MODELO SEMÁNTICO DE EXPANSIÓN DE CONSULTAS PARA LA BÚSQUEDA WEB

IVÁN DARÍO LÓPEZ GÓMEZ

CARLOS ADRIÁN ANDRADE HOYOS

Director: PhD (student). Miguel Ángel Niño Zambrano

Asesor: PhD (c). Carlos Cobos Lozada

UNIVERSIDAD DEL CAUCA

FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES

DEPARTAMENTO DE SISTEMAS

POPAYÁN

2011

MODELO SEMÁNTICO DE EXPANSIÓN DE CONSULTAS PARA LA BÚSQUEDA WEB

IVÁN DARÍO LÓPEZ GÓMEZ
CARLOS ADRIÁN ANDRADE HOYOS

Monografía presentada para optar al título de Ingeniero de Sistemas

Director: PhD (student). Miguel Ángel Niño Zambrano
Asesor: PhD (c). Carlos Cobos Lozada

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE SISTEMAS
POPAYÁN
2011

NOTA DE ACEPTACIÓN

_____

_____

_____

_____

PRESIDENTE DEL JURADO

_____

JURADO

Popayán, 2011

AGRADECIMIENTOS

Queremos agradecer a Dios por permitirnos afrontar este gran reto con nuestra mayor dedicación y responsabilidad para finalizar así una etapa más de formación académica y personal. A nuestros padres, familiares, compañeros y amigos por su constante apoyo y amor incondicional en todos aquellos momentos que compartieron con nosotros. A cada uno de los docentes que a lo largo de estos años nos transmitieron sus conocimientos y valores en pro de nuestro crecimiento académico y profesional. A nuestro tutor, el Ingeniero Miguel Ángel Niño por su permanente disposición y sus contribuciones en el desarrollo de nuestra tesis, igualmente al Ingeniero Carlos Alberto Cobos por sus valiosos aportes dentro del proyecto y a cada una de las personas que de alguna u otra forma participaron en el desarrollo del mismo. Finalmente a nuestra Alma Mater, la Universidad del Cauca, en especial a la Facultad de Ingeniería Electrónica y Telecomunicaciones, que en el marco de la celebración de sus 50 años de labores, brindó el espacio propicio para nuestra formación. Muchas gracias a todos.

Atentamente,

Iván Darío López Gómez

Carlos Adrián Andrade Hoyos

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Capítulo I

1 INTRODUCCIÓN

1.1 Contexto General

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[1] [4] 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 del usuario, el cual está interesado 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, debería 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[2], entre otros [9].

En adición a los problemas mencionados anteriormente, los usuarios de internet continúan empleando motores de búsqueda tradicionales, los cuales ofrecen una visión de la Web que sólo se limita al tratamiento léxico de los documentos, sin tener en cuenta el significado que estos 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 en este proceso examinando cientos de documentos hasta encontrar el adecuado [9].

[1] Se refiere a la utilidad, o potencial uso de los materiales recuperados, con relación a la satisfacción de los objetivos, el interés, el trabajo o los problemas intrínsecos del usuario.

[2] En el campo de la informática, remiten en particular al conocimiento relacionado con la estructura y los contenidos de las bases de datos.

1.2 Declaración del Problema

Teniendo en cuenta lo anterior, el problema sobre el cual se enfoca el presente proyecto, radica en la poca precisión de las búsquedas realizadas por los usuarios, de tal forma que el usuario obtiene resultados que son en muchas ocasiones poco relevantes para su necesidad de información [3]. En este trabajo se presenta un modelo semántico para expansión de consultas que se basa en el uso de Ontologías de dominio y el Perfil del Usuario (PU), de tal manera que la expansión se convierte en una forma personalizada de aproximación a los intereses de información del usuario y así mejorar la precisión de sus búsquedas.

1.3 Escenario de Motivación

"Dado un conjunto de datos jerarquizados o no, encontrar aquella información relevante para el usuario" [10]. La anterior tarea parece sencilla, pero a partir de esta, se ha generado toda una evolución a lo largo de las últimas décadas con el objetivo de mejorar la relevancia de la información que se presenta al usuario.

El alto grado de consolidación de la web está siendo favorecido por el vertiginoso abaratamiento de la tecnología informática, por el desarrollo de las telecomunicaciones y por la facilidad de publicación de cualquier documento que un autor considere interesante, sin tener que pasar por el filtro de los tradicionales círculos editoriales. Cientos de millones de personas utilizan los buscadores Web cada día, de modo que es evidente el rápido crecimiento en la importancia de la Recuperación de Información la cual es un proceso que puede ser comparable a encontrar una aguja en un pajar, ya que, por ejemplo, en Internet la cantidad de información es de unas dimensiones inmanejables.

1.4 Contribuciones

- **Fortalecimiento teórico en las áreas de Recuperación de Información y Web Semántica.** A partir del estudio y el análisis de diferentes investigaciones a través de fichas bibliográficas, en adición a la presente propuesta, se fortalece la base teórica perteneciente a la institución académica en lo referente a la Web Semántica para ser aplicada a la Recuperación de Información, lo cual permite establecer un soporte para futuras investigaciones en estas áreas.

- **Modelo Semántico de Expansión de Consultas para la Búsqueda Web.** La adición del factor semántico en las búsquedas Web mediante el uso de Ontologías de dominio y la personalización que provee el Perfil de Usuario, permiten dotar al sistema de una mayor precisión en las búsquedas que realiza el usuario.

- **Prototipo de meta-buscador Web basado en el modelo propuesto.** Desarrollo de una aplicación Web que permite la recuperación de documentos a partir de la

expansión de consulta basada en el uso de Ontologías y Perfiles de Usuario. Esta aplicación está soportada en el modelo mencionado anteriormente.

- **Artículo de investigación.** Creación del artículo de investigación *"Modelo Semántico de Expansión de Consultas para la Búsqueda Web - MSEC"*, el cual ha sido enviado a la "Revista Facultad de Ingeniería" de la Universidad de Antioquia. Este artículo se encuentra en estado de revisión por parte del personal encargado en esta revista.

- **Monografía del trabajo de grado.** Corresponde al presente documento, donde se describe el proceso seguido en el desarrollo del proyecto, los problemas que se presentaron, las respectivas soluciones, los principales aportes, las conclusiones y recomendaciones para el desarrollo de futuras investigaciones.

1.5 Alcance

El modelo que se propone en este trabajo de grado no se enfoca en abarcar todos los dominios del conocimiento que se encuentran en la Web debido a su gran tamaño y complejidad, por tal motivo en este proyecto se analizan diferentes Ontologías de dominio existentes en la Web y a partir de este análisis se opta por enfocar las búsquedas en el dominio de la Oncología[3], en el cual se aplica el modelo semántico de expansión de consultas y se valida mediante la evaluación que se realiza al prototipo software. Cabe resaltar que para esta validación se hace necesario contar con personal que reúna los conocimientos necesarios en el dominio seleccionado y de esta manera establecer un adecuado proceso de validación.

1.6 Contenido de la Monografía

Capitulo II. MARCO TEÓRICO

En este capítulo se presentan las bases conceptuales que son importantes para el desarrollo del presente trabajo, las cuales hacen referencia a los principales núcleos temáticos como la Recuperación de Información y la Web Semántica, además se realiza el análisis del estado del arte destacando los trabajos de investigación relacionados con este proyecto.

Capitulo III. CREACIÓN DEL MODELO

Esta sección abarca principalmente la creación del Modelo Semántico de Expansión de Consultas (MSEC), aquí se tiene en cuenta cada uno de los elementos y sus relaciones para su construcción, y se definen cada uno de los módulos que lo componen.

[3] La oncología es la especialidad médica que estudia los tumores benignos y malignos, pero con especial atención a los malignos, esto es, al cáncer.

Capitulo IV. IMPLEMENTACIÓN DEL PROTOTIPO

En este capítulo se describe el proceso de implementación del prototipo MSEC Web Search, el cual se basa en el modelo propuesto en el Capítulo 2. Esta implementación está guiada por una metodología de desarrollo de software.

Capitulo V. VALIDACIÓN DEL PROTOTIPO

Este capítulo aborda la evaluación del prototipo descrito en el Capítulo 4 mediante una serie de pruebas compuestas principalmente por estadísticas y medidas de relevancia. Finalmente se presentan los resultados obtenidos a partir de la ejecución de dichas pruebas.

Capitulo VI. CUMPLIMIENTO DE OBJETIVOS

En esta sección se realiza un análisis detallado acerca del cumplimiento de los objetivos del proyecto.

Capitulo VII. CONCLUSIONES, RECOMENDACIONES Y TRABAJO FUTURO

Por último, en este capítulo se analizan los resultados del trabajo realizado, se detallan las principales contribuciones obtenidas en la ejecución del proyecto y se expone un conjunto de recomendaciones importantes para el desarrollo de trabajos futuros.

Capítulo II

2 MARCO TEÓRICO

Este capítulo contiene las bases teóricas sobre las cuales se encuentra enmarcado el presente proyecto, estas contemplan la Recuperación de Información y la Web Semántica. En esta fase se utiliza el modelo de investigación documental planteado por Hoyos y Serrano [11] y por medio de esta metodología se definen los núcleos temáticos y se construyen las respectivas fichas bibliográficas relacionadas con cada documento que ha sido consultado en las principales bases de datos documentales (ACM, IEEE, Science Direct, entre otras). Las fichas bibliográficas se describen en el Anexo A.

2.1 Recuperación de Información

La Recuperación de Información ó RI, es un tema de mucha importancia en nuestra sociedad actual, este se encuentra relacionado estrechamente con la información disponible en la Web y en la necesidad de indagar sobre herramientas que permitan gestionar, recuperar y filtrar esta información [12].

El término recuperación de información es usado muy frecuentemente, sin embargo, su definición puede ser muy amplia, generando cierta confusión. Algunos autores presentan su propio punto de vista, tal como se presenta a continuación:

- Salton [6], indica que la RI trata de la "representación, almacenamiento, organización y acceso a ítems de información".
- Meadow [13], describe la recuperación de la información como "una disciplina que involucra la localización de una determinada información dentro de un almacén de información o base de datos".
- Grossman y Frieder [14], enfatizan en que "la recuperación de información es encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits".
- Greengrass [15], la define como "la disciplina que trata con recuperar datos no estructurados, especialmente documentos de texto, en respuesta a una consulta o tema, la cual también puede estar no estructurada, como una oración u otro documento".
- Manning [10] plantea que la RI "consiste en encontrar material (usualmente documentos) de naturaleza no estructurada (usualmente texto) que satisface una necesidad de información dentro de una gran colección (usualmente servidores de computadora locales o en Internet)".

De acuerdo con los objetivos de este proyecto, la definición que mejor se ajusta al mismo es la propuesta por Salton, en la cual se considera que el principal problema de la RI es

presentar al usuario el subconjunto de documentos más relevantes que estén relacionados con la necesidad de información del usuario, además Baeza-Yates y Ribeiro-Neto [3] toman en cuenta la importancia de combinar diferentes técnicas de personalización como el perfil de usuario entre otras.

2.1.1 Sistema de Recuperación de Información

Tomando como punto de partida la definición de Recuperación de Información concebida por Saltón, un Sistema de Recuperación de Información o SRI, como se puede observar en la Figura 1, es entendido como "un conjunto de ítems de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo (SIMILAR) que determine qué ítems satisfacen las necesidades de información expresadas por el usuario en la petición" [6].

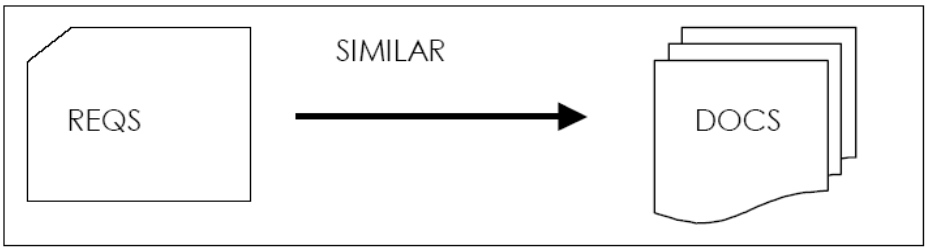

Figura 1. Esquema simple de un SRI.
Fuente Salton, G. and Mc Gill, M.J. Introduction to Modern Information Retrieval. New York: Mc Graw-Hill Computer Series, 1983.

Por otro lado, Baeza-Yates [3] entiende un SRI como se expresa en la siguiente definición: "dada una colección de documentos y una consulta formulada por un usuario en un cierto momento, proporcionar el subconjunto de documentos que es más relevante para la consulta del usuario". Adicionalmente identifica las siguientes funciones en este proceso:

- El usuario introduce una consulta en el sistema. Esta consulta representa sus necesidades de información.
- El sistema procesa dicha consulta. Sé buscan documentos que, de alguna forma, sean coincidentes con los términos que aparecen en dicha consulta.
- El sistema muestra los documentos que son coincidentes con la consulta, ordenándolos de mayor a menor relevancia según el valor proporcionado por una función de ranking[4].

Chowdhury [16] identifica la siguiente secuencia de funciones:

[4] En la RI, hace referencia a la lista de documentos ordenada de acuerdo a la relevancia de cada uno de estos, generalmente esta lista se ordena de mayor a menor relevancia.

- Identificar las fuentes de información relevantes a las áreas de interés de las solicitudes de los usuarios.
- Analizar los contenidos de los documentos.
- Representar los contenidos de las fuentes analizadas de una manera adecuada para compararlas con las preguntas de los usuarios.
- Analizar las preguntas de los usuarios y representarlas de una forma que sea adecuada para compararlas con las representaciones de los documentos de la base de datos.
- Realizar la correspondencia entre la representación de la búsqueda y los documentos almacenados en la base de datos.
- Recuperar la información relevante.
- Realizar los ajustes necesarios en el sistema basados en la retroalimentación con los usuarios.

2.1.1.1 Búsqueda Web

Desde la aparición de la World Wide Web (WWW o W3) en la década de los 90, se ha dado lugar a la creación de un SRI más avanzado y de mayor volumen denominado motor de búsqueda o buscador Web[5], el cual hasta el momento tiene la mayor parte de la responsabilidad en la búsqueda y localización de la información que se encuentra esparcida en internet. En la Figura 2 se presentan los componentes básicos de un buscador Web [17] [18]:

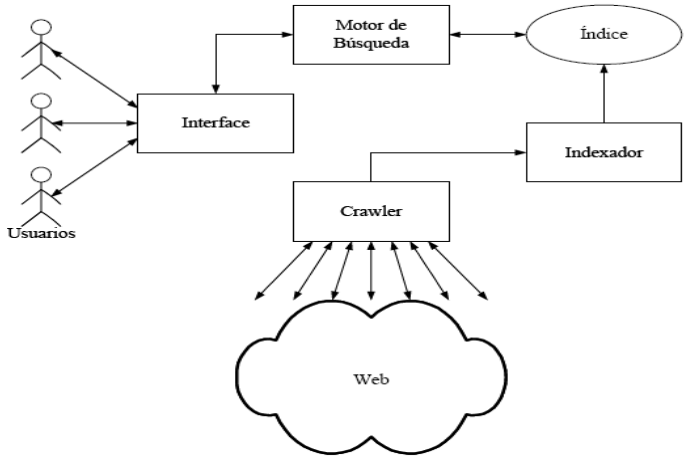

Figura 2. Funcionamiento de un buscador Web.
Fuente K. D. R. Benavides, "Introducción a la Recuperación de Información," Universidad de Costa Rica, 2010.

A continuación se describen los principales componentes de un buscador Web.

[5] Se refiere a los buscadores como Google, Bing, Altavista, entre otros.

- **Crawler.** Es un programa que se ejecuta localmente en la máquina de búsqueda, recorriendo la web por medio de peticiones a los servidores Web, transfiriendo el texto o contenido de las páginas Web que va encontrando. El Crawler comienza con un conjunto de URLs conocidas, dentro de las cuales encuentra direcciones de sitios Web y las adiciona a una lista de direcciones visitadas, conforme crece la lista es importante decidir el orden en que se visitan las paginas, para determinar dicho orden se utilizan políticas de crawling o medidas de popularidad que miden la importancia de una página y se mide como el numero de enlaces que apuntan a la pagina o como la cantidad de veces que la pagina es visitada y con base en estas medidas el objetivo del Crawler es recorrer las paginas que tienen mayor importancia.

- **Indexamiento.** Es el proceso de crear un índice de las páginas visitadas por el Crawler. Por lo general el buscador construye un índice invertido el cual almacena una lista de palabras encontradas en las páginas visitadas por el Crawler y cada palabra almacena los documentos donde esta aparece. Las palabras clave de una consulta son buscadas en el índice y se procede a interceptar sus listas de páginas correspondientes. La Figura 3 muestra un índice invertido.

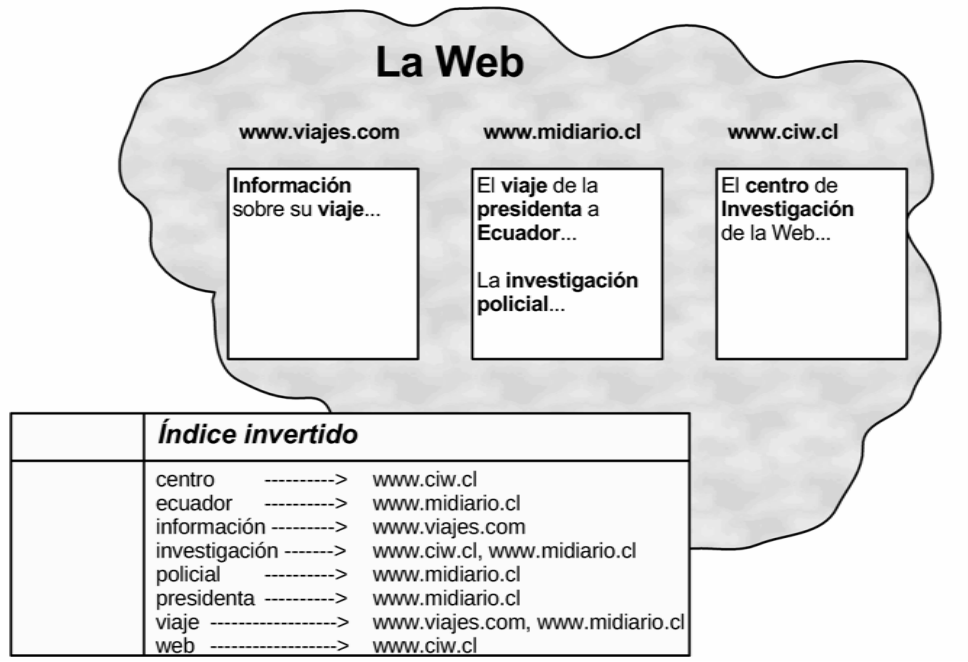

Figura 3. Ejemplo de un índice invertido para tres páginas Web.
Fuente G. Navarro, Como funciona la Web. Santiago de Chile, 2008.

- **Búsqueda.** Su objetivo es analizar la consulta de usuario por medio de operaciones como la eliminación de stopwords o palabras vacías[6] [10], extracción de raíces, entre otras y finalmente obtener las palabras clave de la consulta, las cuales son buscadas

[6] Palabras muy comunes y semánticamente no selectivas.

en el índice, proporcionando al usuario las correspondientes páginas relacionadas a cada palabra clave. Si la consulta tiene uno o más términos, el sistema deberá comparar los documentos relacionados a cada término, realizando una unión o intersección según corresponda. Luego estos documentos se ordenan según la relevancia estimada para cada uno.

Por otra parte, existen muchas formas de calcular la relevancia de los documentos, y en general, los buscadores utilizan fórmulas matemáticas para llevar a cabo este cálculo, teniendo en cuenta algunos de los siguientes aspectos los cuales son mencionados a continuación:

- La relevancia es mayor cuando las palabras aparecen en el título o primeras líneas en el documento.
- Frecuencia o número de apariciones de las palabras de la consulta en la página, a mayor frecuencia se da mayor relevancia, paginas con frecuencia excesiva se descartan.
- Cuanto mayor sea el número de hiperenlaces que apuntan a una página, mejor es esa página, el inconveniente es que no se distingue la importancia de la página donde aparece el hiperenlace.

Los aspectos mencionados anteriormente están limitados, debido a que en la Web hay mucha información que no se puede recuperar en su totalidad mediante la búsqueda de documentos que contengan ciertas palabras de una consulta. Este inconveniente se debe a que no es fácil implementar búsquedas más sofisticadas a gran escala, es por eso que actualmente se están realizando investigaciones que respondan a consultas más complejas a escala de la Web, como por ejemplo preguntas que se puedan inferir sin la necesidad de la cooperación del usuario tales como responder a la consulta: *“¿cuál es la farmacia más cercana que venda un antigripal a un precio inferior a $3.000?”* y *“¿qué universidades dictan una carrera de Diseño Gráfico de 5 años en la Región Metropolitana?”*.

En general, la mayoría de los motores de búsqueda realizan el proceso de RI descrito en la Figura 4.

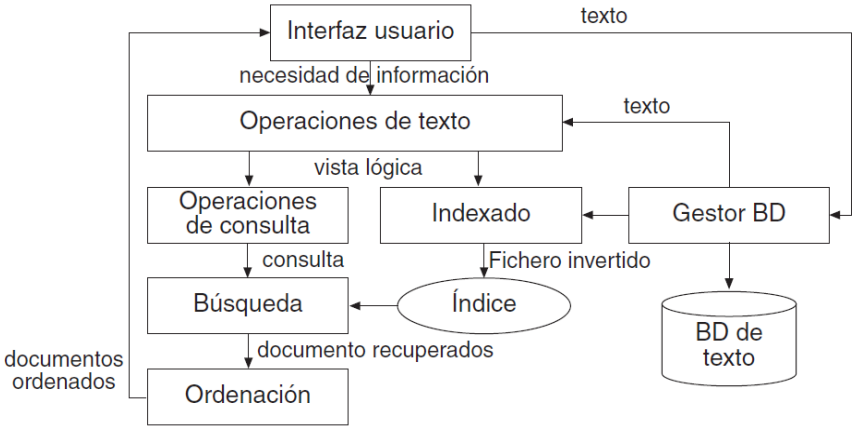

Figura 4. Proceso de Recuperación de Información.
Fuente F. Cacheda, "Introducción a los modelos clásicos de Recuperación de Información," in Revista General de Información y Documentación. vol. 18, 2008, pp. 365-374.

A continuación se describen las etapas del proceso de RI [3]:

- **Operaciones de texto.** En esta etapa se genera la representación o vista lógica de un documento como se muestra en la Figura 5, que por lo general es representado a través de términos indexados o palabras clave. Estos términos son extraídos con el siguiente proceso: Se eliminan y normalizan los términos del documento a ser indexado, luego se descartan las stopwords (la, las, y, los, en, etc.), posteriormente se eliminan acentos, espacios y se reducen las palabras a su raíz gramatical (stemming). Todo esto es necesario para evitar redundancia en el índice y lograr una optimización del mismo.

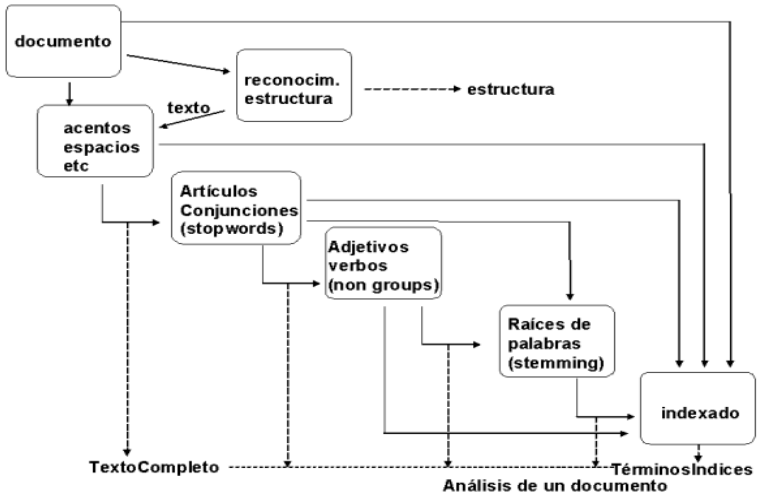

Figura 5. Vista lógica de un documento.
Fuente B. R.-N. R. Baeza-Yates, "Modern information retrieval," in Information Processing & Management, 1999, p. 453.

- **Base de Datos de Texto.** Se define la base de datos de texto y se realiza mediante el "Gestor BD" el cual especifica: los documentos que harán parte de la colección,

operaciones que se realizan sobre el texto de los documentos generando una vista lógica de los mismos y finalmente la estructura del texto y los elementos a recuperar, los cuales conforman el modelo del texto.

- **Indexado.** Después de obtener la vista lógica de los documentos se procede al proceso de indexación, el cual es el mismo que fue descrito anteriormente. Actualmente la estructura de datos más utilizada es el índice invertido.

- **Operaciones de consulta.** Para aplicar las operaciones de consulta es necesario que los documentos de la base de datos ya estén indexados y además es necesario que el usuario especifique su necesidad de información, la cual será analizada y transformada a través de las operaciones de consulta (operaciones lógicas o de conjunto).

- **Búsqueda.** Después de realizar las operaciones de consulta a la necesidad de información del usuario, se procede a realizar un proceso de búsqueda que básicamente es el mismo que fue descrito anteriormente.

- **Ordenación.** Una vez realizado el proceso de búsqueda, los documentos recuperados son ordenados de acuerdo a un criterio de relevancia basado en la consulta original de usuario, generalmente los motores de búsquedas utilizan el modelo booleano[7] o el modelo de espacio vectorial[8] [19].

- Finalmente los documentos ordenados mediante el proceso de ordenación son presentados al usuario el cual examina la información que considere de utilidad, y con base en este conjunto de recursos de interés, se inicia un ciclo de retroalimentación. En este ciclo, el sistema usa los documentos seleccionados por el usuario para modificar la formulación de la consulta con la posibilidad de obtener mejores resultados.

De acuerdo al proceso de RI en los buscadores Web, es importante resaltar la forma básica de ordenamiento de acuerdo a criterios de relevancia que utilizan los modelos de RI sobre los documentos. Considerando que existen librerías software como Lucene .Net [20] que implementan este proceso, y que además tienen en cuenta el modelo vectorial para dicho ordenamiento, es importante para este trabajo de grado utilizar herramientas que hagan uso de estas librerías, ya que facilitan una funcionalidad que brinda las bases fundamentales del proceso de RI, evitando implementarlas desde el inicio.

[7] Se basa en un criterio de decisión binario (pertinente o no pertinente) para saber si un documento tiene relación con una pregunta.

[8] Trabaja asignando pesos no binarios a los términos índice de las preguntas y de los documentos. Estos pesos se utilizan para comprobar el grado de similitud entre un documento guardado en el sistema y la pregunta realizada por el usuario.

2.1.2 Modelos para recuperación de información

Un modelo es una representación abstracta de un proceso, probablemente del mundo real. Actualmente se usan múltiples modelos para estudiar las propiedades de un proceso, obtener conclusiones y, en la mayoría de los casos, hacer predicciones. Se puede deducir que , las predicciones de un modelo son mejores, en cuanto más se ajuste a la realidad [12].

Teniendo en cuenta la definición de un modelo descrita en el párrafo anterior, el diseño de un SRI se realiza bajo un modelo, donde su proceso es definido como "la obtención de las representaciones de los documentos y de la consulta, la estrategia para evaluar la relevancia de un documento respecto a una consulta, los métodos para establecer la importancia (orden) de los documentos de salida y los mecanismos que permiten la retroalimentación por parte del usuario para mejorar la consulta" [21].

Con base en lo anterior, en la Tabla 1 se muestra una de las más completas descripciones de modelos de SRI realizado por Dominich [22] , quien estableció cinco grupos.

MODELO	DESCRIPCIÓN
Modelos clásicos	Incluye los tres más comúnmente citados: booleano, espacio vectorial y probabilístico.
Modelos alternativos	Están basados en la lógica Fuzzy
Modelos lógicos	Desarrollados en la década de los noventa, basados en la Lógica Formal. La RI se entiende como un proceso de inferencia a través del cual se puede estimar la probabilidad de que una necesidad de información de un usuario, expresada como una o más consultas, sea satisfecha ofreciendo un documento como "prueba".
Modelos basados en la interactividad	Incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la relevancia de los documentos recuperados [6].
Modelos basados en la inteligencia artificial	Bases de conocimiento, redes neuronales, algoritmos genéticos y procesamiento del lenguaje natural.

