Modelo Semántico de Expansión de Consultas para la Búsqueda Web (MSEC)

Query Expansion Semantic Model for Web Search (MSEC)

MIGUEL ANGEL NIÑO ZAMBRANO
Ingeniero de Sistemas, Magíster en Informática, Estudiante de Doctorado en Ingeniería Telemática
Profesora Titular, Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones
Miembro del Grupo de I+D en Tecnologías de la Información, Universidad del Cauca
manzamb@unicauca.edu.co
Popayán, Cauca, Colombia

IVÁN DARÍO LÓPEZ GÓMEZ
Ingeniero de Sistemas
Programa de Ingeniería de Sistemas, Universidad del Cauca
Miembro del Grupo de I+D en Tecnologías de la Información, Universidad del Cauca
ivancholg11@gmail.com
Popayán, Cauca, Colombia

CARLOS ADRIAN ANDRADE
Ingeniero de Sistemas
Programa de Ingeniería de Sistemas, Universidad del Cauca
Miembro del Grupo de I+D en Tecnologías de la Información, Universidad del Cauca
ingenierosistemas23@hotmail.com
Popayán, Cauca, Colombia

CARLOS ALBERTO COBOS LOZADA
Ingeniero de Sistemas, Magíster en Informática, Ph.D. (c) en Ingeniería de Sistemas y Computación
Profesor Titular, Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones
Director del Grupo de I+D en Tecnologías de la Información, Universidad del Cauca
ccobos@unicauca.edu.co
Popayán, Cauca, Colombia

RAMON FABREGAT GESA
Doctor en Tecnologías de la Información
Profesor Titular, Departament Arquitectura i Tecnologia de Computadors, Institut d'Informàtica i Aplicacions
Miembro del Grup de Comunicacions en Banda Ampla i Sistemes Distribuïts, Universitat de Girona
ramon@eia.udg.es
Girona, España

Fecha recibido: 31/10/2011
Fecha de aprobación: 15/06/2012

RESUMEN

Internet se ha convertido en el mayor repositorio de conocimiento humano y la cantidad de información almacenada crece cada día más. Esto último repercute en el bajo nivel de precisión que reportan los sistemas de búsqueda Web respecto a los documentos que son recuperados para el usuario. Para enfrentar este problema, una de las estrategias utilizadas es la recuperación personalizada de recursos. Actualmente existen varios proyectos que proponen métodos semánticos para aumentar la relevancia de las búsquedas, a través del uso de ontologías, procesamiento de lenguaje natural, sistemas basados en conocimiento, lenguajes de especificación de consultas y perfil de usuario, entre otras. Los resultados generalmente son mejores que los obtenidos por buscadores que no usan éstas técnicas. Sin embargo, el costo que se paga por estas mejoras en precisión se centra en el uso de algoritmos más complejos en implementación y que consumen más recursos computacionales. Este artículo describe un modelo semántico de expansión de consultas denominado MSEC, el cual está basado principalmente en el concepto de similitud semántica a partir de Ontologías de dominio y en el uso del perfil de usuario para personalizar las búsquedas y así

mejorar la precisión de las mismas. Para evaluar el modelo propuesto se creó un prototipo software. Los resultados experimentales preliminares muestran una mejora respecto al enfoque tradicional de búsqueda. Finalmente se comparó con el mejor buscador semántico del estado del arte, llamado GoPubMed para la colección MEDLINE.

PALABRAS CLAVE. Búsqueda Web, Expansión de Consulta, Ontologías de Dominio, Perfiles de Usuario, Similitud Semántica.

ABSTRACT

Internet has become the largest repository of human knowledge, and the amount of stored information increases day by day. This increase of information affects the levels of precision reported by Web search engines regarding documents retrieved for the user. One strategy being used to address this problem is a focus on a personalized resource recovery. Several projects currently offer semantic methods for improving the relevance of search results through the use of ontologies, natural language processing, knowledge based systems, query specification languages, and user profile, among others. Results are generally better than for web search engines that do not use these techniques. However, the high cost of these improvements in precision relate to use of more complex algorithms in carrying out the search and which are more wasteful of computational resources. This article describes a semantic query expansion model called MSEC, which is based mostly on the concept of semantic similarity, starting from domain ontologies and on the use of user profile in order to customize user searches so to improve their precision. In order to evaluate the proposed model, a software prototype was created. Preliminary experimental results show an improvement compared to the traditional web search approach. Finally the model was compared against the best state of the art semantic search engine, called GoPubMed, for the MEDLINE collection.

KEYWORDS. Web Search, Query Expansion, Domain Ontologies, User Profiles, Semantic Similarity.

1. INTRODUCCIÓN

En los últimos años la enorme cantidad de información disponible en la Web ha crecido de una manera sustancial, convirtiendo la Web en el mayor repositorio de conocimiento humano y en un medio de publicación fácilmente accesible para todos [1-2]. La constante búsqueda de información en la Web hace que los sistemas de recuperación de información (SRI) [3] establezcan nuevos métodos o estrategias que intenten mejorar la calidad de los resultados que se muestran al usuario, convirtiendo la relevancia [4] de los mismos en un factor determinante de éxito [5].

La recuperación de datos, en el contexto de un SRI, consiste principalmente en determinar cuáles documentos contienen colecciones de palabras clave en la consulta de búsqueda; sin embargo, esto no es suficiente para satisfacer la necesidad de información de los usuarios, quienes están interesados en la recuperación de recursos Web acerca de un tema particular, más que de datos aislados.

