XXX aniversario

novática

UPGRADE Revista de la Asociación de Técnicos de Informática

Novática, revista fundada en 1975 y decana de la prensa informática española, es el órgano oficial de expresión y formación continua de **ATI** (Asociación de Técnicos de Informática). ***Novática*** edita también **UP**GRADE, revista digital de **CEPIS** (*Council of European Professional Informatics Societies*), en lengua inglesa, y es miembro fundador de **UP**ENET (**UP**GRADE ***E**uropean **NET**work*)

<http://www.ati.es/novatica/>
<http://www.upgrade-cepis.org/>

ATI es miembro fundador de **CEPIS** (*Council of European Professional Informatics Societies*) y es representante de España en **IFIP** (*International Federation for Information Processing*); tiene un acuerdo de colaboración con **ACM** (*Association for Computing Machinery*), así como acuerdos de vinculación o colaboración con **AdaSpain**, **AI2** y **ASTIC.**

Coordinación Editorial, Redacción Central y Redacción ATI Madrid
Padilla 66, 3º, dcha., 28006 Madrid
Tlfn.914029391; fax.913093685 <novatica@ati.es>
Composición, Edición y Redacción ATI Valencia
Av. del Reino de Valencia 23, 46005 Valencia
Tlfn./fax 963330392 <secreval@ati.es>
Administración y Redacción ATI Cataluña
Ciudad de Granada 131, 08018 Barcelona
Tlfn.934125235; fax 934127713 <secregen@ati.es>
Redacción ATI Andalucía
Isaac Newton, s/n, Ed. Sadiel,
Isla Cartuja 41092 Sevilla, Tlfn./fax 954460779 <secreand@ati.es>
Redacción ATI Aragón
Lagasca 9, 3-B, 50006 Zaragoza.
Tlfn./fax 976235181 <secreara@ati.es>
Redacción ATI Asturias-Cantabria <gp-astucant@ati.es>
Redacción ATI Castilla-La Mancha <gp-clmancha@ati.es>
Suscripción y Ventas
<http://www.ati.es/novatica/interes.html>, o en ATI Cataluña o ATI Madrid
Publicidad
Padilla 66, 3º, dcha., 28006 Madrid
Tlnf.914029391; fax.913093685 <novatica.publicidad@ati.es>
Imprenta
Derra S.A., Juan de Austria 66, 08005 Barcelona.

Depósito legal: B 15.154-1975 -- ISSN: 0211-2124; CODEN NOVAEC

Portada: Antonio Crespo Foix / © ATI 2005
Diseño: Fernando Agresta / © ATI 2005

sumario

Nº 178, noviembre-diciembre 2005, año XXXI

Monografía del próximo número:
"*Factores clave de éxito en Ingeniería de Software*"

Recuperación de información en la Web Semántica

David Vallet Weadon, Miriam Fernández Sánchez, Pablo Castells Azpilicueta
Escuela Politécnica Superior, Universidad Autónoma de Madrid

<{david.vallet,miriam.fernandez,pablo.castells}@uam.es>

Resumen: *la búsqueda semántica ha sido una de las motivaciones principales de la Web Semántica desde sus inicios. En este artículo proponemos un modelo para la explotación de bases de conocimiento orientadas a ontologías para mejorar la búsqueda en grandes repositorios documentales. El modelo de recuperación se basa en una adaptación del modelo vectorial clásico, con un método para la asignación de pesos a la anotación semántica de documentos, y un algoritmo de ranking o clasificación. La búsqueda semántica se combina con una búsqueda basada en palabras clave para conseguir una tolerancia a la incompletitud de las bases de conocimiento. Nuestra propuesta se ha probado en corpus de escala significativa, con resultados prometedores respecto de la búsqueda por palabra clave, y abriendo campo para el análisis y la exploración.*

Palabras clave: *anotación semántica, ontologías, recuperación de información, Web Semántica.*

David Vallet Weadon está realizando el doctorado en la Universidad Autónoma de Madrid (UAM), donde obtuvo el título de Ingeniero en Informática. Su interés investigador se centra en la confluencia de la recuperación de información, modelización de usuario, y modelización de contexto.

