IR+SW: Un sistema de recuperación de información extendido semánticamente

Oswaldo Solarte Pabón*, Martha Millán †

Universidad del Valle, Cali Colombia
*oswaldo.solarte@correounivalle.edu.co, †martha.millan@correounivalle.edu.co

Palabras clave: recuperación de información (*IR*), Web semántica, ontologías, anotaciones semánticas.

Resumen

En este artículo se describe IR+SW (*Information Retrieval and Semantic Web*), un sistema de recuperación de información extendido semánticamente. Este sistema está orientado a mejorar la recuperación de documentos en una biblioteca digital de ciencias de la computación. IR+SW soporta anotación semántica de documentos basándose en una ontología de dominio. En la recuperación de documentos se integran técnicas clásicas de recuperación de información y anotaciones semánticas.

1 Introducción

Muchos sistemas de búsqueda de información se han desarrollado basándose en los modelos clásicos de recuperación de información**.** En estos modelos, los documentos se representan como un conjunto de términos o palabras clave [1]. Los sistemas desarrollados bajo estos modelos le permiten al usuario buscar información mediante palabras clave. Dada una consulta, se retorna un conjunto de documentos que podría satisfacer las necesidades de información de los usuarios [2]. Sin embargo, una de las principales desventajas de estos sistemas es que no tienen en cuenta la semántica de los documentos y por lo tanto, los resultados de una consulta se limitan sólo a la frecuencia de aparición de los términos en los documentos [3].

Una de las propuestas para mejorar esta limitación es usar tecnologías la Web semántica. La Web semántica es una extensión de la Web actual, en la que los datos tienen un significado bien definido, facilitando a las computadoras y a las personas trabajar en cooperación [4]. Una de las ventajas de esta tecnología es que se tiene en cuenta el significado de los términos en los documentos. Por tanto, la búsqueda considera también aspectos semánticos de los datos, de forma que se pueden mejorar los resultados en una consulta.

En este artículo se describe un prototipo de sistema de recuperación de información extendido con tecnologías de la Web semántica. En este prototipo, el proceso de recuperación de información se enriquece semánticamente mediante anotaciones basadas en una ontología de dominio. En IR+SW, la recuperación de información tiene en cuenta tanto las palabras clave expresadas en la consulta del usuario como también su significado, el cual se representa por medio de anotaciones semánticas obtenidas a partir de una ontología. El resto del artículo está organizado así: en la sección 2, se describen los trabajos relacionados, en la sección 3 se presenta un modelo conceptual en el que se basa el sistema propuesto. En la sección 4, se describe detalladamente la implementación del sistema y en sección 5 se presentan las pruebas realizadas y los resultados obtenidos.

2 Trabajos relacionados

De acuerdo con [5], los sistemas de búsqueda semántica se pueden clasificar en dos categorías: los que se enfocan en la recuperación de instancias a partir de una ontología y los que se enfocan en la recuperación de documentos. IR+SW corresponde a la segunda categoría, y tiene como propósito mejorar la recuperación de información sobre documentos de texto. Los sistemas de búsqueda semántica orientados a mejorar la recuperación de documentos se pueden ver como una extensión de la recuperación de información tradicional. En estos sistemas, los documentos se anotan semánticamente con base en una ontología de dominio [6]. El proceso de recuperación se lleva a cabo haciendo coincidir las consultas de los usuarios con las anotaciones semánticas extraídas de los documentos.

Una de las primeras propuestas que busca mejorar el proceso de recuperación de información usando tecnologías de la Web semántica se presenta en [7]. En esta propuesta, los documentos se enriquecen con anotaciones semánticas que se obtienen automáticamente aplicando técnicas de extracción de información. Las anotaciones semánticas obtenidas en este proceso se almacenan en el mismo documento. Una consulta se puede expresar mediante palabras clave o usando el lenguaje de consulta *DQL* (*DAML+OIL Query Language*).

