PROCEDIMIENTO PARA LA CREACIÓN DE ÍNDICES SEMÁNTICOS BASADOS EN ONTOLOGIAS DE DOMINIO.

DIGNORY JIMENA PEREZ URBANO
DIANA MARIBEL PEZO ARTEAGA

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Popayán
2010

PROCEDIMIENTO PARA LA CREACIÓN DE ÍNDICES SEMÁNTICOS BASADOS EN ONTOLOGIAS DE DOMINIO.

DIGNORY JIMENA PEREZ URBANO
DIANA MARIBEL PEZO ARTEAGA

Monografía de trabajo de grado

DIRECTOR: Magíster. Miguel Ángel Niño
ASESOR: PhD(C) Carlos Cobos Lozada

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Popayán
2010

NOTA DE ACEPTACION

_____
_____
_____
_____
_____

_____
Firma de Jurado:

_____
Firma de Jurado:

Popayán, Cauca 10 de diciembre de 2010

AGRADECIMIENTOS

Inicialmente doy gracias a Dios y mis padres Efraín Pérez y Sonia Urbano, por el amor y apoyo incondicional que me dieron a lo largo de mi vida. A mi hermano Nilton Andrés Pérez por ser el ejemplo a seguir y a mi esposo José Andrés Muñoz por ser mi gran apoyo, mi gran amor, mi compañía y por ser la persona que me ha enseñado que todo es posible, que con esmero y dedicación todo se puede alcanzar.
Gracias a Diana Pezo, mi compañera de tesis por su colaboración y disposición en este proyecto de grado.
Igualmente a la Ingeniero Miguel Ángel Niño por sus enseñanzas, asesorías y dirección en la realización de este proyecto.
Gracias al Ingeniero Carlos Cobos por estar siempre dispuesto a resolver nuestras dudas.
A la institución Campestre Americano y la Institución Educativa Alejandro de Humboldt por brindarnos su ayuda en la realización de pruebas de la aplicación.
A mis compañeros de carrera por su compañía y apoyo en cada momento durante la elaboración de este proyecto.
Y cada una de las personas que de alguna forma u otra colaboraron o participaron en el desarrollo de este proyecto, muchas gracias a todos.

Dignory Jimena Pérez Urbano

Doy gracias a Dios, a mi madre Amilvia D. Arteaga por su amor y apoyo constante en mi vida, a mi padre Manuel A. Pezo, que en paz descanse y que aunque no esté físicamente conmigo, siempre ha estado en mi mente y sus enseñanzas han sido mi fuente de energía. A mis tíos y primos por sus palabras de ánimo y alegría en los momentos necesarios.
Agradezco a mi compañera de tesis, Jimena Pérez, por su colaboración y disposición para este trabajo de grado.
Gracias al Ingeniero Miguel Ángel Niño por la dirección de este proyecto, sus asesorías y colaboración en la realización del mismo. Igualmente al ingeniero Carlos Cobos por su disposición y ayuda en los momentos necesarios.
Doy gracias a Javier Gaviria por su apoyo incondicional, fortaleza, y las palabras adecuadas en el momento justo.
A mis compañeros de universidad por su apoyo en los momentos compartidos durante la realización de la investigación.
Al Colegio Campestre Americano y a la Institución educativa Alejandro de Humboldt por su colaboración en la evaluación del presente proyecto.
Y muchas gracias a todas aquellas personas que colaboraron o participaron, de una u otra forma, en la realización de este proyecto.

Diana Maribel Pezo Arteaga

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

INDICES FIGURAS

INTRODUCCIÓN

Desde hace una década, varios proyectos [1-3], entre otros, han propuesto diversas soluciones que mejoran la relevancia [4, 5] en las búsquedas Web, desarrollando o mejorando las técnicas actuales de recuperación de información. Una de las mejoras es la utilización de nuevas metodologías enmarcadas bajo el nombre de búsqueda semántica. Cada uno de estos proyectos utiliza diferentes técnicas con las cuales han obtenido resultados muy favorables, en especial las técnicas que aplican las ontologías y la construcción de índices semánticos; al parecer, éstos últimos han empezado a utilizarse en variados estudios [6-9], en dónde la semántica de los conceptos es el principal problema a resolver. Así, cada estudio plantea su propia forma de implementar índices semánticos, aplicando diferentes criterios, procesos y pasos, que dependen de los autores y sus objetivos. A pesar de la importancia que tienen los índices semánticos en los sistemas de recuperación de la información, los investigadores deben recurrir a un proceso largo de sensibilización y entendimiento en su construcción y uso, dificultando así las nuevas investigaciones en el área en particular.

El presente proyecto se basa en el área de la creación de índices semánticos, con el fin de crear una base teórica que permita construir un procedimiento general para generar índices semánticos en el entorno de la recuperación de la información Web. En el análisis del estado del arte realizado, no se encontró un estudio que plantee un procedimiento general para construir índices semánticos basados en ontologías de dominio, que oriente a los investigadores al momento de su construcción.

Por lo anterior, se analizaron diversos proyectos [3, 8, 10-15] con la perspectiva de abstraer un procedimiento para generar índices semánticos basados en ontologías de dominio, el cual sirve como soporte para el desarrollo de aplicaciones que tengan por objetivo mejorar la relevancia de los resultados obtenidos en la búsqueda Web.

El presente documento se encuentra organizado por capítulos de la siguiente manera:

En el primer capítulo se presentan las bases conceptuales que son importantes para el desarrollo del presente trabajo, así como el análisis del estado del arte.

En el segundo capítulo se realiza la creación del procedimiento teniendo en cuenta cada uno de los elementos y sus relaciones para la construcción, y se definen los pasos y actividades del mismo. A su vez, se describe la instanciación de un índice semántico basada en el procedimiento planteado. Se realiza el análisis de pruebas que se ejecutan para evaluar el procedimiento definido.

En el capítulo tercero se describe la implementación del prototipo de indexación semántica, basado en una metodología de construcción de software.

En el cuarto capítulo se presenta la evaluación del índice, tomando estadísticas y medidas de relevancia con los resultados obtenidos.

En el capitulo seis se presenta el análisis del cumplimiento de los objetivos.

Finalmente, en el capitulo siete, se presentan las conclusiones y recomendaciones del presente trabajo, así como los trabajos futuros.

GLOSARIO

Frames: Los frames (marcos) son una manera de dividir una página web en varios espacios independientes, así, en cada uno se coloca una página distinta que se codifica en un archivo HTML distinto[1]. Una de las ventajas es navegar por los contenidos del sitio web con la barra de navegación visible y sin recargarse en cada página que se visita.

OWL: acrónimo de los términos en ingles: Ontology Web Language, un lenguaje de marcado para publicar y compartir datos usando ontologías en la WWW. OWL tiene como objetivo facilitar un modelo de marcado construido sobre RDF y codificado en XML [16].

RDF Shema: es una extensión semántica de RDF. Un lenguaje primitivo de ontologías que proporciona los elementos básicos para la descripción de vocabularios, permitiendo definir los recursos como instancias de clases. Demás se pueden definir las clases en forma jerárquica [17].

XML Shema: es un lenguaje de esquema utilizado para describir la estructura, el contenido y semántica de los documentos XML de una forma muy precisa [18]

DAML: DARPA Agent Markup Language. Su objetivo es presentar un lenguaje y herramientas para facilitar el los conceptos en la Web Semántica[2]. La versión más reciente es DAML OIL, la cual proporciona un gran conjunto de de constructores para crear ontologías y marcas de información legibles por maquina y más comprensibles.

RDF (Resource Description FrameWork): El Framework para la Descripción de Recursos es un modelo estándar para el intercambio de datos en la Web. Facilita la fusión de los datos incluso si los datos subyacentes son diferentes [19].

OIL (Ontology Interchange Language): presentación basada en la Web y la capa de interferencia para ontologías, que combina las primitivas ampliamente utilizadas para la modelación de los lenguajes basados en imágenes con la

[1] Extraído de http://www.desarrolloweb.com/articulos/791.php.

[2] Más información en http://www.daml.org/about.html

semántica formal y los servicios de motivación aportada para las lógicas de descripción [20].

Representatividad: Especificaciones (léxicas, sintácticas, semánticas) representadas con etiquetas, que permiten identificar el contenido en la Web entre un gran número de artículos disponibles[3]. Al realizar un estudio de representatividad se extraen las características que permiten definir los conceptos, términos o elementos utilizados en determinado ámbito. Esto ayuda a definir la importancia de un término o concepto de acuerdo a los requerimientos de información.

Frase nominal: Es un grupo de palabras organizadas alrededor de un sustantivo (núcleo). La frase nominal puede consistir en un sustantivo solo, o con adjetivos de varios tipos que lo modifican y también puede ser encabezada por un pronombre [21]. Los elementos que acompañan al sustantivo nuclear se llaman adyacentes y cumplen una función atributiva respecto al núcleo [22].

Tokenizador: Realiza la conversión de una secuencia de caracteres en una secuencia de palabras candidatas a ser tomadas para el índice de un sistema de recuperación de información. Identifica las palabras que contienen los documentos [23]. Remoción de caracteres especiales como "/ \ - : ?¿;)(&#", entre otros.

Lematizador: Realiza el análisis morfológico de cada token o palabra, con lo cual, se identifica la raíz, la categoría gramatical y la flexión o derivación que la produce [24], por ejemplo, derivando las palabras en plural a su raíz en singular

Desambiguación: La desambiguación del sentido de las palabras, trata los fenómenos lingüísticos de diversa índole de forma automatizada. Elimina la ambigüedad en las palabras, que surge cuando una estructura gramatical puede ser interpretada de varias maneras y por tanto, puede confundir en el sentido de la oración [25].

[3] Extraído de http://recuperaciondeinformacion.lacoctelera.net/categoria/indexaci-n

1 MARCO TEORICO

En este capítulo se presenta una visión general de los puntos más importantes para la generación de índices semánticos. Para esta fase se utiliza el modelo de investigación documental planteada por Hoyos y Serrano [26]. Siguiendo esta metodología se definen los núcleos temáticos y se construyen las fichas descriptivas relacionadas con cada uno y consultadas en las principales bases de datos (Science Direct, ACM, IEEE, entre otras), las cuales se describen en el Anexo A. Posteriormente se realiza un estudio y análisis de las investigaciones encontradas para la realización de un articulo con el estado del arte actual (Anexo B).

En el presente documento, inicialmente se exponen las bases conceptuales sobre los índices semánticos y los diferentes algoritmos y proyectos que se han implementado. Se analizará cada uno de los proyectos realizados con índices semánticos y ontologías para el mejoramiento de la relevancia en la recuperación de información. Como conclusión de las investigaciones analizadas, se presenta una comparación de los pasos y procedimientos realizados por cada uno de estos para la creación de índices semánticos y posteriormente, se analizaran cada una de las herramientas para la construcción de procedimientos.

1.1 ONTOLOGIAS

Una de las principales herramientas a tener en cuenta en nuestro proyecto son las ontologías, las cuales ayudan en la extracción de la semántica necesaria para las búsquedas en la Web.

1.1.1 DEFINICIONES

El termino ontología tiene diferentes definiciones; desde un punto de vista filosófico la ontología es una antigua disciplina, que se define como un esquema especifico de categorías que reflejan una visión determinada del mundo [27], más específicamente es una rama de la metafísica relacionada con la naturaleza y las relaciones del ser, pero desde un punto de vista informático podemos encontrar diversas definiciones acerca de ontología, donde cada una de ellas refleja la forma en cómo son usadas en un área particular.

De acuerdo a lo anterior, una ontología la podemos definir desde el punto de vista informático como un conjunto de conceptos donde cada uno es representado por una etiqueta, un conjunto de sinónimos de este término, y un conjunto de relaciones que conectan estos conceptos por la relación especifica y la relación de composición [1].

Las ontologías juegan un papel importante en la Recuperación de Información (RI), ya que es una de las técnicas más útiles en la resolución de los problemas semánticos en los términos de la consulta.

Para el tema específico que abarca el presente documento, adoptaremos la siguiente definición de ontología: una ontología proporciona un vocabulario común de un dominio específico y se define más o menos las entidades, clases, propiedades, predicados, funciones, sentido de los términos y algunas de sus relaciones [28].

En el Anexo C se encuentra una información más detallada sobre Ontologías, la cual servirá al momento de la implementación del procedimiento.

1.1.2 PROYECTOS QUE CONSTRUYEN O USAN ONTOLOGÍAS

El uso de ontologías es una de las técnicas más utilizadas en la solución a los problemas de recuperación de información en la Web que requieren un manejo semántico de la información, en el entorno podemos encontrar diferentes proyectos que hacen uso de las mismas. Estos proyectos hacen uso de técnicas de la Web semántica como son las ontologías, motores de búsqueda y repositorios de Ontologías como WordNet.

Se debe tener en cuenta los proyectos como SIRO [10], donde se propone un sistema de recuperación basado en ontologías, en el cual se integran los resultados de los motores tradicionales y los motores de búsqueda de texto basado en ontologías. El propósito del proyecto es usar dos tipos de ontologías: ontologías de dominio, ontología de servicio y WordNet. Este proyecto esta aplicado para un dominio específico (turismo), que permite mejorar la precisión y la relevancia de los documentos devueltos al usuario en sus búsquedas. Además, este trabajo permite la filtración de documentos para mejorar la construcción de ontologías basadas en técnicas de aprendizaje.

Con el fin de mejorar la precisión de las búsquedas mediante el análisis semántico, WI OntoSearch [11] propone conceptos como el vector de peso y el algoritmo de emparejamiento (CWVMA). El algoritmo es el encargado de analizar las palabras de entrada y las palabras claves de los documentos obtenidos, basándose en un conjunto de conceptos. También se determinan las coincidencias encontradas según la influencia de los conceptos establecidos en la ontología semántica y luego se crea un modelo vectorial[4]. Este algoritmo obtendrá

[4] Se conoce como **modelo de espacio vectorial** a un modelo algebraico utilizado para filtrado, recuperación, indexado y cálculo de relevancia de información. Representa documentos en lenguaje natural de una manera formal mediante el uso de vectores (de

finalmente la medida de similitud entre las palabras de entrada y los resultados preliminares. Además, en este proyecto se diseña y se desarrolla un motor de búsqueda de ontologías basado en el algoritmo WI OntoSearch prototype system, el cual puede buscar cerca de 4 millones de páginas Web de Google. Muchos de los resultados obtenidos demuestran que el algoritmo mejora en gran medida la precisión de las búsquedas a través de la ontología.

En An Ontology-Based Information Retrieval Model [12] se propone un modelo para la explotación de la ontología basada en KBs (Knowledge Base) para mejorar la búsqueda en los repositorios de los documentos de gran tamaño. Su enfoque incluye un régimen basado en la ontología para la anotación semiautomática de los documentos y de los sistemas de recuperación. El modelo de recuperación se basa en una adaptación del clásico modelo de espacio vectorial, incluyendo un algoritmo de ponderación de anotación y un algoritmo de clasificación. La búsqueda semántica se combina con la palabra clave de búsqueda con respecto a una función definida por ellos en el archivo KB. Los experimentos realizados muestran una clara mejoría en las búsquedas con respecto a las realizadas con palabras claves.

En los tres proyectos anteriormente mencionados se hace uso de técnicas que sirvieron de ayuda para su desarrollo. Una de las principales fue el uso de ontologías, las cuales en su mayoría son creadas para un dominio especifico, cada domino de las ontologías es caracterizado por una lista de servicios, actividades y tareas. También, se hace uso del modelo vectorial, el cual, es adaptado para sustituir los términos por conceptos, y se tiene en cuenta el peso de los conceptos en vez del peso de los términos.

1.1.3 ONTOLOGIAS DISPONIBLES EN LA WEB

En la actual red podemos encontrar ontologías en diferentes dominios médicos, científicos, informáticos, biomédica, educativos y demás. Cada una de estas nos brindan ventajas a la hora de trabajar en nuestros proyectos. En el caso de nuestro proyecto necesitamos una ontología educativa que se ajuste a las necesidades del proyecto. Ver Anexo D para más información.

1.2 RECUPERACIÓN DE INFORMACIÓN

La recuperación de información comprende el área de la extracción información de los documentos, por lo general de una naturaleza no estructura (como en los

identificadores, por ejemplo términos de búsqueda) en un espacio lineal multidimensional.

textos), para satisfacer las necesidades de los usuarios [29]. La representación, almacenamiento organización y acceso, son algunas de esas necesidades de información para las cuales existen grandes estudios y formas de llevar a cabo una recuperación ágil y eficiente, por ejemplo el uso de modelos (booleano, vector, probabilístico, indexación semántica latente, entre otros [30]), algoritmos y áreas como la probabilística y la lógica.

Para recuperar la mayor cantidad de información, en el menor tiempo y con la mayor eficiencia para el usuario, se han desarrollado los sistemas de recuperación de información (SRI), los cuales son un tipo de sistemas de información, que trabajan con grandes bases de documentos y procesan las consultas de los usuarios, permitiendo acceder a la información relevante en un tiempo relativamente mínimo. Un sistema de recuperación de Información se encarga de encontrar la semejanza de términos de los documentos disponibles en la Web, de acuerdo con la consulta realizada por el usuario para listarlos por orden de relevancia [31], esto implica filtrar los contenidos que pueden ser relevantes y los que no lo son en cada consulta (sentencia que expresa la necesidad de información).

En la Figura 1. Proceso de Recuperación de Información se observa el proceso de RI, donde las colecciones de documentos y las herramientas juegan un papel muy importante en los sistemas de recuperación de información (SRI). Los usuarios realizan la consulta que desean (necesidad de información) y el resultado de esto es un conjunto de documentos relevantes y otro de los no relevantes. Los documentos relevantes son los que se presentan al usuario.

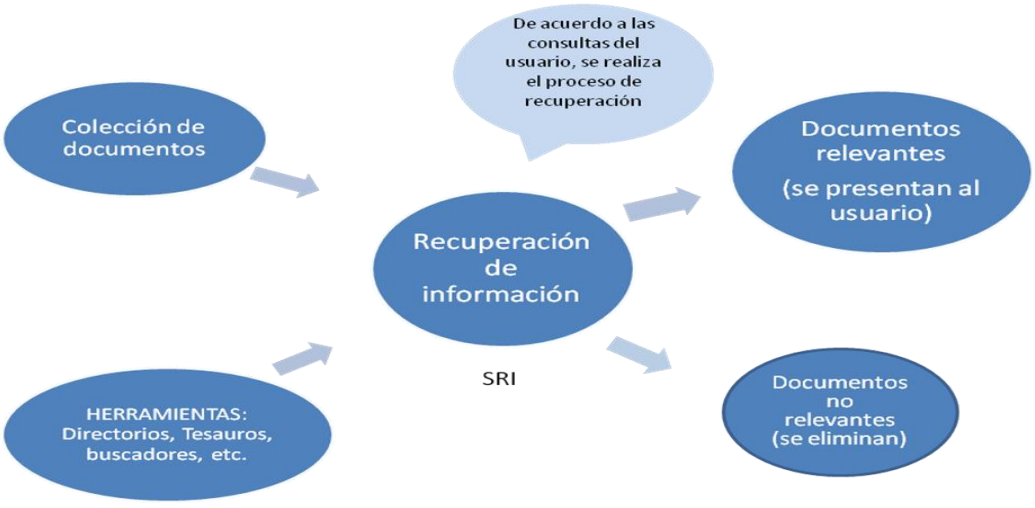

Figura 1. Proceso de Recuperación de Información

Mayor información sobre recuperación de información se encuentra en el Anexo C.

1.3 ÍNDICES SEMÁNTICOS

1.3.1 DEFINICIONES

El término indexar significa registrar ordenadamente datos e informaciones para elaborar su índice (según el diccionario de la Real Academia Española) y así referenciarlo de manera más rápida y eficiente. Se realizan varias técnicas para llevar a cabo la indexación de los documentos en la Web; la más simple es la indexación automática basada en el número de veces que se encuentra una palabra en un documento [32] y de esta manera, determinar la relevancia de ese documento para las búsquedas que contengan la palabra específica.

La indexación semántica va mas allá de buscar la ocurrencia de una palabra en los documentos, se enfoca también en asociar los conceptos con los términos o palabras en las páginas Web. Con ello se busca encontrar patrones en los datos no estructurados (documentos sin descriptores, como palabras clave o etiquetas especiales) [33] y usar los patrones de búsqueda en una mejor clasificación de los datos y precisión en la recuperación de información.

En otras palabras, el uso de índices semánticos en la búsqueda web, significa que los objetos son indexados no solo por los términos empleados, sino también por los conceptos que contienen para representarlos.

1.3.2 CARACTERÍSTICAS DE LOS ÍNDICES SEMÁNTICOS

Según el enfoque dado por Suarez B. Marco [32] , un índice semántico se caracteriza por:

- Un índice semántico es multidimensional, ya que cualquier combinación de propiedades son moldeados en un concepto que definen como el número total de términos que aparecen en un documento, el cual, puede servir como un elemento de indexación.
- Los elementos de indexación son valores de atributos que pueden estar basados en complejas descripciones de objetos relacionados, como un concepto estructurado.
- Un índice semántico es altamente adaptable a las necesidades de cada proyecto. Los conceptos de indexación pueden ser añadidos o eliminados

como se desee, lo cual los hace muy densos y precisos con respecto al interés de un grupo de personas.
- Dado que el índice es en realidad un conjunto de descripciones parciales de los objetos indexados, mucha información se puede extraer del índice solo, sin tener acceso a las descripciones individuales de todo.

Al realizar una indexación semántica, se utiliza la información del concepto que está dentro de los objetos indexados para mejorar la relevancia en la recuperación de información. Así, por ejemplo, la consulta Paris Hilton se asocia con una mujer, en vez de relacionarse con una ciudad y un hotel por separado, de esta manera se brinda al usuario un resultado satisfactorio de las consultas que realiza.

Otro enfoque, es utilizar la representación jerárquica derivada de las ontologías para calcular la distancia entre conceptos o similitud semántica entre las palabras que deben compararse[32].

1.3.2.1 Similitud semántica

La similitud semántica se refiere a la cercanía entre los conceptos en una estructura jerárquica como ontologías, tesauros o bases de datos léxicas. En las investigaciones estudiadas, la similitud es medida a partir de las relaciones o caminos más cortos entre los conceptos de una jerarquía. Por lo general, las funciones de similitud, varían de acuerdo a la estructura jerárquica que se tenga, los recursos con los que se cuente y la decisión propia de investigadores y equipos de desarrollo.

Existen varias medidas de similitud para calcular las relaciones más cercanas o caminos más cortos entre conceptos, varias de ellas, desarrolladas en el marco de la Inteligencia Artificial. A continuación se presentan las investigaciones con las medidas más utilizadas en la literatura encontrada.

Similitud semántica en una taxonomía [34]

Resnik es uno de los pioneros de las funciones de similitud y la primera medida de similitud conocida, expresa la relación de común información que guarda la similitud entre dos conceptos. Formalmente, se define:

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} (-\log p(c))$$

Ecuación 1. Similitud semántica de Resnik.

Donde S(c1, c2) representa el conjunto de conceptos de los cuales, c1 y c2 descienden, mientras que p(c) es la probabilidad del concepto c, medida por la frecuencia de aparición de los términos de ese concepto en los textos.

En esta investigación la medida de similitud semántica toma en cuenta una taxonomía con relaciones “IS-A” (es un), la cual se basa en la información común que se comparte en una taxonomía de este tipo mediante los conceptos más específicos que derivan. Esto significa que si el camino mínimo de la relación “IS-A” entre dos nodos es larga, se debería ir mas alto en la taxonomía, a conceptos más abstractos con el fin de encontrar una mínima cota superior [34].

El cálculo que define dicha medida es el contenido de la información del padre común más cercano (*ccp*) de dos conceptos *c1* y *c2*:

$$sim_{res}(c_1, c_2) = IC(ccp(c_1, c_2))$$

Ecuación 2. Similitud semántica en una taxonomía

Donde IC es el contenido de información del concepto c.

Medida de similitud tomando las propiedades de los objetos en una Ontología [35]

La medida de similitud permite tomar en cuenta las relaciones funcionales entre conceptos, se enfoca en los puntos en común en las relaciones semánticas. En esta investigación se realiza un cálculo del grado de relación entre dos conceptos no jerárquicos, lo cual presenta mayor complejidad que otras medidas de similitud. A su vez, consideran varios tipos de relaciones como “is-a”, “part-of” presentes en una ontología y experimentan con algunos valores constantes (pesos) para ajustarlos a las relaciones no jerárquicas de WordNet.

Similitud semántica para recuperación de información [36]

La medida se basa en los pesos de las relaciones en una ontología asignados por expertos en el dominio especifico de esta. Básicamente definen el factor densidad que denota el grado de un concepto en una jerarquía y el factor profundidad que se refiere a las relaciones de clasificación en una ontología y tienen en cuenta qué tan profundo es el nodo de concepto. Con este enfoque buscan relacionar los pesos en una jerarquía de acuerdo a la densidad y profundidad que tengan los conceptos en ella, así al calcular la similitud se tendrá en cuenta la cercanía entre conceptos en la jerarquía de acuerdo a la mayor especificación de ellos.

Una definición teórica de similitud [37]

En esta investigación, se realiza un estudio de las medidas de similitud y la conceptualización de ellas. Se tienen en cuenta las proposiciones: “La similitud entre conceptos A y B es relacionada con las diferencias entre ellos. Entre mayor sea la diferencia entre ellos, menor es su similitud”. Y “La similitud entre un par de objetos idénticos es 1”. En este trabajo se obtiene de manera teórica una medida de similitud entre dos conceptos:

$$sim(c_1, c_2) = \frac{2 * \log p(c_3)}{\log p(c_1) + \log p(c_2)}$$

Ecuación 3. Similitud semántica de Lin.

Donde c3 es el padre común a c1 y c2 más próximo.

Similitud basada en una ontología y conjuntos de conceptos [38]

En este proyecto implementaron un algoritmo para calcular la similitud entre conjuntos de conceptos pertenecientes a la misma ontología. Su algoritmo implementado es una extensión del algoritmo de Dijkstra, donde puede haber más de un nodo destino y el valor de una ruta es evaluado como el producto de los pesos de las aristas de las rutas. Además proponen que si hay muchos conceptos con alta similitud en dos conjuntos podrían “premiar” la similitud con una función que respete la desigualdad:

$$similarity(o, target, source, lower\_bound) > \frac{\sum_{s \in source} sim\_c(o, target, source, lower\_bound)}{|source|}$$

Ecuación 4. Similitud semántica basada en una ontología

Algoritmo genético para Clustering de documentos usando ontologías [39].

Diferencian las medidas de similitud semántica en dos grupos: Los métodos basados en el conocimiento de conteo (o basadas en diccionario/ tesauro) los métodos basados en teoría de la información (o basadas en el cuerpo del documento). Y proponen un algoritmo genético para clustering de documentos basados en una ontología y evalúan varias medidas de similitud semántica. La medida que utilizan para al cálculo de la similitud en su proyecto está basada en las relaciones entre conceptos de hiponimia (IS-A) y toman en cuenta dos factores: La longitud de la ruta más corta entre dos conceptos y la profundidad de la “subsuma” en la jerarquía. Así, la similitud semántica entre dos conceptos es dada por:

$$sim(c_1, c_2) = f_1(l) \cdot f_2(h),$$

Ecuación 5. Similitud semántica usando ontologías

Donde l es la ruta más corta entre los conceptos c1 y c2, h es la profundidad de la subsuma en la jerarquía de las redes semánticas, es decir, se deriva del cálculo de la longitud más corta de enlaces, sumando desde el concepto raíz de la ontología.