Miriam Fernández Sánchez es Ingeniero en Informática por la Universidad Autónoma de Madrid (UAM), donde está realizando el doctorado. Sus intereses de investigación incluyen la ingeniería de ontologías, búsqueda semántica, y anotación semántica semi-automática.

Pablo Castells Azpilicueta es profesor titular de la Universidad Autónoma de Madrid (UAM) desde 1999. Obtuvo el doctorado en Ingeniería Informática en 1994 en la misma universidad, con una tesis en el campo de la demostración automática de teoremas. En 1994-95 realizó una estancia posdoctoral en la Universidad del Sur de California (EE.UU.). Más recientemente, ha dirigido o participado en varios proyectos nacionales e internacionales en las áreas de la Web Semántica y los sistemas basados en conocimiento, en dominios de aplicación como el periodismo, finanzas y gestión sanitaria. Su investigación se centra actualmente en la recuperación de información, las tecnologías de personalización, y los servicios web semánticos.

1. Introducción

La búsqueda semántica ha sido uno de los beneficios esperados de la Web Semántica desde su emergencia a finales de los 90. Una forma de entender un motor de búsqueda semántica es como una herramienta que recibe consultas basadas en ontologías (p.e. en RDQL, RQL, SPARQL, etc.), las ejecuta contra una base de conocimiento, y devuelve tuplas que satisfacen la consulta [2][3][7]. Estas técnicas típicamente utilizan modelos booleanos de búsqueda, basados en una visión ideal del espacio de información, consistente en piezas formales de conocimiento ontológico sin ambigüedad ni redundancia. Bajo esta perspectiva, un elemento de conocimiento es una respuesta o bien correcta o bien incorrecta procedente de una petición de información, y por ende los resultados de la búsqueda se suponen siempre 100% precisos, de forma que no se contempla la noción de respuesta aproximada a una necesidad de información. Si bien esta concepción de la búsqueda semántica aporta ventajas fundamentales, nuestro trabajo pretende dar un paso más allá. En nuestra forma de ver la recuperación de información en la Web Semántica, un motor de búsqueda devuelve documentos, más que (o además de) valores exactos, en respuesta a las consultas del usuario. Más aún, y como requisito clave para la escalabilidad hacia fuentes masivas de información, el motor debe ordenar los documentos de acuerdo con criterios de relevancia basados en las ontologías.

Un modelo de recuperación basado en ontologías puramente booleano tiene sentido cuando el corpus de información puede ser completamente representado como una base de conocimiento formal, de manera que los resultados de las búsquedas consisten en entidades de la ontología. Pero, como es bien conocido, existen límites respecto al punto hasta donde el conocimiento se puede formalizar de este modo. En primer lugar, debido al enorme volumen de información disponible hoy día en forma de texto y contenidos multimedia no estructurados, convertir esta cantidad de información en conocimiento ontológico con un coste viable es un problema sin resolver en general. En segundo lugar, los documentos tienen un valor por sí mismos, y no son equivalentes a la suma de sus partes. Aunque es útil descomponer documentos en unidades de información menores que puedan ser reutilizadas y ensambladas para diferentes del modelo vectorial clásico de recuperación de información [10], adecuada para una representación basada en ontologías, sobre la cual definimos un algoritmo de *ranking*. El rendimiento de nuestro modelo está en relación directa con la cantidad y calidad de la información en la BC sobre la que opere. Los últimos avances en la automatización del poblado de ontologías y la anotación semi-automática de textos son prometedores [6]. Mientras, si es que algún día sucede, las ontologías y metadatos (y la propia Web Semántica) llegan a ser un recurso de disponibilidad común, la falta o incompletitud de las ontologías y BCs disponibles será una limitación que muy probablemente tendremos que admitir a medio plazo. En consecuencia, la tolerancia a BCs incompletas se establece como un importante requisito en nuestra propuesta.

