Revista EIA, ISSN 1794-1237 / Año XI / Volumen 11 / Edición N.22 / Julio-diciembre 2014 / pp. 51-65
Publicación semestral de carácter técnico-científico / Escuela de Ingeniería de Antioquia —EIA—, Envigado (Colombia)

PROPUESTA PARA EXTENDER SEMÁNTICAMENTE EL PROCESO DE RECUPERACIÓN DE INFORMACIÓN

Oswaldo Solarte Pabón[1]
Martha Elena del Socorro Millán González[2]

RESUMEN

En los modelos clásicos de recuperación de información los documentos se representan mediante un conjunto de términos o palabras clave. Una desventaja de esta representación es que los resultados de una consulta se limitan solo a la frecuencia de aparición de los términos. No se tiene en cuenta el significado de los términos ni las relaciones semánticas que puedan existir entre los documentos. Una de las alternativas para resolver este problema es usar tecnologías de la *web* semántica con el objetivo de asignarle a los datos un significado bien definido. En este artículo se describe una propuesta para extender el proceso de recuperación de información usando tecnologías de la *web* semántica. Los documentos se enriquecen semánticamente por medio de anotaciones que se obtienen a partir de una ontología de dominio. La recuperación de información extendida semánticamente tiene en cuenta tanto las palabras clave expresadas en la consulta del usuario como también su significado, el cual se representa mediante una ontología. La recuperación de información extendida semánticamente mejora los resultados en términos de *precision* y *recall* en comparación con los obtenidos en la recuperación de información clásica.

PALABRAS CLAVE: recuperación de información (IR); *web* semántica; ontologías; anotaciones semánticas.

PROPOSAL TO SEMANTICALLY EXTEND THE INFORMATION RETRIEVAL PROCESS

ABSTRACT

In classical information retrieval models, documents are represented by a set of terms or keywords. A disadvantage of this representation is that the query results are limited only to the frequency of occurrence of terms. No considers the meaning of terms and semantic relationships that may exist between the documents. One alternative to solve this problem is using Semantic Web technologies in order to assign data a well defined meaning. This article describes a proposal to extend the information retrieval process using Semantic Web technologies. Documents are semantically enriched with annotations that are obtained from domain ontology. The extended semantic retrieval takes into account both, keywords expressed in the user query as well, its meaning, which is represented by ontology. The extended semantic retrieval improves results in terms of precision and recall compared with those obtained in classic information retrieval.

KEYWORDS: Information Retrieval (IR); Semantic Web; Ontologies; Semantic Annotations.

[1] Ingeniero de sistemas. Magíster en Ingeniería de Sistemas y Computación de la Universidad del Valle. Profesor de la Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle.

[2] Licenciada en Matemáticas y Física. Magíster en Ingeniería Industrial y de Sistemas de la Universidad del Valle. Magíster en Ingeniería del Conocimiento y PhD. en Informática de la Universidad Politécnica de Madrid - España. Profesora de la Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle.

Autor de correspondencia: Solarte-Pabón, O. (*Oswaldo*). Universidad del Valle, Ed. 331, Segundo piso Ciudad Universitaria Meléndez, . Tel: (572) 321 22 83. Correo electrónico: oswaldo.solarte@correounivalle.edu.co

Historia del artículo:
Artículo recibido:22-IV-2013 / Aprobado: 19-VIII-2014
Disponible online: 30 de agosto de 2014
Discusión abierta hasta diciembre de 2015

DOI: http://dx.doi.org/10.14508/reia.2014.11.22.49-63 reia.2014.11.22.51-65

PROPOSTA PARA ESTENDER SEMÂNTICAMENTE O PROCESSO DE RECUPERAÇÃO DE INFORMAÇÃO

RESUMO

Em modelos convencionais de recuperação de informação os documentos são representados por um conjunto de termos ou palavras-chave. A desvantagem desta representação é que os resultados da consulta são limitados apenas para a frequência de ocorrência dos termos. Não é levado em conta o significado dos termos nem as relações semânticas que possam existir entre os documentos. Uma das alternativas para resolver este problema é a utilização de tecnologias da Web Semântica, com o objetivo de dar aos dados um significado bem definido. Este artigo descreve uma proposta para estender o processo de recuperação de informação utilizando tecnologias de web semântica. Os documentos são semanticamente enriquecidos por anotações que são derivados de uma ontologia de domínio. A recuperação de informação estendida semanticamente leva em conta tanto as palavras-chave expressas na consulta do usuário, bem como o seu significado, que é representado por uma ontologia. A recuperação da informação estendida semântica melhora os resultados em termos de precisão e recall comparados com os obtidos na recuperação de informação convencional.

PALAVRAS-CHAVE: Recuperação da informação (IR); Web semântica; Ontologias; Anotações semânticas.

1. INTRODUCCIÓN

Los modelos clásicos de recuperación de información se usan ampliamente como soporte para el desarrollo de herramientas de búsqueda**.** En estos modelos, los documentos se representan como un conjunto de términos (Baeza-Yates y Ribeiro-Neto, 1999). Los sistemas desarrollados con base en estos modelos le permiten al usuario buscar información mediante un mecanismo de consulta basado en palabras clave. Dada una consulta, se retorna un conjunto de documentos que en cierta forma satisface las necesidades de información del usuario (TrivikRam, 2007). Al momento de realizar una consulta, estos modelos se basan en la frecuencia de aparición de los términos para asignar cierta importancia a los documentos. Los resultados obtenidos se muestran en un orden de relevancia con respeto a los términos de consulta. A pesar de que se han construido muchos sistemas usando los modelos clásicos de recuperación de información, en estos sistemas se usa únicamente la frecuencia de aparición de los términos, sin tener en cuenta su significado (Wei-Wang, Barnaghi y Bargiela 2007).

Una de las propuestas para mejorar la efectividad de los resultados obtenidos en los sistemas de recuperación de información, incluye el uso de tecnologías la *web* semántica. La *web* semántica es una extensión de la *web* actual, en la cual se da un significado, bien definido a los datos, facilitando a las computadoras y las personas trabajar en cooperación (Lee, Hendler y Lassila, 2001). Una de las ventajas que ofrece la *web* semántica, es que se tiene en cuenta el significado de las palabras dentro de los documentos. De esta forma, al ofrecer la posibilidad de buscar información considerando aspectos semánticos de los datos, se pueden obtener mejores resultados en una consulta.

En este artículo se presenta una propuesta para extender el proceso de recuperación de información usando tecnologías de la *web* semántica. Los documentos se enriquecen semánticamente por medio de anotaciones que se obtienen a partir de una ontología de dominio. Las consultas de usuario se expanden a través de las propiedades de anotación e instancias definidas en la ontología. En el proceso de búsqueda de documentos se combina la recuperación de información clásica con la recuperación basada en anotaciones semánticas. La recuperación de información extendida con anotaciones semánticas mejora los resultados en términos de *precision* y *recall*, en comparación con los obtenidos en la recuperación de información clásica.