Tabla 1. Clasificación de los modelos de RI.
Fuente Dominich, S. "A unified mathematical definition of classical information retrieval". Journal of the American Society for information Science, 51 (7), 2000. P. 614-624.

En este proyecto se opta por utilizar los modelos clásicos de RI y los modelos basados en la interactividad que son presentados en la Tabla 1. El modelo de espacio vectorial es de gran importancia ya que es ampliamente usado y es una base para clasificar los documentos recuperados por un buscador Web de acuerdo a la relevancia de los mismos. Por otra parte, los modelos basados en interactividad se utilizan enfocados hacia la realimentación del perfil de usuario de acuerdo a las valoraciones que él mismo brinda a los documentos recuperados.

2.1.2.1 Perfil de Usuario

Un usuario es quien busca información en los sistemas de información. El Perfil de Usuario (PU), es un modelo que busca representar las características de un usuario dentro del sistema, con el fin de adaptar información respecto a dichas características. Entre estas se pueden mencionar los datos básicos, las preferencias, gustos y necesidades de cada usuario [23, 24].

A su vez, la necesidad de modelar el usuario se debe a que cada uno puede tener objetivos e intereses diferentes, incluso un mismo usuario podría tener intereses diferentes dependiendo del contexto en el que se encuentre [25].

Por otro lado, en el área de la RI el perfil de usuario es utilizado como una técnica en la personalización, además de la propia consulta, para estimar los intereses del usuario y de esta forma seleccionar los documentos más relevantes [26], en estos sistemas la consulta representa el interés actual del usuario, mientras que el perfil de usuario representa los intereses a largo plazo.

La forma de representar y extraer el perfil de usuario junto con la forma de combinarlo con la consulta actual, siguen siendo cuestiones abiertas en la literatura [27]. Generalmente el perfil de usuario en sistemas de filtrado de información es considerado la base para filtrar la información que responde a los intereses a largo plazo del usuario, además requiere múltiples interacciones en repetidas sesiones y termina al obtener un numero de referencias a documentos relativamente reducido, o cuando no se pueden realizar mayores filtrados [28].

2.1.2.1.1 Adquisición y Actualización del Perfil de Usuario

Un perfil de usuario puede ser creado o actualizado de dos formas o métodos en la recolección de características; implícita y explícitamente. Estos métodos son descritos a continuación:

- **Implícito.** No requiere participación del usuario ni su consentimiento, transparente (menos intrusivo), es decir la reacción del usuario ante cada documento entrante es guardada [29]. Algunos enfoques para adquirir el perfil de usuario de esta forma realiza análisis de favoritos, marcadores, documentos del historial de navegador y tiempos de acceso, además de analizar las operaciones de búsqueda del usuario como salvar, imprimir, editar, copiar, y realizar un seguimiento de los movimientos de mouse y scrollbar, o incluso el tiempo de navegación entre otros [30].

- **Explicito.** Requiere la participación activa del usuario, es decir controla la información del perfil por medio de formularios y cuestionarios, encuestas, recomendaciones seleccionadas y valoración explícita entre otros [31]. Esta forma de crear el perfil

presenta inconvenientes como lo es la valoración explicita, puesto que el usuario debe calificar manualmente todos los documentos que recibe; esto sólo sería práctico si el usuario recibe pocos documentos [9],| y en cuanto a cuestionarios, algunos usuarios no están dispuestos a invertir tiempo en completar los mismos. Esta forma de crear el perfil de usuario es muy usado, ya que proporciona mayor precisión en los resultados [32].

2.1.2.1.2 Representación del Perfil de Usuario

La mayoría de los sistemas estudiados representan los perfiles mediante términos o páginas favoritas, que luego son utilizados en diversas operaciones de personalización por ejemplo: expansión de la consulta, re-ponderado de términos, asignación de pesos a los documentos, retroalimentación explícita o implícita de usuario, ambigüedad de los términos y re-ranking entre otros [33].

Algunos enfoques representan el PU de diversas formas y son mencionados a continuación:

- Un conjunto de palabras clave ponderadas o conceptos que representan las características de los documentos en los que el usuario está interesado [34].
- Una lista de documentos, donde por cada documento hay una indicación que refleja si el documento es relevante o no para el usuario [34].
- Un conjunto de categorías y cada una tiene una colección de términos ponderados que están relacionados con la categoría [35].
- Un conjunto de consultas anteriores del usuario, almacenadas en su perfil [36].
- Para sistemas de filtrado de información, la estructura más común es la denominada “bag of words”, la cual consiste en un conjunto de palabras clave que representan los intereses del usuario [37].

2.1.2.1.3 Modelos para el Perfil de Usuario

Algunas investigaciones desarrollan modelos que involucran el perfil de usuario al igual que técnicas como la expansión de consulta, los cuales tienen una forma distinta de construirlo, pero en esencia el objetivo es recuperar información relevante para el usuario de acuerdo a sus intereses. Algunos modelos que incluyen el perfil de usuario son mencionados y descritos a continuación.

- Sieg, Mobasher y Burke [38] presentan una aproximación a las búsquedas personalizadas que implica la construcción de modelos de contexto de usuario, como perfiles ontológicos mediante la asignación implícitamente de pesos de interés a conceptos existentes en una ontología de dominio; por lo tanto un algoritmo de activación de propagación es utilizado para mantener los pesos de interés basado en

el comportamiento actual del usuario. Los experimentos en este trabajo muestran que la clasificación de resultados de búsqueda basados en pesos de interés y la semántica evidente en un perfil de usuario ontológico es efectiva en la presentación de los resultados más relevantes para el usuario.

- Chang y Ma [39] proponen un método de expansión de consulta que consiste en una serie de términos relevantes para el usuario, los cuales se almacenan en una Base de Datos, estos sirven como soporte para la expansión de consulta y posteriormente se utiliza el modelo espacio vectorial para clasificar los documentos.

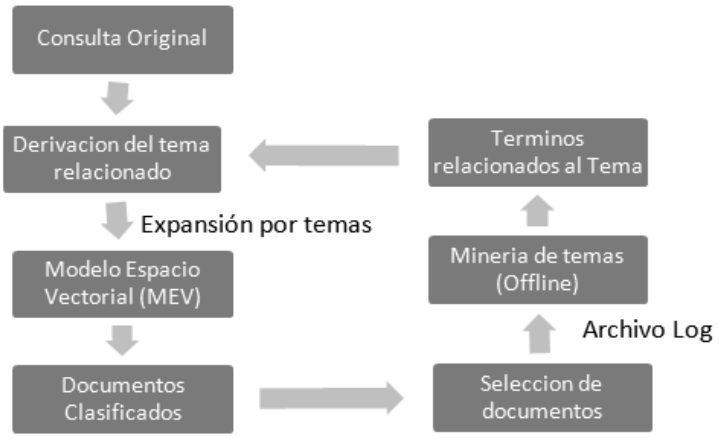

Figura 6. Theme-based Query Expansion by Mining Log Data.
Fuente Peng Chang and H. Ma, "Theme-based Query Expansion by Mining Log Data," in Wireless Communications, Networking and Mobile Computing, 4th International Conference 2008.

La Figura 6 describe los componentes de un framework de expansión de consulta basado en temas. El SRI registra los resultados de búsqueda seleccionados por el usuario. Un método de minería de datos es empleado para encontrar temas relacionados a la consulta dada. Este trabajo es hecho en línea y los resultados extraídos serán guardados en una base de datos. Cuando un nuevo usuario llega, los temas serán derivados por la ocurrencia de temas de términos en la base de datos. Finalmente la consulta original y la expansión de temas serán tratadas como un vector de consultas que será introducido en el modelo de espacio vectorial. Los documentos se clasifican con las medidas de semejanza más altas las cuales son la salida al usuario.

- Porwol [40] usa un algoritmo denominado SQE, el cual es una combinación de búsqueda de texto completo y varias tecnologías de la Web semántica incluidas para la expansión de la consulta. Consiste en la utilización de retroalimentación de relevancia, en la que el usuario mantiene un ciclo de iteraciones con el sistema de recuperación, refinando la consulta con los documentos que el usuario marca como relevantes (retroalimentación explicita).

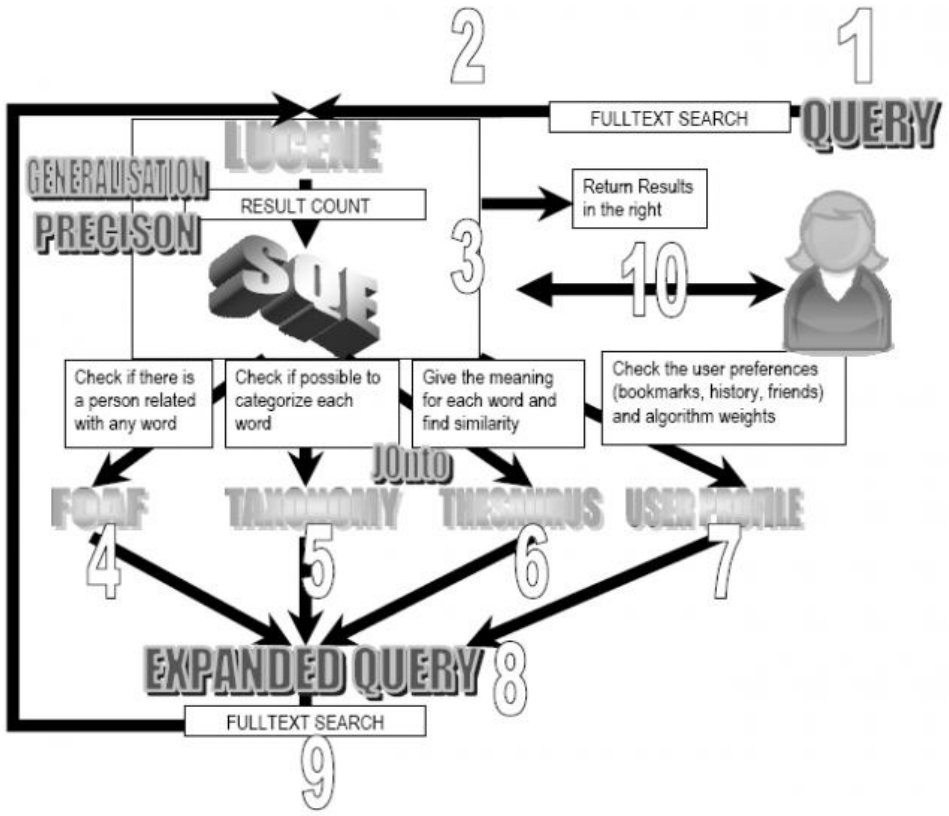

Figura 7. Esquema del Algoritmo para el prototipo SQE implementado.
Fuente L. Porwol, "SQE – Semantic Query Expansion as Search Process Booster " in 2010 Semantic Technology Conference San Francisco, CA, 2009.

En la Figura 7 se muestran los pasos que corresponden al esquema del algoritmo SQE, cada uno de estos se describe a continuación.

1. El usuario envía la consulta original.
2. La consulta es enviada al modulo de búsqueda de texto completo.
3. Los resultados son contados y retornados al usuario con una clasificación asignada.
4. Las entradas personales al repositorio de la red social son asignadas a las palabras de la consulta dada.
5. La taxonomía categoriza cada palabra de la consulta.
6. El tesauro encuentra el significado de cada palabra de la consulta y proporciona su similitud.
7. La expansión es comparada con el historial, con marcadores de usuario[9] y con los marcadores de usuarios amigos. Los elementos que son similares a los marcadores o al historial, consiguen ser clasificados.
8. Se construye la estructura de la consulta y las clasificaciones iniciales son contadas y asignadas a los elementos de la consulta.

[9] Representa las preferencias de usuario.

9. La estructura de la consulta es transformada a una cadena y es enviada al motor de búsqueda.

- Calegari y Pasi [36] proponen un análisis preliminar para la definición de un nuevo gadget[10] [41] de escritorio de Google denominado "Personal Information Context" (PIC) basado sobre una Ontología Fuzzy Local Personalizada (O-FCN)[11] para expandir una consulta semánticamente. Esta técnica de consulta expandida hace uso del contexto de usuario tomando en cuenta consultas de usuario pasadas almacenadas en su perfil.

- Zhu, Xu, Ren, Tian y Li [35] plantean que la construcción del perfil de usuario viene dada cuando el número de páginas Web que un usuario ha navegado es mayor que un umbral especificado, el agente de aprendizaje adquiere perfiles de usuario por medio de un algoritmo de agrupamiento de documentos Web [38] que combina un algoritmo genético con K-means[12] [42]. La tarea del algoritmo es formar jerarquías de categorías que representan los intereses del usuario.

- Wei, Huang y Tan [30] tratan de resolver la desventaja de la falta de información semántica de palabras clave, al diseñar un método de moldeamiento del perfil de usuario soportado en la base de conocimiento de categorías. El modelo y los perfiles de usuario están constituidos por intereses a corto y a largo plazo y dos diferentes mecanismos de creación y actualización se adoptan para estas dos clases de interés.

Para resumir los anteriores trabajos enfocados en modelos de perfil de usuario, se realizó la Tabla 2 que reúne las principales características de los trabajos considerados más importantes.

Propuesta	Objetivo	Valoración Usuario	Perfiles de Usuario
Theme-based Query Expansion by Mining Log Data [39]	Expandir consultas mediante extracción de log archivo (temas conformados de términos derivados de documentos recuperados).	Implícita: Asume que todo documento seleccionado es relevante.	Implícito a partir de la identificación del tema de sesión del usuario mediante los recursos seleccionados
SQE – Semantic Query Expansion as Search Process Booster [40]	Expandir consultas mediante(tesauro, taxonomía, FOAF, perfil de usuario, retroalimentación de relevancia)	Implícito: No especifica el procedimiento o método.	Implícita: No se especifica el procedimiento.

[10] Son objetos en miniatura realizados para ofrecer contenido fresco y dinámico que puede ser colocado en cualquier página en la web.

[11] Por sus siglas en inglés, Personalized Local Fuzzy Ontology.

[12] Es un algoritmo para clasificar o agrupar objetos constituidos de características o atributos dentro de k grupos, donde k es un número entero positivo.

A Personalized Model for Ontology-driven User Profiles Mining [30]	Mejorar la precisión del perfil de usuario mediante una ontología	Implícita: No especifica el procedimiento.	Implícita a partir de la frecuencia de los términos de la consulta en los documentos formando categorías.

Tabla 2. Proyectos que usan modelos que involucran el perfil de usuario para la RI.

De la Tabla 2 se puede concluir, que los trabajos relacionados contemplan la valoración del usuario sobre el recurso recuperado, al igual que la manera de adquirir el perfil, ambas de forma implícita, lo cual permite que este proceso sea transparente al usuario sin la necesidad de una interacción explicita por parte del mismo, lo que puede implicar que se descarte cierta información que posiblemente es importante para refinar el perfil. A pesar de que el perfil y la valoración del usuario se realizan de forma implícita, no se especifica un análisis a profundidad del procedimiento que el sistema tiene que realizar para llevar a cabo tal valoración, a excepción del último trabajo, e igualmente la adquisición del perfil de usuario.

2.1.2.2 Expansión de Consulta

Las consultas realizadas por los usuarios, generalmente son cortas y poco especificas. A pesar de eso, los usuarios esperan que su búsqueda arroje el resultado que ellos quieren, y que estos se encuentren en los primeros resultados mostrados [43], es por eso que se hace necesario la expansión de la consulta.

En el entorno de búsqueda tradicional, el usuario debe dividir su interés de búsqueda en distintos conceptos. Luego debe pensar en cómo los conceptos y los términos asociados con ellos corresponden a la representación de la información almacenada. Una vez que los términos han sido elegidos, pueden ser combinados para formar la consulta.

En ese mismo entorno, el usuario establecía su consulta de búsqueda en distintos conceptos de interés combinados coherentemente, para representar su necesidad de información, pero no siempre un término representa en forma adecuada un concepto de interés para el usuario [2]. Encontrar otros términos similares para expresar un concepto es realizar una expansión de consulta [44], existen métodos para expandir la consulta tales como la retroalimentación por relevancia, que permite seleccionar documentos previamente recuperados como relevantes o irrelevantes por el usuario a partir del resultado de una consulta inicial, de modo que con base en estos, se reformule la consulta y se recuperen más documentos relevantes que irrelevantes [45], otro método es disponer de una estructura de conocimiento que sea independiente del proceso de búsqueda, tales como los siguientes recursos lingüísticos:

- **Diccionario.** Indican los distintos significados de un término y permiten su expansión con sinónimos. Un diccionario muy utilizado como recurso es WordNet[13] [46], que es un sistema de referencia léxica y proporciona diferentes acepciones de un concepto, permitiendo además la expansión de éste con sinónimos[14] [47], merónimos[15], hipónimos[16] [48] y otros tipos de términos relacionados a la acepción elegida.

- **Tesauro.** Representa el conocimiento de un dominio con una colección de términos y un limitado conjunto de relaciones entre ellos. Estos han sido extensivamente usados en sistemas de RI en las últimas dos décadas. Sin embargo, se han establecido deficiencias con los tesauros. El principal problema es la limitación sobre el conjunto de relaciones, esto algunas veces oculta relaciones actuales y establece relaciones ambiguas entre términos [49].

- **Ontología.** Una ontología es una especificación explícita de una conceptualización, es decir proporciona una estructura y contenidos de forma explícita que codifica las reglas implícitas de una parte de la realidad, independientemente del fin y del dominio de la aplicación en el que se usarán o reutilizarán sus definiciones [50].

Un problema en cualquier tipo de expansión de consulta, es cómo definir cuales términos están estrechamente asociados con los términos de la consulta, para ello se consideran dos tipos de expansión de consulta, estos se mencionan a continuación:

- **Expansión de consulta interactiva.** Al usuario se le proponen términos de búsqueda como parte del proceso de reformulación de la consulta. En este tipo de expansión existe una interacción mutua entre el sistema y el usuario en la selección de términos para la expansión, lo cual implica que la responsabilidad para mejorar la consulta proviene de ambas partes. Por el lado del sistema, este sugiere términos y los presenta al usuario; y por el otro, el usuario es el que toma la decisión final sobre la importancia relativa y la utilidad de un término.

- **Expansión de consulta automática.** En este tipo de expansión los términos son adicionados por el sistema de recuperación, las técnicas utilizadas en este tipo de expansión tienen significativas ventajas frente a la expansión de consulta interactiva como lo es la técnica de retroalimentación por relevancia, ya que no requiere esfuerzo por parte del usuario [51].

[13] Es uno de los tesauros más usados en la Red y permite obtener los sinónimos, hipónimos, hiperónimos, etc., de una palabra dada.

[14] Adjetivo que se utiliza para expresar que un vocablo tiene una misma o muy parecida significación que otro.

[15] Palabra cuyo significado constituye una parte del significado total de otra palabra, denominada esta holónimo (p.e. dedo es merónimo de mano, esta es merónimo de brazo, y brazo es holónimo de mano y mano es holónimo de dedo).

[16] Palabra que posee todos los rasgos semánticos de otra más general, su hiperónimo, pero que añade en su definición otros rasgos semánticos que la diferencian de la segunda (p.e. lunes, martes son hipónimos de día).

Algunas de estas técnicas se usan en información contextual como el perfil de usuario, la vista de eventos del historial del usuario, y tareas de usuarios anteriores para expandir la consulta [52, 53].

2.2 La Web Semántica

Desde la introducción de los ordenadores en la segunda mitad del siglo XX y en especial con la creación de la Web, se han establecido nuevas disciplinas para la gestión de la enorme cantidad de información disponible, una de ellas es la recuperación de información, la cual cada vez más, requiere de distintas estrategias que le permitan satisfacer las necesidades de información de cada uno de los millones de usuarios de la Internet [48], es así como el concepto de Web Semántica se convierte en una opción para tal fin.

Uno de los primeros autores que trató con el tema de Web Semántica fue Tim Berners-Lee, considerado el padre de la Internet, quien en 2001 anunció el proyecto que pretende constituir a la Web Semántica como una extensión de la Web actual[54], en la cual los ordenadores sean capaces de reconocer y comprender la semántica ofrecida por el contenido de las páginas Web a través de una estructura que permita describirlo; de igual forma se permitiría una mejor interoperabilidad entre máquinas[17], además del enriquecimiento de la interacción que un usuario puede tener con el ordenador. Berners-Lee supuso que en un futuro no muy lejano, los ordenadores podrían tener acceso a información marcada semánticamente, ontologías que expresaran conceptos y a un conjunto de reglas de inferencia las cuales permitirían llevar a cabo razonamiento automático sobre las páginas Web para que los ordenadores puedan efectuar tareas inteligentes.

Es preciso aclarar que aún en nuestros días, la Web Semántica es solo una visión, más que una realidad; esta visión involucra la capacidad del software para proporcionar al usuario los recursos de la Web que sean realmente relevantes para sus necesidades, de tal forma que la información contenida en dichos recursos pueda ser extraída, integrada e indexada para poder procesarla de una manera efectiva y eficiente, y así ejecutar sofisticadas tareas para los humanos [55, 56]. Para que esta visión se pueda convertir en una realidad, se necesita un gran esfuerzo por parte de toda la comunidad que realiza algún uso de la Web, es decir, las compañías deben cooperar mutuamente, las investigaciones académicas se deben transformar en sistemas prácticos, y en sí, cada individuo debe descubrir cómo puede contribuir para que el proyecto de la Web Semántica pueda seguir expandiéndose de tal forma que llegue a ser parte de nuestra vida diaria [57].

La Web Semántica es actualmente objeto de diferentes investigaciones [58-63] [64] que pretenden resolver multitud de problemas en distintos campos de aplicación como

[17] Ordenadores, Dispositivos Móviles, entre otros.

agentes inteligentes, integración de información, mediación y almacenamiento, infraestructura y metadatos, razonamiento y representación del conocimiento, ontologías y lenguajes; todos estos temas conciernen con el desarrollo y la implementación de nuevos métodos y tecnologías, además de estas áreas, la Web Semántica también está relacionada a aspectos como los socio-culturales, la confiabilidad, los modelos de crecimiento económico, entre otros [65].

2.2.1 Anotación Semántica

Cuando se habla sobre anotación, se alude a muchas definiciones como "el proceso de adaptar comentarios, críticas o notas explicativas" [66], "el acto de añadir notas" [67], "un nuevo nodo de comentario enlazado a un nodo existente" [68]; de esta forma, partiendo de las anteriores definiciones, una anotación puede ser considerada como una información sobre las entidades o conceptos de una ontología que aparecen en un texto y su situación en el mismo, o también las referencias que hay en un texto sobre un repositorio semántico en el que hay más conocimiento como se muestra en la Figura 8.

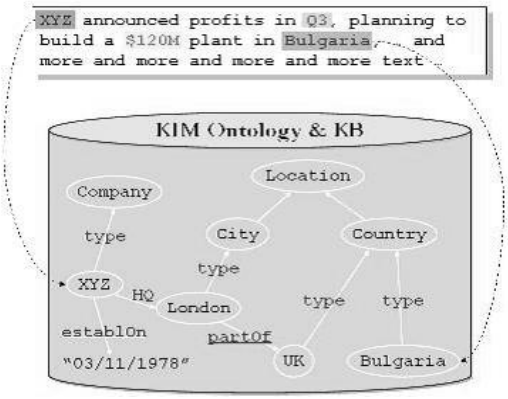

Figura 8. Anotación Semántica de una Ontología.
Fuente: http://www.infor.uva.es/~sblanco/Tesis/Anotaciones%20Sem%C3%A1nticas.pdf

En la actualidad la W3C[18] [69] está trabajando en un sistema experimental de anotación denominado *"Annotea"* [70], este se basa en el formato RDF para describir las anotaciones. El contenido de cada anotación es un pequeño documento XHTML, así mismo estas pueden ser almacenadas y/o cargadas localmente o desde un servidor de Annotea al navegador Web[19] [71].

Se debe aclarar que una anotación no se refiere en todo el sentido de la palabra al concepto de metadatos, puesto que estos describen datos acerca de los datos como por ejemplo el autor o editorial de un libro, y están más ligados al tema del documento, por otro lado las anotaciones están más relacionadas con la forma de pensar, la experiencia y

[18] World Wide Web Consortium. http://www.w3.org
[19] Experimentalmente W3C trabaja con su propio navegador Web denominado "Amaya"

los sentimientos de terceras partes hacia lo que está siendo objeto de la anotación, es decir su propia perspectiva o punto de vista personal [57].

2.2.2 Ontologías

El referente teórico que enmarca el núcleo temático correspondiente a las Ontologías, posee una gran cantidad y variedad de documentación, y buena parte de esta, se encuentra en Internet, especialmente haciendo referencia a su importancia para el adecuado manejo y representación del conocimiento [57, 65].

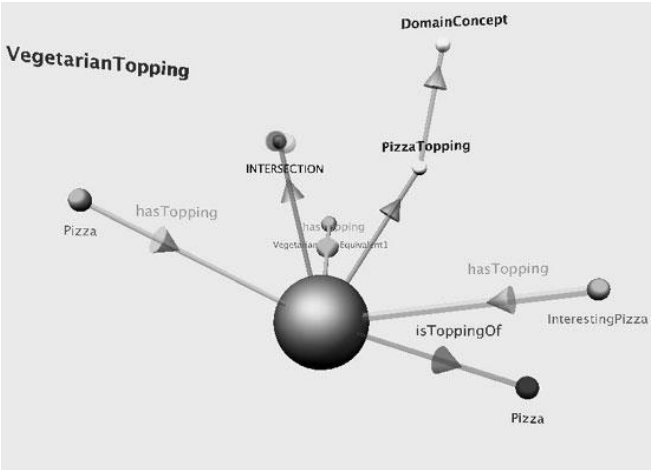

Figura 9. Esquema de una Ontología de dominio.
Fuente: http://www.madrimasd.org/gestion2006/img/Noticias/ciencia-ontologia-UC3M.jpg

Al hablar de una Ontología, como por ejemplo la que se muestra en la Figura 9, se debe hacer referencia a la definición planteada por Gruber [50], en la cual concibe este término como una "especificación formal y explícita de una conceptualización compartida". Para comprender mejor esta definición, una "Conceptualización" es un modelo abstracto de cualquier fenómeno del mundo construido mediante la identificación de los conceptos[20] [72] relevantes de tal fenómeno, y es "Compartida" ya que contiene conocimiento consensuado por parte de una comunidad de expertos en determinado dominio. Cuando se habla de "formal", se hace referencia a que la ontología debe ser expresada en un lenguaje de representación del conocimiento que proporcione una semántica definida, lo que permite que sea procesable a nivel computacional [73, 74]; y es una especificación "explícita" porque no existen ambigüedades al definir los conceptos y las restricciones sobre el uso de los mismos [72, 75].

Teniendo en cuenta el párrafo anterior, una Ontología se convierte en una forma de representar el conocimiento, de modo que la información, almacenada en repositorios de datos, pueda cobrar significado para las máquinas, y así convertirse en conocimiento,

[20] Un concepto se refiere a la representación del significado de una cosa.

posibilitando la realización inferencias a partir de axiomas sobre un dominio en particular [76].

Uno de los principales y más complejos problemas que enfrenta la Web actualmente es la integración semántica. De igual forma que la integración de bases de datos, la integración semántica es un problema que puede ser abordado con la integración espacial de datos [77], no obstante la heterogeneidad semántica [78], o como es denominada por algunos autores, inconsistencia [79], necesita de la intervención de muchas más técnicas y procedimientos que permitan disminuir la ambigüedad en el contenido de los recursos Web. Las Ontologías resultan ser útiles para la integración semántica de información [80] ya que proporcionan un significado común para cada concepto dentro de un determinado contexto, al igual de definir el tipo de propiedades y relaciones que existen entre estos; pero más que estos beneficios, es la lógica quien juega un papel determinante cuando se habla de las Ontologías, pues es esta combinación la que brinda la posibilidad de inferir nuevo conocimiento y así contribuir con el proceso de integración semántica.

2.2.2.1 Componentes de una Ontología

Las Ontologías deben presentar esencialmente dos tipos de componentes: elementos y relaciones entre los mismos, sin ellos no se podría considerar a una especificación de un dominio como una Ontología [48]. A continuación se describen estos componentes.

- **Elementos.**
  - **Conceptos o Clases.** Son entidades del mundo real categorizadas en grupos o conjuntos de objetos con características similares. Estas entidades pueden ser cosas físicas por ejemplo *"Carro"*, *"Árbol"*, *"Libro"*, o conceptuales como *"Teoría"*, *"Idea"*, etc.

