Buscador de la Transparencia

DCC, Universidad de Chile Centro de Investigación de la Web

Senén González
Estudiante de Doctorado en Ciencias de la Computación, DCC, Universidad de Chile. Egresado de Ingeniería Civil en Computación, de la misma Universidad.
sgonzale@dcc.uchile.cl

Mauricio Marín
Investigador del laboratorio Yahoo! Research Latin America, Universidad de Chile. Investigador asociado del Centro de Investigación de la Web. PhD en Computer Science, University of Oxford, UK.
mmarin@yahoo-inc.com

Víctor Sepúlveda
Estudiante de Magíster en Ciencias mención Computación, DCC, Universidad de Chile, bajo la supervisión del profesor Benjamín Bustos. Egresado de Ingeniería Civil en Computación de la misma universidad.
vsepulve@dcc.uchile.cl

1. INTRODUCCIÓN

Durante este año 2009 un equipo de estudiantes y profesores del DCC ha estado participando en el desarrollo del llamado "Buscador de la Transparencia" (http://buscador.chileclic.gob.cl/). Se trata de una herramienta destinada a apoyar la implementación de la nueva Ley de Transparencia Activa promulgada recientemente por el Gobierno de Chile. El objetivo es permitir a cualquier persona buscar información publicada por los distintos sitios Web del Estado y municipalidades.

Los ministerios, municipalidades y otras reparticiones del Estado están obligadas a publicar periódicamente información pertinente a la Ley en forma oportuna. El Buscador proporciona herramientas que permiten a los administradores de los respectivos sitios Web publicar dicha información en un formato estándar, lo cual no sólo abarca el llenado de información vía formularios por parte de los encargados de los sitios Web, sino también la presentación de resultados desde consultas automáticas a sistemas de bases de datos. Dicha estandarización también hace posible supervisar a nivel central el cumplimiento de la normativa de publicación y la actualización de la información de Transparencia.

La primera versión del Buscador entró en operación en mayo de 2009. Desde

entonces ha tenido varias extensiones. Sus inicios no estuvieron exentos de críticas e incomprensión respecto de sus objetivos de diseño y ventajas. La comparación obvia fue con Google, el cual, si bien se puede restringir a dominios específicos, sigue siendo un buscador genérico que al final resulta ser menos efectivo por las siguientes razones:

El Buscador de la Transparencia está centrado en un contexto bien particular, con una problemática asociada a un crecimiento muy dinámico de los sitios Web del Estado. Por ejemplo, cuenta con un sistema diseñado para mantener e incorporar nuevas semillas desde donde hacer nuevas colectas de información. Dichas semillas se pueden agrupar en entidades lógicas que facilitan su administración e indexación. Respecto de las búsquedas puede ser visto como un sistema segmentado. Da la posibilidad de buscar sobre todos los sitios del Estado y municipalidades, considerando por separado tanto las secciones destinadas a publicar información de la Ley de Transparencia Activa como al resto de los contenidos de los distintos sitios Web. Es decir, los resultados de las búsquedas se presentan de manera separada para reducir las interferencias entre ambos tipos de contenidos, y existe separación entre sitios del Estado y municipalidades.

Asimismo, también por construcción, el Buscador permite focalizar las búsquedas solamente dentro de lo que se conoce como ChileClic, "La Guía de Servicios del Estado": un portal administrado por Estrategia Digital (ministerio de Economía) cuyo objetivo principal es entregar información destinada a describir trámites y proporcionar orientación sobre los distintos servicios que el Estado ofrece a chilenos y extranjeros.

Los estudiantes del DCC que han participado en todas las etapas del proyecto de construcción del Buscador de la Transparencia son: Senén González, Víctor Sepúlveda, Eduardo Graells y Mauricio Monsalve. También han participado los profesores Ricardo Baeza-Yates, Claudio Gutiérrez y Mauricio Marín.