En [8] y [9] se describe KIM, un *framework* que también usa un mecanismo automático de anotación semántica de documentos. Las anotaciones se representan como enlaces entre conceptos incluidos en el documento y clases de una ontología. A diferencia de la propuesta anterior, en *KIM* las anotaciones semánticas se almacenan en una base de conocimiento separada de los documentos y se representan mediante tripletas *RDF*[1]. Durante la etapa de anotación se usa

[1] http://www.w3.org/RDF/

una ontología liviana para hacer anotaciones de propósito general. Para facilitar la recuperación, los documentos se indexan adaptando la herramienta *Apache Lucene*[2]. En *KIM* cuando se lanza una consulta, primero se buscan las instancias de la ontología asociadas a los términos de la consulta, luego se recuperan los documentos relacionados con estas instancias.

En [10] y [11] se adapta el modelo de espacio vectorial para permitir la búsqueda semántica de documentos. Además de anotar semánticamente los documentos, en este trabajo se utiliza un algoritmo de *ranking* para calcular el grado de relevancia de las anotaciones semánticas con respecto a los documentos. El grado de relevancia de una anotación depende de la frecuencia de aparición de la clase de la ontología con la cual se anotó el documento. A diferencia del modelo clásico de espacio vectorial, el *ranking* se aplica sobre las anotaciones y no sobre términos. El sistema toma como entrada una consulta expresada en *SPARQL*[3] y retorna una lista de instancias de la ontología. A partir de estas instancias, se expande la consulta explorando las jerarquías de clase en la ontología. Finalmente, los documentos anotados con las instancias de la ontología se recuperan y se ordenan usando un algoritmo de *ranking* que calcula la similitud semántica entre la consulta y un documento. De acuerdo con los autores, la búsqueda semántica mejora los resultados de la búsqueda basada en técnicas clásicas de *IR* en términos de *precison* y *recall*. Sin embargo, la búsqueda semántica puede fallar cuando las anotaciones semánticas son incompletas y no cubren toda la información de un documento.

Por su parte, en [12] y [13] se aplica el concepto de búsqueda híbrida para combinar resultados de la búsqueda basada en técnicas clásicas de *IR* con resultados de la búsqueda basada en anotaciones semánticas. La búsqueda basada en técnicas clásicas de *IR* tiene en cuenta únicamente la frecuencia de aparición de las palabras clave. Por otro lado, la búsqueda basada en anotaciones semánticas puede fallar cuando las anotaciones semánticas son incompletas, la ontología utilizada como base de anotación no cubre toda la semántica de un documento o cuando hay errores en la anotación, si ésta se hace manualmente. La búsqueda híbrida trata estos problemas usando un mecanismo de *ranking* que asigna un nivel de importancia tanto a la búsqueda basada en técnicas clásicas de *IR* como a la basada en anotación semántica.

Tanto en [12] como en [13], se ofrece al usuario un mecanismo de consulta que le permite navegar por la estructura jerárquica de la ontología. El usuario debe seleccionar manualmente las clases de la ontología con las cuales se orientará el proceso de búsqueda de documentos. Este mecanismo de consulta puede ser una desventaja para el usuario ya que, éste necesita invertir una gran cantidad de tiempo seleccionando las clases, además de que se requiere conocer la estructura de la ontología.

[2] http://lucene.apache.org/core/
[3] http://www.w3.org/TR/rdf-sparql-query/

3 Modelo Conceptual

Basándose en las ideas de [10], [12] y [13] se propone el modelo conceptual de la figura 1. Este modelo está enfocado en extender semánticamente el proceso de recuperación de información. Este modelo integra dos componentes: el procesamiento híbrido de documentos que se muestra en la parte superior de la figura y la búsqueda híbrida, en la parte inferior.

