

Servicios para la implementación de un motor de búsquedas basado en SOLR

Antecedentes

Las búsquedas en Internet están en las fronteras del conocimiento humano. Esto se hace visible tanto en la innovación constante como en el esfuerzo de la academia y de las empresas por investigar y generar desarrollos en este campo.

Los conceptos matemáticos de alta complejidad y programación sofisticada que dan soporte a las distintas herramientas de búsqueda resultan invisibles a los usuarios que las disfrutan. Sólo ven la interfaz y hacen uso intensivo de ella. Es que el volumen de información se ha tornado tan amplio y crece a una velocidad tan elevada que es imposible abarcarlo sin asistencia, y son precisamente las herramientas de búsqueda las que brindan esa posibilidad.

Si concebimos un buscador como una herramienta que **ordena** un universo de documentos para devolvernos los de **mayor probabilidad** de ser relevantes para una clave dada, entonces obtendremos un abanico muy amplio de posibilidades de aplicación. Partiendo de la obvia de los resultados de búsqueda, pasando por los temas relacionados y productos vinculados, y llegando a la personalización masiva y automática de contenidos.

Para la mayoría de los sitios Web, que contienen unos cientos, o tal vez algunos miles de documentos, una herramienta de búsqueda de uso gratuito tal como la Búsqueda Personalizada de la empresa Google suele ser una solución muy buena. Provee resultados de búsqueda de alta calidad sin costo económico y con una necesidad ínfima de configuración. Cuando el tamaño de la base de documentos crece, el público se vuelve variado o la diversidad de información y servicios es significativa, las necesidades de un buscador propio se hacen evidentes. Controlar la priorización de resultados, poder administrar los sinónimos, determinar palabras claves propias de la organización y sobre todo, por dar al motor de búsquedas usos más sofisticados que la provisión de una lista de resultados.

SOLR

SOLR es un potente motor de búsquedas, de código abierto, especializado en las búsquedas "corporativas", es decir, en realizar búsquedas que cumplen con dos características:

- ✓ **El universo de documentos es controlado:** independientemente del tamaño, que puede llegar sin inconvenientes a los millones, se trata de los documentos de una organización o conjunto de organizaciones y no de sitios que se publican por terceros sobre los que no se tiene control alguno.
- ✓ **La estructura de los documentos es conocida:** a diferencia de las búsquedas tradicionales de internet, donde los documentos indexados no tienen más estructura que la pequeña información que provee el HTML, SOLR está diseñado

para aprovechar la estructura interna, optimizando los resultados en función de la relevancia de los elementos (títulos vs. descripción) los tipos de los datos (fechas vs. números vs. texto) entre muchos otros.

SOLR es un producto maduro, utilizado ampliamente desde sitios de pequeño y mediano porte hasta algunos de altísimo tráfico y reconocimiento mundial como:

- ✓ WhiteHouse.gov – El sitio de la Casa Blanca
- ✓ [Netflix](http://Netflix.com) - El popular sistema para ver películas y series en línea
- ✓ Internet Archive – El repositorio que contiene la historia de Internet
- ✓ The Smithsonian Institution - la colección de uno de los museos más reconocidos del planeta, con más de 4 millones de ítems.
- ✓ [AOL](http://AOL.com) utiliza intensivamente SOLR para sus distintos servicios

Hay publicada una lista más abarcativa de sitios públicos que utilizan SOLR en <http://wiki.apache.org/solr/PublicServers>.

Objetivo

La presente propuesta tiene como objetivo la implementación de un servidor de búsqueda SOLR para brindar servicios de búsquedas a uno o más sitios Web.

Metodología

La implementación se llevará a cabo en 3 pasos: relevamiento inicial, implementación, depuración y sintonía fina. Si el cliente lo desea, Concreta provee un servicio adicional de seguimiento una vez que la implementación está en régimen, que consiste en la extensión de la etapa 3.

1. Relevamiento inicial

El punto de partida de la implementación de SOLR es la revisión detallada tanto del universo de documentos a indexar así como de los usos que se le dará a los resultados de las búsquedas.

