Educación

Cuaderno Red de Cátedras Telefónica

El Conocimiento en la Web.

Cátedra Telefónica de la Universidad de Extremadura

Trabajo realizado con el patrocinio de la Cátedra Telefónica de la Universidad de Extremadura "Aplicación de las TIC en el Entorno Universitario".

Adolfo Lozano Tello

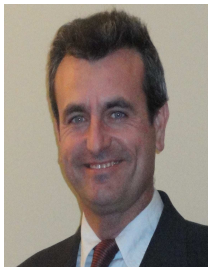
Octubre 2010

Cátedras Telefónica

www.catedras.telefonica.es

Telefónica

Biografía



Adolfo Lozano Tello

Nacido en Mérida en 1968, es profesor Titular de Universidad del Departamento de Ingeniería de Sistemas Informáticos y Telemáticos de la Universidad de Extremadura, y Director de la Cátedra Telefónica de la Universidad de Extremadura desde 2009. Es Licenciado en Informática por la Universidad de Granada en 1993 y Doctor en Ingeniería Informática por la Universidad de Extremadura en 2002 con la tesis "Métrica de Idoneidad de Ontologías". Fue Director del Servicio de Informática de la Universidad de Extremadura en los años 2004 y 2005, y Director del Centro Internacional de Referencia Linux, desde 2006 a 2008.

Índice

1. El Problema Actual de la Representación de la Información en la Web

Hoy día, la Web es un lugar preparado para el intercambio de información, diseñado principalmente para el consumo humano. Las páginas web son creadas por personas para ser entendidas por personas, o generadas dinámicamente para ofrecer a los usuarios información del producto o servicio que se quiere comunicar. Los creadores de páginas web las diseñan pensando en los potenciales usuarios que van a visitarlas, y es evidente que existen innumerables formas de ofrecer esta información.

Los actuales motores de búsqueda realizan la recopilación de información, con más o menos fortuna, mediante palabras clave que aparecen dentro del código de las páginas web, o en documentos con otros formatos, desde los múltiples servidores dispersos en Internet.

Con los estándares web del momento, al realizar búsquedas de información, no se puede diferenciar entre información personal, académica, comercial, etc. Cuando un buscador web realiza una consulta con algunas palabras clave, normalmente aparece información que no siempre es útil porque, a menudo, no se corresponde con lo que queremos encontrar. Además, el grado de detalle en la información sobre un producto, servicio o concepto es muy variable, debido precisamente a que no existe un formato o convenio que nos diga qué contenido debe aparecer sobre cada concepto. Además, los agentes de búsqueda actuales no se diseñan para “comprender” la información que reside en la web, ya que es prácticamente imposible conocer la representación de los datos ubicados en las diferentes páginas y no puede interpretarse semánticamente con qué se relaciona.

Todos sabemos que cuando queremos encontrar información concreta sobre algún producto, servicio o cualquier concepto general, debemos gastar un considerable tiempo en buscar y seleccionar la información que nos puede ser útil, navegando por las referencias URL hasta encontrar, con suerte, lo que estamos buscando. Si queremos, además, comparar dos productos o servicios similares, el tiempo que debemos dedicar se multiplica. Si en la web queremos encontrar por ejemplo un "servicio de reparación de

calefacción en Cáceres" o "una casa rural en la zona del Bierzo" sabemos que, en la mayoría de los casos, vamos a tener que dedicar bastante tiempo en encontrar y, sobre todo, comparar las características y calidades de lo que se nos ofrece. A menudo, en el tiempo que conlleva este proceso interviene la suerte y el arte de acertar con las palabras claves utilizadas.

2. Cómo Dotar de Conocimiento a la Web

Es indudable que las ventajas que ofrece Internet son enormes en la búsqueda e intercambio de información, aunque con la forma de representación actual esta ingente cantidad de información no se puede manejar de forma precisa y no está preparada para realizar deducciones lógicas con ella. Como es sabido, el lenguaje HTML es la base para representar las páginas web, que es procesado por los navegadores para mostrarnos las páginas. Con esta sintaxis se indica el tipo, formato y contenidos de los componentes (como párrafos, tablas, imágenes, formularios, etc.) que aparecen en cada una de las páginas web. Los datos que aparecen en estos componentes carecen totalmente de semántica formal, y no es posible para un agente software diferenciar, por ejemplo, si un número de una tabla se corresponde a un número de teléfono o a un precio de un producto.