A continuación se presenta una descripción general del diseño del Buscador.

2. DESCRIPCIÓN GENERAL

El Buscador de la Transparencia es un sistema que tiene tres componentes principales:

MyWeb: conjunto de herramientas para recolectar sitios, crear índices de búsqueda y realizar consultas sobre esos índices (utiliza tecnología del buscador TodoCL).

Panel de Control: sistema basado en Web, que permite gestionar las herramientas y datos de MyWeb. Realiza operaciones tales como la gestión de URLs que serán recolectados e indexados por el Buscador.

Gestor de Contenidos o CMS: parte pública del Buscador (sitio Web incorporado a ChileClic) que posee un panel de control para gestionar la información que éste publica y la interacción con otros sitios del gobierno.

2.1. MyWeb

MyWeb es considerablemente más rápido y liviano que otros sistemas de dominio público, lo que es una ventaja frente al alto tráfico de consultas. Este era el único componente existente al momento de iniciar la construcción del Buscador. Contiene cuatro componentes:

Recolector (o crawler): que recibe una dirección de inicio, usualmente la página principal de un sitio o dominio. Descarga todas las páginas que encuentra dentro del mismo dominio/sitio que tengan profundidad física mayor o igual a la de la dirección inicial.

- También puede recibir una blacklist de palabras para rechazar ciertas URL.
- Baja páginas HTML y archivos PDF, PPT, XLS, DOC y TXT.

Indexador: crea un índice a partir de una o más colectas realizadas por el crawler. Las colectas pueden ser de sitios distintos, de modo de mezclarlas todas dentro de un único índice. Todos los documentos bajados con éxito por el crawler son indexados.

Demonio de Consultas: componente que está continuamente escuchando en un puerto de la máquina en busca de consultas. Cuando recibe una, busca los documentos relevantes en el índice.

- Los documentos son ordenados por relevancia de texto de acuerdo a las palabras presentes en la consulta.
- Puede responder consultas diferenciando o ignorando acentos.
- AND: busca documentos que sean relevantes para algunas de las palabras de la consulta. Equivalente a buscar +palabra1 +palabra2 en otros buscadores. Los documentos seleccionados contienen todas las palabras de la consulta.
- OR: busca documentos con al menos una de las palabras de una consulta. Equivalente a la búsqueda por omisión en otros buscadores. Los documentos seleccionados pueden contener una o más palabras de la consulta.
- EXACT: busca la frase exacta de las palabras de la consulta. Equivalente a escribir la consulta entre comillas en otros buscadores: ''palabras de consulta''.
- Se pueden correr distintos demonios de manera simultánea, de modo que cada uno responda consultas utilizando distintos índices (los resultados pueden ser mezclados en el CMS).
- Se entregan sugerencias a consultas que tienen pocos o ningún resultado. Estas sugerencias provienen de nuevas consultas que producen una cantidad suficiente de resultados, las cuales son construidas de tal manera que la distancia de edición respecto de la consulta original es pequeña.
- Entrega un máximo de 200 resultados por consulta. Para obtener un mayor número, el usuario puede refinar su consulta incluyendo palabras más

específicas o seguir las sugerencias entregadas por el Buscador.

Procesador de Consultas: componente que recibe la consulta y se comunica con el demonio para obtener los resultados. Este componente se puede modificar para adaptarlo a las necesidades del cliente (el procesador, al igual que los siguientes dos componentes, fue desarrollado específicamente para el Buscador).

2.2. Panel de Control

El Panel de Control se utiliza para gestionar los índices y colectas de MyWeb. Permite crear varias instancias de MyWeb para focalizar el Buscador en distintos segmentos de la Web, tal como ocurre actualmente donde se recolectan e indexan por separado distintos tipos de sitios Web del gobierno (ministerios, municipalidades, etc.). Cuenta con la siguiente funcionalidad:

- Crear/editar/eliminar colectas.
- Especificar semillas para las colectas.
- Agrupar colectas en índices.
- Ver registro (logs) de funcionamiento del Buscador.