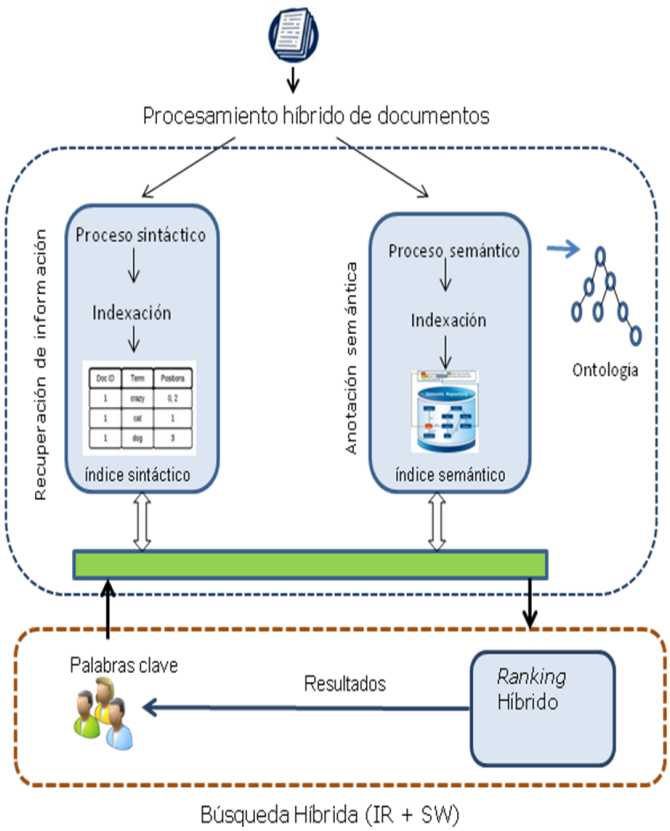

Figura 1. Modelo conceptual del prototipo

El procesamiento híbrido permite indexar los documentos tanto sintáctica como semánticamente. La búsqueda híbrida recupera documentos combinando técnicas clásicas de *IR* y anotaciones semánticas. En este modelo, el usuario expresa las consultas usando sólo palabras clave, no necesita conocer la estructura de la ontología, ni de lenguajes formales de consulta como *SPARQL* para acceder a las anotaciones semánticas.

3.1 Procesamiento híbrido de documentos

Este componente procesa sintáctica y semánticamente un conjunto de documentos de texto. El procesamiento sintáctico se basa en técnicas clásicas de *IR*. Estas técnicas involucran tareas como la división del texto en *tokens*, la eliminación de *stopwords* y la aplicación de algoritmos de *stemming*. Por su parte, el procesamiento semántico permite encontrar las relaciones que existen entre un documento y las clases de una

ontología. El resultado de este proceso es un conjunto de anotaciones semánticas que identifican al documento. El procesamiento semántico representa un documento como un conjunto de anotaciones semánticas. Cuando el procesamiento de documentos finaliza se crean dos índices: el índice sintáctico donde cada documento se representa como un conjunto de términos y el índice semántico, donde los documentos se representan mediante un conjunto de anotaciones semánticas.

3.2 Recuperación híbrida de documentos

En este componente, una consulta se procesa combinando técnicas clásicas de *IR* con técnicas basadas en anotaciones semánticas. Los resultados obtenidos separadamente se mezclan y se muestran al usuario usando un mecanismo de *ranking* híbrido. El *ranking* híbrido se calcula mediante la fórmula (1), combinando el grado de relevancia obtenido en la búsqueda basada en técnicas clásicas de *IR* (***ir-score)*** con el grado de relevancia obtenido en la búsqueda basada en anotaciones semánticas (***semantic-score***).

$$\text{hybrid-score} = \lambda\,(\text{ir-score}) + \omega\,(\text{semantic-score}) \qquad (1)$$

Los factores λ y ω representan el grado de importancia de la búsqueda basada en *IR* y de la búsqueda basada en anotaciones semánticas, respectivamente. Los valores de λ y ω están entre 0.0 y 1.0. Si el valor para λ y ω es 0.5, significa que ambos tipos de búsqueda tienen la misma importancia. En IR+SW, los factores λ y ω se ajustan dependiendo de las condiciones de la consulta. Si la consulta se puede representar completamente con los conceptos de la ontología (clases, propiedades de anotación, instancias), se da mayor importancia a la búsqueda basada en anotaciones semánticas, y en este caso $\omega = 0.7$ y $\lambda = 0.3$. Por el contrario, si la consulta no se puede representar en su totalidad con los conceptos de la ontología, entonces el factor de importancia semántico tendrá un valor más bajo, $\omega = 0.4$ y $\lambda = 0.6$.