  - **Instancias u Objetos.** Son individuos o ejemplares que representan a determinada clase de forma concreta. Por ejemplo *"Roble"* es una instancia de la clase *"Árbol"*.

  - **Propiedades.** Las entidades que pertenecen a una clase poseen atributos determinados, por ejemplo tienen un nombre, un color o un peso. Por tanto, las propiedades consisten en pares de atributo/valor y sirven para describir de forma conveniente las características relevantes de las entidades que forman las clases.

- **Relaciones.**
  - **Relaciones.** Constituyen el enlace entre los conceptos de una Ontología, al igual que la interacción entre los mismos. Algunos tipos de relación pueden ser: subclase-de, parte-de, parte-exhaustiva-de, conectado-a, etc. La relación puede

tener un dominio y un rango específico, el dominio hace referencia a la clase o clases iniciales y el rango a la clase o clases finales formando una relación [81].

- **Funciones.** Son un tipo concreto de relación donde se identifica un elemento mediante el cálculo de una función que considera varios elementos de la ontología. Por ejemplo, pueden aparecer funciones como: asignar-fecha, categorizar-clase, etc.

Además de los dos anteriores tipos de componentes de una Ontología, se deben mencionar también las reglas de restricción [57, 72], las cuales son axiomas o expresiones que son siempre ciertas, estas resultan útiles en varios aspectos como definir el significado de los componentes de la Ontología o definir restricciones complejas sobre los valores de los atributos, argumentos de relaciones, etc.

2.2.2.2 Lenguajes de Marcado para definir Ontologías

Las ontologías pueden ser modeladas con diferentes técnicas de modelado de conocimiento, a lo largo de los años se han diseñado diversos formalismos y lenguajes que permiten modelar de un modo formal el conocimiento, entre las técnicas diseñadas están [82]:

- **Resource Description Framework (RDF).** RDF es un Marco para la descripción de recursos que establece un modelo de datos para objetos (recursos) y relaciones entre ellos. Desarrollado por proporcionando una semántica simple para el mismo. Este tipo de modelo de datos puede ser representado con una sintaxis XML y se basa en la idea de convertir las declaraciones de los recursos en expresiones con la forma sujeto-predicado-objeto[21]. El sujeto es el recurso, es decir aquello que se está describiendo. El predicado es la propiedad o relación que se desea establecer acerca del recurso. Por último, el objeto es el valor de la propiedad o el otro recurso con el que se establece la relación. La combinación de RDF con otras herramientas como RDF Schema y OWL permite añadir significado a las páginas, y es una de las tecnologías esenciales de la Web Semántica [83].

  Si bien RDF permite dar valores a las distintas propiedades de diferentes recursos, no dispone de mecanismos para describir esas propiedades ni las relaciones existentes entre ellas y otros recursos. Para tal fin es necesario un lenguaje que permita definir vocabularios RDF. Dicho lenguaje, construido mediante RDF, es RDF Schema o RDFS [84]. Este lenguaje define clases y propiedades que permiten, a su vez, describir nuevas clases, propiedades y recursos. Sin embargo, tampoco RDF ni RDF Schema son capaces por sí solos de modelar ontologías, por tal razón se comienzan a desarrollar nuevos lenguajes para este fin, con la diferencia de que estos se construyen sobre el estándar RDFS.

[21] Conocidas en términos RDF como tripletes.

- **Web Ontology Language (OWL).** Este lenguaje de ontologías Web, es un lenguaje de etiquetado semántico para publicar y compartir ontologías en la World Wide Web. OWL tiene como objetivo facilitar un modelo de marcado construido sobre RDFS y codificado en XML. Se espera que, junto al entorno RDF y otros componentes, estas herramientas hagan posible el proyecto de la Web Semántica [85]. El lenguaje OWL tiene tres sub-lenguajes que incrementan su expresión, cada uno con nivel de expresividad mayor que el anterior: OWL Lite, OWL DL, y OWL Full[86], cada uno de estos se describe a continuación.

    - **OWL Lite.** Está diseñado para aquellos usuarios que necesitan principalmente una clasificación jerárquica y restricciones simples.

    - **OWL DL.** Está diseñado para aquellos usuarios que quieren la máxima expresividad conservando completitud computacional[22] y resolubilidad[23].

    - **OWL Full.** Está dirigido a aquellos usuarios que necesitan la máxima expresividad y la libertad sintáctica de RDF pero sin garantías computacionales. Permite, por ejemplo, aumentar el significado de vocabulario predefinido (en RDF o en OWL), por lo que es muy improbable que ningún software de razonamiento sea capaz de soportar razonamiento completo para cualquier característica de OWL Full.

2.2.2.3 Inferencia de nuevo conocimiento

Los elementos conceptuales de una Ontología son definidos axiomáticamente, esto permite que se pueda realizar inferencia computacional debido a que reduce la redundancia en la representación de una base de conocimiento y facilita su mantenimiento al no existir la necesidad de afirmar explícitamente lo que ya está especificado en la Ontología; de este modo la habilidad de utilizar los ordenadores en la deducción de conocimiento adicional basado en el contenido axiomático de una Ontología, es de gran valor y genera interés desde la perspectiva de la investigación [87].

2.2.3 Las Ontologías como soporte para la RI

Teniendo en cuenta la importancia de la Web Semántica y específicamente de las Ontologías para la búsqueda Web, en la Tabla 3 se presenta un cuadro comparativo en el cual se hace referencia a algunos proyectos destacados en este ámbito.

[22] Se garantiza que todas las conclusiones sean computables.
[23] Todos los cálculos se resolverán en un tiempo finito.

PROPUESTA	CARACTERÍSTICAS		
	Expansión Semántica de Consulta	**Relevancia de los Documentos Recuperados**	**Inferencia de nuevo Conocimiento**
Meta Web search model based on Ontologies, taxonomies and user feedback [88]	Usa Taxonomías[24] de Conocimiento Generales y Ontologías para encontrar la semejanza semántica[25] entre los documentos recuperados por los motores de búsqueda y la consulta hecha por el usuario	Permite reorganizar y filtrar los resultados recuperados por los buscadores Web; Sin embargo, la calidad de los resultados puede afectarse, puesto que depende de la ontología.	No realiza inferencia
Ontology based semantic information retrieval [89]	La consulta enviada por el usuario es expandida con sinónimos[26] y su similitud semántica[27].	No se menciona	Sólo se limita a los sinónimos de las palabras clave.
An Ontology-Based Information Retrieval Model [90]	El sistema procesa la consulta semántica contra la base de conocimiento, que devuelve un conjunto de instancias; estas pueden ser vistas como una forma de consulta expandida, donde el conjunto de instancias representan un nuevo conjunto de términos para la consulta.	El modelo está basado en una adaptación del modelo vectorial clásico: Esto incluye un algoritmo de tanteo de pesos y un algoritmo de clasificación.	La expansión implícita de la consulta está hecha usando reglas de inferencia.
SIRO: On-line Semantic Information Retrieval using Ontologies [91]	Utiliza Ontologías para reformular las consultas. Se realiza un análisis semántico para determinar si coinciden los conceptos que aparecen en la Ontología del dominio y la consulta de usuario.	Utiliza el modelo vectorial y la distancia del coseno para calcular la similitud entre documentos, los documentos más similares son recuperados.	Se apoya en WordNet[28] que se usa para encontrar sinónimos, e hipónimos[29] de palabras clave en la búsqueda.

Tabla 3. Proyectos que hacen uso de las Ontologías para la RI.

Como se observa en la Tabla 3, los trabajos propuestos analizan la consulta original encontrando la semántica de sus términos individualmente y por ende ninguno considera las interrelaciones semánticas de todos los términos de la consulta original vistos como un todo semántico, es por esto que el presente proyecto involucra el procesamiento semántico de la consulta que permita su refinamiento y evaluar sus resultados por medio de las medidas de evaluación de RI mencionadas en el Capítulo 5.

[24] Conjunto de conceptos organizados jerárquicamente. Las taxonomías definen las relaciones entre los conceptos, pero no los atributos de éstos.

[25] Basada en la distancia del coseno.

[26] Los sinónimos tienen distancia cero (0) y una similitud uno (1).

[27] Dos conceptos son semánticamente similares si comparten la misma super-clase.

[28] http://wordnetweb.princeton.edu/perl/webwn

[29] Palabra que posee todos los rasgos semánticos de otra más general.

2.2.4 Similitud Semántica en las Ontologías

La similitud semántica representa un caso especial de relación semántica[30] que se establece entre dos elementos con significado [92] y se orienta hacia qué tanto tienen en común ambos elementos, es decir, qué tan relacionados están [93]. En la Figura 10 se puede apreciar la jerarquía de conceptos en una Ontología, intuitivamente se identifica que los conceptos “Avión” y “Helicóptero” son semánticamente más similares por el hecho de ser tipos de vehículos aéreos, sin embargo, la pareja de conceptos “Avión” y “Automóvil” no se encuentran tan estrechamente relacionados. Dado lo anterior, para determinar el grado en que dos conceptos se encuentran relacionados en una Ontología, se hace necesario introducir una medida de similitud que permita cuantificar adecuadamente tal nexo.

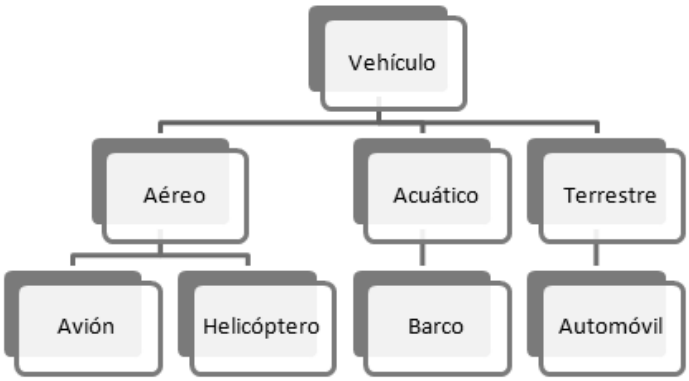

Figura 10. Representación jerárquica de una Ontología.

2.2.4.1 Medidas de Similitud Semántica

La utilización de medidas de similitud semántica se enmarca en un campo de investigación con muchos años de experiencia, en 1965 se destaca el trabajo de Rubenstein y Goodenough [94] en el que se aborda la similitud entre términos del idioma inglés dependiendo del contexto en el que aparecen. Esta propuesta sentó una base para los posteriores estudios, tal es el caso de los experimentos realizados por Miller y Charles [95] en los cuales se mide la similitud semántica sobre el diccionario global WordNet. En la Tabla 4 se presentan las medidas más destacadas que se han considerado para analizar la similitud semántica entre los conceptos de una Ontología.

Medida Similitud Semántica	Características
R. Rada [96] $dist_{rad}(c_1, c_2) = len(c_1, c_2)$ **Ecuación 1**.	Considera la similitud semántica como la menor distancia semántica entre dos conceptos, $\boldsymbol{len}$ representa la longitud de la ruta más corta que los une en la jerarquía. La distancia se mide por el número de enlaces.

[30] Las relaciones semánticas más comunes son: meronimia / holonimia, hiponimia / hiperonimia y sinonimia / antonimia

P. Resnik [92] $sim_{res}(c_1, c_2) = CI(pcc(c_1, c_2))$ **Ecuación 2.**	Asocia un peso a cada nodo. Dicho peso representa el contenido de información (CI) del concepto, los conceptos más especializados en la jerarquía tienen mayores pesos que los más generales. La similitud está dada por el CI del pariente común más cercano (PCC)[31].
Jiang & Conrath [97] $dist_{JC}(c_1, c_2) = CI(c_1) + CI(c_2) - 2 * CI(pcc(c_1, c_2))$ **Ecuación 3.**	Este enfoque pretende realizar una combinación entre los métodos basados en enlaces y los basados en nodos. Se basa principalmente en la medida planteada por Rada, con el factor adicional de la incorporación de pesos en los nodos como factor de decisión.
G. Hirst & D. St-Onge [98] $rel_{HS}(c_1, c_2) = C - len(c_1, c_2) - k * d$ **Ecuación 4.**	Dado que esta propuesta tiene en cuenta varios tipos de relación en una Ontología como meronimia/holonimia, hiponimia/hiperonimia y sinonimia/ antonimia; se enfoca hacia una medida más general de relación semántica. También se involucran constantes como $\boldsymbol{C}$, $\boldsymbol{k}$ y $\boldsymbol{d}$ para ajustar la distancia de la ruta y los cambios de dirección de los enlaces.

Tabla 4. Medidas destacadas de Similitud Semántica.

Cabe destacar que la mayoría de los anteriores estudios se enfocan principalmente en Ontologías de tipo jerárquico como WordNet, ODP [99] o MeSH [100], a excepción de algunos trabajos [98, 101, 102] en los cuales se consideran algunas relaciones no jerárquicas, lo que implica una mayor complejidad en el cálculo de la similitud semántica. Tomando en cuenta este hecho, han sido muy pocos los trabajos que se han centrado en este tipo de relaciones Ontológicas, estos se basan en la asignación explícita de pesos a los conceptos o a las relaciones de una Ontología por parte de especialistas en un dominio específico [103]; por otra parte, también se han considerado valores experimentales para dichos pesos que han sido determinados por medio de diferentes pruebas, dichas pruebas sólo se han enfocado en WordNet, mas no en una Ontología OWL o RDF de un dominio más específico [101].

2.3 Estado actual de la Recuperación de Información y la Web Semántica.

Actualmente la RI se realiza basado en modelos definidos por algunos autores, pero el más utilizado es el modelo de espacio vectorial, el cual tiene como función principal clasificar los documentos generados por una consulta en orden descendente de acuerdo a la relevancia de cada recurso recuperado, en estos modelos los documentos y la consulta de usuario son representados como un conjunto de palabras clave necesarios

[31] En una jerarquía, representa a aquel concepto antecesor más cercano que es pariente de ambos conceptos.

para el proceso de clasificación de documentos, y a su vez estos son evaluados mediante mediciones tales como la precisión, proporción recuerdo entre otras, las cuales proporcionan una forma de medir la relevancia de los documentos respecto a la necesidad de información del usuario.

Con el fin de aumentar la precisión de las búsquedas Web, se han desarrollado nuevas técnicas dentro del área de la RI, representadas en modelos que pueden hacer uso de agentes inteligentes, ontologías, perfiles de usuario, expansión de consulta y técnicas de minería de datos entre otras.

Actualmente el perfil de usuario como técnica de personalización en la RI contempla la evaluación implícita o explícita que proporciona el usuario a los documentos recuperados desde un buscador Web; además, el perfil es representado por palabras claves, un conjunto de términos que representan un concepto de interés para el usuario, una colección de términos a corto y a largo plazo, por categorías de términos o una serie de consultas anteriores hechas por el usuario.

Por otra parte la Web Semántica ha cobrado una importancia significativa, ya que permite agregar un significado a los términos utilizados por los tradicionales sistemas de RI, brindando la posibilidad de generar mayor precisión en la información recuperada, este proceso se aprecia en la Figura 11, para que este proceso sea posible la Web Semántica se apoya en diferentes técnicas mencionadas, como ontologías para el refinamiento de consulta que solo tienen en cuenta el análisis semántico de cada término individualmente, pero la gran mayoría de estos trabajos no incluyen todos los términos de la consulta original vistos conjuntamente.

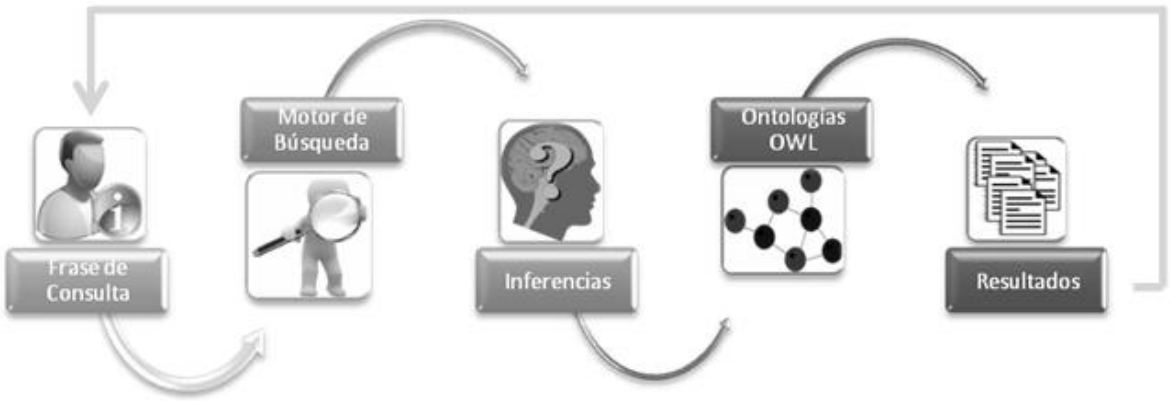

Figura 11. Proceso de la búsqueda semántica.

Capítulo III

3 CREACIÓN DEL MODELO

Usualmente las personas utilizan instintivamente modelos a la hora de tomar decisiones acerca de determinados aspectos de la realidad. En este orden de ideas, en el proceso de toma de decisiones se elige una entre varias acciones posibles, teniendo en cuenta el efecto que cada una de estas vaya a producir; por tanto la relación que liga las posibles acciones con sus efectos es el modelo del sistema [104].

En este capítulo es presentado un compendio de información, a partir del cual se propone el Modelo Semántico de Expansión de Consultas (MSEC), el cual está basado en el componente semántico que proporcionan la Ontologías de dominio y los recursos léxicos, además de la realimentación que ofrece el manejo de Perfiles de Usuario en las búsquedas Web, con el objetivo de mejorar la precisión de los resultados que son mostrados al usuario.

Para llevar a cabo la creación del modelo, es necesario distinguir tres fases [105]:

- **Fase de Conceptualización.** Determinar el ámbito del modelo.
- **Fase de Formulación.** Definir el modelo.
- **Fase de Evaluación.** Diseño de la prueba preliminar del modelo.

3.1 Fase de Conceptualización

El fundamento de toda actividad de diseño, manejo y monitoreo de proyectos es un modelo conceptual del proyecto. Este constituye un instrumento flexible para ser adaptado a la situación objeto de estudio.

Un modelo conceptual es la base para una buena planificación de proyectos, permite ver explícitamente la forma en que distintos factores están vinculados entre sí y por consiguiente la mejor forma de planificar y manejar el proyecto. Además muestra los posibles obstáculos o dificultades que pueden presentarse e ilustra la forma en que las intervenciones planificadas pueden afectar la condición de interés. Un buen modelo conceptual permite identificar los datos apropiados y necesarios que se requerirán para un monitoreo efectivo y eficaz del proyecto [105].

3.1.1 Modelo conceptual

La palabra Modelo se refiere a una representación simplificada de la realidad. La palabra conceptual se refiere a creencias teóricas. Un modelo conceptual es por lo tanto, una representación de las creencias teóricas en cuanto a un proyecto.

El modelo conceptual es la expresión en términos de conceptos y una serie de relaciones de ciertos factores que se cree impactan o conducen a la condición de interés que se desea describir y/o representar. A continuación se explica en forma breve aquellos conceptos y relaciones que interactúan en el modelo conceptual, los cuales han sido extraídos en su mayoría a partir del análisis documental realizado en el Capítulo 2.

- **Usuario.** Representa a los usuarios que manipulan el SRI.
- **Perfil de Usuario.** Está reflejado en los intereses de información del usuario.
- **Recursos.** Representan los documentos recuperados que se muestran al usuario. Estos recursos pueden ser representaciones electrónicas de texto.
- **Consulta.** Equivalente a la necesidad de información del usuario formulada por el mismo y que sirve de entrada para el SRI.
- **Consulta Expandida.** Representa la consulta original refinada semánticamente mediante un proceso de expansión.
- **Ontología.** Representa las relaciones entre conceptos en un dominio particular.
- **Término.** Representa las palabras que conforman la consulta original.

La Figura 12 presenta los conceptos y relaciones abstraídos del entorno de la personalización en la búsqueda Web, los cuales permiten establecer una hipótesis inicial enfocada hacia la obtención de resultados con mayor precisión para el usuario en el proceso de RI.

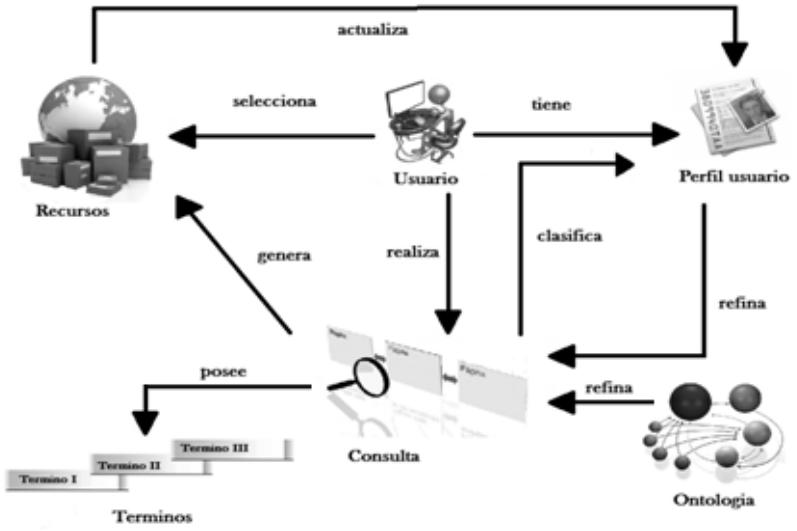

Figura 12. Modelo conceptual.

El anterior modelo conceptual describe un usuario el cual tiene una necesidad de información, esta es representada como un conjunto de términos que conforman una consulta, que al inicio es construida completamente por el usuario. Para obtener una expansión de esta consulta, se recurre a la ontología de dominio con el fin de realizar un refinamiento semántico a partir de los términos de la consulta original, además se puede hacer uso del Perfil de Usuario para refinar aún más dicha consulta. Una vez que se tiene

la consulta refinada ya sea por medio del Perfil o la Ontología, se generan los resultados a partir de los que ofrecen los buscadores Web y se presentan al usuario, el cual puede seleccionar aquellos que sean de su interés estableciendo a cada recurso seleccionado una evaluación explícita que le proporciona un porcentaje de relevancia.

Con la unión y/o combinación de cada uno de estos conceptos y sus relaciones, el presente modelo conceptual se formula como un instrumento fundamental para acompañar, direccionar y orientar la creación del modelo formal MSEC que se describe más adelante.

3.2 Fase de Formulación

En esta fase se representan los elementos manejados en la fase anterior por medio de un lenguaje formal. Para ello es necesario establecer diagramas formales al igual que las ecuaciones que se manejarán en el modelo [105].

A continuación se detalla el Modelo Semántico de Expansión de Consultas (MSEC). Para una mejor comprensión por parte del lector, primero se define la forma en que el presente proyecto concibe y maneja el Perfil de Usuario, puesto que este es un componente muy importante para el mismo.

3.2.1 Perfil de Usuario manejado en MSEC

La idea de involucrar el concepto de personalización en la búsqueda Web dentro de un meta-buscador, se debe a que las técnicas de personalización proporcionan resultados de interés para el usuario, teniendo en cuenta su perfil que actúa como filtro, reduciendo la información que no se considera relevante.

El PU en esta propuesta está representado como un conjunto de conceptos pertenecientes a la Ontología de dominio, los cuales son identificados a partir de una consulta de usuario y tienen asociado un peso que brinda una prioridad a cada uno de estos con la finalidad de utilizarlos en una futura expansión de consulta. Lo anterior se muestra en la Tabla 5, donde cada concepto C_i (i toma los valores desde 1 hasta el numero de conceptos identificados, n) posee un peso, el cual en este proyecto se define como el peso del concepto en los documentos relevantes para el usuario (Wru), y se obtiene con base al esquema de pesos abordado en el área de la RI, donde se define como la importancia de un término en un documento [3] y la calificación que proporciona el usuario a cada documento que considera relevante para su necesidad de información.

Concepto	C_i	,…,	C_n
Peso del concepto en los documentos relevantes para el usuario	Wru_i	,…,	Wru_n

Tabla 5. Representación del Perfil de Usuario.

3.2.2 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 mediante sus búsquedas. Este modelo consta de 5 módulos:

- Módulo de Consulta, que permite la interacción del usuario con el SRI a través de las consultas de búsqueda.

- Módulo de Expansión de Consulta, el cual realiza el procesamiento léxico y semántico de la consulta, ampliándola con nuevos conceptos extraídos a partir de Ontologías generales[32] y de dominio, además del uso del PU en tal proceso.

- Módulo de Recuperación de Documentos, que permite obtener los primeros documentos devueltos por el ranking de varios buscadores Web y presentarlos al usuario.

- Módulo de Evaluación de Documentos, el cual permitirle al usuario calificar aquellos documentos de su interés de acuerdo a su criterio de relevancia.

- Módulo de Gestión del Perfil de Usuario, que realimenta el PU con nueva información de las búsquedas del usuario y además realiza un ajuste de esta información almacenada, a los intereses actuales del usuario.

Cada uno de estos módulos se muestra en la Figura 13 y son descritos a continuación.

[32] El diccionario global WordNet también puede ser considerado como una Ontología general.

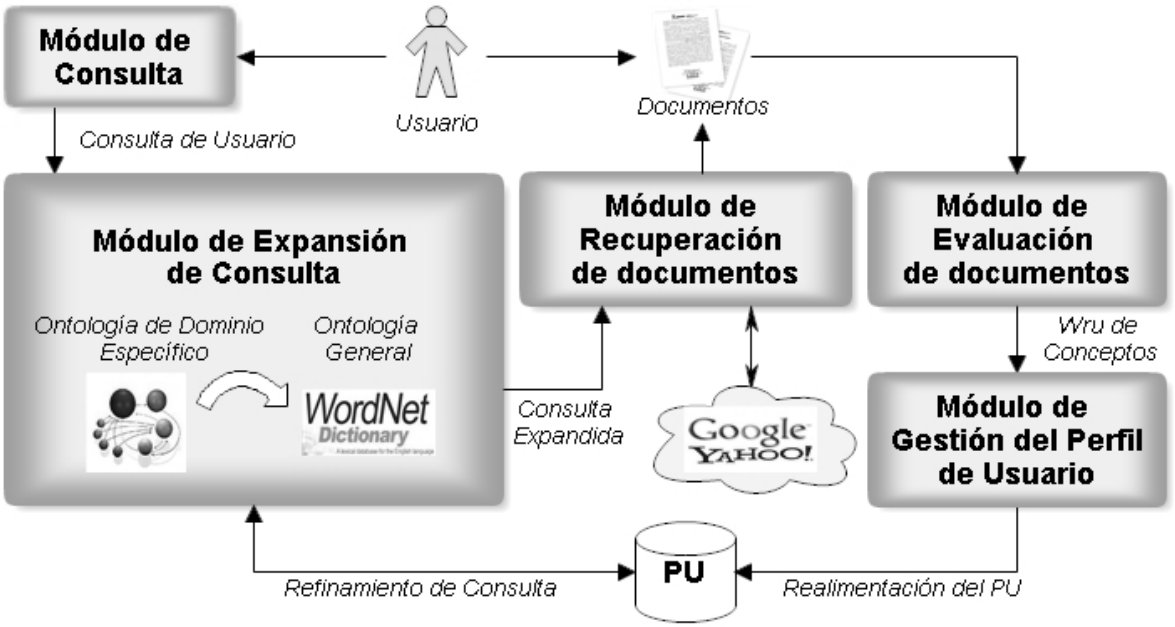

Figura 13. Modelo Semántico de Expansión de Consultas – MSEC.

3.2.2.1 Módulo de Consulta

Proporciona al usuario el servicio de consulta de documentos. En este módulo, el usuario accede al sistema mediante su login y su password, una vez adentro 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.

3.2.2.2 Módulo de expansión de consulta

Este módulo es uno de los que más aporta a esta propuesta, debido a que busca 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 representada como un conjunto de palabras clave, se muestra en la Figura 14, este se apoya en el uso de ontologías de dominio específico y general, las cuales brindan soporte para llevar a cabo procedimientos como la identificación de conceptos, la extracción de sinónimos, además del cálculo de la similitud semántica entre pares de conceptos.

Por otra parte, este proceso de expansión también se apoya en la información almacenada en el PU, siempre y cuando, esté relacionada con la necesidad de información del usuario, de tal forma que el PU se transforme en un medio para la extracción de conceptos apropiados para la expansión.

