

## Introducción a los Motores de recuperación de documentos XML/RDF

En un principio la World Wide Web fue ideada para uso humano, por lo que la recuperación y organización de la información contenidos en ella estaba sujeta al difícil proceso de automatización de búsquedas satisfactorias para los usuarios. Fue necesario dotar a las páginas Web de metadatos, es decir, información sobre los datos contenidos en el documento, como medio de describir e informar sobre los recursos ofrecidos por la Web.

Disponiendo ya de webs que proporcionan información en los metadatos acerca de sus contenidos, surgió la necesidad de automatizar el proceso de recuperación de información que describa los diferentes recursos. Como respuesta a esto se realizó la especificación XML/RDF, dejando como último hito para conseguir una recuperación eficaz de información sobre los contenidos la implementación de motores de recuperación de documentos XML/RDF

El ámbito de búsqueda de estos motores no es la World Wide Web tradicional, sino una extensión de la misma denominada Web Semántica, es decir, un entorno al que se le han añadido datos semánticos. Estos, expresados en un lenguaje formal como XML/RDF, permiten describir el contenido, el significado y la relación de los datos, facilitando su procesamiento automático.

La adición de semántica permitirá dotar a la Web de una base de conocimiento que satisfará de forma exacta las solicitudes de información de los usuarios: Supongamos que un usuario utiliza en la actualidad alguno de los motores de recuperación de información para encontrar los vuelos entre Madrid y Londres que salen esta tarde. Los buscadores actuales devuelven un amplio abanico de resultados, desde webs de aerolíneas, información sobre Madrid o Londres, y demás información descontextualizada. La única posibilidad para el usuario pasa por refinar su búsqueda sobre esos resultados, o incluso redefinir la consulta. La adición de semántica y su utilización por parte de los motores ofrecería a los usuarios una respuesta exacta: vuelos que salen esta tarde de Madrid a Londres. Gracias a la semántica palabras como tarde podrían ser interpretadas y el origen geográfico podría omitirse al detectarse y contextualizarse adecuadamente.

Por tanto, la ventaja de la dotar a la Web de contenido semántico es que permite ofrecer soluciones a problemas habituales de la recuperación y organización de la información, al servirse de una infraestructura mediante la cual la transmisión y el procesamiento de información se realizan de forma sencilla. La información no se procesa por los motores de recuperación en términos de entradas y salidas, sino en función de la semántica y apoyándose en una redefinición tanto de los operadores como de los datos.

Las siguientes secciones ofrecen una profundización en el concepto de Web semántica, los motores de recuperación utilizados en la misma y enlaces a documentación adicional.

## La Web Semántica

Esta sección profundiza en el concepto de Web Semántica, estableciendo las bases tecnológicas, que facilitan una recuperación y organización de la información más óptima que la proporcionada por los motores de recuperación actuales que realizan una exploración del contenido de la web, en función de temáticas variadas, y construyen índices que permiten agilizar las búsquedas posteriores.

Sin embargo, la inmensidad de la información indexada por los buscadores puede producir un efecto no deseado, conocido como sobrecarga de información, que se traduce en el desaliento de aquél que solicita a un buscador que le ofrezca la mayor cantidad de sitios que traten un tema específico y se encuentra con cientos, quizá miles de fuentes posibles, para los que no hay tiempo material de comprobar si su contenido le satisface. Para evitar este efecto de sobrecarga de información surge la idea de la Web Semántica, que va a ser analizada en este documento describiendo sus principales componentes.

## Estándares de la Web Semántica

La Web Semántica se sirve de diferentes estándares y herramientas para describir la función y relación entre cada uno de sus componentes, los más importantes son:

- XML (eXtensible Markup Language): es un estándar de representación que proporciona la sintaxis elemental para expresar metadatos sobre cualquier recurso digital.
- RDF: modelo de datos para los recursos y las relaciones que se puedan establecer entre ellos. Aporta una semántica básica para este modelo de datos que puede representarse mediante XML.
- OWL (Ontology Web Language): Lenguaje basado en RDF que sirven para definir ontologías que proporcionan vocabularios más apropiados para modelar todos estos conceptos.
- SPARQL: protocolo y lenguaje de consulta para fuentes de datos de la Web Semántica. Está siendo estandarizado por el "RDF Data Access Working Group" (DAWG) del "World Wide Web Consortium" (W3C).

## **Recuperación y organización de la información en la Web Semántica**

Los elementos descritos anteriormente permiten aumentar la utilidad de la WWW interconectando e informando sobre los recursos mediante:

Servidores que exponen sistemas de datos usando XML/RDF y SPARQL. Pueden proporcionar información bien usando conversores a RDF, o directamente documentos de ese tipo.

- Documentos etiquetados con información semántica generada automáticamente.
- Ontologías.
- Servicios web que proporcionan información a agentes automáticos.

### **Motores de recuperación de documentos XML/RDF**

Estos motores de recuperación constituyen una herramienta con la que buscar documentos de forma eficiente en la Web Semántica: XML/RDF. El usuario realiza una consulta de forma usual, tras lo cual se transfiere a un agente automático que mide la relevancia entre diferentes ontologías y le devuelve los resultados.

A continuación se describen las características y el funcionamiento de dos de los principales motores de recuperación de información de la Web Semántica: Swoogle y SWSE.

### **SWOOGLE**

Entre los motores de recuperación de documentos XML/RDF se puede destacar Swoogle, que es fruto de un proyecto de investigación del "Computer Science and Electrical Engineering Department at the University of Maryland, Baltimore County".

Es un motor de recuperación especializado que descubre, analiza e indexa conocimiento codificado en documentos publicados en la Web Semántica. Swoogle "razona" sobre estos documentos y las partes que los componen y almacena metadatos significativos sobre ellos.

La recuperación de información que realiza el buscador se basa en el análisis de la semántica de la búsqueda, proporcionando resultados para consultas manuales o automáticas realizadas por software. El motor ha sido también utilizado por diversas organizaciones para gestionar y mantener su base de conocimiento (documentación RDF).

Swoogle ofrece un algoritmo personalizable inspirado en el PageRank de Google pero adaptado para la utilización de patrones que se encuentran presentes en los documentos que conforman la Web Semántica. Éste emula un agente racional adquiriendo conocimiento sobre la Web Semántica usando los hipervínculos proporcionados.

## **SWSE: The Semantic Web Search Engine**

Este buscador se define como uno de los motores de recuperación y búsqueda de datos de la Web Semántica, y presume de proporcionar resultados más acertados que los buscadores tradicionales. Se presenta a través de una interfaz HTML, si bien se advierte que no es operativo para el buscador Internet Explorer por problemas de compatibilidad de JavaScript.

SWSE implementa las funcionalidades típicas de los motores de recuperación de documentos XML/RDF: búsqueda a través de semánticas RDF u OWL, cuyas ontologías y vocabularios permiten afinar las búsquedas.

El contenido a indexar proviene de la exploración de la web mediante su framework MultiCrawler le permite recopilar RDF, HTML y XML, convirtiendo estos dos últimos tipos en XML/RDF antes de añadirlos al índice.