Procedimiento para la construcción de Índices Semánticos basados en ontologías de dominio específico[1]

Procedure for building semantic indexes based on domain-specific ontologies

Procedimento para a construção de índices semânticos baseados em ontologias de domínio específico

Miguel Ángel Niño Zambrano
Magíster en Informática de la Universidad Industrial de Santander. Profesor Titular del Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca. Popayán, Colombia
manzamb@unicauca.edu.co.

Dignory Jimena Pérez
Ingeniera de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca. Ingeniera de Desarrollo de la Empresa Carvajal.
dicnory7@hotmail.com

Diana Maribel Pezo
Ingeniería de Sistemas, Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca. Consultora de Proyectos Software de SCAD Colombia.
mariluna0212@hotmail.com

Carlos Alberto Cobos Lozada
Magíster en Informática de la Universidad Industrial de Santander. Profesor Titular del Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca. Popayán, Colombia.
ccobos@unicauca.edu.co

Gustavo Adolfo Ramírez González
PhD en Ingeniería Telemática de la Universidad Carlos III de Madrid-España. Profesor Titular del Departamento de Telemática, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca.
gramirez@unicauca.edu.co

Resumen

Los sistemas de búsqueda web actual, aún están lejos de ofrecer respuestas completamente contextualizadas y precisas a los usuarios, ya que éstos deben hacer esfuerzos adicionales de filtrado y evaluación de la información proporcionada. Una forma de mejorar los resultados, es mediante la creación de índices semánticos, los cuales incorporan conocimiento y procesamiento inteligente de los recursos. Sin embargo, al momento de implementar los índices semánticos, existen variadas investigaciones con procedimientos propios y con procesos largos de conceptualización, implementación y afinación. Es así, como se vuelve importante definir una herramienta que permita crear este tipo de estructuras de una manera más estructurada y eficiente. El presente trabajo propone un procedimiento que permite crear índices semánticos a partir de ontologías de dominio específico. La metodología utilizada fue la creación de un estado del arte de las diferentes propuestas existentes y posteriormente la abstracción de un procedimiento general que incorpore las mejores prácticas de creación de índices semánticos. Posteriormente, se creó un índice semántico el dominio de las plantas y sus componentes. Los resultados permiten establecer que el proceso definido es una buena herramienta para guiar la implementación de este tipo de estructuras con un alto grado de personalización. Sin embargo, también evidenció que el proceso depende otras variables al momento de construir y trabajar con el índice y por lo tanto se debe reevaluar el diseño hasta obtener los resultados deseados.

Palabras clave

Indexación semántica, Ontología, recuperación de información, marcado colaborativo.

Fecha de recepción: 25 - 11- 2012 **Fecha de aceptación: 03 - 02 - 2013**

Abstract

The current on-line search systems are still far from providing users with contextualized and accurate answers because users have to make additional efforts to filter and evaluate information supplied to them. One of the ways to improve the results is to create semantic indexes that incorporate knowledge and intelligent processing of resources. When it comes to the implementation of semantic indexes, however, there is a wide range of research studies with their own procedures and lengthy conceptualization, implementation, and refinement processes. Thus, it becomes of the utmost importance to define an instrument that allows creating these kinds of structures in a more structured and efficient manner. This work proposes a procedure that makes it possible to create semantic indexes based on domain-specific ontologies. The methodology entailed creating a state of the art of the various existing proposals and drawing a general procedure that incorporates the best practice for creating semantic indexes. Then, a semantic index was created of the domain of plants and their components. The results demonstrate that the defined process is a good instrument that guides implementation of these kinds of structures with a high degree of customization. Nevertheless, it also shows that the process depends on other variables in building and processing the index, so the design needs to be re-examined until the desirable results are obtained.

Keywords

Semantic indexing, Ontology, information retrieval, collaborative market.