La Recuperación de Información (RI) consiste en la representación, almacenamiento, organización y acceso a ítems de información [6]. La representación y organización de estos ítems deben proporcionar al usuario un fácil acceso a la información en la cual está interesado. Pero hoy en día esto no es así debido a varios inconvenientes como la sobrecarga de información [7], la heterogeneidad semántica [8] y el uso inapropiado de la meta-información [9], entre otros.

Los usuarios de Internet continúan empleando motores de búsqueda tradicionales, los cuales ofrecen una visión de la Web que se limita al tratamiento léxico de los documentos sin tener en cuenta el significado que éstos representan para el usuario. Es decir su contenido semántico. Esto dificulta la búsqueda de información útil y obliga a los usuarios a pasar un mayor tiempo examinando gran cantidad de documentos hasta encontrar los adecuados o abandonar la búsqueda [9].

Teniendo en cuenta lo anterior, el problema sobre el cual se enfoca el presente trabajo busca mejorar la precisión de los resultados entregados a los usuarios cuando hacen búsquedas web [3]. Para lo anterior, en este artículo se presenta un modelo semántico para la expansión de la consulta que se basa en el uso de Ontologías de dominio y Pérfil de Usuario (PU), de tal manera que la expansión se convierte en una forma personalizada de aproximación a las necesidades de información del usuario y así permite mejorar la precisión de sus búsquedas.

A continuación en la sección 2 se hace un resumen de los principales trabajos relacionados con el presente proyecto. Luego en la sección 3 se explica en detalle MSEC, el modelo propuesto. Después en la sección 4, se muestra la experimentación y los resultados de la misma. Finalmente, se presentan algunas conclusiones y el trabajo futuro que el grupo de investigación espera desarrollar alrededor de la temática.

2. TRABAJOS RELACIONADOS

Tradicionalmente los métodos de expansión de consulta se enmarcan en el análisis de documentos globales [10], análisis de documentos locales [11] y el análisis del contexto local [12].

El análisis de documentos globales se basa en la extracción de palabras de expansión a partir de repositorios léxicos como diccionarios o tesauros globales al igual que ontologías, entre otros [10]. Este tipo de análisis ofrece la posibilidad de expandir la consulta de usuario mediante palabras relacionadas tales como sinónimos o hiperónimos. Pero este método en ocasiones provoca problemas de ambigüedad ya que ciertas palabras pueden tener diferentes significados dependiendo del contexto.

El análisis de documentos locales corresponde a la obtención de las palabras clave de los documentos que han sido recuperados a partir de la consulta del usuario y adicionar posteriormente ciertas palabras que expanden la consulta original del usuario [11]. Este método es uno de los más utilizados actualmente. Sin embargo, algunas veces tiene la desventaja de seleccionar palabras que no se encuentran relacionadas con el tema central de la consulta de usuario, produciendo resultados que no satisfacen su necesidad de información.

Finalmente el análisis del contexto local pretende evitar que las palabras no relacionadas con la consulta sean adicionadas como palabras de expansión. Para tal fin calcula una serie de valores que determinan qué tan relacionada se encuentra una palabra con una determinada consulta y las palabras más relacionadas se consideran como palabras de expansión [12]. Este método depende en gran medida de los resultados recuperados inicialmente, ya que si no son documentos relacionados, la precisión tiende a decaer para las búsquedas siguientes.

Existe una gran variedad de investigaciones enfocadas en la expansión de consulta como soporte para la mejora de la precisión de los resultados de la búsqueda Web. En [13] se describe el concepto de perfil ontológico, que es una extensión semántica de una ontología donde se da a cada concepto una descripción en términos de un vector de palabras clave ponderado. Estos perfiles ontológicos son utilizados para la expansión de consulta. La evaluación muestra resultados prometedores en comparación con la búsqueda tradicional basada en palabras clave. También se describe la noción de contexto en un entorno de recuperación de información y se indica cómo se pueden combinar la semántica y el contexto en la búsqueda basada en expansión de consulta. Sin embargo, el proceso de expansión con conceptos está limitado a los conceptos vecinos y no se consideran otras relaciones que pueden ser importantes.

Por otra parte, en [14] se propone un modelo de expansión de consulta semántico basado en ontologías. Mediante el uso de conocimiento del dominio en la ontología, el sistema de recuperación puede mejorar la comprensión semántica de los documentos recuperados y da la posibilidad de que el usuario ingrese su solicitud de información en una forma más precisa. Para llevar a cabo la expansión semántica, este trabajo establece un algoritmo de expansión semántica y varias funciones de expansión. Una desventaja de este trabajo es que no utiliza la realimentación que proporciona el PU para realizar el proceso de expansión de consulta.

El uso de la similitud semántica es muy común para determinar los términos más relacionados con una consulta. En [15] se expone un modelo de RI basado en ontologías de dominio que utiliza similitud entre los documentos y las consultas para mejorar la exactitud de los resultados en el proceso de búsqueda. Los términos definidos en la ontología se utilizan como metadatos para el marcado de contenido en la Web. Estas marcas semánticas son términos de índices semánticos en la RI y permiten obtener las clases equivalentes de dichos términos utilizando un razonador de descripción lógica. Sin embargo, no se hace un uso más detallado de la similitud semántica entre conceptos.

Con relación a las anteriores propuestas, en este trabajo se presenta un modelo semántico para expansión de consultas que se basa en el uso de ontologías de dominio y la realimentación del PU, de tal manera que la expansión se convierte en una forma personalizada de aproximación a los intereses de información del usuario y así se mejora la precisión de sus búsquedas.