Así, como las descritas anteriormente, hay varias funciones y medidas de similitud definidas por sus investigadores con buenos resultados. En el presente proyecto se utilizaron relaciones jerárquicas entre los conceptos, y una medida de similitud que asigna pesos a los conceptos de la jerarquía dependiendo de su ubicación en ella. Además, se tiene en cuenta mayor peso para los hijos (en vez de los padres) de cada concepto, porque, los descendientes en la jerarquía son más específicos y por tanto suponen mayor similitud con el concepto padre. Con los pesos asignados, se calculan los conceptos más relacionados semánticamente y luego se realiza una ponderación para construir el índice semántico.

1.3.3 HERRAMIENTAS

En la indexación semántica se puede utilizar algunas herramientas que proporcionan cierto grado de facilidad a la hora de crear los índices semánticos. A continuación se presentan las herramientas más importantes que han aportado a la creación uso de índices semánticos.

En la Tabla 1 se muestra un cuadro comparativo extraído de la investigación de C. Cobos [40] de las herramientas más utilizadas en recuperación de la Información. En esta tabla comparativa se observa una gran ventaja de la herramienta Lucene respecto a las otras, debido a la indexación incremental y a la búsqueda realizada por cualquier campo que el usuario decida. Además es una herramienta multiplataforma.

	LUCENE	**TERRIER**	**XAPIAN**	**LEMUR**
Multiplataforma	Si	No	Si	Si
Lenguaje de implementación	Java	Java	c++	c++
Soporte para otros lenguajes	perl, phyton, **c#**, ruby y c++	No	perl, python, php, tcl, c# y ruby	java y c#
Archivos que indexa	pdf, word, html, htm, txt, xml, rtf, entre otros	html , pdf, word, xls, ppt, txt	pdf, word, html, htm, txt, xml, rtf, entre otros	pdf, word, html, ppt, txt, xml, rtf, entre otros
Stemming para varios idiomas	Si	Si	Si	Si

Búsqueda mientras actualiza índice	Si	No	No	No
Indexación incremental	Si	No	No	Si
Modelo de representación	Espacio vectorial	Probabilístico	Probabilístico	Probabilístico
Búsquedas por cualquier campo	Si	No	No	No
Tecnologías enSource	Si	Si	Si	Si
Ultima actualización	Versión: Lucene.Net 2.3.2 Fecha: 24/07/2009	Versión: Terrier 2.2.1 Fecha: 29/01/2009	Versión: Xapian 1.0.16 Fecha: 10/09/2009	Versión: Lemur 4.10.1 Fecha: 28/07/2009

Tabla 1. Comparación de herramientas en indexación semántica.

1.3.3.1 El Prototipo de Espacio Intermedio

El Prototipo de espacio intermedio (Interspace Prototype) [41] es desarrollado por un grupo de investigadores en el programa de Gestión de la Información de DARPA. Es un entorno de análisis para la indexación semántica de la información multimedia en un banco de pruebas de colecciones reales, basadas en el concepto de indexación semántica y agrupamiento semántico para navegar entre conceptos[13].

1.3.3.2 InfoReuser: Motor de indización y búsqueda semántico

La herramienta infoREUSER [42] es un módulo de indexación y recuperación semántica de contenidos ofimáticos (Semántico). Permite la recuperación de archivos de texto, ya sean documentos de texto plano, RTF, MS Word, PDF, HTML, MS Power Point, o MS Excel. Algunas de sus características son:

- **Lematización**: Esta característica permite recuperar documentos que contengan términos escritos en singular o plural, o con verbos conjugados.
- **Propagación por ontologías**: Esto permite que a los términos del usuario, se unan otros que el sistema entienda similares. Esta tarea se aplica a sustantivos, adjetivos y a verbos.
- **Indexación basada en relaciones**: Esto hace que el elemento clave de la indexación y la búsqueda no sean las palabras clave, sino cómo éstas se relacionan unas con otras.

1.3.4 ALGORITMOS

A continuación se describen algunos de los algoritmos más utilizados en los procesos de indexación semántica.

1.3.4.1 Algoritmo Espacio Conceptual

El algoritmo espacio conceptual ha sido usado para generar e integrar múltiples índices semánticos y está basado en correlaciones estadísticas del contexto dentro de los documentos.

En el proceso de este algoritmo se pueden llevar a cabo dos procesos.

- **Extracción de frase nominal**: se encarga de extraer las frases cuyo núcleo es un sustantivo y se realiza en tres fases: Tokenización, Etiquetar "Part-Of-Speech", que comprende el análisis léxico y el contextual; e Identificación de la frase nominal.
- **Análisis de co-ocurrencia** [43]: permite calcular la información de frecuencia en que aparece la frase nominal, la cual es usada para calcular pesos para cada frase nominal en los documentos. Es calculado basándose en una función de similitud asimétrica.

1.3.4.2 Algoritmo Desambiguación del Sentido (concepto) de las Palabras WSD

El algoritmo de desambiguación WSD (Word Sense Disambiguation) permite realizar una desambiguación semi-completa pero precisa, de las palabras que se reciben en una consulta realizada por el usuario [2]. Se usa en las fases de indexación y búsqueda. En el primer caso, este algoritmo permite la desambiguación de las palabras del cuerpo y en el segundo, resuelve las ambigüedades en las palabras de consulta [44].

La Desambiguación del Sentido (concepto) de las Palabras, puede expresarse en un conjunto más amplio de técnicas llamado Procesamiento de Lenguaje Natural (PLN), que básicamente "trata los fenómenos lingüísticos de diversa índole de forma automatizada mediante computadores" [25].

WSD es una fase necesaria para la consecución de acciones como el análisis sintáctico o la interpretación semántica en tareas del PLN, así como para el desarrollo de aplicaciones finales, tanto de recuperación de información, como

de clasificación de textos, análisis de discurso, traducción automática o análisis gramatical, entre otras [25].

1.3.5 PROYECTOS CON INDEXACION SEMANTICA

Desde hace más 10 años se ha investigado y construido índices semánticos para la recuperación de información, a partir de técnicas, algoritmos y herramientas que permiten un buen manejo de índices basados en la semántica de los documentos.

Los proyectos que han utilizado indexación semántica, se pueden clasificar así:

- **Por algoritmos**, los que han utilizado en su proceso principal algoritmos como el de desambiguación, y de espacio conceptual.
- **Por Modelos**, se encuentran los que han utilizado el modelo vector y el de creación del espacio conceptual.
- **Por utilización de herramientas**, corresponde a los proyectos más relevantes que utilizan herramientas ya desarrolladas de indexación como Lucene y lémur, entre otras.

1.3.5.1 Clasificación De Proyectos Por Algoritmos

En el área de la recuperación de información, existen varios proyectos que han logrado una mejora en la relevancia de los documentos recuperados, a través de la implementación de algoritmos específicos. Los más destacados, son los algoritmos de desambiguación, principalmente el WSD (Word Sense Desambiguation) y el algoritmo de espacio conceptual. A continuación se describen algunos proyectos que los utilizan.

El proyecto de indexación semántica ***Semantic Indexing Using Wordnet Senses*** [2] se basa en la implementación de un prototipo que combine la indexación basada en palabras y basada en sentidos o conceptos, utilizando WordNet. El proyecto se realiza en tres etapas que comprenden:

1. **El módulo WSD** en el que cada palabra es reemplazada con un nuevo formato: “Pos|Stem|POS|O.f.f set”. La información obtenida del modulo WSD es usada para el principal proceso de indexación donde la palabra raíz y la ubicación están indexados junto al Synset (conjunto de sinónimos) de WordNet (si existe).
2. **Módulo de indexación.** Indexa los documentos y luego son procesados por el módulo WSD. El segmento Stem y, separadamente el Offset|POS

son adicionados al índice. El proceso de indexación toma un grupo de archivos de documentos y produce un nuevo índice.

3. **Módulo de recuperación**. Rescata documentos basados en una consulta de entrada.

El algoritmo utilizado [45] comprende 10 pasos en los cuales se aplican los procedimientos definidos por ellos y se resumen a continuación:

Procedimiento 1: Pre-procesar el texto con etiquetas utilizando el etiquetado de "part-of-speech" [44].
Procedimiento 2: Inicializar el conjunto de palabras desambiguadas (SDW), con las palabras de entrada (Ambiguas, SAW).
Procedimiento 3: Identificar nombres propios en el texto.
Procedimiento 4: Identificar las palabras que solo tienen un sentido en WordNet.
Procedimiento 5: desambiguar palabras basadas en su ocurrencia.
Procedimiento 6: Identificar el conjunto de nombres.
Procedimiento 7: Identificar sinónimos en SAW y en SDW.
Procedimiento 8: sinonimia entre palabras de SAW.
Procedimiento 9: Identificar palabras de SAW con distancia máxima de 1 respecto a las palabras SDW.
Procedimiento 10: distancia máxima entre palabras de SAW.

Una vez terminados los procesos descritos anteriormente, crean un Benchmark con 52 textos para probar el método de desambiguación.

En sus resultados se observa la eficiencia del método WSD para dominios abiertos, lo cual demuestra que el algoritmo aplicado es adecuado en tareas de indexación semántica. Esta indexación ofrece una mejora en las técnicas actuales de recuperación de información.

La investigación realizada en ***Semantic Indexing For A Complete Subject Discipline*** [13] permitió el desarrollo de una técnica estadística que pertenece a la semántica escalable, la cual indexa grandes colecciones para búsquedas profundas. Para llevar a cabo este experimento, utilizaron los registros bibliográficos de la Biblioteca Nacional de Medicina (NLM) de Estados Unidos.

En el proceso de indexación se utiliza el **algoritmo espacio conceptual** adoptado en varios estudios [46-48] y usado para generar e integrar múltiples índices semánticos. Se realizan etapas intermedias en el proceso, las cuales se describen a continuación.

1. **Extracción de la frase nominal:** Se utilizaron algunas reglas de identificación de frases nominales y fraseo para el desarrollo de un extractor de frases. El fraseo opera en tres fases: Tokenización, Marcado de "Part of Speech" (parte de la palabra) basándose en el etiquetador de *Brill* [49] e Identificación de la frase nominal con AZ Noun Phraser [50]. Se guarda la frecuencia de la frase nominal para calcular sus pesos en los documentos.
2. **Análisis de co-ocurrencia**: Este análisis se calcula basándose en una función de similitud asimétrica. Su resultado es una matriz que representa una red de frases nominales y sus probables relaciones.

Para indexar MEDLINE realizaron dos pasos:

- Dividir en segmentos los subdominios para navegar por ellos.
- Utilizar la ordenación de Medical Subject Headings (MeSH) [51], que consiste en una estructura alfabética y jerárquica llamada MeSH. Esta tiene las propiedades de un tesauro y un sistema de clasificación.

Cabe agregar, que el presente proyecto también hace parte de la clasificación por Modelos en "Espacio Conceptual", puesto que construyen un prototipo de espacio conceptual, basados en el algoritmo descrito anteriormente. La construcción del espacio se realiza para *MEDSPACE: Semantic Indexing experiment For A Medical Discipline* (ítem 4 del documento), lo cual constituye la creación de su prototipo para obtener los resultados experimentales en su investigación.

MEDSPACE, es un experimento de indexación semántica para una disciplina médica, en el cual escogieron conjuntos de frases desde MeSH que describen bien los artículos o documentos y luego se creó el Prototipo Interespacio. Al finalizar el trabajo comenzaron a evaluar la utilidad de los sistemas vs. las necesidades de información.

Para nuestro proyecto se utilizaron técnicas de extracción de conceptos (ontología) pero en la indexación no se realizó una extracción de frases nominales. Posteriormente, en el procesamiento de la consulta (de usuario), se realizó una combinación de términos que permite encontrar frases nominales en la búsqueda.

1.3.6 CLASIFICACIÓN DE PROYECTOS POR MODELOS

1.3.6.1 Espacio Vectorial

Otra forma de proporcionar más eficiencia en las búsquedas Web, es basándose en los modelos actuales de recuperación de información con algunas modificaciones en dichos modelos y arquitecturas. A continuación se describen algunos proyectos que lo usan.

El proyecto ***A Novel Approach to Semantic Indexing Based on Concept*** [52] describe el método de indexación basado en un "Concept Vector Space", es decir, el espacio vectorial de conceptos, a través del cual se representa el contenido semántico de un documento.

Para la extracción de conceptos utilizaron cadenas léxicas con los vectores de concepto y vectores de texto, así se calculan los índices semánticos y su grado de importancia semántica.

El sistema propuesto tiene cuatro componentes:

- Construcción de cadenas léxicas[53].
- Ponderación de cadenas y nombres.
- Reponderación del término basada en el concepto.
- Extracción del índice del término semántico.

En los dos primeros, se discriminan las cadenas representativas de las cadenas léxicas. Las cadenas representativas son cadenas delegadas para representar un concepto característico de un documento. Se asume además que los conceptos son independientes entre sí, sin considerar su similitud. En el segundo bloque, las cadenas léxicas son empleadas para la extracción de conceptos. Son formadas usando WordNet y relaciones asociadas entre palabras. Las cadenas tienen cuatro relaciones:

Sinónimos (synsets)	Son palabras que tienen un significado similar o idéntico entre sí, y pertenecen a la misma categoría gramatical. Ej. Carro tiene como sinónimos: coche, auto, automóvil.
Hiperónimo (hypernyms)	**E**l término más alto en una jerarquía terminológica, que incluye a otros. Es el género superordinado (nivel más alto) respecto a sus especies. Ej. El hiperónimo de carro podría ser vehículo, pues está en un nivel superior en la jerarquía respecto a carro, coche, auto, etc
Hipónimos (hyponyms)	Término específico y subordinado a otro más general. El hipónimo tiene al menos un atributo más que el hiperónimo o término superior a él, que lo especifica y le da identidad propia [54]. Ej. Hipónimos de carro serían: camión, camioneta, autobús, puesto que son carros pero

	tienen características más específicas como el tamaño y el uso de ellos.
Merónimo (meronyms)	Palabra que representa un miembro de, que forma parte de, ó que es sustancia de algo [55]. Ej. Merónimos de carro son sus partes: sillas, cajuela, motor y volante, entre otras.

Tabla 2. Relaciones entre cadenas

El índice semántico y el peso son extraídos de acuerdo al valor numérico de la cantidad de información y "Ratio" (proporción) de información, definidas en el proyecto así:

- **Cantidad de información**: Cantidad semántica de un texto, concepto o palabra en todo el documento. Esta magnitud es generada por la composición de todos los conceptos
- **Ratio de Información**: Es la proporción de la cantidad de información de una etiqueta comparativa con la cantidad de información de un texto, concepto o palabra. Esta medida denota la proporción de información de una palabra con respecto al concepto en el cual está incluido.

Siguiendo el proceso anterior, realizaron la comparación evaluativa con otros métodos de indexación como el estándar de frecuencia de términos (Stándard TF) y la extracción del peso semántico. Según su experimento, demostraron gráficamente que los resultados de la extracción por el peso semántico se acercan mucho a los de la extracción de índices de términos semánticos, realizada en este proyecto.

1.3.6.2 Espacio Conceptual

En el proyecto ***Performance And Implications Of Semantic Indexing In A Distributed Environment*** [14], se desarrolla un prototipo que contiene un amplio conjunto de clases y relaciones de datos para el módulo de indexación semántica, construido en un entorno distribuido de análisis. El desarrollo del prototipo se llevó a cabo en dos fases que se describen a continuación:

Fase 1: Se realiza el pre-procesamiento necesario de los documentos. Los pasos de esta fase son:

- Extracción de sintagmas (frases) nominales [21] de cada documento en una colección, los cuales se convierten en los conceptos.
- Estos conceptos son puestos en una lista global, así como una lista que pertenece a cada documento. Y se recogen varias estadísticas para cada concepto.

Fase 2: Se distribuyen las tareas de indexación a diferentes máquinas en el entorno y se utilizan una función de similitud [56] para la asociación de conceptos.

- Se inicializan los procesos en el depósito. Se utilizan mensajes a procesos maestros y esclavos.
- Luego se genera la indexación real, realizada por diferentes equipos de cómputo. Cada proceso esclavo se inicia mediante una asignación de carga de trabajo.
- El paso final es la recopilación de los resultados y el trabajo de limpieza del proceso.

Con lo anterior construyen su prototipo Espacio Intermedio, “Interspace” [57], el cual es la base de indexación semántica (la creación de un espacio conceptual). El espacio conceptual se basa en un cálculo estadístico que determina las relaciones entre los conceptos de una colección de documentos.

Como resultado del proyecto, se obtuvo un mejor rendimiento en la indexación semántica de documentos en un entorno distribuido (cinco máquinas). Concluyen que la indexación semántica en la actualidad, se puede llevar a cabo con pocos recursos en comparación con la indexación realizada en el pasado.

En el anterior proyecto se mencionan dos etapas importantes en la realización de la indexación semántica en documentos que son el procesamiento de los documentos y tareas de indexación a diferentes maquinas. Pero en otros proyectos se propone métodos como la creación de un índice plano utilizando alguna herramienta y la especialización de este mismo adicionando relaciones semánticas entre conceptos, para nuestro proyecto es necesario seguir estas segundas técnicas ya que ajustan más a la necesidad del mismo.

1.3.7 CLASIFICACIÓN DE PROYECTOS POR USO DE HERRAMIENTAS

La utilización de algunas herramientas para la indexación y sus respectivas modificaciones ha sido un factor importante para mejorar las búsquedas, teniendo en cuenta la semántica de las palabras. De esta manera se logra encontrar con mayor precisión lo que realmente se busca. Lucene es la herramienta que observamos es líder en el proceso, puesto que ha demostrado que al trabajar en una indexación semántica, se acopla mucho a las necesidades de información. A continuación se describe uno de los proyectos realizados con éxito utilizando dicha herramienta.

El proyecto ***The effect of Semantic Index in Information Retrieval development*** [3], propone un sistema basado en la recuperación de información con índices semánticos denominado: Semantic Information Retrieval System (SIRS), el cual

incluye dos módulos importantes: Semantic Indexer SI (indexador semántico) y Query Searcher QS (buscador de consultas). En la Figura 2 se observa el modelo descrito.

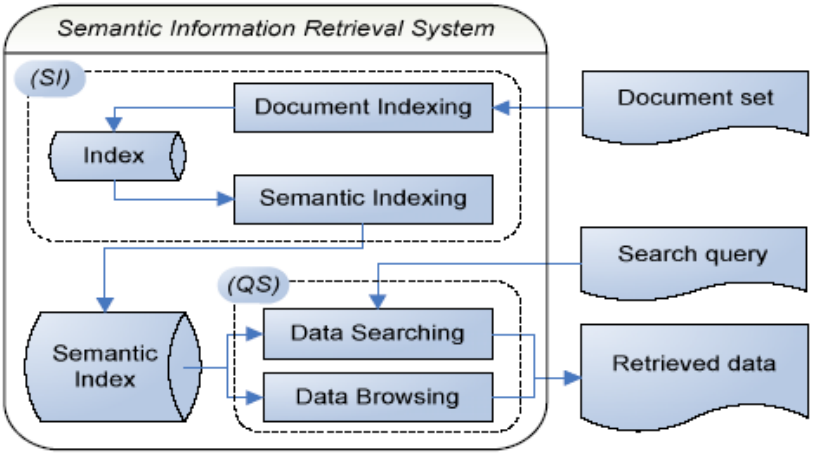

Figura 2. Modelo propuesto por SIRS [3]

Para el primer módulo se presentan dos etapas importantes:

- Indexación de documentos. Se realiza una indexación básica de documentos con la herramienta Lucene [58].
- Indexación semántica (para crear el índice semántico añadiendo más información al índice creado en el primer paso). Los usuarios pueden recuperar los datos por navegación.

El índice semántico no solo se encarga de almacenar todos los registros ordenados sino que también apoya la búsqueda de mecanismos para proporcionar los datos pertinentes.

La investigación de este proyecto se enfoca en lo siguiente:
Un sistema SI: se presenta un SI de formación heurística (llamado indexación semántica) que incluye como primer paso elaborar un índice normal (de recuperación de información) con Lucene y el segundo paso añadir algunas relaciones semánticas, particiones y clases semánticas para construir el índice semántico.

Fusión de dos índices semánticos: puede ser utilizada para combinar índices semánticos de recuperación de información o de bibliotecas digitales para crear nuevos índices. Este proceso es útil para hacer índices de varios dominios específicos y se realiza en dos pasos:

- Fusión de dos índices semánticos con el mismo atributo y estableciendo la relación.
- Fusión de dos índices semánticos con diferentes atributos y estableciendo la relación.

Extracción de sub índice semántico: Se realiza con el fin de crear un índice de dominios específicos a partir de un índice multi-dominio. Los casos para esta actividad son:

- E1: Extracción del sub índice semántico con el sub conjunto de atributos y sus relaciones.
- E2: Extracción de sub índice semántico con la restricción del documento Fuente.

En la experimentación de indexación semántica se trabajó en dos pasos:

- Crear un índice normal utilizando las herramientas de Lucene.
- Crear el índice semántico, añadiendo un poco más de información en la base de entradas en el índice creado en el primer pasó.

Como conclusión de este trabajo, se pudo observar que el tamaño de los índices no es muy alto pero el tiempo de indexación es bastante elevado, por lo tanto las tareas para su trabajo futuro es mejorar la indexación y el algoritmo de búsqueda con el fin de reducir el tiempo de procesamiento.

En este proyecto se siguen técnicas como la realización de índices planos los cuales se especializan adicionado relaciones semánticas, este fue una de las técnicas que se tuvieron en cuenta a la hora de la creación de nuestro índice semántico.

1.4 PROYECTOS CON ONTOLOGÍAS E ÍNDICES SEMÁNTICOS

Como observamos anteriormente, las ontologías y los índices semánticos son dos técnicas muy importantes en la resolución de los problemas de recuperación de información, cada una de ellas ha brindado grandes ventajas a la hora de resolver problemas de este tipo, por ello, este documento se enfoca en la utilización conjunta de éstas dos técnicas.

En los últimos 10 años se han llevado a cabo proyectos que han hecho uso de estas técnicas obteniendo resultados muy favorables. A continuación nombraremos los más importantes.

1.4.1 CREACIÓN DE ÍNDICES SEMÁNTICOS CON ENRIQUECIMIENTO DE ONTOLOGÍA.

Un problema detectado, es que muchos conceptos extraídos de un documento y que pertenecen a determinado contexto, no están presentes en la ontología de dominio. Por lo cual, en el proyecto ***Ontology enrichment and indexing process*** [7], tienen como objetivo principal construir un índice de estructura de las páginas Web de acuerdo a una ontología, la cual proporciona la estructura del índice. Para llevar a cabo la construcción del índice, ellos proponen cuatro pasos generales, con los cuales se logra la construcción del mismo con estructura como se muestra en la Figura 3.

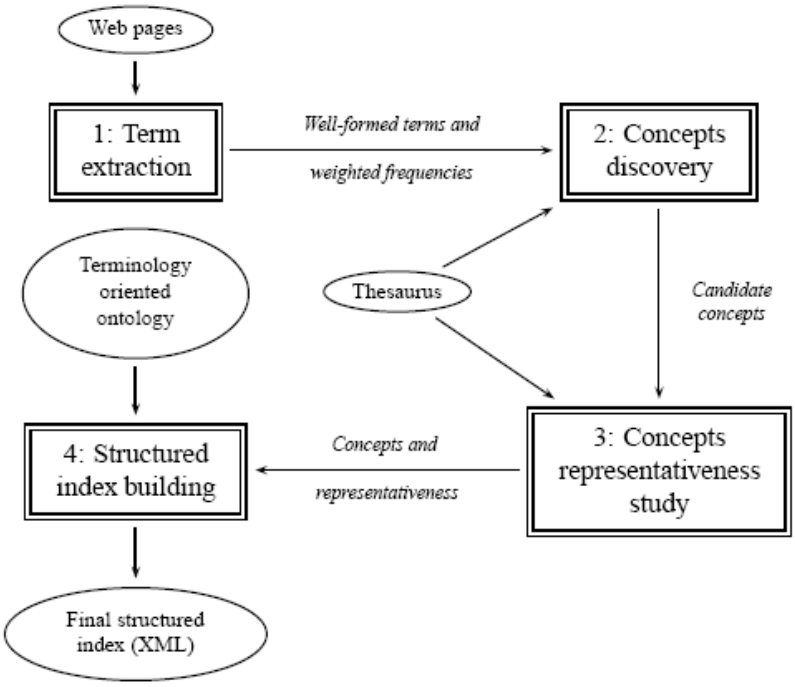

Figura 3. Proceso de Indexación según Desmontils, C.J., L. Simon [7]

- Como primer paso, a cada página se le construye un índice plano de los términos, en donde cada término es relacionado con su frecuencia ponderada.
- Utilizan el tesauro de WordNet [59] para generar los conceptos candidatos que pueden ser etiquetados por un término del índice anterior.
- Cada concepto candidato se estudia para determinar su representatividad en el contenido de la página Web. Esta evaluación se basa en la frecuencia ponderada y en las relaciones con los otros conceptos.

- Entre los conceptos candidatos, se aplica un filtro a través de la ontología y la representatividad de los conceptos. Un concepto seleccionado es un concepto que pertenece a la ontología y tiene una alta representatividad de los contenidos de la página.

Como segundo objetivo del proyecto es el enriquecimiento de la ontología, para lo cual tuvieron en cuenta dos criterios:

- El enriquecimiento de perfeccionamiento (o especialización), que trata de regresar a una ontología más especializada.
- Enriquecimiento mediante la abstracción, que trata de hacer que la ontología sea más general (por la ampliación del campo o suprimiendo conceptos muy específicos).

Se propone un método semiautomático del enriquecimiento de la ontología [60], que ofrece a los expertos un medio de comunicación de gran alcance para gestionar el dominio cubierto por la ontología. Para esto realizan cuatro pasos:

- Un índice estructurado para añadir los conceptos de la ontología que son útiles.
- Un post- tratamiento basado en una poda de la estructura del índice final.
- Una herramienta de la validación de la ontología.
- Enriquecimiento de la ontología utilizando un diccionario de sinónimos.

Para el enriquecimiento de la ontología durante la construcción del índice utilizan WordNet como diccionario de sinónimos, lo cual hace posible el proceso de agregar a la ontología algunos conceptos que están presentes en el índice plano, pero que no pertenecen a la ontología. Para determinar estos conceptos utilizan la heurística basada en rutas de hiperonimia asociadas con los conceptos en WordNet y en las relaciones “IS-A” asociada a los conceptos de la ontología.

El enriquecimiento de la ontología no es completamente automático. Un experto humano tomará la decisión final de añadir o no un nuevo concepto de la ontología, mediante una herramienta de visualización desarrollada, para ayudar en el proceso de indexación y control de la desviación de la ontología potencial.

El proceso propuesto por este proyecto trae consigo una serie de **ventajas** con relación a otros procesos de indexación tradicional. Éstas son:

1. Las páginas seleccionadas contienen, a su vez, las palabras claves y los conceptos necesarios.
2. De estos conceptos, son más representativos los temas tratados en las páginas seleccionadas.
3. Las Páginas puede comprender los conceptos necesarios y los más específicos.
4. La importancia de un concepto no sólo depende de su frecuencia del término, sino también en los marcadores de HTML y también sus relaciones con los otros conceptos de la página.

1.4.2 INDEXACION SEMANTICA UTILIZANDO TESAURO Y ONTOLOGIA

La investigación realizada en ***Semantic Indexing of Technical Documentation*** [8], se basa en una extensión del modelo vectorial y proponen un modelo de indexación semántica que explota las estructuras lógicas y el contenido semántico de los documentos. Los documentos técnicos son fuertemente estructurados y se componen de etiquetas como títulos, secciones, párrafos, capítulos, etc. Teniendo en cuenta esto, utilizan la extensibilidad del lenguaje XML que permite representar simultáneamente el contenido y la estructura lógica de los documentos.

Realizan un pre-procesamiento de los documentos, lo cual, permite identificar los términos candidatos por medio de un tesauro.