El resto del artículo está organizado de la siguiente manera: en la sección 2 se presentan algunos trabajos relacionados, en los cuales se utilizan tecnologías de la *web* semántica para mejorar el proceso

de recuperación de información. En la sección 3, se describe un modelo para extender semánticamente el proceso de recuperación de información mediante el uso de anotaciones semánticas. En en la sección 4, un mecanismo automático para anotar semánticamente documentos de texto. En la sección 5 se describe una estrategia para buscar documentos enriquecidos semánticamente. En la sección 6, la implementación de un prototipo y las pruebas realizadas. Por último, en la sección 7 se presentan las conclusiones y trabajo futuro.

2. TRABAJOS RELACIONADOS

En los últimos años, diferentes trabajos se han desarrollado buscando mejorar la recuperación de información usando tecnologías de la *web* semántica (Vallet-Weadon, Fernández-Sánchez y Castells-Azpilicueta, 2005; Castells-Azpilicueta, Fernández-Sánchez y Vallet-Weadon, 2007; Bhagdev, *et al*., 2008*)*. Estos trabajos se enmarcan dentro del campo de la búsqueda semántica, haciendo referencia a sistemas que usan tecnologías de la *web* semántica para mejorar las diferentes partes del proceso de recuperación de información. La búsqueda semántica, de acuerdo con Mangold (2007), se define como el proceso de recuperación de documentos que aprovecha el conocimiento de un dominio, y que se puede formalizar mediante una ontología. Para Wei-Wang, Barnaghi y Bargiela (2008), la búsqueda semántica tiene como objetivo mejorar las técnicas y los métodos convencionales de recuperación de información. Por su parte Nagypal (2007), clasifica los sistemas de búsqueda 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. Este artículo se enfoca en la segunda categoría, y tiene como propósito mejorar la recuperación de información sobre documentos de texto, usando ontologías y anotaciones semánticas. Los sistemas de búsqueda semántica orientados a mejorar la recuperación de documentos se pueden ver como una extensión de la recuperación de información clásica. En estos sistemas, los documentos se anotan semánticamente con base en una ontología de dominio. El proceso de recuperación se lleva a cabo haciendo coincidir las consultas de los usuarios con las anotaciones semánticas extraídas de los documentos (Wei-Wang, Barnaghi y Bargiela, 2008).

2.1 Criterios para analizar sistemas de búsqueda semántica

Varios autores, Manglod (2007), Wei-Wang, Barnaghi y Bargiela (2008) y Strasunskas y Tomassen (2010) han clasificado diferentes sistemas de búsqueda semántica basándose en un conjunto de criterios que permiten analizar sus características más importantes. Con el objetivo de facilitar el análisis de sistemas de búsqueda semántica orientados a la recuperación de documentos, se seleccionaron los siguientes criterios: el nivel de transparencia del sistema, el lenguaje de representación del conocimiento (*RDF, OWL, DAML+OIL*), el mecanismo de anotación semántica y la forma en que el usuario hace las consultas. El nivel de transparencia se refiere a la forma de interacción del usuario con el sistema de búsqueda semántica. Esta puede ser invisible, si las capacidades semánticas están ocultas para el usuario, interactiva, si el sistema le pide al usuario retroalimentación para hacer cambios en la consulta, o híbrida, si es una combinación de las dos anteriores. Por su parte, el mecanismo de anotación puede ser manual, semiautomático o automático. La anotación manual es un proceso difícil y costoso en términos de tiempo y personas que se necesita para realizarlo (Corcho, 2006). En la anotación semiautomática se requiere una mínima intervención del usuario y las anotaciones se hacen con base en las sugerencias realizadas por algún proceso automático (Oren, *et al.,* 2006). Por su parte, en la anotación automática, las anotaciones semánticas se hacen casi sin intervención del usuario y reducen los costos en términos de tiempo y personas requeridos para realizarlas.

La forma en que los usuarios hacen las consultas al sistema de búsqueda semántica puede ser: basada en palabras clave, en formularios, en lenguaje natural o basada en un lenguaje formal de consulta *(*e.g., SPARQL, es un acrónimo recursivo del inglés SPARQL *Protocol and RDF Query Language*, se trata de un lenguaje estandarizado para la consulta de grafos RDF, normalizado por el RDF *Data Access Working Group* (DAWG) del World Wide Web Consortium (W3C). Para ampliar más información en: http://www.w3.org/TR/rdf-sparql-query/).

Los sistemas basados en palabras clave se caracterizan por la facilidad de uso que ofrecen a los usuarios. Los que se basan en formularios, presentan gráficamente al usuario partes de la estructura de la

ontología para que él seleccione las clases con las cuales realizará la búsqueda. Estos sistemas tienen una desventaja: son poco flexibles ya que el usuario sólo puede seleccionar los elementos que le muestra el formulario (Uren, *et al.*, 2007). Además, el usuario consume mucho tiempo navegando por la estructura de la ontología. Por su parte, los sistemas basados en lenguaje natural, ofrecen respuestas precisas a las consultas del usuario (*Query Answering*). Por último, algunos sistemas de búsqueda semántica requieren que la consulta se exprese usando un lenguaje formal de consulta. Esto puede ser una desventaja, ya que representa un alto nivel de complejidad para los usuarios.

2.2 Propuestas que usan tecnologías la *web* semántica para mejorar la recuperación de información

Una de las primeras propuestas que busca mejorar el proceso de recuperación de información usando tecnologías de la *web* semántica se presenta en Shah, *et al.*, (2002). En esta propuesta, los documentos se enriquecen con anotaciones semánticas que se obtienen automáticamente aplicando técnicas de extracción de información. Las anotaciones se almacenan dentro del mismo documento y una consulta se puede expresar mediante palabras clave o usando el lenguaje de consulta *DQL* (*DAML+OIL Query Language*). En Popov, *et al.* (2004) y Kiryakov, *et al.* (2005), 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, las anotaciones semánticas se almacenan en una base de conocimiento separada de los documentos y se representan mediante tripletas *RDF* (*Resource Description Framework* en inglés, modelo estándar para intercambio de datos en la web, más información en www.w3.org/RDF). 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 anotados con estas instancias. En *KIM* no se tiene en cuenta la relevancia de las anotaciones, por lo tanto es difícil aplicar un algoritmo de *ranking* que permita mostrar ordenadamente los documentos de acuerdo a la relevancia de las anotaciones semánticas.