UIS Ingenierías
REVISTA DE LA FACULTAD DE INGENIERÍAS FISICOMECÁNICAS

3. MODELO SEMÁNTICO DE EXPANSIÓN DE CONSULTAS – MSEC

En esta sección se describe el modelo semántico de expansión de consultas denominado MSEC, en el cual se propone la expansión de consulta teniendo en cuenta la interacción entre el componente semántico de las ontologías y la retroalimentación que proporciona el usuario en cada una de sus búsquedas. Este modelo consta de 5 módulos. Cada uno de estos módulos se muestra en la Figura 1 y son descritos a continuación.

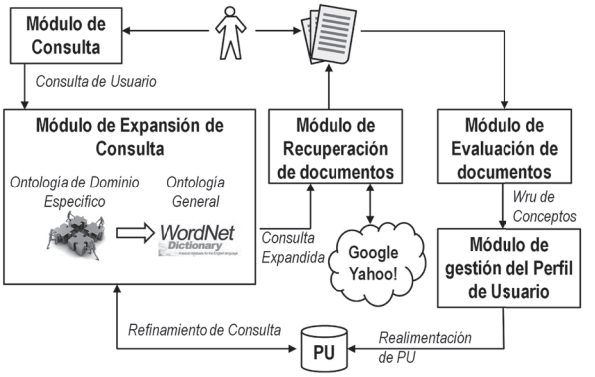

Figura 1. *Modelo Semántico de Expansión de Consultas – MSEC*

3.1 Módulo de Consulta

Proporciona al usuario el servicio de consulta de documentos. En este módulo, el usuario accede al sistema mediante una cuenta de usuario y una clave. Una vez dentro del sistema procede a digitar textualmente su consulta en la interfaz de búsqueda Web de forma análoga a como lo hace en la interfaz de un motor convencional de búsqueda, como por ejemplo en Google, Yahoo! o Bing.

3.2 Módulo de Expansión de Consulta

Este módulo se encarga de buscar la mejor forma de expandir la consulta, de tal manera que permita obtener documentos más relevantes para el usuario a través del meta buscador Web.

El proceso para expandir la consulta de usuario se muestra en la Figura 2. El objetivo de éste módulo es expandir los términos originales de la consulta del usuario con el fin de que la consulta enviada a los buscadores retornen mayor cantidad de recursos relacionados a las necesidades de información del mismo. Esto se logra adicionando conceptos relacionados a los términos originales, los cuales se obtienen a través de ontologías generales y de dominio. El cálculo de la similitud semántica se utiliza para encontrar más conceptos relacionados en la misma ontología y finalmente se añaden conceptos de la información almacenada en el PU. Dada la importancia de este módulo, a continuación se describen con más detalle sus principales componentes.

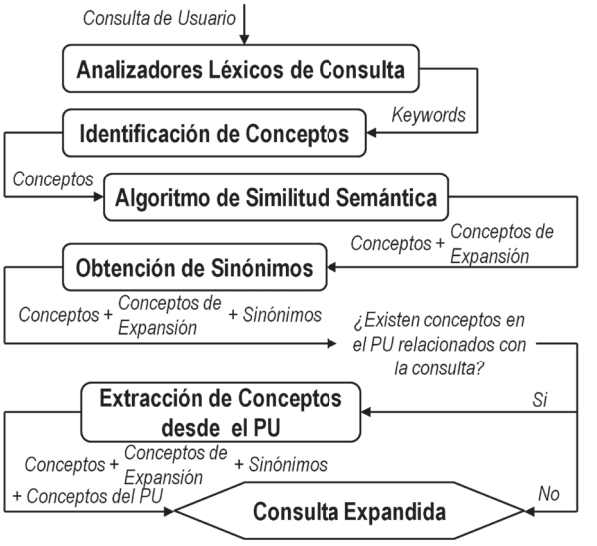

Figura 2. *Módulo de Expansión de Consulta*

El primer paso es el análisis léxico de la consulta. En este paso se busca obtener las palabras clave (keywords) de la consulta, por medio de la eliminación de palabras vacías, la eliminación de caracteres especiales, la conversión de los caracteres a minúsculas y el uso de stemming [3]. Posteriormente se procede a la identificación de conceptos. Se toma como base el método propuesto por Baziz [16] aplicándolo a la Ontología específica, identificando los tipos de conceptos contenidos en la consulta (específicamente conceptos simples (Cs), frases simples (Fs) [17] y términos (t) que no se encuentren en la ontología), y los conceptos relacionados por medio de relaciones de restricción en la ontología (Cr). Como resultado se obtiene un vector de conceptos para cada tipo identificado. El siguiente paso es el cálculo de similitud semántica. Ésta se hace entre pares de conceptos que pertenecen a una ontología y puede ser usada con el objetivo de obtener conceptos adicionales que contribuyen a enriquecer la semántica de la consulta de usuario [18]. Estos conceptos han sido denominados conceptos de expansión por similitud semántica (Ces).

Teniendo en cuenta lo anterior, para determinar la similitud semántica entre dos conceptos este proyecto toma como base el trabajo propuesto en [19] que se basa

en el método de conteo de relaciones. MSEC utiliza esta medida por su simplicidad, rendimiento y expresividad.