Resumo

Os atuais sistemas de busca na web, estão ainda longe de fornecer respostas plenamente contextualizadas e precisas aos usuários, uma vez que eles devem fazer esforços extras de filtragem e avaliação das informações fornecidas. Uma forma de melhorar os resultados é através da criação de índices semânticos, que incorporam conhecimento e processamento inteligente dos recursos. No entanto, no momento de implementar os índices semânticos, existem variadas investigações com procedimentos próprios e com longos processos de conceituação, implementação e ajuste. É assim que se torna importante definir uma ferramenta que permita criar este tipo de estruturas de uma maneira mais estruturada e eficiente. Este artigo propõe um procedimento que permite criar índices semânticos a partir de ontologias de domínio específico. A metodologia usada foi a criação de um estado de arte das diferentes propostas existentes e posteriormente a abstração de um procedimento geral que incorpore as melhores práticas de criação de índices semânticos. Posteriormente, foi criado um índice semântico de masterização das plantas e seus componentes. Os resultados permitem estabelecer que o processo definido é uma boa ferramenta para orientar a implementação deste tipo de estruturas com um alto grau de personalização. No entanto, também revelou que o processo depende de outras variáveis no momento de construir e trabalhar com o índice e, portanto, o projeto deve ser reavaliado até obter os resultados desejados .

Palavras-chave

Indexação semântica, Ontologia, recuperação de informação, marcado colaborativo.

Introducción

Desde hace más de una década, varios proyectos de investigación (Desmontils & Jacquin, 2002; Mihalcea Rada, 2000; Thanh Nguyen, 2008), han propuesto diversas soluciones para mejorar la relevancia (Molina, 2009; N., Salto, & Pérez, 2009) de los documentos recuperados en sistemas de búsqueda Web, desarrollando o mejorando las técnicas actuales de recuperación de información - RI. Uno de los enfoques utilizados se ha denominado búsqueda semántica - BS. En este enfoque se han utilizado diferentes técnicas, destacándose el uso de ontologías y la construcción de índices semánticos - IS; estos últimos se han empleado en diversos estudios (Desmontils, Jacquin, & Simon, 2003; Samaneh Chagheri, 2009; Shahrul Azman Noah, 2004; Song Jun-feng, 2005), en donde la semántica de los conceptos es el principal problema a resolver.

A pesar de la importancia que tienen los índices semánticos en los sistemas de recuperación de la información - SRI actuales, los investigadores deben recurrir a un proceso largo de sensibilización y entendimiento en su construcción y uso, dificultando así el desarrollo de nuevas investigaciones en el área particular.

Por lo anterior, en esta investigación se analizaron diversos proyectos (Aufaure, Soussi, & Baazaoui, 2007; Chung, He, Powell, & Schatz, 1999; Gao, Liu, & Chen, 2005; Nguyen & Phan, 2008; Samaneh Chagheri, 2009; Tumer, 2009; Vallet, Fernández, & Castells, 2005) con la perspectiva de abstraer un procedimiento para

generar IS, el cual sirve como soporte para el desarrollo de aplicaciones que tengan por objetivo mejorar la relevancia de los resultados obtenidos en la RI Web.

A continuación, en la sección 2 se presentan los fundamentos teóricos de la investigación, se inicia con conceptos básicos de recuperación de información, se explica luego la forma clásica de indexación y termina con un conjunto representativo de trabajos de investigación relacionados con la presente propuesta. En la sección 3, se muestra en detalle el procedimiento para la creación de los índices semánticos y en la sección 4 una plantilla que facilita la instanciación del procedimiento a un caso de estudio específico. La sección 5 muestra un caso de estudio del procedimiento propuesto en el entorno de las plantas y la sección 6 presenta los resultados de la evaluación del sistema de búsqueda que usa el índice semántico específico que se construyó. Finalmente se revelan las conclusiones de la investigación y el trabajo futuro que el grupo de investigación espera desarrollar.

1. Fundamentación teórica

Esta sección presenta los principales aspectos teóricos que se tuvieron en cuenta para la definición del procedimiento. Primero se abordan conceptos básicos de RI e Indexación y luego se analizan los trabajos previos más representativos relacionados con la investigación. Dado que el estudio de los trabajos previos se terminó a finales del 2010, no se dan resultados de investigaciones reportadas en fechas posteriores.