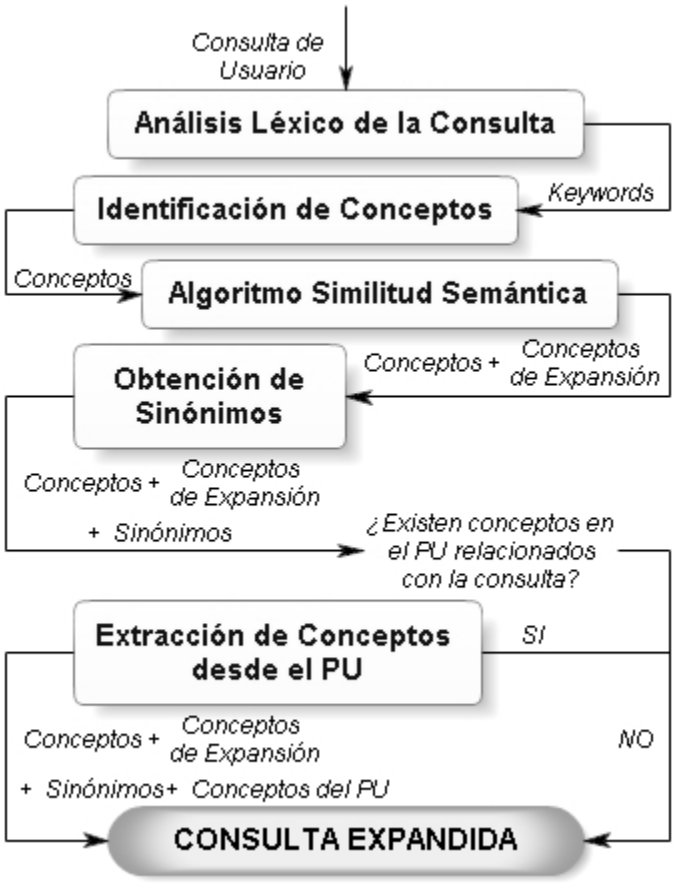

Figura 14. Módulo de Expansión de Consulta.

Cabe resaltar que la salida resultante de este modulo es una consulta de conceptos de expansión, en la cual dichos conceptos poseen un orden de prioridad de acuerdo a su tipo, como se presenta en la Tabla 6.

Acrónimo	Concepto	Descripción	Proceso en el que se utiliza	Prioridad
Fs	Frase Simple[33] [106]	Término que corresponde a una sola noción, pero que está constituido por más de una palabra la cual se encuentra en cualquier ontología manejada en el modelo. Por ejemplo "flor de lis" en el caso de utilizar una ontología de botánica	Identificación de conceptos	1, ya que es mas especifico, y por ende refleja la idea principal o parte de la necesidad de información del usuario.
Csi	Concepto Simple Individual	Concepto simple que hace parte de una determinada Frase simple y que además está ubicado en una distinta línea jerárquica de la Ontología que la de la Frase Simple a la que pertenece.	Identificación de conceptos	2, puesto que al estar ubicado en diferente línea jerárquica en la Ontología se puede interpretar con un sentido semántico diferente del que provee la frase simple a la que pertenece

[33] Esta denominación se propone en el trabajo realizado en [106] S. Liu, F. Liu, C. Yu, and W. Meng, "An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* NY, USA, 2004, pp. 266 - 272., en el cual se establecen cuatro tipos de frases identificadas a partir de una consulta de usuario.

Cs	Concepto Simple[107]	Término constituido por una sola palabra, la cual se encuentra en cualquier ontología manejada en el modelo. Por ejemplo “adrenalina” en el caso de utilizar una ontología de Medicina.	Identificación de conceptos	3, ya que es digitado por el usuario y se encuentra en la Ontología de dominio.
Cr	Concepto de Restricción	Concepto que está relacionado a otro concepto simple o compuesto en la Ontología mediante una regla de restricción o Axioma [57].	Identificación de conceptos	4, Debido a que se encuentra directamente relacionado con un Fs o Cs
t	Término	Son los términos que no se encuentran en las ontologías de dominio específico ni general utilizadas en este modelo.	Identificación de conceptos	5, ya que son directamente términos digitados por el usuario, y por defecto son importantes para el mismo
Csn	Concepto Sinónimo	Adjetivo que se utiliza para expresar que un vocablo tiene una misma o muy parecida significación que otro. Con base en esta definición un concepto de la consulta puede tener sinónimos pertenecientes a WordNet. Por ejemplo educar es sinónimo de enseñar.	Extracción de Sinónimos para expansión de consulta	6, puesto que es el sinónimo más comúnmente usado de un término.
Ces	Concepto de Expansión por Similitud Semántica	Concepto simple o compuesto que se extrae a partir de la mayor similitud semántica de un par de conceptos, considerando sólo los conceptos que se encuentren en la misma línea jerárquica o el concepto antecesor común a los dos conceptos en la Ontología.	Cálculo de similitud semántica para expansión de consulta	7, puesto que no son propiamente digitados por el usuario, pero están relacionados semánticamente con la necesidad de información

Tabla 6. Orden de prioridad de los conceptos de expansión.

La organización y la forma de utilizar los tipos de conceptos en los diferentes procesos identificados en la tabla anterior para la expansión de consulta, se describen en detalle más adelante en este módulo.

3.2.2.2.1 Análisis léxico de la consulta

En este paso se busca obtener las palabras clave de la consulta, por medio de la eliminación de palabras vacías, la eliminación de caracteres especiales y la conversión de los caracteres a minúsculas [3].

3.2.2.2.2 Identificación de Conceptos

Este procedimiento busca identificar los tipos de conceptos contenidos en la consulta específicamente términos, frases y conceptos simples, por medio de ontologías, generando cinco vectores con cada tipo de concepto identificado, los cuales son descritos en la Tabla 7.

Vector	Descripción
$\overrightarrow{V_t} = (t_i, \dots, t_n)$	Vector de términos el cual está compuesto por los términos t_i (donde ***i*** toma los valores de 1 hasta ***n***).
$\overrightarrow{V_{cs}} = (Cs_j, \dots, Cs_m)$	Vector de conceptos simples identificados en la ontología de dominio, conformado por los Cs_j (donde ***j*** toma los valores de 1 hasta el número de conceptos ***m***).
$\overrightarrow{V_{fs}} = (Fs_l, \dots, Fs_s)$	Vector de frases simples identificadas en la ontología de dominio, conformado por los Fs_l (donde ***l*** toma los valores de 1 hasta el número de conceptos ***s***).
$\overrightarrow{V_{ces}} = (Ces_x, \dots, Ces_t)$	Vector de conceptos de expansión por similitud semántica, conformado por los Ces_x (donde ***x*** toma los valores de 1 hasta el número de conceptos ***t***).
$\overrightarrow{V_{cr}} = (Cr_k, \dots, Cr_r)$	Vector de conceptos de restricción identificados en la ontología de dominio a partir de un Fs o un Cs, conformado por los Cr_k (donde ***k*** toma los valores de 1 hasta el número de conceptos ***r***).

Tabla 7. Vectores según el tipo de concepto identificado.

Para realizar la identificación de conceptos con el propósito de obtener los vectores descritos en la Tabla 7, se toma como base el estudio de Baziz [108], que consiste en utilizar una oración o frase que análogamente en el presente proyecto corresponde a la consulta digitada por el usuario, de tal manera que abarque todos sus términos de izquierda a derecha, los cuales son considerados como Frase Simple siempre y cuando, exista en una ontología, ya sea de dominio o general. El mismo proceso se repite con los mismos términos a excepción del último, hasta que la consulta quede con un solo término. Luego con los conceptos simples representados en $\overrightarrow{V_{cs}}$ y frases simples representadas en $\overrightarrow{V_{fs}}$ los cuales se identificaron en la Ontología de dominio, se descartan aquellos que estén contenidos en otros conceptos también identificados y que a la vez en la ontología sean hiperónimos, pues con base a los estudios de Navigli [109] y Bechara [110], los hiperónimos producen un efecto limitado debido al ruido que generan en la RI.

Para aclarar la manera de descartar conceptos simples y compuestos se realiza un ejemplo en la Figura 15; suponiendo que los conceptos identificados son *"Student"* y *"UndergraduateStudent"*, se descarta *"Student"* debido a que está contenido en *"UndergraduateStudent"* y además es Hiperónimo del mismo. Ahora, si los conceptos identificados son *"Student"* y *"GraduateStudent"* no se descartan, porque a pesar de que *"Student"* está contenido en *"GraduateStudent"*, no es Hiperónimo de este.

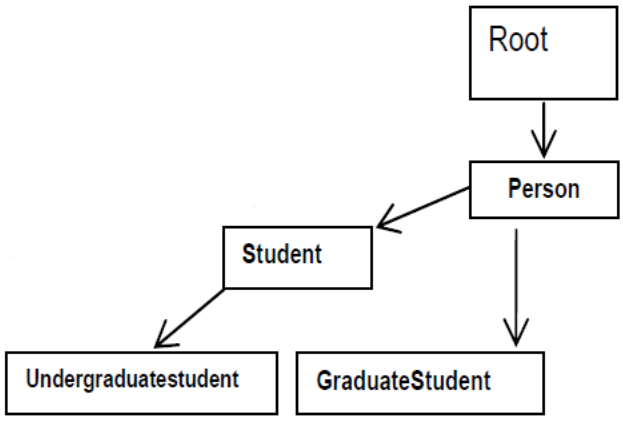

Figura 15. Vista parcial de la ontología Univ-Bench.
Fuente T. Slimani, B. B. Yaghlane, and K. Mellouli, "A New Similarity Measure based on Edge Counting," In Proceedings of world academy of science, engineering and technology, vol. 17, 2006.

También es preciso aclarar que para cada concepto perteneciente a $\overrightarrow{V_{cs}}$ y a $\overrightarrow{V_{fs}}$ se obtienen sus posibles conceptos relacionados a estos por medio de axiomas o reglas de restricción y son representados en $\overrightarrow{V_{cr}}$ y además aquellos términos que no se encuentren en la Ontología son representados en el $\overrightarrow{V_t}$ mostrado en la Tabla 7.

El proceso anterior se repite con los mismos términos de la consulta a excepción del primero hasta que finalmente la consulta quede con un solo término a identificar.

3.2.2.2.3 Cálculo de similitud semántica para expansión de consulta

La similitud semántica entre pares de conceptos que pertenecen a una ontología, puede ser usada con el objetivo de obtener conceptos adicionales que contribuyan a enriquecer la semántica de la consulta de usuario [111].

Tomando en cuenta lo anterior, para determinar la similitud semántica entre dos conceptos, este proyecto toma como base el estudio de Slimani [112], que se basa en el método de conteo de relaciones y consiste en una adaptación de la medida original de Wu y Palmer [113] con el propósito de obtener valores de similitud más precisos para pares de conceptos no localizados en una misma línea jerárquica. Esta propuesta utiliza esta medida, puesto que presenta algunas ventajas como su simplicidad, su rendimiento y su expresividad.

La anterior medida se aplica para proporcionar valores de similitud entre pares de conceptos de la consulta identificados a partir de la Ontología de dominio, considerando solamente aquellos pares con los mayores valores de similitud, los cuales podrían involucrar conceptos adicionales denominados conceptos extraídos por similitud

semántica (*Ces*) para expandir la consulta. La forma de obtener dichos conceptos adicionales depende de los siguientes casos:

- **El par de conceptos se encuentran en la misma línea jerárquica.** En este caso se obtienen los conceptos intermedios del par y son adicionados a la consulta, ya que se espera que estos Ces estén relacionados semánticamente con los dos conceptos anteriores.

- **El par de conceptos se encuentran en distinta línea jerárquica.** Se Obtiene el antecesor común al par el cual es adicionado a la consulta. Ya que el antecesor nos da el contexto común a los conceptos y podemos encontrar documentos relacionados a ese contexto. En este punto solo se extrae un concepto antecesor, ya que los estudios [109, 110], mencionan que al consultar más antecesores, se crea una divergencia en los conceptos, desviando el contexto de la consulta.

- **Dos o más pares de conceptos tienen el mismo valor de similitud semántica y cada par pertenece a distinta línea jerárquica.** Se extraen conceptos intermedios de todos los pares para agregarlos a la consulta, teniendo en cuenta las razones del primer caso.

- **Dos o más pares de conceptos tienen el mismo valor de similitud semántica y cada concepto del par pertenece a distinta línea jerárquica.** En este caso se extrae el antecesor común de cada par con el objetivo de agregarlos a la consulta, teniendo en cuenta las razones del segundo caso.

En cualquiera de los casos anteriores, los *Ces* se representan en el vector $\overrightarrow{V_{ces}}$ como se presentó en la Tabla 7.

A continuación se propone un ejemplo para entender mejor la inclusión del concepto de la similitud semántica para expandir la consulta de usuario.

En la Figura 16 se observa un fragmento de una Ontología de dominio. Por ejemplo, si el par de conceptos con mayor similitud semántica es *"Automóvil"* y *"Barco"*, el concepto de expansión a tener en cuenta es su antecesor común, *"Medio de Transporte"*. Por otro lado, si el mayor valor de similitud semántica fue obtenido a partir del par *"Automóvil"* y *"Medio de Transporte"*, el concepto intermedio *"Terrestre"* es utilizado en la expansión de consulta.

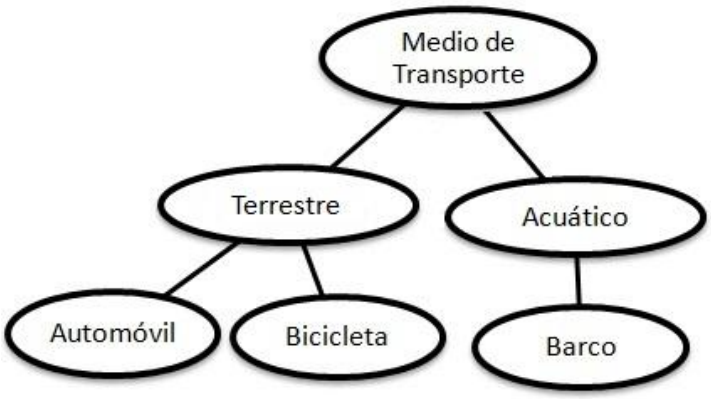

Figura 16. Fragmento de una Ontología de Dominio Específico.

3.2.2.2.4 Extracción de Sinónimos para expansión de consulta

Este procedimiento considera un aspecto, el cual consiste en extraer desde la ontología general WordNet, el sinónimo de uso más común por cada término perteneciente al vector $\overrightarrow{V_t}$ descrito en la Tabla 7.

3.2.2.2.5 Extracción de conceptos del PU para expansión de consulta

El perfil de usuario es utilizado como una técnica en la personalización, además de la propia consulta, para estimar los intereses del usuario [26], 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 el módulo de gestión del Perfil de Usuario.

Este procedimiento recurre al PU con la finalidad de encontrar los conceptos obtenidos a través de la similitud semántica entre dos conceptos (sección 3.2.2.2.3), a partir de estos conceptos se seleccionan sólo aquellos con mayor valor de Wru para ser adicionados a la consulta expandida.

3.2.2.3 Módulo de Recuperación de Documentos

En este modulo se busca recuperar los documentos relacionados a la consulta expandida enviándola como entrada a dos de los Buscadores Web más populares como Google y Yahoo!, estos motores de búsqueda retornan un conjunto de resultados, de los cuales se extraen los 10 primeros de cada buscador [114], dichos documentos son ordenados y posteriormente presentados al usuario en una lista de links con su respectivo título, url y resumen.

Este proyecto plantea un proceso para llevar a cabo lo descrito anteriormente, el cual consiste en una serie de pasos que están organizados en las siguientes secciones.

3.2.2.3.1 Formato de la consulta expandida

El uso de operadores lógicos de consulta cobra gran importancia cuando se utilizan apropiadamente en un SRI. Por otra parte diversos estudios indican que un 90% de los usuarios que utilizan los buscadores Web, utilizan consultas extremadamente simples, mientras que únicamente el 10% utiliza operadores de consulta avanzados [114]. Teniendo en cuenta lo anterior, la consulta expandida puede beneficiarse del uso de operadores lógicos de consulta; para tal fin se establece un formato adecuado que permita organizar cada concepto de forma que al ser enviada a los buscadores Web, estos permitan optimizar la búsqueda al utilizar herramientas que los mismos proveen como lo son los operadores de consulta.

En esta sección se establece un formato para los conceptos de expansión pertenecientes a los vectores que se presentaron en la Tabla 7, para tal fin se considera evaluar precisión de los resultados de búsqueda que se obtienen mediante el uso del operador lógico *OR* y su combinación con *AND*, en este último caso, la consulta expandida se agrupa de la siguiente manera.

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

El anterior formato de consulta expandida está organizado de acuerdo a la prioridad establecida en la Tabla 6. Se puede observar que el operador lógico *AND* sólo está ubicado entre la frase o concepto simple y sus respectivos conceptos de restricción debido a que las frases o conceptos simples están relacionados con sus conceptos de restricción, de esta manera se pueden recuperar documentos que contengan cada par de conceptos involucrados por relación como se muestra a continuación.

$$\left(Fs_1\ AND\ (Cr_{11} \dots OR\ Cr_{1t})\right) \equiv (Fs_1\ AND\ Cr_{11} \dots\ OR\ Cc_1\ AND\ Cr_{1t})$$

Para los restantes conceptos se ha establecido el operador lógico *OR*, porque se pueden recuperar documentos que también sean relevantes aún cuando no respondan rigurosamente a los términos utilizados por el usuario. Este formato de expansión es automático y permite aumentar la cantidad de documentos a recuperar [2].

Adicionalmente se propone hacer uso solamente del operador *OR* para todos los conceptos de la consulta expandida por la razón mencionada en el párrafo anterior. Este segundo formato se propone con el objetivo de evaluar la precisión de sus resultados frente al formato de la combinación *OR-AND* y establecer cuál de los dos utilizar con base al de mayor precisión para su aplicación en el modelo propuesto[34].

3.2.2.3.2 Recuperación de documentos

Con el formato aplicado a la consulta expandida, se procede a enviarla al meta-buscador Web el cual se soporta en varios buscadores para la RI [115] proporcionando ventajas como la mejora del factor de recall y la precisión, entre otras [116]. Por otra parte con base a los estudios [2, 109], en los cuales se analizaron distintos buscadores encontrando que Google tiene la limitación de 10 palabras por consulta, y una estrategia de expansión de consulta compleja puede llegar a tener muchas más, en contraste se analizó Yahoo! y se observó que no tiene esta limitación. Por lo anterior se podría descartar la utilización de Google en este proyecto, pero no es suficiente argumento ya que este motor de búsqueda es uno de los más utilizados en la actualidad[35] haciendo que sus resultados sean de gran importancia para el prototipo que se describe en el siguiente capítulo.

Por otra parte el estudio realizado por Wang [117] indica que la cantidad de palabras de expansión influye en la precisión de los resultados de búsqueda, que mediante un experimento se logro comprobar que adicionar máximo 9 palabras de expansión a la consulta inicial de usuario contribuye a mejorar la precisión, ahora bien, tomando como base este enfoque, el presente proyecto considera limitar la longitud de las consulta de expansión dejando solo máximo 9 conceptos de expansión adicionales a la consulta original de usuario. Una vez realizado lo anterior finalmente los documentos recuperados se presentan al usuario.

3.2.2.4 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, de igual forma, el usuario puede calificar cada documento con el fin de reflejar su nivel de interés sobre dicho recurso teniendo en cuenta su propio criterio de relevancia. A continuación se describen los principales componentes de este módulo.

[34] La evaluación de estos dos formatos de consulta expandida se describe en el Capítulo 5 referente a la validación del prototipo.

[35] De acuerdo a los reportes realizado por OneStat.com, líder mundial en analítica Web en tiempo real. www.onestat.com

3.2.2.4.1 Evaluación de documentos

Luego de que los resultados sean presentados al usuario, este 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, pues ha sido ampliamente demostrado que la realimentación por relevancia explicita mejora la precisión y el recall en la RI [118], por lo anterior, este proyecto opta por este tipo de realimentación, la cual consiste en un rango de números discretos de cero a cinco, donde el usuario puede seleccionar un único valor, y reflejar su nivel de interés sobre el documento de forma más precisa.

El proceso anterior se repite hasta que el usuario finalice su sesión. Al finalizar la sesión se obtienen los documentos seleccionados y su respectiva calificación en términos de porcentaje como se muestra en la Tabla 8.

Documento seleccionado	$Doc_i, \ldots, Doc_n$	Doc_i representa un determinado documento seleccionado, (donde ***i*** toma los valores de 1 hasta el número de documentos seleccionados ***n***)
Porcentaje de calificación	$Cal_i, \ldots, Cal_n$	Cal_i representa la calificación de un determinado documento ***i*** seleccionado, (donde ***i*** toma los valores de 1 hasta el número de documentos seleccionados ***n***)

Tabla 8. Representación de los documentos calificados.

3.2.2.5 Módulo de Gestión del Perfil de Usuario

En este módulo se busca crear el PU por medio de los conceptos que conforman una determinada consulta expandida, asociándole a cada uno 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. Los conceptos de la consulta expandida se almacenan en el PU con su respectivo peso. Finalmente la actualización de la información almacenada en el PU pretende establecer los intereses actuales del usuario, para ello este módulo se basa en los conceptos extraídos desde el PU utilizados en la expansión, con el fin de ajustar sus valores Wru a tales intereses.

3.2.2.5.1 Cálculo del peso del concepto en los documentos relevantes para el usuario

La importancia o el peso de un término clave *i* para un documento *j* se basa en la tradicional fórmula TF-IDF en el modelo clásico de Espacio Vectorial [3] (Ecuación 5).

$$Wd_{ji} = tf_{ji} * idf_i$$

Ecuación 5.

Donde:

$\boldsymbol{tf_{ji}}$ Cuantifica el grado de similitud de un término $\boldsymbol{i}$ en un documento j a través de su frecuencia de aparición.

$\boldsymbol{idf_i}$ Cuantifica el grado de disimilitud de un término $\boldsymbol{i}$ en el conjunto de documentos recuperados, este grado se calcula a través de su frecuencia inversa la cual equivale al número de documentos sobre la frecuencia del término en el total de documentos.

Tomando como base la Ecuación 5, este proyecto toma el peso de un determinado concepto el cual es calculado para cada documento seleccionado y lo relaciona con la respectiva calificación del documento, de tal manera que el peso se ajuste a la calificación que proporciono el usuario al documento, con la finalidad de determinar el peso del concepto en los documentos relevantes para el usuario (Wru).

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

Ecuación 6.

Donde:

$\boldsymbol{Cal_j}$ Representa la calificación de un documento j seleccionado por el usuario, (donde j toma los valores de 1 hasta el número de documentos seleccionados n).

$\boldsymbol{Wd_{ji}}$ Representa el peso de un concepto i en un documento j seleccionado por el usuario, (donde i es un concepto de $\overrightarrow{V_c}$).

La Ecuación 6 se aplica individualmente a los conceptos de la consulta expandida, tal como se observa en la Tabla 9:

Documento	**Calificación Documento**	**Conceptos de la Consulta Expandida**			**Cálculo del Wru de los conceptos de $\overrightarrow{V_c}$**		
		C_i	, …. ,	C_n	C_i	, …. ,	C_n
Documento j	Cal_j	Wd_{ji}	, …. ,	Wd_{jn}	$Wd_{ji} * Cal_j$	, …. ,	$Wd_{jn} * Cal_j$
⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮
Documento n	Cal_n	Wd_{ni}	, …. ,	Wd_{nn}	$Wd_{ni} * Cal_n$	, …. ,	$Wd_{nn} * Cal_n$
Peso del concepto en los documentos relevantes para el usuario					Wru_i	, …. ,	Wru_n

Tabla 9. Peso del concepto en los documentos relevantes para el usuario.

3.2.2.5.2 Ajuste del Wru de conceptos de expansión extraídos del PU

Los conceptos extraídos desde el PU utilizados en la expansión, poseen un Wru que se relaciona con un nuevo Wru, el cual fue recalculado como se observó en la Tabla 9, con el objetivo de ajustar el primero de acuerdo al valor del segundo mediante una relación directamente proporcional, es decir que si el valor del segundo se incrementa, el primero también, o si el valor del segundo disminuye, también lo hace el primero.

Para efectuar el ajuste mencionado anteriormente esta propuesta plantea la siguiente relación (Ecuación 7) denominada ajuste del Wru de conceptos de expansión extraídos del PU ($Wrua$):

$$Wrua = \alpha\, Wrup + \beta\, Wrur$$

Ecuación 7.

Donde:

- $\boldsymbol{Wrup}$ Representa el peso histórico del Wru perteneciente al concepto de expansión almacenado en el PU.

- $\boldsymbol{Wrur}$ Representa el peso del nuevo Wru perteneciente al respectivo concepto de expansión de PU utilizado en la actual consulta expandida, el cual fue calculado como se observó en la Tabla 9.

Los valores α y β son constantes de ajuste, las cuales han sido establecidas con α=0.25 y β=0.75 a partir de las pruebas experimentales que se describen más adelante en el capítulo 5. 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.

Finalmente el valor $Wrup$ de cada concepto del $\overrightarrow{V_{cp}}$ se reemplaza por el $Wrua$ con el objetivo de reflejar el interés actual del usuario hacia dichos conceptos, es decir, los pesos de los conceptos tienden hacia ítems muy buscados, de modo que si el valor de $Wrup$ aumenta respecto al anterior, significa que aumenta la probabilidad de ser seleccionado nuevamente en una posterior expansión, de lo contrario disminuye dicha probabilidad.

3.3 Fase de Evaluación

La fase de evaluación permite realizar un análisis del modelo así como su sometimiento a criterios de aceptabilidad como ensayos mediante simulación de las hipótesis sobre las que se asienta el modelo y su consistencia, además del análisis de sensibilidad para estudiar la dependencia de las conclusiones extraídas del modelo con las variaciones de los parámetros que aparecen en el mismo [105]. Teniendo en cuenta que a partir del modelo se desarrolla un prototipo software, la fase de evaluación del modelo está ligada a

la implementación y posterior validación del prototipo (Capítulo 4 y Capítulo 5). Vale la pena aclarar que el modelo MSEC presentado en la fase de formulación, es producto de una serie de iteraciones en su construcción, las cuales surgieron como parte de la evaluación del prototipo, el cual permitió adicionar mejoras a MSEC y a la vez eliminar ciertos elementos que hacían parte de éste en un principio pero que no ofrecieron el comportamiento adecuado en pro del desempeño del sistema.

Capítulo IV

4 IMPLEMENTACIÓN DEL PROTOTIPO

Para el desarrollo del prototipo se emplea la metodología UP Ágil (Agile Unified Process), utilizando sus fases y sólo los artefactos que se consideren necesarios. Las fases son las siguientes:

4.1 Fase de Inicio

En esta fase se definen los casos de uso más relevantes a un alto nivel con su respectivo diagrama de casos de uso, esto con el fin de establecer los requisitos mínimos para desarrollar el prototipo de acuerdo al modelo establecido. Posteriormente se identifica y se define a nivel general una arquitectura potencial que permite guiar el proceso de construcción. En esta fase también, se realiza un estudio detallado de las ontologías existentes en un dominio en particular, con el fin de establecer cuál de ellas es susceptible para ser utilizada en la creación del prototipo (ver Anexo C).

La herramienta software principal para llevar a cabo la implementación del proyecto es Visual Studio 2008 con Framework 3.5 de Microsoft .NET en el lenguaje C# y el motor de bases de datos SQL Server Express Edition 2005. Se desarrolla una aplicación web que sirve como medio de interacción con el usuario para realizar las búsquedas respectivas.

Además de la anterior herramienta, se utiliza Protégé en sus versiones 3.4.3 y 4.1 Beta. Protégé permite el manejo de la Ontología en un lenguaje OWL. En el presente proyecto se utiliza la API de Jena para la plataforma .NET (LinkedDataTools.JenaDotNet), que proporciona métodos necesarios para el manejo de la ontología tales como la obtención de clases, super-clases, sub-clases y axiomas de la ontología. También se utiliza el API de Lucene para .NET con el fin de realizar el análisis léxico de la consulta original de usuario, específicamente la eliminación de palabras vacías y la posterior extracción de palabras clave.

Finalmente se emplea el diccionario léxico WordNet a través de las API's IKVM.OpenJDK.ClassLibrary, jaws-bin para la obtención de sinónimos.

Se realizó un estudio de las ontologías más completas existentes y disponibles en la web. Se escogió la ontología *"nciOncology.owl"* del Instituto Nacional del Cáncer de los Estados Unidos [119] que almacena información principalmente relacionada con el dominio de la medicina, específicamente sobre el Cáncer. La información sobre el estudio de la misma se encuentra en el Anexo C.

