Recuperación y Organización de la Información

Motores de recuperación de documentos XML RDF

El objetivo de la web es tratar el papel de los motores de recuperación de documentos XML y RDF dentro de la recuperación y organización de la información. Para ello se realizará una breve introducción al concepto de web semántica y al por qué de la necesidad de estos motores de recuperación para después analizar dos de ellos: swoogle y semanticwebsearch.

Introducción a la web semántica

Hasta ahora la WWW (World Wide Web) estaba concebida más para la recuperación y organización de la información para los humanos que para manipular datos o procesar información de manera automática. Sin embargo, se pensó en una web que permitiera programar agentes que navegaran la infinitud de sitios pudiendo obtener la información que necesitamos sin tener que indicarles de dónde obtenerla o qué significado debe tener cada recurso, transformando esa información posteriormente a un formato entendible por nosotros.

Este es, por tanto, el objetivo de la web semántica: crear un medio universal para el intercambio de información basado en representaciones del significado de los recursos de la web, de una manera inteligible para las máquinas. Con esto se pretende aumentar la interoperabilidad entre los sistemas informáticos y disminuir la mediación de operadores humanos en los procesos inteligentes del flujo de información.

Elementos básicos de la web semántica

- XML (eXtensive Markup Language)
- RDF (Resource Description Framework)

Para saber más sobre estos tipos de documentos ver la página sobre metadatos y documentos XML y RDF

- Ontologías – Colecciones de enunciados redactados en un lenguaje, como el RDF, que define las relaciones entre conceptos y especifica reglas lógicas para razonar con ellos.
- Agentes – Encargados de realizar las tareas para los usuarios de la web semántica.

La recuperación de la información en la web semántica

Con los actuales motores de recuperación de Internet la única información que podemos obtener son contextos descontextualizados, es decir, si en un buscador se introduce la palabra “Margarita” aparecerán resultados sobre personas con ese nombre, páginas sobre flores e incluso sobre la isla situada en Venezuela.

Con la web semántica se podrán realizar búsquedas precisas del tipo “quiero el viaje con menor coste entre Madrid y París teniendo en cuenta que soy no fumador y me gusta ir en el pasillo”.

El gran problema que presenta la recuperación y organización de la información en esta web semántica es no existen motores de recuperación de carácter general que permitan búsquedas basadas en documentos RDF por toda la web. Ni siquiera los grandes motores de recuperación como google indizan RDF, ni basan su recuperación de la información en ontologías.

Por este motivo la web semántica en la actualidad está formada por:

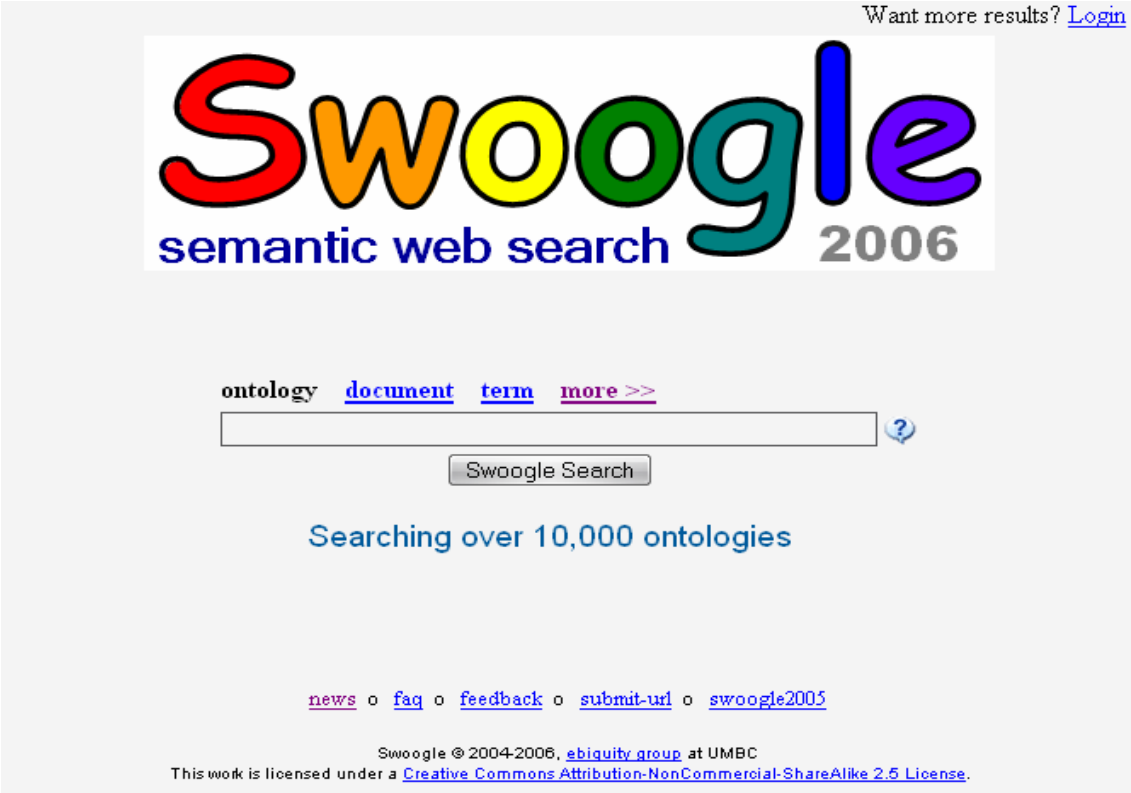
- Un conjunto de sitios o dominios particulares que utilizan sus ontologías ad-hoc y desarrollan sus propios motores de recuperación.
- Web 2.0, un conjunto de aplicaciones que comprenden algunas notaciones semánticas (RSS, FOAF) y permiten añadir y difundir contenidos en un contexto dirigido al usuario.