Por su parte, en Vallet-Weadon, Fernández-Sánchez y Castells-Azpilicueta (2005) y Castells-Azpilicueta, Fernández-Sánchez y Vallet-Weadon (2007) se adapta el modelo de espacio vectorial para facilitar la búsqueda semántica de documentos basándose en la relevancia de las anotaciones. Los autores proponen un algoritmo de *ranking* para calcular el grado de relevancia de las anotaciones semánticas. El grado de relevancia depende de la frecuencia de las clases de la ontología con las cuales se anotaron los documentos. Este algoritmo de *ranking* ordena los documentos de acuerdo a la relevancia de las anotaciones semánticas. El sistema toma como entrada una consulta expresada en *Sparql* 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 de acuerdo a la similitud entre la consulta y las anotaciones semánticas. En consonancia con los autores, la búsqueda semántica mejora los resultados con respecto a la búsqueda que se basa únicamente en modelos clásicos de recuperación de información. Sin embargo, la búsqueda semántica puede fallar cuando las anotaciones semánticas son incompletas y no cubren toda la información de un documento. Una desventaja de esta propuesta es que la consulta se debe expresar en *Sparql*, lo cual puede representar un nivel de complejidad alto para los usuarios del sistema.

En Bhagdev, *et al.* (2008), y Bikakis, *et al.* (2010), se aplica el concepto de búsqueda híbrida que combina los resultados de la búsqueda basada en modelos clásicos de *IR,* con resultados de la búsqueda basada en anotaciones semánticas. La búsqueda basada en modelos clásicos de *IR* (e.g., Modelo de espacio vectorial) tiene en cuenta únicamente la frecuencia de aparición de las palabras clave. Por otro lado, la búsqueda basada en anotaciones semánticas puede fallar cuando la ontología utilizada como base de anotación no cubre toda la semántica de un documento. La búsqueda híbrida trata estos problemas combinando la búsqueda basada en modelos clásicos de *IR* con la que se basa en anotaciones semánticas. Esto es una ventaja ya que se mejoran los resultados en términos de *precision* y *recall* en el promedio de los casos. Sin embargo, una desventaja de esta propuesta es la forma de interactuar con el sistema; el usuario debe navegar por la estructura jerárquica

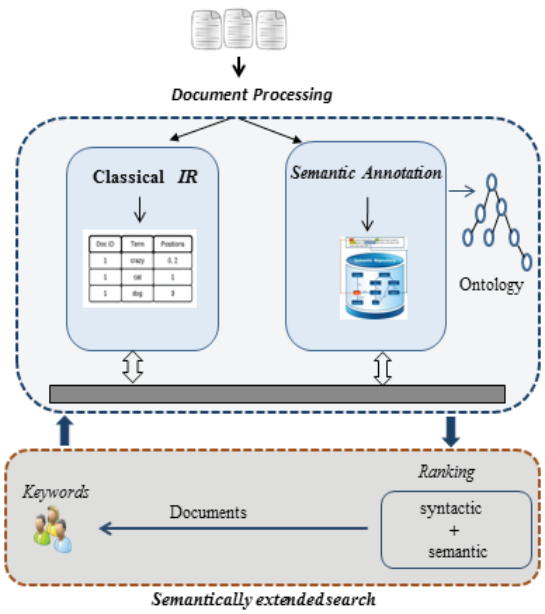
de la ontología y seleccionar manualmente las clases con las cuales se orientará el proceso de búsqueda. El usuario necesita invertir una gran cantidad de tiempo seleccionando las clases, además, se requiere conocer la estructura de la ontología.

Por último, en Rodríguez-García, *et al.* (2014a, 2014b), se describe una plataforma para enriquecer semánticamente el descubrimiento de servicios en la nube *(Cloud services*). Esta plataforma usa la descripción de los servicios en la nube como documentos y, a partir de estos, se crean anotaciones semánticas automáticamente. En el proceso de anotación se pueden usar múltiples ontologías y formatos de documentos y las anotaciones semánticas se indexan adaptando el modelo clásico de espacio vectorial. Para cada documento se crea un vector donde cada dimensión representa el nivel de relevancia de un concepto de la ontología para dicho documento. La plataforma permite la búsqueda semántica de documentos por medio de palabras clave, lo cual facilita la interacción con el usuario. Uno de los aspectos más interesantes en esta propuesta, es que se sugiere un módulo para soportar la evolución de ontologías. Con esto se busca enriquecer, mejorar y aumentar el conocimiento representado en las ontologías. Para soportar este proceso, se propone un algoritmo que busca información en *Wikipedia* de aquellos términos que no están representados en la ontología. Con estos términos se busca en *Wikipedia* los artículos que coincidan con los términos de búsqueda; luego se crea un nuevo concepto que contiene sinónimos tanto en inglés como en español, y añade a la ontología.

3. MODELO PARA EXTENDER SEMÁNTICAMENTE EL PROCESO DE RECUPERACIÓN DE INFORMACIÓN

De acuerdo con Baeza-Yates y Ribeiro-Neto (1999), el proceso de recuperación de información incluye etapas como el modelado de la información, la indexación, la consulta y el *ranking* de los resultados. En el modelo de la **Figura 1**, se extiende este proceso mediante una etapa de anotación semántica con el fin de darle significado a los términos de un documento. De acuerdo con Bontcheva, *et al.* (2006), las anotaciones semánticas permiten crear enlaces entre las entidades presentes en un texto y sus descriptores definidos en una estructura semántica como una ontología. Este modelo se diferencia de otras propuestas de búsqueda semántica en la forma en que el usuario interactúa con el sistema. Bajo este modelo, el usuario expresa las consultas usando solo palabras clave, no necesita seleccionar manualmente clases de la ontología, ni conocer su estructura. Tampoco necesita conocer lenguajes formales de consulta para acceder a las anotaciones semánticas. De acuerdo con Tran-Duc Than, *et al.* (2009), el usuario está acostumbrado a expresar las consultas mediante interfaces de consulta usables, las cuales, generalmente, se basan en palabras clave. El modelo de la **Figura 1** está formado por dos componentes principales: El procesamiento de documentos (*Document processing)*, que se muestra en la parte superior de la figura y la búsqueda extendida semánticamente (*Semantically extended search)*, en la parte inferior.

Figura 1. Modelo para extender semánticamente la recuperación de información



El procesamiento de documentos se divide en dos módulos: el primero (*classical IR*), donde los documentos se procesan por medio de la aplicación de técnicas como la división del texto en *tokens,* la eliminación de *stopwords* y la aplicación de algoritmos de *stemming*. En este módulo, cada documento se representa mediante un conjunto de términos usando el modelo clásico de espacio vectorial (Salton, Wong-Andrew y Yang-

Chungshu 1975). Una descripción detallada de este módulo se presenta en la sección 4. El segundo módulo (*Semantic Annotation),* se encarga de hacer anotaciones semánticas con base en una ontología de dominio. En el módulo de anotación semántica, los documentos se representan mediante un conjunto de anotaciones, las cuales tienen un significado bien definido en la ontología. En la sección 5 se describe detalladamente el proceso de anotación semántica.

