BUSCAMEDIA
HACIA UNA ADAPTACIÓN SEMÁNTICA DE MEDIOS DIGITALES MULTIRRED- MULTITERMINAL
PROYECTO CENIT-E

Búsqueda semántica a través del Procesamiento de Lenguaje Natural

José L. Martínez-Fernández, José C. González, Pablo Suárez y Paloma Martínez
DAEDALUS S.A., Grupo LABDA - Universidad Carlos III de Madrid

INFORMACIÓN DEL ARTÍCULO

Publicado el 30 de septiembre del 2010

Palabras clave:

Búsqueda semántica, búsqueda por palabras, recuperación de información, búsqueda de respuestas, búsqueda facetada

RESUMEN

En los últimos años la tecnología de búsqueda semántica está en boca de todos como solución a los problemas de los sistemas de recuperación de información actuales. El objetivo es ir más allá de una simple búsqueda por palabras, teniendo en cuenta algo más que la aparición o no de unas cuantas palabras en un texto. La búsqueda semántica pretende comprender las expresiones proporcionadas por el usuario en su consulta, desambiguando su significado si es el caso. Entre los motivos para ir más allá se encuentra tanto el deseo de aumentar la precisión de los resultados de la búsqueda, como la necesidad de facilitar al usuario la especificación de sus consultas. Así, el Procesamiento de Lenguaje Natural (PLN) es imprescindible, no sólo a la hora de simplificar la interacción del usuario con el buscador sino también como herramienta para interpretar el significado de búsquedas y contenidos web. Por otra parte, el PLN permite trabajar con metadatos asociados a contenidos multimedia (video, audio, imágenes), cada vez más presentes en la web.

Los buscadores tradicionales como Google, Yahoo! o Bing no interpretan el significado de la expresión de búsqueda por el usuario; se limitan a comprobar si las palabras aparecen o no en los contenidos web, la frecuencia con que lo hacen y cómo se enlazan las páginas

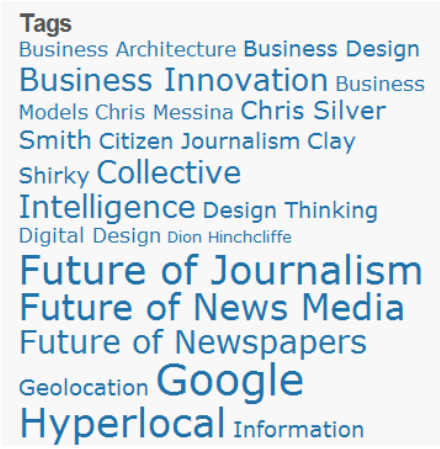
en que aparecen. En ellos, la búsqueda se lleva a cabo estrictamente en función de los términos introducidos por el usuario, sin consideraciones relacionadas con el sentido de las cosas. La única diferencia entre Google, Yahoo! y Bing se encuentra en los rangos estadísticos que utilizan para posicionar mejor o peor un determinado resultado obtenido [3].

Existen varios inconvenientes en este tipo de buscadores, entre ellos [1], [2]:

- Dificultad del usuario para expresar la consulta que desea realizar.
- Poca precisión de los resultados.
- Sensibilidad de los resultados frente a los términos exactos introducidos.
- Aparición de resultados ruidosos, es decir, páginas que consiguen relevancia por técnicas optimización de buscadores (*Search Engine Optimization, SEO*) a través de términos clave que no tienen en realidad nada que ver con su contenido.
- Uso de procedimientos estadísticos, que ponderan positivamente lo más usado por los visitantes del buscador (lo más popular), pero pierden eficacia con relación a lo minoritario o específico.

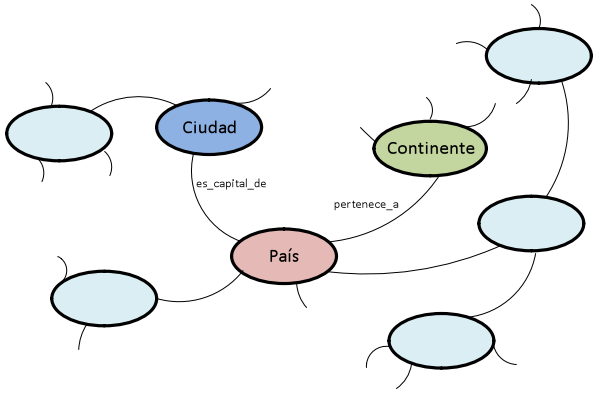
La introducción de tecnología semántica en el proceso de búsqueda puede contribuir a superar algunos de los problemas mencionados pero, ¿qué es exactamente la búsqueda semántica? Existen muchas definiciones al respecto, algunos [4] opinan que se trata de una búsqueda sobre metadatos, mientras otros [5] hablan de interpretar y utilizar el significado de los términos de búsqueda para ordenar los resultados. En la actualidad, un buscador semántico es aquel que incorpora una o varias de las siguientes características:

- Permite realizar búsquedas por campos, es decir, sobre metadatos asociados a los contenidos.
- Tiene capacidad para extender los términos de la consulta mediante sinónimos o palabras relacionadas.
- Reconoce entidades nombradas, como nombres de empresas, organizaciones o personas, que se emplean como tales en el proceso de búsqueda.
- Emplea técnicas de agrupamiento para construir categorizaciones de contenidos sobre los que buscar o para agrupar términos clave. Es el caso de las nubes de etiquetas que muestran los términos clave de un sitio web según su importancia.