1.1. Conceptos básicos de recuperación de la información

Según la ISO 5963 (ISO, 2000), un documento es cualquier fuente de información – en este caso digital -, que se puede catalogar e indizar. A su vez, la norma define un concepto como una unidad de pensamiento, el cual se puede expresar como una combinación de otros conceptos. Así mismo, un tema es representado por un conjunto de conceptos y los documentos pueden tratar diversos temas. La norma también define la indización como la "Acción de describir o identificar un documento en relación con su contenido", es decir, la indización se entiende como la forma de representar el contenido de un documento mediante un conjunto de términos (o conceptos), que especifican los temas de los que trata el documento. Además, el diccionario de la Real Lengua Española (RAE, 2011), define que la indización también se conoce como indexación y ésta como la acción o efecto de indexar, a su vez indexar consiste en "Hacer índices" y "Registrar ordenadamente datos e informaciones, para elaborar su índice".

Teniendo en cuenta las definiciones anteriores, para la presente investigación, el hecho de crear un índic, implica hacer un proceso de representación de las fuentes documentales mediante un conjunto de términos o conceptos, de tal forma que se pueda registrar ordenadamente los temas de los que trata dicho documento con el fin de permitir una clasificación y consulta rápida de los mismos. En este caso, la indexación semántica implica que adicionalmente se utilizan herramientas de la gestión del conocimiento (técnicas de procesamiento de lenguaje natural, vocabularios controlados, tesauros y ontologías) y herramientas de la Web Semántica (anotación semántica, marcado social, entre otros), para encontrar los conceptos que representen con mayor precisión los documentos, de tal forma que su indexación permita consultas con un alto grado de relevancia (utilidad) para el usuario.

A continuación se definen los principales conceptos clave que serán utilizados en las siguientes secciones:

- ***Representatividad:*** **Se refiere a especificaciones (léxicas, sintácticas o semánticas) representadas con etiquetas, que permiten identificar el contenido en la Web entre un gran número de artículos disponibles (Barite, 2000). 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 con los requerimientos de información.**

- ***Frase Nominal:*** **Es un grupo de palabras organizadas alrededor de un sustantivo (núcleo) (Barite, 2000). La frase nominal puede consistir en un único sustantivo, o en adjetivos de varios tipos que lo modifican, y también puede ser encabezada por un pronombre. Los elementos que acompañan al sustantivo nuclear se llaman adyacentes y cumplen una función atributiva respecto al núcleo.**

- ***Tokenizador:*** **Es un algoritmo que realiza la conversión de una secuencia de caracteres a una secuencia**

de palabras candidatas (tokens) a ser tomadas para el índice de un sistema de recuperación de información. Identifica las palabras que contienen los documentos (Pardo & Ferro, 2010). Realiza la remoción de caracteres especiales como "/ \ - : ?¿;)(&#", entre otros procesos.

- **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 (Roma-Ferri & Palomar, 2005), 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 confundirse en el sentido de la oración (Leal, 2009).**

- **Análisis de Co-ocurrencia: Método automático que permite, a través del análisis de documentos, establecer el número y grado de apariciones simultáneas de palabras o grupos de palabras en conjuntos de documentos, así como la distancia a la que ocurren. La co-ocurrencia permite establecer términos de indización en proporción directa a la frecuencia de aparición de los mismos (Barite, 2000).**

- **Espacio conceptual: Se puede entender como una red de términos cuyas asociaciones se encuentran ponderadas. Este concepto se ha utilizado, entre otras cosas, para crear tesauros conceptuales automáticos (Chen *et al.*, 1996). Por otro lado se habla también de los espacios conceptuales como un conjunto de "atributos de calidad" (Gärderfors, 2004) derivados de los mecanismos de percepción, con los cuales se puede representar diversos tipos de información para el aprendizaje de conceptos. La red de conceptos es modelada a partir de representaciones geométricas a diferencia de las representaciones simbólicas o conexionistas que presenta la teoría cognitiva. Sin embargo existen otros autores (Chang & Schatz, 1999) que consideran el espacio conceptual como un índice de una colección que utiliza las estadísticas del documento para capturar las relaciones entre los conceptos, desarrollando algoritmos que utilizan correlaciones estáticas en los mismos documentos para encontrar las interrelaciones entre los conceptos.**