4.1.1 Análisis de requerimientos

Se desarrollará una aplicación Web que permita realizar búsquedas de acuerdo a las necesidades de información de los usuarios que utilicen la aplicación.

- Se pide que los resultados de la búsqueda muestren una precisión aceptable (de acuerdo a lo expuesto en [10]) de las páginas que serán presentadas al usuario.
- La interfaz que se presente al usuario debe ser un entorno amigable de fácil interacción y de fácil acceso al mismo.
- Utilización de ontologías para la expansión de consulta de la cual hace uso el meta-buscador Web.
- Utilización de los servicios de WordNet para la expansión de consulta, específicamente la adición de sinónimos.
- El idioma que se maneja para la ontología es el inglés, ya que en las referencias investigadas no se encontró una ontología lo suficientemente robusta y completa que manejara el idioma español.

4.1.2 Diagrama de Casos de Uso

Un Diagrama de Casos de Uso muestra la relación entre los actores y los casos de uso del sistema. Representa la funcionalidad que ofrece el sistema en lo que se refiere a su interacción externa. Después de la definición de los requerimientos del prototipo software se realizaron los diagramas de casos de uso para describir la interacción entre los usuarios (actores). En la Figura 17 se muestra el diagrama principal.

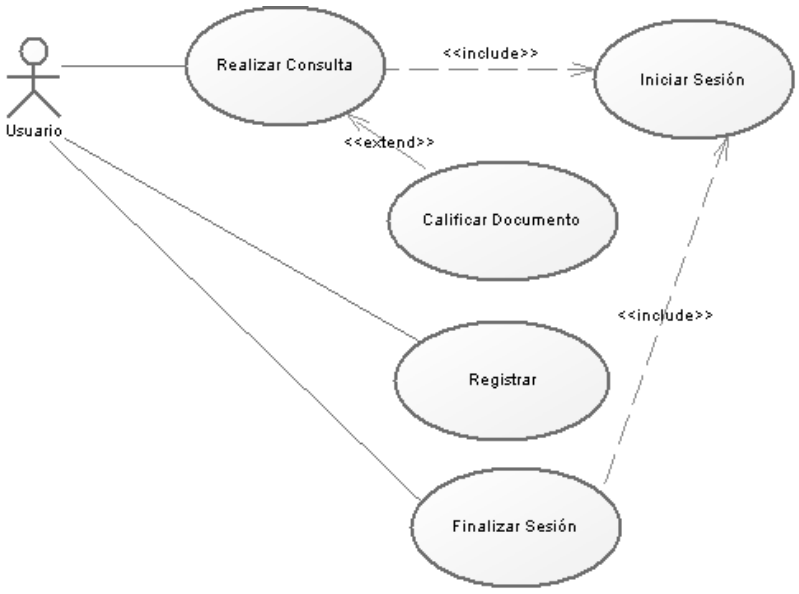

Figura 17. Diagrama de Casos de Uso

4.1.3 Casos de Uso en Formato Compacto

En ingeniería del software, un caso de uso es una técnica para la captura de requisitos potenciales de un nuevo sistema o una actualización de software. Cada caso de uso proporciona uno o más escenarios que indican cómo debería interactuar el sistema con el usuario o con otro sistema para conseguir un objetivo específico [120]. En la Tabla 10 se presentan los casos de uso en formato compacto.

CASO DE USO	Iniciar Sesión
ACTOR	Usuario
DESCRIPCIÓN	El usuario desea iniciar sesión para conectarse al sistema.

CASO DE USO	Registrar
ACTOR	Usuario
DESCRIPCIÓN	El usuario desea registrarse en el sistema.

CASO DE USO	Realizar Consulta
ACTOR	Usuario
DESCRIPCIÓN	El usuario podrá realizar una consulta en la web, digitando la frase o frases que describan su pregunta. Con esto se pretende obtener una respuesta relevante a sus necesidades de información.

CASO DE USO	Calificar Documento
ACTOR	Usuario
DESCRIPCIÓN	El usuario desea ver el contenido del documento que ha seleccionado y posteriormente calificarlo.

CASO DE USO	Finalizar Sesión
ACTOR	Usuario
DESCRIPCIÓN	El usuario desea finalizar su sesión para desconectarse del sistema.

Tabla 10. Casos de Uso en formato compacto.

4.2 Fase de elaboración

En esta fase se busca definir la arquitectura del sistema y realizar el modelado de la misma.

4.2.1 Arquitectura de la aplicación

Para el desarrollo del prototipo de meta-buscador MSEC Web Search, se definió una arquitectura tres capas, pues permite llevar a cabo el desarrollo en varios niveles, lo cual brinda ventajas en la construcción de la aplicación para la descomposición de actividades, flexibilidad y escalabilidad de la misma. En la Figura 18 se muestra el diagrama general de la arquitectura tratada.

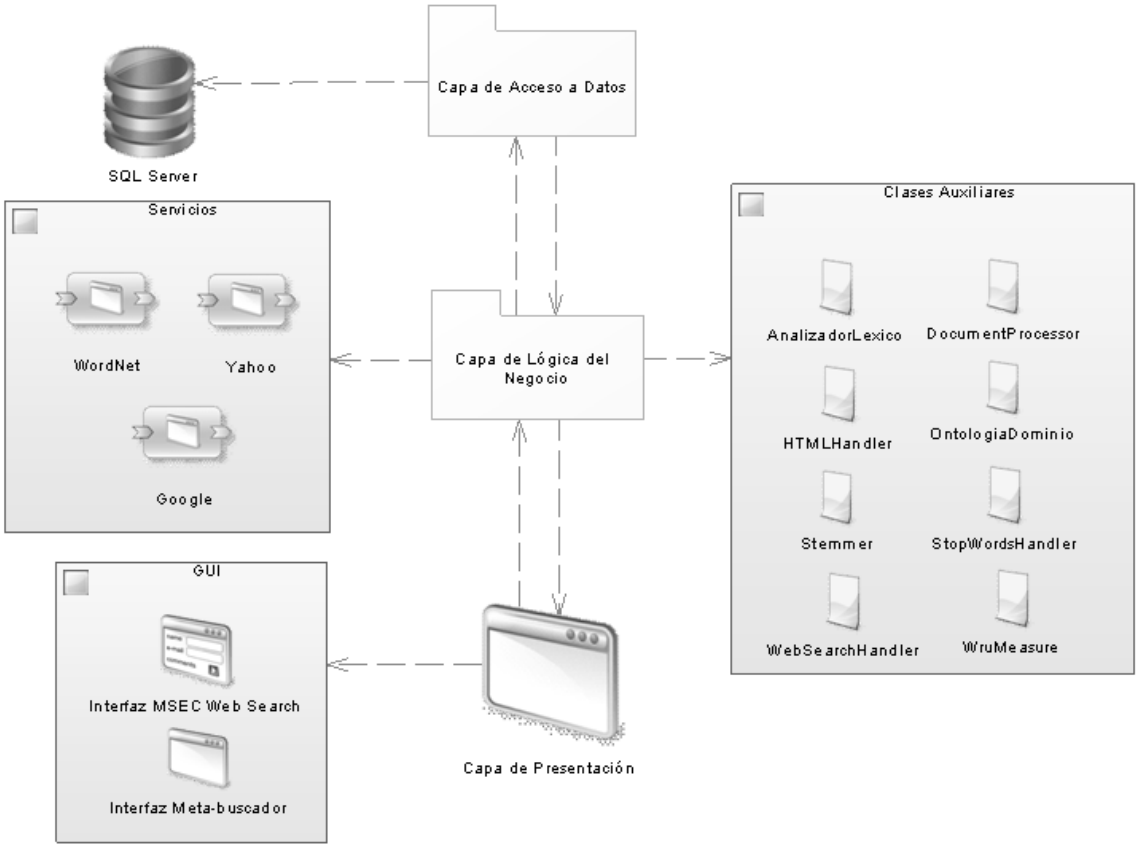

Figura 18. Arquitectura tres capas del prototipo MSEC Web Search.

Para la creación del prototipo se ha seleccionado una herramienta de desarrollo Microsoft .NET junto con el uso del Framework de desarrollo 3.5 de .Net, el cual le brinda al desarrollador la facilidad de crear aplicaciones para cualquier tipo de dispositivo. Además ofrece una independencia de lenguaje a los programadores ofreciendo así una plataforma más madura, documentada, más amplia en su ámbito de actuación y tiene un diseño consistente. Brinda la posibilidad de integrar componentes escritos en numerosos lenguajes de programación y desarrollar características de productividad tales como soporte para emular dispositivos inteligentes, diseño visual de formularios, control de la interfaz del usuario, soporte remoto para la detección de errores y despliegue simplificado de las aplicaciones. Para el proyecto se ha elegido el lenguaje de programación C#.

La importancia de la arquitectura tres capas, es que permite la separación de la capa de presentación de la capa de negocio y la capa de datos; de esta manera se facilita el mantenimiento del software en caso de cambios y además permite distribuir el trabajo de creación de una aplicación por niveles, de este modo, cada grupo de trabajo estará completamente abstraído del resto de niveles enfocándose en una tarea específica de su respectiva capa. A continuación se describe cada una.

- **Capa de presentación.** Es la capa que ve el usuario, le presenta el sistema y le comunica información, además, captura la información del usuario en un mínimo proceso. Esta etapa se comunica únicamente con la capa de negocio. Esta capa también es conocida como interfaz grafica y debe tener la característica de ser amigable para el usuario.

  En la aplicación se construye una interfaz (MSEC Web Search) que proporciona interacción con los usuarios para realizar las pruebas correspondientes y evaluación del prototipo; las interfaces principales corresponden al inicio de sesión (Figura 19), la página de búsqueda donde se despliegan los resultados (Figura 20) y la interfaz en la que se muestra el documento al usuario para su calificación (Figura 21).

Figura 19. Interfaz de inicio de sesión.

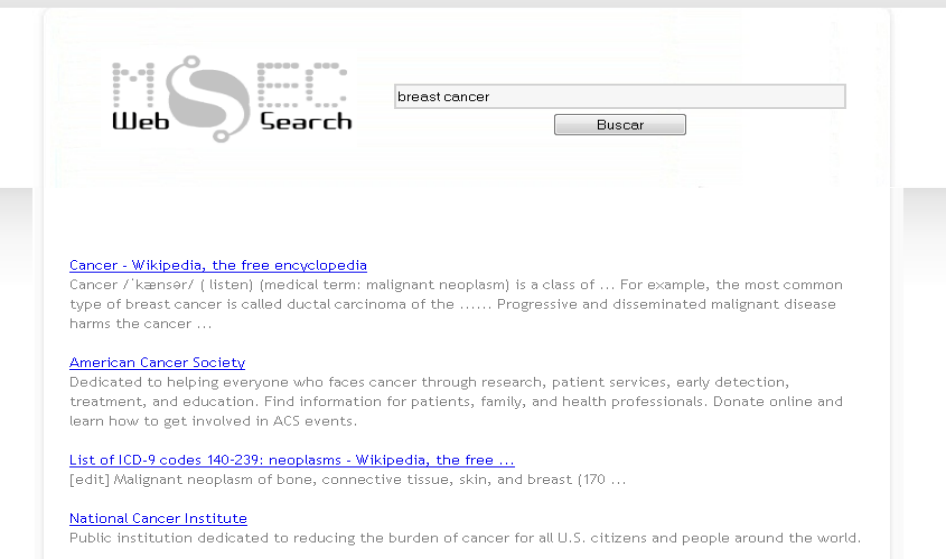

Figura 20. Interfaz del meta-buscador MSEC Web Search.

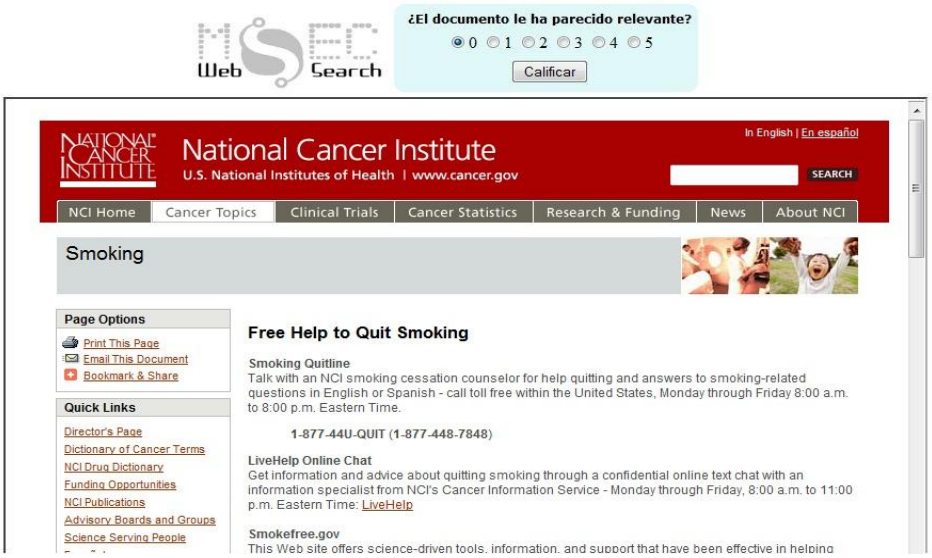

Figura 21. Vista de un documento seleccionado por el usuario y su calificación.

- **Capa de lógica del negocio.** En esta capa se reciben las peticiones del usuario y se le envían las respuestas tras el proceso. Se denomina capa de lógica del negocio porque aquí es donde se establecen todas las reglas que deben cumplirse. Esta capa se comunica con la capa de presentación para recibir las solicitudes y presentar los resultados, además también se comunica con la capa de datos para solicitar al gestor de la base de datos, almacenar o recuperar datos del mismo[36].

  Los servicios se muestran como una sub-capa, los cuales pueden ser de datos propios del sistema o datos expuestos por sistemas externos (Servicios Web externos, etc.) [121], como en el caso de esta aplicación. Los servicios brindan la información necesaria para la expansión de la consulta en un dominio general o específico, accediendo por medio de las API's a sus servicios. Los servicios Web que intervienen en la búsqueda como por ejemplo WordNet y los servicios de búsqueda de dos de los más populares buscadores Web como Google y Yahoo. La capa de lógica del negocio se considera el corazón de la aplicación ya que esta es la que se comunica con todas las demás capas para llevar a cabo las tareas [122].

  En el prototipo de este proyecto, se incluyen las clases y métodos que interactúan con la capa de lógica del negocio y con la interfaz de presentación al usuario. Las clases principales son: *AnalizadorLexico*, *DocumentProcessor*, *HTMLHandler*, *OntologiaDominio*, *Stemmer*, *StopWordsHandler*, *WebSearchHandler*, *WruMeasure*. Por medio de estas clases se maneja la información obtenida desde la interfaz de búsqueda del usuario y la información obtenida al realizar la búsqueda. La capa de lógica del negocio se muestra en la Figura 22.

[36] Extraído de http://ceisuss.wordpress.com/2008/06/23/programacion-por-capas/

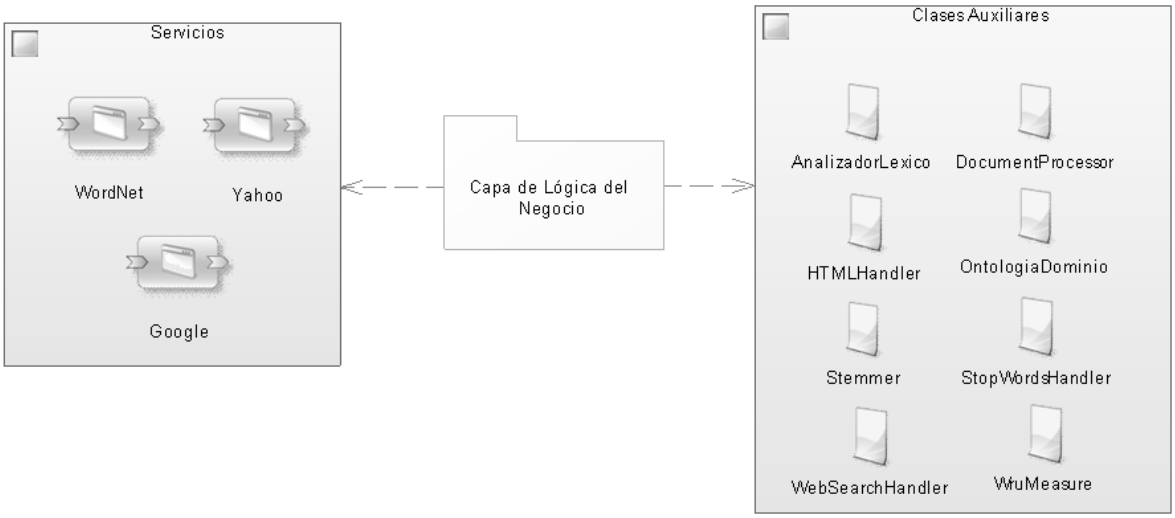

Figura 22. Capa de lógica del negocio.

- **Capa de acceso a datos.** En la lógica de acceso a datos se tienen las clases que permiten la recuperación de datos relacionados con el PU el cual está almacenado en una base de datos de SQL Server Express Edition 2005 como se muestra en la Figura 23.

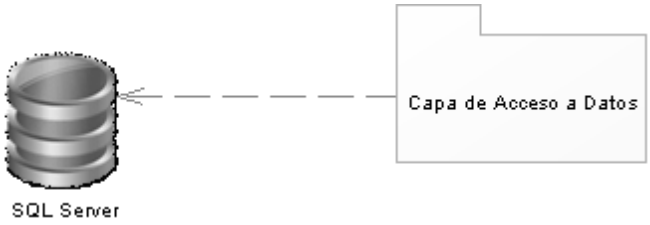

Figura 23. Capa de acceso a datos.

La especificación de la arquitectura junto con los patrones de software usados se describe en detalle en el Anexo D.

4.3 Fase de construcción

En esta fase se definen en forma detallada los casos de uso, se realizan los diseños de los diagramas de clase, de secuencia, de despliegue y el modelo de base de datos los cuales, permiten mostrar al usuario el funcionamiento del prototipo a ser desarrollado. Además en este proceso se implementa el prototipo software que posteriormente será integrado al meta-buscador GruWeb[37].

[37] En la actualidad el grupo de investigación GTI cuenta con una versión mejorada de GruWeb denominada Minerva, en este caso la integración se realizará con esta nueva versión. La aplicación Minerva está disponible online en http://spar.unicauca.edu.co/minerva.

4.3.1 Casos de Uso en Formato Extendido

Los casos de uso en formato extendido son de gran importancia puesto que es un formato que es más detallado que el formato de alto nivel, además tiene una sección del curso normal de los eventos. Este formato se usa durante la especificación de requisitos para los más importantes o de mayor influencia. Dado que estos casos de uso son mucho más extensos que los casos de uso de alto nivel, se describen con mayor detalle en el Anexo D.

4.3.2 Diagramas de Secuencia

Los diagramas de secuencia son utilizados para modelar la interacción entre objetos en una aplicación a través del tiempo y se modelan para cada caso de uso. Estos diagramas se muestran en la Figura 24, Figura 25, Figura 26, Figura 27 y Figura 28.

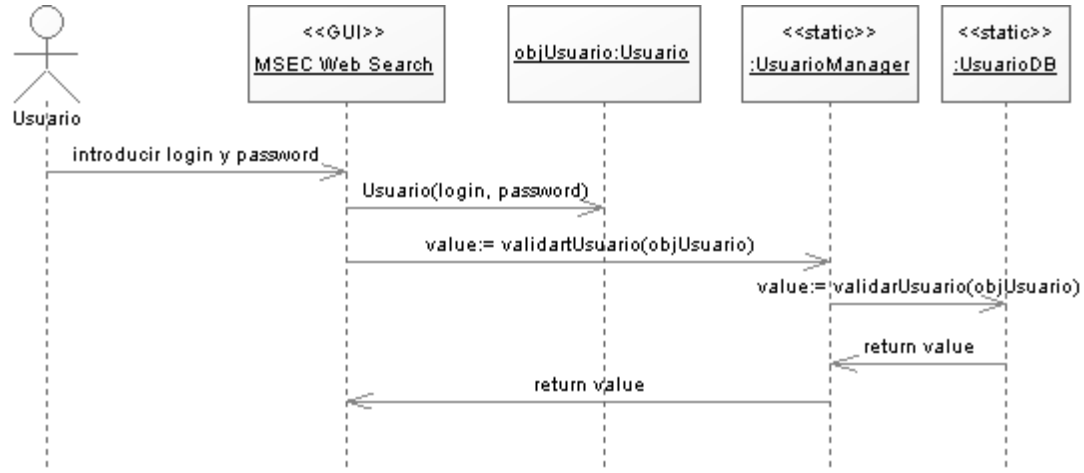

Figura 24. Diagrama de Secuencia Iniciar Sesión.

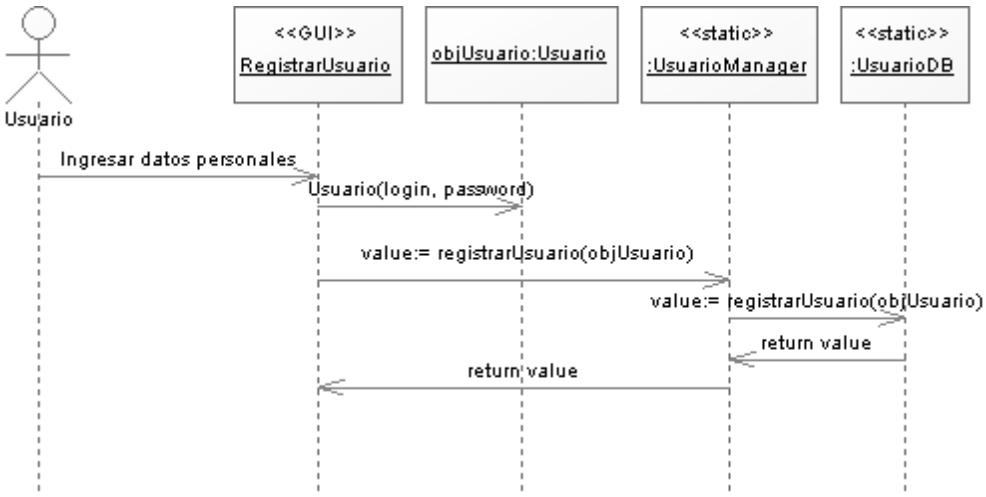

Figura 25. Diagrama de Secuencia Registrar Usuario.

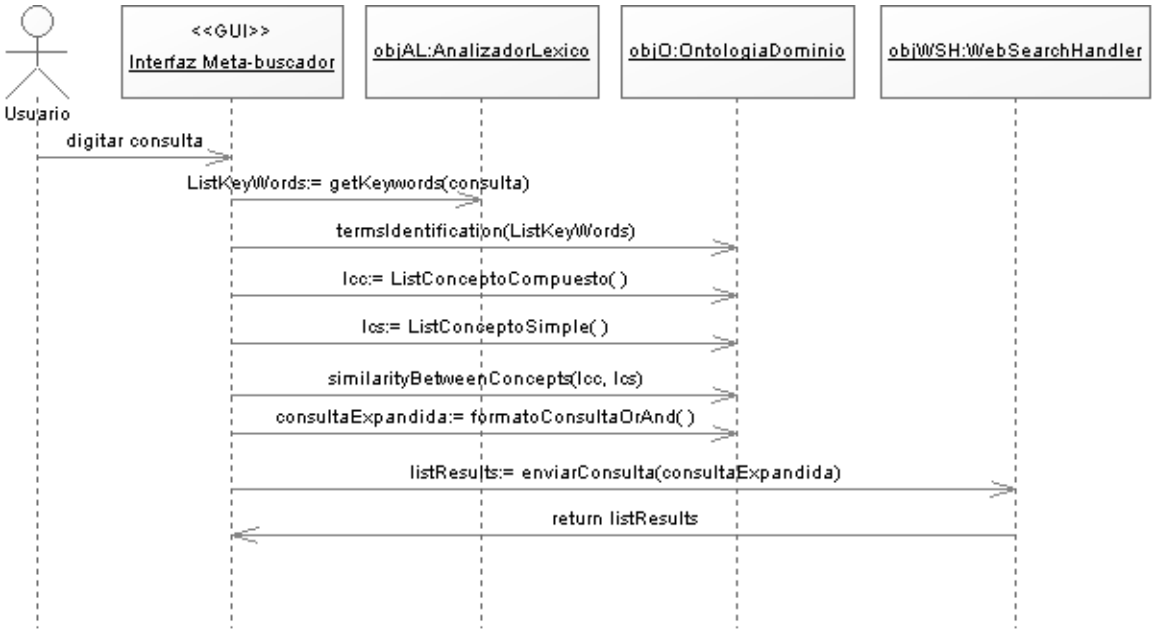

Figura 26. Diagrama de Secuencia Realizar Consulta.

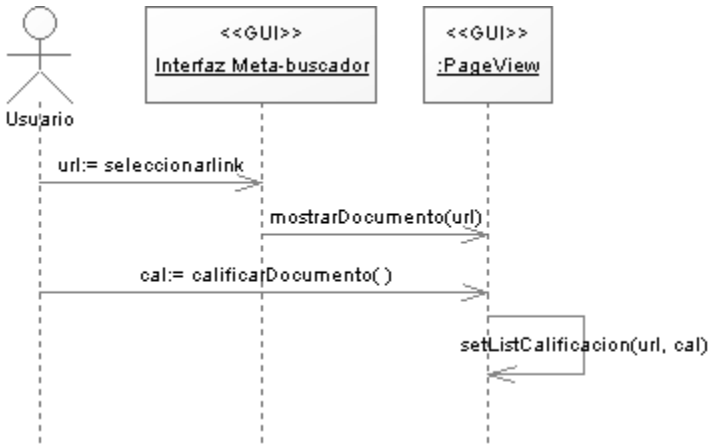

Figura 27. Diagrama de Secuencia Calificar Documento.

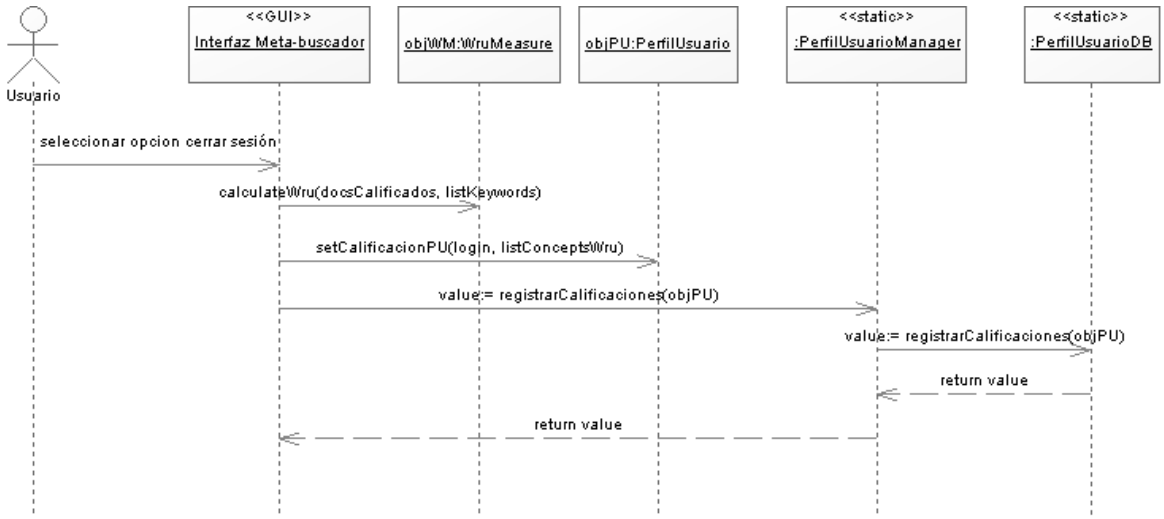

Figura 28. Diagrama de Secuencia Finalizar Sesión.

4.3.3 Diagrama de Clases

El diagrama de clases de la aplicación muestra las relaciones entre las clases necesarias para la construcción del prototipo. A continuación, en la Figura 29 se muestra el diagrama general.