Ejemplo nube de tags

- Detecta relaciones entre términos de búsqueda y palabras que aparecen en los contenidos basándose en modelos de conocimiento representados a través de ontologías.
- Ofrece la posibilidad de emplear lenguaje natural para expresar consultas e incluso preguntas factuales, para las que se obtienen respuestas concretas. Por ejemplo, respondiendo a preguntas como *¿Cuál es la capital de Croacia?*



Ejemplo de representación gráfica de ontología

Hasta la fecha no se han encontrado buscadores comerciales que incluyan todas estas capacidades, aunque sistemas como Hakia o Wolfram Alpha siguen esta línea. En cualquier caso, ninguno de los buscadores comerciales o experimentales ofrece estas capacidades semánticas en entornos multilingües.

Por otro lado, los contenidos de la web actual incluyen cada vez una mayor proporción de videos, imágenes y audios, por lo que los buscadores deben ir más allá del tratamiento del texto. En este aspecto, la posibilidad de trabajar con metadatos permite incluir contenidos multimedia en el proceso de búsqueda.

Desde nuestro punto de vista, un buscador, para denominarse semántico, debe contemplar una hibridación de las técnicas mencionadas, proporcionando así una comprensión completa de las consultas del usuario y unos resultados mucho más adaptados a sus necesidades.

Desde el punto de vista del proyecto BUSCAMEDIA, se marca como objetivo el análisis de estas tecnologías semánticas con el fin de integrarlas en un proceso de búsqueda único. Cada uno de los socios involucrados en el proyecto tiene experiencia en alguno de los ámbitos mencionados, constituyendo el mejor equipo posible para abordar la combinación de todas ellas en un proceso de búsqueda semántica.

Como modelo de conocimiento que sirva de base para este tratamiento semántico, se trabaja en la creación de una ontología multilingüe, multidominio y multimedia, denominada M3.

En DAEDALUS disponemos de recursos lingüísticos y productos software para el Procesamiento de Lenguaje Natural desde una perspectiva multilingüe. Clientes como Grupo PRISA (El País), Unidad Editorial (El Mundo), Vocento (ABC), lainformacion.com, el Instituto Cervantes o Yell Publicidad (Páginas Amarillas, 11888), entre muchos otros, confían en nuestra tecnología para mejorar los contenidos que generan y los procesos de clasificación y búsqueda que forman parte de su operativa diaria. En el marco del proyecto BUSCAMEDIA, DAEDALUS persigue la aplicación de su tecnología en el ámbito del tratamiento semántico, de cuyos resultados se beneficiarán nuestros clientes.

Referencias

[1]. Abián, M. A.: *El futuro de la Web*, 2005

[2]. Abián, M. A.: *Buscadores semánticos: Comprender para encontrar*, 2009

[3]. Arrieta, A.: *La nueva frontera: buscadores semánticos*, 2009

[4]. Starr, B.: Semantic Search: An Interview with Peter Mika., Yahoo! Research, 2010

[5]. Pérez, J.C.: Google Rolls out Semantic Search Capabilities, IDG News, 2009

Currículum vitae del/os autor/es

José Luis Martínez Fernández es Doctor Ingeniero de Telecomunicación por la Escuela Técnica Superior de Ingenieros de Telecomunicación (ETSIT) de la Universidad Politécnica de Madrid (UPM), 2010. Es profesor asociado del Departamento de Informática de la Universidad Carlos III de Madrid desde 2002, concretamente en el Grupo de Bases de Datos Avanzadas. Compagina su labor docente con la Dirección de Consultoría de DAEDALUS, compañía de la que es socio. Entre 2000 y 2001 trabajó en SGI (Soluciones Globales de Internet), unidad de negocio del grupo GMV Sistemas S.A. Ha dirigido proyectos de investigación en el ámbito de los Sistemas Inteligentes, contando con numerosas publicaciones internacionales.

José Carlos González Cristóbal, Doctor Ingeniero de Telecomunicación por la Universidad Politécnica de Madrid (1989). Profesor Titular de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la UPM, donde trabaja desde 1985. Socio fundador, en 1998, de DAEDALUS, S.A., empresa especializada en Tecnologías de la Lengua, Gestión de Conenidos e Inteligencia de Negocio. Investigador en el área de los Sistemas Inteligentes (Razonamiento Aproximado, Sistemas Multiagente, Ingeniería Lingüística, Gestión del Conocimiento, Aprendizaje Automático, etc.), ha dirigido múltiples proyectos de I+D nacionales y europeos, con financiación pública y privada. Entre 1996 y 1998, ha sido Presidente del Consejo Técnico de CITAM A.I.E. (Centro de Investigación en Tecnologías y Aplicaciones Multimedia), así como Adjunto al Director de la ETSIT-UPM para Cooperación Institucional. Presidente del Capítulo Español de la Computer Society, IEEE, desde enero de 2002 hasta 2009.

Pablo Suárez García es Doctor en Filología por la Universidad de Oviedo (2007), Licenciado en Ciencias Matemáticas por la UNED (2007) e Ingeniero de Telecomunicación por la Universidad Politécnica de Valencia (1996). Entre 1997 y 2006 trabajó como analista para IECISA e INDRA. En 2007 fue investigador contratado por la Universidad Politécnica de Madrid. Desde 2008 es investigador senior en DAEDALUS.

Paloma Martínez Fernández, Licenciada en Informática por la Universidad Politécnica de Madrid desde 1992 y Doctora en Informática por la misma Universidad desde 1998. Actualmente es Profesora Titular del Área de Ciencias de la Computación e Inteligencia Artificial del Departamento de Informática de la Universidad Carlos III de Madrid y responsable del Grupo de Bases de Datos Avanzadas. Sus líneas de interés son la tecnologías del lenguaje humano (recuperación y extracción de información multilingüe en distintos dominios, búsqueda de respuestas, reconocimiento de entidades y tratamiento de información temporal) y accesibilidad web.

Proyecto cofinanciado por el Centro para el Desarrollo Tecnológico Industrial (CDTI)