

Proyecto de trabajo de Iniciación a la investigación

Miriam Fernández

Tutor: Pablo Castells Azpilicueta

Universidad Autónoma de Madrid, Escuela Politécnica Superior
Ciudad Universitaria de Cantoblanco, c/ Tomás y Valiente 11, 28049 Madrid
miriam.fernandez@uam.es

Resumen. Uno de los principales objetivos perseguidos dentro del campo de la Web Semántica radica en la mejora de las actuales técnicas de recuperación de información mediante el uso de nuevas metodologías englobadas bajo el nombre Búsqueda Semántica. Este trabajo se centra en la implementación de un nuevo modelo de búsqueda semántica enfocado a la recuperación de información sobre grandes repositorios de documentos. El modelo consta fundamentalmente de tres partes diferenciadas: Un modelo de anotación semi-automático que enlaza las entidades semánticas con los documentos donde aparecen, un modelo de recuperación de información que devuelve los documentos que se adaptan a las necesidades del usuario, y un modelo de ranking que se encarga de gestionar la ordenación final de documentos en base a diferentes criterios.

1 Introduction

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 consiste en 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 [10] 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 para 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, este trabajo pretende dar un paso más allá. En esta 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 canti-

dad 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 propósitos, es a menudo apropiado mantener los documentos originales en el sistema. En tercer lugar, allí donde los valores de las ontologías contienen texto libre, la búsqueda booleana realiza una búsqueda en texto completo dentro de las cadenas de caracteres, y por ende, en la medida en que estos fragmentos sean de mayor tamaño, la hipótesis de una “equiparación exacta” se vuelve discutible. En estas condiciones, si no se proporciona un criterio de ranking claro, el sistema de búsqueda puede resultar inoperante en la práctica si el espacio de búsqueda es demasiado grande.

En este trabajo se propone un modelo de recuperación de información orientado a la explotación de ontologías del dominio y bases de conocimiento de pleno desarrollo, para dar soporte a la búsqueda semántica en grandes repositorios documentales [12]. En contraste con los sistemas booleanos de búsqueda semántica, en esta perspectiva se devuelven documentos, más que (o además de) valores de ontología específicos de una Base de Conocimiento, en respuesta a las necesidades de información del usuario. Para soportar fuentes de información de gran escala, se propone una adaptación 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 este modelo estará en relación directa con la cantidad y calidad de la información en la Base de Conocimiento sobre la que opere. Los últimos avances en la automatización y poblado de ontologías y la anotación semi-automática de textos son prometedores. Mientras, la falta o incompletitud de las ontologías y Bases de Conocimiento disponibles será una limitación que muy probablemente se tendrá que admitir a medio plazo. En consecuencia, la tolerancia a Bases de Conocimiento incompletas se establece como un importante requisito en esta propuesta.

2 Trabajo Relacionado

La visión del problema de recuperación semántica propuesta en este trabajo es muy próxima a las propuestas de KIM [4 5]. Sin embargo, mientras que KIM se centra en el poblado de ontologías y la anotación automática de textos, esta propuesta se centra en los algoritmos de ranking para la búsqueda semántica. Junto con TAP [1], KIM es una de las propuestas más completas publicadas hasta la fecha, para la construcción de Bases de Conocimiento y la anotación automática a gran escala. Este 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 centrado en la ponderación de anotaciones o enlaces entre los conceptos de las Bases de Conocimiento y los documentos almacenados en el repositorio.

Los llamados portales semánticos [6] 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 [1], 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), este trabajo pretende abordar la ordenación de los documentos anotados por dichas respuestas.

Por último, esta propuesta comparte con Mayfield y Finin [7] 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, se utilizará la inferencia para completar el conocimiento y explotar información implícita en las BCs.