2. "Estado del arte"

Nuestra visión del problema de recuperación semántica es muy próxima a las propuestas de KIM [6]. Mientras que KIM se centra en el poblado de ontologías y la anotación automática de textos, nuestro trabajo se ocupa de los algoritmos de *ranking* para la búsqueda semántica. Junto con TAP [5], KIM es una de las propuestas más completas publicadas hasta la fecha, en nuestro conocimiento, para la construcción de BCs y la anotación automática a gran escala. Nuestro trabajo complementa al de KIM y TAP con un algoritmo de *ranking* específicamente diseñado para un modelo de recuperación basado en ontologías, utilizando un sistema de indexado semántico basado en la ponderación de las anotaciones.

Los llamados portales semánticos [2][3][7] típicamente proporcionan funcionalidades sencillas de búsqueda que más podrían caracterizarse como recuperación semántica de datos que como recuperación semántica de información. Las búsquedas devuelven instancias de una ontología más que documentos y por lo general no se proporciona un método de *ranking*. En algunos sistemas, se añaden enlaces a documentos referenciados por las instancias, junto a cada instancia devuelta en la respuesta a la consulta [3], pero ni las instancias ni los documentos están ordenados por relevancia.

El problema del *ranking* se ha retomado en [11] y más recientemente en [9]. Mientras

> El rendimiento de nuestro modelo está en relación directa con la cantidad y calidad de la información en la Base de Conocimiento sobre la que opere

KIM es una de las propuestas más completas publicadas hasta la fecha, en nuestro conocimiento, para la construcción de BCs y la anotación automática a gran escala. Nuestro trabajo complementa al de KIM y TAP con un algoritmo de *ranking* específicamente diseñado para un modelo de recuperación basado en ontologías, utilizando un sistema de indexado semántico basado en la ponderación de las anotaciones.

Los llamados portales semánticos [2][3][7] típicamente proporcionan funcionalidades sencillas de búsqueda que más podrían caracterizarse como recuperación semántica de datos que como recuperación semántica de información. Las búsquedas devuelven instancias de una ontología más que documentos y por lo general no se proporciona un método de *ranking*. En algunos sistemas, se añaden enlaces a documentos referenciados por las instancias, junto a cada instancia devuelta en la respuesta a la consulta [3], pero ni las instancias ni los documentos están ordenados por relevancia.

El problema del *ranking* se ha retomado en [11] y más recientemente en [9]. Mientras que estos dos trabajos se ocupan de la ordenación de las respuestas a las consultas (i.e. instancias de las ontologías), nosotros abordamos la ordenación de los documentos anotados por dichas respuestas. Puesto que nuestras respectivas técnicas se aplican en fases consecutivas del proceso de recuperación, sería interesante experimentar con la integración de la función de relevancia de resultados propuesta por estos trabajos en nuestras medidas de relevancia de documentos.

Por último, compartimos con Mayfield y Finin [8] la idea de que la búsqueda semántica sea un complemento de la búsqueda por palabra clave mientras no haya suficientes ontologías y metadatos disponibles. Igual que ellos, utilizamos la inferencia para completar el conocimiento y explotar información implícita en las BCs.

3. Base de conocimiento y base documental

En nuestra aproximación a la recuperación semántica de información, suponemos que una BC ha sido construida y asociada a las fuentes de información (la base documental), utilizando una o varias ontologías de dominio que describen conceptos que aparecen en el texto de los documentos. Nuestro sistema puede funcionar con cualquier ontología del dominio, esencialmente sin restricciones, excepto algunos requisitos menores, que básicamente consisten en adoptar un conjunto de clases raíz para las ontologías. Éstas se muestran en la **figura 1**.

