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 de 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 un 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.

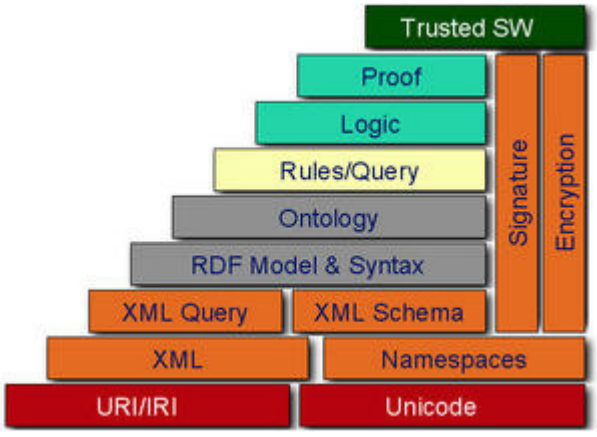
La Web Semántica

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,** y diferentes aplicaciones actuales surgidas con el "boom" de la Web 2.0

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/RDF**: XML proporciona la base sintáctica para la estructuración de documentos, sin imponer ninguna semántica. RDF es un modelo de datos para definir recursos y las relaciones entre ellos, pudiendo ser expresado mediante un documento XML.
- **OWL:** Desarrolla temáticas o vocabularios específicos asociados a los recursos. Es por tanto un lenguaje para definir ontologías, es decir, el conjunto de términos de un área de conocimiento, las relaciones que existen y la forma de operar con ellos.
- **SPARQL**: Es un 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.

Aplicaciones de la Web Semántica

La llegada de la Web 2.0 ha supuesto la aparición de aplicaciones destinadas a los usuarios, algunas de las cuales han integrado distintas funciones con contenido semántico como:

- **RSS**: Vocabulario XML/RDF que permite clasificar información de forma que sea posible encontrar la información adaptada al perfil de cada usuario.
- **FOAF**: Proyecto de Web Semántica basado en **RDF**, que permite describir personas y relaciones entre ellas, permitiendo que un **motor de búsqueda** encuentre información sobre una persona en concreto y las comunidades de las que es miembro.

El futuro de la Web Semántica pasa por la implementación de **motores de recuperación de documentos XML/RDF**. De esta forma será posible responder a las necesidades de información de los usuarios de forma precisa.

Motores de recuperación de documentos XML/RDF

Estos **recuperadores de información** constituyen la herramienta con la que buscar documentos de la Web Semántica: XML/RDF. El usuario realiza una consulta de la 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

Swoogle es un motor de recuperación para la Web Semántica, fruto de un proyecto de investigación del "Computer Science and Electrical Engineering Department at the University of Maryland, Baltimore County". 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**).

Los contenidos indexados por el motor, unos 1.4 millones, son **documentos** escritos en **XML/RDF** y **OWL,** o que incluyan fragmentos de **XML/RDF**, recogidos de la World Wide Web. La indexación se basa en la solicitud manual y una búsqueda de metas similar a la de Google.

El posicionamiento se realiza en base a un algoritmo configurable también basado en el PageRank de Google. Éste emula un agente racional adquiriendo conocimiento sobre la web semántica usando los hipervínculos proporcionados.

El sitio de Swoogle ofrece una completa documentación, con artículos sobre Swoogle, búsqueda y clasificación en la Web Semántica y una completa F.A.Q., así como información actualizada sobre los datos de indexación. Por otra parte tiene una función de archivo, dando la opción de acceder a contenido cacheado por el buscador.

The Semantic Web Search Engine - SWSE

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 o 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.

La **recuperación de documentos XML/RDF** se realiza de la siguiente forma:

- Se introduce la palabra de búsqueda.
- Se elige uno de los resultados obtenidos o bien permite refinar la búsqueda usando el filtro que ofrece el buscador: archivos de Wikipedia, FOAF Document, FOAF Person, RSS, etc.