3 Objetivos

En el modelo de recuperación de información semántica propuesto se asume la existencia de Bases de Conocimiento asociadas a las fuentes de información o repositorio de documentos. Se propone un sistema flexible que pueda trabajar con cualquier ontología de dominio con apenas restricciones, excepto por unos requisitos mínimos que básicamente consisten en la generación de un conjunto de clases raíz que permitirán realizar el proceso de anotación o indexación conceptual de forma semi-automática.

Aunque en este trabajo no se enfoca en el problema de la extracción de conocimiento de texto [1, 4, 5] se propondrán técnicas sencillas de anotación así como el uso de herramientas actualmente existentes en el área de la Web Semántica. Las anotaciones se utilizarán posteriormente durante el proceso de recuperación de información y ranking de documentos. Para enfocar el ranking se pretende realizar una adaptación del modelo vectorial [10], como explicaremos a continuación. En el modelo vectorial clásico se asigna un peso a las palabras que aparecen dentro de los documentos reflejando la capacidad de discriminación o identificación de un documento mediante el uso de una determinada palabra clave. Esta idea se adquiere en el modelo propuesto donde a las anotaciones se les asigna un peso que refleja en qué grado la instancia o el concepto representa la semántica del documento. Estos pesos se calculan automáticamente mediante una adaptación del algoritmo TF-IDF [10], basado en la frecuencia de las instancias en cada uno de los documentos, ponderándolo a la frecuencia máxima dentro repositorio [12].

Esta propuesta de recuperación de información puede verse como una evolución del clásico modelo vectorial, solo que en lugar de indexar por palabras clave, se indexa por conceptos e instancias representadas en Bases de Conocimiento. El proceso completo de recuperación de información puede verse en la Fig. 1.

El sistema toma como entrada una query formal, expresada en lenguajes como por ejemplo RDQL. Esta query puede generarse mediante una consulta basada en palabras clave [3, 9], una consulta basada en lenguaje natural [1], una interfaz de formulario [6], o técnicas de interfaz de usuario más sofisticadas. Una vez obtenida la consulta se procede a recuperar la información que mejor se adapta a las necesidades del usuario. Este proceso puede verse representado en dos fases principales: en una primera fase la consulta formal se ejecuta contra una Base de Conocimiento y se devuelve una lista de instancias o tuplas que cumplen los requisitos de la consulta. Seguidamente, en una segunda fase, se utilizan las anotaciones de dichas instancias con los documentos del repositorio para recuperar el conjunto de documentos que satisfacen la consulta del usuario.

Tras este proceso de recuperación, los documentos son ordenados y presentados al usuario siguiendo una adaptación del modelo vectorial que utiliza los pesos de las anotaciones para dilucidar el orden final y presentar al usuario en primer lugar aquellos documentos que contienen la semántica que mejor responde a su necesidad de información.