Por su parte, la búsqueda extendida semánticamente, permite buscar documentos combinando la recuperación de información clásica, con una búsqueda basada en anotaciones semánticas. Los resultados obtenidos separadamente se mezclan y se muestran al usuario usando un *ranking* híbrido que se muestra en la fórmula 1. El *ranking* híbrido se calcula combinando la relevancia de la búsqueda basada en técnicas clásicas de *IR* (*ir-score*), con la relevancia de la búsqueda basada anotaciones semánticas (*semantic-score*).

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

Los factores λ y ω representan el grado de importancia de la búsqueda basada en técnicas clásicas *IR* y de la 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. El uso de estos factores se explica detalladamente en la sección 5, donde se aborda la búsqueda de documentos extendida semánticamente.

3.1 Procesamiento de documentos basado en técnicas clásicas de IR

En esta propuesta, las técnicas clásicas de *IR* se refieren al uso del modelo de espacio vectorial para representar los documentos como vectores de términos (Salton, Wong-Andrew y Yang-Chungshu 1975) y además que la relevancia de un término depende sólo de su frecuencia de aparición en los documentos. Este procesamiento se divide en dos fases: el análisis y la indexación; en la primera, se aplica una serie de técnicas como la división del texto en *tokens,* la eliminación de acentos y *stopwords* y la reducción de los términos a su raíz, usando el algoritmo de *stemming* propuesto en Porter (1997). La fase de indexación toma como entrada el conjunto de términos que se obtuvieron en la fase anterior y los representa mediante el modelo de espacio vectorial. Bajo este modelo, un documento se representa como un vector de términos. Cada término tiene asociado un grado de relevancia que se calcula usando el algoritmo TF-IDF (Manning, Raghavan y Schütze, 2008). La relevancia depende únicamente de la frecuencia del término (TF) en el documento, y la frecuencia inversa del documento (IDF), es decir, la ocurrencia del término en la colección de documentos. El resultado de esta fase es un índice donde a cada documento le corresponde un vector y cada componente del vector representa el grado de relevancia que tiene un término para el documento.

4. ANOTACIÓN SEMÁNTICA DE DOCUMENTOS

El proceso de anotación semántica se dividió en dos partes: en la primera parte se implementó la ontología que se usa como base para anotar los documentos de texto. En la segunda parte se desarrolló un mecanismo automático de anotación semántica basado en el *API* de la herramienta *GATE* (el significado de estas siglas en inglés, *General Architecture for Text Engineering*, más información http://gate.ac.uk/.

API (*Application Programming Interface* por su sigla en inglés. Conjunto de subrutinas, funciones y procedimientos — o métodos, en la programación orientada a objetos— que ofrece cierta biblioteca para ser utilizado por otro *software* como una capa de abstracción).

A continuación se describe cada una de estas partes.

4.1 Ontología base para el proceso de anotación semántica

Durante el proceso de anotación semántica se usó la ontología propuesta por ACM (como se puede observar en www.computer.org/portal/web/publications/acmtaxonomy), que describe el dominio de ciencias de la computación. Esta se implementó en *Protegé* 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 usa para describir sinónimos o formas alternativas de representar textualmente

Tabla 1. Algunas propiedades de anotación definidas en la ontología

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

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 los valores en esta tabla están relacionados semánticamente. Durante el proceso de anotación semántica se tienen en cuenta todas las propiedades de anotación definidas previamente y también las instancias de clase, 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.

4.2 Anotación automática de documentos basada en GATE

GATE es un *framework* para el procesamiento de texto y se puede acceder a sus funcionalidades por medio de una interfaz gráfica, o usando un *API* que permite integrarlo a otras aplicaciones. En esta propuesta se usa el *API* de forma embebida para crear un mecanismo automático de anotación de documentos. Para realizar el proceso de anotación semántica se crearon recursos del lenguaje y recursos de procesamiento. Los recursos del lenguaje permiten definir la ontología que guiará el proceso de anotación y los documentos que serán anotados. Los recursos de procesamiento analizan los documentos y crean las anotaciones semánticas. La **Figura 2** muestra los recursos de procesamiento que definieron en *GATE* para realizar el proceso de anotación automáticamente. El recurso *Sentence splitter* divide el texto en sentencias, que pueden ser oraciones o frases y, el recurso *Tokenizer* divide el texto en *tokens. POS Tagger* y *Morphological Analyzer* se usan para asignar a cada palabra en el texto su categoría gramatical (e.g., verbo, artículo, adverbio, pronombre). Finalmente, el recurso *Onto-Root Gazetter* permite asociar los conceptos encontrados en el texto con las clases, de la ontología. El resultado de este proceso es una lista de anotaciones semánticas que representan enlaces entre el texto y las clases de la ontología.

La **Figura 3** muestra el algoritmo que se utiliza para hacer anotación semántica de forma automática. Este algoritmo recibe como entrada una lista con los documentos que serán anotados (*CorpusDocuments*) y un archivo (*gateApp*) que describe los recursos de procesamiento que se definieron previamente en *GATE*. Para cada documento, se obtiene su identificador (*docID*) y luego se ejecuta el proceso de anotación definido en el archivo *gateApp* (línea 5). El método *gateApp.execute* devuelve el conjunto de anotaciones semánticas realizadas al documento. Luego, se analiza este conjunto (línea 7-12), y para cada anotación se obtienen atributos de la ontología como: la clase (*concept*), la *uri* y la propiedad de anotación con la cual se anotó el documento. El algoritmo retorna una lista que contiene las anotaciones de cada uno de los documentos procesados.

En el proceso de anotación semántica presentado en la **Figura 3**, un documento se puede anotar con varias clases en la ontología y una clase se puede usar para anotar diferentes documentos. Dada esta situación, algunas anotaciones semánticas pueden ser más relevantes que otras. Por esta razón, después de que el algoritmo de la

Figura 2. Recursos de procesamiento usados en GATE para hacer anotación semántica

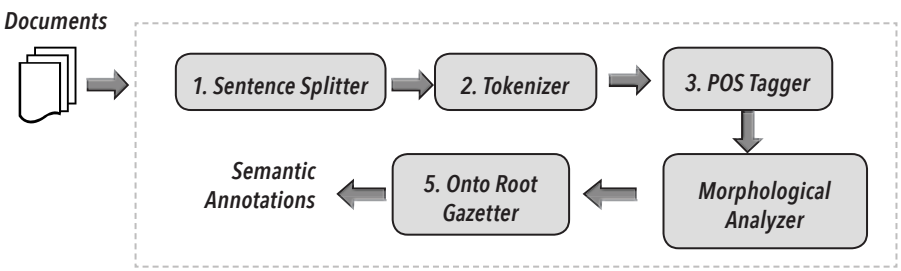


Figura 3. Algoritmo para hacer anotaciones semánticas usando el API de GATE

```
1  PROCEDURE  semanticAnnotator (CorpusDocuments, gateApp)
2
3      FOR EACH document  IN CorpusDocuments
4          docId  =  document.getID ();
5          gateAnnotations = gateApp.execute(document);
6
7          FOR EACH  annotation  IN gateAnnotations
8              concept = annotation.getConcept ();
9              uri =  annotation.get Uri ();
10             property =  annotation.getProperty ();
11             annList.add  (docID, concept, uri, property);
12         END
13     END
14     return annList;
15 END
```