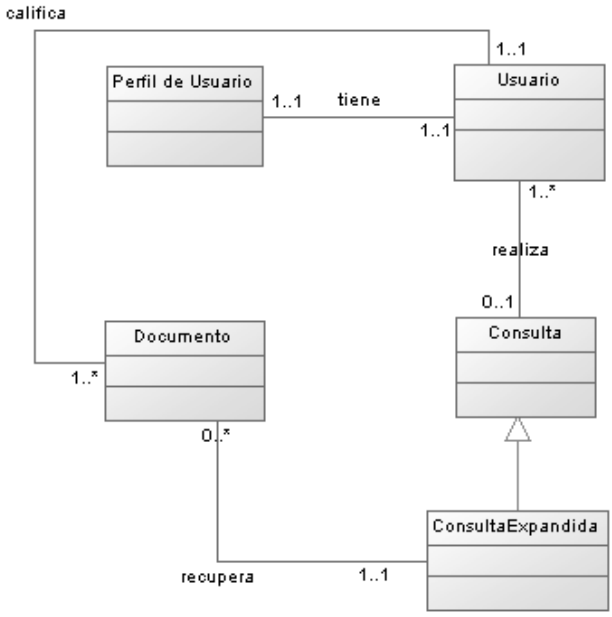

Figura 29. Diagrama de Clases.

4.3.4 Diagrama de Despliegue

El diagrama de despliegue contiene la forma como se muestran los componentes de la aplicación, teniendo en cuenta el servidor donde se implanta y los clientes que pueden acceder a su uso.

El diagrama mostrado en la Figura 30 representa la forma en que los componentes del prototipo software MSEC Web Search son desplegados en los diferentes elementos hardware. El primer nodo representa el equipo del usuario que accede a la aplicación a través de un navegador Web. El segundo nodo constituye el servidor Web donde se encuentran las capas de la arquitectura que se mostraron en la Figura 18, la capa de acceso a datos y la lógica de negocio de la aplicación, en ella se encuentra cada una de las clases que realizan los procesos de recuperación de los paginas a partir de la utilización de los servicios que ofrecen los motores de búsqueda, el manejo de la ontología, el cálculo del valor *Wru* de los conceptos, así como también el proceso de expansión de la consulta. Además de las dos capas mencionadas anteriormente, se encuentra la capa encargada de la lógica de presentación del proyecto y donde se implementa la interfaz grafica del usuario. En el Servicio Web de WordNet se realiza todo el proceso del manejo del diccionario de WordNet para el procesamiento de las consultas del usuario. Finalmente se muestra el servidor de bases de datos que aloja el motor SQL Server y en el cual se almacena la información correspondiente al perfil de usuario.

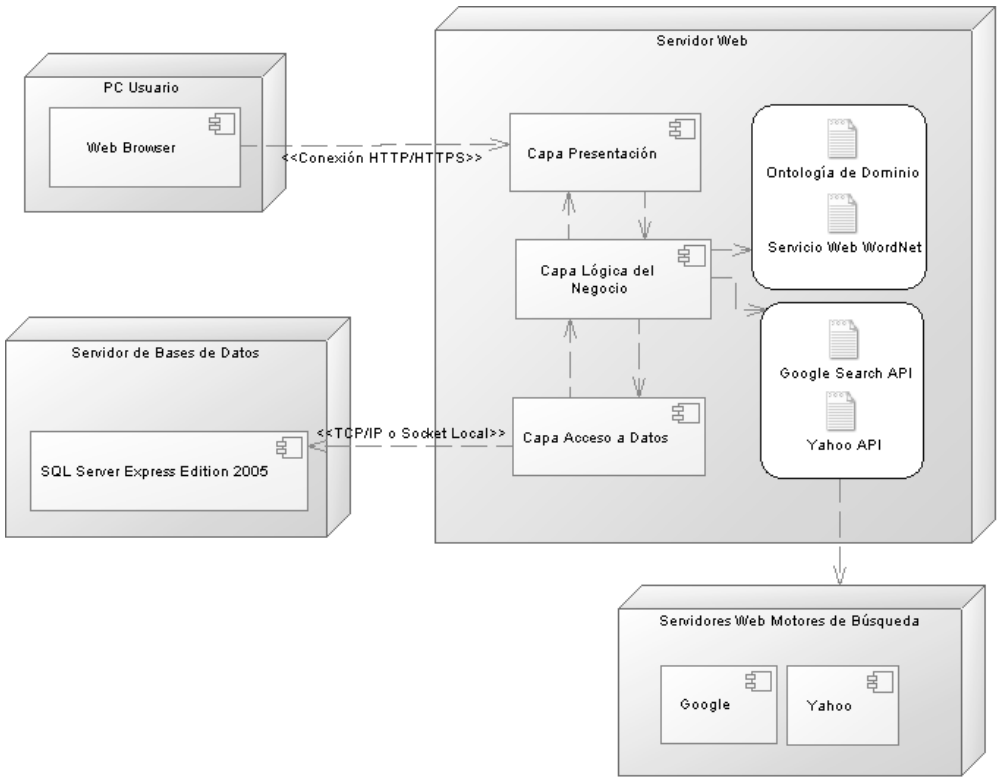

Figura 30. Diagrama de Despliegue.

4.3.5 Modelo de Base de Datos

En la Figura 31 se muestra el modelo relacional de la Base de Datos el cual involucra 3 tablas relacionales las cuales se describen en la Tabla 11.

Nombre de la Tabla	Descripción
TBL_CONCEPTO	Almacena todos los conceptos de la Ontología de dominio.
TBL_USUARIO	Almacena todos los usuarios del sistema.
TBL_PERFILUSUARIO	Asociado entre las tablas TBL_CONCEPTO y TBL_USUARIO, la cual almacena principalmente los valores ***Wru*** relacionados a un determinado concepto identificado en las consultas de un usuario.

Tabla 11. Descripción de las tablas de la Base de Datos.

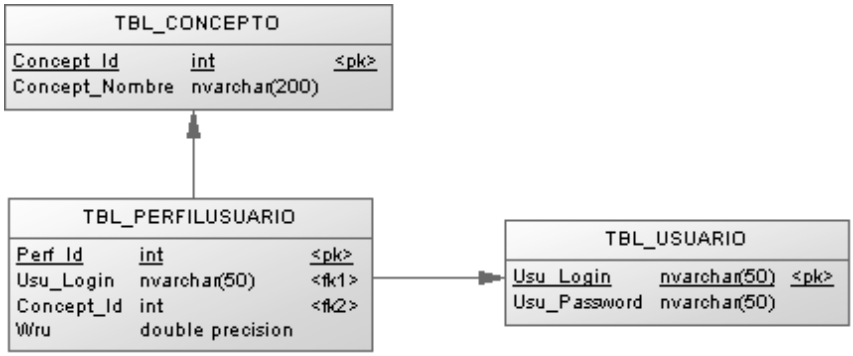

Figura 31. Modelo de Base de Datos.

4.3.6 Desarrollo del prototipo software

4.3.6.1 Iteración 1

En esta iteración se desarrolla la interfaz del prototipo software, para ello se tiene en cuenta la familiarización de un usuario típico con la interfaz de un buscador Web. Por esta razón las interfaces de usuario están orientadas hacia el estándar manejado en una interfaz gráfica tradicional de búsqueda, es decir una caja de texto en la que el usuario introduce su consulta, un botón que inicia el proceso de búsqueda y un logo distintivo con el nombre del buscador, esto fue mostrado en la Figura 20 al describir la capa de presentación de la arquitectura tres capas. En esta misma interfaz se muestran además los resultados de búsqueda con el título, y el resumen de cada documento. De la misma forma se debe contar con otra interfaz para mostrar el documento seleccionado por el usuario, aquí se cuenta con un sistema de calificación para que el usuario establezca que tan relevante es el documento para su necesidad de información, esto se presentó en la Figura 21.

4.3.6.2 Iteración 2

Aquí se implementa la funcionalidad del prototipo software y se realizan las pruebas preliminares del mismo. Para esto se tienen en cuenta las herramientas mencionadas al inicio de este capítulo y los artefactos resultantes de las anteriores fases. Las pruebas se describen con mayor detalle en el Anexo E.

4.3.6.3 Iteración 3

En esta iteración, el prototipo software se despliega en un servidor Web[38] y se incorpora al meta-buscador *GruWeb* con el objetivo de realizar las estadísticas y registro de datos que se utilizarán en el análisis de pruebas. En la actualidad el grupo de investigación GTI cuenta con una versión mejorada de *GruWeb* denominada *Minerva*, en este caso la

[38] En este caso el servidor Web es *Prometeo* perteneciente a la Universidad del Cauca y la url de MSEC Web Search es: http://prometeo.unicauca.edu.co/msec/ContentPages/InterfazBuscador.aspx.

integración se realiza con esta nueva versión por medio de de un servicio Web que ofrece la búsqueda de MSEC Web Search, al cual se envía una consulta de usuario y este retorna un conjunto de enlaces a páginas Web con su respectivo título, resumen y url para que estas puedan ser visualizadas en otra aplicación Web que haga uso de este servicio. Los detalles de cómo están desplegados los componentes de la aplicación se encuentran en la vista física mostrada en Anexo D.

4.4 Fase de transición

El enfoque principal de la fase de transición son las pruebas beta, las cuales se encargan de validar el prototipo software implementado respecto a las expectativas del usuario, para tal fin se entrenaron o capacitaron a los usuarios en el uso del sistema con el objetivo de ajustar factores como la usabilidad y el funcionamiento general del sistema. Con base a los resultados proporcionados por dichas pruebas, se realizan las correspondientes correcciones y modificaciones al sistema.

Las pruebas de usabilidad se realizaron teniendo en cuenta que el dominio de la aplicación esta relacionado con el área de la salud, por este motivo se seleccionaron usuarios pertenecientes a esta área, específicamente estudiantes de fisioterapia de noveno semestre de la Universidad del Cauca[39], los cuales accedieron voluntariamente a hacer las pruebas de la aplicación, de tal manera que cada estudiante realizó una prueba de usabilidad (Ver Anexo E para más información), donde se refleja su grado de satisfacción con la aplicación (flexibilidad, diseño, ayuda) y a su vez, se evalúa la relevancia de documentos retornados según las búsquedas que ellos realizaron. Se tuvieron en cuenta los aspectos:

- Visibilidad del estado del sistema
- Relación entre sistema y mundo real
- Consistencia y estándares
- Reconocer en lugar de recordar (reconocimiento del sitio donde se encuentran)
- Recuperación de Información (de acuerdo a algunas consultas realizadas por ellos)
- Ayuda y documentación
- ¿Cómo califica globalmente el sitio Web analizado?

Para cada aspecto se realizó una serie de preguntas a las que se debía responder, según el grado de satisfacción del evaluador, como: Excelente, Bueno, Neutro, Regular, Deficiente. Las preguntas y respuestas se pueden ver en detalle en el Anexo E.

Al terminar la prueba se realiza la ponderación de resultados que se muestra en la Figura 32.

[39] Al realizar estas pruebas no se contó con la disponibilidad de estudiantes de medicina de esta Universidad.

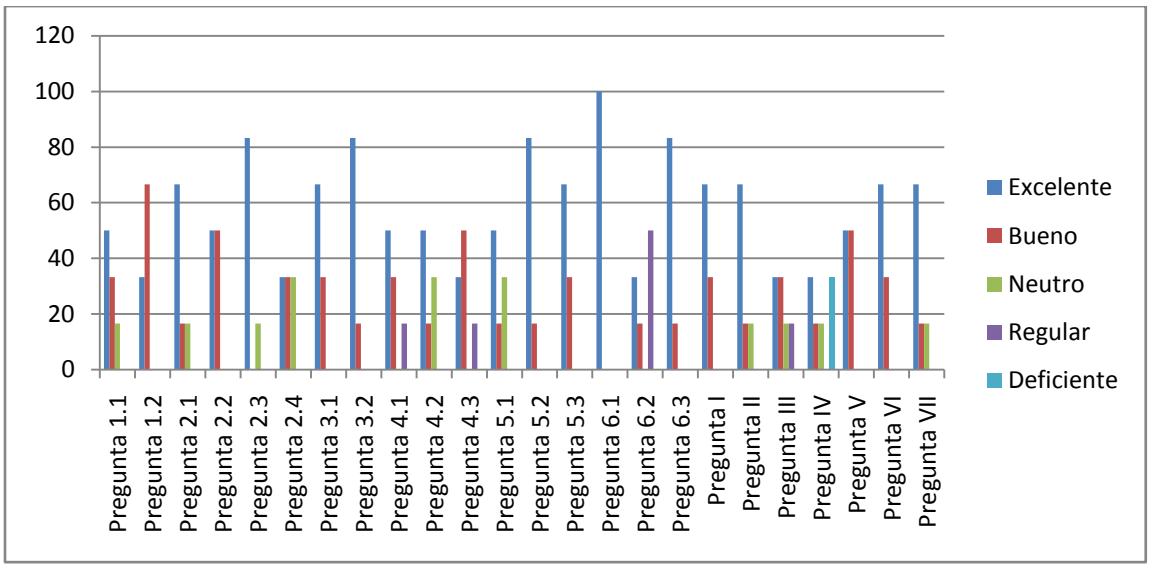

Figura 32. Resultados del test de usabilidad.

Como se observa en la Figura 32, los mayores resultados de calificación de la aplicación corresponden a los porcentajes de excelente y bueno, con lo que se concluye que los resultados son satisfactorios para los usuarios en cuanto a la usabilidad y al funcionamiento general del sistema. Los porcentajes y detalles de los resultados, así como las preguntas realizadas, se pueden consultar en el Anexo E.

Capítulo V

5 VALIDACIÓN DEL PROTOTIPO

"La relevancia es un concepto subjetivo, en el sentido que es la satisfacción de la necesidad humana, la última meta, solo que en ocasiones las personas tienen diferente punto de vista sobre la relevancia de un mismo documento. Lo que una persona considera relevante, para otra puede no serlo. Por lo que se puede decir que un documento es relevante para una consulta, usuario y colección en particular. Para otro usuario, otro tipo de colección o en otra consulta, el documento puede no tener la misma relevancia" [123].

Medir la efectividad de los SRI es un factor de suma importancia para conocer el desempeño del sistema, enfocándose en la calidad de la recuperación de documentos. Para llevar a cabo la evaluación del modelo propuesto, se desarrolló un prototipo de meta-buscador Web como fue descrito en el capítulo anterior, el cual ha sido denominado MSEC Web Search, este hace uso de la Ontología del Instituto Nacional del Cáncer de los Estados Unidos [119], 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 (ver Anexo C para mayores detalles acerca de esta Ontología).

De esta manera, para determinar el desempeño de MSEC, se plantean dos fases de evaluación. En estas fases se considera aplicar las medidas de Precision-Recall [10], Precision At k, Índice Mean Average Precision (MAP) [124, 125] y estadísticas Kappa [10]; todas estas medidas son ampliamente usadas en la evaluación de los SRI.

El objetivo de la primera fase de pruebas es determinar el formato adecuado para el texto de consulta que será enviada a los buscadores Web en la segunda etapa. Por este motivo y teniendo en cuenta que las medidas de Precisión y Recuerdo requieren de colecciones cerradas de documentos[40] y que Internet es un repositorio gigantesco de información, en el cual no se cuenta con los apropiados juicios de relevancia para los documentos, se opta por utilizar una colección cerrada de documentos para el cálculo de dichas medidas en la primera etapa de pruebas. Debido a que 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, se hace uso de la colección *MED* [126] compuesta por documentos cuya temática es la Medicina.

Mediante la segunda fase de evaluación se pretende evaluar la precisión de los documentos recuperados por MSEC Web Search respecto a los obtenidos con dos buscadores Web tradicionales (Google y Yahoo!), además de esto, se pretende comparar

[40] La relevancia de cada documento de acuerdo a una consulta específica ha sido previamente evaluada por uno o más expertos en el dominio.

los resultados del prototipo software propuesto, con los resultados de un buscador Web semántico especializado en el área de la salud como es el caso de GoPubMed [127].

5.1 Precisión

La precisión se puede definir como la proporción de materia recuperado realmente relevante, del total de los documentos recuperados [6]. Esta medida toma valores entre 0 y 1, donde valores cercanos a 1 representan un adecuado nivel en la precisión de la recuperación. Por consiguiente, la precisión está representada en la Ecuación 8.

$$Precisión = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ total\ de\ documentos\ recuperados}$$

Ecuación 8.

5.2 Recall

A esta medida también se le denomina “Recuerdo” o “Exhaustividad”, y representa la proporción de material relevante recuperado, del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no [6, 128]. De acuerdo a esto, el Recall se define en la Ecuación 9.

$$Recall = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ total\ de\ documentos\ relevantes}$$

Ecuación 9.

Lo ideal sería que ambas medidas se aproximen al 100%, sin embargo estas evolucionan inversamente en los sistemas reales, ya que para una consulta al retornar varios documentos, la relevancia de los mismos decrece y al aumentar la precisión disminuye el número de documentos retornados, excluyendo textos que posiblemente son relevantes en los resultados, por ende lo ideal es equilibrar la precisión y el recuerdo mediante el grado de detalle por parte del usuario en la formulación de sus consultas [9], sin embargo, Salton y McGill hacen especial hincapié en que debería ser el propio sistema el encargado de establecer mecanismos destinados a aumentar la precisión, sin disminuir por ello la tasa de rememoración. A continuación se describen otras medidas de evaluación.

5.3 Curva Precision-Recall

La curva de Precisión vs. Recall es ampliamente utilizada en el área de la RI, porque permite evaluar la eficacia de los resultados devueltos por un SRI respecto a un conjunto de consultas realizadas por el usuario. Esta medida permite establecer un balance entre la precisión y el recall, teniendo en cuenta que un sistema no será adecuado si presenta

una alta precisión con un bajo nivel de recall y viceversa [128]. La medida de Precision-Recall involucra ambas medidas en una sola gráfica para lo cual es necesario realizar una interpolación de los valores de precisión ajustándolos a cada punto de recall.

En la realización de la primera fase de las pruebas de relevancia se seleccionaron quince consultas de la colección de documentos de prueba (MED), es decir el 50% de las consultas ya que esta colección posee 30 consultas de prueba; cada una de estas se envía al meta-buscador MSEC Web Search para su expansión, por cada una de las consultas se han establecido cinco formatos de texto que son aplicados a las mismas antes de realizar el proceso de búsqueda sobre la colección de documentos[41]. Estos formatos de texto se describen en la Tabla 12.

Formato de Texto de Consulta	Descripción
Consulta sin Expansión	Consulta de usuario en su forma original, es decir, tal como fue digitada por el usuario a través de la interfaz gráfica del meta-buscador
Expansión con operadores (OR AND)	Consulta de usuario que ha sido expandida por medio del módulo de expansión de consulta descrito en el Capítulo 3 y a la cual se adicionan operadores lógicos OR y AND entre cada uno de los conceptos que la componen. La forma de adicionar estos operadores también fue descrita en el Capítulo 3.
Expansión con operadores (OR)	Consulta de usuario que ha sido expandida por medio del módulo de expansión de consulta y a la cual se adicionan únicamente operadores lógicos OR.
Expansión con operadores (OR AND) y PU	Consulta de usuario que ha sido expandida por medio del módulo de expansión de consulta y a la cual se adicionan operadores lógicos OR y AND, además, en el PU existen conceptos relacionados con la consulta los cuales son usados para refinar la expansión.
Expansión con operadores (OR) y PU	Consulta de usuario que ha sido expandida por medio del módulo de expansión de consulta y a la cual se adicionan únicamente operadores lógicos OR, además, en el PU existen conceptos relacionados con la consulta los cuales son usados para refinar la expansión.

Tabla 12. Formatos de texto aplicados a una consulta de usuario.

Por cada formato que se aplica a una consulta se recuperan una serie de documentos y posteriormente se calcula la precisión y el recuerdo en cada punto en que un documento relevante es recuperado, de esta manera, teniendo varios puntos de precisión por cada consulta se calcula además, la precisión promedio [123] que es utilizada más adelante para determinar el índice MAP. Los detalles de las pruebas realizadas para la evaluación de la RI en MSEC Web Search se presentan en el Anexo F.

En la Tabla 13 se muestra la precisión promedio obtenida por cada uno de los cinco formatos de texto que se aplicaron a tres consultas de ejemplo seleccionadas de las 15 que fueron realizadas en MSEC Web Search.

[41] En esta primera etapa de pruebas de relevancia se ha indexado MED Collection a través de Lucene .NET y la búsqueda se realiza de forma local, es decir sin el uso de los buscadores Web Google y Yahoo.

Consulta	Formatos de Texto de Consulta	No. Documentos Relevantes Recuperados	Precisión Promedio
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	Consulta sin Expansión	8	0,512
	Expansión con operadores (OR AND)	8	0,455
	Expansión con operadores (OR)	5	0,342
	Expansión con operadores (OR AND) y PU	9	0,649
	Expansión con operadores (OR) y PU	6	0,376
method for experimental production of and known cause of hydrocephalus in animal and human	Consulta sin Expansión	10	0.562
	Expansión con operadores (OR AND)	9	0.483
	Expansión con operadores (OR)	7	0.415
	Expansión con operadores (OR AND) y PU	12	0.579
	Expansión con operadores (OR) y PU	8	0.432
palliation of cancer patients by using drugs, x-ray, physical phenomena or properties	Consulta sin Expansión	7	0,410
	Expansión con operadores (OR AND)	8	0,577
	Expansión con operadores (OR)	7	0,392
	Expansión con operadores (OR AND) y PU	10	0,735
	Expansión con operadores (OR) y PU	7	0,460

Tabla 13. Precisión promedio de los cinco formatos de texto aplicados a tres consultas de usuario.

5.3.1 Resultados medida Precision-Recall

Una vez realizadas las pruebas necesarias para calcular la precisión y el recuerdo de los documentos retornados por cada formato de texto aplicado anteriormente, se procede a realizar las curvas de precision-recall por cada consulta. A manera de ejemplo, en la Figura 33, Figura 34 y Figura 35 se muestran las tres consultas seleccionadas anteriormente. En cada una de estas figuras se observan cinco curvas que corresponden a los cinco formatos de texto aplicados a la respectiva consulta, estos formatos se pueden observar en la Figura 36.

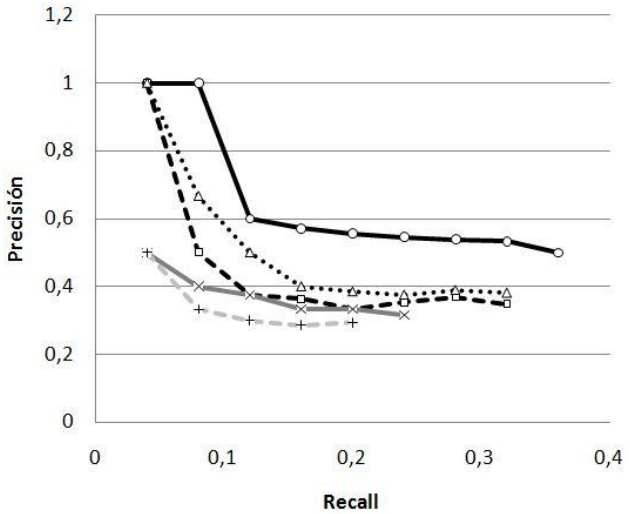

Figura 33. Curvas Precisión-Recuerdo para la primera consulta.

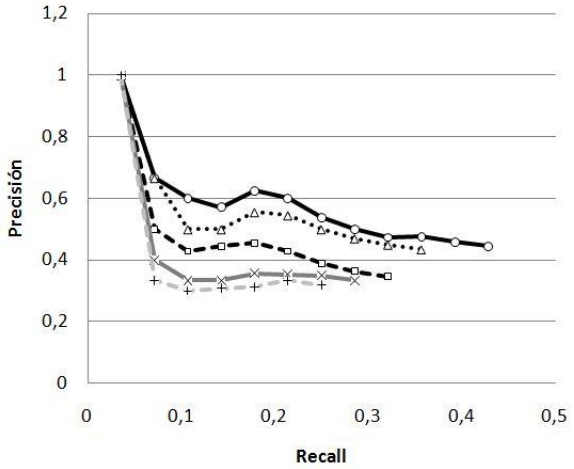

Figura 34. Curvas Precisión-Recuerdo para la segunda consulta.

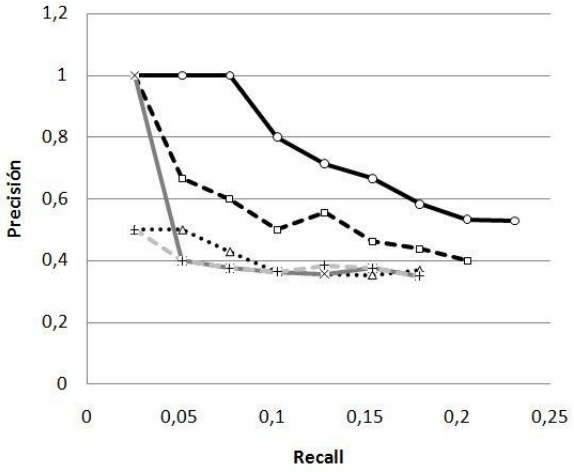

Figura 35. Curvas Precisión-Recuerdo para la tercera consulta.

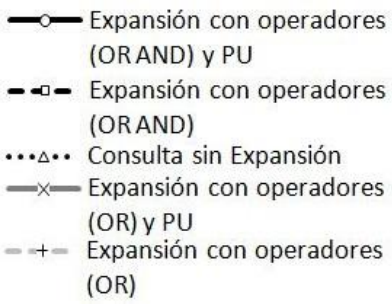

Figura 36. Formatos de texto aplicados a una consulta expandida.

En las tres gráficas de Precisión-Recuerdo 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. Adicionalmente, de las 15 consultas seleccionadas, 11 de estas presentaron un mejor comportamiento de este formato de texto de consulta. Lo anterior indica que existen más documentos

posicionados en los primeros lugares del ranking que se muestra al usuario gracias al uso de este formato. A partir de estos resultados experimentales 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 más documentos relevantes en las primeras posiciones del ranking que se muestra al usuario.

5.4 Cálculo de índice MAP para formatos de texto de consulta

Finalmente, la medida Mean Average Precision (MAP) [124, 125], es el promedio del valor de precisión media de un conjunto de consultas, es decir, la medida de evaluación MAP ofrece la posibilidad de conocer el comportamiento del sistema en forma global por medio de la ubicación de los documentos recuperados que son relevantes para el usuario en un conjunto de consultas, generando un promedio para cada consulta de acuerdo a la ubicación de cada recurso relevante sobre la lista de resultados; finalmente se calcula el promedio total del conjunto de consultas lo que permite establecer el comportamiento general del sistema.

Para este cálculo se tiene en cuenta la precisión promedio en cada consulta realizada anteriormente, es decir, la precisión media de las quince consultas. En la Tabla 14 se presenta el índice MAP para cada formato de texto aplicado a cada una de las consultas. Para mayores detalles, en el Anexo F se muestra el cálculo del índice MAP con el que se obtuvieron estos resultados.

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

Tabla 14. Índice MAP por cada formato de texto aplicado a las consultas.

El valor MAP de la consultas expandidas mediante los operadores lógicos *OR-AND* combinado con el PU presenta un promedio total en las 15 consultas 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 [10]. Con base en lo anterior se opta por utilizar este formato para la segunda fase de evaluación que es presentada en las siguientes secciones, esta evaluación se realiza esta vez en la Web, es decir, sin tomar una colección cerrada de documentos para la evaluación.

5.5 Precisión en los k primeros resultados

Esta evaluación se realizó comparando los resultados de MSEC Web Search con los resultados de dos motores de búsqueda tradicionales (Google y Yahoo). En este proceso se contó con 15 usuarios pertenecientes al programa de fisioterapia de la Universidad del Cauca y se llevó a cabo mediante la técnica de evaluación a ciegas, en la cual los usuarios no saben cual sistema están evaluando a través de una sola interfaz, la cual oculta las características que hacen alusión a los buscadores utilizados y lógicamente a MSEC Web Search.

La Tabla 15 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. En el Anexo F se muestran los juicios de relevancia respectivos para el cálculo de la precisión en k resultados.

Documento	Google		Yahoo!		MSEC Web Search	
	% Exa	% Pre At -k	% Exa	% Pre At -k	% Exa	% Pre At -k
1	100	100	86,66	86,66	100	100
2	100	100	93,33	89,99	100	100
3	100	100	86,66	87,77	100	100
4	93,33	98,33	73,33	84,43	100	100
5	93,33	98,33	66,66	83,10	100	100
6	80	96,11	66,66	83,10	93,33	98,88
7	60	93,25	46,66	80,24	86,66	97,92
8	46,66	91,58	33,33	78,57	80	97,1
9	26,66	89,36	33,33	78,57	66,66	95,61
10	13,33	88,02	6,66	75,90	46,66	93,61

Tabla 15. Resultados de Precisión en K=10

La columna (% Exa) indica la exactitud en cada uno de los documentos recuperados, la cual se define por (número de personas que consideraron el ítem como relevante / número de personas que realizaron la prueba (15)) * 100. La columna (% Pre At-k) muestra la precisión At–k en cada uno de los documentos recuperados en los sistemas evaluados.