1. Utilizan herramientas de procesamiento de lenguaje natural como tokenizadores, lematizadores y otros analizadores para obtener el lema de las palabras.
2. Proyectan los documentos en WordNet para extraer los términos que coinciden con ellos.
3. Para los términos compuestos (frases nominales) buscan una combinación más grande de palabras (máximo siete) que coincidan con una entrada en WordNet.
4. Calculan los pesos de los términos de acuerdo a su frecuencia en los elementos lógicos en los que aparecen (títulos, secciones, etc.) y teniendo encuentra la importancia de su elemento en el documento.

El índice del documento es compuesto de varios vectores de términos ponderados: uno por tipo de elemento. Luego unen todos los vectores de términos ponderados, que describen un elemento lógico, en un solo vector. Posteriormente realizan la detección de conceptos utilizando una ontología.

5. Realizan un mapeo entre los términos candidatos y los conceptos de la ontología, obteniendo un conjunto de conceptos por cada término candidato.

6. Escogen para cada término, el mejor concepto dentro del conjunto, el cual representa la semántica del documento. Asignan una puntuación para cada concepto en el mapeo semántico y escogen el concepto con la más alta puntuación.
7. Utilizan una función de similitud que mide el camino más corto entre conceptos.
8. Así, el índice del documento es un conjunto de pesos por cada concepto encontrado.

Una tarea compleja encontrada es la conversión de todo el cuerpo de los documentos en formato XML para lograr el etiquetado de la estructura. Sin embargo la indexación semántica realizada presenta muy buenos resultados según sus pruebas y la utilización de las dos herramientas proporciona ventajas en la extracción de conceptos.

1.4.3 CREACIÓN DE ÍNDICE SEMÁNTICO BASADO EN LA FRECUENCIA PONDERADA, Y CÁLCULO DE REPRESENTATIVIDAD DE CONCEPTOS

El siguiente articulo presenta una nueva idea para indexar un sitio Web, haciendo uso de ontologías y técnicas del lenguaje natural para la recuperación de información en la Internet. Esta propuesta es presentada en la investigación ***Indexing a Web Site with a Terminology Oriented Ontology*** [1], y cuyo objetivo es realizar un proceso semiautomático, que ofrece un índice basado en el contenido de un sitio Web, donde utilizan las técnicas del lenguaje natural.

El proceso de indexación que este proyecto, es similar al descrito en el anterior[6], pues ellos se basan en esta investigación. Sin embargo se presentan a continuación las especificaciones de éste, y en la Figura 4 se muestra el mismo.

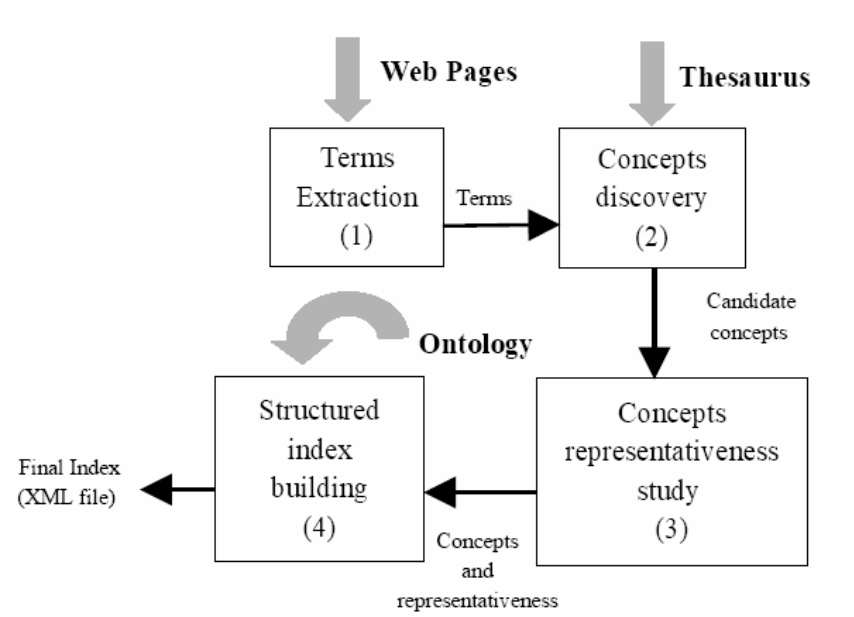

Figura 4. Proceso de Indexación [1]

En este proceso, se determinan todas las etiquetas candidatas de un concepto. Se basa en un diccionario de sinónimos y utiliza un número de heurísticas similares como los propuestos por Microkosmos [61].

Para la construcción del índice tienen en cuenta dos pasos esenciales:

- Condiciones de extracción de páginas Web y el cálculo de la frecuencia ponderada.
- Determinación de los conceptos candidatos y el cálculo de representatividad de un concepto.

Estos dos conceptos son puntos importantes a tener en cuenta en la construcción del índice, pero además cada una de ellas tiene sus respectivas partes.

Para las condiciones de extracción:

· Eliminación de los marcadores de HTML de las páginas Web.
· Dividir el texto en frases independientes.
· Lematización de las palabras incluidas en las páginas. A continuación, las páginas Web se anotan con parte de las etiquetas de voz, utilizando el etiquetador de BRILL [62].

Proceso para generar conceptos candidatos
El proceso de generación de conceptos candidatos se hace con WordNet [63], a continuación a cada concepto candidato se le calcula la representatividad de acuerdo a la frecuencia ponderada y su similitud acumulada del concepto, con relación a los otros conceptos de la página. La similitud acumulada se basa en la similitud entre dos conceptos, lo cual permite evaluar la distancia semántica entre ellos. Esta medida se define relativamente a un tesauro y la relación entre hiperónimos.

Asociación de conceptos y conjunto de sinónimos (synsets)
Los conceptos candidatos se corresponden con los conceptos de la ontología. Si un concepto está en la ontología y en la página Web, la dirección URL de esta página y su representatividad, se añade a la ontología. El proceso de evaluación permite valorar la adecuación entre las páginas y la ontología y así adoptar estrategias diferentes en función del valor de los coeficientes.

En conclusión el proceso que proponen otorga una serie de ventajas sobre los métodos de indexación tradicional e incluso sobre los métodos de anotación Web. Además, los resultados presentados pueden ser utilizados en diversas aplicaciones. Actualmente se estudian otras relaciones genéricas y relaciones específicas, con el fin de mejorar el proceso de extracción de conceptos. Hoy día este proceso está siendo incorporado en el sistema Bomon Multiagente [64], para buscar información relevante en Internet.

En ***Towards Building Semantic Rich Model for Web Documents Using Domain Ontology*** [6], se enfocan en la construcción de modelos de la Web semántica, para documentos que emplean el análisis del lenguaje natural y un conjunto de ontologías de un dominio especifico (en este caso el ámbito medico). Estos enfoques se utilizan para realizar el análisis textual, que se traduce no solo en la identificación de conceptos importantes presentados en el documento, sino también las relaciones entre estos conceptos. En este proyecto se sigue el enfoque general de la construcción del índice propuesto por Desmontils y Jacquin.

El proceso llevado a cabo es el siguiente:

Análisis de documentos: Para este análisis se realizan los siguientes pasos:

- Se toma los documentos HTML en proceso de transformación, para ser codificados y generar archivos de tipo ASCII, los cuales son documentos libres de etiquetas HTML.

- Luego los documentos son sometidos a un proceso de análisis de palabras.
- En el documento filtrado, todas las palabras vacías serán eliminadas y los conceptos seleccionados se derivan a su raíz para ser ordenados de acuerdo a la frecuencia de aparición en el documento.
- Se dividen los documentos en párrafos y luego en frases, las cuales se almacenaran en un repositorio.
- Los conceptos con alta frecuencia previamente obtenidos en el proceso de análisis de palabras, son comparados con las frases almacenadas en el repositorio, con el fin de seleccionar las frases candidatas para ser utilizadas en el NLA (Natural Lenguaje Analysis)[5].

Los resultados del proceso de análisis de documentos, es una lista de los posibles conceptos candidatos, y la lista de las frases en donde los conceptos fueron encontrados.

Análisis del lenguaje natural: este análisis se define en las siguientes etapas:

- **Morfología y proceso de acceso de análisis semántico:** se analizarán las frases de entrada (oraciones que contienen conceptos candidatos) previamente almacenadas en el repositorio de frases en un árbol de análisis utilizando el Analizador Apple Pie Parser [65].
- **Análisis semántico:** Extrae las relaciones semánticas entre los conceptos seleccionados. Se realiza ya sea por la ontología de dominio específico o por la explotación de la estructura semántica de las oraciones analizadas con la ayuda del usuario.
- **Modelo global de la semántica del documento:** Los modelos de los documentos semánticos serán almacenados y sometidos a un proceso de integración, para la creación de un modelo de documento de semántico global. Este será usado para la recuperación y la navegación semántica.

Se concluye que las ontologías de dominio juegan un papel importante en las tareas de clasificación y organización de documentos. En este proyecto se combinó una ontología de dominio con las técnicas del lenguaje natural, donde éste no sólo sirvió para extraer conceptos importantes, sino también para construir el contenido semántico de los documentos Web.

1.4.4 ALGORITMO DE RANKING Y MODELO VECTORIAL

Uno de los principales objetivos perseguidos en el campo de la Web semántica, radica en la mejora de las técnicas actuales de recuperación de información,

[5] Análisis del lenguaje natural

mediante el uso de nuevas metodologías englobadas bajo el nombre de búsqueda semántica. ***El Proyecto de trabajo de iniciación a la investigación*** [66], se centra en este objetivo, por lo cual, realiza la implementación de un nuevo modelo de búsqueda semántica enfocada en la recuperación de información sobre grandes repositorios de documentos. Este modelo de recuperación se basa a una ontología de dominio y en bases de conocimiento.

Para este proyecto, se tienen en cuenta las propuestas de KIM [67, 68] y TAP [69] que son las más completas publicadas hasta la fecha, para la construcción de bases de conocimiento y la anotación automática a gran escala. Pero este trabajo complementa a KIM y TAP con un algoritmo de ranking, específicamente diseñado para un modelo de recuperación de información basado en ontologías, utilizando un sistema de indexado semántico centrado en la ponderación de anotaciones entre los conceptos de las bases de conocimiento y los documentos almacenados en el repositorio. Además, el presente proyecto se basa en la idea de que "la búsqueda semántica sea un complemento de la búsqueda por palabra clave mientras no haya suficientes ontologías y metadatos disponibles"[70]. El proceso propuesto por este proyecto se puede ver en la Figura 5.

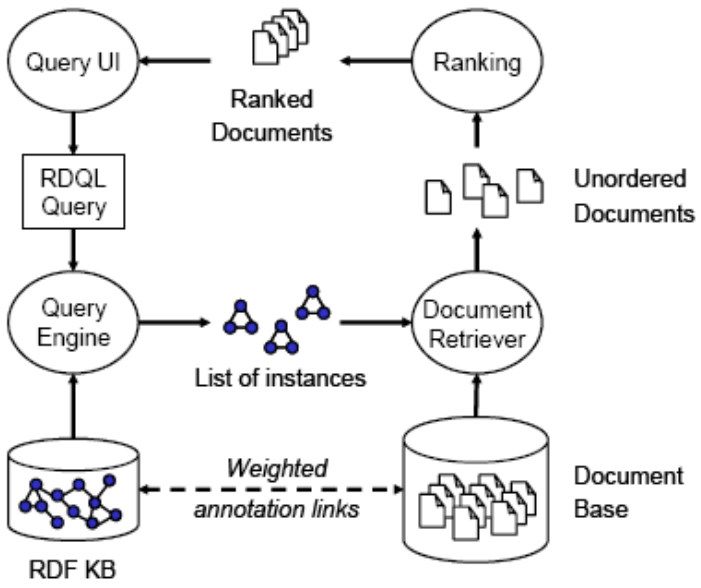

Figura 5. Vista de Modelo de Recuperación de Información Basado en Ontologías [66]

El sistema sigue estos pasos:

- Toma como entrada una pregunta formal expresada en lenguajes como RDQL (Lenguaje de consultas que permite extraer metadatos de archivos disponibles en una URL) [71] o una interfaz de formulario [72].
- La pregunta puede generarse mediante una consulta basada en palabras clave, basada en lenguaje natural, una interfaz de formulario o técnicas de usuario más sofisticadas.
- Se procede a recuperar la información que mejor se adapta a las necesidades del usuario. Este proceso se puede ver en dos fases: la primera es la consulta formal ejecutada contra una base de conocimiento, se devuelve una lista de instancias o tuplas que cumplen los requisitos de la consulta. Como segunda fase, se utilizan las anotaciones de dichas instancias con los documentos del repositorio para recuperar el conjunto de documentos que satisfacen la consulta del usuario.
- Los documentos son ordenados y presentados al usuario siguiendo una adaptación del modelo vectorial que utiliza los pesos de las anotaciones, para aclarar el orden y así presentar al usuario los documentos que contienen la semántica que mejor responde a la necesidad del mismo.

El modelo de este trabajo puede verse como una evolución del modelo vectorial clásico [73], donde los índices basados en palabras claves, son reemplazados por bases de conocimiento fundamentadas en ontologías.

1.5 COMPARACION DE PROCEDIMIENTOS

Con la revisión anterior de los proyectos se puede realizar una comparación de sus procedimientos y algunos de los pasos que tienen en común. Para esto se describe a continuación una tabla comparativa de cada proyecto y la descripción de los pasos que utiliza. Se marcan los pasos que realiza cada uno en la casilla correspondiente.

Nota: En el proyecto *Towards building Semantic rich model for web documents using domain Ontology* [6] realizan una tokenización y lematización de términos pero con la característica de que obtienen conceptos.

PROYECTOS	PROCEDIMIENTO											
	1. Extracción términos	2. Extraer cadenas léxicas	3. Tagger o anotador (Part-Of-Speech)	4. calculo de frecuencia	5. Extracción de conceptos			6. Análisis de conceptos		7. calculo de frecuencia		8. Asociación conceptos y Synsets
	Tokenizar, lematizar, análisis léxico	Frases nominales	Anális léxico y contextual	Ponderación de frecuencia de términos	WordNet	Ontología	Otro	Espacio Conceptual	Relaciones Conceptos	Representatividad	Reponderación	Asociación de conceptos
[52]. Indexación semántica basada en el concepto.					X			X	X	X	X	
[2]. Indexación semántica usando los sentidos de WordNet.	X	X	X		X				X	X		X
[13]. Indexación semántica para una disciplina específica.	X	X	X	X			X	X	X	X		
[14]. Indexación semántica en un entorno distribuido.		X					X	X	X	X		
[3]. Indexación semántica en la Recuperación de Información.	X.						X					
[1]. Indexación de un sitio web con ontologías.	X	X	X	X.	X	X			X	X		X
[7]. Enriquecimiento de una ontología y proceso de indexación.	X.	X	X	X	X	X			X	X		X
[6]. Modelo rico en semántica usando una ontología de dominio.	X	X	X	X		X	X					

Tabla 3. Comparación procedimientos. Generación de índices semánticos.

La comparación anterior permite observar cuales son los pasos más comunes en la creación de índices semánticos. La extracción de términos y frases nominales se realiza en varias investigaciones de manera conjunta, lo cual, según sus autores, genera buenos resultados en la recuperación de información.

Por otra parte, el estudio de relaciones entre conceptos y/o el cálculo de representatividad, son necesarios en la construcción de dicho índice, pues proporcionan los valores numéricos de los conceptos y sus relaciones para determinar lo que el usuario realmente necesita. Las herramientas más utilizadas son WordNet (tesauro) y las ontologías, los cuales proporcionan relaciones semánticas más precisas para la extracción de conceptos, como relaciones jerárquicas (sinonimia, hiponimia[6], hiperonimia[7]) y de inferencia.

1.6 HERRAMIENTAS PARA LA CONSTRUCCIÓN DE PROCEDIMIENTOS

Como parte del diseño para la creación del procedimiento se cuenta con herramientas que facilitan la abstracción de los pasos necesarios en la creación de un índice semántico. Estas abstracciones permiten obtener la estructura, reglas, restricciones y objetos presentes en dicho procedimiento. Se describen a continuación varias herramientas utilizadas para definir procedimientos, para mayor información, puede ver el Anexo C.

1.6.1 META-MODELOS Y ESTÁNDARES

Existen algunos estándares y meta-modelos útiles como el estándar BPMN, el IDEFØ y el meta-modelo SPEM para la creación de métodos, procesos y procedimientos, los cuales apoyan el modelado y diseño del objetivo. Un meta-modelo de procesos “describe un conjunto de conceptos genéricos y sus interrelaciones” [74], los cuales son básicos en la definición de modelos de procesos. A continuación se describe brevemente cada uno.

El estándar **BPMN** (Business Process Modeling Notation), orientado a toda organización, es muy reconocido y utilizado en la “diagramación y especificación de procesos conceptuales y lógicos” [75], incluyendo diseños de procesos orientados a tecnología BPM: Workflow.

El estándar **IDEFØ** (y sus derivados) es un método diseñado para el modelado de decisiones, acciones y actividades de una organización o sistema. Es muy útil en

[6] Relación de inclusión de un significado respecto de otro. Ejem: “perro” está incluido dentro de “animal”. Extraído de http://faculty.ksu.edu.sa/belaichi/Clases/SEM%C3%81NTICA/HIPONIMIA.pdf.

[7] Relación inversa de inclusión. Términos más generales incluyen los particulares.

la gestión de procesos y el modelado de funciones integradas, tanto en gestión como en reingeniería de procesos de servicios y ayuda a establecer el alcance de un análisis. Los modelos construidos basándose en este estándar, ayudan a organizar el análisis de un sistema y a promover la buena comunicación entre el analista y el cliente [76].

El metamodelo **SPEM** [77] (Software and Systems Process Engineering Metamodel Specification), se utiliza en el modelado de métodos y procesos de software, fué creado por El OMG (Object Management Group). SPEM no es un lenguaje de modelado de procesos general ni provee conceptos para modelado del comportamiento, sin embargo, incluye mecanismos para realizar un adecuado modelado (diagramas de actividad de UML 2, BPMN, entre otros) [78].

Para el desarrollo y creación del procedimiento actual, no es necesario seguir el meta-modelo, dado que en este caso, incrementa la complejidad del procedimiento, al requerir aprendizaje extra para manejar un entorno de trabajo y herramientas que no son indispensables en la creación de dicho procedimiento.

1.6.2 DIAGRAMAS DE PROCESOS

1.6.2.1 Diagrama de flujo

Los diagramas de flujo (o flujo-gramas) son gráficas que representan la secuencia de operaciones, pasos, etapas y/o actividades que ocurren durante un proceso. Se utilizan símbolos gráficos para su representación. Puede incluir, además, la información que se considera deseable para el análisis, por ejemplo el tiempo necesario entre las etapas [79]. Favorecen la comprensión del proceso pues se define el inicio, el final, las relaciones entre las actividades y los puntos de decisión durante el proceso [80].

1.6.2.2 Diagrama de actividad

El Diagrama de Actividad es un diagrama de flujo del proceso multi-propósito que hace parte del estándar UML (*Unified Modeling Language*)[8], el cual, se usa para modelar el comportamiento del sistema y también para modelar un Caso de Uso, o una clase, o un método complejo[9]. Un diagrama de actividades representa los flujos de trabajo, por pasos, del negocio y operaciones del sistema, en otras palabras, muestra el flujo de control general. Se utiliza para modelar un proceso de flujo de trabajo y/o secuencias de acciones y condiciones dentro de un proceso

[8] Es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema. UML ofrece un estándar para describir un "plano" del sistema (modelo), incluyendo aspectos conceptuales tales como procesos de negocio y funciones del sistema. Extraído de http://www.grupoinformatica.com/biblioteca-articulos/1459-uml-lenguaje-unificado-de-modelado.html.

[9] Extraído de http://www.ibiblio.org/pub/linux/docs/LuCaS/Tutoriales/doc-modelado-sistemas-UML/multiple-html/x291.html

[81]. Los diagramas de actividades se componen principalmente de los elementos flujos y Nodos, (Ver Anexo C, para más detalle).

En la presente investigación, se hizo uso de esta herramienta para modelar el procedimiento de creación de índices semánticos, los cual permite observar los flujos de procesos y/o actividades a seguir en cada paso del mismo.

1.6.3 MAPA CONCEPTUAL Y MENTAL

Los mapas conceptuales son diagramas (bidimensionales) que muestran relaciones significativas entre conceptos en determinado ámbito o área de estudio y se expresan como proposiciones. Una proposición consta de dos o más términos conceptuales unidos por palabras (palabras de enlace) para formar una unidad semántica [82]. Los mapas conceptuales pueden representarse en forma de estrella, con un concepto o palabra central que irradia otros conceptos relacionados (también llamado mapa mental), o en forma de árbol invertido, con el nodo raíz en la parte superior (concepto principal) y el resto de forma jerárquica descendente.

El mapa mental es un diagrama que permite organizar, representar y analizar la información sobre un ámbito, con el propósito de facilitar los procesos de aprendizaje, administración y planeación organizacional así como la toma de decisiones. Mediante estas herramientas se pueden representar nuestras ideas utilizando de manera armónica las funciones cognitivas de los hemisferios cerebrales [83]. Facilitan además, la organización lógica y estructurada de los contenidos de aprendizaje, ya que permiten seleccionar, extraer y separar la información significativa o importante de la información poco útil dependiendo del área de investigación, y a su vez, se pueden insertar nuevos conceptos en la propia estructura de conocimiento [84].

Existen varias herramientas para construir mapas mentales y conceptuales, una de las más utilizadas es FreeMind.

FreeMind [85] es una herramienta para la elaboración y manipulación de mapas conceptuales. Permite organizar y estructurar las ideas, los conceptos, su relación entre ellos y su evolución. Puede ser utilizada para cualquier área o ámbito de investigación como mecanismo o forma de plasmar lluvia de ideas para su posterior reutilización.

En el presente proyecto se utilizó la herramienta FreeMind con el fin de exponer y clarificar los conceptos, pasos, relaciones y operaciones en las investigaciones estudiadas. Esto permite tener una visión general y un tanto específica para la definición del procedimiento.

2 CREACIÓN DEL PROCEDIMIENTO

En la investigación documental anterior, se observó que muchos proyectos tienen como fin dar una solución a los problemas de recuperación de información ya que es el principal inconveniente para que la Web evolucione adecuadamente a lo que se ha denominado la Web Semántica. Con esto se espera que sus servicios permitan a los computadores ser capaces de interpretar y procesar automáticamente no sólo la información, sino los conocimientos circundantes en la Web.

Para mejorar la relevancia en la información recuperada, se han desarrollado varias técnicas y modelos que obtienen (de manera parcial) buenos resultados de acuerdo a las consultas realizadas por el usuario. Las investigaciones analizadas se centran en la técnica de Indexación semántica que proporciona una búsqueda más relevante, puesto que tiene en cuenta no solo las palabras clave en los documentos, sino también los conceptos asociados a ellos. Cada proyecto ha realizado un proceso o metodología para la extracción de los índices y sus resultados han sido favorables.

A pesar de los beneficios encontrados con la Indexación semántica, actualmente no se cuenta con un procedimiento general que pueda ser utilizado como guía para construir índices semánticos, cada uno de los proyectos analizados propone su propio procedimiento teniendo en cuenta diferentes aspectos, formas, modelos y demás criterios para construir sus índices.

La generación de un procedimiento también es importante para las nuevas aplicaciones en Web semántica, ya que muchos de estos proyectos utilizarán índices semánticos y necesitarán de una estructura con la cual se puedan basarse para construirlos. Por lo tanto es necesario plantear un procedimiento bien definido que permita guiar en la generación índices semánticos.

Para este fin, se tienen en cuenta diferentes aspectos de los proyectos que fueron analizados anteriormente, los cuales, son una base importante para construir dicho procedimiento general para crear índices semánticos.

Antes de comenzar con los elementos básicos de un índice semántico cabe recordar la definición de un índice: Un índice es una lista de frases (o palabras) ordenadas en una estructura de datos que permite acceder a los datos de manera

fácil y rápida [86]. Los índices crecen cada vez que la colección de documentos y páginas se extiende y son básicos en los buscadores tradicionales, pues gracias a la indexación, los usuarios pueden encontrar de manera ágil las respuestas a muchas de sus consultas.

A continuación se muestran las diferentes etapas que se llevan a cabo en la creación del procedimiento, para el cual se utilizó la metodología que se basa en la propuesta hecha por Miguel Ángel Niño [87]. El proceso seguido se describe en secciones así:

En la primera sección se determina el ámbito del procedimiento y la indexación tradicional así como los elementos que se deben tener en cuenta a la hora de construir un índice semántico. Luego se identifican las operaciones respectivas que se deben realizar con los elementos descritos y las relaciones que se presentan entre ellos, y posteriormente se muestra un mapa conceptual, el cual es un compendio de los pasos y las características encontrados en las investigaciones realizadas en este proyecto.

En la segunda sección, después de aplicar las abstracciones necesarias, se expone la definición del procedimiento para crear índices semánticos, teniendo en cuenta los pasos analizados y sus actividades, también se presenta en un diagrama de actividades que lo resume. Y posteriormente se presenta una descripción general, en forma de tabla, de los pasos y actividades con el lenguaje de representación matemática y los ejemplos correspondientes.

En la tercera sección se presenta una plantilla de instanciación del procedimiento que se realiza para este proyecto, incluyendo una descripción de las principales herramientas usadas en la implementación del mismo. Además se presenta la descripción de la evaluación que se llevará a cabo después de construido el índice semántico.

2.1 INDEXACIÓN TRADICIONAL

Para llevar a cabo el proceso de indexación semántica se debe realizar un pre-procesamiento que corresponde a la indexación tradicional, en la cual interfieren elementos básicos para extraer el índice de los documentos.

Este proceso inicia con el análisis léxico de los documentos donde se incluye la eliminación de signos de puntuación, guiones y se decide sobre el tratamiento de mayúsculas, nombres propios, y espacios en blanco.

Como segundo paso, la eliminación de palabras vacías, muy frecuentes y poco frecuentes permite reducir el número de términos con poco valor en la recuperación [88] de información como artículos, preposiciones, conjunciones, etc.

Posteriormente se pasa a la lematización, en la cual se eliminan prefijos y sufijos. En esta fase se extrae el lexema o raíz de cada término o palabra extraída del paso de análisis léxico, por ejemplo, el verbo sin conjugar o la palabra en singular de dicho término [89].

A continuación se produce la selección de términos mediante el cálculo de la frecuencia en los documentos, es decir, el número de veces que aparece un término en el contenido del documento. Posteriormente se construyen los vectores de pesos de cada término Índice. En la Figura 6 se observa el proceso a seguir en la indexación.