Los valores de λ y ω se obtuvieron a partir de pruebas supervisadas construidas sobre un conjunto de documentos y a partir de consultas de usuario representadas completa y parcialmente usando los conceptos de la ontología.

En modelo de la figura 1, el usuario expresa las consultas en forma de palabras clave. De esta forma se busca ocultar la complejidad y la estructura de la ontología al usuario. También se busca evitar que las consultas se expresen usando algún lenguaje formal de consulta porque esto podría ser complejo para el usuario. Según [14] y [15], los usuarios están acostumbrados a expresar sus necesidades de información usando interfaces de consulta sencillas y fáciles de usar, las cuales generalmente se basan en palabras clave.

4 Implementación del sistema

El prototipo IR+SW está integrado por 3 módulos: indexación sintáctica, anotación semántica y el módulo de consulta.

4.1 Indexación sintáctica de documentos

Este módulo utiliza técnicas clásicas de *IR* para el procesamiento de texto. Los documentos se representan usando el modelo de espacio vectorial, cada documento tiene asociado a un vector de términos. Para implementar este módulo se usó la herramienta *Apache Lucene*.

4.2 Anotación semántica de documentos

Durante el proceso de anotación semántica se implementó la ontología propuesta por ACM[4], que describe el dominio de ciencias de la computación, ésta se implementó en *Protegé*[5] usando el lenguaje *OWL*. Cada clase de la ontología representa un área de conocimiento en este dominio. Para cada clase se definieron varias propiedades de anotación: *english_name*, *spanish_name* y *related_content*. Las dos primeras se usan para asociar a cada clase una etiqueta tanto en inglés como en español, respectivamente. La propiedad *related_content* se usó para describir sinónimos o formas alternativas de representar textualmente los conceptos en el dominio. En la tabla 1 se muestra un ejemplo de las propiedades de anotación y sus valores, definidos para la clase "*#Association_rules*". Todos estos valores están relacionados semánticamente.

Propiedad	Valor
english_name	Association rules
spanish_name	Reglas de asociación
reletad_content	Algoritmo Apriori
reletad_content	Algoritmo FP-growth

Tabla 1: Propiedades de anotación definidas en la ontología

En IR+SW, la anotación semántica se hace en dos etapas: en la primera, se usa la herramienta *GATE*[6] de forma embebida para crear anotaciones automáticamente. Estas anotaciones crean enlaces entre los conceptos del dominio encontrados en el texto y las clases de la ontología. En la segunda etapa se usa la estructura jerárquica de la ontología para descubrir nuevas clases con las cuales se relacionan los documentos.

Etapa 1: En esta etapa las anotaciones se hacen usando *GATE*, se tiene en cuenta las propiedades de anotación y las instancias de clase para crear anotaciones semánticas. Es decir, se tienen en cuenta las diferentes formas de representar un concepto del dominio en el texto para crear enlaces entre los documentos y las clases de la ontología. Las anotaciones semánticas que se hagan con instancias de clase también crean enlaces entre el documento y la clase a la cual corresponde dicha instancia.

Como un documento se puede anotar varias veces con la misma clase de la ontología, en esta etapa también se calcula

[4] http://www.computer.org/portal/web/publications/acmtaxonomy
[5] http://protege.stanford.edu/
[6] http://gate.ac.uk/

la relevancia de las anotaciones para identificar cuáles clases son más importantes para cada documento. El cálculo de la relevancia de las anotaciones se basa en [10], donde se hace una adaptación del modelo clásico de espacio vectorial. A cada anotación se le asigna un valor que representa el grado de relevancia semántico con respecto a un documento. El grado de relevancia semántico se calcula de acuerdo con la frecuencia de las anotaciones que tiene un documento con respecto a una clase de la ontología. Una anotación semántica se representa mediante tres atributos: *doc_id*, *ontology_class* y *relevance.* El atributo *doc_id*, es el identificador del documento, *ontology_class* es la clase de la ontología con la cual se hizo la anotación y *relevance* es el grado de relevancia entre la clase y el documento. Después de calcular la relevancia, cada documento tiene sólo una anotación por cada clase de la ontología y su respectivo grado de relevancia. Finalmente, las anotaciones semánticas se almacenan en forma de tripletas *RDF* usando el *framework Jena*.