Dentro del relevamiento se incluye:

- ✓ **Sitios a indexar:** definición de todos los dominios que se incluirán en el índice.
- ✓ **Estructura de los documentos:** descripción de la estructura interna de los distintos tipos de documentos. Por ejemplo, una ficha de producto que contiene un nombre, una descripción, un precio y una foto. Una noticia por su parte puede contener un título, un abstract, un cuerpo, un autor y una fecha.
- ✓ **Búsquedas a realizar:** en qué sitios y en qué contextos se realizarán búsquedas
- ✓ **Otros usos para las búsquedas:** utilización de resultados de búsquedas para temas relacionados, otros productos de interés o cualquier otro uso de resultados de búsqueda distintos de listado tradicional de páginas que coinciden con una clave dada.
- ✓ **Plantillas:** listado de las plantillas o formatos de salida de los resultados, que puede variar en función del uso de los resultados y de los sitios en los que se utiliza.

- ✓ **Dimensionamiento de la infraestructura:** en función del tamaño del universo de documentos y de la cantidad de búsquedas esperada, una estimación de la infraestructura de hardware que será necesaria para la implementación.

Entregable

El entregable de la etapa de relevamiento inicial se compone de dos ítems: el esquema SOLR y el plan de trabajo.

SOLR indexa los documentos en base a un esquema de campos y es a partir de esta definición que permitirá después ponderar y filtrar las búsquedas. Por ejemplo, si hay un campo de precio entonces se podrán realizar búsquedas por precio, por rangos de precios, filtrar los productos que tengan un precio mayor que o menor que, etcétera. El esquema SOLR es el punto de partida de cualquier implementación y compone entregable de la etapa 1.

En función de los datos relevados, Concreta entregará el plan de trabajo para las etapas siguientes, indicando las tareas a realizar, los plazos y los responsables de realizarlas (Concreta, el cliente, terceros). El plan de trabajo será realizado de común acuerdo con la contraparte del cliente.

Duración Estimada: 15 días.

2. Implementación

Durante la etapa de implementación se realizarán todas las tareas necesarias para poner en producción el servidor de búsquedas y hacerlo disponible a los sitios Web que lo utilizarán. Esta etapa incluye las siguientes tareas:

- 1. Infraestructura:** el cliente deberá hacer disponible la infraestructura necesaria para la implementación de SOLR.
- 2. Bocetos de plantillas:** Concreta proveerá los bocetos de todas las plantillas definidas en la etapa anterior. Concreta interactuará con el cliente hasta ajustar las plantillas a
- 3. Instalación:** Concreta realizará la instalación inicial de SOLR, incluyendo las configuraciones básicas y la definición del esquema SOLR.
- 4. Carga inicial de datos:** a partir de los datos en formato XML que provea el cliente Concreta realizará la carga inicial de datos, con todos los documentos que hayan sido definidos como parte del universo de búsquedas.
- 5. Implementación de la carga periódica:** ya sea como un proceso continuo o como un proceso que se ejecuta periódicamente Concreta y el cliente definirán cómo se incorporan los nuevos documentos y se realizan las actualizaciones necesarias sobre los viejos, de modo que el índice esté siempre al día.
- 6. Implementación de las búsquedas:** el cliente, con el apoyo de Concreta, incorporará los distintos formularios y plantillas a los sitios Web, para que las búsquedas queden disponibles para los usuarios, proveyendo la programación, scripts y HTML que hagan falta.
- 7. Puesta en el aire:** una vez que las búsquedas estén operativas, se coordinará con el cliente la puesta en el aire del motor de búsqueda y de los distintos formularios y plantillas.

Entregable

El entregable principal de la segunda etapa es el conjunto de formularios y plantillas funcionando. Es importante señalar que no es posible alcanzar este resultado sin la participación activa del cliente.

Como entregable secundario, Concreta entregará la documentación de la implementación, del esquema y del mecanismo de actualización de la información.

Duración Estimada: asumiendo que la infraestructura y los desarrollos presenten un grado de complejidad medio o bajo, se prevé un plazo de 75 días para esta etapa. En el caso de que la infraestructura o los desarrollos requieran plazos adicionales, la duración se ajustará con el cliente en la etapa 1, al definir el plan de trabajo.

3. Depuración y sintonía fina

Una vez puesto en el aire, Concreta realizará un seguimiento detallado de las búsquedas, los resultados que arrojan y otros factores relevantes para realizar los ajustes que sean necesarios, a los efectos de eliminar resultados poco relevantes y mejorar la calidad y relevancia de los resultados previstos.