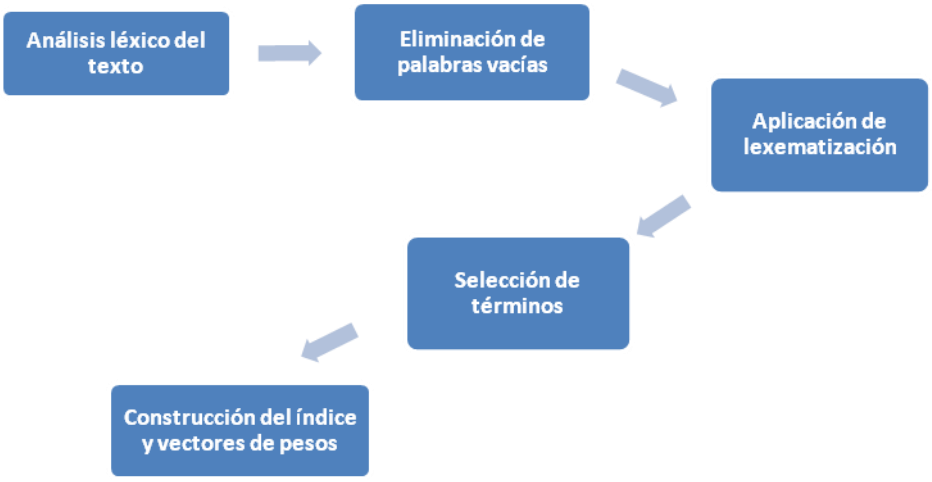

Figura 6. Indexación tradicional

2.2 ELEMENTOS IMPORTANTES PARA CONSTRUIR UN ÍNDICE SEMÁNTICO

Sabemos que un índice semántico se enfoca en asociar los conceptos con los términos o palabras en las páginas Web. Con ello se busca encontrar patrones en los datos no estructurados (documentos sin descriptores, como palabras clave o etiquetas especiales) y usar los patrones de búsqueda en una mejor clasificación de los datos y precisión en la recuperación de información. Para la construcción de éstos, es necesario tener en cuenta ciertos elementos, que nos van a permitir

construir de una forma más fácil y rápida, nuestro índice semántico. Los elementos son los siguientes:

Términos	Un término es una palabra que se encuentran en uno o varios documentos. Estos se comparan directamente con los términos de las consultas realizadas por el usuario con el fin de determinar los documentos relacionados a la consulta.
Términos índice	Son términos candidatos para la indexación que se encuentran en los documentos y además representan sustantivos, verbos o adjetivos que pueden ser introducidos en una consulta de usuario.
Consultas	Corresponden a la frase o frases que digita un usuario de acuerdo a la necesidad de información que desee suplir.
Lista de términos	Es importante construir un vocabulario con determinadas palabras, este permitirá detectar información más relevante a la hora de hacer la búsqueda
Conceptos candidatos	Se trata de un conjunto de conceptos o significados posibles que son seleccionados por su relevancia a la hora de realizar la consulta y están estrechamente relacionados a los términos de la consulta.
Frecuencia ponderada	A los términos índice de las consultas normalmente se les asigna valores numéricos, los cuales corresponden a la frecuencia con la que los términos aparecen en los documentos, colecciones o subconjunto de colecciones de documentos, para que estos sean considerados relevantes en una consulta. La frecuencia de un Término Indice K_i en un documento d_j, se representa así: $freq_{i,j}$. y se refiere al número de veces que aparece el término índice en el texto del documento.
Representatividad	Cuando un término aparece muchas veces en un documento es indicativo de que éste es representativo en el contenido del mismo, pero siempre y cuando no aparezca con una frecuencia muy alta en todos los documentos [90].
Peso de un término	El peso del término se calcula en función de su frecuencia de aparición en los documentos, lo cual indica la importancia del término en el documento. Se representa en un vector o matriz así: W_i, que significa el peso del término *i* ó $W_{i,j}$ que sería el peso del término *i* en el documento *j*.
Tesauro	"Es una colección de vocabularios seleccionados (términos preferidos o descriptores), con enlaces hacia términos, sinónimos, equivalentes, genéricos, específicos o relaciones"[91]. Los tesauros son utilizados en conjunto con algoritmos, para convertir el lenguaje natural de los documentos en un lenguaje controlado. Igualmente para la recuperación de documentos que son relevantes a la consulta del usuario, gracias a las relaciones entre conceptos que se encuentran en éstos. Los tesauros en el contexto de la recuperación de la información tiene dos objetivos fundamentales los cuales son: controlar el vocabulario lo cual significa identificar dentro de un campo semántico todos los conceptos que son representados por un término, lo cual hace posible minimizar la perdida de información en las búsquedas realizadas en un sistema documental automatizado

	[95]. El segundo objetivo es conocer todos los términos relacionados con un concepto determinado, lo que ayuda a añadir más términos adecuados para enriquecer tanto los análisis del contenido de los documentos como las estrategias de búsqueda para recuperación de información [96].
Ontologías	Una ontología es "una especificación explicita de una conceptualización compartida" [28], en otras palabras, proporciona una estructura y contenidos específicos de una abstracción del mundo que se desea representar para algún propósito. En una ontología se define un vocabulario controlado de un área, mediante conceptos y relaciones (específicas y de composición) que los conectan entre sí.
Vectores	Un vector está compuesto por los pesos de los términos indexados en una consulta. Se pueden representar como una tupla así: q = (W_1, W_2,… W_n), donde q representa la consulta escrita por el usuario o el documento analizado y n corresponde al n-ésimo peso del término Índice de la consulta o documento.

Tabla 4. Elementos para la construcción de índices semánticos

2.3 IDENTIFICACION DE OPERACIONES Y RELACIONES

Teniendo los elementos anteriormente descritos para construir un índice semántico, se puede hablar de operaciones y relaciones entre ellos, las cuales representan su interacción en el proceso de indexación semántica. A continuación se describe lo más importante encontrado hasta el momento.

En la indexación es importante analizar el texto del documento para determinar qué términos pueden utilizarse como términos índice. Es necesario mencionar que se comienza por un documento de texto completo, con lo cual se conseguirá al final del proceso un conjunto numeroso de términos Índice que lo representan[88]. Otro aspecto importante a tener en cuenta es la técnica de generación de un espacio conceptual, el cual proporciona las bases para definir un método de representación y cálculo de similitud. Además, se propone la presentación de cada documento mediante un vector, cuyos componentes son los pesos asociados a los conceptos utilizados [92]. Para realizar el proceso de indexación mediante estos aspectos hay que tener en cuenta las siguientes fases y sus operaciones (actividades) según la Figura 7.

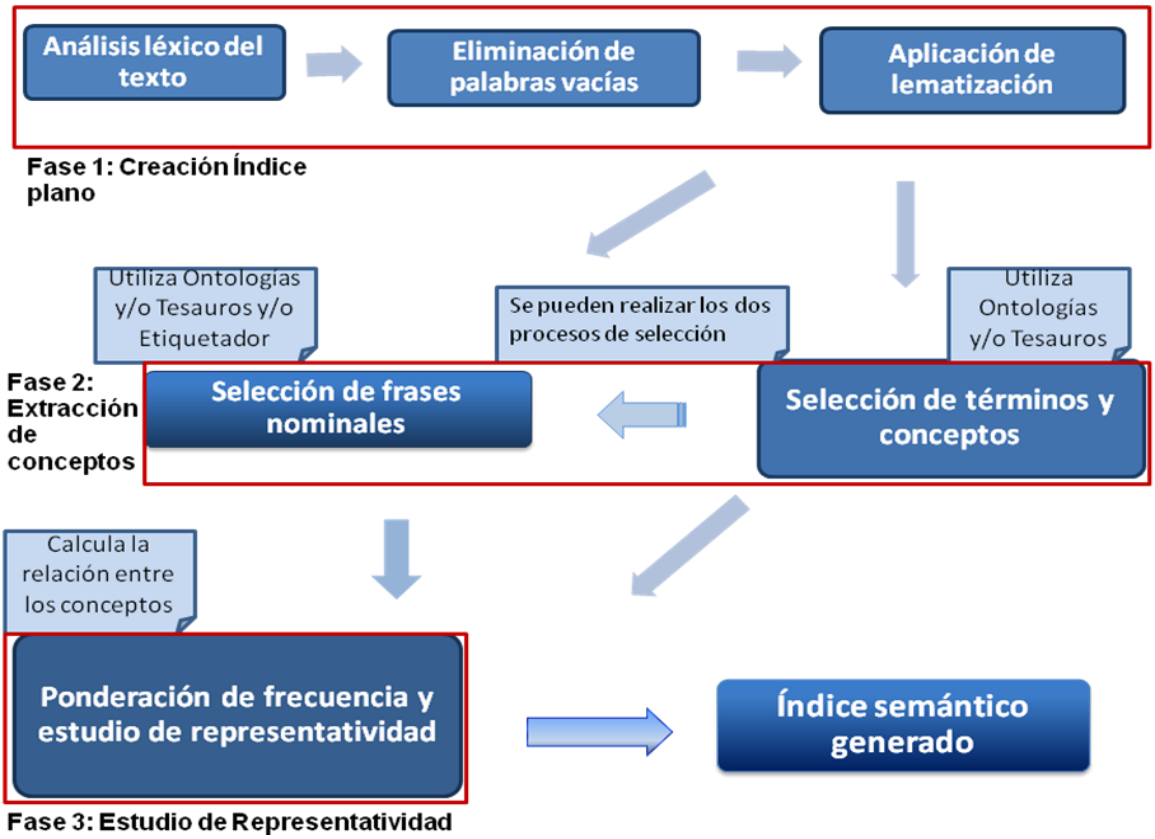

Figura 7. Indexación semántica

- **Análisis léxico, Eliminación de palabras vacías y Lematización:** Como se expresó anteriormente, su objetivo es convertir la cadena de entrada en un conjunto de palabras y determinar el tratamiento que se realizara sobre números, signos de puntuación, etc. A continuación, La eliminación de palabras vacías y la Aplicación de Lexematización, se encargan de reducir el número de términos y encontrar las variaciones morfológicas, lo cual produce un único término a partir de varias palabras [93]. Este proceso se realiza de manera similar en la indexación tradicional.

- **Extracción de términos candidatos:** Como en la indexación tradicional, se debe aplicar una selección de términos candidatos, es decir, los términos relevantes en los documentos y que son opcionados para extraer sus conceptos.

- **Selección de frases nominales:** Este proceso es general para varios proyectos puesto que es muy útil en la extracción de conceptos y relaciones entre los términos. Una frase nominal es un conjunto de términos y

constituye en sí misma un nuevo término (concepto) de indexación[92] Este paso permite, no solo extraer los conceptos de los términos aislados, sino también diferencias en sus relaciones semánticas entre si. Para esto se utilizan herramientas como Tesauros y Ontologías para comparar las relaciones jerárquicas de estas con las frases extraídas.

- **Ponderación de frecuencia y estudio de representatividad:** Finalmente, se realiza un estudio o evaluación de representatividad de los conceptos y sus relaciones en los documentos. Si se realizó una ponderación de frecuencia previa, este estudio se basará en ella. En este proceso, la medida de similitud y ponderación de frecuencia con TF-IDF, son los cálculos más usados para representar los conceptos más relevantes en los textos respecto a una consulta realizada por el usuario. Con esto, se pueden retornar los documentos que el usuario realmente necesita.

- **Análisis de co-ocurrencia**: Otro análisis que se lleva a cabo en algunas investigaciones es el análisis de concurrencia, para el cual, se guarda la frecuencia de la frase nominal, que será usada para calcular los pesos de cada una en los documentos. El análisis de co-ocurrencia se calcula basándose una función de similitud asimétrica.

- **Construcción del índice semántico:** Después del estudio de representatividad de conceptos y/o análisis de co-ocurrencia, se finaliza la indexación semántica, la cual genera una estructura de datos con el índice creado.

2.4 MAPA CONCEPTUAL DE PROCEDIMIENTOS PARA CREAR INDICES SEMANTICOS

Para clarificar los conceptos, relaciones y operaciones anteriormente descritos, se realizó un mapa conceptual con toda la información escogida para la construcción de un procedimiento genérico que permita crear índices semánticos.

El mapa conceptual se realizó en la herramienta FreeMind [85] con el propósito de mostrar cada paso con sus posibles acciones y/o actividades teniendo en cuenta las investigaciones estudiadas, los autores que manejan el tema y la literatura existente, al respecto, en la Web.

La Figura 8 muestra el mapa en resumen, la expansión del mismo se deja como anexo (archivo llamado “Mapa Conceptual _Procedimiento Índice S.pdf”).

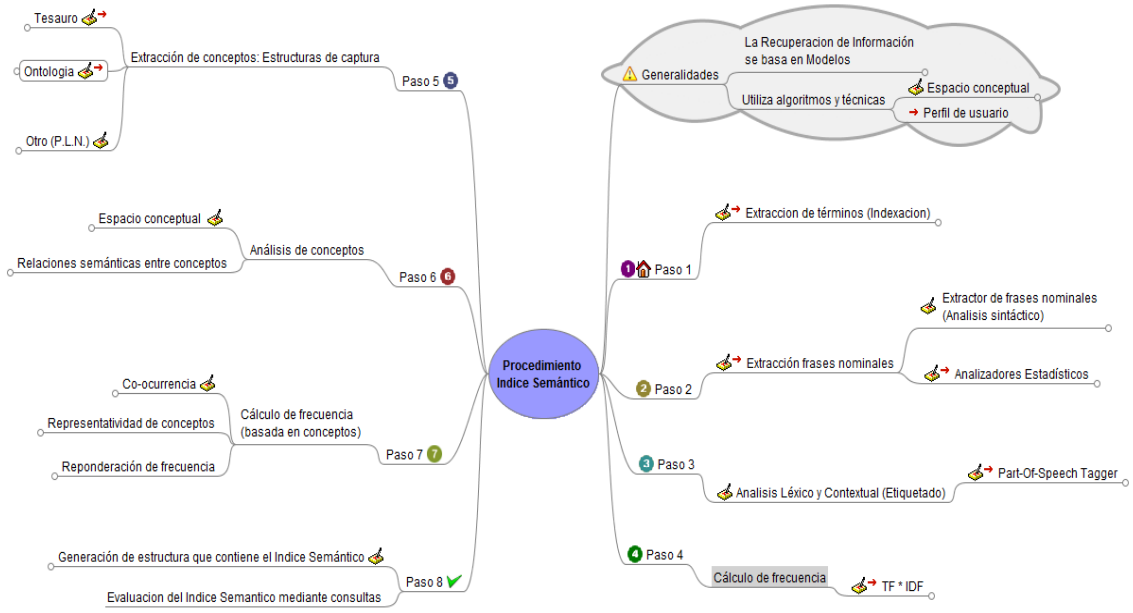

Figura 8. Mapa Conceptual del procedimiento para crear índices semánticos

Inicialmente se encuentra una nube que representa generalidades de los procesos encontrados sobre la indexación semántica en las investigaciones. Posteriormente se describen 8 pasos generales, los cuales presentan unas posibles variaciones en la construcción del índice, y se despliegan en el mismo mapa (observadas en el anexo). Estas variaciones dependen del implementador, las herramientas con las que se cuente y la necesidad de realizar cálculos o subprocesos más específicos que permitan una indexación semántica con resultados más relevantes.

La realización de este mapa (en conjunto con la previa investigación) permitió obtener un mayor análisis, comprensión y abstracción para definir los pasos de un procedimiento genérico para crear índices semánticos basados en ontologías de dominio.

2.5 PASOS PARA LA CREACIÓN DE ÍNDICES SEMÁNTICOS

La creación de índices semánticos se lleva a cabo siguiendo una serie de pasos que dependen de las necesidades del implementador y ámbito en el que se utilice la indexación. En este caso, se describen los pasos y actividades generales que se pueden seguir para realizar una indexación semántica en el área de la recuperación de información. La descripción y decisión del procedimiento genérico, se tomó del análisis de las investigaciones previamente descritas en este documento.

1. Para la creación de un índice semántico se inicia con la construcción de un índice plano[10], el cual se puede generar mediante algoritmos de pre-procesamiento de documentos (permiten preparar los contenidos de los archivos para obtener el índice), o utilizando algunas de las herramientas que apoyan este proceso (como Lucene) y que siguen el proceso a continuación :

    1.1 Conversión de los textos a un formato plano (.txt) de tipo ASCII que consiste en la eliminación de las etiquetas HTML.

    1.2 Utilizar alguna herramienta de proceso de lenguaje natural como tokenizadores, los cuales separan la secuencia de caracteres de los textos para generar una secuencia de palabras que posteriormente pueden ser los términos índice del documento. Remoción de caracteres especiales, como "/ \ - : ?¿;)(&#", entre otros.

    1.3 Remoción de palabras vacías de paso[11] (StopWords removal) como pronombres, partículas interrogativas y ciertas preposiciones. Entre los artículos están por ejemplo: "un, la, los, el, ellos", las partículas interrogativas son: "what, when, who, how, where", entre otros. Algunas preposiciones son: "con, desde, entre, hasta, por, según", etc.

    1.4 Lematizadores (Stemmer), que permiten la reducción de las palabras a su forma básica o raíz, por ejemplo, eliminando las partes no esenciales de los términos como prefijos y sufijos o derivando las palabras en plural a su raíz en singular.

Con el proceso descrito anteriormente podemos obtener un índice plano que puede ser optimizado con el fin de generar un índice semántico, para lo cual se tendrán en cuenta los siguientes pasos:

2. Extracción de frases nominales (opcional). Se dividen los documentos en párrafos y luego en frases independientes llamadas frases nominales, las cuales serán utilizadas, junto con los términos generados en el paso 1, para la extracción de conceptos candidatos. La extracción de estas frases es opcional y depende del implementador realizarlo. Este paso se realiza por medio de un etiquetador o analizador léxico- sintáctico, lo cual permite extraer nombres simples o compuestos de los documentos. Por ejemplo: Gerente (simple), director de operaciones (compuesto).

[10] Un índice plano se basa en los términos, no en los documentos.
[11] Extraído de http://basesdatos.uc3m.es/fileadmin/Miembros/Mayte/Sesion2B_2.pdf

3. Se debe decidir (opcional), si realiza o no una indexación semántica basada en la construcción de un espacio conceptual, donde se procesan los términos y las frases nominales extraídas con el fin de formar un conjunto n-dimensional de vectores de conceptos.

    3.1. Si escogió la opción de realizar un espacio conceptual, el siguiente paso es la ponderación del mismo teniendo en cuenta cálculos como Ratio de Información y Cantidad de información[52], Medidas de similitud[13] o matriz de Co-ocurrencia[14]. Este paso corresponde al estudio de representatividad de conceptos.

4. Si no escogió la realización de un espacio conceptual, continua con el paso de Ponderación de la frecuencia de los términos obtenidos (incluyendo las frases nominales, si se extrajeron). Al realizar este paso se obtienen los cálculos de frecuencia de términos, generalmente, según el modelo vectorial TF-IDF (frecuencia de términos – frecuencia inversa del documento)[12][94], lo cual permite definir el número de veces que aparece cada término en los documentos.

5. Extracción de conceptos candidatos: para llevar a cabo este paso se puede elegir en la utilización de una ontología, un tesauro o el uso de ambas herramientas. Esto dependerá de los objetivos del índice a construir y la funcionalidad especial de cada opción, además del dominio o ámbito en que se trabaje. Para esto se debe tener en cuenta la funcionalidad de los tesauros y las ontologías por separado.

    - ¿**Cuándo usar tesauros?:** los tesauros se utilizan generalmente cuando se manejan varios dominios. Se debe tener en cuenta que este se encarga de agrupar un conjunto de palabras de un idioma particular, las cuales representan un concepto del conocimiento humano. Además Los tesauros se pueden usar para ayudar a convertir el lenguaje natural de los documentos en un lenguaje controlado, ya que representan el contenido de éstos [95] por medio de unas relaciones semánticas y genéricas. Estas relaciones son básicamente: Las Jerárquicas (estructuras todo/parte), de equivalencia (Sinonimia, Hiponimia, Hiperonimia) y asociativas (ayudan a reducir la poli-jerarquía entre los términos).

[12] Este peso es una medida estadística para evaluar la importancia de un termino en el documento, colección o corpus. Extraído de http://en.wikipedia.org/wiki/Tf-idf.

- **¿Cuándo usar ontologías?** [96]: las ontologías se recomienda utilizarlas cuando se maneje un dominio o área del conocimiento específico. Estas presentan un desarrollo semántico más profundo para las relaciones del tipo clase/subclase, y para las relaciones cruzadas [97], que los tesauros. Dan la posibilidad de trabajar en sistemas heterogéneos, al describir formalmente objetos en el mundo sus prioridades y las relaciones entre ellos. Las ontologías añaden valor a los tesauros tradicionales a través de una semántica más especializada, así como unas enriquecidas relaciones entre clases y conceptos.

Se realiza una comparación entre los términos candidatos y conceptos de la ontología de dominio o tesauro, con lo cual se define un vocabulario de palabras permitidas para estudiar los conceptos representativos del contenido de los documentos. Como resultado de este paso es la selección de los conceptos candidatos a indexar.

6. Extracción de conceptos utilizando las dos herramientas: es importante mencionar que se puede optar por la utilización de las dos herramientas para asegurar un índice semántico mejor establecido para ciertos dominios, en el orden que se requiera. Sin embargo la mayoría de investigaciones que utilizan las dos herramientas, realizan el mapeo en el siguiente orden.

    - **Tesauro**: En primer lugar se utiliza un tesauro para extraer los conceptos candidatos, realizando un mapeo con los términos anteriormente extraídos. Este mapeo permite construir un conjunto de conceptos diferenciado por varios dominios, puesto que los tesauros generalmente manejan varios entornos o ámbitos del aprendizaje.
    - **Ontología**: Luego de realizar el mapeo de términos y conceptos con el tesauro, se procede a realizar un mapeo de conceptos candidatos con los conceptos de la ontología de dominio. En este paso se estudia la representatividad de los conceptos de acuerdo al dominio y se realiza un cálculo de frecuencia de estos conceptos mediante ponderación de frecuencias (TF-IDF), funciones de similitud acumulativa y/o distancia semántica entre conceptos [1].

    Al realizar estos pasos se obtienen los conceptos más representativos de los documentos y se genera el índice semántico de acuerdo a ello.

7. Estudio de la representatividad de los conceptos en el contenido de las páginas (documentos). Esto se realiza a través de la frecuencia ponderada calculada en el paso 4 y las relaciones que tienen los conceptos generados

con los de la ontología o tesauro. Si no se realizó una ponderación previa de términos, se procede a calcular la frecuencia de los conceptos en este paso.

El cálculo de la ponderación se puede realizar utilizando alguno de los siguientes cálculos:

- Cálculo de la distancia semántica entre conceptos. Esta medida se puede utilizar por ejemplo para calcular coeficientes representativos entre los conceptos como en [1].
- Función de similitud. Estas medidas se pueden realizar de variadas formas, las más utilizadas y referenciadas por los expertos son: La medida de similitud planteada por Resnik [34], en la que se calcula la similitud entre dos conceptos de una taxonomía. La medida de Mazuel [35], que se enfoca en los puntos en común en las relaciones semánticas de una ontología. También se pueden utilizar medidas desarrolladas en el contexto de la inteligencia artificial como la que define Wei Song [39], la cual se utiliza con ontologías. Además de estas funciones se pueden utilizar otras como las definidas en [36-38] o utilizadas en [7, 8, 13] para medir la similitud entre conceptos en un tesauro y/u ontología.
- Matriz de Co-ocurrencia. Se puede llevar a cabo la construcción de esta matriz con los términos y frases nominales (si se extraen), por ejemplo basándose en medidas probabilísticas entre otras [13] [14].
- Cálculo basado en TF-IDF. Esta medida es una de las más utilizadas en la recuperación de información y motores de búsqueda, no solo para los índices semánticos sino también para la indexación tradicional [8, 98, 99].

La ponderación basada en el cálculo TF-IDF se representa así:

$$TF = frecuencia\ del\ concepto$$

$$IDF = \log\frac{d}{df_j}\quad Frecuencia\ inversa\ del\ documento.$$

- $d = número\ total\ de\ documentos$
- $df_j = número\ de\ documentos\ que\ contienen\ el\ concepto\ j$

8. Construcción de la matriz de conceptos. Como resultado del paso anterior, se genera una estructura de datos, generalmente una matriz, la cual contiene los pesos de los conceptos más representativos en los documentos. Esto con el fin de almacenar el índice semántico para posteriores consultas.

9. Evaluación del índice semántico. En este paso se realizan una serie de consultas utilizando el índice creado para verificar la relevancia de las páginas resultantes.

   9.1 Realización de consultas para verificar la relevancia de resultados obtenidos con el índice semántico.

   9.2 Depurar la Consulta (opcional). Se puede realizar un procesamiento a la consulta de manera que se obtengan conceptos relevantes en ella. Esto se hace por medio un mapeo de los términos de la misma con los conceptos de la ontología (dominio específico).

   Si no encuentra los conceptos de la consulta en la ontología, se realizará un mapeo semántico con un tesauro, un vocabulario controlado o una base de datos léxica que es de dominio general. Al realizar esta comparación se extraerán los conceptos que se asemejen (sinónimos) a los términos de la consulta y así, se comparan con los documentos relevantes de acuerdo al índice construido.

   9.3 Para observar los resultados que se obtienen con la utilización del índice se realizan unas pruebas exhaustivas, las cuales deben ser evaluadas con algunas medidas como el índice MAP [100], Precisión-recuerdo [101], estadísticas Kappa [29, 102]. Con estas medidas se verifica la precisión y relevancia de los resultados obtenidos.

   9.4 Realizar una retroalimentación al procedimiento si no se logran los resultados esperados, esto significa que, si al verificar la precisión de los resultados obtenidos con el índice no son los requeridos, se revisan los procesos (pasos) realizados en la construcción del mismo para mejorar lo que se considere necesario.

A continuación se presenta un diagrama de actividades, Figura 9, el resumen de los pasos y/o actividades a realizar en el proceso de indexación semántica descritos anteriormente. El procedimiento para crear los índices puede tomar diversos caminos. La decisión depende de las necesidades del implementador, dominio u objetivos del mismo, como se mencionó anteriormente.

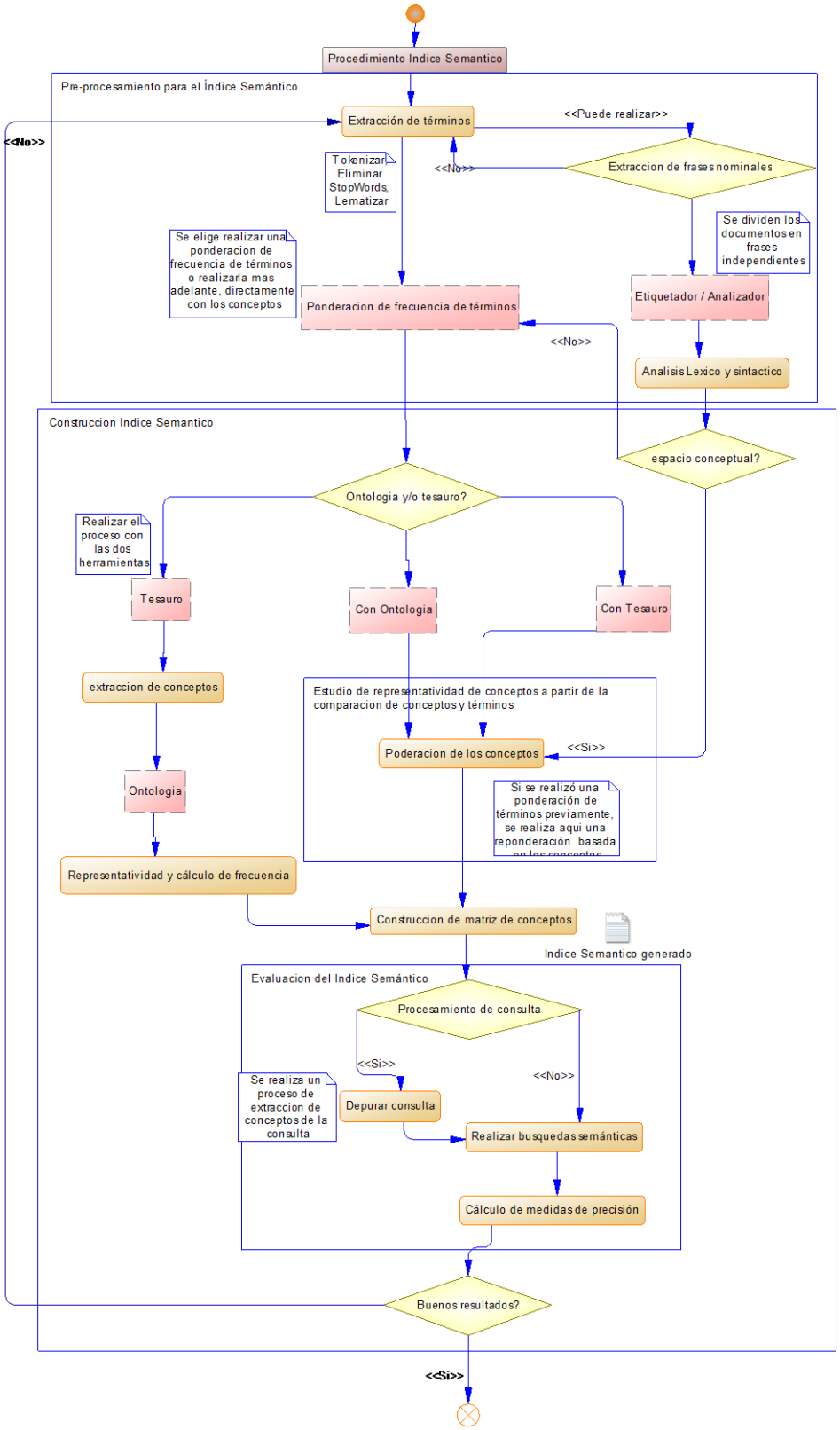

Figura 9. Diagrama de actividades: Procedimiento para crear Índices Semánticos

2.6 DEFINICIÓN DEL PROCEDIMIENTO A IMPLEMENTAR

El procedimiento anteriormente mostrado se puede definir de forma matemática, con el fin de formalizar el procedimiento propuesto.

