención de nuevos objetivos que trasciende lo meramente académico

Referencias

[1] "Diseño de un motor de recuperación de la información para uso experimental y educativo" de Carlos G. Figuerola, José Luis Alonso Berrocal y Ángel Francisco Zazo Rodríguez.

[2] Rijsbergen, C. J. van. (1979). Information retrieval. London: Butterworths.

[3] Lucene in Action Manning

[4]" Self-indexing inverted files for fast text retrieval", ACM Transactions on Information Systems (TOIS). Alistair Moffat Univ. of Melbourne, Victoria, Australia Justin Zobel RMIT, Victoria, Australia

[5] "Document Ranking and the Vector-Space Model", DIK L. LEE, HongKongUniversity of Science and Technology

Datos de Contacto:

Epifanio Tula, Luis Gerónimo
Universidad Tecnológica Nacional Facultad Regional Córdoba
Carvajal y Saravia 4178
Tel. 0351-4565100
luisepifanio@yahoo.com.ar, info@faelsoft.com.ar

Medeot, Matías Daniel
Universidad Tecnológica Nacional Facultad Regional Córdoba
Tel 0358-154185191
matiasmedeot@yahoo.com.ar,
desarrollomdm@yahoo.com.ar