El panel es de uso interno. Sólo pueden acceder a él los administradores del sistema mediante una interfaz Web (utiliza PHP y MySQL para administrar la base de datos).

2.3. CMS

El gestor de contenidos tiene dos partes:

Frontend: es el sitio del Buscador, accesible desde http://buscador.chileclic.cl. Recibe las búsquedas ingresadas en el sitio y en las cajas de búsqueda externas y muestra los resultados.

- Permite a usuarios del gobierno registrarse para acceder al backend.
- Permite presentar una nube de consultas más frecuente. En esta nube las consultas se realizan por "frase exacta" (por

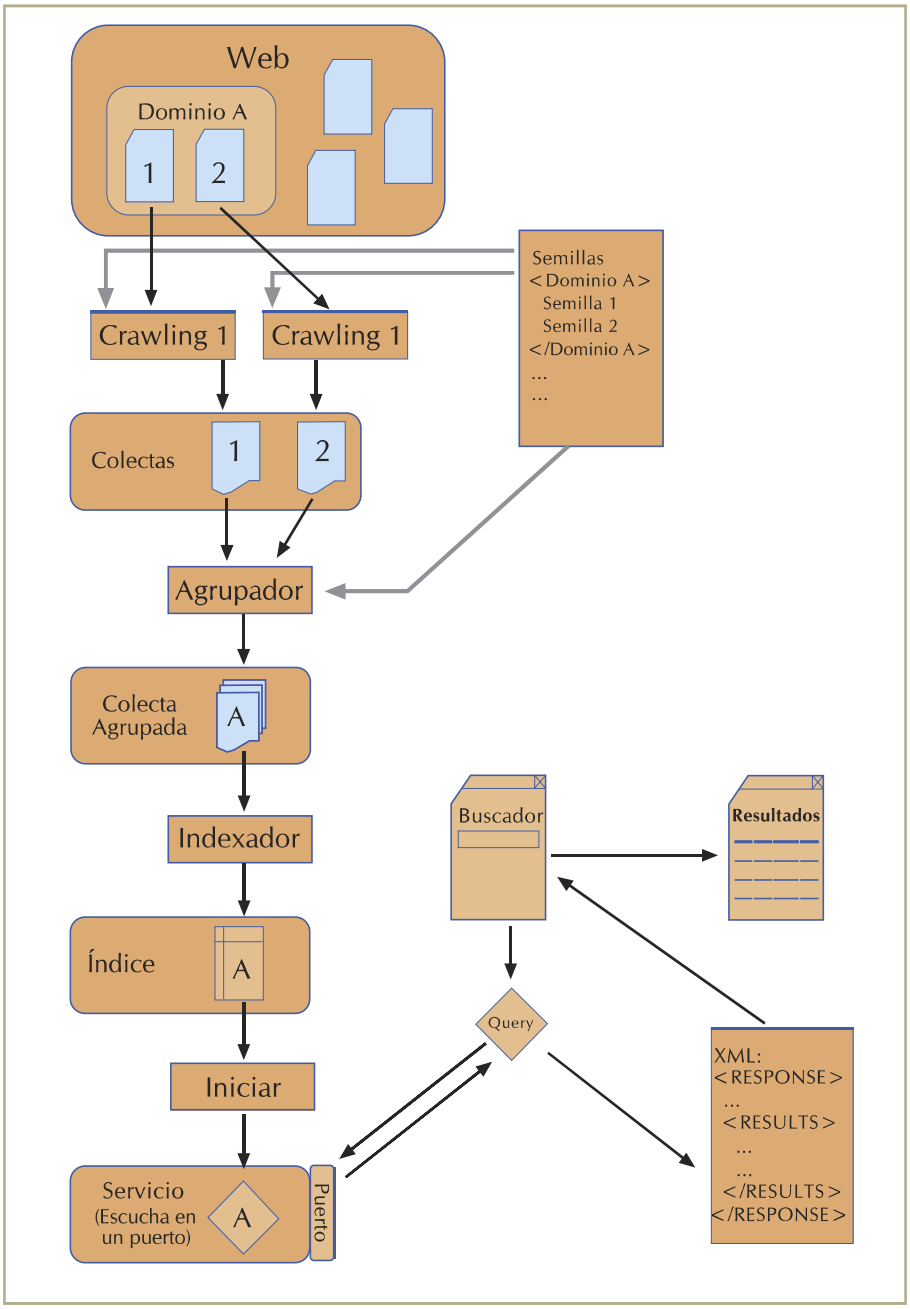

Figura 1: Funciones principales del Buscador.

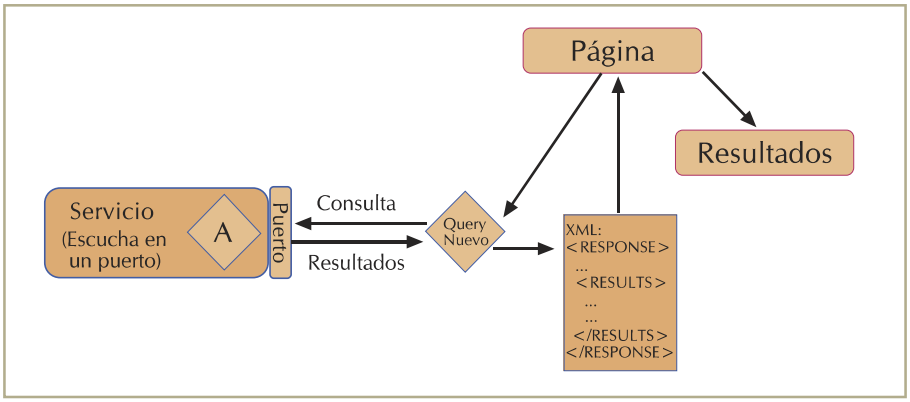

Figura 2: Sistema de respuestas.

ejemplo: "licencia de conducir"). En caso de no encontrar resultados, se utiliza el filtro "todas las palabras".

Backend: es el panel interno del gestor de contenidos. Entre otras cosas, admite:

- Gestionar la presentación y contenidos de la nube de consultas.
- Generar reportes de uso: consultas por tipo y sitio, y cantidad de consultas en los últimos períodos (día/semana/mes).
- Crear cajas de búsqueda que llevan al usuario al sitio del Buscador.
- Gestionar usuarios con acceso a los paneles de gestión de nubes de consulta, reportes de uso y de cajas de búsqueda.
- Crear nubes de consultas para las consultas más frecuentes. Se pueden aplicar filtros de número mínimo de consultas, blacklist de palabras y período de tiempo.

2.4. Detalles Adicionales

El esquema general del funcionamiento del Buscador de la Transparencia se muestra en la Figura 1. El crawler examina y descarga una parte de la Web, los sitios de interés para el sistema, y los almacena para su posterior análisis. Luego el Indexador es activado para revisar las páginas recolectadas y construir con su información índices que serán utilizados durante la posterior fase de consulta. El módulo de consultas consta de un conjunto de sistemas independientes que permanecen activos, respondiendo los requerimientos hechos con los índices que se han construido en la etapa anterior.

Tal como se muestra en la Figura 2, cada sistema de respuesta es un servicio al que se le ha asignado un puerto para que reciba consultas y entregue respuestas. La página utiliza uno de los programas de consulta para comunicarse con dicho puerto, realizar las preguntas ingresadas por el usuario y desplegar los resultados de la forma más adecuada. El sistema de procesamiento de consultas responde con un archivo XML

Figura 3: Interfaz de servicios.

Figura 4: Interfaz de Índices.

