Swish-e

Alternativa de solución a la administración de contenido digital.

Rodrigo Bello
rodrigobello2001@yahoo.es

Jhonattan Prieto
jhonapri@yahoo.es

RESUMEN

En las unidades de información se presenta el reto de minimizar la brecha digital que ha supeditado a los países en vía de desarrollo a través de la recuperación de contenidos digitales en la Web, para ello se hace necesaria la vinculación de herramientas tecnológicas que faciliten la administración y la recuperación de estos contenidos de una forma accesible a todas las comunidades. Por esta razón es ineludible la disposición de software libre que permita la realización de dichas labores enfocadas a la preservación y disposición del conocimiento en una civilización. Se presenta entonces, como opción a esta problemática el sistema de indexación Swish-E como alternativa de búsqueda y recuperación de contenidos completos, se hace una aproximación a la experiencia realizada por el Boletín oficial del Estado de España con el software Simple Web Indexing System for Humans – Enhanced, se muestran sus características más relevantes, ventajas, desventajas y se pretende motivar su uso como herramienta clave en la administración de contenidos digitales.

PALABRAS CLAVES

Swish-e, Boletín Oficial del Estado (BOE), Recuperación de información, Software libre, indexación.

ABSTRACT

The intelligence units is the challenge of minimizing the digital divide that has been subject to the developing countries through the recovery of digital content on the Web, it is necessary for linking technological tools that facilitate the administration and recovery of these contents in a manner accessible to all communities. Therefore it is imperative the provision of free software that permits the realization of these tasks focused on the preservation and disposal of knowledge in a civilization. It comes then, as an option to this problem the indexing system Swish-E as an alternative search and retrieval of full content, it is an approximation to the experiment conducted by the Official Gazette of the State of Spain with the software Simple Web Indexing System for - Enhanced Humans, showing its most important characteristics, advantages, disadvantages and is intended to motivate its use as a key tool in the management of digital content.

KEYWORKS

Swish-e, Official State Bulletin (BOE), Information retrieval, Free, indexing.

1. Experiencia.

Con la *necesidad de ofrecer servicios de información a los usuarios el gobierno español adopto el Boletín Oficial del Estado (BOE), concebido como el diario oficial del Estado Español, es, decir, el órgano de publicación de las leyes, disposiciones y actos de inserción obligatoria.* Contiene además las leyes producidas en el seno de las Cortes Generales, las disposiciones emanadas del Gobierno de la Nación y las disposiciones generales de las Comunidades Autónomas.

Además del BOE, existen también Boletines Oficiales del resto de administraciones territoriales (de cada Comunidad autónoma y de cada Provincia), al mismo tiempo que otros Boletines, como el de las Comunidades Europeas y de las Cámaras Parlamentarias.[1]

Por la importancia de este boletín y la ferviente consulta por parte de los usuarios, el estado español vio necesaria la comunicación oficial entre el pueblo y el estado on-line dando origen a la utilización del recurso Internet, el cual permitió dicha conexión con los usuarios quienes por ley merecían conocer los adelantos legislativos que se iban presentando.

---

[1] *http://www.boe.es/*. [En línea]. Consultado el 25 de octubre de 2007.

En el año 1999 el boletín oficial disponía de un aplicativo para la búsqueda e indexación de contenidos que utilizaban los usuarios, desafortunadamente, después de realizar varias pruebas esta solución mostró sus desventajas al intentar satisfacer la cantidad simultanea de usuarios reales viéndose las siguientes debilidades:

- Recuperación ineficiente en índices.
- Lenguaje de generación de frontends lento en la ejecución.
- Dificultad en la ordenación de registros por campos diferentes a los indexados.
- Exceso de uso de la memoria con relación a los procesos de búsqueda.
- Soporte insuficiente por parte del proveedor.

De esta forma los servidores colapsaron al intentar ejecutar dichas demandas de información. Estos problemas, unidos a otros detectados en los procesos de indexación, desaconsejaban la utilización de la herramienta para un previsible acceso de múltiples consultas simultáneas desde Internet.

2. Software libre como solución

Ante la situación presentada en el punto anterior se hacía necesario utilizar otro tipo de herramienta de indexación y búsqueda de contenidos aplicable a los contenidos a distribuir por el Boletín Oficial del Estado en Internet.

Dicha solución no precisaba complejos elementos de gestión de contenidos ni una funcionalidad similar a las muy costosas soluciones existentes en el mercado. Su principal requerimiento era la velocidad de respuesta en las búsquedas.

