

Buscador Semántico para la Documentación de la Administración Pública del Principado de Asturias

Diego Berrueta and Luis Polo

Fundación CTIC, Departamento de I+D+i
Parque Científico y Tecnológico de Gijón
33203, Gijón, España
{diego.berrueta,luis.polo}@fundacionctic.org,
<http://www.fundacionctic.org/actuaciones/idi>

Resumen El Departamento de I+D+i de la Fundación CTIC (Centro para el desarrollo de las Tecnologías de la Información y Comunicación en Asturias) impulsa una línea de especialización tecnológica en el ámbito de la Web Semántica. Dentro de esta actividad, un equipo multidisciplinar¹ está actualmente trabajando en el desarrollo de servicios semánticos para el Principado de Asturias y, en particular, en un buscador semántico para el Boletín Oficial del Principado de Asturias (BOPA).

1. Introducción y Motivación

El Boletín Oficial del Principado de Asturias (BOPA) es una publicación diaria emitida por la administración autonómica, a través de la cual comunica a los ciudadanos todo tipo de información: nuevas leyes, notificaciones, convocatorias de subvenciones o empleo público, resoluciones, etc. Teóricamente, cualquier ciudadano está capacitado para acceder a la información que aparece en los boletines oficiales. Sin embargo, el volumen de textos y la opacidad del lenguaje técnico administrativo y jurídico dificultan esta comunicación entre el ciudadano y la Administración Pública.

El área de web semántica de la Fundación CTIC trabaja en un proyecto que facilita el acceso a la información legal y administrativa mediante la construcción de una herramienta de búsqueda semántica. El objetivo que se persigue consiste en habilitar a las personas sin conocimientos jurídicos ni del vocabulario específico (es decir, la mayoría de ciudadanos) para que encuentren la información que buscan en el Boletín Oficial. Para el desarrollo de este proyecto se emplean herramientas *open source* como Protégé 3.2. (editor de ontologías), Lucene 1.4 (buscador sintáctico) y OWL-API (API para tratar y manejar ontologías OWL[1]).

¹ El equipo de este proyecto está compuesto por tres informáticos y tres lingüistas (I. Frade, J.M. Álvarez, D. Berrueta, E. Rubiera, M. Cuevas, L. Polo), en colaboración con los profesores J.E. Labra, A. Cernuda, E. del Teso, G. Lorenzo, R. Bosch y N. Cueto de la Universidad de Oviedo, bajo la dirección de A. Campos.

2. Desarrollo y Estado del Proyecto

El primer paso consistió en extraer la información (artículos del boletín) de distintas bases de datos y representarla en XML. A continuación, se desarrolló la labor semántica con un enfoque basado en ontologías. El modelo que se ha desarrollado consiste en usar las ontologías como bases de conocimiento complejas que sirven de interfaz entre las intenciones de búsqueda del usuario y la base documental que conforma el BOPA. Se han construido varias ontologías de dominio sobre las que se han implementado tres servicios: un buscador semántico, una suscripción a alertas y un navegador de conceptos.

2.1. Las Ontologías

Para formalizar las ontologías, este proyecto utiliza el lenguaje OWL-DL[2]. Se han identificado dos categorías diferentes de ontologías, según su propósito:

- Ontología administrativa: formaliza la estructura básica del BOPA y de la Administración Pública regional.
- Ontologías de dominio: se corresponden con ámbitos concretos y reconocibles por el usuario. Teniendo en cuenta que los dominios que abarca un Boletín Oficial son muy amplios, inicialmente el proyecto se concentró en un dominio particular (*comunidades de vecinos*) y, actualmente, está expandiéndose para abarcar nuevos dominios.

La arquitectura de la ontología de dominio es modular, construida a partir de micro-ontologías reutilizables. Cada micro-ontología abarca un ámbito de conocimiento que es competencia de un organismo de la Administración. De este modo, el diseño es más escalable y permite acometer futuras ampliaciones. Por otro lado, la ontología administrativa sirve como filtro de la ontología de dominio, dirigiendo la búsqueda hacia ciertos tipos de disposiciones (leyes, resoluciones, notificaciones, decretos, etc.) y organismos de la administración (Consejería de Sanidad, Educación, etc.).

La construcción de las ontologías sigue el modelo semántico que propone la ontología funcional DOLCE[3], desarrollada en el proyecto WonderWeb. Esto provee un marco de integración para ontologías de distinto tipo. El alineamiento entre DOLCE y las categorías de WordNet da pie a pensar que es posible llegar a crear un triángulo entre las tres bases de conocimiento: las ontologías de dominio, DOLCE y WordNet (EuroWordNet[4]).

2.2. Funcionamiento del Buscador Semántico

Una aproximación convencional, utilizando etiquetado de los artículos con conceptos de la ontología, presenta dos problemas prácticos: el tamaño de la base documental, que asciende a decenas de miles de artículos cada año, y el mantenimiento conforme se vayan incorporando nuevas ontologías de dominio. En cambio, este proyecto realiza un enfoque distinto basado en ontologías para

acceder a los documentos a través de un complejo híbrido de búsqueda sintáctica convencional, reglas de inferencia y un algoritmo de activación de conceptos (*spreading activation algorithm*[5,6]).