En la Tabla 15 se pueden observar los resultados de la precisión en los K primeros resultados (k=10) para tres diferentes SRI (Google, Yahoo! y MSEC Web Search), estos resultados muestran que Google como punto máximo de exactitud tiene 100% y como punto mínimo 13,33% con lo que se establece una precisión para k=10 que oscila entre el 88,02% y el 100%. Para Yahoo! se obtiene un punto máximo de exactitud de 93,33% con un punto mínimo de 6,66%, esto conlleva a obtener una precisión para k=10 en el rango de 75,9% a 89,99%. Por último para MSEC Web Search obtiene un punto máximo de exactitud en 100% y un punto mínimo en 46,66%, además la precisión para k=10 oscila entre 93,61% y 100%, esto demuestra que los 10 primeros resultados mostrados por MSEC Web Search son más relevantes que los retornados por los dos buscadores con los que se compara.

La Figura 37 presenta la precisión en k documentos recuperados, para cada uno de los sistemas y permite compararlos con el modelo propuesto.

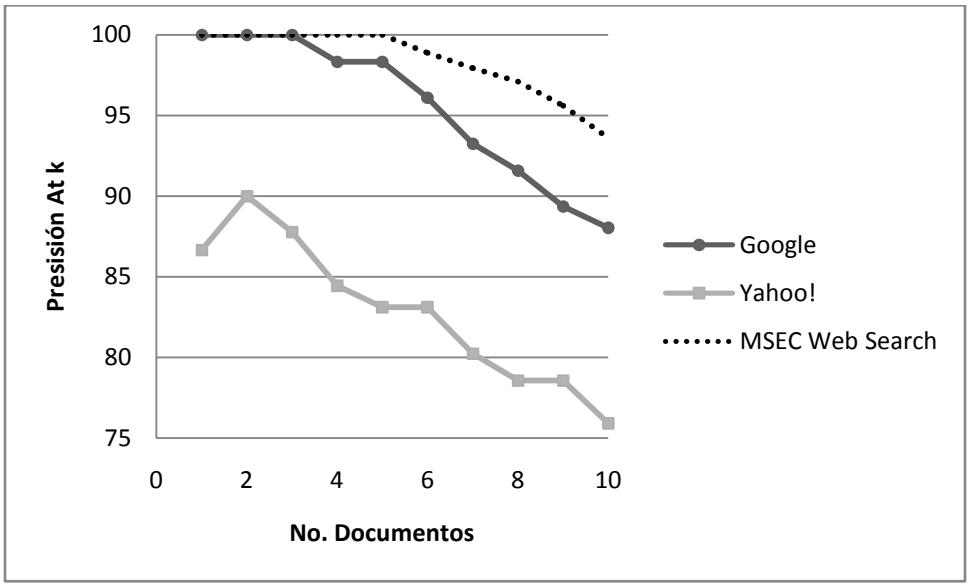

Figura 37. Precisión At k (k=10) para Google, Yahoo! y MSEC Web Search

La Figura 37 muestra como resultado, la precisión promedio en diferentes valores de k, con resultados comprendidos entre el 93,61% y 100% para MSEC Web Search, mostrando así el alto grado de precisión del modelo propuesto, 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.

5.6 Cálculo de índice MAP para varios SRI

En la Tabla 16 se muestran los resultados de la medida general de precisión (Mean Average Precision, MAP), la cual es utilizada para medir de una forma más general la precisión de los sistemas de búsqueda.

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

Tabla 16. Comparativo del índice MAP entre los SRI evaluados

La tabla anterior presenta los resultados del índice MAP tomando como base las consultas realizadas por los usuarios en los tres sistemas comparados en la segunda etapa de pruebas. Los valores corresponden a los promedios de Precisión at k reportados en la Tabla 15 y de dos consultas mas que se encuentran consignados en el Anexo F.

Vale la pena 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!), en un mínimo de 3,5% que se logra frente a Google y de 8,8% que se consigue frente a Yahoo!.

5.7 Estadísticas Kappa

Para evaluar los resultados obtenidos por MSEC Web Search con el índice Kappa, se tuvo en cuenta que dos aspectos distintos entran a formar parte del estudio de fiabilidad: por una parte, la concordancia entre observadores, es decir, hasta qué punto los observadores no coinciden en su medición, y por otra parte el sesgo entre observadores definido como la tendencia de un observador a dar consistentemente valores mayores que el otro.

Dado lo anterior se utilizó el *índice Kappa de Fleiss* [129, 130] que trabaja con cualquier número de observadores que proporcionan grados categóricos (en este caso R = Relevante, I = No relevante o Irrelevante), a un número fijo de documentos (en este caso de k=1 hasta 10 documentos recuperados de la web). Esta medida puede ser interpretada como el grado de concordancia observada entre varios observadores con una medida de consistencia la cual puede tomar valores entre -1 y +1. Mientras más cercano a +1, mayor es el grado de concordancia entre los observadores, por el contrario, mientras más cercano a -1, mayor es el grado de discordancia entre dichos observadores. Esta medida se calcula basada en las siguientes ecuaciones.

Primero se calcula $\mathrm{p_j}$ que representa la proporción de todos los jueces que se inclinaron por una determinada categoría, es decir, Relevante o No relevante (Ecuación 10).

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$$

Ecuación 10.

Ahora se calcula $\mathrm{P_i}$, este indica cuantas parejas juez-juez están de acuerdo, relativo al número de todos los posibles pares juez-juez (Ecuación 11).

$$P_i = \frac{1}{n(n-1)}\left[\left(\sum_{j=1}^{k} {n_{ij}}^2\right) - n\right]$$

Ecuación 11.

Posteriormente se calcula $\bar{P}$ que representa la media de los P_i (Ecuación 12) y la probabilidad de acuerdo entre dos jueces por azar (Ecuación 13).

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

Ecuación 12.

$$\bar{P}_e = \sum_{j=1}^{k} {p_j}^2$$

Ecuación 13.

Finalmente se calcula el índice Kappa de Fleiss (Ecuación 14).

$$Kappa_f = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Ecuación 14.

Se calculó el índice Kappa de Fleiss, en los diez primeros documentos recuperados, para ser evaluados con base en el juicio de Relevancia y No Relevancia marcada por los 15 observadores.

En la Tabla 17 se muestra el juicio de los 15 observadores en los primeros 10 documentos recuperados para una misma consulta.

DOCUMENTO	OBSERVADORES (USUARIOS)														
	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
1	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
2	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
3	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
4	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
5	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
6	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
7	I	R	R	I	R	R	I	I	I	R	R	I	I	I	R
8	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
9	I	I	I	R	I	I	R	R	I	I	I	I	I	I	I
10	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R

Tabla 17. Apreciación de los observadores en k=10 resultados

En la Tabla 18 se puede observar que teniendo como base el primer documento (k=1), el 100% de los observadores (jueces) lo consideran relevante. Cuando se evalúan los 10 primeros documentos (k=10) se tiene que un 86,6% de los jueces considera los documentos como relevantes mientras que el restante 13,3% los considera como no relevantes. El índice Kappa para los primeros 10 documentos muestra que el número de personas que están de acuerdo es considerable y se ubica dentro de una concordancia sustancial de acuerdo a los rangos establecidos por Landis y Koch [131] (índice de Kappa de Fleiss de 0,6208 de tal manera que se encuentra dentro del rango 0,61 – 0,80).

Documento	Total: R	Total: I	*Pi*
1	15	0	1
2	15	0	1
3	15	0	1
4	15	0	1
5	15	0	1
6	15	0	1
7	7	8	0,46666667
8	15	0	1
9	3	12	0,65714286
10	15	0	1
Totales	130	20	9,12380952
pi	0,86666667	0,13333333	
pi^2	0,75111111	0,01777778	
$\overline{P}$	0,91238095		
$\overline{P}e$	0,76888889		
Kappa	0,62087912		

Tabla 18. Resultado del índice kappa en k=10

Estas pruebas se repitieron para dos consultas más, obteniéndose buenos resultados respecto a la concordancia de los observadores. Los detalles se encuentran consignados en el Anexo F.

5.8 Precisión de MSEC Web Search vs. GoPubMed

Dada la evidente superioridad de los resultados de precisión obtenidos con el prototipo desarrollado en este proyecto, MSEC Web Search, frente a los resultados de dos de los buscadores tradicionales más utilizados como Google y Yahoo!, se hace necesario comparar dicho prototipo con un buscador que utilice características similares en su búsqueda. Cabe resaltar que estas pruebas se realizan como una validación adicional del modelo propuesto y se encuentran por fuera de los objetivos de este proyecto. La razón para incluir estas pruebas en el presente documento, es mostrar que tan buenos son los resultados de precisión de MSEC Web Search respecto a un buscador que utilice la semántica como base fundamental en sus búsquedas y que además realice dichas búsquedas en el mismo dominio que se ha establecido en este trabajo (Medicina, específicamente la Oncología).

Para establecer la comparación que se menciona en el párrafo anterior, se realizó una selección de los buscadores que cumplieran con las características expuestas anteriormente. Actualmente existen muy pocos buscadores semánticos que cumplan con dichas características; el buscador semántico más adecuado fue GoPubMed [127], dado que se especializa en la recuperación de textos biomédicos de la base de datos de MEDLINE [132] por medio del uso de la Ontología *Gene Ontology* (GO) [133] y el vocabulario controlado *MeSH* (Medical Subject Headings) [100], este sistema fue desarrollado en la Universidad Técnica de Dresden por Michael Schroeder y su equipo de Transinsight [134] y ha sido reconocido con el “Red Dot Award 2009” [135], un prestigioso

premio en la categoría de diseño de comunicaciones, interfaces gráficas de usuario y herramienta interactiva, además del premio a la industria Alemana 2010 [134]. Es importante mencionar que de acuerdo a los estudios de [140], GoPubMed presentó una precisión del 78,9% y un recall del 83,3%.

Para realizar esta evaluación se utiliza la medida de Precisión y el índice MAP. Dado que GoPubMed trabaja sobre la base de datos MEDLINE, se hace necesario adaptar a MSEC Web Search para que realice sus búsquedas en esta misma colección de documentos, para ello se obtienen los textos biomédicos por medio del software RefNavigator [136], el cual permite obtener los artículos médicos de MEDLINE, y posteriormente se realiza una indexación de estos documentos con Lucene .NET para finalmente realizar la búsqueda con MSEC Web Search.

Para dar una muestra de esta evaluación, en La Figura 38 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)"*.

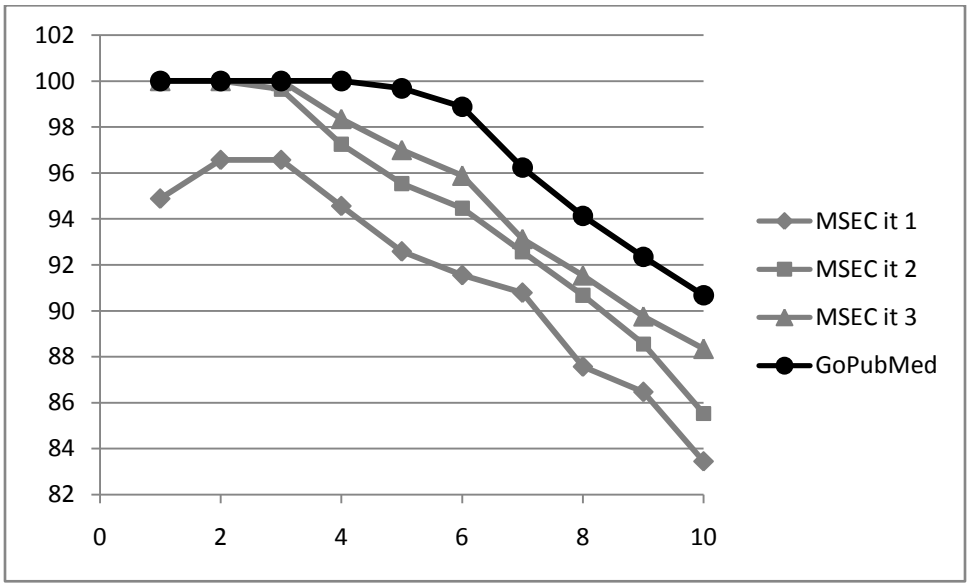

Figura 38. Precisión de GoPubMed y MSEC Web Search

En la Figura 38 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. Si bien en las siguientes iteraciones se incrementaron los valores de Precisión-Recuerdo para los documentos recuperados por el prototipo software de este proyecto, sólo se logró una aproximación a la precisión de GoPubMed. Se realizó el

mismo procedimiento con 15 consultas obteniendo resultados similares, los cuales se muestran en el Anexo F.

Finalmente, en la Tabla 19 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).

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

Tabla 19. Comparativo del índice MAP entre GoPubMed y MSEC Web Search

El índice MAP refleja los valores generales de precisión para ambos SRI, en donde se establece una diferencia del 8,6% de GoPubMed sobre MSEC Web Search. Si bien existe una superioridad del primer SRI en cuanto a la precisión de sus resultados, no es tan amplia respecto al segundo, si se tiene en cuenta que estos resultados pueden deberse a que GoPubMed maneja una Ontología muy orientada a MEDLINE, además de manejar un vocabulario más especializado como la terminología MeSH cuando realiza la expansión de las consultas, el cual ha sido desarrollado especialmente para trabajar con la base de datos de MEDLINE.

5.9 Ejemplo de expansión de consulta y realimentación del PU

A continuación se muestra el proceso de expansión de una consulta de ejemplo por medio de la ontología de dominio y de la realimentación del PU, de igual forma se muestra la variación de los valores *Wru* de los conceptos de dicha consulta.

Consulta original del usuario	**Consulta expandida**	**Valores Wru en el PU**	
		Concepto	**Wru**
"breast cancer"	*"breast cancer OR malignant breast neoplasm (breast AND tissue)"*	*breast cancer*	0,715
		malignant breast neoplasm	0,423
		breast	0,825
		tissue	0,156
"treatment for breast cancer"	*"cancer treatment OR (breast carcinoma therapy) OR (breast cancer OR malignant breast neoplasm)"*	*breast cancer*	0,203
		malignant breast neoplasm	0,728
		breast	0,825
		tissue	0,039
		cancer treatment	0,589
		breast carcinoma therapy	0,492
"innovations for cancer treatment"	*"cancer treatment OR breast carcinoma therapy OR (innovations*	*breast cancer*	0,050
		malignant breast	0,182

	AND laser surgery research)"	*neoplasm*	
		breast	0,899
		tissue	0,009
		cancer treatment	0,715
		breast carcinoma therapy	0,597
		laser surgery research	0,368
"technology for surgery of malignant tumors"	*"cancer OR malignant tumors AND (cancer treatment OR laser surgery research OR technology)"*	*breast cancer*	0,012
		malignant breast neoplasm	0,045
		breast	0,224
		tissue	0,000
		cancer treatment	0,845
		breast carcinoma therapy	0,149
		laser surgery research	0,598
		cancer	0,618
"future technology for cancer cure"	*"cancer OR malignant tumors AND (laser surgery research OR cancer treatment OR technology) AND (new research OR future)"*	*breast cancer*	0,003
		malignant breast neoplasm	0,011
		breast	0,056
		tissue	0,000
		cancer treatment	0,887
		breast carcinoma therapy	0,037
		laser surgery research	0,680
		cancer	0,836
		new research	0,328

Capítulo VI

6 CUMPLIMIENTO DE OBJETIVOS

Para determinar el cumplimiento de los objetivos, a continuación se presenta un modelo de indicadores que permite evaluar de manera objetiva el cumplimiento de los mismos en el presente proyecto, de esta manera se exponen los lineamientos de conformación e interpretación de los indicadores propuestos.

6.1 Lineamientos de conformación e interpretación de los indicadores

Con el fin de expresar los resultados finales de cada uno de los objetivos se presenta a continuación una explicación de los tipos de indicadores utilizados en la evaluación de los resultados y la forma correcta de interpretarlos.

Los indicadores de desempeño que se evalúan, básicamente adoptan la forma de un cociente, en el cual, el denominador es un valor numérico que ayuda a efectuar la comparación con el logro obtenido así:

$$Indicador = \left(\frac{Numerador}{Denominador}\right) * FactorEscala$$

Ecuación 15.

De esta forma se definen los siguientes modelos de indicadores que se deben personalizar y aplicar a los actores, productos, funciones, etc. dependiendo del contexto del objetivo evaluado.

- **Indicador de Cobertura (IC).** Determina la cantidad de elementos cobijados por un producto o estrategia.

$$Cobertura = \left(\frac{Número\ de\ nodos\ beneficiados\ con\ el\ servicio}{Número\ de\ nodos\ que\ se\ espera\ servir}\right) * 100$$

Ecuación 16.

- **Indicador de Eficacia (IE).** Permite analizar el cumplimiento con los requisitos definidos.

$$Eficacia = \left(\frac{Recursos\ Ejercidos}{Recursos\ Asignados}\right) * 100$$

Ecuación 17.

- **Indicador de Eficiencia (IF).** Permite identificar la relación que existe entre las metas alcanzadas, el tiempo y los recursos consumidos con respecto a un estándar. Representa el buen uso de los recursos.

$$Eficiencia = \left(\frac{Metas\ alcanzadas}{Recursos\ consumidos}\right) * 100$$

Ecuación 18.

- **Indicador de Calidad (IQ).** Están orientados a medir la satisfacción de los beneficiarios.

$$Eficiencia = Calificación\ entre\ (1: Mala\ (0\%), 2: Regular\ (50\%), 3: Buena(75\%), 4: Excelente(100\%))$$

Ecuación 19.

Con el modelo de indicadores aquí presentado, se desarrolló un conjunto de indicadores que permiten evaluar adecuadamente el nivel de cumplimiento de cada uno de los objetivos. A continuación se presenta la evaluación realizada.

6.2 Descripción y alcance del cumplimiento de los objetivos

En la Tabla 20**¡Error! No se encuentra el origen de la referencia.**, Tabla 21, y Tabla 22, se especifican los objetivos comprometidos en el proyecto, los productos esperados derivados de cada objetivo, los resultados obtenidos, los indicadores que evalúan el objetivo, los medios de verificación de los resultados y finalmente, unas observaciones que permiten aclarar los resultados en cada objetivo.

Se desarrolla una tabla por cada objetivo específico comprometido en la propuesta del proyecto *"Modelo Semántico de Expansión de Consultas para la Búsqueda Web"*.

No. Objetivo	1
Descripción del objetivo	Definir un modelo semántico de expansión de consultas con las siguientes características: • La expansión de la consulta es implícita y para ello utilizará una Ontología de dominio que permita analizar la consulta vista como una sola unidad semántica, partiendo de los términos originales digitados por el usuario. • El perfil de usuario debe permitir almacenar, procesar y recuperar la información del usuario que hace la consulta, con el fin de integrarlo a un Sistema de búsqueda Web.
Productos esperados	1. Documento de especificación de cada modulo del modelo semántico de expansión de consultas, basado en ontologías de dominio incluyendo el perfil de usuario. 2. Artículo de investigación sobre la definición del modelo semántico enviado a una revista nacional y/o internacional donde se describa la investigación realizada.
Resultados obtenidos	1. Documento de especificación de cada modulo del modelo semántico de expansión de consultas, basado en ontologías de dominio incluyendo el perfil de usuario. 2. Realización del articulo denominado: "Modelo Semántico de Expansión de

	Consultas para la Búsqueda Web - MSEC" en la "Revista Facultad de Ingeniería" de la Universidad de Antioquia. Estado: en evaluación.
Indicadores (Escala * 100)	**Eficacia** $$IE1 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{2}{2} * 100 = 100\%$$ **Calidad** *IQ1 = ¿El modelo propuesto permite realizar la expansión de la consulta de forma implícita al analizar la consulta vista como una sola unidad semántica, partiendo de los términos originales digitados por el usuario?* *R* = El modelo permite ver la consulta como una unidad semántica mediante la combinación de una ontología de dominio y recursos léxicos para encontrar conceptos adicionales relacionados semánticamente a los conceptos identificados en la consulta original, además de esto se aplican técnicas como la similitud semántica y el perfil de usuario para refinar aún más la consulta. $IQ1 = 4 = 100\%$ *IQ2 = ¿El Perfil de Usuario utilizado en el modelo propuesto permite almacenar, procesar y recuperar la información del usuario que hace la consulta con el fin de que dicho perfil pueda ser integrarlo a un Sistema de búsqueda Web?* *R* = El modelo propuesto maneja un Perfil de Usuario en el cual se almacena, se procesa y se recupera la información del usuario específicamente aquella concerniente a los conceptos de interés que son establecidos a través de su historial de búsqueda. $IQ2 = 4 = 100\%$ *IQ3 = ¿El impacto del modelo semántico fue el esperado?* *R* = El modelo se ha realizado de acuerdo a la propuesta, sin embargo, se debería probar en más dominios del conocimiento y ser aceptado por la comunidad científica, para obtener el impacto esperado. $IQ3 = 3 = 75\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 1 = \frac{100+100}{2} = 100\%$
Medios de verificación	1. En el Capítulo 3 de la presente investigación se encuentra el modelo semántico propuesto para la expansión de consulta. Además se encuentra la descripción de cada modulo del mismo. 2. El Artículo sobre MSEC se encuentra en el Anexo B.
Estrategias, problemas y/o observaciones	Se ha definido un modelo semántico para la expansión de la consulta basado en Ontologías de dominio, las cuales permiten utilizar técnicas como la similitud semántica y el perfil de usuario como medio de expansión de la consulta, con el propósito de recuperar información relevante para el usuario. El modelo propuesto, identifica una serie de módulos, que describen el proceso para la expansión de consulta como módulo principal, dicho proceso está conformado por módulos los cuales están descritos en el Capitulo 3. Respecto a los inconvenientes encontrados durante la creación de MSEC se tiene la gran cantidad de información dispersa existente sobre el tema, descartando una parte de información importante que posteriormente se descubrió cuando el modelo estaba construido al llevar a cabo varias iteraciones sobre las fases de construcción, por ende esto condujo a realizar varias iteraciones más en la fase de formulación del modelo y de esta manera adaptarlo a los nuevos temas investigados.

Tabla 20. Cumplimiento del primer objetivo específico.

No. Objetivo	2
Descripción del objetivo	Desarrollar un prototipo software, basado en el modelo anterior, integrándolo al meta-buscador GruWeb[42] con el fin de incrementar su funcionalidad hacia la gestión de perfiles de usuario.
Productos esperados	1. Código fuente del prototipo software. 2. Documentos de Análisis y Diseño del prototipo. 3. Interfaz de GruWeb interactuando con el prototipo software desarrollado.
Resultados obtenidos	1. Código fuente del prototipo software denominado MSEC Web Search. 1.1. Meta-buscador semántico con interfaz propia para acceder desde la web. 2. Documentos de trabajo generados para construir el meta-buscador Web. 3. Interacción con el meta-buscador GruWeb para realizar búsquedas semánticas cuando el usuario lo requiera.
Indicadores (Escala * 100)	**Eficacia** $IE1 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{3}{3} * 100 = 100\%$ **Calidad.** *IQ1 = ¿Se desarrolló el prototipo software basado en el modelo propuesto?* *R* = El prototipo software fue desarrollado siguiendo el ciclo de vida de desarrollo software bajo la metodología UP Ágil y siguiendo cada uno de los módulos definidos en el modelo propuesto. Para ello, se tienen en cuenta los pasos y sus actividades de acuerdo al dominio y entorno en que fue construido. $IQ1 = 4 = 100\%$ *IQ2 = ¿Se realizó la integración del prototipo con el meta-buscador planteado?* *R* = El prototipo desarrollado se integró con el meta-buscador planteado, después de realizar la evaluación del mismo con los expertos en el dominio de la salud. $IQ2 = 4 = 100\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 2 = \frac{100}{1} = 100\%$
Medios de verificación	1. En el capítulo 4 de este documento se describe el proceso de desarrollo del prototipo software. Se puede acceder al meta-buscador semántico en: ***http://prometeo.unicauca.edu.co/msec/ContentPages/Usuarios/InterfazBuscador.aspx*** 2. El código fuente del meta-buscador es anexado digitalmente. 3. En las secciones 4.1, 4.2, 4.3 y 4.4 del presente documento se describe el proceso de análisis, diseño, implementación y despliegue del prototipo creado, además en el Anexo D se encuentran la documentación extendida de dicho proceso. 4. La integración con el meta-buscador GruWeb y el enlace directo al buscador semántico se encuentra en ***http://spar.unicauca.edu.co/gruweb*** o su versión mejorada ***http://spar.unicauca.edu.co/minerva***.
Estrategias, problemas y/u observaciones	El meta-buscador semántico es el prototipo desarrollado basado en el Modelo Semántico de Expansión de Consultas - MSEC. Es una herramienta de búsqueda disponible en la Web, que permite realizar búsquedas principalmente en idioma inglés sin embargo admite búsquedas en español sobre ciencias de la salud específicamente la Oncología. Está dirigido a estudiantes de educación superior pertenecientes a programas afines a las ciencias de la salud, sin embargo puede ser utilizado por cualquier usuario que busque y necesite resultados en esa área. La evaluación de dicho prototipo se realizó mediante la interfaz de búsqueda desarrollada, no la del meta-buscador GruWeb, para encontrar resultados en cuanto a usabilidad y eficiencia neta de la aplicación creada. Posteriormente se realizó la

[42] El meta-buscador se encuentra disponible en http://spar.unicauca.edu.co/groupweb. Puede utilizarse, en su defecto, el meta-buscador que está en proceso de desarrollo y es una versión mejorada del anterior; está disponible en http://spar.unicauca.edu.co/minerva.

	integración con el meta-buscador GruWeb. Debido a que la integración del prototipo software se realizó con Minerva y no con GruWeb, vale la pena aclarar que estos meta-buscadores actualmente ya cuentan con la gestión de perfiles de usuario. En el momento de definir la propuesta de este trabajo de grado, estos no contaban con esta funcionalidad, de modo que la integración del prototipo MSEC Web Search se realiza únicamente a través de un servicio Web el cual a partir de una consulta retorna un conjunto de enlaces a páginas Web y este servicio puede ser utilizado por Minerva activándolo en su panel de configuración de búsqueda.

Tabla 21. Cumplimiento del segundo objetivo específico.

No. Objetivo	3
Descripción del objetivo	Determinar la precisión de los resultados de búsqueda obtenidos por el prototipo software, evaluándolo a través de medidas como radio precisión[43], proporción recuerdo[44] e Índice MAP[45] [123]
Productos esperados	1. Documento de evaluación y resultados con medidas y estadísticas definidas.
Resultados obtenidos	1. En el capítulo 5 de la presente investigación y el Anexo F, se muestra la evaluación y los resultados para cada ítem propuesto, además de la comparación en relevancia de resultados con los buscadores Web tradicionales.
Indicadores (Escala * 100)	**Eficacia** $IE1 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{1}{1} * 100 = 100\%$ **Calidad** *IQ1 = ¿Se realizaron pruebas para verificar el funcionamiento del prototipo software?* *R* = Las primeras pruebas para verificar el funcionamiento del prototipo fueron las pruebas alfa que representan la primera etapa de pruebas de relevancia (sección 5.3 y 5.4) y beta (sección 4.4), las cuales fueron realizadas por el equipo de desarrollo. Las anteriores pruebas determinaron la funcionalidad actual, teniendo en cuenta los resultados arrojados, con lo cual se definieron las medidas: Índice MAP y curva Precisión-recuerdo. Posteriormente se realizó la validación del prototipo con estudiantes de fisioterapia de la Universidad del Cauca con el fin de evaluar la eficiencia del sistema en el entorno al que está dirigido, en esta etapa de pruebas se aplicaron medidas como Precision At K, índice MAP y estadísticas Kappa. En este caso se realizó la respectiva documentación y se definieron las medidas necesarias. $IQ1 = 4 = 100\%$ *IQ2 = ¿Las pruebas realizadas se tomaron en cuenta para las mediciones correspondientes de relevancia?* *R* = En la validación del prototipo se realizaron las pruebas para evaluar y medir la relevancia de los resultados mediante Índice MAP y Curva Precisión-recuerdo. Luego, la evaluación realizada con los estudiantes de fisioterapia permitió calcular la Precisión en los K primeros documentos, la comparación del índice MAP perteneciente a MSEC Web Search, Google y Yahoo!, además se realizaron las estadísticas Kappa y se compararon los resultados de

[43] El indicador utilizado será Radio de Precisión, el cual corresponde a una calificación del usuario del número de páginas bien calificadas sobre el número de páginas consultadas.