- **Similitud Semántica: Es un proceso cognitivo en el cual se establece qué tan cercanos son dos conceptos, en el cual se supone que existe un distancia entre ellos y cuanto más corta sea dicha distancia, los conceptos son más similares. A su vez, la representación de los conceptos puede definir varias dimensiones y generar una estructura, esta estructura se puede formalizar con un espacio conceptual (Gärderfors, 2004). En la RI, la similitud semántica se utiliza fundamentalmente para establecer la distancia de conceptos en un tesauro, ontología o entre los términos de la consulta original y los términos que representan los documentos (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Lin, 1998; Lv, Zheng, & Zhang, 2009; Mazuel & Sabouret, 2008; Resnik, 1995; Song, Li, & Park, 2009).**

1.2. Conceptos básicos de Indexación

Es necesario diferenciar la indexación semántica de la indexación tradicional, esta última aplica generalmente variaciones de modelos clásicos de RI (Modelo booleano, Espacio vectorial y Probabilístico) (Dominich, 2000). Los primeros SRI se enfocaron en calcular la relevancia de los documentos, a través del cálculo de los pesos locales o globales de todos los términos encontrados en los documentos (Salton & McGill, 1986), para finalmente compararlos con los de la consulta y así entregar una lista de documentos ordenados descendentemente por relevancia, sin realizar un análisis de los términos, para clasificarlos y encontrar sus temas y/o conceptos, lo cual sí se hace en la indexación semántica.

En la Figura 1 se presenta el proceso general de indexación clásica, 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 (stopwords), palabras muy frecuentes en la colección de documentos, permitiendo de esta manera reducir el número de términos con poco valor en la recuperación de información, entre las palabras vacías están: artículos, preposiciones, conjunciones, entre otras. Posteriormente se pasa a la lematización, en la cual se eliminan prefijos y sufijos, en esta fase se extrae el lexema (steam) o raíz de cada término o palabra extraída del paso de

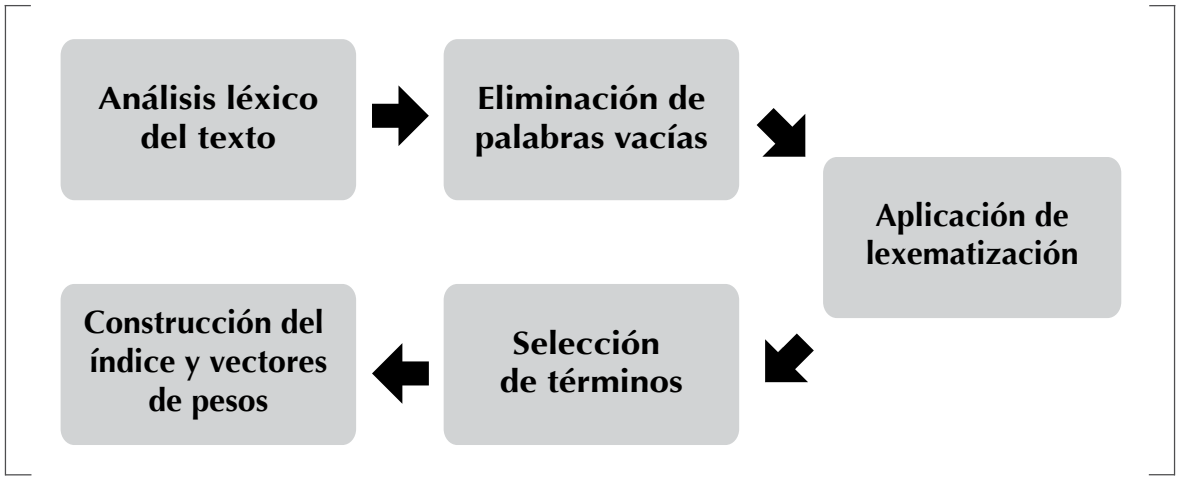


Figura 1. Proceso de indexación clásica **Fuente:** Presentación propia de los autores

análisis léxico, por ejemplo, el verbo sin conjugar o la palabra en singular de dicho término. A continuación se produce la selección de términos mediante el cálculo de su 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 utilizando normalmente el modelo de espacio vectorial, u otras medidas derivadas como Okapi BM25 (Jones, Walker, & Robertson, 2000). Finalmente se construye el índice invertido que contiene los términos, pesos y documentos relacionados, el cual puede ser consultado por las aplicaciones de búsqueda de información para seleccionar los documentos más relevantes con respecto a una consulta dada.