Tabla 2. Anotaciones semánticas expresadas en tripletas

Sujeto	Propiedad	Objeto
annotation_1	*linked_to*	*document_3*
annotation_1	*ontology_class*	*Database_ Systems*
annotation_1	*weight*	*0,25*

Figura 3 finaliza su ejecución, se procede a calcular la relevancia de las anotaciones semánticas. La relevancia permite identificar qué clases de la ontología son más importantes para cada documento. El grado de relevancia semántico se obtiene mediante la **Fórmula 2**, y se basa en la propuesta de Castells-Azpilicueta, Fernández-Sánchez y Vallet-Weadon (2007).

$$W_{i,j} = \frac{freq_{i,j}}{max_l\ freq_{l,j}} * \log \frac{N}{n_i} \qquad (2)$$

El peso $W_{i,j}$ es el grado de relevancia que tiene la anotación ***i*** para el documento ***j***. Para calcular este peso, primero se calcula la frecuencia de la clase ($f_{i,j}$), que es el número de anotaciones semánticas que tiene el documento ***j*** con respecto la clase ***i*** de la ontología. Esta frecuencia se normaliza dividiéndola entre la máxima frecuencia ($maxl\ freq_{l,j}$). La frecuencia de la clase ($f_{i,j}$), se multiplica por la frecuencia inversa del documento, donde N es el número de documentos en la colección y n_i es el número de documentos anotados con la clase ***i***. Una anotación semántica se representa mediante tres atributos: el identificador del documento (*doc_id),* la clase de la ontología con la cual se hizo la anotación (*ontology_class*) y el grado de relevancia entre la clase y el documento (*relevance)*. 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*.

La **Tabla 2** muestra un ejemplo donde se ha creado una anotación semántica (*annotation_1*), que se enlaza con el documento (*document_3*), a través de la propiedad "*linked_to*". De esta forma, se puede decir que *document_3* fue anotado semánticamente con la clase *"Database_Systems"* y esta anotación tiene un grado de relevancia de 0,25.

4.3 Anotación semántica extendida

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 Nesić, *et al.* (2010), 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 caso se asume que cada nodo representa una clase en la ontología. Para descubrir las nuevas clases, se parte de aquellas que se obtuvieron con la herramienta *GATE*, a las cuales se les llama clases base. A partir de las clases base, se explora la estructura jerárquica de la ontología y mediante la propiedad *rdf:subClassOf* se obtienen las clases antecesoras a la clase base. Con las nuevas clases descubiertas 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. Cuanto 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, cuanto más grande sea la distancia semántica, menor será el grado de relevancia de la nueva anotación.

Por otro lado, la jerarquía de conceptos en la ontología puede ser muy grande, lo que implica que haya distancias muy largas entre un concepto y otro. Para evitar hacer anotaciones entre conceptos que son muy distantes en la estructura jerárquica, es necesario definir un límite en la distancia semántica. En esta propuesta, el límite se puede especificar de dos formas: la primera es definir un valor manualmente cada vez que se inicia el proceso de anotación; la segunda es configurar un valor límite por defecto. En Samper-Zapater, *et al.* (2008), se recomienda usar un límite menor o igual a tres, ya que los conceptos cuya distancia es mayor a tres, no representan una relación semántica bien definida debido a la forma en que la jerarquía de conceptos se construye. La **Figura 4** muestra una parte de la ontología que se usó durante el proceso de anotación semántica.

Si se toma como clase base a "*Transaction_procesing*", identificada con la etiqueta C_0, entonces la distancia semántica entre C_0 y C_1 es menor que la distancia semántica que hay entre C_0 y C_3. Basándose en esta distancia semántica, se puede decir que para la clase "*Transaction_procesing*" es más relevante la clase "*Database_Systems*" que la clase "*Information_Tecnology_and_Systems*". La anotación semántica extendida crea nuevas anotaciones con su respectivo grado de relevancia, el cual se calcula con base en la **Fórmula 3**. El grado de relevancia de una clase relacionada (*Wrc*), depende de la distancia semántica (*SemDistance*) y del grado de relevancia de la clase base (*Wbc*), el cual se obtuvo previamente usando la **Fórmula 2** (Sección 4.2). El grado de relevancia *Wrc* será más alto entre más pequeña sea la distancia semántica que hay entre las clases.

$$Wcr = Wbc * \beta^{-SemDistance}; \qquad \textbf{(3)}$$

5. CONSULTAS DE USUARIO EXTENDIDAS SEMÁNTICAMENTE

Una consulta se procesa en tres fases, tal como se muestra en la **Figura 5**. En la fase 1, se recuperan los documentos con base en el modelo de espacio vectorial y la métrica *TD-IDF* (Técnicas clásicas de *IR*). En la fase 2, la recuperación se hace con base en 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

Figura 4. Estructura jerárquica de la ontología

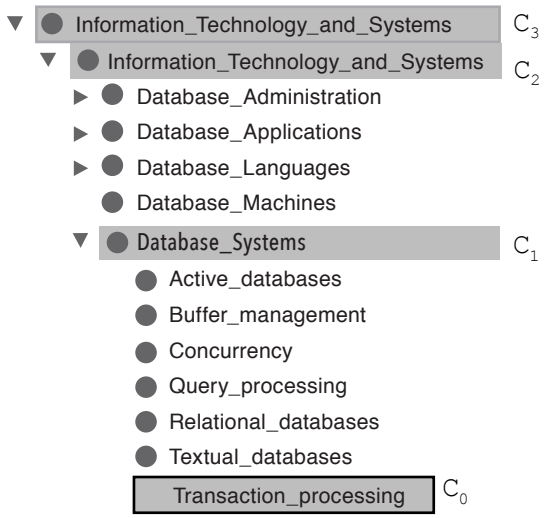
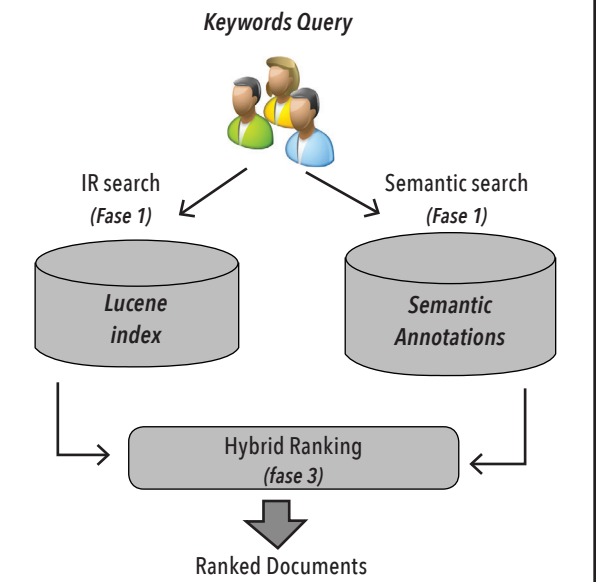


Figura 5. Consultas de usuario extendidas semánticamente



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 técnicas clásicas de *IR*. Los documentos se representan como vectores y se recuperan teniendo en cuenta solo la frecuencia de los términos. El resultado es una lista ordenada de documentos, cuya relevancia se calcula comparando la similitud entre el vector consulta y los vectores de cada documento. Para implementar esta fase se usó el *API* de *Apache Lucene* (es una API de código abierto para recuperación de información, originalmente implementada en Java por Doug Cutting, para ampliar este concepto http://lucene.apache.org/core/) , una herramienta *open source* que implementa el modelo de espacio vectorial.