Swoogle

Entre los motores de recuperación de documentos XML/RDF se puede destacar Swoogle. 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.

Swoogle nace de un proyecto de investigación (aún en curso ya que finaliza en diciembre de 2006) del grupo del Computer Science and Electrical Engineering Department de la Universidad de Maryland, en Estados Unidos.

Este motor de recuperación proporciona un acceso web para facilitar la búsqueda de documentos tanto a personas como a sistemas software (por ejemplo agentes).



Posee una documentación acerca de como utilizar el buscador, estadísticas y otras características como la posibilidad de enviar la url de documentos propios para que sean indexados.

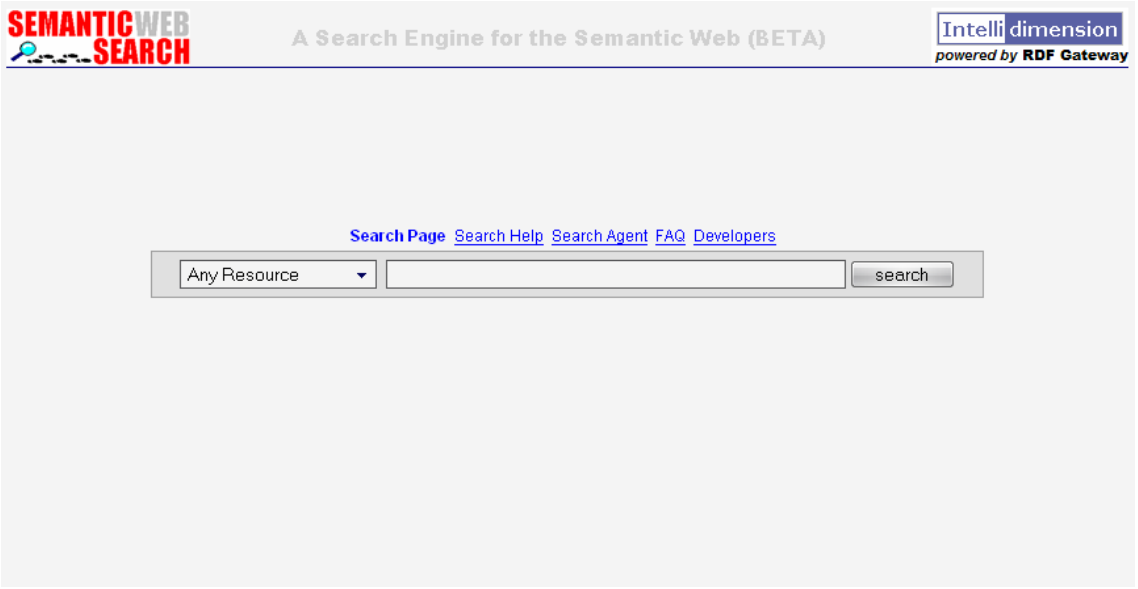
Swoogle ofrece un algoritmo personalizable inspirado en el PageRank de Google pero adaptado para la utilización de patrones que se encuentran presentes en los documentos que conforman la web semántica.

Actualmente este motor de recuperación cuenta con 1.3 millones de documentos indexados y sigue creciendo. Además para proporcionar servicios de recuperación de información en la web semántica, Swoogle ha sido utilizado por diversos proyectos para el mantenimiento y gestión de colecciones especializadas de documentos RDF.

Sin embargo, y a pesar de todo lo anterior, Swoogle no es aún un motor de recuperación dirigido al usuario final para encontrar recursos, si no que se puede considerar más bien un “parabuscador” para buscar, clasificar e incluso validar documentos y vocabularios de la web semántica.

SemanticWebSearch

SemanticWebSearch es un motor de recuperación para la web semántica. En su página (htpp://www.semanticwebsearch.com) se proporciona una interfaz estándar para un motor de recuperación (como puede ser la de google) ampliada para permitir al usuario introducir la descripción de la información que desea obtener.



Por ejemplo, con un motor de recuperación tradicional se puede buscar información sobre una persona usando como palabra clave su nombre (pongamos "Margarita"). Con esta búsqueda el motor puede producir multitud de resultados diferentes, desde páginas sobre islas en el caribe, páginas sobre plantas y flores o, lo que realmente se está buscando, personas de nombre Margarita. Será tarea del usuario el discriminar los resultados que no le interesen. Sin embargo, con el motor de recuperación SemanticWebSearch se permite la búsqueda específica de personas mediante documentos FOAF (foaf:Person) que tienen como nombre "Margarita" (foaf:firstName). A partir de ahí se pueden realizar más búsquedas para obtener artículos (rss:item) que fueron creados (dc:creador) por esa persona. El motor de recuperación permite introducir además cualquier parámetro en la búsqueda que sea utilizado en la web semántica. Por ejemplo, para obtener todos los documentos creados por Margarita se podría introducir "[dc:creador]=Margarita", obteniendo sólo los resultados que cumplan exactamente la búsqueda.

En cuanto a los agentes, el motor de recuperación SemanticWebSearch proporciona un servicio web con capacidades similares al anteriormente descrito. Los agentes software inteligentes envían *queries* describiendo la información que necesitan para llevar a cabo su tarea. El motor de recuperación devolverá la información que se corresponda exactamente con la solicitud en un formato entendible por el agente. La documentación necesaria para el uso de estos servicios por los agentes se puede encontrar en http://www.semanticwebsearch.com/reference.rsp.