Ante la escasez de tiempo, se estudiaron las posibilidades que el software libre ofrecía. En aquel momento existían varios indexadores de contenidos susceptibles de ser utilizados (freeWAIS-sf, htdig, swish-e, etc.) pero ninguno de ellos cumplía plenamente los requerimientos buscados. Básicamente, el aplicativo debía cumplir unos mínimos imprescindibles:

- Alto rendimiento en búsqueda ante el volumen esperado de consultas.
- Alto rendimiento en indexación ya que los contenidos se modificaban a diario.
- Disponibilidad de librería de programación (API).
- Ordenación de resultados por diversos criterios (campos).
- Indexación de caracteres nacionales (ISO-8859) y soporte de tablas de conversiones.
- Utilización de filtros externos para poder indexar diferentes tipos de contenidos (texto, html, pdf, etc.).

Ninguna de las soluciones basadas en software libre cumplía todos los requisitos anteriores. Sin embargo, una de ellas, aunque prácticamente abandonada en su desarrollo por sus autores, era una implementación muy sencilla de un indexador, estaba bien documentada y su licencia, basada en GPL, permitía su modificación y adaptación. El paquete en cuestión era swish-e en su versión 1.3 y podía ser un buen punto de partida al implementar un índice invertido y los procesos básicos necesarios de indexación y búsqueda.[2]

3. Swish-e en su versión 2.0

A partir de la versión de swish-e 1.3 se procede a desarrollar y adaptar el aplicativo a los requerimientos básicos del BOE. Para ello se le añadió la funcionalidad necesaria de la que carecía:[3]

- Nueva gestión de índices. Se añade un índice hash al índice invertido para acelerar las búsquedas.

- Nuevo motor de indexación más rápido utilizando tablas hash en lugar de las originales listas enlazadas. De esta manera se podían indexar decenas de miles de documentos en un tiempo razonable.

- Almacenamiento en los índices de las posiciones de las palabras para permitir la búsqueda de frases.

- Inclusión de una librería C y de su correspondiente API. De esta manera se hace posible escribir CGIs de búsqueda en lenguaje C.

- El formato de Base de Datos nativo se hace "portable" entre plataformas hardware y software diferentes. Se puede indexar en un sistema operativo diferente al que se vaya a utilizar como explotación para las búsquedas.

Todo el desarrollo se realizó en entornos de software libre (sistema operativo linux, compilador y herramientas de desarrollo GNU gcc) y se comprobó su correcto funcionamiento en varios sistemas UNÍX, tanto comerciales (SUN solaris e IBM AIX), como libres (Linux, FreeBSD).

Básicamente el aplicativo se compone de 2 partes, un indexador y un buscador. El indexador se encarga de analizar los documentos y extraer toda la información necesaria que permita crear la base de datos de índices. Una vez

---

[2] Ibíd.

[3] Michael Chilli. Busqueda de escritorio en Perl, Ve a por el.[En linea]. Consultado el 15 de octubre de 2007.

obtenida esta, se puede localizar la información a través del buscador. Además del buscador se dispone de una librería abierta que permite a los usuarios crear sus propios desarrollos.

Esta estructura se mantiene prácticamente idéntica en la actualidad.

Una vez realizadas las correspondientes pruebas se pone el aplicativo en explotación en el BOE y se ofrece por Internet la Base de Datos IndiBOE (sumarios del BOE desde 1995 hasta la actualidad) que incluye cuatro índices:

- Sección I (Disposiciones Generales)
- Sección II (Autoridades y Personal)
- Sección III (Otras Disposiciones)
- Sección V (Anuncios)

Dado el origen del paquete, basado en licencia GPL, con el cual aún compartía gran parte del código, y acorde a sus condiciones, se procedió a ofrecer el nuevo paquete con la misma licencia, como versión 2.0. Durante el año 2000 se corrigieron algunos errores y se añadió alguna funcionalidad menor siendo la versión 2.0.5 la última de la rama 2.0.[4]

4. Nuevas versiones.

A partir de la versión 2.0 el desarrollo del paquete cobra más interés por parte de la comunidad de usuarios y se decide ampliar su funcionalidad. En ese momento se inicia un desarrollo nuevo desde diversos partes del mundo y sé comienza a ampliar notoriamente la funcionalidad del aplicativo. Para coordinar todos estos esfuerzos, se decide alojar el proyecto en sourceforge.net y, mediante la herramienta libre CVS, se procede a coordinar todas las fases del desarrollo (control de versiones, corrección de errores, incorporación de mejoras, etc). Así, se crea la versión de desarrollo swish-e 2.1 a la que progresivamente y, desde diversas fuentes, se le ha ido ampliando su funcionalidad.

5. Planes de futuro

Actualmente, a muy corto plazo, los planes incluyen la liberación de lo que será la versión 2.2 que incluye todas las funcionalidades anteriormente descritas.