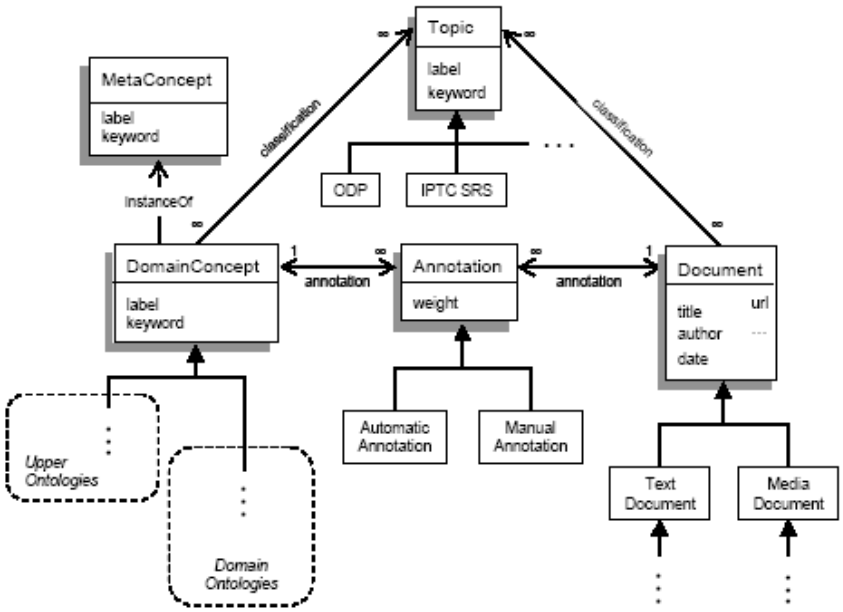

Figura 1. Clases raíz de las ontologías.

Los conceptos e instancias de la BC se asocian a los documentos mediante anotaciones explícitas, almacenadas externamente a los documentos. Aunque no abordamos aquí el problema de la extracción de conocimiento a partir de textos [3][6] o contenidos multimedia, proporcionamos un vocabulario y algunos mecanismos sencillos para ayudar a la anotación semi-automática de documentos de texto. El procedimiento de anotación automática está basado en un 'mapeo', o correspondencia, de los conceptos e instancias del dominio, definidos en la BC, a palabras clave (cadenas de texto), de modo similar a otros sistemas como KIM [6] y TAP [5]. Este 'mapeo' es utilizado por nuestro anotador automático para encontrar apariciones de conceptos e instancias en los textos, en cuyo caso se crea una anotación (un enlace bidireccional entre el concepto y el documento). Por supuesto, se utilizan técnicas más elaboradas para tratar las complejidades (polisemia, etc.) de este proceso (véase [12] para una explicación más detallada).

Las anotaciones son utilizadas por el módulo de recuperación y *ranking*, como se explicará en la siguiente sección. El algoritmo de *ranking* está basado en una adaptación del modelo de espacio vectorial [10]. En el modelo vectorial, se asigna un peso a las palabras clave que aparecen en un documento, reflejando que algunas palabras son mejores que otras para discriminar documentos entre sí. De forma similar, en nuestro sistema, se asigna un peso a las anotaciones, que refleja cuán importante se estima que es una instancia para el significado del documento al que anota. Los pesos se computan automáticamente mediante una adaptación del algoritmo TF-IDF [10], basada en la frecuencia de aparición de las instancias en cada documento. Más concretamente, el peso d_x de una instancia x para un documento d se calcula como:

$$d_x = \frac{frec_{x,d}}{\max_y frec_{y,d}} \cdot \log \frac{|D|}{n_x}$$

Donde $frec_{x,d}$ es el número de apariciones de x en d, $\max_y frec_{y,d}$ es la frecuencia de la instancia que más se repite en d, n_x es el número de documentos anotados por x, y D es el conjunto de todos los documentos en el espacio de búsqueda. El número de apariciones de una instancia en un documento se determina mediante el 'mapeo' concepto / palabras clave antes mencionado. Se remite al lector a [12] para más detalles sobre este aspecto.