Esta tarea se desarrollará en coordinación con el equipo contraparte del cliente, de modo que a la vez que se depura y mejora la herramienta de búsquedas, se transfiere el conocimiento básico para mantenerla y monitorearla.

Entregable

El entregable de esta etapa lo constituye un documento con los cambios realizados que Concreta entregará cada 15 días.

Duración Estimada: 90 días.

Seguimiento Opcional

De forma opcional, el cliente podrá optar continuar con el apoyo de Concreta a la implementación realizada, extendiendo la modalidad de trabajo de la etapa 3. En este caso, en base a la experiencia recogida se acordará el paquete de tareas a realizar, la forma de coordinación y se establecerá que los informes se realicen cada 30 días.

Responsabilidades del cliente

- ✓ **Contraparte:** nombrar una contraparte a todos los efectos del relacionamiento, con autoridad para dar la aprobación final a los entregables.
- ✓ **Transformación de los datos a XML:** el cliente deberá proveer los datos de los documentos que se indexarán en formato XML incluyendo los campos que se hayan definidos para la priorización de las búsquedas (por ejemplo: título, abstract, cuerpo, precio, categoría, etc.)
- ✓ **Hardware y software de base:** todo el equipamiento y software distinto de SOLR que sea necesario para la implementación, es responsabilidad del cliente.
- ✓ **Participación activa:** el objeto de los servicios de la presente propuesta no es posible sin la participación activa del cliente. Es éste quien conoce sus sitios, sus visitantes y clientes y su infraestructura de TI. El cliente deberá garantizar la participación activa de todos los técnicos y actores clave para el cumplimiento exitoso de los objetivos.

Confidencialidad

Toda la información que el cliente proporcione a los especialistas y técnicos de Concreta, a cualquier otro de sus empleados o personal de terceros contratados por ésta a los efectos del objeto de la presente propuesta será tratada como Confidencial, con excepción de aquella información que antes, durante o después del trabajo sea hecha pública por el cliente y que por lo tanto es considerada de dominio público. La confidencialidad no caduca cuando terminan las tareas objeto de la presente propuesta.

Condiciones económicas

El costo de los servicios incluidos en la presente propuesta es de \$169.800 (ciento sesenta y nueve mil ochocientos pesos uruguayos) por todo concepto.

La forma de pago es la siguiente:

1. 10% con la aceptación de la propuesta.
2. 10% con la aceptación de los entregables de la etapa 1- relevamiento inicial
3. 50% con la aceptación de los entregables de la etapa 2 - implementación
4. 30% a la finalización de la etapa 3 - depuración y sintonía fina

Concreta facturará una vez aceptado cada entregable final. El cliente deberá efectuar los pagos en los 10 días corridos siguientes a la fecha de recepción de la factura. Los precios no incluyen Impuesto al Valor Agregado ni ningún otro impuesto que pudiera aplicarse según la Ley.

Seguimiento opcional

En el caso de que el cliente opte por contratar el servicio de seguimiento, Concreta facturará a mes vencido el importe correspondiente al 6% del total. Este precio mensual se reajustará en junio, diciembre y/o cuando el aumento del IPC acumulado supere el 20%, según las estadísticas de IPC provistas por el Instituto Nacional de Estadísticas de Uruguay. El cliente deberá efectuar los pagos en los 10 días corridos siguientes a la fecha de recepción de la factura. Los precios no incluyen Impuesto al Valor Agregado ni ningún otro impuesto que pudiera aplicarse según la Ley.

El cliente y Concreta podrán cancelar unilateralmente el servicio de seguimiento cuando lo consideren conveniente, sin necesidad de expresión de motivo alguno, sin que esto genere ninguna otra obligación para las partes que el pago del mes en curso, en caso de que la cancelación sea decisión del cliente.

Otros trabajos

La realización de trabajos adicionales involucrará costos adicionales que serán pactados de común acuerdo entre Concreta y el cliente. Concreta no facturará ninguna hora adicional sin previa orden escrita por parte del cliente para la realización de trabajos adicionales al asesoramiento motivo de la presente propuesta.