Sin embargo, algunos motores de búsqueda utilizan algoritmos personalizados como Google con el PageRank (Page, Brin, Motwani, & Winograd, 1999), el cual le permite tener formas adicionales para jerarquizar y evaluar la relevancia de los documentos Web teniendo en cuenta las referencias que le hacen. Otros motores utilizan términos META, los cuales son agregados en los documentos HTML por parte de sus creadores, dando algún tipo de aprovechamiento de la anotación semántica.

Por otro lado, en los últimos años han surgido una serie de motores de búsqueda denominados buscadores semánticos, los cuales realizan el rastreo de la información teniendo en cuenta el significado del grupo de palabras de la consulta y su relación con los documentos de una forma inteligente (Javier, 2011).

Un índice es una lista de palabras o indicadores que permite la ubicación de información, ya sea en un libro o publicación, con la finalidad de disminuir el tiempo de búsqueda de información. Un índice invertido es un archivo o estructura de datos que contiene una lista de palabras (vocabulario) que aparecen en todos los textos de la base de datos documental en orden lexicográfico, cada elemento del vocabulario tiene asociados los documentos en dónde aparecen y opcionalmente las posiciones que ocupan en cada documento (lista de ocurrencias). La principal diferencia con los índices normales es que las palabras apuntan hacia todos los documentos que las contienen y no sólo su posición en un documento particular, lo que permite disminuir el tiempo de búsqueda de información (Benavides, 2011).

La creación de un índice invertido presenta varios retos, directamente relacionados con la granularidad (Yates & Neto, 1999) del índice; específicamente debe tener en cuenta lo siguiente:

- **Disminuir al máximo el tiempo de su construcción, actualización y búsqueda de información.**
- **Capacidad de almacenar tanto el vocabulario, como las listas de ocurrencias, estas últimas pueden consumir gran cantidad de espacio.**
- **Tolerancia a los fallos.**

1.3. Trabajos relacionados

A continuación se describen los trabajos previos más representativos con respecto a la forma de crear un índice semántico:

- **La investigación realizada en *Semantic Indexing For A Complete Subject Discipline* (Chung *et al.*, 1999) 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, emplearon 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 y usado para generar e integrar múltiples índices semánticos. Se realizan etapas intermedias en el proceso, como extracción de frases nominales y análisis de coocurrencia.**
- **En el proyecto *Performance And Implications Of Semantic Indexing In A Distributed Environment* (Chang & Schatz, 1999), 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: En la fase 1 se realiza el pre-procesamiento necesario de los documentos y en la fase 2 Se distribuyen las tareas de indexación a diferentes máquinas en el entorno y se utiliza una función de similitud para la asociación de conceptos.**
- **El proyecto *Semantic Indexing Using Wordnet Senses* (Rada & Dan, 2000) se basa en la implementación de un prototipo que combina la indexación basada en palabras y basada en sentidos o conceptos, utilizando WordNet. El proyecto se realiza en tres etapas que comprenden: 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 módulo 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). El módulo de indexación, indexa los documentos que 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. Finalmente el módulo de recuperación rescata documentos basados en una consulta de entrada.**
- **En la investigación titulada *Indexing a Web Site with a Terminology Oriented Ontology* (Desmontils & Jacquin, 2002), se realiza un proceso semiautomático, que ofrece un índice basado en el contenido de un sitio Web, donde utilizan las técnicas del lenguaje natural. En primer lugar se hacen una eliminación de los marcadores de HTML de las páginas Web, se divide el texto en frases independientes y se realiza una lematización de las palabras incluidas en las páginas. A continuación, las páginas Web se anotan con parte de etiquetas de voz. Luego, se lleva a cabo un proceso de generación de conceptos candidatos con WordNet (University, 2009). Se calcula la representatividad de acuerdo con la frecuencia ponderada y su similitud acumulada del concepto. Posteriormente, se asocian los conceptos de la ontología y conjunto de sinónimos (synsets). 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. Este proceso está siendo incorporado en el sistema Bomon Multiagente (S Cazalens, 2000), para buscar información relevante en Internet.**
- **El proyecto *A Novel Approach to Semantic Indexing Based on Concept* (Kang, 2003) 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, ponderación de cadenas y nombres, reponderación del término basada en el concepto y extracción del índice del término semántico.**
- **En el proyecto *Ontology enrichment and indexing process* (Desmontils *et al.*, 2003), se construye 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 proponen cuatro pasos generales: Primero se construye un índice plano de los términos y se pondera la frecuencia, luego se generan los conceptos candidatos mediante el tesauro WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1993), posteriormente se determina la representatividad de los conceptos en el contenido de la página Web, luego se utiliza una ontología para seleccionar los conceptos con una alta representatividad en los contenidos de la página. Finalmente realizan el**