Etapa 2: En esta etapa se usa la estructura jerárquica de la ontología y el concepto de distancia semántica para descubrir nuevas clases con las cuales se relaciona un documento. De acuerdo con [16], la distancia semántica se puede entender como el número de saltos que se debe dar en la ontología para llegar de un nodo a otro. En este trabajo se asume que cada nodo representa una clase en la ontología.

Para descubrir las nuevas clases, se parte de las clases que se obtuvieron con la herramienta *GATE*, a las cuales se les llama clases base. A partir de las clases base, se explora la ontología y se obtienen las clases antecesoras a éstas, mediante la propiedad *rdf:subClassOf*. Con las nuevas clases descubiertas también se crean anotaciones semánticas a las cuales se les asigna un grado de relevancia que depende de la distancia semántica y del grado de relevancia de la clase base. Entre más pequeña sea la distancia que hay entre la clase base y la nueva clase descubierta, mayor será el grado de relevancia de la nueva anotación. Por el contrario, entre más grande sea la distancia semántica, menor será el grado de relevancia de la nueva anotación.

4.3 Módulo de consultas de usuario

En este módulo el usuario expresa una consulta en forma de palabras clave. El sistema procesa esta consulta en tres fases: en la fase 1, se recuperan los documentos basándose en técnicas clásicas *IR*, en la fase 2, la recuperación se hace con base en las anotaciones semánticas. Las dos fases anteriores se pueden ejecutar concurrentemente ya que se procesan sobre repositorios separados. En la fase 3, los resultados obtenidos en las dos primeras fases se mezclan usando un algoritmo de *ranking* híbrido y se muestran al usuario.

***Fase 1*:** A esta fase se le denomina búsqueda tradicional ya que se basa en las técnicas clásicas de *IR*. Los documentos se recuperan teniendo en cuenta únicamente la frecuencia de los términos de la consulta. El resultado es una lista de documentos ordenada con respecto a la frecuencia de los términos.

***Fase 2*:** En esta fase, la búsqueda de documentos se basa en las anotaciones semánticas y se hace en tres pasos: transformación de las palabras clave en un conjunto de clases de la ontología, búsqueda de los documentos anotados con estas clases y ordenamiento de los documentos recuperados de acuerdo al grado de relevancia de las anotaciones semánticas.

La transformación de palabras clave a clases de la ontología, se hace por medio de un índice de conceptos que almacena parejas de la forma (*class*, *text*). El elemento *class* corresponde a la *URI* de la clase en la ontología, el elemento *text* contiene el texto que se ha extraído de las propiedades de anotación que tiene la clase. El índice de conceptos se crea previamente haciendo un pre-procesamiento de la ontología y permite obtener automáticamente las clases que orientarán el proceso de búsqueda. Los usuarios expresan sus consultas mediante palabras clave, mientras que las anotaciones semánticas se almacenan en forma de tripletas *RDF*. Teniendo en cuenta que la forma de representar las consultas de usuario y las anotaciones semánticas son diferentes, fue necesario crear el índice de conceptos, que permite interpretar una consulta expresada en palabras clave y relacionarla con las clases de la ontología.

Por ejemplo, si la consulta expresada por el usuario es "*Reglas de asociación en minería de datos*", se obtienen las clases de la ontología que se muestran en la tabla 2. Las clases con grado de relevancia más alto son aquellas que están más relacionados con la consulta de usuario. El grado de relevancia se tendrá en cuenta al momento de calcular la similitud entre la consulta y los documentos recuperados. Las clases obtenidas se adicionan a un vector de consulta que se usará al momento de calcular la similitud. Si el usuario hubiera expresado esta consulta en inglés, se obtendrían las mismas clases de la ontología de la tabla 2. Esto gracias a las propiedades de anotación *english_name y spanish_name* que tiene cada clase.