4. Procesamiento de consultas y *ranking* de los resultados

Nuestra aproximación a la recuperación de información basada en ontologías puede verse como una evolución de las técnicas clásicas de recuperación basada en palabras clave, donde el índice de claves se substituye por una base de conocimiento semántica. El proceso global de recuperación se ilustra en la **figura 2**. Nuestro sistema toma como entrada una consulta formal en RDQL (*RDF Data Query Language*). Ésta se podría generar a partir de una consulta por palabras clave [5][9], una consulta en lenguaje natural [3], una interfaz basada en formularios [7], u otras técnicas de interfaz [4], pero no abordaremos aquí este aspecto. La consulta RDQL se ejecuta contra la BC, que devuelve una lista de tuplas de instancias que satisfacen la consulta. Finalmente, los documentos anotados con estas instancias son recuperados, ordenados, y presentados al usuario.

La consulta RDQL puede expresar condiciones que involucren tanto instancias de la ontología de dominio como propiedades de los documentos (tales como *autor*, *fecha*, *editor*, etc.). La ejecución de la consulta devuelve un conjunto de tuplas que la satisfacen. Es función del recuperador de documentos obtener todos los documentos que corresponden a las tuplas de instancias. Si las tuplas sólo están formadas por instancias del dominio, el recuperador sigue todos los enlaces salientes de anotación desde las instancias y recopila todos los documentos del repositorio anotados con estas instancias. Si las tuplas contienen instancias de clases de documentos (porque la consulta incluía condiciones directas sobre los documentos), se sigue el mismo procedimiento, pero restringido a los documentos de las tuplas, en lugar de todo el repositorio.

Nuestro sistema aplica, además, mecanismos de inferencia para llevar a cabo una expansión implícita de la consulta, basada en jerarquías de clases (p. ej., pigmentos orgánicos pueden satisfacer una consulta sobre colorantes), y reglas (p. ej., una que permita inferir el ganador de un encuentro deportivo a partir del tanteo). Una vez formada la lista de documentos, el motor de búsqueda calcula un valor de similitud semántica entre la consulta y cada documento, como sigue. Sea O el conjunto de todas las clases e instancias de la ontología, y Δ el conjunto de todos los documentos del espacio de búsqueda. Sea q una consulta RDQL, y sea V_q el conjunto de variables en la cláusula SELECT de q. Sea

$$T_q \subset O^{|V_q|}$$

la lista de tuplas en el conjunto resultado de la consulta, donde para cada tupla $t \in T_q$ y cada $v \in V_q$, se tiene $t_v \in$ O. Representamos cada documento del espacio de búsqueda como un *vector de documento* $d \in D$, donde d_x es el peso de la anotación del documento por el concepto x para cada $x \in$ O, si tal anotación existe, y cero en otro caso. Definimos el *vector extendido de consulta* q como aquél dado por

$$d_x = \frac{freq_{x,d}}{\max_y freq_{y,d}} \cdot \log\frac{|\mathcal{D}|}{n_x}$$

es decir, la coordenada del vector de consulta correspondiente a x es el número de variables en la consulta RDQL para las que existe una tupla t donde la variable es instanciada por x. Si x no aparece en ninguna tupla, asignamos $q_x = 0$. Finalmente, la medida de similitud entre un documento d y la consulta q se calcula como[1]:

$$\mathrm{sim}(d,q) = \frac{d \bullet q}{|d| \cdot |q|}$$

Allí donde el conocimiento de la BC sea incompleto, el algoritmo de *ranking* semántico tendrá muy poca efectividad: las consultas RDQL devuelven menos resultados de lo esperado, y los documentos relevantes no serán recuperados, o recibirán un valor de similitud mucho menor del que debieran. Por limitada que pueda ser, la búsqueda por palabra clave puede funcionar mejor en estos casos. Es por ello que nuestro modelo de *ranking* combina la medida de similitud semántica con la de un algoritmo basado en palabras clave. El valor de *ranking* final se calcula como $s \cdot \mathrm{sim}(d,q) + (1 - s)\,\mathrm{ksim}(d,q)$, donde ksim es el valor calculado por el algoritmo de palabra clave, y hemos fijado $s = 0.5$ empíricamente.