enriquecimiento de la ontología mediante WordNet como diccionario de sinónimos.

- En *Towards Building Semantic Rich Model for Web Documents Using Domain Ontology* (Noah *et al.*, 2004), 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). El proceso llevado a cabo es el siguiente: Para el análisis de documentos: primero se toman los documentos HTML para eliminar las etiquetas. Luego se eliminan todas las palabras vacías y los conceptos seleccionados se derivan a su raíz. Posteriormente, se dividen los documentos en párrafos y luego en frases, las cuales se almacenarán en un repositorio. Los conceptos con alta frecuencia previamente obtenidos, 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)[2]. Este proceso da como resultado una lista de los posibles conceptos candidatos, y la lista de las frases en donde los conceptos fueron encontrados. Para el análisis del lenguaje natural se definen las etapas: Morfología y proceso de acceso de análisis semántico, Análisis semántico, Modelo global de la semántica del documento.
- El proyecto *The effect of Semantic Index in Information Retrieval development* (Nguyen & Phan, 2008), 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). Presenta un primer módulo que hace indexación de documentos a través de la herramienta Lucene, luego se puede enriquecer la información por parte de los usuarios. Crea finalmente herramientas de navegación entre los conceptos del índice y utilizades de acceso a los índices semánticos.
- La investigación realizada en *Semantic Indexing of Technical Documentation* (Samaneh Chagheri, 2009), se basa en una extensión del modelo vectorial y propone un modelo de indexación semántica que explota las estructuras lógicas y el contenido semántico de los documentos. Utilizan la extensibilidad del lenguaje XML que permite representar simultáneamente el contenido y la estructura lógica de los documentos. Primero identifican los términos candidatos por medio de un tesauro y luego obtienen el lema de las palabras mediante herramientas de procesamiento de lenguaje natural. Posteriormente, proyectan los documentos en WordNet para extraer los términos que coinciden y calculan sus pesos de acuerdo con la frecuencia en los elementos lógicos en los que aparecen (títulos, secciones, etc.). A continuación, obtienen un conjunto de conceptos mediante la ontología, por cada término candidato asignan una puntuación para cada concepto (se escoge el concepto con la más alta puntuación). Finalmente utilizan una función de similitud que mide el camino más corto entre conceptos, así, el índice del documento es un conjunto de pesos por cada concepto encontrado.

En la Tabla 1 se resumen y comparan cada uno de los elementos utilizados por las investigaciones relacionadas a la presente investigación.

Aunque se intentó colocar los elementos en el orden en que los proyectos los utilizan, los encabezados de la Tabla 1 no deben entenderse como pasos consecutivos que definen cada uno de los proyectos, ya que algunos de ellos no siguen éstos en el orden presentado.

La Tabla 1 permitió definir los pasos que son relevantes a la hora de crear los índices semánticos, además de particularidades que dependen del objetivo de construcción del índice semántico.

El análisis de cada proyecto permitió además tener en cuenta las herramientas y recursos más utilizados para la creación de los índices semánticos, presentando elementos necesarios para realizar la propuesta de procedimiento general para crear índices semánticos, la cual se presenta en el siguiente apartado.

Para la elaboración de índices semánticos debe existir una herramienta (tesauros y/o ontologías) de clasificación y cálculo de representatividad de los términos, además de la posibilidad de utilizarlas para hallar nuevos conceptos o conceptos relacionados a la temática de la que trata el documento.