Clase de la ontología	**Grado de relevancia**
#Association_rules	1.00
#Data_mining	0.90
#Mining_methods_and_algorithms	0.75
#Text mining	0,57
#Web_mining	0,52

Tabla 2. Transformación de una consulta a clases

Después de transformar las palabras clave a clases de la ontología se construye automáticamente una consulta *SPARQL*. Cada clase obtenida en el paso anterior se agrega a la clausula WHERE de la consulta. En la figura 2 se muestra una parte de la consulta generada. La consulta *SPARQL* retorna una lista de anotaciones semánticas que coinciden con

las clases de la ontología del vector consulta. Cada elemento de la lista contiene el identificador del documento, la clase de la ontología y el grado de relevancia de esta anotación semántica.

```
SELECT  ?Annot  docID  ?Weight WHERE{

 {?Annot   Uv:concept 'Association_rules '  .
   ?Annot  Uv:Weight  ?Weight.
   ?Annot  Uv:doc_id  ?docID .
 }
 UNION
{ ?Annot  Uv:concept 'Data_mining' .
  ?Annot  Uv:Weight  ?Weight.  ?
  ?Annot  Uv:doc_id  ?docID .
}
UNION
{ ?Annot   Uv:ontology_concept
           'Mining_methods_and_algorithms ' .
  ?Annot   Uv:Weight  ?Weight.
  ?Annot   Uv:doc_id  ?docID .
}
…
ORDER BY DESC(?Weight)
```

Figura 2: Consulta en SPARQL generada automáticamente

La recuperación basada en anotaciones semánticas expande las consultas por medio de las propiedades de anotación y de las instancias de clase. La consulta "*Reglas de asociación en minería de datos*" se expande semánticamente con palabras clave como "*Algoritmo Apriori*" y "*Algoritmo FP-growth*", ya que estos conceptos están relacionados semánticamente con la clase "*#Association_rules*", como se muestra en la tabla 1. La búsqueda basada en anotaciones semánticas recupera información teniendo en cuenta las clases de la ontología con las cuales se anotaron los documentos. Finalmente, para calcular la relevancia, tanto la consulta como los documentos se representan como vectores. Cada posición del vector corresponde a una clase de la ontología. La relevancia entre la consulta y los documentos se calcula por medio de la similitud coseno. En IR+SW también se expande la consulta con clases relacionadas en la ontología. Partiendo de las clases del vector consulta se buscan las clases que están relacionadas semánticamente en la ontología. Esta expansión permite ofrecer más posibilidades de búsqueda al usuario, como la búsqueda de documentos relacionados o la recomendación de documentos.

***Fase 3*:** En esta fase se combinan los documentos que se obtuvieron en la búsqueda basada en técnicas clásicas de *IR* con los documentos obtenidos en la búsqueda basada en anotaciones semánticas (Fase 1 y 2). Después de combinar los documentos se muestran al usuario. La combinación de documentos se realiza aplicando un mecanismo de *ranking* híbrido en el que se tiene en cuenta un factor de importancia para cada tipo de búsqueda, como se mostró en la fórmula 1.

5 Pruebas

El escenario de pruebas está compuesto por una colección de documentos, las consultas de usuario expresadas en palabras clave, la ontología de dominio y un conjunto de anotaciones semánticas. La colección de documentos consta aproximadamente de 1.200 ejemplares en diversas áreas de ciencias de la computación. La ontología de dominio contiene alrededor de 840 clases. Las pruebas incluyeron tanto consultas que se podían representar completamente con la información de la ontología como también aquellas que sólo se podían representar parcialmente o no se podían representar en la ontología. Cada consulta se ejecutó usando tres estrategias de búsqueda: la búsqueda tradicional (*IR based search*), la búsqueda basada en anotaciones semánticas (*Semantic based search*) y la búsqueda híbrida (*Hybrid based search*), que combina los resultados de las dos anteriores. En la figura 3 se muestra el resultado para el promedio de las consultas.