5. Pruebas experimentales

Hemos probado el sistema sobre un corpus de 145.316 documentos de la CNN, <http://dmoz.org/News/Online_Archives/CNN.com>. Como ontología de dominio hemos utilizado la ontología KIM [6] y su BC, disponibles públicamente en <http://www.ontotext.com/kim>, con algunas extensiones y ajustes menores. La BC resultante incluye 281 clases, 138 propiedades, 35.689 instancias y 465.848 sentencias. A partir del 'mapeo' concepto / palabras clave disponible en KIM, hemos creado automáticamente más de tres millones de anotaciones por el procedimiento mencionado en la **sección 3**.

El algoritmo de recuperación se ha probado sobre un conjunto de consultas, comparando los resultados con una búsqueda basada sólo en palabras clave, utilizando la librería de Jakarta Lucene, <http://lucene.apache.org>. La **figura 3** muestra una comparación promedio del rendimiento de nuestro sistema sobre veinte consultas, tales como "*bancos que cotizan en NASDAC, con un ingreso fiscal neto superior a dos billones de dólares*" y otras similares.

6. Discusión

El valor añadido de la recuperación semántica de información con respecto a la tradicional basada en palabras clave, tal como se contempla en nuestro enfoque, radica en la información explícita adicional (tipo, estructura, jerarquía, relaciones, reglas) almacenada en la BC sobre los conceptos referenciados en los documentos, frente a los índices planos clásicos por palabras clave. La búsqueda semántica introduce un paso adicional con respecto a los modelos clásicos de recuperación de información: en lugar del barrido de un índice simple por palabras, la búsqueda semántica procesa una consulta semántica contra la BC, y devuelve un conjunto de instancias. Esto puede verse como una forma de expansión de la consulta, donde las instancias devueltas representan un nuevo conjunto de términos de búsqueda, que conduce a un mayor nivel de recuperación (en inglés, *recall*). Esta expan-

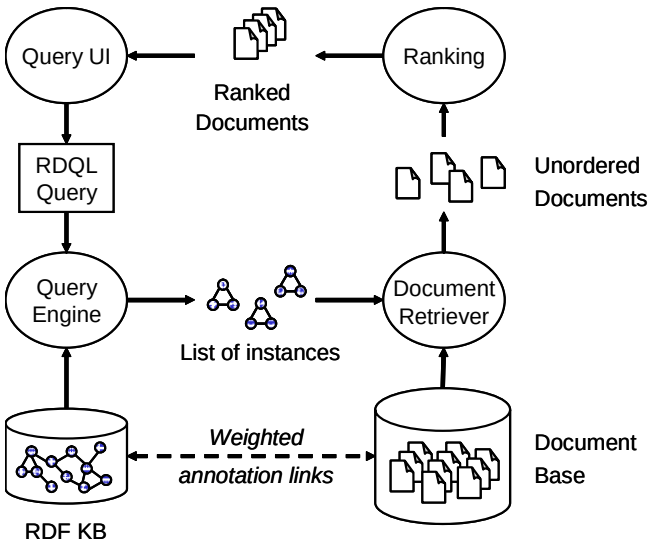

Figura 2. Nuestra visión de la recuperación de información basada en ontologías.

sión de la consulta se amplía también mediante la inferencia sobre reglas, jerarquías y relaciones. En resumen, nuestra propuesta da lugar a las siguientes mejoras con respecto a la búsqueda por palabra clave:

■ Mayor recuperación en las consultas por clase. Por ejemplo, la consulta "*Empresas británicas que cotizan en el NYSE*" devolvería documentos que mencionan a *Barclays*, *Vodafone*, y otras empresas como éstas, aún cuando las palabras "británica" y "NYSE" no estén presentes en los documentos.