***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 las 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 es diferente, 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 3**. Con las clases obtenidas se crea un vector que representa la consulta y este se usará más adelante, al momento de calcular la similitud entre los documentos y la consulta. 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.

Después de transformar las palabras clave a clases de la ontología se genera automáticamente una consulta *Sparql*. Cada clase obtenida en el paso anterior se agrega a la cláusula *WHERE* de la consulta. En la **Figura 6** se muestra una parte de la consulta generada. La consulta *Sparql* retorna una lista de anotaciones semánticas, donde cada elemento de la lista contiene el identificador del documento, la clase de la ontología y el grado de relevancia de la anotación semántica. Por ejemplo, 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 3**. Después de retornar la lista de anotaciones que coinciden con la búsqueda expresada por el usuario, se calcula la relevancia entre la consulta y los documentos. Tanto la consulta como los documentos se representan como vectores, cada posición del vector corresponde a una clase de la ontología. La relevancia se calcula por medio de la similitud coseno (Tan, Pang-Ning, Steinbach y Kumar, 2006). La recuperación basada en anotaciones semánticas también 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 que se basa en

Tabla 3. Transformación de una consulta a clases

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

Figura 6. Consulta Sparql generada a partir de palabras clave

```
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)
```

técnicas clásicas de *IR* con los documentos obtenidos en la búsqueda basada en anotaciones semánticas (Fase 1 y 2). 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. Los factores λ y ω se ajustan dependiendo de las condiciones de la consulta:

Condición 1: 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 ($\omega > \lambda$). En este caso $\omega = 0.7$ y $\lambda = 0.3$

Condición 2: La consulta no se puede representar en su totalidad con los conceptos de la ontología. El factor de importancia semántico tendrá un valor más bajo, entonces ($\omega < \lambda$). En este caso $\omega = 0.4$ 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. La verificación de la consulta con respecto a la información de la ontología permite darle más valor a la búsqueda semántica en aquellos casos donde la consulta del usuario se puede representar completamente con los conceptos de la ontología y disminuir este valor cuando no se puede representar en su totalidad. Finalmente, después de combinar los documentos se muestran al usuario.

6. IMPLEMENTACIÓN Y PRUEBAS

La **Figura 7** muestra un diagrama de diseño del prototipo que se desarrolló para mejorar la búsqueda documentos en una biblioteca digital de ciencias de la computación. En la implementación se utilizó el lenguaje de programación Java y se integraron algunas herramientas *open source* como *Apache Lucene*, *Apache Tika*,

Figura 7. Diagrama de diseño del prototipo

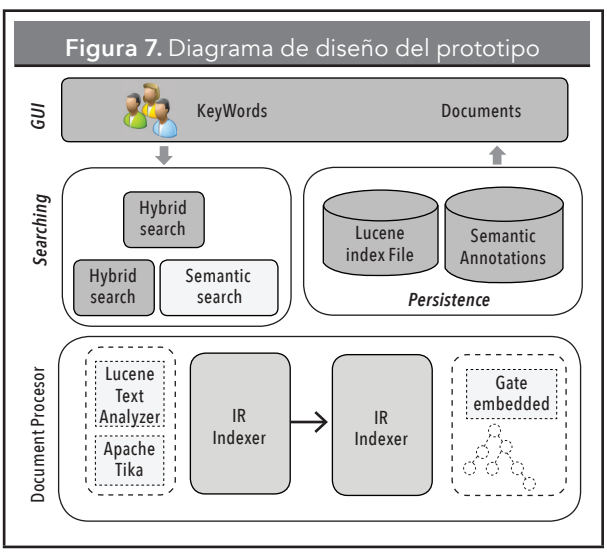


Tabla 4. Resultados de algunas consultas en términos de *Precision* (P) y *Recall* (R)

#	Consulta	Classical IR		Semantic		Hybrid	
		P	R	P	R	P	R
1	Reglas de asociación en minería de datos	0,54	0,59	0,81	0,9	0,72	0,84
2	Estimación de costos en Ingeniería de software	0,57	0,61	0,77	0,89	0,70	0,84
3	Sistemas operativos	0,60	0,65	0,87	0,92	0,75	0,81
4	Minería de datos aplicada a la medicina	0,57	0,63	0,35	0,43	0,59	0,63
5	Algoritmos de polinización de abejas artificiales	0,54	0,61	0,30	0,41	0,59	0,67

GATE y *Apache Jena.* El prototipo está integrado por cuatro componentes: *Document processor,* procesa los documentos, como se describió en las secciones 3 y 4. *Persistence* almacena la información obtenida durante el procesamiento de documentos en dos repositorios: "*Lucene index file"* y *"Semantic annotations"*. En el primero, los documentos se representan como vectores de términos. El segundo repositorio contiene las anotaciones semánticas obtenidas durante el proceso de anotación, las cuales se guardan en forma de tripletas *RDF.* El componente *Searching* incluye las funciones de búsqueda ofrecidas al usuario y está integrado por tres módulos: *IR search,* que procesa las consultas de usuario basándose en técnicas clásicas de recuperación de información, *Semantic search,* que las procesa basándose en las anotaciones semánticas, e *Hybrid search,* que combina los resultados de las dos anteriores. Por último, el componente *GUI* le permite al usuario buscar documentos enriquecidos semánticamente por medio de palabras clave.