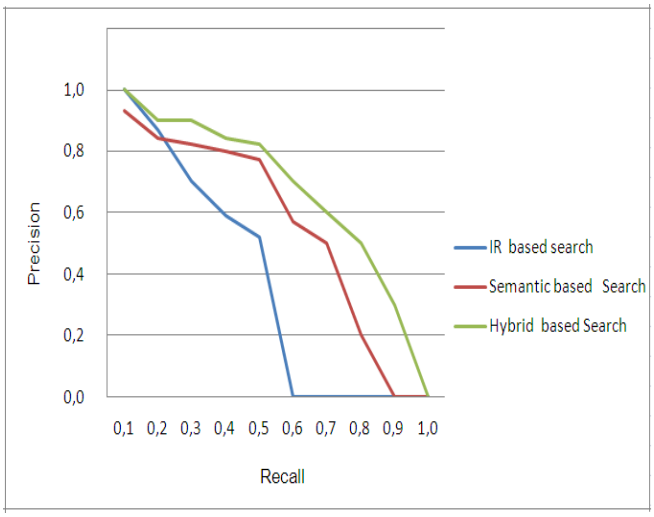

Figura 3: Resultados obtenidos, *Precision* Vs *Recall*

De acuerdo con la figura 3, la búsqueda tradicional tiene un desempeño inferior que la búsqueda basada en anotaciones semánticas y que la búsqueda híbrida. Por ejemplo, en la búsqueda tradicional, para niveles de *recall* cercanos a 0.5, la métrica *precision* también toma valores aproximados a 0.5. A partir de este nivel de *recall*, la métrica *precision* decrece rápidamente hasta llegar a cero. De otro lado, en la búsqueda basada en anotaciones semánticas, para valores de *recall* cercanos a 0.5 se obtuvieron valores de *precision* cercanos a 0.8. Además, en búsqueda semántica se obtienen mejores niveles de *recall* porque las consultas se expanden por medio de las propiedades de anotación y las instancias de clase de la ontología. En la búsqueda semántica se obtuvo valores de *recall* cercanos a 0.9 mientras que en la búsqueda tradicional sólo se obtienen valores cercanos a 0.6.

Por su parte, la búsqueda híbrida presenta un desempeño superior que aquella que se basa sólo en anotaciones

semánticas. La búsqueda basada en anotaciones semánticas funciona muy bien cuando el usuario expresa consultas que se pueden interpretar completamente con la información de la ontología, sin embargo, su rendimiento es inferior cuando no hay anotaciones semánticas que permitan interpretar completamente una consulta de usuario. La búsqueda híbrida funciona mejor en el promedio de los casos ya que aprovecha las ventajas tanto de la búsqueda tradicional como de la búsqueda basada en anotaciones semánticas. Esto se logra por que el factor de importancia semántico en el cálculo del *ranking* híbrido se configura dependiendo de las condiciones de la consulta de usuario.

6 Conclusiones y trabajo futuro

En este trabajo se presentó un sistema de recuperación de información extendido semánticamente. Este sistema presenta un mejor desempeño en términos de *precision y recall*, que un sistema que se basa sólo en técnicas clásicas de *IR*. Las anotaciones semánticas y la ontología de dominio son una parte fundamental de este sistema. Cuando la ontología de dominio es completa y cubre totalmente las necesidades de información del usuario, la búsqueda basada en anotaciones semánticas ofrece mejores resultados que la búsqueda tradicional. Sin embargo, si la ontología es incompleta la búsqueda semántica puede fallar porque las anotaciones no cubren toda la semántica de un documento. Por esta razón, esta propuesta se basa en el paradigma de la búsqueda híbrida que aprovecha las ventajas de recuperación de información clásica y de la recuperación basada en anotaciones semánticas.

Las ontologías ayudan a mejorar los resultados obtenidos en los sistemas de recuperación de información. Estas, generalmente se implementan en *OWL* o *RDF* y se consultan usando lenguajes como *SPARQL*. Como en IR+SW, el usuario expresa las consultas usando palabras clave, fue necesario crear un índice de conceptos para relacionar una consulta de usuario con las clases representadas en la ontología. El índice de conceptos permite a los usuarios acceder a las anotaciones semánticas y a la información representada en la ontología, sin necesidad de conocer lenguajes de consulta como *SPARQL*. Esta propuesta se diferencia de otros trabajos como [10], donde la consulta se expresa mediante *SPARQL*, lo cual puede representar un alto nivel de complejidad para los usuarios.