Para intentar mejorar estas carencias, existen propuestas desde hace varios años para transformar la red desde un espacio de información a un espacio de conocimientos. La idea tomó impulso cuando Tim Berners-Lee, uno de los inventores de la Web, defendía el desarrollo de la Web con conocimientos [1], y organizaciones como *SemanticWeb.org* fomentan la estandarización de lenguajes y herramientas para hacer efectiva la **web semántica**. Pero, ¿qué se puede hacer en la web semántica?

La idea es que los datos puedan ser utilizados y “comprendidos” por las ordenadores sin necesidad de supervisión humana, de forma que los agentes web puedan ser diseñados para tratar la información situada en las páginas web, o en ficheros adjuntos a documentos, de manera automática.

1 Berners-Lee T., Hendler J. and Lassila O, “The Semantic Web”, Scientific American, Volume 284, Number 5 (May, 2001), pages 34-43.

Consiste en anotaciones de datos introducidas dentro del código HTML de las páginas web o en ficheros asociados a documentos, siguiendo algún esquema de anotación común. Se trata de convertir la información en conocimiento, referenciando los datos que pueden estar situados dentro de las páginas web a un esquema común consensuado sobre algún dominio (por ejemplo, consensuar el esquema de datos de la información que debe aparecer sobre casas rurales en las páginas web de la comarca del Bierzo). Estos metadatos, consistentes en información sobre los datos que se sitúan en algún servidor, no sólo describen el esquema de datos sobre la información que debe aparecer en cada caso particular, sino que además contienen información adicional que permiten hacer deducciones con ellos, es decir, axiomas que podrán aplicarse en los diferentes dominios para obtener nueva información relacionada con otros esquemas de datos. Por ejemplo, si busco el teléfono de un médico concreto (primero en su página web), y existe una relación sobre el lugar de trabajo (con información anotada en esa página web), un axioma podría llevar a un agente web semántico a deducir que un posible teléfono de contacto podría ser el del lugar del trabajo, buscando el teléfono en la página web del hospital de la relación.

La idea básica de la web semántica consiste en diseñar esquemas de metadatos (llamados **ontologías**) donde se definen los conceptos de un dominio, sus atributos, relaciones y axiomas, situados en servidores de Internet, y que puedan ser referenciados desde las páginas web que guarden relación con cada uno de estos esquemas. Tomando un ejemplo muy sencillo, supongamos que el Colegio de Médicos (o una asociación médica, o una compañía de seguros médicos, etc.) ha definido un esquema propio y lo ha colocado en su servidor. En ese esquema se representará la taxonomía de médicos especialistas, con los atributos que se consideran relevantes (como el número de colegiado, nombre, teléfono, especialidad, etc), y relaciones entre estos conceptos o con los de otro esquema de metadatos existente (por ejemplo de hospitales, compañías de seguros, etc). En una web concreta (por ejemplo de un médico particular) se pueden insertar referencias a esta ontología de médicos y completar, incrustado dentro de este código HTML, la información correspondiente a cada atributo. Los navegadores web actuales mostrarán la información habitual del código HTML, siendo invisibles para estos navegadores las anotaciones añadidas. En cambio, un navegador web semántico, podrá buscar las anotaciones insertadas dentro de este código, pudiendo identificar exactamente cada dato y tratar esta información para obtener consultas más completas y

precisas. En un entorno ideal, donde la información de todas las páginas web (o todas las de un determinado dominio) tuvieran anotaciones con referencia a ontologías consensuadas, permitiría realizar consultas totalmente precisas, y con capacidad para inferir conocimiento buscando en otros servidores. En este supuesto idealizado, se permitiría a los usuarios realizar consultas de gran complejidad como "Búscame un dentista que esté a menos de 5 kilómetros de mi casa y que tenga hueco libre antes del martes". Poder realizar consultas con este grado de precisión supondría tener anotada información detallada en cada una de las páginas web con referencias a los mismos esquemas de datos.