Como planes de futuro se encuentran:

[4] Ibíd.

- Nuevo gestor de índices que permita una manera más flexible de añadir documentos dinámicamente a un índice ya creado. Ello conllevará un rediseño total del sistema de base de datos nativo de la aplicación.
- Creación de un Servidor de Base de Datos que permita, entre otras cosas, llevar una cache de las búsquedas.
- Soporte de otras Bases de Datos para mantener los índices, por ejemplo Berkeley DB.
- Gestión transaccional de las actualizaciones

6. Algunas cifras de rendimiento

Las siguientes cifras muestran el comportamiento del indexador y del buscador sobre un servidor Dell PowerEdge (Intel Pentium III, 2 GB RAM, 4 discos 72 GB SCSI) con RedHat Linux 7.2 como sistema operativo.

Prueba de indexación
La prueba de indexación se compone de 100000 documentos XML (codificación ISO-8859), y un volumen total de 1.56 GB. Tiempo de indexación: 21 minutos.[5]

Pruebas de búsqueda
Se realizan varias pruebas de búsqueda sobre el anterior índice en el mismo sistema.

En todos los casos, se muestran solamente los 20 primeros resultados (la ordenación se realiza sobre el número total de resultados).

Búsqueda	Núm. resultados	Criterio de ordenación	Tiempo ejecución	Tiempo busqueda
Búsqueda de una palabra muy común ("resolucion")	17928	Por relevancia	0.056 seg.	0.033 seg.
		Por fecha de publicación	0.066 seg.	0.043 seg.
		Por departamento	0.057 seg.	0.034 seg.
Búsqueda de una palabra poco común ("informática")	528	Por relevancia	0.024 seg.	0.001 seg.
		Por fecha de publicación	0.029 seg.	0.006 seg.
		Por departamento	0.028 seg.	0.005 seg.
Búsqueda de una frase ("ley 30/1984")	45	Por relevancia	0.034 seg.	0.010 seg.
		Por fecha de publicación	0.038 seg.	0.015 seg.
		Por departamento	0.037 seg.	0.014 seg.

7. Referencias: Algunos sitios donde se usa swish-e

[5] Michael Chilli. Busqueda de escritorio en Perl, Ve a por el.[En linea]. Consultado el 15 de octubre de 2007.

Al estar el código fuente disponible, swish-e se encuentra funcionando en la totalidad de los sistemas UNÍS (Linux, FreeBSD, Solaris, AIX, HPUX, etc), Mac OSX y plataformas windows, existiendo binarios para gran cantidad de dichos sistemas.

Adicionalmente, al haber sido construido de forma totalmente abierta, el tipo de contenidos que se están indexando varia desde los clásicos PDF o HTML hasta MP3. Como ya se ha mencionado, el Boletín Oficial del Estado utiliza extensamente dicho buscador. Las Bases de Datos que ofrece con este sistema son:

- IndiBOE: Ya mencionada, incluye 4 índices. Cada índice incluye más de 100000 documentos en formato XML. Lenguaje de programación de búsqueda utilizado en el web: perl

- Web: Documentos fuente HTML. Lenguaje de programación de búsqueda utilizado en el web: perl
- Tienda del BOE: Documentos fuente en Base de Datos Relacional. Lenguaje de programación de búsqueda utilizado en el web: PHP
- Dictámenes del Consejo de Estado: Documentos fuente en Base de Datos Relacional. Lenguaje de programación de búsqueda utilizado en el web: perl
- Jurisprudencia del Tribunal Constitucional: Documentos fuente en formato XML. Lenguaje de programación de búsqueda utilizado en el web: perl[6]

A continuación se incluyen otros lugares significativos que utilizan swish-e:

- Apache Web Server site (http://apache.org)

Berkeley Digital Library Sunsite (http://sunsite.berkeley.edu/cgi-bin/search.pl)

- Librarians Index to Internet (http://lii.org)
- http://swish-e.org/sites.html

---

6 Michael Chilli. Busqueda de escritorio en Perl, Ve a por el.[En linea]. Consultado el 15 de octubre de 2007.

Bibliografía.

Michael Chilli. Búsqueda de escritorio en Perl, Ve a por él.[En línea]. Consultado el 15 de octubre de 2007.

http://swish-e.org/.[En línea]. Consultado el 25 de octubre de 2007.

http://sunsite.berkeley.edu/SWISH_E. [En línea]. Consultado el 25 de octubre de 2007.

www.senado.es/boletines/D0296.html. [En línea]. Consultado el 25 de octubre de 2007.

http://www.boe.es/. [En línea]. Consultado el 25 de octubre de 2007.