2.6.1 LENGUAJE DE REPRESENTACIÓN

La definición del lenguaje se muestra mediante una representación matemática en la recuperación de información tradicional. Esta representación se utiliza en la definición de los pasos del procedimiento para crear índices semánticos y se consigna en las casillas de la última columna de la Tabla 5.

Pasos		Descripción	Ejemplo	Representación
Paso 1	**Actividad**			
construcción de índice plano[13]	1.1 Eliminación de etiquetas HTML (obligatorio)	Conversión de los textos a un formato plano (.txt) de tipo ASCII que consiste en la eliminación de las etiquetas HTML	Las etiquetas: <title>, <html>, <head>, <body>, etc., se eliminan del contenido de la pagina, quedando el texto como un archivo ".txt"	$Cd = [d_1, d_2, \dots d_n]$ $se\ convierte\ en\ D$ n = número de documentos de la Web D = documento de tipo .txt en el que se guardan los documentos sin etiquetas
	1.2 Tokenización (obligatorio)	Utilizar alguna herramienta de proceso de lenguaje natural como tokenizadores. Remoción de caracteres especiales.	Se genera una secuencia de palabras. Eliminan caracteres como / \ - : ?¿;)(&#, etc.	$D = [d_1, d_2, \dots\ d_j, \dots d_n]$ $d_j = T_1, T_2 \dots T_n$ T= términos
	1.3 Eliminacion de Stopwords (obligatorio)	Remoción de palabras vacías o de paso[14] (StopWords removal) como pronombres, partículas interrogativas y ciertas preposiciones.	Entre los artículos están por ejemplo un, la, los, el, ellos, las partículas interrogativas son: what, when, who, how, where, entre otros. Algunas preposiciones son: con, desde, entre, hasta, por, según, etc.	$d = [T_1, T_2, \dots T_n]$ $SW = [T_1, T_2 \dots T_j]$ $(d\ U\ SW) - (d \cap SW) = dd = [T_1, T_2, \dots T_m]$ n = numero de términos en un documento. SW = conjunto de palabras vacías j = numero de palabras vacías. m = número de términos en el documento sin StopWords.

[13] Un índice plano se basa en los términos, no en los documentos.

[14] Extraído de http://basesdatos.uc3m.es/fileadmin/Miembros/Mayte/Sesion2B_2.pdf

				dd = documento sin palabras vacías.
	1.4 Lematización (obligatorio)	Lematizadores, que reducen de las palabras a su forma básica o raíz	Por ejemplo, se eliminan las partes no esenciales de los términos como prefijos y sufijos o derivando las palabras en plural a su raíz en singular.	$d = [T_1, T_2, \dots T_n]$ $T = [t, t_2 \dots t_j]$ Comparar las palabras raíces y los términos para llega a: $dl = [T_1, T_2, \dots T_m]$ n = numero de términos en un documento. j = numero de lemas (palabras raíces). m = número de términos raíces en el documento.
Paso 2				
Extracción de frases nominales	2.1 División de documentos en frases (opcional)	Se dividen los documentos en frases independientes que contienen un sustantivo como núcleo (frases nominales)	"orquídea siberiana"	$d = [fn_1, fn_2, \dots fn_n]$ fn_i = frase nominal
Paso 3				
DECISIÓN de construcción de espacio conceptual	3.1 Decidir la creación de espacio conceptual (obligatorio)	Se debe escoger la opción de realizar un espacio conceptual o continuar con la ponderación de términos y/o frases nominales		
Paso 4				
Ponderación de términos	4.1 ponderación de términos y frases extraídos (opcional)	Se realiza el cálculo de frecuencia de términos obtenidos y frases nominales (si se extrajeron)	Número de veces que se encuentra un término o frase en los documentos	
Paso 5				
DECISIÓN en la Extracción de conceptos candidatos.	5.1 Elección de la herramienta semántica (obligatorio)	Se puede escoger entre varias opciones: Utilizar un tesauro, una ontología, o las dos herramientas. En este proyecto se utiliza una Ontología de dominio. La Ontología identifica relaciones de superclases, subclases	Ejemplo de relación de composición: "una flor es componente de una planta", por lo tanto existe relación al buscar "flor" y "planta"	Ontología = ***O***

		(hipónimos, hiperónimos), sinónimos y relaciones de inferencia		
	5.2 Comparación de términos y conceptos (obligatorio)	Se define un vocabulario de palabras permitidas para estudiar los conceptos representativos del contenido de los documentos. Un concepto tiene asociados varios términos y así un documento tendría asociado un vector de conceptos.	Ejemplo, en un dominio de plantas: Concepto: estructura de planta. Términos: tallo, célula, flor. Ejemplo, en un dominio de Administración de empresas. Concepto: director de operaciones Términos: director ejecutivo, CEO, ejecutivo de negocio.	$C = [T_1, T_2 \dots T_k]$ $d_j=(W_{1,j}, W_{2,j}, \dots, W_{t,j})$ d_j tiene varios términos que se comparan con *C*, que es el conjunto de conceptos.
Paso 6				
Utilización de dos herramientas semánticas	6.1 Utilizar dos herramientas: tesauro y ontología (opcional)	Se utilizan las dos herramientas semánticas para la extracción de conceptos		
Paso 7				
Estudio de representatividad de los conceptos.	7.1 Ponderación de frecuencias (obligatorio)	Se mide la distancia semántica entre conceptos y ponderan la frecuencia de los mismos en los documentos. La ponderación puede ser utilizando una función de Similitud: [103], distancia semántica: [104], creación de matriz de co-ocurrencia [14] o un cálculo basado en TF-IDF[8, 98, 99]	Al utilizar una ponderación de frecuencia con TF-IDF (modelo vector), se modifica tomando en cuenta los conceptos en vez de los términos y se estudia la frecuencia de ellos en los documentos.	$TF = Frecuencia\ del\ concepto$ Frecuencia inversa $IDF = \log \frac{d}{dfj}$ $Fórmula\ utilizada$ $TF * IDF$
Paso 8				
Construcción de la matriz o estructura de datos	8.1 Estructura de datos del índice semántico (obligatorio)	Se construye la matriz de conceptos por documentos. En la matriz se encuentra la frecuencia de los conceptos en cada documento	Por ejemplo el concepto "*Flor*" tiene frecuencia 10 en el documento "*Guía de plantas*". En la matriz se calcula esta frecuencia y se determinan los pesos de cada concepto en los documentos indexados.	$SI = [d_{11}, d_{12}, \dots d_{1n}]$ $[d_{21}, d_{22}, \dots d_{2n}]$ $[d_{m1}, d_{m2} \dots d_{mn}]$ $SI = [m][n]\ conceptos \times documentos$
Paso 9				
Evaluación del índice semántico generado	9.1 Consultas de prueba (obligatorio)	Realización de consultas para probar los resultados que	Las consultas para este proyecto se realizan en ingles, pues la ontología está en dicho idioma.	Consulta = Q[ti] (términos de la consulta)

		retorna el índice semántico construido.		
	9.2 Depuración de la consulta (opcional)	Se realiza un mapeo de los términos de la consulta con los conceptos de la Ontología 5.2.1. Si no se encuentran los conceptos en la ontología se procede a buscarlos en un tesauro (dominio general). 5.2.2 Se compara con el índice construido y se extraen los conceptos asociados.	Consulta: “estructura de la célula en una planta”. Conceptos encontrados: “célula de la planta”, “estructura de la planta”. Consulta: “florescencia ” Concepto relacionado encontrado en el tesauro: “Flor”.	Q[ti] se compara con los conceptos de la matriz SI[i][j] Q[ti] se compara con el tesauro
	9.3 Medidas de precisión (obligatorio)	Se realizan medidas de precisión como: Estadísticas Kappa, proporción recuerdo, índice Map.	Como ejemplo de las medidas, se calculan promedios de documentos relevantes vs. Documentos recuperados.	$k = \frac{(P(A) - P(E))}{1 - P(E)}$ $Precisión = \frac{DRelRec}{tot\ DRec}$ $Recall = \frac{DRelRec}{tot\ DRel}$ DRelRec = Documentos Relevantes Recuperados
	9.4 Revisión y mejoramiento del índice creado (opcional)	Si los resultados de la evaluación muestran poca relevancia de acuerdo a las consultas, se procede a una revisión y mejoramiento del índice creado	Volver al paso 1 del procedimiento para revisar el proceso y continuar hasta el paso 7 revisando las posibles falencias	

Tabla 5. Procedimiento para crear índices semánticos.

2.7 PLANTILLA DE INSTANCIACIÓN

A continuación se presenta una plantilla, la cual es una herramienta que puede ser utilizada por el desarrollador para realizar la aplicación del procedimiento. Esta plantilla presenta los pasos que se deben seguir así como sus respectivas actividades.

Pasos		¿Se realiza?		Observaciones y/o Comentarios
Paso 1	**Actividad**	**Si**	**No**	
construcción de índice plano[15]	1.1 Eliminación de etiquetas HTML (obligatorio)			
	1.2 Tokenización (obligatorio)			.
	1.3 Eliminación de Stopwords (obligatorio)			
	1.4 Lematización (obligatorio)			
Paso 2				
Extracción de frases nominales	2.1 División de documentos en frases (opcional)			
Paso 3				
DECISIÓN de construcción de espacio conceptual	3.1 Decidir la creación de espacio conceptual (obligatorio)			
Paso 4				
Ponderación de términos	4.1 ponderación de términos y frases extraídos (opcional)			
Paso 5				
DECISIÓN en la Extracción de conceptos candidatos.	5.1 Elección de la herramienta semántica (obligatorio)			
	Ontología			
	Tesauro			
	5.2 Comparación de términos y conceptos (obligatorio)			
Paso 6				
Utilización de dos herramientas semánticas	6.1 Utilizar dos herramientas: tesauro y ontología (opcional)			
Paso 7				
Estudio de representatividad de los conceptos.	7.1 Ponderación de frecuencias (obligatorio)			
	Distancia semántica			
	Medida de similitud			
	Matriz de co-ocurrencia			
	Calculo TF - IDF			
Paso 8				
Construcción	8.1 Estructura de			

[15] Un índice plano se basa en los términos, no en los documentos.

de la matriz o estructura de datos	datos del índice semántico (obligatorio)			
Paso 9				
Evaluación del índice semántico generado	9.1 Consultas de prueba (obligatorio)			
	9.2 Depuración de la consulta (opcional)			
	9.3 Medidas de precisión (obligatorio)			
	9.4 Revisión y mejoramiento del índice creado (opcional)			

Tabla 6. Plantilla de instanciación

3 IMPLEMETACIÓN DEL INDICE SEMANTICO

Al realizar los respectivos análisis para la definición del índice semántico, se procede a realizar la implementación del mismo. Para esto se inicia llenando la plantilla en donde se especifican los pasos, las actividades y los comentarios respectivos de cómo se llevo a cabo el proceso de construcción del índice semántico. Posteriormente se presentan cada una de las fases donde se describe el proceso del desarrollo del procedimiento. Para las fases se utilizó la metodología de desarrollo de software UP Ágil.

3.1 PLANTILLA DE INSTANCIACIÓN

Para la abstracción de la información sobre el dominio (plantas) que se maneja en la creación del índice semántico, se utiliza Delicious[16], un Servicio de la Web 2.0, característico del marcado social o en ingles Social Bookmarking. Delicious es un servicio de gestión de marcadores sociales en la web, el cual permite agregar los marcadores que clásicamente se guardan en los navegadores y categorizarlos con un sistema de etiquetado denominado folkcsonomias (tags). Este sitio no solo puede almacenar direcciones de sitios web si no que también permite compartirlos con otros usuarios de Delicious[17]

La importancia del servicio radica en que permite hacer marcado colaborativo, el cual hace una jerarquización de los enlaces dependiendo de la cantidad de usuarios que los anexen, haciendo que estos recursos aumenten su relevancia por sí mismos. Cabe mencionar que maneja documentos estructurados y gracias a todo lo anterior, se obtiene una gran relevancia en los resultados del índice. Para la obtención de las URL con la información sobre plantas se utilizó el método GetUrlsByDelicious, proporcionado por ScottWater.Boss.

Para esta solución se especifica que el número de URLs recuperadas en Delicious sean 50 URLS, la cuales son obtenidas la primera vez que se carga el índice semántico.

La Tabla 7 presenta la plantilla de instanciación con los elementos que se tienen en cuenta para la construcción del índice semántico de este proyecto.

Pasos		¿Se realiza?		Observaciones y/o Comentarios
Paso 1	Actividad	Si	No	

[16] http://www.delicious.com/

[17] Extraído de http://tecnologiaedu.uma.es/materiales/web20/archivos/cap8Web20_marcadores_sociales.pdf

construcción de índice plano[18]	1.1 Eliminación de etiquetas HTML (obligatorio)	**X**		
	1.2 Tokenización (obligatorio)	**X**		.
	1.3 Eliminación de Stopwords (obligatorio)	**X**		
	1.4 Lematización (obligatorio)	**X**		
Paso 2				
Extracción de frases nominales	2.1 División de documentos en frases (opcional)		**X**	
Paso 3				
DECISIÓN de construcción de espacio conceptual	3.1 Decidir la creación de espacio conceptual (obligatorio)	**X**		Al llegar a este paso se decidió no realizar la creación del espacio conceptual y continuar por el otro camino en el procedimiento, es decir por la extracción de conceptos, mediante otras técnicas.
Paso 4				
Ponderación de términos	4.1 ponderación de términos y frases extraídos (opcional)		**X**	
Paso 5				
DECISIÓN en la Extracción de conceptos candidatos.	5.1 Elección de la herramienta semántica (obligatorio)			Para este paso se escogió una ontología de dominio particular. En la actual red podemos encontrar ontologías en diferentes dominios: médicos, científicos, informáticos, biomédica, educativos y demás. En el caso de nuestra aplicación fue necesaria una ontología educativa. Para realizar el proceso de selección de la ontología, se hicieron diferentes búsquedas de ontologías educativas, sobre el dominio de las ciencias naturales, encontrando pocas ontologías que manejaran este tema. La ontología escogida es PlantOntology: **Plantontology [105]** Su principal objetivo es desarrollar vocabularios controlados (ontologías) que describen las estructuras de las plantas así como su crecimiento y etapas de desarrollo, proporcionando un marco de semántico de las consultas a través de especies significativas entre las bases de datos. Ver Anexo D. Es importante mencionar que la ontología es manejada en lenguaje OWL y el manejo de esta misma fue Protege 3.1.1. Después de haber seleccionado la ontología, procedemos a utilizar los diferentes métodos para hacer el tratamiento respectivo de la misma, para tal caso se utilizó la librería de OWLAPI la cual permitió obtener las súper clases, subclases y los respectivos axiomas de la ontología.
	Ontología	**X**		
	Tesauro		**X**	
	5.2 Comparación de términos y conceptos (obligatorio)	**X**		Al tener los conceptos que se manejan en la ontología de dominio de plantas, se procede a almacenar estos conceptos en un diccionario, opción que es brindada por la herramienta de .NET, cabe decir que estos se

[18] Un índice plano se basa en los términos, no en los documentos.

				almacenan en un diccionario para facilitar el proceso de comparación de conceptos. Para realizar la comparación de estos conceptos se utiliza un proceso llamado GetCountOcuncurrence el cual se encargara de realizar las comparaciones de los conceptos de la ontología y los respectivos documentos que se recuperaron, además de esto realiza un conteo para almacenar el número de ocurrencias de los conceptos en cada uno de los documentos
Paso 6				
Utilización de dos herramientas semánticas	6.1 Utilizar dos herramientas: tesauro y ontología (opcional)		**X**	
Paso 7				
Estudio de representatividad de los conceptos.	7.1 Ponderación de frecuencias (obligatorio)			Para nuestro proyecto se utilizo el cálculo de TF –IDF (frecuencia de términos – frecuencia inversa del documento), donde medimos la frecuencia de los conceptos candidatos en los documentos que fueron recuperados, estos resultados son posteriormente guardados en una matriz. Para el cálculo de la similitud en el índice se realizó la asignación de pesos a los conceptos de la ontología, teniendo en cuenta la jerarquía de la misma. Los pesos fueron asignados consecutivamente dando más peso a los hijos que los padres. Realizando este proceso se consiguió mejorar los resultados obtenidos por el índice semántico creado.
	Distancia semántica			
	Medida de Similitud	**X**		
	Matriz Co-ocurrencia			
	Calculo TF-IDF	**X**		
Paso 8				
Construcción de la matriz o estructura de datos	8.1 Estructura de datos del índice semántico (obligatorio)	**X**		Se construye la matriz de ponderación de conceptos y es el lugar donde queda almacenado el índice semántico como producto de la indexación.
Paso 9				
Evaluación del índice semántico generado	9.1 Consultas de prueba (obligatorio)	**X**		Para probar nuestra aplicación se realizaron pruebas con dos colegios de la ciudad (como se especifican en el ítem 4.3)
	9.2 Depuración de la consulta (opcional)	**X**		Para nuestro caso se realizo una depuración de la consulta realizada por el usuario, en el caso de que se digitara un concepto que no se encontraba dentro del domino de la ontología que estábamos utilizando, se procedía a compararlo con un tesauro de domino general. Para este caso se tomo el tesauro WordNet.
	9.3 Medidas de precisión (obligatorio)	**X**		Para medir la precisión de nuestro procedimiento planteado realizamos las siguientes medidas de precisión: precisión-recuerdo, índice Map y Estadísticas Kappa.
	9.4 Revisión y mejoramiento del índice creado (opcional)	**X**		En nuestro caso se realizó una revisión de nuestro índice creado, en donde la primera vez se noto poca relevancia en los resultados que eran retornados con respecto la consulta del usuario. Al observar este problema se revisó de nuevo nuestro índice y se hicieron los respectivos cambios, obteniendo con esto una gran mejoría en los resultados que se presentaban a los usuarios.

Tabla 7. Plantilla instanciada de la creación del procedimiento

Para realizar la implementación del índice semántico, se sigue la metodología UP Ágil (Agile Unified Process) con sus respectivas fases. La fase de construcción del software se lleva a cabo en dos iteraciones puesto que se necesita realizar retroalimentaciones de los módulos creados. Las otras fases del proceso no requieren más de una iteración, esto porque se define el inicio (requerimientos) y la elaboración (diseño) del prototipo de manera completa y secuencial antes de la construcción, así también la fase de transición se lleva a cabo mediante las pruebas correspondientes en una sola iteración.

3.2 FASE DE INICIO

En la fase de inicio se desarrolla una descripción del producto final y se presenta el análisis del negocio para el producto. Para llevar a cabo esto se tuvo en cuenta los siguientes aspectos:

Acuerdo del Alcance: con el grupo de trabajo se analizó y se especificó el alcance del proyecto.

Definición Inicial de Requerimientos: se realiza la captura de requerimientos por parte del equipo de trabajo para el desarrollo de la aplicación, la cual permite realizar búsquedas de acuerdo a las necesidades de los usuarios. Los requerimientos son los siguientes:

- Se pide que los resultados de la búsqueda muestren una precisión aceptable (según la escala en sistemas de recuperación de información que está entre 0-1) en 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 en la creación del índice semántico.
- Utilización de los servicios de WordNet para procesar las consultas realizadas por el usuario.
- El idioma que se maneja para la ontología es el inglés, ya que en las referencias investigadas no se encontró una ontología que manejara el idioma español.

Aceptación del proceso: se realiza una aceptación por parte del equipo para llevar a cabo la metodología UP Ágil.

Viabilidad: se observa que el proyecto de la creación del índice semántico tiene sentido desde la perspectiva técnica y operacional.

Herramientas:
La herramienta principal para llevar a cabo la implementación del proyecto es Visual Studio 2008 con Framework 3.0 de Microsoft .NET[19] en el lenguaje C#. Se realizó una aplicación web que sirve de interacción con el usuario, para realizar las búsquedas respectivas.

Otra herramienta utilizada fue protégé en su versión 3.1.1 la cual permite el manejo de la ontología en un lenguaje OWL[20]. En nuestro proyecto se utilizó la API de la OWL (owlapi-bin), la cual proporciona los métodos necesarios para el manejo de la ontología por ejemplo para la obtención de clases, súper-clases, sub-clases y axiomas de la ontología.

Delicious una herramienta que genero un impacto importante en la obtención de la información así como también en los resultados obtenidos en el índice.

3.3 FASE DE ELABORACIÓN

Se especifica en detalle los casos de uso, diagramas conceptuales, diagramas de clase, de interacción, los cuales, permiten mostrar al usuario el funcionamiento del prototipo a crear y se diseña la arquitectura base.

3.3.1 DIAGRAMA DE CASOS DE USO

El Diagrama de Casos de Uso muestra la relación entre los actores y los casos de uso del sistema. Después de la definición de los requerimientos del prototipo software se realizaron los diagramas de casos de uso del presente proyecto, para describir la interacción entre los usuarios (actores). En la Figura 10 se muestra el diagrama principal.

[19] Se puede descargar en: http://www.microsoft.com/downloads/es-es/details.aspx?FamilyID=10cc340b-f857-4a14-83f5-25634c3bf043

[20] Extraído de: http://www.w3.org/2007/09/OWL-Overview-es.html

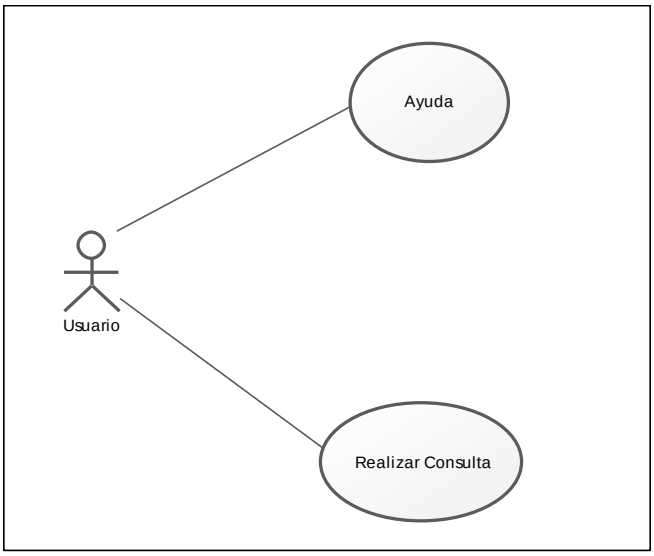

Figura 10. Diagrama de casos de uso

A continuación se presenta la definición de los actores que se identificaron para interactuar con el sistema y la descripción general de los casos de uso. Los detalles (formato expandido y casos de uso reales) de los mismos, así como los diagramas de secuencia, se encuentran en el Anexo E.

Caso de uso: Obtener Ayuda

Actor ACT-01	Usuario
Autores	Dignory Jimena Pérez Diana Maribel Pezo
Fuentes	Análisis de requerimientos del prototipo software.
Descripción	Un usuario tendrá la opción de consultar una ayuda con la descripción general de la utilización del software y los posibles resultados.
Comentarios	Esta opción también indica los nombres de los autores que realizaron la aplicación y un correo electrónico para contactarlos en caso de ser necesario.

Tabla 8. **Descripción caso de uso Obtener ayuda**

Caso de uso. Realizar Consulta

Actor ACT-02	Usuario
Autores	Dignory Jimena Pérez Diana Maribel Pezo
Fuentes	Análisis de requerimientos del prototipo software

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.
Comentarios	El usuario puede consultar los conceptos o preguntas de manera general, como en los buscadores actuales de la web.

Tabla 9. Descripción caso de uso Realizar Consulta

3.3.2 DIAGRAMA DE CLASE

El diagrama de clases de la aplicación muestra las relaciones entre las clases necesarias para la construcción del prototipo. Como principales clases se tienen las siguientes:

- **ScootWaterManager y HTMLParseManager:** se encuentran los métodos que permiten la eliminación de etiquetas HTML de los documentos retornados por Delicious.
- **LuceneManager:** se encuentran los métodos para la eliminación de StopWords, caracteres especiales y llevar las palabras a su forma raíz.
- **TFIDFMeasure:** se encarga de realizar los cálculos correspondientes a la ponderación de frecuencias de los conceptos que se extraen.
- **OntologyManager:** se encuentran los métodos que permiten la manipulación de la ontología.
- **SemanticIndexManager:** se definen los métodos correspondientes para la creación del índice Semántico.

A continuación, en la Figura 11 se muestra el diagrama general.