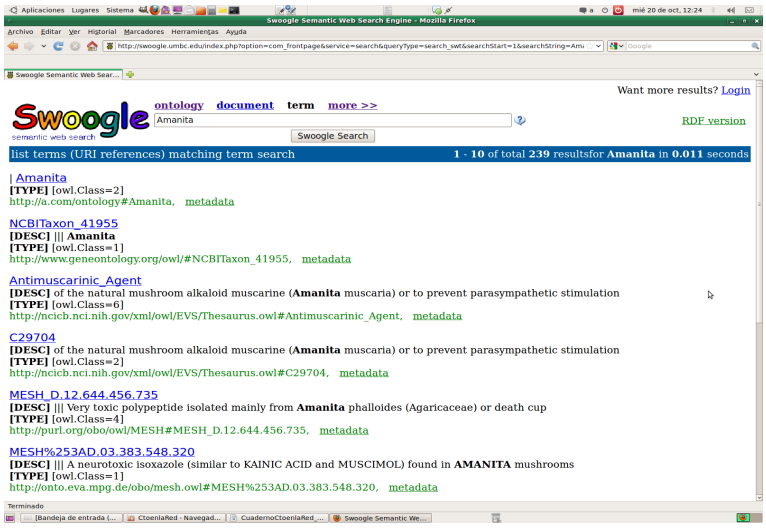
3. Posibilidades y Dificultades de la Web con Conocimientos

En la actualidad, las ontologías son la forma más extendida de representar el conocimiento, principalmente en la Web. Son utilizadas por empresas y grupos de usuarios para acordar términos, taxonomías, relaciones entre términos, atributos, restricciones en los valores de los atributos y otras posibilidades de representación del conocimiento. Los ámbitos de aplicación son enormemente variados y existen muchos ejemplos en entornos empresariales, dominios médicos, para comercio electrónico, en Ingeniería del Software, para representación de tesauros, y en la mayoría de los campos científicos y técnicos, siendo un medio transversal de representación para consensuar y formalizar el conocimiento. En la Cátedra Telefónica de la Universidad de Extremadura se están utilizando ontologías sobre domótica como base para el desarrollo de un sistema inteligente [2] que aprende de los hábitos de los usuarios con el objetivo de mejorar el ahorro energético de una instalación.

En la aceptación de las ontologías como la forma habitual de representación del conocimiento han contribuido varias causas. La estandarización del

2 Lozano-Tello, A. and Valiente, P. "Ontology and SWRL-Based Learning Model for Home Automation Controlling", International Symposium on Ambient Intelligence-ISAmI10 Advances in Soft Computing. Springer Berlin / Heidelberg. Vol. 72/2010.

lenguaje OWL (*Ontology Web Language*) [3] recomendado por la W3C, ha unificado la sintaxis de representación de ontologías. La consolidación de Protégé [4] como herramienta de edición de ontologías ha permitido mejorar esta aplicación, y se ha trabajado en el desarrollo de *plugins* para explotar las ontologías representadas. Además, hay otra característica que debe tenerse en cuenta: las ontologías suelen estar disponibles en Internet, muy frecuentemente, de forma gratuita. Esta es una diferencia fundamental con los sistemas inteligentes desarrollados hace algunos años, que tenían métodos propios y privados para representar el conocimiento. La tendencia actual es que el esfuerzo que se realiza para adquirir y organizar el conocimiento de una materia se ofrezca para que otros usuarios lo reutilicen y lo enriquezcan. Por poner un ejemplo, una ontología sobre micología desarrollada para un sistema inteligente de clasificación de setas podría ser reutilizada para un sistema diagnóstico sobre intoxicaciones. En este sentido, existe cierto paralelismo con el software libre, donde hay una comunidad de usuarios que desarrollan y prueban ciertas aplicaciones software. En este caso, puede hablarse de **conocimiento libre** en el que, a partir de un esfuerzo original de alguna empresa o grupo de usuarios de adquirir, organizar y representar el conocimiento sobre un dominio quizás para el uso de un sistema concreto, las ontologías obtenidas se ofrecen de manera gratuita a otros usuarios para que no tengan que representar desde cero ese conocimiento y, como ventaja, ayuden a mantener y completar la información representada.