La anterior medida se usa para proporcionar valores de similitud entre pares de conceptos de la consulta identificados a partir de la ontología de dominio. Aquellos pares con los mayores valores de similitud se usan para involucrar conceptos adicionales como los conceptos intermedios (si el par se encuentra en la misma línea jerárquica) o el antecesor común al par (si el par se encuentra en diferente línea jerárquica).

El siguiente paso es la extracción de sinónimos (Csn), el cual corresponde a aquellos términos que no hayan sido identificados en la ontología de dominio específico y son identificados en la ontología general, en este caso WordNet, para extraer el sinónimo de uso más común.

Finalmente, se realiza la extracción de conceptos desde el PU. Éste es utilizado como una técnica en la personalización, además de la propia consulta, para estimar los intereses del usuario [20]. De esta manera, conceptos que hayan sido utilizados en anteriores búsquedas, pueden complementar la consulta siempre y cuando se relacionen con la misma. La forma de almacenar y actualizar estos conceptos en el PU se trata con más detalle en los módulos de construcción y actualización del PU. Este procedimiento recurre al PU con la finalidad de encontrar los conceptos obtenidos a través de la similitud semántica entre dos conceptos. A partir de estos conceptos se seleccionan sólo aquellos con mayor valor de peso del concepto en los documentos relevantes para el usuario (Wru) para ser adicionados a la consulta expandida.

3.3 Módulo de Recuperación de Documentos

En este modulo se busca recuperar los documentos relacionados a la consulta expandida enviándola como entrada a los Buscadores Web más populares como por ejemplo Google y Yahoo!. Estos motores de búsqueda retornan un conjunto de resultados, de los cuales se extraen los 10 primeros de cada buscador.

Es preciso mencionar que en este módulo se aplica un formato al texto de consulta que es enviado a los buscadores teniendo en cuenta los tipos de conceptos identificados en el módulo anterior. Para tal fin se considera hacer uso primordialmente del operador lógico OR y su combinación con el operador AND entre los conceptos de la consulta expandida (OR si los conceptos son sinónimos o se encuentran en una relación hiperónimo-holónimo, y AND entre los conceptos de la consulta que estén estrechamente relacionados, por ejemplo en el caso de que un concepto en la ontología presente una restricción hacia otro), de tal manera que los nuevos conceptos adicionados a la consulta puedan servir como una posibilidad adicional para encontrar más documentos relevantes al usuario. En la Figura 3 se muestra el formato de texto que es aplicado a la consulta expandida.

$$Fs_1 \textbf{ OR } (Csi_{11} \ldots \textbf{ OR } Csi_{1m}) \textbf{ OR } (Fs_1 \textbf{ AND } (Cr_{11} \ldots \textbf{ OR } Cr_{1t})) \ldots \textbf{ OR } (Fs_n \textbf{ OR } (Csi_{n1} \ldots \textbf{ OR } Csi_{nm})) \textbf{ OR }$$
$$(Fs_n \textbf{ AND } (Cr_{n1} \ldots \textbf{ OR } Cr_{nr})) \textbf{ OR }$$
$$(Cs_1 \textbf{ AND } (Cr_{11} \ldots \textbf{ OR } Cr_{1t})) \ldots \textbf{ OR } (Cs_n \textbf{ AND } (Cr_{n1} \ldots \textbf{ OR } Cr_{nr})) \textbf{ OR }$$
$$(t_1 \textbf{ OR } Csn_{11}) \ldots \textbf{ OR } (t_n \textbf{ OR } Csn_{n1}) \textbf{ OR }$$
$$(Ces_1 \ldots \textbf{ OR } Ces_n)$$

Figura 3. *Formato de texto aplicado a la consulta expandida.*

3.4 Módulo de Evaluación de Documentos

En este módulo el usuario puede seleccionar cada uno de los documentos recuperados en el anterior módulo con el objetivo de visualizarlos y calificar cada documento con el fin de reflejar su nivel de interés sobre dicho recurso teniendo en cuenta su propio criterio de relevancia.

Después de que los resultados son presentados al usuario, éste puede seleccionar uno o más documentos del total de resultados que han sido recuperados por el meta-buscador Web. Además cada documento seleccionado puede ser calificado por el usuario de forma explícita seleccionando un único valor natural entre un rango de cero a cinco y reflejar de esta forma, su nivel de interés sobre el documento de forma más precisa.

3.5 Módulo de Gestión del Perfil de Usuario

Este módulo se encarga de gestionar el PU por medio de los conceptos que conforman una determinada consulta expandida, asociándole a cada concepto un valor de peso el cual es definido como la importancia de un término en un documento según el esquema de pesos abordado en el área de la RI [3]. Tomando como base esta definición, este módulo determina el peso de un concepto específico en los documentos relevantes para

el usuario (Wru). Es posible actualizar la información almacenada en el PU para establecer los intereses actuales del usuario. Como se explica más adelante, usando los conceptos extraídos desde el PU utilizados en la expansión se ajustan sus valores definidos en el Wru a tales intereses.