[44] Denominado también como Recall.

[45] Esta medida nos da una idea global del sistema a través de un conjunto de consultas. Para ello se calcula el promedio de las precisiones promedio para ese conjunto de consultas.

	concordancia de los jueces en cuanto a relevancia de documentos. $IQ2 = 4 = 100\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 3 = \frac{100}{1} = 100\%$
Medios de verificación	1. En el capítulo 5 de este documento se encuentran las medidas: Curva de precisión-recuerdo (sección 5.3) Índice MAP (sección 5.4 y 5.6) y estadísticas Kappa (sección 5.7). En el Anexo F se muestran los formatos para la realización de estas pruebas.
Estrategias, problemas y/o observaciones	Para cumplir con este objetivo, fue necesario revisar la bibliografía existente sobre las mediciones de sistemas de recuperación de información y así calcular e interpretar los resultados arrojados. Las pruebas fueron diseñadas para medir la usabilidad y relevancia de resultados obtenidos de acuerdo a las consultas realizadas por cada usuario. Además se realizaron para medir los acuerdos entre jueces (estudiantes de fisioterapia) sobre cada resultado específico de acuerdo a una consulta realizada.

Tabla 22. Cumplimiento del tercer objetivo específico.

Capítulo VII

7 CONCLUSIONES, RECOMENDACIONES Y TRABAJO FUTURO

Este capítulo describe inicialmente las principales conclusiones del trabajo realizado a las que se llegó durante su desarrollo, posteriormente presenta las recomendaciones, y finalmente propone los trabajos futuros.

7.1 Conclusiones

- Se propuso el Modelo Semántico de Expansión de Consultas (MSEC), el cual se centra en la combinación entre recursos léxicos, la similitud semántica sobre Ontologías de dominio y la realimentación que ofrece Perfil de Usuario, de tal manera que las búsquedas de un usuario estén enfocadas a sus intereses particulares de información incrementando la relevancia de los documentos que le son recuperados.

- 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.

- Con base en el modelo propuesto, se desarrolló un prototipo de meta-buscador Web denominado MSEC Web Search, el cual puede ser visto como una instanciación del modelo. Este prototipo utiliza recursos léxicos tales como un diccionario global y una Ontología de dominio en el área de la medicina, específicamente la rama de la Oncología[46] [120] para enriquecer semánticamente una consulta de usuario mediante el proceso de expansión.

- Dado que un meta-buscador Web se basa en la utilización de los servicios de búsqueda de varios buscadores Web, el prototipo MSEC Web Search se apoyó en dos de los motores de búsqueda más populares actualmente como los son Google y Yahoo!, de este modo se cuenta con la ventaja de disponer de los resultados más relevantes proporcionados por ambos buscadores, los cuales son obtenidos mediante una previa expansión de la consulta de usuario.

[46] La oncología es la especialidad médica que estudia los tumores benignos y malignos, pero con especial atención a los malignos, esto es, al cáncer.

- Los resultados experimentales muestran que el modelo propuesto establece una mejora importante respecto al enfoque tradicional en el que sólo se tiene en cuenta la coincidencia de palabras clave sobre los documentos indexados. MSEC proporciona algunas ventajas importantes tales como la recuperación de documentos relevantes que en el esquema tradicional de búsqueda eran omitidos; además de un mejor ranking de estos documentos en el conjunto de resultados que son mostrados al usuario.

- Los resultados experimentales de precisión de MSEC Web Search se aproximaron a los obtenidos con el buscador semántico GoPubMed, el cual maneja un enfoque mucho más complejo y especializado como la utilización de la ontología Gene Ontology y el vocabulario controlado MeSH además de estar basado en conocimiento.

- Mediante la investigación desarrollada durante el transcurso del proyecto, la creación del modelo, la implementación del prototipo y la validación del mismo, se realizó el artículo de investigación "Modelo Semántico de Expansión de Consultas para la Búsqueda Web - MSEC", el cual fue enviado a la "Revista Facultad de Ingeniería" de la Universidad de Antioquia y se encuentra en evaluación.

- La utilización de Ontologías en el proceso de expansión de consulta es de gran importancia para la RI, puesto que permite el enriquecimiento semántico de la consulta en torno a un interés particular de información, es decir, las búsquedas que realiza el usuario no se limitan a la simple coincidencia de palabras clave sobre documentos indexados, sino que además, al tener en cuenta el significado de los conceptos que representan a la consulta original, es posible adicionar otros que al estar relacionados semánticamente, posibiliten incrementar la precisión de los documentos que son recuperados.

- La integración de las diferentes conclusiones de las investigaciones relacionadas con la Expansión de Consulta y el Perfil de Usuario en las decisiones de cómo armar la estructura de la consulta expandida permitió obtener mejores resultados desde el principio de esta investigación.

7.2 Recomendaciones

Para proporcionar una expansión de consulta más efectiva, es necesario tener en cuenta el tipo de relaciones entre conceptos pertenecientes a una Ontología, o bien a un diccionario o cualquier otro recurso léxico, puesto que existen diferentes conceptos que si bien se encuentran relacionados semánticamente con los conceptos de la consulta original, no brindan el adecuado componente semántico que se ajuste a toda la necesidad de información del usuario, por el contrario se podría desviar la búsqueda de una manera muy drástica, afectando la calidad de los resultados que se recuperan.

7.3 Trabajo Futuro

El grupo de investigación espera evaluar el modelo propuesto en otro dominio particular del conocimiento, pues es ahí donde proyecta su potencial de recuperación, además de considerar ajustes al modelo con base en los resultados. Tales ajustes se pueden traducir en la elaboración de una estructura más compleja del PU, que permita detectar en mayor medida los intereses del usuario y además de esto, el modelo de expansión se puede combinar con la indexación semántica [137] y con el procesamiento del lenguaje natural [138] para obtener mayor efectividad en la RI.

Otro aspecto importante a tener en cuenta para un trabajo futuro es la realimentación que brinda el usuario por medio de sus búsquedas, esta podría ser implícita [139] sin que exista la necesidad de calificar documentos y de este modo hacer transparente todo el proceso al usuario facilitando su tarea de búsqueda.

8 REFERENCIAS

[1] R. Dhanapal, "An intelligent information retrieval agent," *Knowledge-Based Systems,* vol. 21, pp. 466-470, 2008.

[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," in *Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos.*, F. d. F. y. Letras, Ed. Mendoza, Argentina, 2005, pp. 35-46.

[3] R. Baeza-Yates and B. Ribeiro-Neto, "Modern information retrieval," in *Information Processing & Management*, 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: an International Journal,* vol. 26, pp. 755-776, 1990.

[5] K.-M. Kim, J.-H. Hong, and S.-B. Cho, "A semantic Bayesian network approach to retrieving information with intelligent conversational agents," *Information Processing & Management,* vol. 43, pp. 225-236, 2007.

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

[7] Y. Marcano and R. Talavera, "Gestión de la información a través de la Web Semántica: Iniciativas y dificultades," in *Revista Venezolana de Gerencia (RVG)*, 2006, p. 36.

[8] P. Mitra, N. Noy, and A. Jaiswal, "Ontology Mapping Discovery with Uncertainty," in *Fourth International Conference on the Semantic Web*, 2005, p. 15.

[9] D. G. Avello, "Web Cooperativa," Oviedo: Universidad de Oviedo, 2002, p. 67.

[10] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2007.

[11] C. Serrano and C. Hoyos, "Un modelo de investigación documental," S. Editora, Ed., 2000, p. 67.

[12] F. CACHEDA, "Introducción a los modelos clásicos de Recuperación de Información," in *Revista General de Información y Documentación* Coruña: Universidad A Coruña, 2008, pp. 365-374.

[13] C. Meadow, *Text Information retrieval Systems*. San Diego, 1993.

[14] D. Grossman and O. Frieder, "Information retrieval," in *algorithms and heuristics*, 1998.

[15] E. Greengrass, "Information Retrieval," A. Survey, Ed., 2000, p. http://www.cs.umbc.edu/cadip/readings/IR.report.120600.book.pdf.

[16] G. Chowdhury, *Introduction to modern information retrieval*. London, 1999.

[17] G. Navarro, *Como funciona la Web*. Santiago de Chile, 2008.

[18] K. D. R. Benavides, "Introduccion a la Recuperación de Información," Universidad de Costa Rica, 2010.

[19] D. M. C. Ramírez, "Recuperación y Organización de la Información", [en línea] http://modelos-recuperacion.50webs.com/. [Consulta: 04 de julio 2010].

[20] Lucene, "Lucene.Net", [en línea] http://Lucene.apache.org/lucene.net. [Consulta: 06 de Julio 2010].

[21] V. Ramón, "Sistemas de Recuperación de Información" [En línea]. Valladolid: Departamento de Ingeniería de Sistemas Telemáticos, Universidad. <http://www.mat.upm.es/~jmg/doct00RecupInfo.pdf> [Consulta: 06 de junio de 2010]

[22] S. Dominich, "A unified mathematical definition of classical information retrieval," *Journal of the American Society for Information Science,* vol. 51, pp. 614 - 624, 2000.

[23] J. Carmichael and J. Kummerfeld, "Consistent Modelling of Users, Devices and Sensors in a Ubiquitous Computing Environment," *User Modeling and User-Adapted Interaction,* vol. 15, pp. 197-234, 2005.

[24] A. orozco, J. Cárdenas, L. Florez, and A. Carrillo, "Modelo de Preferencia de Actividades para la Definición de un Perfil," in *IV congreso Colombiano de Computación* Bucaramanga,Colombia, 2009.

[25] L. Tamine and W. Bahsoun, "Définition d'un profil multidimensionnel de l'utilisateur," in *Actes de la Conférence francophone en Recherche d'Information et Applications(CORIA 2006)* Lyon France, 2006, pp. 225-236.

[26] P.-M. Chen and F.-C. Kuo, "An information retrieval system based on a user profile," *Journal of Systems and Software,* vol. 54, pp. 3-8, 2000.

[27] M. Vallace and G. Stamou, "Towards a Context Aware Mining of User Interests consumption of Multimedia Documents " in *IEEE International Conference on Multimedia (ICME)* Lausanne, Switzerland, 2002.

[28] J. Callan, "Learning while filtering documents," in *21st annual international ACM SIGIR conference on Research and development in information retrieval* Melbourne, Australia, 1998, pp. 224-231.

[29] K. D. Bollacker, S. Lawrence, and C. L. Giles, "Discovering relevant scientific literature on the web," *IEEE Intelligent Systems,* pp. 42-47, 2000.

[30] C. Wei, C. Huang, and H. Tan, "A Personalized Model for Ontology-driven User Profiles Mining " in *Intelligent Ubiquitous Computing and Education, 2009 International Symposium* Chengdu, 2009, pp. 484 - 487

[31] anonimo, "Adaptación y personalización de sitios web," Alicante: Universidad de Alicante, 2003.

[32] D. Oard and G. Marchionini, "A Conceptual Framework for Text Filtering," University of Maryland, Maryland, USA 1996.

[33] A. Micarelli and F. Sciarrone, "Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System," *User Modelling and User-Adapted Interaction,* vol. 14, pp. 159-200, 2004.

[34] R. Cöster, "The Architecture and Implementation of a System for Collaborative and Content-based Filtering," 2002.

[35] Zhengyu ZHU, Jingqiu XU, Xiang REN, Yunyan TIAN, and L. LI, "Query Expansion Based on a Personalized Web Search Model," in *Third International Conference on Semantics, Knowledge and Grid* Shan Xi 2007.

[36] Silvia Calegari and G. Pasi, "Personalized Ontology-based Query Expansion," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* Sydney, NSW 2008.

[37] M. L. Rodríguez, "Modelos de recuperación de información basados en Información Linguística Difusa y Algoritmos Evolutivos.Mejorando la Representacion de las Necesidades de Información " in *Departamento de Ciencias de la computacion e inteligencia artificial* Granada: Universidad de Granada, 2005.

[38] S. Ahu, M. Bamshad, and B. Robin, "Web search personalization with ontological user profiles," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* Lisbon, Portugal: ACM, 2007.

[39] Peng Chang and H. Ma, "Theme-based Query Expansion by Mining Log Data," in *Wireless Communications, Networking and Mobile Computing, 4th International Conference* 2008.

[40] L. Porwol, "SQE – Semantic Query Expansion as Search Process Booster " in *2010 Semantic Technology Conference* San Francisco, CA, 2009.

[41] Wikipedia-Gadget, "Gadget", [en línea] http://es.wikipedia.org/wiki/Gadget. [Consulta: 06 de julio 2010].

[42] K. Teknomo, "K-Mean Clustering Tutorials", [en línea] http://people.revoledu.com/kardi/tutorial/kMean/WhatIs.htm. [Consulta: 06 de Julio 2010].

[43] H. Cui, J. Wen, J. Nie, and W. Ma, "Query Expansion by Mining User LOGS," *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, pp. 829-839, 2003.

[44] E. N. Efthimiadis, "Query Expansion," in *Annual Review of Information Systems and Technology (ARIST)*, 1996, pp. 121-187.

[45] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback. ," *Journal of the American Society for Information Science,* pp. 288-297, 1990.

[46] G. Miller, "A lexical database for English," *Communication of the ACM,* pp. 39-41, 1995.

[47] Definición.de, "Definición de sinónimo", [en línea] http://definicion.de/sinonimo/. [Consulta: 07 de julio 2010].

[48] R. Pedraza, L. Codina, and C. Rovira, "Web semántica y ontologías en el procesamiento de la información documental," in *El profesional de la información*. vol. 16, 2007.

[49] Khosravi, Fariborz, Ghadimi, and Narges, "Trilingual cultural thesaurus (ASFA)," 3 ed Tehran, Iran: National Library of Iran, 2005.

[50] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *International Journal of Human Computer Studies,* pp. 907–928, 1995.

[51] I. Seher, "QUERY EXPANSION IN PERSONAL QUERIES," NSW, Australia.: University of Western Sydney, 2006.

[52] P. Appan and H. Sundaram, "Networked multimedia event exploration," in *Proceedings of the 12th annual ACM international conference on Multimedia*, A. Press, Ed., 2004, pp. 40-47.

[53] J. Wen, "Probabilistic model for contextual retrieval," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, A. Press, Ed., 2004, pp. 57-63.

[54] T. B. Lee, J. Hendeler, and O. Lassila, "The Semantic Web," in *Scientific American Magazine*, 2001.

[55] S. Cranefield, "UML and the Semantic Web," in *SWWS'01: The First Semantic Web Working Symposium* Sardinia, Italia, 2001.

[56] C. Anutariya, "Semantic Web Modeling and Programming with XDD," in *SWWS'01: The First Semantic Web Working Symposium* Sardinia, Italia, 2001.

[57] T. B. Passin, "Explorer's Guide to the Semantic Web," T. Taylor, Ed. Bruce Park Avenue, Greenwich: Manning Publications Co., 2004.

[58] L. Han and G. Chen, "A fuzzy clustering method of construction of ontology-based user profiles," *Advances in Engineering Software,* vol. 40, pp. 535-540, 2009.

[59] F. Rui-xue, Y. Xin, S. Ming, and X. Zhan-hong, "An architecture of knowledge management system based on agent and ontology," *www.sciencedirect.com/science/journal/10058885,* 2008.

[60] G. Solskinnsbakk and J. A. Gulla, "Combining ontological profiles with context in information retrieval," *www.elsevier.com/locate/datak,* p. 10, 2009.

[61] M. Nakatsujia, M. Yoshidab, and T. Ishida, "Detecting innovative topics based on user-interest ontology," *Web Semantics: Science, Services and Agents on theWorldWideWeb,* 2009.

[62] X. Jiang and A.-H. Tan, "Learning and inferencing in user ontology for personalized Semantic Web search," *Information Sciences,* 2009.

[63] Q. Guo and M. Zhang, "Question answering based on pervasive agent ontology and Semantic Web," *Knowledge-Based Systems,* 2009.
[64] S.-S. Weng and H.-L. Chang, "Using ontology network analysis for research document recommendation," *Expert Systems with Applications,* 2007.
[65] U. Visser, "Intelligent Information Integration for the Semantic Web," J. G. C. a. J. Siekmann, Ed. Berlin Heidelberg: Springer Science + Business Media, Inc., 2005.
[66] J. P. Pickett, "American Heritage," in *American Heritage Dictionary*, 3 ed, H. Mifflin, Ed., 1994, p. 960.
[67] WordNet, "WordNet Search - 3.0", [en línea] http://wordnetweb.princeton.edu/perl/webwn. [Consulta: 01 de Junio de 2010].
[68] FOLDOC, "Free On-Line Dictionary Of Computing , annotation", [en línea] http://foldoc.org/annotation. [Consulta: 01 de Junio de 2010].
[69] W3Consortium, "The World Wide Web Consortium (W3C)", [en línea] http://www.w3.org/. [Consulta: 09 de Junio de 2010].
[70] W3CAnnoteaProject, "Annotea Project", [en línea] http://www.w3.org/2001/Annotea/. [Consulta: 09 de Junio de 2010].
[71] W3CAmayaBrowser, "W3C's Editor/Browser", [en línea] http://www.w3.org/Amaya/. [Consulta: 09 de Junio de 2010].
[72] J. Breis, "Un Entorno de Integración de Ontologías para el Desarrollo de Sistemas de Gestión de Conocimiento," in *Departamento de Ingeniería de la Información y las Comunicaciones* Murcia, España: Universidad de Murcia, 2003.
[73] S. Grimm, P. Hitzler, and A. Abecker, "Knowledge Representation and Ontologies: Logic, Ontologies and SemanticWeb Languages," in *Semantic Web Services: Concepts, Technologies, and Applications*, B. Springer, Ed., 2007, pp. 51–105.
[74] D. Oberle, "Ontologies," in *Semantic Management of Middleware*, Springer, Ed., 2006, pp. 33–53.
[75] A. Gómez-Pérez and D. Manzano-Macho, "An overview of methods and tools for ontology learning from text," *The knowledge engineering review,* vol. 19, pp. 187-212, 2005.
[76] S. B. Suárez, "Biblioteca Semántica de WEBQUEST," in *Departamento de Informática* Valladolid: Universidad de Valladolid, 2004.
[77] A. Vckovski, *Interoperable and Distributed Processing in GIS*. Londres, Inglaterra, 1998.
[78] M. F. Worboys and S. M. Deen, *Semantic heterogeneity in distributed geographical databases. SIGMOID Record*, 1991.
[79] I. Shepherd, "Information integration in gis," in *Geographical Information Systems: Principles and applications*. vol. 1, M. F. G. D. J. Maguire, and D. W. Rhind, editors, Ed. Longman, London, UK, 1991, pp. 337–360.
[80] M. Grüninger and M. Uschold, "Ontologies and semantic integration," *Government report on the state of the art and future predictions for agent technology,* 2002.
[81] H. A. F. Fernández, "Construcción de ontologías OWL," in *VINCULOS 7*, 2008, p. 33.
[82] D. Gasevic, D. Djuric, V. Devedic, and B. Selic, "Model Driven Architecture and Ontology Development," Springer, 2006.
[83] Wikipedia, "Resource Description Framework," [en línea] http://es.wikipedia.org/wiki/Resource_Description_Framework. [Consulta: 10 de Diciembre de 2009].
[84] D. Brickley and R. Guha, "RDF vocabulary description language 1.0: RDF Schema. W3C Recommendation," [en línea] http://www.w3.org/TR/rdf-schema/]. [Consulta: 01 de Junio de 2010].

[85] S. Bechhofer, F. V. Harmelen, and J. Hendler, "OWL Web ontology language " *Recomendación W3C,* 2004.

[86] G. P. C. Ríos, "Propuesta y construcción de una ontología para lenguajes de modelado gráfico," 2008.

[87] M. Hepp, J. Cardoso, and M. Lytras, "Ontology Management. Semantic Web, Semantic Web Services, and Business Applications," in *SEMANTIC WEB AND BEYOND. Computing for Human Experience*, R. Jain and A. Sheth, Eds. New York, NY: Springer, 2008.

[88] G. B. Vidal, E. P. Hernández, and C. C. Lozada, "Meta web search model based on ontologies, taxonomies and user feedback," 2009.

[89] J. Mustafa, S. K. han, and K. Latif, "Ontology based semantic information retrieval," in *4th International IEEE Conference*, 2008, pp. 22-14-22-19.

[90] D. Vallet, M. Fernandez, and P. Castells, "An Ontology-Based Information Retrieval Model," 2005.

[91] M. A. Aufaure, R. Soussi, and H. Baazaoui, "SIRO: On-line semantic information retrieval using ontologies," in *2nd International Conference*, 2007, pp. 321-326.

[92] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1* Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995.

[93] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*: Morgan Kaufmann Publishers Inc., 1998.

[94] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM,* vol. 8, pp. 627-633, 1965.

[95] G. Miller and W. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes,* vol. 6, pp. 1-28, 1991.

[96] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *Systems, Man and Cybernetics, IEEE Transactions on* vol. 19, pp. 17 - 30, 1989

[97] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. on International Conference on Research in Computational Linguistics* Taiwan, 1997, pp. 19--33.

[98] G. Hirst and D. St-Onge, "Lexical chains as representation of context for the detection and correction malapropisms," in *Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press*, 1998, pp. 305–332.

[99] DMOZ, "DMOZ - Open Directory Project," [en línea] http://www.dmoz.org/. [Consulta: 01 de Diciembre de 2010].

[100] BiKE-Laboratory, ""MeSH Ontology in OWL format"," [en línea] http://bike.snu.ac.kr/?q=node/207. [Consulta: 30 de Noviembre de 2010].

[101] L. Mazuel and N. Sabouret, "Semantic Relatedness Measure Using Object Properties in an Ontology," in *Proceedings of the 7th International Conference on The Semantic Web* Karlsruhe, Germany: Springer-Verlag, 2008, pp. 681-694.

[102] A. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, and A. Vespignani, "Algorithmic Computation and Approximation of Semantic Similarity," *World Wide Web,* vol. 9, pp. 431-456(26), 2006.

[103] G. Lv, C. Zheng, and L. Zhang, "Text Information Retrieval Based on Concept Semantic Similarity," in *Semantics, Knowledge and Grid, 2009. SKG 2009. Fifth International Conference on* Zhuhai, China, 2009, pp. 356 - 360

[104] C. Perez, "Modelado de Sistemas Dinámicos. Aplicaciones," 1 ed: Editorial Club Universitario, 2005, p. 15.

[105] Departamento_de_Ingeniería_de_Sistemas_y_Automática, "Introducción al Modelado de Sistemas", [en línea] http://www.isa.uma.es/C17/Presentaciones%20de%20Clase%20(ppt)/Document%20Library/INTRODUCION%20AL%20MODELADO%20DE%20SISTEMAS.pdf. [Consulta: 17 de Mayo de 2011].

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

[107] M. Barite, "Clasificación, Indización, Terminología - Letra T," in *KO Dictionary - Diccionario de Organización y Representación del Conocimiento.*, 2000.

[108] M. Baziz, M. Boughanem, and N. Aussenac-Gilles, "Evaluating a Conceptual Indexing Method by Utilizing WordNet," *Lecture Notes in Computer Science,* vol. 40, pp. 238 - 246, 2006.

[109] R. Navigli and P. Velardi, "An Analysis of Ontology-based Query Expansion Strategies," in *14th European conference on machine learning (ECML 2003)* Dubrovnik, Croatia, 2003.

[110] A. Bechara, M. L. C, Machado, and V. Braganholo, "Applying Biomedical Ontologies on Semantic Query Expansion," *Nature Precedings* 2009.

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

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

[113] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* Las Cruces, New Mexico: Association for Computational Linguistics, 1994.

[114] N. Lahkar and S. K. Deka, "Impact of Query Operators on Web Search Engine Results : An Evaluative Study," in *2nd Convention PLANNER - 2004* Manipur Uni., Imphal, 2004.

[115] C. Sherman, "The Future Revisited: What's New with Web Search", [en línea] http://www.onlineinc.com/onlinemag/OL2000/sherman5.html. [Consulta: 18 de Octubre 2010].

[116] J. A. Aslam and M. Montague, "Models for Metasearch," in *Proceedings of 24th Internationsl ACM SIGIR* New Orleans, Louisiana, USA, 2001, pp. 276-283.

[117] H. Wang, J. Qin, and H. Shao, "Expansion Model of Semantic Query Based on Ontology," in *Web Mining and Web-based Application, 2009. WMWA '09. Second Pacific-Asia Conference on* Wuhan, China 2009.

[118] G. Salton, *The SMART Retrieval System---Experiments in Automatic Document Processing*: Prentice-Hall, Inc., 1971.

[119] NCI, "National Cancer Institute. U.S. National Institutes of Health", [en línea] http://www.cancer.gov. [Consulta: 26 de Abril de 2011].

[120] Wikipedia, "Caso de uso", [en línea] http://es.wikipedia.org/wiki/Caso_de_uso. [Consulta: 26 de Abril de 2011].

[121] C. d. l. Torre, M. Ramos, and J. Calvarro, "Guía de Arquitectura N-Capas orientada al dominio con .Net," in *K. Consulting, Editor* Madrid: Microsoft Ibérica.

[122] R. J. Vargas and J. P. Maltés, "Programación en Capas," in *Universidad de Costa Rica*, 2007, p. 5.

[123] M. A. R. Barragán, "Un Método para Recuperación de Información en Documentos Orales basado en Codificación Fonética," Tonantzintla,Puebla: Instituto Nacional de Astrofísica,Optica y Electronica, 2008.
[124] J. J. Yepes, "Ontology Refinement for Improved Information Retrieval in the Biomedical Domain." vol. PhD Thesis Castellón: Universitat Jaume, 2009.
[125] B. Croft, D. Metzler, and T. Strohman, "Search Engines: Information Retrieval in Practice," 1 ed USA: Addison-Wesley, 2009, p. 552.
[126] MEDLARS, "MEDical Literature Analysis and Retrieval System", [en línea] http://www.uninet.edu/do/MEDLARS.html. [Consulta: 26 de Abril de 2011].
[127] GoPubMed, "GoPubMed, searching is now sorted!", [en línea] http://www.gopubmed.com/. [Consulta: 23 de Junio de 2011].
[128] R. Gómez, "La evaluación de la recuperación de la información," *Hipertext.net - Anuario académico sobre documentación digital y comunicación interactiva, Departamento de Comunicación - Grupo de Investigación DIGIDOC,* 2003.
[129] F. Martínez, Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet, in Información y Documentación. 2002, Universidad de Murcia: Murcia, España. p. 283.
[130] F. Cacheda, V. Formoso, and V. Carneiro, Performance Analysis of Distributed Web Information Retrieval Systems. Latin America Transactions, IEEE (Revista IEEE America Latina), 2007. 5(6): p. 479-485.
[131] J. Landis and G. Koch, "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, 1977, pp. 159–174.
[132] MEDLINE, "MEDLINE/PubMed Resources Guide", [en línea] http://www.nlm.nih.gov/bsd/pmresources.html. [Consulta: 27 de Junio de 2011].
[133] Gene_Ontology, "The Gene Ontology", [en línea] http://www.geneontology.org/. [Consulta: 27 de Junio de 2011].
[134] Transinsight, "German Industry Prize 2010", [en línea] http://test.transinsight.com/company. [Consulta: 27 de Junio de 2011].
[135] Dexigner, "Red Dot Award", [en línea] http://www.dexigner.com/news/22215. [Consulta: 27 de Junio de 2011].
[136] RefNavigator, "RefNavigator", [en línea] http://www.refnavigator.com/. [Consulta: 27 de Junio de 2011].
[137] M. Suárez and K. Salinas, "An Approach to Semantic Indexing and Information Retrieval," in *Revista Facultad de Ingeniería Universidad de Antioquia*, 2009, pp. 174-187.
[138] P. Jackson and F. Schilder, "Natural Language Processing: Overview," *Encyclopedia of Language & Linguistics, Elsevier,* pp. 503 – 518, 2006.
[139] D. Hardtke, M. Wertheim, and M. Cramer, "Demonstration of Improved Search Result Relevancy Using Real-Time Implicit Relevance Feedback," in *Understanding the user - Logging and interpreting user interactions in information search and retrieval* SIGIR 2009, 2009.
[140] Hakenberg, J., Royer, L., Plake, C., Strobelt, H., and Schroeder, M. Me and my friends: gene mention normalization with background knowledge. In *Proceedings 2nd BioCreAtIvE Challenge Evaluation Workshop*, 2007.