■ Mejor precisión por la utilización de consultas semánticas estructuradas. Éstas permiten describir necesidades de información más precisas, dando lugar a respuestas más específicas. Por ejemplo, en un sistema basado en palabras clave, no es fácil distinguir la consulta "*jugadores estadounidenses en equipos europeos*" de la consulta *"jugadores europeos en equipos estadounidenses*", lo que en cambio sería sencillo con una consulta semántica.

■ Mayor recuperación mediante la utilización de jerarquías de clases y reglas. Por ejemplo, una consulta por *DeportesAcuáticos* en *España* puede devolver resultados sobre *Submarinismo*, *Windsurf*, y otras subclases, en *Cádiz*, *Málaga*, *Almería*, y otros lugares en España (por transitividad de la relación *localizadoEn*).

■ A pesar de la separación del espacio de contenidos (documentos) y el espacio de conceptos, es posible combinar condiciones sobre conceptos y contenidos. Por ejemplo, en la consulta "*críticas de cine publicadas en el año en curso sobre películas japonesas de ciencia ficción*", las condiciones de tipo (crítica de cine) y fecha (año en curso) se aplican al documento, mientras que el resto se refiere a un concepto (una película), que no pertenece al espacio de los documentos, pero que anota el documento.

■ Las mejoras de nuestro método con respecto a la búsqueda por palabra clave crecen con el número de condiciones en (i.e. la especificidad de) la consulta formal. Esto no es sorprendente, dado que cuanto más compleja es la necesidad de información, más información se pierde en su trascripción por palabras clave.

■ El grado de mejora de nuestro modelo de recuperación semántica depende de la completitud y calidad de la ontología, la BC, y el 'mapeo' de conceptos a palabras. Para asegurar la robustez, el sistema recurre a la búsqueda por palabra clave cuando la BC devuelve insuficientes resultados.

La combinación del *ranking* por palabra clave y por semántica es un aspecto delicado. Hemos observado que ocasionalmente una buena puntuación por *ranking* semántico se deteriora por un valor bajo por palabras clave. Una solución sencilla sería establecer un umbral mínimo para que la relevancia por palabra clave se tenga en cuenta. De cualquier modo, estos casos, aunque infrecuentes, sugieren que es preciso investigar métodos más elaborados que la simple combinación lineal de ambos valores, para mejorar nuestros resultados iniciales.

7. Conclusión

Nuestro enfoque se puede ver como una evolución del modelo vectorial clásico, donde los índices por palabras se reemplazan por una BC basada en ontologías, y un método semi-automático de anotación ponderada es el equivalente del proceso extracción de palabras clave e indexación. Hemos mostrado que es posible desarrollar un algoritmo de *ranking* consistente sobre esta base, dando lugar a mejoras medibles con respecto a la búsqueda por palabra clave, sujetas a la calidad y masa crítica de los metadatos. Nuestra propuesta hereda todos los problemas, bien conocidos, de la construcción y compartición de ontologías, el poblado de grandes BCs, y el 'mapeo' de palabras a conceptos. La investigación reciente en estas áreas está consiguiendo resultados prometedores [6]. Es nuestro objetivo proporcionar un modelo consistente por el cual cualquier avance en estos problemas se traduzca en mejoras para la búsqueda semántica.

Existe un amplio espacio para la mejora y la investigación más allá de nuestros resultados actuales. Por ejemplo, nuestro modelo de ponderación de las anotaciones no aprovecha aún las diferencias de relevancia de los campos de documentos estructurados (p.e. *título* es más importante que *cuerpo*). La anotación de documentos con sentencias, además de instancias, es otra posibilidad interesante por explorar. Asimismo, estamos extendiendo actualmente nuestro modelo con un perfil de intereses del usuario para dar lugar a una búsqueda personalizada [1].