La importancia (o peso) de un término clave i para un documento j se basa en la tradicional fórmula TF-IDF en el modelo Espacio Vectorial de representación de documentos [3]:

$$Wd_{ji} = tf_{ji} * idf_i \quad (1)$$

Tomando como base la ecuación anterior, este proyecto toma el peso de un determinado concepto, el cual es calculado para cada documento seleccionado, y lo relaciona con la respectiva calificación hecha por el mismo usuario del documento, estas calificaciones y pesos de cada uno de los documentos se definen como el peso del concepto en los documentos relevantes para el usuario que en esta propuesta se define como *Wru*.

$$Wru_i = \sum_{j=0}^{n} (Wd_{ji} * Cal_j) \quad (2)$$

Donde Cal_j representa la calificación del documento *j* (es un porcentaje equivalente al rango de calificación de números discretos de cero a cinco, es decir 0 representa el 0% y 5 el 100%) que ha sido seleccionado por el usuario y Wd_{ji} el peso de un concepto *i* en el mismo documento. Adicionalmente, la ecuación (2) se aplica de forma individual a los conceptos de la consulta expandida obteniendo así un valor *Wru* para cada uno. El valor Wru es finalmente almacenado en el PU junto con el respectivo concepto al cual pertenece.

La actualización de la información almacenada en el PU establece los intereses actuales del usuario y éstos pueden cambiar en el tiempo. Para ello, este módulo realiza un ajuste a los valores *Wru* de los conceptos extraídos desde el PU utilizados en la expansión de consulta, de tal forma que, al calcular nuevamente estos valores, se ponderen con los anteriores y así se obtiene un nuevo valor *Wru* denominado peso ajustado (o *Wrua*) del concepto en los documentos relevantes para el usuario que es más acorde con el interés que presenta el usuario hacia el concepto en su actual necesidad de información. Para efectuar el ajuste mencionado anteriormente, esta propuesta plantea la siguiente relación denominada Ajuste de *Wru* de conceptos de expansión extraídos del PU.

$$Wrua = \alpha Wrup + \beta Wrur \quad (3)$$

Donde el peso histórico es representado por Wrup y el peso de la nueva evaluación es representado por Wrur. Los valores α y β son valores de ajuste y han sido establecidos en α=0.25 y β=0.75 a partir de las pruebas experimentales que se describen más adelante. Se asigna un mayor valor a β por ser la constante que acompaña a Wrur, el cual está más relacionado con el actual interés de información del usuario.

4. EXPERIMENTACIÓN Y ANÁLISIS DE RESULTADOS

Para llevar a cabo la evaluación del modelo propuesto se desarrolló un prototipo de meta-buscador Web el cual ha sido denominado MSEC Web Search. Éste hace uso de la ontología del Instituto Nacional del Cáncer de los Estados Unidos [21] la cual ha sido seleccionada en este proyecto por ser una ontología lo suficientemente especializada y robusta para proveer el componente semántico al proceso de expansión descrito anteriormente.

Para determinar el desempeño de MSEC se desarrollaron dos fases de evaluación. En la primera se usó la curva de Precision-Recall (precisión-recuerdo) y el índice Mean Average Precision (MAP) [23][24] sobre una colección cerrada de documentos donde se cuenta con juicios previos realizados por expertos acerca de la relevancia de un documento en un tema particular. En el prototipo software desarrollado en este proyecto se recurre a una ontología referente al cáncer en el dominio de las ciencias de la salud. Por esta razón se hace uso de la colección MED [22] compuesta por documentos cuya temática es la Medicina. En la segunda fase de evaluación se utilizaron las medidas de Precision at k y el índice MAP pero con la diferencia de que se realizan tomando la Web como repositorio de documentos.

Tomando como base las 30 consultas de pruebas preestablecidas en la colección utilizada (MED), 15 de éstas fueron seleccionadas como consultas de usuario en MSEC Web Search y el desempeño del mismo puede ser mostrado a través de la curva de Precisión- Recall que es calculada a partir de los resultados experimentales. En la Figura 4 se presenta un ejemplo para la consulta "*mycoplasma (infection or presence) in embryo, fetus, newborn infant or animal, or in pregnancy, gynecologic diseases, or as related to chromosomes or chromosome abnormaliti, or microanatomy*". Para cada consulta se han establecido cinco formatos de texto con los cuales la consulta se envía al meta-buscador Web.

En la gráficas de Precision-Recall se observa que la curva con mejor comportamiento para valores de precisión es la correspondiente a la consulta expandida mediante la

combinación de operadores OR-AND haciendo uso del PU. Este formato mostró mejores valores de precisión en 11 de las 15 consultas seleccionadas. Estos resultados indican que existen más documentos posicionados en los primeros lugares del ranking que se muestra al usuario.

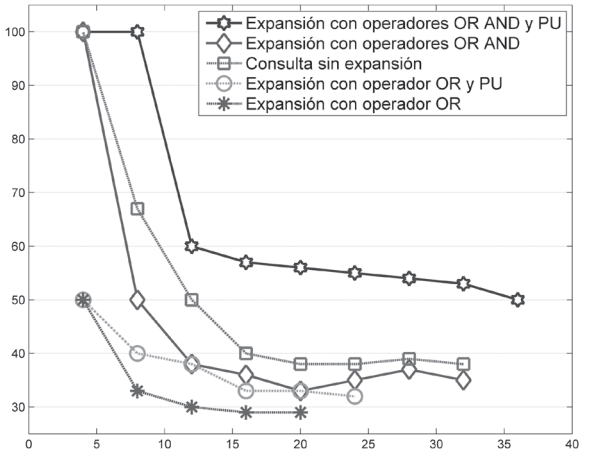

Figura 4. *Precisión vs. Recall para una consulta*

El índice MAP, que es el promedio del valor de precisión media de un conjunto de consultas, se resumen en la Tabla 1 para las tres primeras consultas.