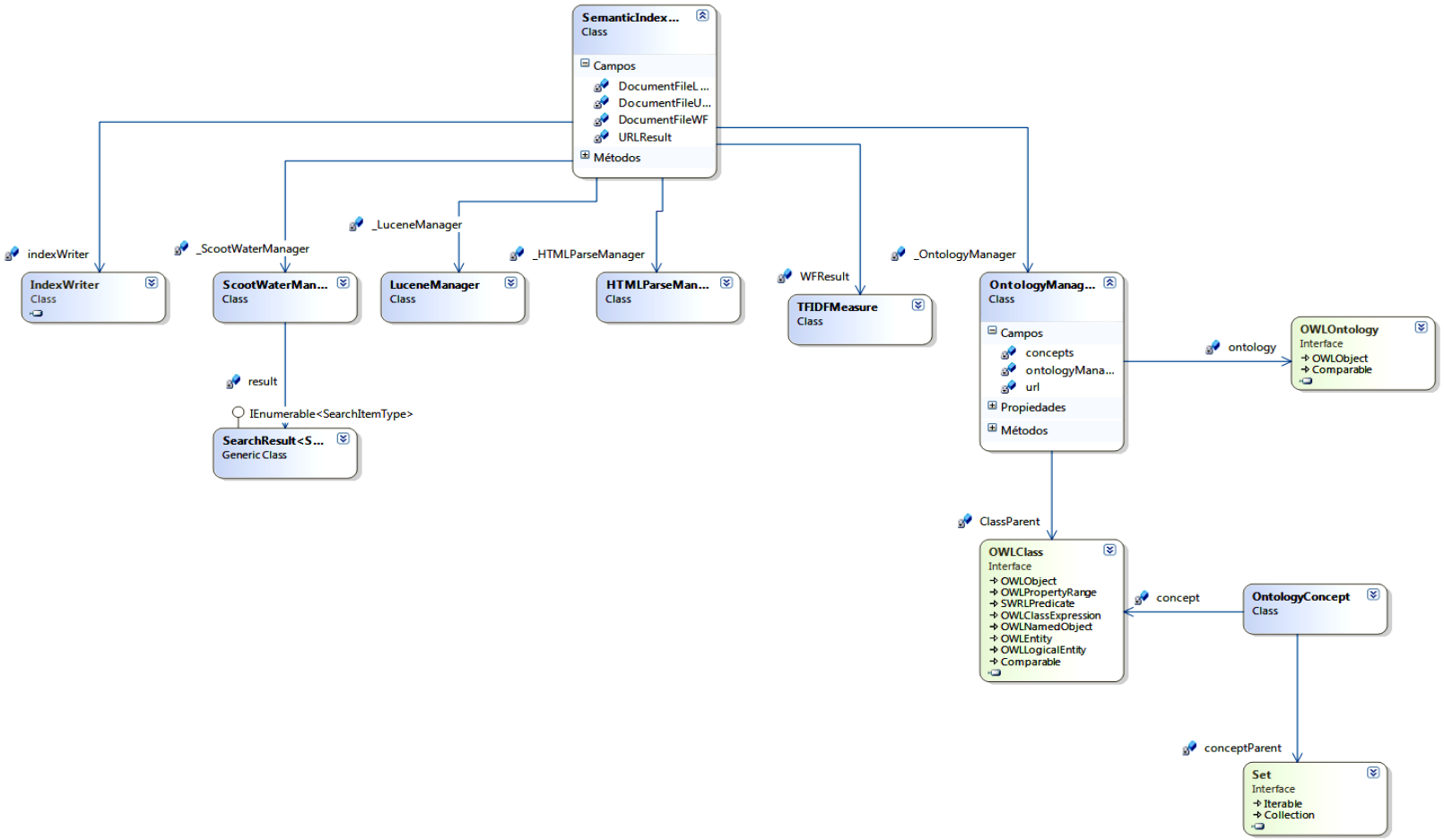

Figura 11. Diagrama de clases general

Después de haber implementado los objetos y la lógica de negocio que se utilizara en el índice semántico y siguiendo el diagrama de clases anterior, se hizo necesario implementar un Servicio que utilice los recursos de WordNet. Para ello se realizo un diagrama de clases (Figura 12) de acuerdo a su implementación. A continuación se describen las clases que permiten el manejo de WordNet:

- **Concept:** tiene los métodos necesarios para el manejo de los conceptos en la jerarquía como los hiperónimos, hipónimos y sinónimos.
- **WordNetDataBase:** esta clase es abstracta y es propia de la API de WordNet, la cual proporciona toda la base de datos de la misma.
- **WordNetManager:** contiene los métodos necesarios para el manejo de las relaciones entre los conceptos y la extracción de los mismos.

En el Anexo E se encuentran los diagramas de las clases implementadas por otros autores y que se utilizaron como recurso para el presente proyecto.

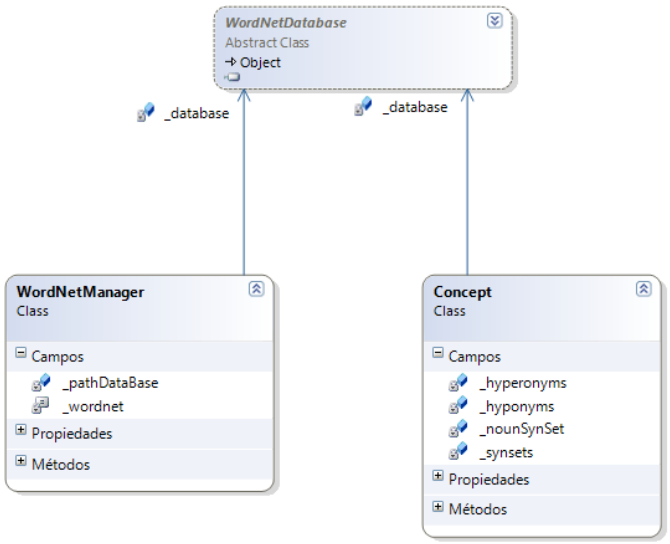

Figura 12. Diagrama de clases para utilizar WordNet

3.3.3 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.

La Figura 13 presenta el diagrama de despliegue del sistema. En primer lugar se presenta SemanticIndex la cual tiene 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 en Delicious, el manejo de la ontología, el cálculo de la representatividad de los conceptos, así como también el

proceso de la consulta en el índice. A continuación se encuentra SemanticIndexWeb encargada de la lógica de presentación del proyecto y donde se implementa la interfaz grafica del usuario. A su vez se conecta con el WebService, el cual proporciona el servicio para utilizar el índice en otro buscador. En el WebService WordNet se realiza todo el proceso del manejo del tesauro de WordNet para el procesamiento de las consultas del usuario.

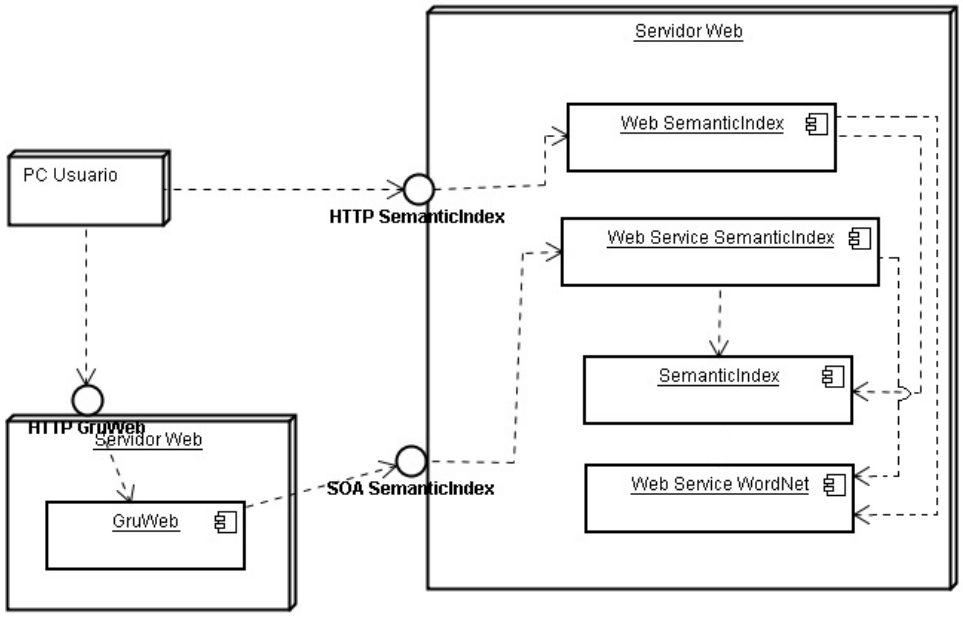

Figura 13. Diagrama de despliegue

3.3.4 ARQUITECTURA DE LA APLICACIÓN

Para el desarrollo de la aplicación de índices semánticos 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 14 se muestra el diagrama general de la arquitectura tratada.

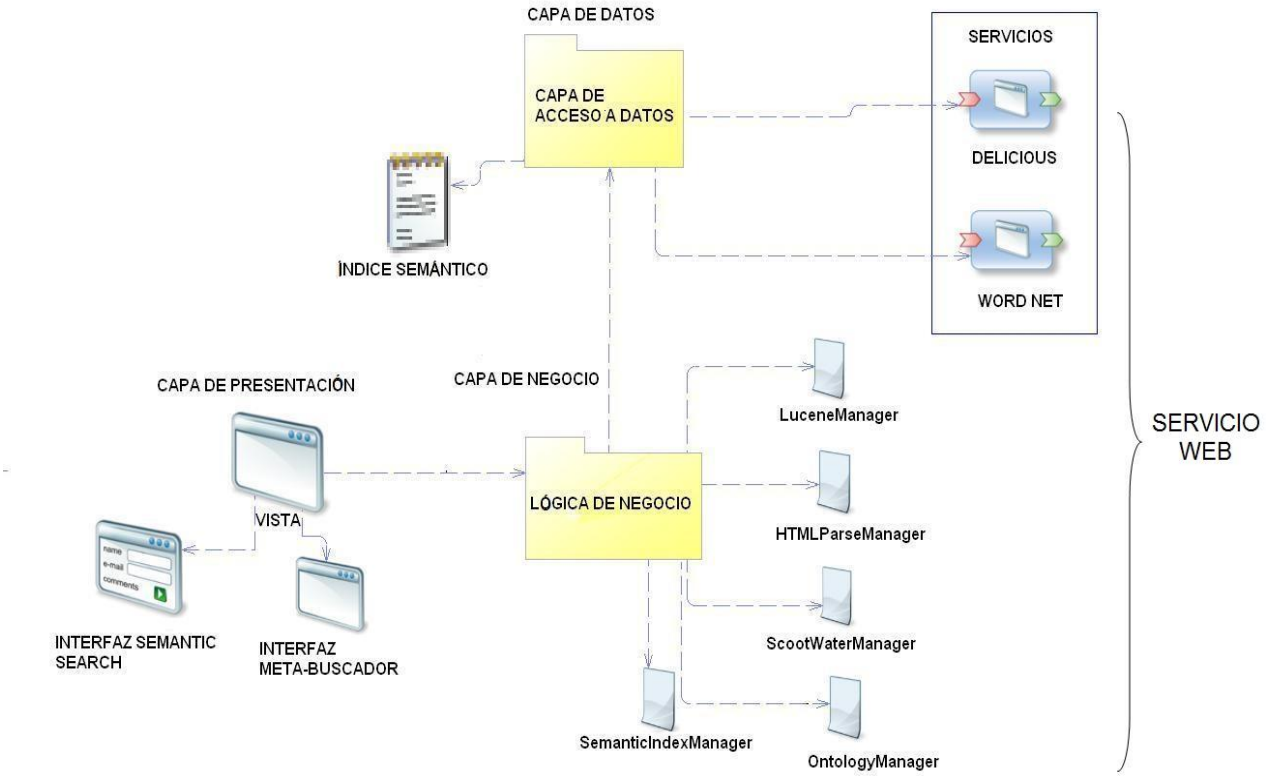

Figura 14. Arquitectura tres capas

Para la aplicación de la creación del índice semántico se ha seleccionado una herramienta de desarrollo Microsoft .NET compact junto con el uso de un Framework de desarrollo 3.0 de .Net, el cual, ofrece una plataforma madura, documentada y con un diseño consistente para desarrollar características de productividad. 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.

- Capa de presentación
- Capa lógica de negocios
- Capa de datos

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 (Semantic Search) que proporciona interacción con los usuarios para realizar las pruebas correspondientes y evaluación del prototipo. Además para que nuestro índice semántico sea utilizado por los motores de búsqueda se genera un servicio Web. Para este caso se realizo con el fin de integrarlo al meta-buscador GruWeb.

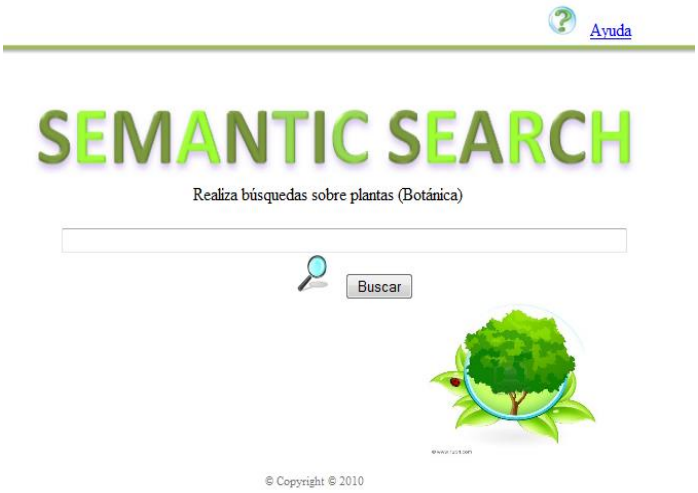

Figura 15. Capa de presentación

- **Capa lógica de negocios:** parte donde se reciben los programas que se ejecutan, se recibe las peticiones del usuario y se le envían las respuestas tras el proceso. Se denomina capa de 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, y con la capa de datos, para solicitar al gestor de la base de datos o servicios, almacenar o recuperar datos de él[21]. 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 [106].

  En el prototipo de este proyecto, se incluyen las clases y métodos que interactúan con la capa de acceso a los archivos del índice semántico y servicios Web, y con la interfaz de presentación al usuario. Las clases principales son: HTMLParseManager, LuceneManager, OntologyManager, ScootWaterManager, SemanticIndexManager. Por medio de estas clases

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

se maneja la información obtenida de los servicios Web y se genera el índice semántico mediante el mapeo de la Ontología con los documentos.

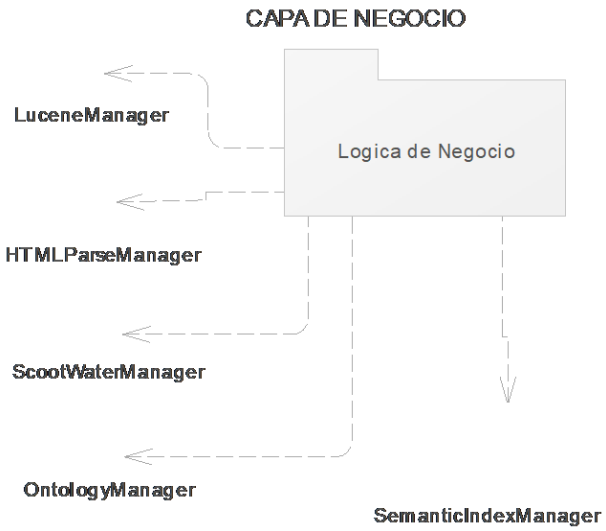

Figura 16. Lógica de Negocio

- **Capa de datos:** en la lógica de acceso a datos se tienen las clases que permiten la recuperación de información del buscador Delicious y la obtención, manipulación y almacenamiento de los documentos obtenidos de las páginas Web. En esta capa se genera un archivo plano (.txt) del Índice Semántico obtenido al llevarse a cabo el proceso de indexación.

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.) [107], como en el caso de esta aplicación. Los servicios brindan la información necesaria para la construcción del índice sobre un dominio dado (Delicious), accediendo por medio de las API's a sus servicios y luego proporcionar información para evaluar del índice construido (WordNet).

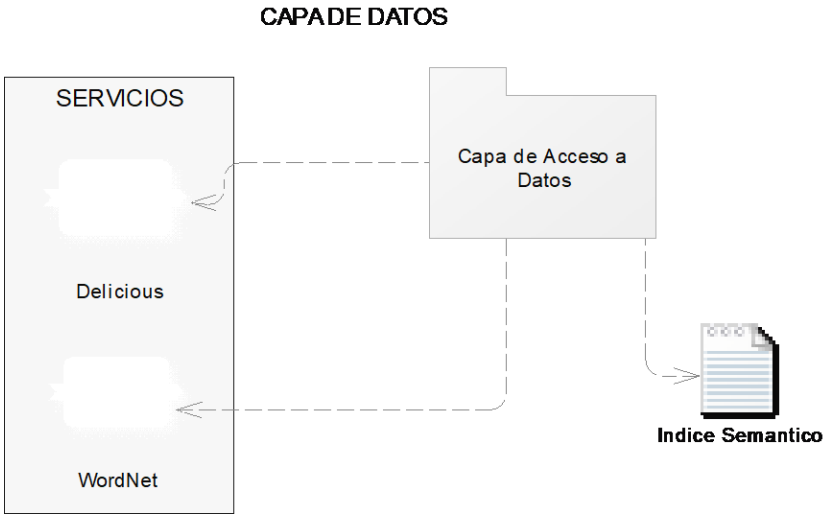

Figura 17. Capa de Datos

Debido a los requerimientos de la aplicación, no se hace necesario el almacenamiento y manipulación de grandes cantidades de información, por lo cual no se hace necesario la conexión a bases de datos, servicios Web, ni servicios SMTP para correos electrónicos. Esta aplicación solo se basara en el uso de archivos planos.

La especificación de los patrones de software usados se describe en detalle en el Anexo F.

3.4 FASE DE CONSTRUCCIÓN

En esta fase se realiza la construcción del prototipo que permite construir el índice semántico a partir de la ontología seleccionada. Posteriormente será utilizado por un buscador Web, lo cual se inicia con la definición de la arquitectura

3.4.1 ITERACIÓN 1

Teniendo la arquitectura base ya definida, se procede a la construcción del índice semántico en la plataforma de .NET teniendo en cuenta las clases establecidas y las funcionalidades a tener en cuenta para la generación del índice semántico. Al generarse este índice se guarda en un archivo plano que se encuentra en la capa de datos y posterior a esto se realizan las primeras validaciones del prototipo como se especifica a continuación.

3.4.1.1 Pruebas preliminares

La realización de las primeras pruebas del prototipo se llevo a cabo consultando algunos conceptos (“Flower”, “Seendling”, “Plant Structure”, “Seed”, “Leaf” y “Vascular System”) en el índice y probando el número de ocurrencias (frecuencia)

en todos los documentos (50) indexados. Esto con el fin de probar el funcionamiento de la aplicación y, siguiendo la metodología utilizada (UP Ágil), se toma como referencia el enfoque de caja negra, en la primera iteración de la construcción del software. Las pruebas fueron realizadas por los implementadores del sistema en el ambiente de desarrollo y se evaluó la existencia de las consultas escogidas y su frecuencia en cada URL retornada. Además se probó la indexación correcta de los conceptos seleccionados los cuales fueron

A continuación, se describen dichos conceptos con sus correspondientes frecuencias en cada URL retornada.

URL	Conceptos					
	flower	Seedling	Plant structure	seed	Leaf	Vascular system
http://shop.ebay.com/i.html?_nkw=plants	5	0	0	0	1	0
http://www.enchantedlearning.com/themes/plants.shtml	48	0	0	7	21	0
http://waynesword.palomar.edu/indxwayn.htm	10	0	0	10	2	0
http://shopping.yahoo.com/plants	4	0	0	2	0	0
http://www.discoverplants.com/plant-types.php	2	0	0	0	0	0
http://www.floridata.com/index.cfm	1	0	0	3	3	0
http://www.wikihow.com/Grow-a-Tomato-Plant	0	2	0	2	1	0
http://www.biology4kids.com/files/plants_main.html	0	0	0	0	0	0
http://www.wilderness-survival.net/plants-1.php	0	0	0	0	4	0
http://www.proplants.com/	0	0	0	0	0	0
http://www.theodorepayne.org/index.html	0	0	0	0	0	0
http://www.botany.com/	3	0	0	0	1	0
http://www.thefind.com/garden/info-these-plants	1	0	0	8	1	0
http://plants.web-indexes.com/	0	0	0	0	3	0
http://www.bizrate.com/flowers-plants/christmas-plants/	4	0	0	1	0	0
http://www.target.com/s/plants	0	0	0	0	0	0
http://www.internet4classrooms.com/science_elem_plants.htm	0	0	0	0	0	0
http://www.crescentbloom.com/Plants/default.htm	0	0	0	0	0	0
http://www.buzzle.com/articles/plants/	2	2	1	5	9	0
http://www.abcteach.com/directory/basics/science/plants/	7	0	0	4	12	0
http://www.1800flowers.com/houseplants	3	0	0	0	0	0
http://perennial-plants.suite101.com/	6	0	0	0	0	0
http://www.naturephoto-cz.com/plants.html	0	0	0	0	1	0
http://www.ucmp.berkeley.edu/plants/plantae.html	0	0	0	0	0	0
http://plants.usda.gov/java/factSheet	1	0	0	0	0	0
http://www.qrg.northwestern.edu/projects/marssim/simhtml/info/Whats-a-plant.html	0	0	0	0	0	0
http://dir.yahoo.com/Science/Biology/Botany/Plants/	2	0	0	0	0	0
http://en.wikipedia.org/wiki/Flowering_plant	16	1	0	14	1	0
http://www.theteachersguide.com/plantsflowers.htm	35	0	1	0	3	0
http://www.atozteacherstuff.com/Themes/Plants/	3	0	0	2	0	0

http://www.enchantedlearning.com/subjects/plants/plant/	2	0	1	0	3	0
http://www.teleflora.com/plants/plants-collection-509_807c.asp	1	0	0	0	0	0
http://www.stokestropicals.com/	0	0	0	0	0	0
http://crimsonsage.com/	1	0	0	0	1	0
http://library.thinkquest.org/3715/	1	0	0	0	0	0
http://www.neok12.com/Plants.htm	3	1	1	1	0	0
http://www.newworldencyclopedia.org/entry/Plant	0	0	0	31	3	1
http://www.tms.riverview.wednet.edu/lrc/plants.htm	4	1	1	3	6	0
http://www.givingplants.com/	1	0	0	2	0	0
http://www.dmoz.org/Home/Gardening/Plants/	3	0	0	0	0	0
http://www.basic-info-4-organic-fertilizers.com/plants.html	4	0	0	2	1	0
http://www.webshots.com/explains/outdoors/plants.html	21	0	0	32	1	0
http://www.proflowers.com/house-plants-pbs	3	0	0	0	0	0
http://www.gardenguides.com/plants/	2	0	0	1	1	0
http://www.1800flowers.com/plantdelivery	5	0	0	0	0	0
http://www.ftd.com/plants-ctg/product-plants	1	0	0	0	0	0
http://www.ftd.com/	1	0	0	0	0	0
http://www.answers.com/topic/plant	1	2	0	26	5	1
http://plants.usda.gov/	1	0	0	0	0	0
http://en.wikipedia.org/wiki/Plant	0	0	0	15	3	0

Tabla 10. Frecuencia de conceptos en cada URL.

Al revisar la frecuencia en cada documento, se observa la correspondencia de ésta con el número que se presenta en la prueba realizada.

3.4.2 ITERACIÓN 2

Después de tener una implementación estable del índice semántico, se continuó con la realización de las pruebas para el registro de datos utilizado en la comprobación del funcionamiento e interacción del sistema. Se realizaron las pruebas alfa de la aplicación.

3.4.2.1 Pruebas alfa

Además de las anteriores pruebas teniendo en cuenta la frecuencia de conceptos en los documentos, se realizaron las búsquedas por medio de la aplicación y de una base de recursos previamente calificados y clasificados con respecto a algunas consultas, con el fin de comparar los resultados por medio de la aplicación y de forma manual con cada url de acuerdo a la búsqueda.

Estas pruebas se realizaron teniendo en cuenta las pruebas alfa en la ingeniería del software y tomando como base la metodología utilizada. En las primeras se evalúa el funcionamiento del sistema tomando como referencia la búsqueda de conceptos de forma manual, con los desarrolladores y un cliente de la aplicación, que en este caso fue el director del proyecto. Estas pruebas corresponden a las

que se realizan en la categoría de pruebas de integración, enfoque caja negra, y basadas en la metodología UP Ágil.

La prueba manual se realizó con cada una de las 50 páginas indexadas buscando cada concepto especificado en ellas. La prueba con la aplicación se refiere a la comprobación de las páginas que retorna el buscador y que realmente corresponden al concepto buscado. Las consultas realizadas para cada prueba fueron: “Flower”, “Seedling”, “Plant structure”, “seed” y “leaf”.

La Tabla 11 presenta dichas observaciones. **M**: prueba manual y **A**: prueba con la aplicación.

N°	URL	Flower		seedling		Plant structure		seed		Leaf	
		M	A	M	A	M	A	M	A	M	A
1	http://www.enchantedlearning.com/themes/plants.shtml	X	X					X	X	X	X
2	http://www.theteachersguide.com/plantsflowers.htm	X	X			X	X			X	X
3	http://www.webshots.com/explains/outdoors/plants.html	X	X					X	X	X	X
4	http://www.wikihow.com/Grow-a-Tomato-Plant			X	X			X	X	X	X
5	http://www.bizrate.com/flowers-plants/christmas-plants/	X	X					X	X		
6	http://en.wikipedia.org/wiki/Flowering_plant	X	X	X	X			X	X	X	X
7	http://www.biology4kids.com/files/plants_main.html										
8	http://waynesword.palomar.edu/indxwayn.htm	X	X					X	X	X	X
9	http://www.teleflora.com/plants/plants-collection-509_807c.asp	X	X								
10	http://www.thefind.com/garden/info-these-plants	X	X					X	X	X	X
11	http://www.wilderness-survival.net/plants-1.php									X	X
12	http://www.gardenguides.com/plants/	X	X					X	X	X	X
13	http://www.abcteach.com/directory/basics/science/plants/	X	X					X	X	X	X
14	http://www.proplants.com/										
15	http://www.naturephoto-cz.com/plants.html									X	X
16	http://www.theodorepayne.org/index.html										
17	http://www.tms.riverview.wednet.edu/lrc/plants.htm	X	X	X	X	X	X	X	X	X	X
18	http://www.theodorepayne.org/index.html										
19	http://www.1800flowers.com/houseplants	X	X								
20	http://plants.web-indexes.com/									X	X
21	http://www.1800flowers.com/plantdelivery	X	X								
22	http://www.basic-info-4-organic-fertilizers.com/plants.html	X	X					X	X	X	X
23	http://www.target.com/s/plants										
24	http://www.botany.com/	X	X							X	X
25	http://www.buzzle.com/articles/plants/	X	X	X	X	X	X	X	X	X	X
26	http://www.enchantedlearning.com/subjects/plants/plant/	X	X			X	X			X	X
27	http://www.internet4classrooms.com/science_elem_plants.htm										
28	http://perennial-plants.suite101.com/	X	X								
29	http://www.proflowers.com/house-plants-pbs	X	X								
30	http://www.crescentbloom.com/Plants/default.htm										
31	http://www.answers.com/topic/plant	X	X	X	X			X	X	X	X
32	http://www.ucmp.berkeley.edu/plants/plantae.html										

33	http://shopping.yahoo.com/plants	X	X					X	X		
34	http://www.qrg.northwestern.edu/projects/marssim/simhtml/info/Whats-a-plant.html										
35	http://www.atozteacherstuff.com/Themes/Plants/	X	X					X	X		
36	http://www.neok12.com/Plants.htm	X	X	X	X	X	X	X	X		
37	http://www.stokestropicals.com/										
38	http://www.dmoz.org/Home/Gardening/Plants/	X	X								
39	http://www.discoverplants.com/plant-types.php	X	X								
40	http://plants.usda.gov/java/factSheet	X	X								
41	http://www.stokestropicals.com/										
42	http://dir.yahoo.com/Science/Biology/Botany/Plants/	X	X								
43	http://www.ftd.com/plants-ctg/product-plants	X	X								
44	http://crimsonsage.com/	X	?							X	?
45	http://www.ftd.com/	X	X								
46	http://www.floridata.com/index.cfm	X	X					X	X	X	X
47	http://library.thinkquest.org/3715/	X	X								
48	http://www.newworldencyclopedia.org/entry/Plant							X	X	X	X
49	http://en.wikipedia.org/wiki/Plant							X	X	X	X
50	http://www.givingplants.com/	X	?					X	X		

Tabla 11. Comparación de búsquedas con la aplicación y manualmente.

En la Tabla 12 se presentan los resultados de las comparaciones anteriores con la búsqueda manual y por medio del prototipo construido.

Conceptos	Total paginas manualmente	Total paginas retornadas aplicación	
		Correctas	Incorrectas
Flower	35	33	2
seedling	6	6	0
Plant structure	5	5	0
Seed	20	20	0
Leaf	21	21	0
Vascular system	2	2	0
TOTAL	89	87	2

Tabla 12. Resultados de recuperación de conceptos

Se observa una variación mínima de dos páginas incorrectas por 87 correctas teniendo en cuenta todos los conceptos y los resultados retornados con la aplicación y manualmente. Esto permite obtener una apreciación adecuada con respecto a la eficiencia de la aplicación, pues se obtienen, desde la aplicación, las páginas esperadas al realizarse la búsqueda manual. Por otra parte, las páginas retornadas por el servicio Delicious, permiten obtener mejores resultados debido al marcado social que se realiza en dicho servicio.