De este modo, cuando un usuario realiza una búsqueda, primero se reconocen los conceptos subyacentes a la cadena de texto introducida (palabras) y se contextualizan, es decir, se selecciona la ontología adecuada. Después se calcula un superconjunto de conceptos semánticamente relacionados, cuya combinación trata de reflejar las intenciones de búsqueda del usuario. Cada concepto es anotado con un valor numérico que pondera su relación con estas intenciones, o lo que es lo mismo, la probabilidad de que el concepto sea relevante para la consulta del usuario.

A partir de esta información (los conceptos activados y su relevancia), y utilizando conocimiento lingüístico (morfología y sinonimia), se prepara una consulta sintáctica sobre los artículos indexados con el motor de búsqueda Lucene. Aunque en último término se utilice un buscador sintáctico, el valor semántico del buscador radica en que la consulta está guiada por las relaciones entre los conceptos de la ontología.

De esta forma, una consulta por «accesibilidad» nos remite a artículos vinculados con el concepto *accesibilidad* y con conceptos semánticamente próximos como *minusvalía*, *discapacidad*, *barrera arquitectónica* o *sordera*. Se obtienen resultados muy interesantes cuando el usuario introduce varios conceptos. Por ejemplo, la consulta «vacaciones del portero» genera como resultado el convenio laboral del sector de empleados de fincas, ya que es en este documento en el que se regulan los días de vacaciones de los que puede disfrutar el portero de una finca urbana (véase Figura 1).

2.3. Arquitectura e Interfaz

La aplicación está construida en la plataforma J2EE utilizando el *framework* del Principado de Asturias (FW-PA). El interfaz de usuario se ajusta a los estándares del W3C. Además de los lenguajes XHTML y CSS, se utiliza RDF para describir los metadatos y SVG para representar gráficamente las relaciones entre los conceptos. La funcionalidad se expone también a través de servicios web, abriendo la puerta a la integración con otras aplicaciones. Se presta especial atención a los aspectos relacionados con la accesibilidad. Actualmente está en construcción un interfaz de voz usando VoiceXML.

2.4. Líneas de Investigación

Actualmente se continúa desarrollando el proyecto, trabajando en el ajuste y refinamiento de las búsquedas. En particular, nuestras líneas de investigación son:

1. Correspondencia entre las palabras y los conceptos.
2. Relación entre los conceptos y los artículos.
3. Ajuste del algoritmo de activación de conceptos.

PRINCAST.ES
GOBIERNO DEL PRINCIPADO DE ASTURIAS

Inicio: Buscador Sintáctico: Buscador semántico jueves, 23 / marzo / 2006

Inicio Boletín de hoy Acceso por fecha Buscador Lista de conceptos Lista de contextos

Buscador

Término(s) de búsqueda: [Ayuda sobre la búsqueda](#) [Cambiar al formulario de búsqueda avanzada](#)

Búsqueda por términos Búsqueda semántica en el contexto «Particular empleado de fincas»

Mostrando 100 primeros resultados (encontrados 1.320 artículos en 1,681 seg.)

Artículos

Relevancia	Título	Fecha
1	RESOLUCION de 26 de abril de 2005, de la Consejería de Industria y Empleo, por la que se ordena la inscripción del Convenio colectivo del sector de Empleados de Fincas Urbanas, en el Registro de Convenios Colectivos de la Dirección General de Trabajo.	21/05/05
2	RESOLUCION de 26 de agosto de 2005, de la Consejería de Industria y Empleo, por la que se ordena la inscripción del Convenio colectivo del sector de Empleados de Fincas Urbanas, en el Registro de Convenios Colectivos de la Dirección General de Trabajo.	24/09/05
3	DECRETO 99/2005, de 23 de septiembre, por el que se regula el régimen jurídico y retributivo del personal docente e investigador contratado laboral por la Universidad de Oviedo.	3/11/05

Conceptos consultados

- Vacaciones ESUB
- Portero SVB

Conceptos relacionados

- Contrato laboral SVB
- Vacaciones ESUB
- Empleado de fincas

Figura 1. Captura de pantalla de los resultados del buscador semántico ante la consulta «vacaciones del portero». Los dos primeros resultados corresponden a sendos convenios laborales del sector de empleados de fincas urbanas.

- Evaluación de la herramienta mediante métricas de calidad y realimentación de los usuarios, a través de un periodo de pruebas en la red de Telecentros del Principado de Asturias.

Referencias

- McGuinness, D.L., van Harmelen, F.: Owl web ontology language overview. Technical report, W3C (2004)
- Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook: Theory, Implementation, and Applications. In Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: Description Logic Handbook, Cambridge University Press (2003)
- Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: The WonderWeb Library of Foundational Ontologies (D18). Laboratory for Applied Ontology - ISTC-CNR. (2003)

4. Vossen, P.: Eurowordnet, general document. Technical report, University of Amsterdam (1999)
5. Rocha, C., Schwabe, D., de Aragão, M.P.: A hybrid approach for searching in the semantic web. In: WWW. (2004) 374–383
6. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review (1997) 453–482