Tabla 1. *Índice MAP por cada formato de texto aplicado a las consultas.*

Formato de Texto de Consulta	MAP
Consulta sin Expansión	0,503
Expansión con operadores (OR AND)	0,530
Expansión con operadores (OR AND) y PU	**0,573**
Expansión con operadores (OR)	0,443
Expansión con operadores (OR) y PU	0,451

El valor MAP de las consultas expandidas mediante los operadores lógicos OR-AND y el perfil de usuario presenta un promedio de 0.573 lo que indica que es un buen promedio de acuerdo a los resultados que usualmente se registran para un SRI, los cuales varían entre 0.1 y 0.7 [25].

A partir de estos primeros resultados experimentales se optó por utilizar la combinación de los operadores lógicos OR y AND en la segunda fase de evaluación. Estos operadores se aplican al texto de la consulta expandida.

En la segunda fase de evaluación se utiliza la medida de Precision At k, la cual permite conocer la precisión en los k primeros resultados en el ranking de un SRI. Por otra parte, para poder realizar una comparación a nivel general de estos SRI, se aplica nuevamente el índice MAP pero esta vez sobre los resultados de precisión pertenecientes a MSEC Web Search, Google y Yahoo!.

La Tabla 2 muestra la relación de la precisión en k resultados tomado los diez primeros documentos de los principales motores de búsqueda (Google y Yahoo!) y el modelo propuesto a través de MSEC Web Search.

La Figura 5 presenta la precisión en k documentos recuperados (k=10) para cada uno de los sistemas y permite compararlos con el modelo propuesto. Se puede observar la precisión promedio en diferentes valores de k con resultados comprendidos entre el 93,61% y 100% para MSEC Web Search. Esto muestra un alto grado de precisión comparado con los dos motores de búsqueda seleccionados. Además en la gráfica se observa que los resultados de MSEC Web Search mantienen valores de precisión superiores a los que proporcionan los buscadores tradicionales para todos los valores de k.

Tabla 2. *Resultados de Precisión en k=10.*

Doc	Precision At –k		
	Google %	Yahoo! %	MSEC Web Search %
1	100	86,66	100
2	100	89,99	100
3	100	87,77	100
4	98,33	84,43	100
5	98,33	83,10	100
6	96,11	83,10	98,88
7	93,25	80,24	97,92
8	91,58	78,57	97,1
9	89,36	78,57	95,61
10	88,02	75,90	93,61

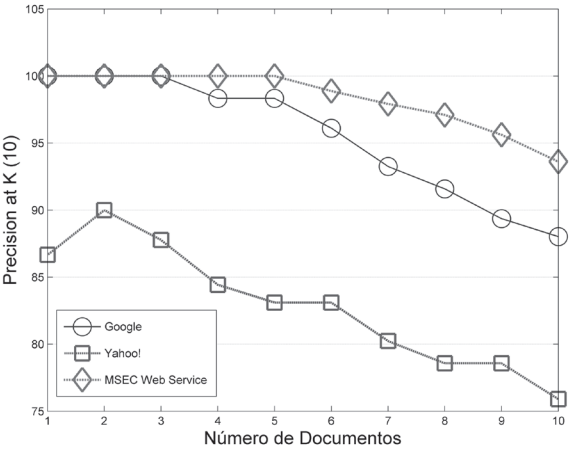

Figura 5. *Precisión At k (k=10) para Google, Yahoo! y MSEC Web Search*

En la Tabla 3 se muestran los resultados del índice MAP, el cual es utilizado para medir de una forma más general la precisión de los sistemas de búsqueda.

Tabla 3. *Comparativo del MAP entre los SRI evaluados.*

MAP		
Google	**Yahoo!**	**MSEC Web Search**
91,6%	86,3%	95,1%

Es importante destacar que el modelo propuesto en este proyecto presenta un índice MAP de 95,1%, que es superior al de los buscadores tradicionales (Google y Yahoo!), con el 3,5% que se logra frente a Google y de 8,8% que se consigue frente a Yahoo!.

Teniendo en cuenta que es de esperar que los resultados de un buscador con técnicas semánticas fueran superiores a los tradicionales, el proyecto contempló además la realización de un análisis comparativo de la precisión de los documentos recuperados por el prototipo MSEC Web Search y GoPubMed [26] que es un buscador semántico basado en conocimiento. GoPubMed se especializa en la recuperación de textos biomédicos de la base de datos de MEDLINE [27] por medio del uso de la Ontología Gene Ontology (GO) [28] y el vocabulario controlado MeSH (Medical Subject Headings) [29]. Para realizar esta evaluación se utilizó la medida de Precisión y el índice MAP. En la Figura 6 se presentan los resultados de precisión en los diez primeros resultados retornados por cada uno de los dos SRI para la consulta "*hemophilia and christmas disease, especially in regard to the specific complication of pseudotumor formation (occurrence, pathogenesis, treatment, prognosis)*".