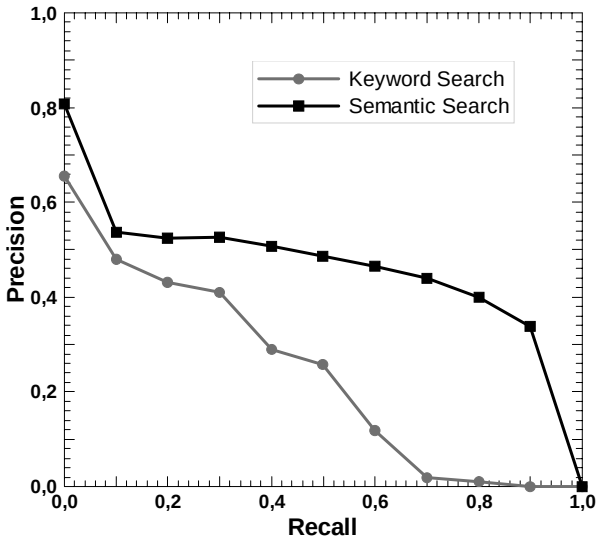

Figura 3. Curva de precisión promedio vs. recuperación para una batería de veinte. consultas.

Referencias

[1] P. Castells, M. Fernández, D. Vallet, P. Mylonas, Y. Avrithis. Self-Tuning Personalized Information Retrieval in an Ontology-Based Framework. *1st IFIP International Workshop on Web Semantics* (SWWS 2005). LNCS Vol. 3532 (2005) 455-470.

[2] P. Castells, F. Perdrix, E. Pulido, M. Rico, V.R. Benjamins, J. Contreras, J. Lorés. Neptuno: Semantic Web Technologies for a Digital Newspaper Archive. *1st European Semantic Web Symposium* (ESWS 2004). LNCS Vol. 3053 (2004) 445-458.

[3] J. Contreras, V.R. Benjamins, et al. A Semantic Portal for the International Affairs Sector. 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004). LNCS Vol. 3257 (2004) 203-215.

[4] E. García-Barriocanal, M.A. Sicilia. User Interface Tactics in Ontology-Based Information Seeking. *Psychnology e-journal* 1:3 (2003) 243-25.6

[5] R.V. Guha, R. McCool, E. Miller. Semantic search. *12th International World Wide Web Conference* (WWW 2003). Budapest, Hungary (2003) 700-709.

[6] A. Kiryakov et al: Semantic Annotation, Indexing, and Retrieval. *Journal of Web Sematics* 2:1 (2004) 49-79.

[7] A. Maedche et al. SEmantic portAL: The SEAL Approach. In: Fensel et al (eds.): *Spinning the Semantic Web*. MIT Press, Cambridge London (2003) 317-359.

[8] J. Mayfield, T. Finin. Information retrieval on the Semantic Web: Integrating inference and retrieval. Workshop on the Semantic Web at the *26th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2003). Toronto, Canada, 2003.

[9] C. Rocha, D. Schwabe, M.P. de Aragão. A Hybrid Approach for Searching in the Semantic Web. *International World Wide Web Conference* (WWW 2004), New York (2004) 374-383.

[10] G. Salton, M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[11] N. Stojanovic, R. Studer, L., Stojanovic. An Approach for the Ranking of Query Results in the Semantic Web. *2nd International Semantic Web Conference* (ISWC 2003). LNCS Vol. 2870 (2003) 500-516.

[12] D. Vallet, M. Fernández, P. Castells. An Ontology-Based Information Retrieval Model. *2nd European Semantic Web Conference* (ESWC 2005). LNCS Vol. 3532 (2005) 455-470.

Nota

[1] Por concisión, omitimos aquí algunos detalles técnicos menores: pesos de las variables RDQL, factores de normalización, funciones de corrección, etc.