En IR+SW, se ofrece la posibilidad de buscar semánticamente documentos sin necesidad que el usuario conozca la estructura de la ontología que se usó durante el proceso de anotación. La recuperación de información es transparente para el usuario, éste sólo expresa un conjunto de palabras clave y el sistema selecciona automáticamente las clases de la ontología que orientarán el proceso de búsqueda. En este sentido, el prototipo presentado se diferencia de otros trabajos como [12] y [13], donde es el usuario quien debe conocer la estructura de la ontología y seleccionar manualmente las clases que se usarán en la búsqueda de documentos.

Como trabajo futuro se plantea incluir múltiples ontologías de dominio en el proceso de anotación y recuperación de documentos, también trabajar con ontologías que ofrezcan más relaciones, además de las jerárquicas, con el objeto de poder soportar consultas de usuario más complejas. Finalmente, se deben seguir explorando técnicas que permitan al usuario acceder a la información almacenada en las ontologías de manera usable y natural, sin que éste tenga que conocer de lenguajes formales de consulta.

Referencias

[1] R.A. Baeza-Yates, B.A Ribeiro-Neto, "Modern Information retrieval*",* *ACM Press/ Addison-Wesley*, (1999)

[2] I. TrivikRam. "Hybrid Approach to retrieving Web documents and Semantic data". *Phd. Dissertation*, Wright State University, (2007).

[3] W. Wei., P.M. Barnaghi, and A. Bargiela. "Semantic-Enhanced Information search and Retrieval", *Advanced Language Processing and Information Technology*, pp. 218-223, (2007).

[4] T. Lee, J. Hendler, O. Lassila., "The semantic web, Scientific American", (2001).

[5] G. Nagypal "Possibly imperfect ontologies for effective information retrieval". *PhD dissertation*, University of Karlsruhe, Germany, (2007).

[6] W. Wei, M. PAYAM, A. Bargiela. "Search with Meanings: An Overview of Semantic Search Systems". *International Journal of Communications of SIWN*, pp. 76-82 (2008).

[7] U. Shah, T. Finin, Joshi. Anupam, R. S.Cost, and J. Matfield. "Information retrieval on the semantic web". *Proceedings of the Eleventh Intern. conference on Information and knowledge management*, pp. 461-468. New York, NY, USA. (2002).

[8] B. Popov., A. Kiryakov. , D. Ognyanoff , D. Manov, A. Kirilov. "KIM – a semantic platform for information extraction and retrieval", *Journal of Natural Language Engineering*. Vol 10, No. 3-4, pp. 375-392, (2004).

[9] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, and K. Miroslav. "Semantic annotation, indexing, and retrieval." *Journal of web Semantics*, pp. 49-79. (2005)

[10] D. Vallet, M. Fernández, P Castells. "An Ontology-Based Information Retrieval model" *The Semantic Web: Research and Applications, Springer* , pp. 103-110, (2005).

[11] P. Castells, M. Fernández, D. Vallet, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval" *IEEE Transactions and Data Engineering*, vol. 19, no.2, pp. 261-272, (2007).

[12] R. Bhagdev, S. Chapman, F. Ciravegna, D. Lanfranchi Petrelli. "Hybrid Search: Effectively Combining keywords and Semantics Searches", *Springer Verlag Berlin ESWC* , pp. 554–568, (2008).

[13] N. Bikakis, Giannopoulos G., T. Dalamagas. "Integrating keywords and semantics on document annotation and search", *Move to Meaningful Internet Systems* Springer,pp. 921-938, (2010).

[14] T. Tran, P. Cimiano, S. Rudolph, R. Studer "Ontology-based interpretation of keywords for semantic search," *in ISWC/ASWC, pp. 523–536,* (2007).

[15] Y. Lei, Uren, V., E. Motta. "Semsearch: A search engine for thesemantic Web", *Conference 15th on Knowledge engineering and Knowledge Management* pp 238–245, (2006).

[16] Nesić S., Jazayeri M., Crestani, F. "Concept-Based Semantic Annotation Indexing and Retrieval of Office-Like Document units*", 9th International conference on of Heterogeneous Information, RIAO,* pp. 234–23, Paris (2010).