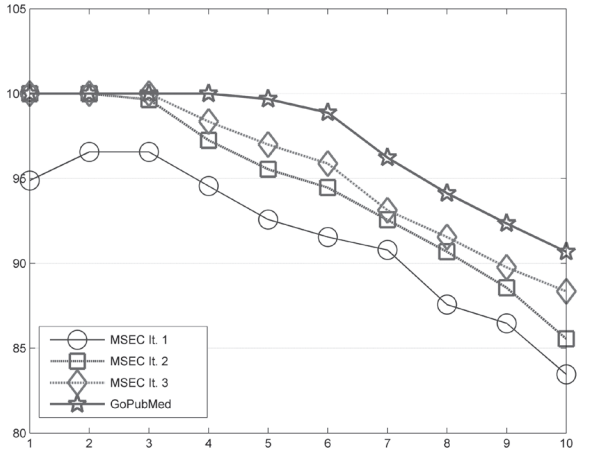

Figura 6. *Precisión de GoPubMed y MSEC Web Search*

En la Figura 6 se observa que la curva perteneciente a GoPubMed establece mayores valores en la precisión respecto a los pertenecientes a MSEC Web Search en la primera iteración de búsqueda. Teniendo en cuenta esto, se procedió a realizar nuevas iteraciones de búsqueda a partir de la misma consulta. Esto fue posible dado que el perfil del usuario en MSEC Web Search permite el refinamiento de la consulta en pro de obtener resultados más precisos. Las siguientes iteraciones incrementaron los valores de Precisión-Recuerdo para los documentos recuperados por MSEC Web Search, lográndo una aproximación a la precisión de GoPubMed. Se realizó el mismo procedimiento con 15 consultas obteniendo resultados similares.

Finalmente, en la Tabla 4 se presentan los valores del índice MAP calculados a partir de la precisión promedio obtenida con las 15 consultas para cada SRI (GoPubMed y MSEC Web Search).

Tabla 4. *Comparativo del índice MAP entre GoPubMed y MSEC Web Search.*

MAP	
GoPubMed	**MSEC Web Search**
71,12%	62,52%

El índice MAP refleja los valores generales de precisión para ambos SRI, en donde se establece una diferencia de 8,6% de GoPubMed sobre MSEC Web Search. Si bien los resultados son mejores para el primer SRI en cuanto a la precisión de sus resultados, se debe tener en cuenta que estos resultados pueden deberse a que GoPubMed utiliza técnicas basadas en conocimiento y además maneja una Ontología muy orientada a MEDLINE. Mientras que MSEC Web Search no usa un vocabulario tan especializado (terminología MeSH, diseñada especialmente para trabajar con la base de datos de MEDLINE). Por el contrario, se logran precisiones similares con MSEC Web Search y que mejoran en el tiempo, aplicando técnicas más simples en la expansión de la consulta y el manejo de un perfil de usuario sencillo.

Teniendo en cuenta el proceso de evaluación realizado, se puede apreciar que una expansión de consulta combinada con el refinamiento de la misma producido por la retroalimentación del PU produce mejores resultados en la búsqueda. Es decir, cuando se conocen de antemano los intereses del usuario, se proporciona un punto de referencia para recuperar documentos más relevantes en las primeras posiciones del ranking que se muestra al usuario.

5. CONCLUSIONES Y TRABAJO FUTURO

La búsqueda semántica proporciona grandes ventajas en comparación con la búsqueda Web tradicional puesto que la primera se enfoca en identificar relaciones existentes entre conceptos y en determinar el sentido más apropiado a una frase, a diferencia de la segunda que sólo se basa en la coincidencia de palabras clave.

El modelo semántico de expansión de consultas propuesto (MSEC) se centra en el uso de la similitud semántica sobre ontologías de dominio y la realimentación del PU de tal manera que las búsquedas de un usuario estén enfocadas a sus intereses particulares de información.

Dentro del modelo se propuso un nuevo algoritmo para la realimentación del perfil de usuario a través de las medidas Wru y Wrua, las cuales permiten establecer la importancia o el peso de un determinado concepto en los documentos relevantes para el usuario, de tal forma que dicho concepto se ajuste a sus intereses actuales en la búsqueda de información.

Se realizó una propuesta de expansión avanzada mediante la combinación de los operadores lógicos OR-AND los cuales resultaron ser más efectivos para la recuperación de documentos relevantes que el uso individual de estos mismos operadores.

Los resultados experimentales muestran que el modelo propuesto establece una mejora inicial (3,5 % con respecto a Google) con relación al enfoque tradicional en el que sólo se tiene en cuenta la coincidencia de palabras clave sobre los documentos indexados.

Además, MSEC proporciona algunas ventajas importantes tales como la recuperación de documentos relevantes que en el esquema tradicional de búsqueda eran omitidos, un mejor ranking de estos documentos en el conjunto de resultados que son mostrados al usuario y la mejora de la precisión en el tiempo al calcular los intereses del usuario.

6. TRABAJO FUTURO

Es conveniente evaluar el modelo propuesto en otro dominio particular del conocimiento (entre ellos, su uso en librerías digitales de múltiples conocimientos y como apoyo a la búsqueda de recursos en sistemas de gestión de aprendizaje.), con el fin de aportar más a la validez del mismo y realizar los ajustes requeridos con base en los resultados obtenidos. Tales ajustes se pueden traducir en la elaboración de una estructura más ajustada del PU que permita detectar en mayor detalle los intereses del usuario También se espera combinar el modelo de expansión con la indexación semántica [30] y con técnicas de procesamiento de lenguaje natural [31] para obtener mayor efectividad en la RI, pero buscando un equilibrio entre la efectividad de las técnicas, su sencillez y su costo computacional. Finalmente, realizar una comparación del modelo propuesto contra el propuesto por Rocchio y otras funciones de relevancia [32].

7. REFERENCIAS

[1] R. Dhanapal, "An intelligent information retrieval agent," Knowledge-Based Systems, vol. 21, No. 6, August 2008, pp. 466-470.

[2] C. Deco, C. Bender, J. Saer, and M. Chiari, "Expansión de consultas utilizando recursos lingüísticos para mejorar la recuperación de información en la web," Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos, Editorial de la Facultad d. Filosofía y Letras, Mendoza Argentina, 2005, pp. 35-46.