Las segundas pruebas realizadas se llevan a cabo con algunos usuarios a los que está dirigida la aplicación y corresponden a las pruebas de aceptación del software, siguiendo la metodología propuesta. Se realizan en el Colegio Campestre Americano con la asesoría de los desarrolladores y se prueba la

funcionalidad del sistema teniendo en cuenta la relevancia esperada por los usuarios. Estas pruebas se documentan en el capítulo de "Validación del prototipo" con el fin de calcular las medidas correspondientes (Véase sección 4.3); esto debido a la necesidad de realizar cálculos y estadísticas basados en el funcionamiento del software. Estas pruebas permitieron la retroalimentación de la aplicación para mejorar su funcionamiento.

3.5 FASE DE TRANSICIÓN

La fase de transición se centra en las pruebas de aceptación según la metodología, siguiendo el enfoque de pruebas beta del prototipo, con el fin de realizar los ajustes necesarios en cuanto a la usabilidad y al funcionamiento general del sistema.

En esta fase se realizaron **pruebas de usabilidad** en el colegio Campestre Americano de la ciudad, con el fin de determinar el grado de satisfacción de los usuarios con el prototipo creado, incluyendo la relevancia de resultados al realizar algunas consultas.

A dos grupos de estudiantes de cuarto y quinto grado (básica primaria), de 5 estudiantes cada uno, se les realizó una prueba de usabilidad (Ver Anexo G 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 realizaron unas 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 ven en detalle en el Anexo G.

Al terminar la prueba se realiza la ponderación de resultados que se muestra en la Figura 8.

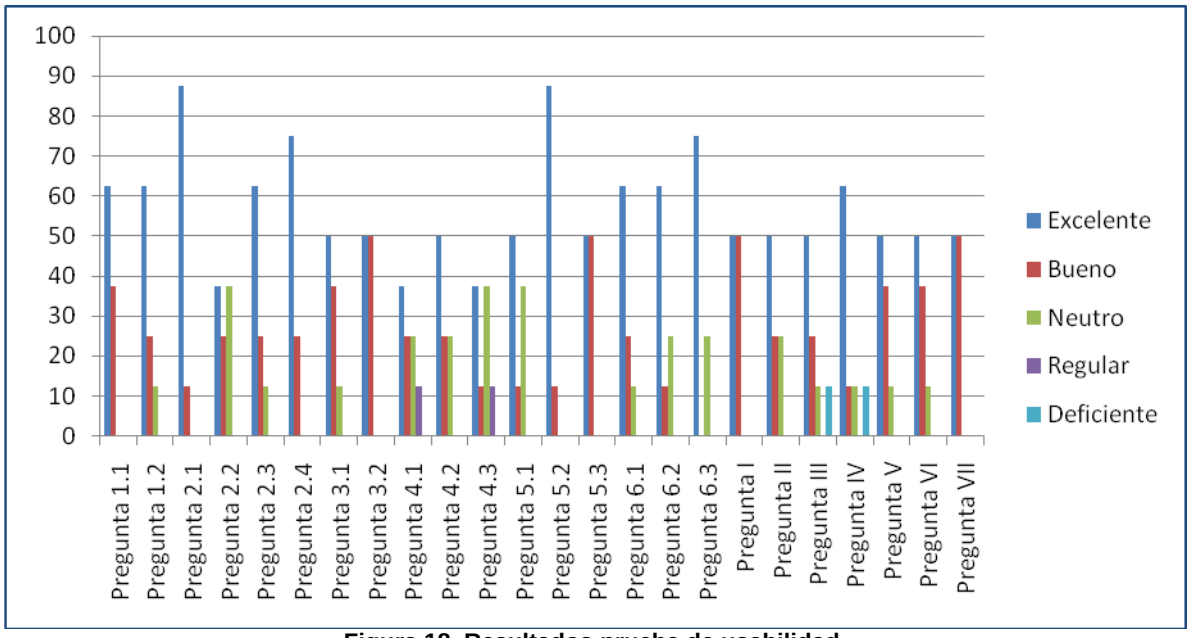

Figura 18. Resultados prueba de usabilidad

Como se observa, se realizaron varias preguntas sobre cada aspecto mencionado y se obtuvieron, en un porcentaje alto, los resultados entre bueno y excelente en la calificación de la aplicación. Por lo anterior, no se hizo necesario realizar de nuevo este tipo de pruebas con el segundo colegio visitado.

Los porcentajes y detalles de los resultados, así como las preguntas realizadas, se pueden consultar en el Anexo G.

4 VALIDACION DEL PROTOTIPO

Para la validación del prototipo se llevaron a cabo medidas de precisión-recuerdo, Índice MAP y estadísticas Kappa. Estas últimas se realizaron gracias a la colaboración del colegio Campestre americano y la institución educativa Alejandro de Humboldt sede Yanaconas.

Las pruebas realizadas se presentan a continuación en el respectivo orden Curva precisión-recuerdo (4.1), Indice MAP (4.2), Pruebas para estadísticas Kappa (4.3) con sus respectivos cálculos.

4.1 CURVA DE PRECISIÓN-RECUERDO

La curva Precision-Recall es una medida utilizada en los sistemas de recuperación de información como los motores de búsqueda, que permite evaluar la eficacia [29] de resultados respecto a varias consultas realizadas por el usuario. Para ello se tiene en cuenta la cantidad de documentos recuperados y la cantidad de documentos recuperados que son relevantes de acuerdo a una consulta. La precisión y recall (recuerdo) se expresan de la siguiente manera.

Precisión: para una consulta dada, la precisión es la razón entre el número de documentos relevantes recuperados y el total de documentos recuperados [102].

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

Recall: es la proporción de documentos relevantes que fueron recuperados con respecto al total de documentos relevantes que se deberían recuperar.

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

4.1.1.1 Average Precision (Precisión Promedio)

Promedio de los valores de precisión en los puntos en que se recupera cada documento relevante [101].

Para realizar las pruebas se tomaron 5 conceptos y se mide para cada uno, la precisión y recuerdo en las URLs retornadas. A continuación se realizan consultas y se calcula la curva de precisión-recall. Posteriormente se promedia la precisión, la cual será necesaria para el cálculo del índice MAP. Para mayor información de los resultados obtenidos y los cálculos realizados en cada uno, puede dirigirse al Anexo H.

Concepto buscado	Documentos relevantes	Promedio precisión (Average precision)
flower	34	0.686
seedling	7	0.201
plant structure	5	0.205
seed	20	0.5822
leaf	21	0.712

Tabla 13. Resumen de Precisión en las consultas

4.1.1.2 Resultados de medida Precision-Recall

Los resultados obtenidos de la medida de Precision-Recall permiten observar que la precisión en los resultados relevantes es buena a pesar del incremento de los documentos. La precisión disminuyó en algunas consultas por la falta de documentos recuperados (en cantidad) que fueran relevantes para cada una.

Por otra parte, el número de documentos relevantes determina en parte, la precisión en las búsquedas, esto debido al servicio que presta Delicious, el cual cuenta con un marcado social de sus páginas y contenidos que permite obtener mayor precisión en lo que se consulta.

4.2 INDICE MAP

El Mean Average Precision (MAP) [100, 102] es el promedio del valor de precisión media de un conjunto de consultas.

Para este cálculo se tiene en cuenta la ***precisión promedio*** en cada consulta realizada anteriormente, es decir, la precisión media de cinco (5) consultas:

$$Mean\ average\ precision = (0.686 + 0.201 + 0.205 + 0.5822 + 0.712)/5$$

$$= \frac{2.3862}{5} = 0.48$$

El Índice MAP presenta un buen promedio, teniendo en cuenta los resultados en general en un sistema de recuperación de información, los cuales varían entre 0.1

y 0.7 [29]. Esto significa una buena precisión en general de acuerdo a las consultas realizadas.

Además de las pruebas anteriores, en la validación del prototipo se llevaron a cabo cálculos de estadísticas Kappa con la colaboración del Colegio Campestre Americano, el cual nos permitió evaluar la aplicación con los estudiantes de grados cuarto, quinto (básica primaria) y sexto (secundaria), y la colaboración de la Institución Educativa Alejandro de Humboldt sede Yanaconas, con el grado cuarto (básica primaria).

En el primer colegio se evaluó la aplicación con un grupo de estudiantes (7), de grado quinto (2 estudiantes) y sexto (5 estudiantes). En el segundo colegio se evaluó la aplicación con un grupo de estudiantes (6) de grado cuarto (básica primaria) y dos docentes de la institución.

Previo aviso a las directivas de los Colegios mencionados, se procedió a realizar las pruebas de relevancia y calcular las medidas Kappa para el primero y segundo colegio.

4.3 PRUEBAS PARA ESTADISTICAS KAPPA

Para los estudiantes del grado sexto y docentes del Colegio Campestre Americano se realizaron unas pruebas que permitieron verificar la concordancia entre dos personas (jueces) sobre la relevancia de las páginas retornadas. Con esta prueba se calcularon las estadísticas Kappa.

La prueba se realizó de la siguiente manera: Dos usuarios realizaron la búsqueda de dos conceptos y verificaron en cada página (Url) retornada, la relevancia de sus resultados teniendo en cuenta que la página observada sea interesante o relevante respecto a su consulta y de acuerdo a su propio criterio. Cada juez anotó su decisión en un formato entregado con la consulta a realizar y la lista de URLs retornadas para cada búsqueda (Ver Anexo H).

Las consultas realizadas por los estudiantes fueron: *Embryo, Seed coat, Plant cell y Xylem.*

Al realizar estas pruebas por primera vez en el colegio, se detectó un inconveniente con el índice semántico generado puesto que la relevancia no era totalmente visible para varios estudiantes en la mayoría de consultas. Esto también debido a que se buscaba la palabra clave dentro de las páginas

retornadas sin interesarles el contenido y su relación con la búsqueda. Otro inconveniente tuvo que ver con el manejo del idioma, pues, aunque tenían cierto dominio, no era el suficiente para revisar todo el contenido de las páginas y determinar si eran relevantes para la consulta que estaban realizando.

Por lo anterior, se realizó un cambio en el Índice semántico, tomando los conceptos más relacionados con la búsqueda hacia niveles inferiores en la jerarquía. Antes de esto se realizaba tomando los conceptos más generales, es decir, los niveles superiores (padres). Con este cambio se buscaba retornar búsquedas más específicas sobre las consultas y así, encontrar en los documentos más relación con los conceptos buscados.

En la Tabla 14Tabla 14 se muestran los resultados (resumidos) de las pruebas realizadas en dicho colegio y se observa la relevancia de los documentos para los jueces escogidos, en las 47 URLs revisadas.

No. PRUEBA	CONSULTA	JUEZ 1		JUEZ 2	
		Si	No	Si	No
Prueba 1	Embryo	4	6	4	6
	Seed coat	3	5	3	5
	Subtotal	7	11	7	11
Prueba 2	Plant cell	7	3	3	7
	Xylem	7	2	2	7
	Subtotal	14	5	5	14
Prueba 3 (docente y estudiante)	Plant cell	9	1	7	3
	TOTAL	30	17	19	28
	Total URLs revisadas	47		47	

Tabla 14. Pruebas Colegio Campestre Americano

Con los resultados obtenidos se realizaron los cálculos de las estadísticas Kappa para determinar el acuerdo de relevancia entre los jueces (sección 4.3.1.1).

Posterior al cambio y retroalimentación en el índice semántico se realizaron otras pruebas en la Institución Educativa Alejandro de Humboldt sede Yanaconas, con estudiantes del grado cuarto (básica primaria). Esto con el fin de comparar las pruebas anteriores teniendo en cuenta los cambios realizados, y así determinar la relevancia de los resultados para cada juez escogido.

En este caso las consultas realizadas por los estudiantes fueron: *Embryo, Seed coat, Plant cell, Xylem, Cuticle, Flower, Seed y Fruit*. Al igual que los estudiantes del Colegio Campestre Americano, los estudiantes de grado cuarto de la mencionada institución, realizaron la consulta de dos conceptos (cada juez) y anotaron su decisión en el formato entregado (Ver Anexo H) para tal fin.

Para resolver un poco el inconveniente del manejo del idioma, se procedió inicialmente a dar una "clase" de botánica con la explicación de lo que necesitaban consultar, gráficamente, y enseñándoles en inglés, las búsquedas a realizar. Además, se contó con el apoyo de la profesora de inglés del Colegio, quien recordó a los estudiantes cierto vocabulario enseñado.

La Tabla 15Tabla 15 presenta los resultados (resumidos) de las pruebas realizadas en dicho colegio y se observa la relevancia de los documentos para los jueces escogidos, en las 77 URLs revisadas.

No. PRUEBA	CONSULTA	JUEZ 1		JUEZ 2	
		Si	No	Si	No
Prueba 1	Embryo	4	6	4	6
	Cuticle	4	6	4	6
	Subtotal	8	10	8	10
Prueba 2	Xylem	4	5	8	1
	Seed coat	5	3	5	3
	Subtotal	9	8	13	4
Prueba 3	Plant cell	8	2	8	2
	Fruit	9	1	10	0
	Subtotal	17	3	18	2
Prueba 4 (docentes)	Flower	10	0	9	1
	Seed	10	0	8	2
	Subtotal	20	0	17	3
	TOTAL	54	23	56	21
	Total urls revisadas	77		77	

Tabla 15. Institución Educativa Alejandro de Humboldt, sede Yanaconas

4.3.1 ESTADÍSTICAS KAPPA

Cada persona que evalúe el sistema puede emitir juicios diferentes respecto a la relevancia de los resultados, esto no solo depende de su objetividad sino también de su cultura o idiosincrasia. Por ello, no se considera solo una opinión a la vez acerca de los documentos retornados, pues esto no sería fiable.

La estadística Kappa busca considerar y medir la cantidad de acuerdos que tienen los jueces en sus opiniones de relevancia, está diseñado para juicios categóricos y corrige una tasa de acuerdo simple por una tasa de concordancia por azar [29]. La estadística Kappa se expresa matemáticamente de la siguiente manera:

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A) es la proporción de las veces que los jueces están de acuerdo y P(E) es la proporción de las veces que se espera llegar a un acuerdo por casualidad. El cálculo de este último se realiza, por lo general, de la siguiente manera [29]:

$$P(E) = P(\text{no-relevantes})^2 + P(\text{relevantes})^2$$

Para este proyecto, los cálculos para estadísticas Kappa se realizaron como sigue: Se toman dos Jueces (estudiantes y/o docentes) que realizaron la prueba de las URLs para dos consultas iguales y se consignan en una tabla sus opiniones respecto a la relevancia de cada resultado retornado.

4.3.1.1 Cálculos Kappa Colegio Campestre Americano

Para el cálculo de las estadísticas Kappa, primero se toman los jueces con sus respectivas consultas y por cada prueba se realiza el cálculo.

La relevancia de los documentos encontrada por el Juez 1 se consigna en filas, la del Juez 2 se muestra en las columnas. El total de urls revisadas por todos es 47. Los cálculos de todas las consultas se detallan en el Anexo H.

Total de documentos revisados: juez 1 y juez 2

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	19	11	30
	NO	0	17	17
	TOTAL	19	28	47

Tabla 16. Relevancia total según jueces de Campestre Americano

Estadística Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.765 - 0.50)}{(1 - 0.50)} = 0.53$$

Se observó un acuerdo en un 53% teniendo en cuenta el total de los jueces y consultas. Esto debido a los inconvenientes mencionados anteriormente.

4.3.1.2 Cálculos Kappa Institución Educativa Alejandro de Humboldt, sede Yanaconas

Para el cálculo de las estadísticas en la institución, se toman primero las dos consultas de cada prueba y las decisiones de sus jueces, con el subtotal de cada prueba según la Tabla 15Tabla 15. El total de URLs revisadas es 77. (Ver detalles Anexo H).

Total documentos revisados: juez 1 y juez 2

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	54	0	54
	NO	2	21	23
	TOTAL	56	21	77

Tabla 17. Relevancia total según jueces de Alejandro de Humboldt

$$Estadística\ Kappa$$

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.974 - 0.795)}{(1 - 0.795)} = 0.872$$

El acuerdo en este caso fue mayor que en el total calculado en el primer colegio evaluado y mayor que en varios casos específicos, lo cual es un buen indicador del funcionamiento de la indexación semántica de acuerdo a las necesidades de información de los usuarios.

4.3.1.3 Comparación de los resultados.

La comparación presentada a continuación se realizó mediante los cálculos de las mismas consultas realizadas en el colegio Campestre Americano y en la Institución Alejandro de Humboldt.

Campestre Americano

		Relevancias Juez 2		
		SI	NO	TOTAL
Relevancias Juez 1	SI	12	9	21
	NO	0	16	16
	TOTAL	12	25	37

Alejandro de Humboldt

		Relevancias Juez 2		
		SI	NO	TOTAL
Relevancias Juez 1	SI	21	0	21
	NO	4	12	16
	TOTAL	25	12	37

Tabla 18. Comparación de resultados en los dos colegios evaluados

Alejandro de Humboldt

Proporción observada de las veces en que los jueces estuvieron de acuerdo.

$$P(A) = \frac{12 + 16}{37} = 0.756$$

$$P(no - relevante) = \frac{16 + 25}{37 + 37} = 0.554$$

$$P(relevantes) = \frac{21 + 12}{37 + 37} = 0.445$$

Probalidad de acuerdo entre dos jueces (por azar)

$$P(E) = P(no - relevantes)^2 + P(relevantes)^2 = 0.306 + 0.198 = 0.504$$

Estadística Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.756 - 0.504)}{(1 - 0.504)} = 0.508$$

Campestre Americano

Proporción observada de las veces en que los jueces estuvieron de acuerdo.

$$P(A) = \frac{21 + 12}{37} = 0.891$$

$$P(no - relevante) = \frac{12 + 16}{37 + 37} = 0.378$$

$$P(relevantes) = \frac{25 + 21}{37 + 37} = 0.621$$

Probalidad de acuerdo entre dos jueces (por azar)

$$P(E) = P(no - relevantes)^2 + P(relevantes)^2 = 0.142 + 0.386 = 0.529$$

Estadística Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.891 - 0.529)}{(1 - 0.529)} = 0.768$$

Al observar los resultados de la comparación se infiere un mayor acuerdo de relevancia entre los jueces de la institución Alejandro de Humboldt, lo cual es un indicador del buen funcionamiento de la indexación semántica en el prototipo y por ende, mayor relevancia en los documentos retornados al usuario. Además, la introducción en inglés dada a los estudiantes, sobre botánica, proporcionó mejor análisis de los textos recuperados por parte de los estudiantes de la institución.

5 CUMPLIMIENTO DE OBJETIVOS

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

5.1 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 sencilla 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:

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

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.

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

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

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

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.

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

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%))*

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.

5.2 DESCRIPCIÓN Y ALCANCE DEL CUMPLIMIENTO DE LOS OBJETIVOS

En la Tabla 19Tabla 19, Tabla 20, y Tabla 21, se especifican de arriba hacia abajo 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 concerniente al "*Procedimiento para la creación de índices semánticos basados en ontologías de dominio*".

No. Objetivo	1
Descripción del objetivo	**Establecer un procedimiento para crear índices semánticos basados en ontologías de dominio específico, que defina los elementos, los requisitos y pasos a tener en cuenta para su construcción. Para ello se creará un Survey[22] en el área de la generación de índices semánticos a partir de ontologías de dominio, con una ventana de observación de los últimos cinco años, como punto partida para la definición del mismo.**
Productos esperados	1. Documento de especificación del procedimiento para crear índices semánticos, basado en ontologías de dominio. 2. Survey sobre la creación de índices semánticos enviado a una revista nacional y/o internacional donde se describirá la investigación realizada.
Resultados obtenidos	1. Documento de especificación y diagrama del procedimiento para la creación de índices semánticos basados en ontologías de dominio. 2. Realización del articulo denominado: "Survey sobre creación de Índices Semánticos" en la revista "UIS Ingenierías" perteneciente a la Universidad Industrial de Santander. Estado: en evaluación.
Indicadores (Escala * 100)	**Eficacia**

22 En ciencias de la computación, es un documento de investigación que presenta una visión general de los puntos de vista, modelos, estrategias y/o algoritmos usados por la comunidad científica para enfrentar y solucionar un problema o tema de investigación.

	$IE1 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{2}{2} * 100 = 100\%$ **Calidad** *IQ1 = ¿El procedimiento propuesto permite construir un Índice semántico basado en ontologías de dominio, utilizando técnicas de recuperación de información?* *R* = El procedimiento abstrae los pasos y actividades esenciales en la creación de índices semánticos utilizando técnicas de recuperación de información, lo cual permite al equipo de desarrollo elegir el proceso que se adecue a sus necesidades teniendo en cuenta el dominio que maneja.. $IQ1 = 4 = 100\%$ *IQ2 = ¿El impacto del procedimiento fue el esperado?* *R* = El procedimiento se ha realizado de acuerdo a la propuesta, sin embargo, se deberían realizar varias instancias de Índices Semánticos y ser aceptado por la comunidad científica, para obtener el impacto esperado. $IQ2 = 3 = 75\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 1 = \frac{100}{1} = 100\%$
Medios de verificación	1. En el capítulo 2 de la presente investigación se encuentra el procedimiento propuesto para la creación de Índices Semánticos. Además se encuentra la plantilla de definición y diagrama del mismo. 2. El Survey sobre la creación de Índices Semánticos se encuentra en el Anexo B.
Estrategias, problemas y/o observaciones	Se ha construido un procedimiento para la creación de Índices Semánticos basados en Ontologías de dominio, descrito en dos capítulos. El primero contiene las bases conceptuales para la creación y el segundo presenta la propuesta del procedimiento. El capítulo de las bases conceptuales (Anexo C) toma como referencia las investigaciones realizadas para la creación de índices semánticos, más relevantes, con el fin, de abstraer e integrar los pasos más comunes para obtener un procedimiento genérico, que permita crear índices semánticos teniendo en cuenta ontologías de dominio y permitiendo elegir un camino de creación de acuerdo a las necesidades del equipo de desarrollo o implementador. El procedimiento propuesto, identifica una serie de pasos y actividades dentro de ellos, que permiten, a los equipos de desarrollo y/o implementadores, crear Índices Semánticos en un dominio específico. Para este procedimiento se describen los pasos y sus actividades en forma de diagrama, una tabla y en prosa, lo cual permite identificar los pasos y sus actividades con la respectiva descripción y opciones a tomar en cada flujo presentado para la creación de dicho Índice. Un inconveniente encontrado fue la gran cantidad de información dispersa existente sobre el tema, esto llevó a consultar cada vez más información para abstraer los elementos necesarios sobre creación de índices semánticos. Y por lo tanto, el tiempo estimado para la definición del procedimiento se extendió un poco.

Tabla 19. Cumplimiento del primer objetivo específico

No. Objetivo	2
Descripción del objetivo	**Crear un índice semántico en un entorno particular de la educación básica primaria, que se base en el procedimiento propuesto, integrándolo al meta-buscador Group Web[23].**
Productos esperados	1. Índice Semántico basado en una ontología de dominio para un entorno educativo en el área de botánica. 2. Código fuente del Índice Semántico y aplicaciones realizadas. 3. Documentos de Análisis y Diseño del prototipo. 4. Interfaz de GruWeb interactuando con el Índice semántico creado.
Resultados obtenidos	1. Índice Semántico basado en una ontología de dominio llamado Buscador Semántico. 2. Código fuente del Índice semántico. 2.1. Buscador semántico con interfaz propia para acceder desde la web. 3. Documentos de trabajo generados para construir un índice semántico presentado como un buscador en la Web. 4. Interacción con el meta-buscador GruWeb para realizar búsquedas semánticas cuando el usuario lo requiera.
Indicadores (Escala * 100)	**Eficacia** $IE1 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{5}{4} * 100 = 125\%$ **Calidad.** *IQ1 = ¿Se construyó el Índice Semántico basado en el procedimiento propuesto?* *R* = El Índice Semántico fue desarrollado siguiendo el ciclo de vida de desarrollo software bajo la metodología UP Ágil y siguiendo uno de los caminos posibles definido en el procedimiento 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 creado se integró con el meta-buscador planteado, después de realizar la evaluación del mismo en los colegios y entorno al que está dirigido. La evaluación del buscador se realizó con interfaz propia para medir su eficiencia neta y relevancia de resultados. $IQ2 = 4 = 100\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 2 = \frac{100}{1} = 100\%$
Medios de verificación	1. En el capítulo 3 de este documento se describe la realización del Índice Semántico. Se puede acceder al buscador semántico en: ***http://prometeo.unicauca.edu.co/BuscadorSemantico/SemanticIndexSearch.aspx*** 2. El código fuente del buscador y las aplicaciones adicionales, producto de la creación del Índice Semántico, son anexados digitalmente. 3. En las secciones 3.2 y 3.3 del presente documento se describe el proceso de análisis y diseño del prototipo creado, además en el Anexo F se encuentran la documentación extendida de dicho proceso. 4. La integración con el meta-buscador GruWeb y el enlace directo al buscador

23 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/gruweb/.

	semántico se encuentra en **http://spar.unicauca.edu.co/gruweb**.
Estrategias, problemas y/o observaciones	El buscador semántico es el prototipo desarrollado bajo el procedimiento para crear Índices Semánticos. Es una herramienta de búsqueda disponible en la Web, que permite realizar búsquedas en idioma inglés sobre botánica dirigido a estudiantes de básica primaria pero que puede ser utilizado por cualquier usuario que busque y necesite resultados en esa área. La evaluación de dicho índice 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 integración son el meta-buscador.

Tabla 20. Cumplimiento del segundo objetivo específico

No. Objetivo	3
Descripción del objetivo	**Realizar una evaluación del índice creado, midiendo la relevancia de los resultados obtenidos a través de las medidas proporción recuerdo, Índice Map y estadísticas Kappa. Para esto se escogerá un grupo especifico[24] de profesores y estudiantes de escuelas del convenio CPE región sur pacifico.**
Productos esperados	1. Documento de evaluación y resultados con medidas y estadísticas definidas.
Resultados obtenidos	1. En el capítulo 4 de la presente investigación y Anexos G y H, se muestra la evaluación y resultados para cada ítem propuesto, además de la comparación en relevancia de resultados, basada en la evaluación en dos colegios de la ciudad y en el entorno al que fue dirigida la aplicación.
Indicadores (Escala * 100)	**Eficacia** $IE1 = \frac{NoProductosObtenidos}{NoProductosAObtener} = \frac{1}{1} * 100 = 100\%$ **Calidad** *IQ1 = ¿Se realizaron pruebas para verificar el funcionamiento del Índice Semántico?* *R* = Las primeras pruebas para verificar el funcionamiento del Índice fueron las preliminares (sección 3.4.1.1), alfa (sección 3.4.2.1) y beta (sección 3.5), las cuales fueron realizadas por el equipo de desarrollo y un experto en el área (director del proyecto). Ellos 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 en un colegio bilingüe de la ciudad con el fin de evaluar la eficiencia del sistema en el entorno al que está dirigido. En este caso se realizó la respectiva documentación y se definieron las medidas necesarias, encontrando un inconveniente en los resultados arrojados. Se realizó una mejora de la indexación y se llevó a cabo otra evaluación con un grupo de estudiantes y docentes de un colegio perteneciente al grupo objetivo, al cual está dirigida la aplicación, y que hace parte del convenio Computadores para Educar. $IQ1 = 4 = 100\%$ *IQ2 = ¿Las pruebas realizadas se tomaron en cuenta para las mediciones correspondientes de relevancia?*

[24] El grupo específico será seleccionado y definido en el transcurso del proyecto, ya que es necesario tener primero el procedimiento creado, el cual define los requisitos a tener en cuenta para crear los índices semánticos.