Figura 5: Mantenedor de colectas (general).

que contiene los resultados de la consulta. Los datos contenidos en los campos XML se utilizan para construir la página Web que se presenta al usuario como respuesta a su consulta.

En la Figura 3 se muestra la interfaz que permite levantar el servicio de respuesta a las consultas, pudiéndose levantar un servicio de respuestas por cada índice creado. Las acciones disponibles son iniciar o detener servicio.

En la Figura 4 se exhibe la interfaz que admite crear los índices a partir de las colectas realizadas. Sólo es posible crear los que han sido definidos en la configuración y que no están vacíos o corruptos. Además es posible borrarlos. También se pueden crear índices globales (que contengan todas las semillas del sistema), e índices de grupo que contienen los índices de algunos de los dominios generados en la interfaz de configuración. En esta Figura se muestra la interfaz que permite controlar las colectas y la Figura 5 describe los detalles de colectas individuales.

Es en la interfaz de configuración donde se configura el sistema añadiendo, borrando o modificando grupos, dominios o semillas. Cada semilla representa un sitio de gobierno que se desea colectar. Estas se agrupan en dominios, por ejemplo ministerio de Salud, el que contendría todas las semillas del Ministerio. Luego los dominios se congregan en grupos generales, los cuales sirven para diferenciar los tipos de dominio que son. Por ejemplo, si son sitios Web de organismos gubernamentales, de Transparencia o si son de municipalidades. Pueden existir muchas clasificaciones. Por cada grupo o dominio se genera un índice. El dominio contiene el índice de todas sus semillas y el de grupo contienen un índice de todas las semillas de sus dominios. La interfaz que permite realizar la gestión de semillas se muestra en la Figura 7.

Dentro de la configuración del sistema existen los enlaces; asociaciones de dominios entre los distintos grupos. De esta manera se pueden asociar los sitios de los organismos con sus páginas de Transparencia. La interfaz se muestra en la Figura 8. BITS

Figura 6: Mantenedor de colectas (detalles).

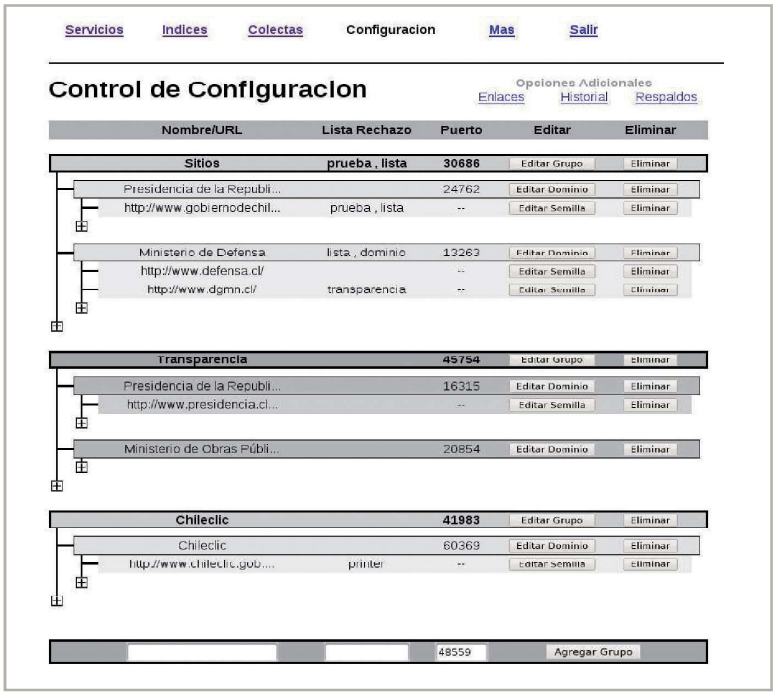

Figura 7: Interfaz de configuración.

Figura 8: Interfaz de configuración (enlaces).