[3] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval, 2nd Edition, ACM Press Books, USA., 1999, p. 453.

[4] L. Schamberg, B. Einseberg, and S. Nilo, "A re-examination of relevance: toward a dynamic, situational definition," Information Procesing and Management, vol. 26, No. 6, 1990, pp. 755-776.

[5] K. Kim, J. Hong, and S. Cho, "A semantic Bayesian network approach to retrieving information with intelligent conversational agents," Information Processing & Management, vol. 43, No. 1, January 2007, pp. 225-236.

[6] G. Salton, Introduction to modern information retrieval, McGraw-Hill, New York, 1983, p. 448.

[7] Y. Marcano and R. Talavera, "Gestión de la información a través de la Web Semántica: Iniciativas y dificultades," Revista Venezolana de Gerencia (RVG), vol. 11, No. 36, October 2006, p. 36.

[8] P. Mitra, N. Noy, and A. Jaiswal, "Ontology Mapping Discovery with Uncertainty," Fourth International Conference on the Semantic Web, Galway Ireland 6th – 10th November 2005, p. 15.

[9] D. Avello, Web Cooperativa (Trabajo de Investigación), Universidad de Oviedo, 2002, p. 67.

[10] J. X. Xu and W. B. Croft, "Query expansion using local and global document analysis", Proceedings of the 19th Annual International

SIGIR Conference on Research and Development in Information Retrieval, New York 1996, pp. 4 – 11.

[11] R. Attar and A. S. Fraenkel, "Local feedback in full-text retrieval systems", Journal of the ACM, vol. 24, No. 3, 1977, pp. 397 – 417.

[12] J. X. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," ACM Transactions on Information Systems, vol. 18, No. 1, 2000, pp. 79 – 112.

[13] G. Solskinnsbakk and J. Gulla, "Combining ontological profiles with context in information retrieval," Data & Knowledge Engieering, vol. 69, No. 3, 2010, p. 10.

[14] H. Wang, J. Qin and H. Shao, "Expansion Model of Semantic Query Based on Ontology," 2009 Second Pacific-Asia Conference on Web Mining and Web based Application, Wuhan China 6th – 7th June 2009, pp. 86 – 90.

[15] J. Mustafa, S. Han, and K. Latif, "Ontology based semantic information retrieval," 4th International IEEE Conference, Varna 6th – 7th September 2008, pp. 22 – 19.

[16] M. Baziz, M. Boughanem, and N. Aussenac-Gilles, "Evaluating a Conceptual Indexing Method by Utilizing WordNet," Lecture Notes in Computer Science, vol. 40, No. 22, 2006, pp. 238 – 246.

[17] S. Liu, F. Liu, C. Yu, and W. Meng, "An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases," Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, NY USA 2004, pp. 266 - 272.

[18] V. Cordi, P. Lombardi, M. Martelli and V. Mascardi, "An Ontology-Based Similarity between Sets of Concepts," 2005, p. 6.

[19] T. Slimani, B. B. Yaghlane, and K. Mellouli, "A New Similarity Measure based on Edge Counting," Proceedings of world academy of science, engineering and technology, vol. 17, 2006, p. 5.

[20] P. Chen and F. Kuo, "An information retrieval system based on a user profile," Journal of Systems and Software, vol. 54, No. 1, 2000, pp. 3 – 8.

[21] National Cancer Institute. U.S. National Institutes of Health. Available: http://www.cancer.gov [citado 26 de Abril de 2011].

[22] MEDLARS. MEDical Literature Analysis and Retrieval System. Available: http://www.uninet.edu/do/MEDLARS.html [citado 26 de Abril de 2011].

[23] J. J. Yepes, Ontology Refinement for Improved Information Retrieval in the Biomedical Domain [PhD Thesis]. Universitat Jaume, Castellón, 2009.

[24] B. Croft, D. Metzler, T. Strohman, Search Engines: Information Retrieval in Practice, first edition, Addison-Wesley, USA, 2009, p. 552.

[25] C. D. Manning, P. Raghavan, H. Schütze, An Introduction to Information Retrieval, first edition, Cambrige University Press, Cambridge, 2008, p. 581.

[26] GoPubMed. GoPubMed, searching is now sorted!. Available: http://www.gopubmed.com/ [citado 23 de Junio de 2011].

[27] MEDLINE. MEDLINE/PubMed Resources Guide. Available: http://www.nlm.nih.gov/bsd/pmresources.html [citado 27 de Junio de 2011].

[28] Gene_Ontology. The Gene Ontology. Available: http://www.geneontology.org/ [citado 27 de Junio de 2011].

[29] BiKE-Laboratory. MeSH Ontology in OWL format. Available: http://bike.snu.ac.kr/?q=node/207 [citado 30 de Noviembre de 2010].

[30] M. Suárez, K. Salinas, "An Approach to Semantic Indexing and Information Retrieval," Revista Facultad de Ingeniería Universidad de Antioquia, No. 48, 2009, p. 174-187.

[31] P. Jackson, F. Schilder, "Natural Language Processing: Overview," Encyclopedia of Language & Linguistics, Elsevier, 2006, pp. 503 – 518.

[32] C. Cobos, E. Estevez, M. Mendoza, L. Gomez and E. León, "Algoritmos de Expansión de Consulta basados en una Nueva Función Discreta de Relevancia," Revista UIS Ingenierías, vol 10, No. 1, 2011, pp. 9-22.