	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 en los colegios permitió calcular las estadísticas Kappa y comparar los resultados en cada uno. $IQ2 = 4 = 100\%$ **Total Cumplimiento del Objetivo (promedio eficacia)** $Objetivo\ 3 = \frac{100}{1} = 100\%$
Medios de verificación	1. En el capítulo 4 de este documento se encuentran las medidas: Curva de precisión-recuerdo (sección 4.1) Índice MAP (sección 4.2) y estadísticas Kappa (sección 4.3.1). En el Anexo G y H se muestran los formatos presentados en los colegios para cada prueba y mayor información de sobre los resultados obtenidos.
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. Se contó con dos colegios de la ciudad: Campestre Americano (bilingüe) y la Institución Educativa Alejandro de Humboldt sede Yanaconas para cumplir la evaluación con los estudiantes y docentes de dichas instituciones. 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 y docentes) sobre cada resultado específico de acuerdo a una consulta realizada.

Tabla 21. **Cumplimiento del tercer objetivo específico**

6 CONCLUSIONES, RECOMENDACIONES Y TRABAJO FUTURO

6.1 CONCLUSIONES

- Se realizó la construcción de un procedimiento para generar índices semánticos, lo cual permite a los investigadores e implementadores, decidir los pasos a seguir para realizar una indexación semántica de documentos en la web.

- Con la investigación realizada se lograron identificar las diferentes técnicas que se encuentran en el entorno de la recuperación de la información, brindándonos los enfoques necesarios con el fin de generar nuestro procedimiento para la creación de índices semánticos. Este procedimiento permite a los investigadores e implementadores, decidir los pasos a seguir para realizar una indexación semántica de documentos en la web.

- El estudio de herramientas y metodologías utilizadas para la construcción de procedimientos nos permitió una mejor abstracción e interpretación en la utilización de las mismas. A su vez, las herramientas usadas proporcionaron mayor claridad para abstraer los elementos, relaciones y pasos a seguir en la construcción del procedimiento genérico. Como producto de lo anterior fue posible establecer los pasos, requisitos y elementos necesarios e importantes que deben tenerse en cuenta a la hora de crear un índice semántico, además se señalan los posibles caminos que puede tomar un desarrollador a la hora de construir su índice semántico.

- Se observó que las funciones de similitud semántica se pueden extraer de muchas formas, como desde las bases de datos distribuidas, recuperación de la información, integración de datos y procesamiento del lenguaje natural.

- La realización del Survey titulado: “Survey sobre Creación de Índices Semánticos”, nos permitió ganar mayor capacidad de abstracción, síntesis y análisis de los procesos realizados por la comunidad científica en la indexación semántica. Este survey contiene el estado del arte actual y los procedimientos (específicos en cada proyecto) encontrados, para realizar índices semánticos; fue enviado a la revista “UIS Ingenierías” de la Universidad Industrial de Santander, y se encuentra en evaluación.

- Se pudo concluir que el uso de ontologías en la creación de índices semánticos, es de gran ayuda para mejorar los procesos de recuperación de información por sus relaciones semánticas. Por tal motivo en la

construcción del índice, se utilizó una ontología de dominio particular llamada PlantOntology (creada en inglés), para la extracción de las relaciones semánticas necesarias en la recuperación de información sobre el tema de botánica. Esto proporcionó relevancia en los resultados de las búsquedas realizadas por los diferentes usuarios de la aplicación.

- La utilización de un servicio Web de marcado social, Delicious, nos proporcionó una buena ayuda en la extracción de información relevante para las búsquedas sobre el dominio dado. La ventaja de utilizar este servicio es el impacto adecuado para lograr una buena relevancia en los resultados obtenidos del índice semántico creado.

- Se construyó un índice semántico como prototipo de la investigación que se basó en el procedimiento para la creación de índices semánticos planteado y permitió la instanciación del mismo. La implementación del procedimiento nos proporcionó una visión clara del funcionamiento del mismo y permitió comprender con mayor profundidad el manejo de las herramientas utilizadas en la recuperación de información. Con esto se proporcionó a la sociedad un sistema que permite el manejo de la información de acuerdo a un área determinada (botánica) y se dirige especialmente a los estudiantes de básica primaria.

- Al llevar a cabo la validación del prototipo con algunas medidas en la recuperación de información, nos permitió observar y considerar la eficacia de los resultados en la aplicación creada para analizar matemáticamente el grado de satisfacción que tienen los usuarios respecto al prototipo. Esta validación se realizó con las medidas Índice MAP y la curva de precisión-recuerdo, lo cual mostró buenos resultados al recuperar información para las consultas dadas. Además se realizó un estudio en dos colegios de la ciudad para calcular las estadísticas kappa, lo cual generó buenos resultados al final de las evaluaciones realizadas.

6.2 RECOMENDACIONES

Para realizar una indexación semántica en sistemas de recuperación de información, es necesario verificar las herramientas de implementación con las que se trabaje, pues el procedimiento proporciona los pasos y actividades opcionales a seguir según las necesidades del investigador y/o implementador, pero está fuera del dominio de este proyecto, indicar las herramientas de implementación para cada etapa del proceso. Solo se expresan las herramientas utilizadas en para el prototipo creado como experimentación de este proyecto.

Si se enfoca la creación de índices semánticos en la educación, sería bueno indexar con ontologías en el idioma español para estudiantes de habla hispana, pues esto facilita la recuperación de información teniendo en cuenta la relevancia y relaciones de sus resultados con sus consultas. Esto también ayudaría si la enseñanza de otro idioma en algunas instituciones educativas no es intensiva.

6.3 TRABAJO FUTURO

Para un trabajo futuro sería interesante dar más opciones de escoger el dominio (tema) entre varias ontologías disponibles para incrementar el entorno de trabajo con los usuarios. Así, permitiría la ampliación de búsquedas y resultados en otros temas que los usuarios requieran.

Puede ampliarse el índice semántico tomando como base no solo una ontología sino también un tesauro para obtener resultados e información en otros dominios o ámbitos del conocimiento.

Sería interesante realizar la construcción de un índice semántico basado en un repositorio de documentos no estructurados y comparar los resultados con el índice creado en esta investigación. Esto proporcionaría una estimación adecuada del valor real al usar Delicious como servicio de marcado social para obtener mejores resultados.

7 REFERENCIAS

1. Jacquin, E.D.C., *Indexing a Web Site with a Terminology Oriented Ontology, Extraido el 10 de Enero 2010*. 2002, IRIN, Université de Nantes 2: Standford University. p. 181-198.
2. Mihalcea Rada, M.D., *Semantic Indexing using WordNet Senses* in *Department of Computer Science and Engineering*. 2000, Southern Methodist University: In Proceedings Of Acl Workshop On Ir & Nlp, Hongkong. p. 11.
3. Thanh Nguyen, T.P., *The effect of Semantic Index in Information Retrieval development*, in *International Conference on Information Integration and web-based Applications and Services*. iiWAS 2008, ACM: Austria. p. 438-441.
4. José M. Diaz N., F.S., Mario Pérez. *Recuperación de Información.* . 2009 [cited 14 de julio de 2010]; Available from: http://sites.google.com/site/glosariobitrum/Home/recuperacion-de-informacion.
5. Molina, M.P. *Búsqueda y Recuperación de Información*. 2009 [cited 27 de abril de 2010]; Available from: http://www.mariapinto.es/e-coms/recu_infor.htm.
6. Shahrul Azman Noah, L.Z., Arifah Che Alhadi, Tengku Mohd Tengku Sembok, Saidah Saad, *Towards Building Semantic Rich Model for Web Documents Using Domain Ontology, extraido el 10 de marzo de 2010*, in *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*. 2004, National University of Malaysia. p. 769 - 770.
7. E. Desmontils, C.J., L. Simon, *Ontology enrichment and indexing process*. 2003, Institut de Recherche en Informatique de Nantes 2, rue de la Houssinire. p. 18.
8. Samaneh CHAGHERI, C.R., Sylvie CALABRETTO, Cyril DUMOULIN, *Semantic Indexing of Technical Documentation*, in *Laboratoire d'InfoRmatique en Image et Systèmes d'information*. 2009, Université de LYON: Toulouse, France. p. 12.
9. Song Jun-feng, Z.W., Xiao W., Li G., Xu Z, *Ontology-Based Information Retrieval Model for the Semantic Web*, in *Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05) on e-Technology, e-Commerce and e-Service*. 2005, IEEE Computer Society: Washington, DC, USA. p. 152 - 155.
10. Marie Aude Aufaure, R.S., Hajer Baazaoui, *SIRO: ON-LINE SEMÁNTICA INFORMATION RETRIEVAL USING ONTOLOGIES, Extraido el 10 de octubre de 2010.* IEEE, 2007.

11. Mingxia Gao, C.L., Furong Chen, *An Ontology Search Engine Basedon Semantic Analysis, Extraido el 20 de octubre de 2010.* IEEE, 2008.
12. David Vallet, M.F., Pablo Castells, *An Ontology-Based Information Retrieval Model.* IEEE.
13. Yi-Ming Chung, Q.H., Kevin Powell and Bruce Schatz, *Semantic Indexing for a Complete Subject Discipline, Extraido el 10 de febrero de 2010*, in *Proceedings of the fourth ACM conference on Digital libraries* 1999, University of Illinois at Urbana-Champaign, Champaign, IL 61820: International Conference on Digital Libraries, Berkeley, California, United States p. 39-48.
14. Conrad T. K. Chang, B.R.S., *Performance and Implications of Semantic Indexing in a Distributed Environment, Extraido el 10 de marzo de 2010*, in *Proceedings of the eighth international conference on Information and knowledge management* 1999, ACM, New York: Kansas City, Missouri, United States. p. 391-398.
15. Duygu Tümer, M.A.S., Yiltan Bitirim *An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia.* IEEE, 2009: p. 51-55.
16. W3C, C. *OWL Web Ontology Language. Extraido el 2 de febrero de 2010*. 2004 [cited; Available from: http://www.w3.org/TR/owl-features/.
17. García, G.G. *RDF y RDF schema*. 2003 [cited 14 de abril de 2010]; Available from: http://www.matem.unam.mx/~grecia/semantic_web/rdf.html.
18. W3C, C. *XML Schema*. 2007 [cited 21 de mayo de 2010]; Available from: http://www.w3.org/XML/Schema.
19. W3C, C. *Resource Description Framework (RDF). Extraído el 10 de febrero de 2010*. 2004 [cited; Available from: http://www.w3.org/RDF/.
20. Cover, R. *Ontology Interchange Language (OIL). Extraido el 10 de febrero de 2010*. Technology Reports 2000 [cited; Available from: http://xml.coverpages.org/oil.html.
21. Tuggy, D. *Lecciones para un curso del náhualt moderno* 2002 [cited 12 de mayo de 2010]; Available from: http://www.sil.org/~tuggyd/nahuatllecciones/l07/lecc_07_nlv.htm.
22. López, J.F. *Orden de Palabras en la Frase Nominal*. 2009 [cited; Available from: http://culturitalia.uibk.ac.at/hispanoteca/Gram%C3%A1ticas/Gram%C3%A1tica%20alemana/Orden%20de%20palabras%20-%20frase%20nominal.htm.
23. Miguel A. Alonso Pardo, J.V.F. *Introducción a la Recuperación de Información*. 2010 [cited 9 de junio de 2010]; Available from: http://www.grupolys.org/docencia/ln/biblioteca/ir.pdf.
24. M. Teresa ROMÁ-FERRI, M.P., *Interoperabilidad Semántica de Ontologías Basada en Técnicas de Procesamiento del Lenguaje Natural.* ISKO. CAPÍTULO ESPAÑOL. CONGRESO 7°, 2005: p. 534-548.
25. Leal, E.T. *La Desambiguación del Sentido de las Palabras: revisión metodológica*. Revista multidisciplinar sobre diseño, personas y tecnología

2009 [cited 10 de marzo de 2010]; Available from: http://www.nosolousabilidad.com/articulos/desambiguacion.htm.
26. C., H., *Un modelo de investigación documental*, ed. S. Editora. 2000. 67.
27. Lyn, G., *Integracion de las ontologias, Extraido el 5 de febrero de 2010.* Blog Lyn, 2008.
28. López, S.E.S., *Modelo de indexación de formas en sistemas VIR basado en ontologías: Ontologías y su Representación Jerárquica*, in *Departamento de Computación, Electrónica y Mecatrónica*. 2007, Universidad de las Américas Puebla: Cholula, Puebla. p. 13, cap. 4.
29. Christopher D. Manning, P.R., Hinrich Schütze, *An Introduction to Information Retrieval, Extraído el 5 de mayo de 2010*. 2009, Cambrige University Press: Cambrige. p. 581.
30. Ricardo Baeza-Yates, B.R.-N., *Modern Information Retrieval, Extraído el 4 de mayo de 2010*, A.-. Wesley, Editor. 1999, ACM Press: New York.
31. Miguel A. Alonso, J.G., Jesús Vilares, *Recuperación de Información en Internet. Extraído el 10 de mayo de 2010*, Universidad de Coruña: Coruña. p. 18.
32. Suárez Barón Marco, S.V.K., *An Approach to Semantic Indexing and Information Retrieval, Extraido 10 de diciembre de 2009.* Revista Facultad de Ingeniería Universidad de Antioquia, 2009. **48**: p. 14.
33. Dr. Clara Yu, D.J.L.C., Aaron Coburn. *The Semantic Indexing Project* knowledgesearch 2003 [cited 12 de Diciembre de 2009]; Available from: http://www.knowledgesearch.org/.
34. Resnik, P. (1995) *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. **Volume**, 6
35. Laurent Mazuel, N.S., *Semantic Relatedness Measure Using Object Properties in an Ontology*, in *Proceedings of the 7th International Conference on The Semantic Web*. 2008, ACM: Karlsruhe, Germany. p. 681-694.
36. Gang Lv, C.Z., Li Zhang, *Text Information Retrieval Based on Concept Semantic Similarity.* 2009 Fifth International Conference on Semantics, Knowledge and Grid, 2009: p. 356-360.
37. Lin, D., *An Information-Theoretic Definition of Similarity.* Proc 15th International Conference on Machine Learning, 1998: p. 296-304.
38. Valentina Cordı, P.L., Maurizio Martelli and Viviana Mascardi (2005) *An Ontology-Based Similarity between Sets of Concepts*. **Volume**,
39. Wei Song, C.H.L., Soon Cheol Park, *Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures.* Expert Systems with Applications, 2009. **36**: p. 9095–9104.
40. Cobos L. Carlos A., A.R.J.K., Constaín D. William A., *Hibridación De La Mejor Búsqueda Armónica Global Y El Algoritmo K-Means Para El Clustering De Documentos Web*, in *Departamento de sistemas*. 2009, UNIVERSIDAD DEL CAUCA: Popayan. p. 85.

41. Powell, K.R. *The Interspace Prototype: An anaysis Environment for Semantic Interoperability*. 1998 [cited 10 mayo de 2010]; Available from: http://www.canis.uiuc.edu/INTERSPACE/.
42. Company, R. *infoREUSER - Motor de búsqueda semántico*. 2010 [cited 12 de abril de 2010]; Available from: http://www.reusecompany.com/producto.aspx?id=4.
43. Gómez, E.E., *Una nota metodológica sobre los análisis cualitativos. El análisis de las relaciones entre los elementos: el análisis de las frecuencias y co-ocurrencias.* Sistema de Información Científica, Redalyc, 2009. **18**: p. 57.
44. Moreno, F.S.R.S.F.A.A.A.L.P., *Nueva Propuesta de Desambiguación de Sentidos de Palabras para nombres en un sistema de Búsqueda de Respuestas.* Procesamiento del Lenguaje Natural, 2006. **36**: p. 47-53.
45. Rada, M.D.I.M., *An Iterative Approach to Word Sense Disambiguation*, in *Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference* 2000, AAAI Press: Orlando, FL. p. 219 - 223.
46. H. Chen, B.S., D. Ng, J. Martinez, A. Kirchhoff, C. Lin., *A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996. **18**(8): p. 39.
47. Hsinchun Chen, A.H., R. Sewell, y B. Schatz., *Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques.* Journal of the American Society for Information Science, 1998. **49(7)**: p. 582-603.
48. Cherukuri Aswani Kumar, S.S., *Latent semantic indexing using eigenvalue analysis For efficient information retrieval.* Int. J. Appl. Math. Comput. Sci, 2006. **16**: p. 551-559.
49. Brill, E., *A Corpus-Based Approach to Language Learning*, in *The Institute For Research In Cognitive Science*. 1993, University of Pennsylvania: Philadelphia. p. 166.
50. Translations, L. *NPtool, a detector of English noun phrases*. 1993 [cited 4 de mayo de 2010]; Available from: http://www2.lingsoft.fi/doc/nptool/.
51. Stuart Nelson: Head, M.S.H. *Medical Subject Headings*. 2010 [cited 14 de abril de 2010]; Available from: http://www.nlm.nih.gov/mesh/.
52. Kang, B.-Y., *A Novel Approach to Semantic Indexing Based on Concept, Extraido el 15 de marzo de 2010*, in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*. 2003, Association for Computational Linguistics: Sapporo, Japan. p. 44-49.
53. Jane Morris, G.H., *Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text*, in *Computational Linguistics*. 1991. p. 21 - 48.
54. Barite, M. *Diccionario de Organización y representación del Conocimiento: Clasificación, Indización, terminología.* 2000 [cited 10 de mayo de 2010]; Available from: http://www.eubca.edu.uy/diccionario/letra_h.htm.

55. Costa, J. *Meronimia*. 2006 [cited 10 de mayo de 2010]; Available from: http://www.solodisenio.com/que-es-meronimia/.
56. Hsinchun Chen, K.J.L., *Automatic Construction of Networks of Concepts characterizing Document Databases.* IEEE Transactions on Systems, Man, and Cybernetics, 2002. **22**(5): p. 885 - 902.
57. Schatz, B.R., *Information Retrieval in Digital Libraries: Bringing Search to the Net.* Science - Bioinformática, 1997. **275**: p. 327 - 334.
58. Cutting, D. *The Apache Lucene*. Actualizado 2010 [cited 14 de abril de 2010]; Available from: http://lucene.apache.org.
59. George A. Miller, R.B., Christiane Fellbaum, Derek Gross, and Katherine Miller, *Introduction to WordNet: An On-line Lexical Database.* International Journal of lexicography, 1993. **3**: p. 235-244.
60. Joerg-Uwe Kietz, A.M., Raphael Volz, *A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet*, in *Swisslife Information Systems Research Lab, Zuerich, Switzerland*. 2000, AIFB, Univ. Karlsruhe: Karlsruhe, Germany. p. 2-6.
61. T. O'Hara, K.M., and S. Niremburg, *Lexical Acquisition with WordNet and Microkosmos Ontology*, in *Proceedings of the ACL Workshop on the Use of WordNet in NLP*. 1998. p. 94-101.
62. Brill, E., *Transformation-based error-driven learning and natural language processing: a case study in Part-of-speech Tagging* Computational Linguistics, 1995. **21**(4): p. 543-565.
63. University, P. *WordNet 3.0* Princenton University 2009 [cited 2 de febrero de 2010]; Available from: http://wordnet.princeton.edu/wordnet.
64. S Cazalens, E.D., C Jacquin, and P Lamarre, *A Web Site Indexing Process for an Internet Information Retrieval Agent System* in *Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00)*. 2000: Hong Kong , China. p. 254-258.
65. Satoshi Sekine, R.G. *Apple Pie Parser - Proteus Project*. 2002 [cited 21 de abril de 2010]; Available from: http://nlp.cs.nyu.edu/app/.
66. Miriam Fernández, T.P.C.A., *Proyecto de trabajo de Iniciación a la investigación*, in *Escuela Politécnica Superior*, Universidad Autónoma de Madrid: Madrid. p. 5.
67. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., *KIM – A Semantic Platform for Information Extaction and Retrieval, Extraido el 3 de marzo de 2010.* Journal of Natural Language Engineering, 2004. **10**: p. 375-392.
68. Kiryakov, A., Popov, B., Terziev, I., Manov, Ognyanoff, D., *Semantic Annotation, Indexing, and Retrieval. Journal of Web Sematics* Journal of Web Semantics, 2004. **2**: p. 49-79.
69. Castells, P., Fernández, M., Vallet, D., Mylonas, P., Avrithis, Y., *Self-Tuning Personalized Information Retrieval in an Ontology-Based Framework, Extraido el 20 de febrero de 2010*, in *1st IFIP International Workshop on Web Semantics (SWWS 2005)*. 2005: Agia Napa, Cyprus. p. 977-986.

70. James Mayfield, T.F., *Information retrieval on the Semantic Web: Integrating inference and retrieval*, in *In Proceedings of the SIGIR Workshop on the Semantic Web Workshop - ACM*. 2003, The Johns Hopkins University and University of Maryland: Toronto, Canada. p. 7.
71. Junior Sinche, J.F. *Servicios Web Semánticos.* . 2008 [cited 12 de mayo de 2010]; Available from: http://www.slideshare.net/guesta5bc77/servicios-web-semnticos-presentation.
72. Alexander Maedche, S.S., Nenad Stojanovic, Rudi Studer, and York Sure, *SEmantic portAL — The SEAL approach*, in *Spinning the Semantic Web*. 2001. p. 27.
73. Gerard Salton, M.J.M., *Introduction to Modern Information Retrieval*. 1986, McGraw-Hill: New York. p. paginas 400.
74. Francisco Ruiz, J.V., *Guía de Uso de SPEM 2 con EPF Composer*. 2008, Universidad de Castilla-La Mancha.
75. Shapiro, R. *Busines Process Modeling Notation y estándares de portabilidad en procesos*. 2009 [cited 9 de julio de 2010]; Available from: http://www.club-bpm.com/BPMday/BPMDaySeminarioBPMN3Dic09.pdf.
76. *Estándares IDEF0, IDEF1, IDEF3*. 2008 [cited 10 de agosto de 2010]; Available from: http://www.pdca.es/pruebas/idef.html#.
77. Group, O.M., *Software & Systems Process Engineering Meta-Model Specification*. 2008.
78. Ruiz, F., *Introducción a la Ingeniería de Procesos Software*. 2008, Universidad de Cantabria.
79. *Diagrama de flujo de proceso*. 2009 [cited 10 de agosto de 2010]; Available from: http://148.202.148.5/cursos/id209/mzaragoza/unidad2/unidad2tres.htm.
80. Vazquez, A.M. *Herramientas organizacionales. Diagrama de flujo*. 2010 [cited 6 de septiembre de 2010]; Available from: http://www.estrucplan.com.ar/Producciones/entrega.asp?IDEntrega=526.
81. Huanca, J.C. *Lenguaje Unificado de Modelado. Diagramas*. 2010 [cited 18 de agosto de 2010]; Available from: http://www.grupoinformatica.com/biblioteca-articulos/1459-uml-lenguaje-unificado-de-modelado.html.
82. Educación, M.d. *FreeMind: mapas conceptuales*. 2009 [cited 11 de agosto de 2010]; Available from: http://recursostic.educacion.es/observatorio/web/es/software/software-general/716-freemind-mapas-conceptuales.
83. Gómez, L. *Software para elaborar mapas mentales y conceptuales*. 2008 [cited 11 de agosto de 2010]; Available from: http://manantialdevida.obolog.com/software-elaborar-mapas-mentales-conceptuales-59333.
84. Ecourban. *Mapas conceptuales*. [cited 10 de agosto de 2010]; Available from: http://www.ecourban.org/profesores/didactica/mapasconceptuales/index.html.

85. Jörg Müller, D.P. *FreeMind - free mind mapping software*. 2010 [cited 3 de septiembre de 2010]; Available from: http://freemind.sourceforge.net/wiki/index.php/Main_Page.
86. Sánchez, P.H. *Lenguaje de Definición de Datos II: Definición de Indices. Extraido el 25 de mayo de 2010*. 2008 [cited; Available from: http://www.devjoker.com/contenidos/catss/12/Indices.aspx.
87. Carlos A. Cobos L., M.A.N.Z., *Metamodelo de evaluación para la educación en linea*, in *SITI 2003*. 2003: Popayán, Colombia. p. 10.
88. Vadim P. Madrid G., Á.F.Z., Carlos G. Figuerola, José L. Alonso B., *Librerías Lucene y dotLucene para Recuperación de Información. Estudio y desarrollo de casos prácticos. Extrído el 5 de mayo de 2010*, in *departamento de Informática y automática*. 2007, Universidad de Salamanca. p. 34.
89. Losada, D.E., *Recuperación de Información*, D.d.E.y. Computación, Editor, Universidad de Santiago de Compostela.
90. Sequeira, Y.C., *Procesamiento del lenguaje natural para recuperar Información. Extraido el 22 de mayo de 2010.* 2008.
91. Gastaminza, F.d.V., *Diseño y Desarrollo de Thesaurus. Extraído el 5 de mayo de 2010*, Universidad Complutense de Madrid: Madrid.
92. Lucene. *Indexación de Información. Lucene. Extraído el 10 de abril de 2010*. [cited; Available from: http://trevinca.ei.uvigo.es/~pcuesta/sm/practicas/Lucene.pdf.
93. Microsoft, S. *Prácticas recomendadas para elegir un idioma al crear un índice de texto completo*. 2008 [cited.
94. Broncano, R.G. *Modelos de Recuperación, Recuperación y Organización de la Información.* 2006 [cited 10 de mayo de 2010]; Available from: http://modelosrecuperacion.tripod.com/.
95. Lapuente, M.J.L., *Hipertexto: El nuevo concepto de documento en la cultura de la imagen. Extraído el 9 de junio de 2010*. 2010, Universidad Complutense de Madrid: Madrid.
96. Carrascal, C. *Tesauros y Ontologías*. 2004 [cited 7 de abril de 2010]; Available from: http://personales.upv.es/ccarrasc/doc/2003-2004/TesaurosOnto/principal.html.
97. Jiménez, A.G., *INSTRUMENTOS DE REPRESENTACIÓN DEL CONOCIMIENTO: TESAUROS VERSUS ONTOLOGÍAS. Extraído el 9 de junio de 2010.* ANALES DE DOCUMENTACION, 2004: p. 79-95.
98. Avello, D.G., *Una nueva técnica para procesamiento de texto no estructurado mediante vectores de n-gramas de longitud variable con aplicación a diversas tareas de tratamiento de lenguaje natural*, in *Departamento de Informática*. 2005, Universidad de Oviedo. p. 241.
99. Herrera, D.A.G.L., *Modelos de Sistemas de Recuperación de Información Documental Basados en Información Linguística Difusa*, in *Departamento de Ciencias de la Computación e Inteligencia Artificial*. 2006, Universidad de Granada. p. 255.
100. Wesley, A., *Search Engines: Information Retrieval in Practice*. 2008.

101. Joydeep Ghosh, D.L., *Performance Evaluation of Information Retrieval Systems*, Univ. of Science and Tech.
102. José Jimeno Yepes, R.B.L., *Ontology Refinement for Improved Information Retrieval in the Biomedical Domain*, in *Depto. de Lenguajes y Sistemas Informáticos*. 2009, Universitat Jaume: Castellón. p. 165.
103. Amorós, D.J.F., *Anotación Semántica no supervisada*, in *Departamento de Lenguajes y Sistemas Informáticos*. 2004, Universidad Nacional de Educación a Distancia: Madrid.
104. Hale, M.L.M. (1998) *A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity*. In Proceedings of the Second International Conference on Information and Knowledge Management (CIKM), **Volume**,
105. *Plant Ontology*. 2010 [cited 10 de agosto de 2010]; Available from: http://www.plantontology.org/].
106. Ricardo J. Vargas Del Valle, J.P.M.G., *Programación en Capas*. 2007, Universidad de Costa Rica. p. 5.
107. César de la Torre L., U.Z.C., Miguel Ángel Ramos B., Javier Calvarro N., *Guía de Arquitectura N-Capas orientada al dominio con .Net*, K. Consulting, Editor, Microsoft Ibérica S.R.L.: Madrid.