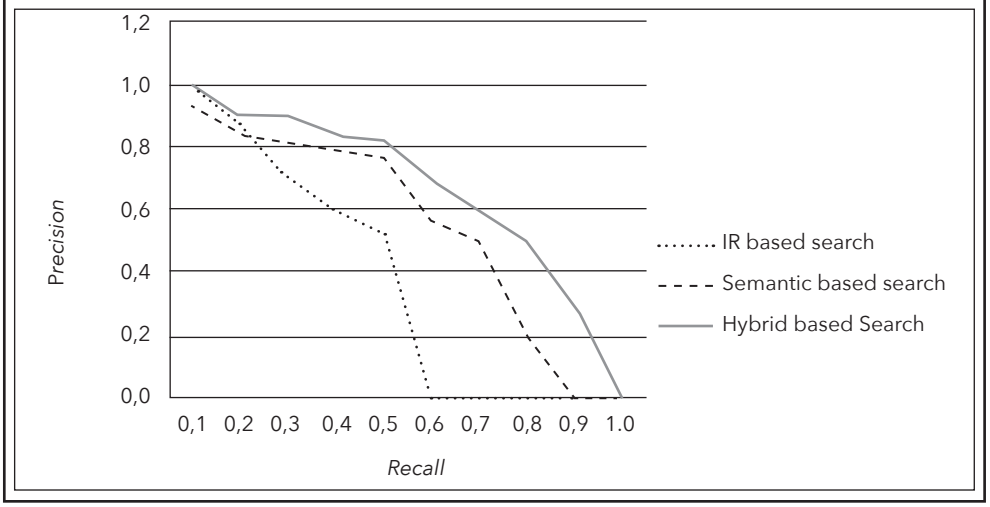
El usuario no necesita conocer la estructura de la ontología o de lenguajes formales de consulta para recuperar semánticamente los documentos; el proceso de búsqueda es trasparente para el usuario.

El escenario de pruebas está compuesto por una colección de documentos, las consultas de usuario expresadas en palabras clave, la ontología de dominio y las anotaciones semánticas. La colección de documentos tiene aproximadamente 2.000 ejemplares 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 que se basa en técnicas clásicas de IR (*IR based search*), la semántica o basada en anotaciones semánticas (*Semantic based search*) y la búsqueda híbrida (*Hybrid based search*). Los resultados obtenidos en cada una de estas estrategias se analizaron con base en las medidas *precision y recall*. Para cada consulta se tomaron 10 niveles de recall, (10 %, 20 %, 30 %,...100 %) y para cada nivel de *recall* se calculó la medida *precision.*

La **Tabla 4** muestra algunos ejemplos de las consultas que se usaron en las pruebas. Las primeras tres consultas corresponden a ejemplos donde todas las palabras expresadas por el usuario se pueden interpretar por medio de los conceptos representados en la ontología. En este caso, la búsqueda basada en anotaciones semánticas tiene un desempeño superior a la búsqueda clásica. Esto se debe a que los documentos se han enriquecido con las anotaciones semánticas y estas tienen en cuenta las propiedades de anotación y las instancias asociadas a las clases de la ontología. Por otro lado, las dos últimas filas de la tabla son consultas que no se pueden interpretar completamente con los conceptos de la ontología. Algunas palabras como "*medicina*", y "*polinización de abejas*", no están representadas en la ontología que se usó en el proceso de anotación semántica. En este caso, la búsqueda semántica falla porque no existen anotaciones que estén relacionadas con todas las palabras expresadas por el usuario. En la consulta 4, la búsqueda semántica se hace teniendo en cuenta solo el concepto "*Minería de datos*", lo cual afecta su rendimiento, como se puede ver en la **Tabla 4**.

De acuerdo con la **Figura 8**, en el promedio de los casos, la búsqueda clásica (*IR based search*) tiene un desempeño inferior que la búsqueda semántica y que la búsqueda híbrida. Por ejemplo, en la búsqueda clásica, para niveles de *recall* cercanos a 0,5, la métrica *precision* también toma valores aproximados a 0,5. A partir de

Figura 8. *Precision* Vs *Recall* para el promedio de los casos



este nivel, *precision* decrece rápidamente hasta llegar a cero. Por otro lado, en la búsqueda semántica, 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. Por su parte, la búsqueda híbrida tiene un desempeño superior que la búsqueda semántica en el promedio de los casos. La búsqueda semántica funciona muy bien cuando el usuario expresa consultas que se pueden interpretar completamente con la información representada en 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, ya que aprovecha las ventajas de la búsqueda clásica y de la búsqueda semántica. Esto se logra porque el *ranking* híbrido se calcula dependiendo de las condiciones de la consulta, como se describió en la fase 3 de la sección 5.

7. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se desarrolló un prototipo de 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 solo en técnicas clásicas de *IR*. Las anotaciones semánticas y la ontología de dominio son una parte fundamental de este sistema. Cuando las consultas de usuario se pueden interpretar completamente con los conceptos de la ontología, la búsqueda basada en anotaciones semánticas ofrece mejores resultados que la búsqueda clásica. Sin embargo, si la ontología es incompleta, la búsqueda semántica puede fallar porque las anotaciones no cubren toda la semántica de los documentos. 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. Otras propuestas como la de Rodríguez-García, *et al*., (2014b) y Bikakis, *et al*. (2010) se basan solo en anotaciones semánticas que se obtienen a partir de las ontologías.

En términos de *precision* y *recall*, un sistema de recuperación de información extendido semánticamente ofrece mejores resultados que un sistema que se basa solamente en la aplicación de técnicas clásicas de *IR*. La medida *precision* se mejoró porque en la búsqueda de documentos se tienen en cuenta los conceptos representados en una ontología, los cuales tienen un significado explícitamente definido. La medida *recall* se mejoró porque las consultas se expanden por medio de instancias y propiedades de anotación asociadas a las clases de la ontología. El sistema extendido semánticamente recupera documentos no solo con las palabras expresadas por el usuario, sino que, se extiende con los conceptos que están relacionados semánticamente en la ontología.

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 esta propuesta 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 el de Castells-Azpilicueta, Fernández-Sánchez y Vallet-Weadon (2007), donde la consulta se expresa mediante *Sparql*, lo cual puede representar un alto nivel de complejidad para los usuarios.

En esta propuesta, 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, este 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, la propuesta presentada se diferencia de otros trabajos como Bhagdev, *et al.* (2008) y Bikakis, *et al*. (2013), 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.

En el desarrollo de este trabajo se integraron herramientas que provienen del campo de la recuperación de información y también de la *web* semántica. La integración de estas áreas de conocimiento ofrece grandes ventajas para mejorar la efectividad y el desempeño de los sistemas orientados a la recuperación de documentos. Las pruebas se realizaron con una ontología de ciencias de la computación; sin embargo, el prototipo permite configurar una ontología de cualquier dominio. El mejoramiento de los resultados depende de la calidad y la completitud de la información representada en la ontología.

Trabajo futuro

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. También es necesario seguir explorando técnicas que permitan al usuario acceder a la información almacenada en las ontologías de manera usable y natural, sin que este tenga que conocer de lenguajes formales de consulta. Otro aspecto es el relacionado con la escalabilidad de las herramientas de búsqueda basadas en anotaciones semánticas. Se deben buscar mecanismos que permitan acceder a las anotaciones semánticas en entornos a gran escala con tiempos de respuesta mínimos. En este trabajo no se evalúo este requerimiento, pero se debe tener en cuenta si se quiere implementar un sistema donde muchos usuarios concurrentes accedan a las anotaciones semánticas.