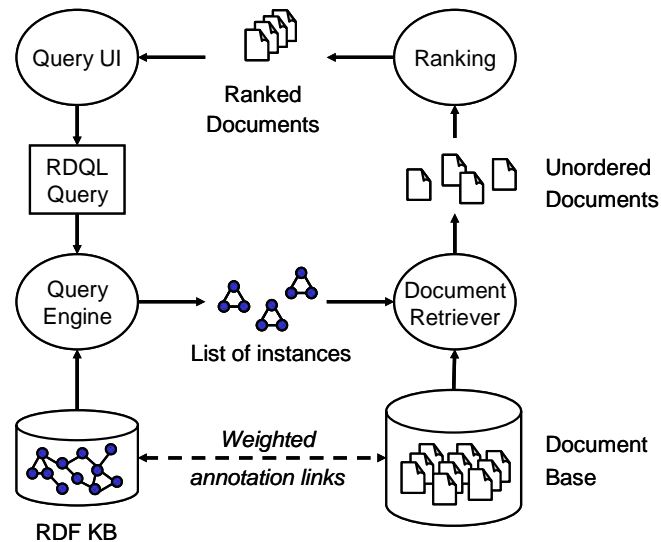


Fig.1. Vista del modelo de Recuperación de Información basado en Ontologías

4 Contribuciones Esperadas

El modelo de este trabajo puede verse como una evolución del modelo vectorial clásico, donde los índices basados en palabra clave son reemplazados por Bases de Conocimiento basadas en ontologías. El proceso de anotación y peso se equipara al proceso de extracción de palabras clave e indexado. El objetivo principal es mostrar que es posible el desarrollo de algoritmos de ranking consistentes en esta base que mejoran considerablemente los actuales algoritmos basados en palabra clave, considerando la calidad y el volumen de metadatos disponibles. Recientes investigaciones en estas áreas presentan resultados muy prometedores [4]. Mi principal objetivo es, por tanto, proveer un modelo consistente donde el avance de estos problemas pueda reflejarse en una mejora de la búsqueda semántica.

5 Metodología

Se pretende realizar una metodología de desarrollo cíclica, partiendo de una sólida investigación previa del estado del arte que permita detectar aquellos puntos en los que los modelos de búsqueda semántica necesitan mayor soporte.

Así mismo, se buscarán y/o generarán posibles Bases de Conocimiento, y repositorios de documentos sobre los que testear el rendimiento de la aplicación.

Para la evaluación de resultados se seleccionarán un grupo de consultas que abarquen un amplio conjunto de tópicos y escenarios. Los resultados de cada consulta serán evaluados por una colección de usuarios seleccionados según sus características.

Los algoritmos de recuperación de información y ranking serán contrastados con las búsquedas tradicionales por palabra clave utilizando la librería Lucene desarrollada por Yakarta y considerando medidas de evaluación clásicas como precisión y recall.

Este proceso se irá refinando progresivamente hasta la obtención de un prototipo final testeado con varias Bases de Conocimiento y Repositorios donde pueda obtenerse una evaluación precisa del rendimiento del nuevo modelo en contraste con técnicas ya conocidas.

Bibliografía

1. Castells, P., Fernández, M., Vallet, D., Mylonas, P., Avrithis, Y.: 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. Contreras, J., Benjamins, V. R., 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
3. Guha, R. V., McCool, R., and Miller, E.: Semantic search. 12th International World Wide Web Conference (WWW 2003). Budapest, Hungary (2003) 700-709
4. Kiryakov, A., Popov, B., Terziev, I., Manov, Ognyanoff, D.: Semantic Annotation, Indexing, and Retrieval. Journal of Web Semantics 2:1 (2004) 49-79
5. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A.: KIM – A Semantic Platform for Information Extaction and Retrieval. Journal of Natural Language Engineering 10:3-4 (2004) 375-392
6. Maedche, A., Staab, S., Stojanovic, N., Studer, R., Sure, Y.: SEMantic portAL: The SEAL Approach. In: Fensel, D., Hendler, J. A., Lieberman, H., Wahlster, W. (eds.): Spinning the Semantic Web. MIT Press, Cambridge London (2003) 317-359
7. Mayfield, J., Finin, T.: 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)
8. M. Fernández, D. Vallet, P. Castells. Probabilistic Score Normalization for Rank Aggregation. 28th European Conference on Information Retrieval (ECIR 2006). London, April 2006. Springer Verlag Lecture Notes in Computer Science, Vol. 3936, pp. 553-556.
9. Rocha, C., Schwabe, D., de Aragão, M. P.: A Hybrid Approach for Searching in the Semantic Web. International World Wide Web Conference (WWW 2004), New York (2004) 374-383
10. Salton, G., McGill, M. Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
11. Stojanovic, N., Studer, R., Stojanovic, L.: 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. Vallet, D., Fernández, M., Castells, P.: An Ontology-Based Information Retrieval Model. 2nd European Semantic Web Conference (ESWC 2005). LNCS Vol. 3532 (2005) 455-470

Propuesta de tribunal de Evaluación

- Pablo Castells Azpilicueta
- Enrique Alfonseca Cubero
- O'Donnell Mick