A pesar de que en el presente no está arraigada la idea del trabajo colaborativo para el desarrollo de ontologías en Internet por una comunidad amplia y abierta de usuarios, tal y como existen en las forjas en las comunidades de software libre, hay iniciativas en desarrollo

3 http://www.w3.org/TR/owl-features/

4 http://protege.stanford.edu/

para almacenar repositorios reutilizables de ontologías, como OpenOntologyRepository [5] o TONES Project [6]. Como herramientas útiles que fomentan la reutilización, cabe destacar los buscadores de ontologías Swoogle [7] y Watson [8] que permiten realizar búsquedas en ontologías y documentos que tengan información semántica anotada. También debe comentarse el creciente uso de los microformatos [9] como esquemas simples de datos, aunque con reducida capacidad de representación, pero con la utilidad de poder anotar fácilmente información relacionada con datos personales, con calendarios, sobre opiniones, etc.

Se puede afirmar que las ontologías son utilizadas actualmente por la mayoría de los sistemas inteligentes que usan información consensuada y que necesitan realizar razonamiento, especialmente en Internet. La gran capacidad expresiva del lenguaje OWL (derivado de RDF) para representar ontologías, las mejoras del editor Protégé con gran cantidad de *plugins*, o la creación de marcos de desarrollo para sistemas con ontologías como Jena [10], han hecho que la tendencia en el desarrollo de sistemas inteligentes tengan su base en las ontologías. El uso de las ontologías para anotar la información interna de páginas web permite hacer consultas precisas sobre la información buscada, con el consiguiente ahorro de tiempo para los usuarios. Los agentes web que incluyan motores de inferencia para ontologías permitirán deducir información a través de axiomas y de anotaciones que residan en otras páginas web. La aplicación generalizada de esta idea supondría una revolución en el tratamiento de la información en Internet y enormes mejoras para las consultas de los usuarios y, sobre todo, en el desarrollo de agentes web que podrían tratar la información anotada para resolver gran cantidad de tareas y servicios. Sin embargo, existen varios impedimentos que están obstaculizando el desarrollo de la web semántica.

5 http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository_Requirement
6 http://www.tonesproject.org/

7 http://swoogle.umbc.edu/

8 http://kmi-web05.open.ac.uk/WatsonWUI/

9 http://microformats.org/

10 http://jena.sourceforge.net/index.html

A pesar de los beneficios que podría suponer el uso generalizado de la web semántica, es cierto que hasta la fecha las ontologías son usadas de manera local o restringida a grupos de usuarios y redes internas de empresas. Esto puede ser debido a varias causas. En primer lugar, no es trivial consensuar el conocimiento sobre un dominio. Pensemos por ejemplo que se desea crear una ontología sobre aparatos electrónicos. Hacer que las empresas fabricantes, distribuidoras, comercios, y demás agentes se pongan de acuerdo sobre las clasificaciones de estos aparatos, atributos, valores de estos atributos y relaciones entre los productos puede ser una tarea muy complicada. Otro problema es cómo convertir la web actual en web con anotaciones semánticas. Los creadores de páginas web deberían emplear un tiempo considerable en adaptar el software de sus servidores para realizar esta tarea o, en el peor de los casos, utilizar herramientas de anotación para realizar el etiquetado individualizado para cada producto o servicio que tengan visible en la web.

No se puede perder la perspectiva de otro detalle. Si se consigue convertir la web en web semántica, las consultas devolverán los resultados sin ambigüedad. Esto supondría cambiar el modelo publicitario del que disponen los actuales buscadores, ya que presuponen que se va a estar un tiempo seleccionando y probando los enlaces sugeridos donde, con mayor o menor agresividad, se priorizan ciertos enlaces o se muestran la publicidad de empresas patrocinadoras.

Como conclusión, puede afirmarse que las ontologías se han consolidado como la forma de representación del conocimiento que se usa en los sistemas inteligentes actuales. Suelen estar disponibles en Internet para que puedan ser reutilizadas y mejoradas por otros desarrollos, por lo que se está mejorando la representación del conocimiento universal en la web en muchas áreas del conocimiento. Utilizar esta capacidad para anotar las páginas web existentes supondría una revolución en Internet que lograría obtener búsquedas precisas, y se podrían desarrollar agentes con motores de inferencia que permitirían realizar un innumerable conjunto de servicios. Por los problemas indicados, quizás sólo se consiga en los próximos años una parte de las posibilidades que supondría dotar a la web de conocimientos.