REFERENCIAS

Baeza-Yates, R.; Ribeiro-Neto, B.A. (1999). Modern Information Retrieval. ACM Press/New York, Addison-Wesley.

Bikakis, N.; Giannopoulos, G.; Dalamagas, T.; Sellis, T. (2010). Integrating Keywords and Semantics on Document Annotation and Search. *Springer-Verlag* Berlin, Heidelberg, pp.921-938.

Bhagdev, R.; Chapman, S.; Ciravegna, F.; Lanfranchi, V.; Petrelli, D. (2008). Hybrid Search: Effectively Combining keywords and Semantics Searches. *Springer Berlin* Heidelberg, LNCS 5021, pp. 554–568.

Bontcheva, K.; Cunningham, H.; Kiryakov, A.; Tablan, V. (2006). Semantic Annotation and Human Language Technology. Semantic Web Technologies: Trends and Research in Ontology-based Systems, Davies, J.;Studer, R.; Warren, P. John Wiley & Sons, Ltd, pp. 29-50.

Castells-Azpilicueta, P.; Fernández-Sánchez, M.; Vallet-Weadon, D. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Transactions on Knowledge and data Engineering*, 19(2), pp. 261-272.

Corcho, O. (2006). Ontology Based Document Annotation: Trends and Open Research Problems. *Inderscience Publishers*, 1(1), pp. 47–57.

Kiryakov, A.; Popov, B.; Ognyanoff, D.; Manov, D.; Terziev, I. (2004). Semantic Annotation, Indexing, and Retrieval. J*ournal of Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1), pp. 49-79.

Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web, *Scientific American*, 284(5), pp. 34-43.

Lei-Yuangui.; Uren, V.; Motta, E. (2006). Semsearch: A Search Engine for the Semantic Web. *Springer Berlin* Heidelberg, 4248, pp.238-245.

Mangold, C. (2007). A Survey and Classification of Semantic Search Approaches. *International Journal of Metadata, Semantics and Ontologies*, 2(1), pp.23-34.

Manning, C.D.; Raghavan, P.; Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Nagypal, G. (2005). Possibly Imperfect Ontologies for Effective Information Retrieval. PhD thesis, University of Karlsruhe, 3762, pp.780-789.

Nesić S.; Jazayeri M.; Crestani, F.; Gašević, D. (2010). Concept-Based Semantic Annotation Indexing and Retrieval of Document Units. In *Proceedings of the 9th International conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*. Paris, France, RIAO, pp.234–237.

Oren, E.; Moller, K.; Scerri, S.; Handschuh, S.; Sintek, M. (2006). What are Semantic Annotations. Technical report, DERI Galway .

Popov, B.; Kiryakov, A.; Ognyanoff, D.; Manov, D.; Kirilov A.; Goranov, M. (2004). KIM Semantic Platform for Information Extraction and Retrieval Journal . *Natural Language Engineering*, 10(3-4), pp. 375-392.

Porter, M.F. (1997). An algorithm for suffix stripping. In Readings in Information Retrieval, Sparck, K. and Willett, P. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 313-316.

Rodríguez-García, M.A.; Valencia-García, R.; García-Sánchez, F.; Samper-Zapater, J.J. (2014a). Creating a Semantically-Enhanced Cloud Services Environment through Ontology Evolution. *Future Generations in Computer Systems*, 32, pp. 295–306.

Rodríguez-García, M.A.; Valencia-García, R.; García-Sánchez, F.; Samper-Zapater, J.J. (2014b). Ontology-based Annotation and Retrieval of Services in the Cloud. *Knowledge-Based Systems*, 56, pp. 15-25.

Salton, G.; Wong-Andrew.; Yang-Chungshu (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), pp. 613–620.

Samper-Zapater, J.J.; Adell-Aguilar, F.J.; Van den Berg, L.; Martínez, J.J. (2008). Improving Semantic Web Service Discovery. *Journal of Networks* (JNW), 3(1), pp.35-42.

Shah, U.; Finin, T.; Joshi, A.; Scott Cost, R.; Matfield, J. (2002). Information Retrieval on the Semantic Web. *International Conference on Information and Knowledge Management*, New York, pp. 461-468.

Strasunskas, D.; Tomassen, S. (2010). On Variety of Semantic Search Systems and Their Evaluation Methods. International Conference on Information Management and Evaluation. South Africa. A*cademic Conferences Publishing*, pp. 380-387.

Tan, Pang-Ning; Steinbach M.; Kumar, V. Introduction to Data Mining. Addison-Wesley, chapter 2, pp 74.

Tran-Duc Than.; Wang-Haofen; Rudolph, S.; Cimiano, P. (2009). Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF). *Data Engineering Conference, IEEE 25th International*, pp. 405-416.

TrivikRam, I., (2007). A Hibrid Approach to Retrieving Web Documents and Semantic data. Phd. tesis Wright State University, pp. 30.

Vallet-Weadon, D.; Fernández-Sánchez, M.; Castells-Azpilicueta, P. (2005). An Ontology-Based Information Retrieval Model. *The Semantic Web: Research and Applications. Springer Berlin* / Heidelberg, 3532, pp. 455-470.

Wang-Haofen; Zhang-Kang.; Liu-Qiaoling.; Tran-Thanh.; Yu-Yong. (2008). Q2semantic: A Lightweight Keyword Interface to Semantic Search. *The Semantic Web: Research and Applications. Springer* Berlin Heidelberg, 5021, pp. 584–598.

Wei-Wang.; Barnaghi, P.M.; Bargiela, A. (2007). Semantic-Enhanced Information search and Retrieval. *Conference on Advanced Language Processing and Web Information Technology,* Luoyang, Henan, China, pp. 218-223.

Wei-Wang, W.; Barnaghi, P.M..; Bargiela, A. (2008). Search with Meanings: An Overview of Semantic Search Systems. *Inter. Journal of Communications of SIWN*, 3, pp. 76-82.

Uren, V.; Lei-Yuangui.; López, V.; Liu-Haiming; Motta, E.; Giordano, M. (2007). The Usability of Semantic Search Tools: A Review. *The Knowledge Engineering Review*, 22(4), pp.361-377.

Zhou-Qi.; Wang-Chong; Xiong-Miao; Wang-Haofen; Yu-Yong (2007). SPARK: Adapting Keyword Query to Semantic Search. In *Proceedings of the 6th international The semantic Web. Springer-Verlag*, Berlin, Heidelberg, 4825, pp. 694-707.

PARA CITAR ESTE ARTÍCULO / TO REFERENCE THIS ARTICLE / PARA CITAR ESTE ARTIGO /

Solarte-Pabón, O.; Millán-López, M.E.delS. (2014). Propuesta para extender semánticamente el proceso de recuperación de información. Revista EIA, 11(22) julio-diciembre, pp. 51-65. [Online]. Disponible en: http://dx.doi.org/10.14508/reia.2014.11.22.51-65