De estos proyectos se puede concluir que las ontologías de dominio juegan un papel importante en las tareas de clasificación y organización de documentos, no sólo sirven para extraer conceptos importantes, sino también para construir el contenido semántico de los documentos Web.

PROYECTOS	ELEMENTOS UTILIZADOS PARA CREAR EL INDICE SEMANTICO (PROCEDIMIENTO)											
	Extracción términos	Extraer cadenas léxicas	Anotador (Part-Of-Speech)	Cálculo de frecuencia por términos	Extracción de conceptos			Análisis de conceptos		Cálculo de frecuencia por conceptos		Asociación conceptos y Synsets
	análisis léxico, Tokenizar, lematizar,	Frases nominales	Análisis léxico y contextual	Ponderación de frecuencia de términos	WordNet	Ontología	Otro	Espacio Conceptual	Relaciones entre Conceptos	Representatividad	Reponderación	Asociación de conceptos
(Chung et al., 1999)	X	X	X	X			X	X	X	X		
(Chang & Schatz, 1999)		X					X	X	X	X		
(Rada & Dan, 2000)	X	X	X		X				X	X		X
(Desmontils & Jacquin, 2002)	X	X	X	X.		X			X	X		X
(Kang, 2003)					X			X	X	X	X	
(Desmontils et al., 2003)	X.	X	X	X	X	X			X	X		X
(Noah et al., 2004)	X	X	X	X		X	X					
(Nguyen & Phan, 2008)	X.						X					
(Samaneh Chagheri, 2009)	X		X	X	X	X			X	X		

Tabla 1. Comparación de elementos para la creación de índices semánticos **Fuente:** presentación propia de los autores

2. Procedimiento para la Creación de Índices Semánticos

La indexación semántica va más 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) (Yu, L., Cuadrado, & Coburn, 2003) y usar los patrones de búsqueda para una mejor clasificación de los datos y precisión en la recuperación de información.

Según el enfoque dado por Suarez B. Marco (2009), un índice semántico se caracteriza por:

- **Ser multidimensional:** **Un concepto se puede modelar como un conjunto de propiedades relacionadas, que a su vez son otros conceptos, estas propiedades se pueden entender como diferentes dimensiones con los que se puede valorar la semántica de los conceptos encontrados en los documentos. Los elementos de indexación son valores de atributos que pueden estar basados en complejas descripciones de objetos relacionados.**

- **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.

- **Información disponible:** Dado que el índice es en realidad un conjunto de descripciones parciales de los objetos indexados, mucha información se puede extraer directamente del índice, sin tener acceso a los documentos para procesamientos posteriores.

Para la indexación semántica, se reutilizan varios tipos de procesamientos mencionados en la indexación clásica, ellos se usan en una fase inicial denominada: "Preprocesamiento de la Base Documental".

La siguiente fase denominada: "Extracción de conceptos", busca que los términos encontrados sirvan para llegar a los conceptos que se tratan en los documentos, estos conceptos se obtienen de estructuras de almacenamiento de conocimiento, como los tesauros o las Ontologías.

Posteriormente una fase denominada "Estudio de representatividad" mediante la aplicación de técnicas estadísticas o algebraicas se ponderan los conceptos con respecto a los documentos.

Finalmente, en una fase llamada "Construcción del índice semántico", se procede a crear la estructura de datos, que normalmente corresponde a un árbol invertido, pero la idea es utilizar estructuras computacionales que permitan búsquedas rápidas, así mismo la fase final establece una evaluación al índice creado, con esta información se puede retroalimentar nuevamente el proceso para afinar y mejorar el índice construido (Ver Figura 2).

A continuación se muestra el procedimiento descrito mediante un diagrama de actividades, el cual permite entender mejor el flujo de decisiones que se deben tomar en la construcción de un índice semántico. Debido a su tamaño se ha dividido en cuatro fases. Los pasos se representan en gris claro / línea delgada y los productos obtenidos en gris oscuro/línea gruesa, las líneas discontinuas son productos o pasos opcionales.

En la Figura 3 se muestra el pre-procesamiento que se le hace a la base de datos documental. Las salidas corresponden a: un índice plano previo o una base

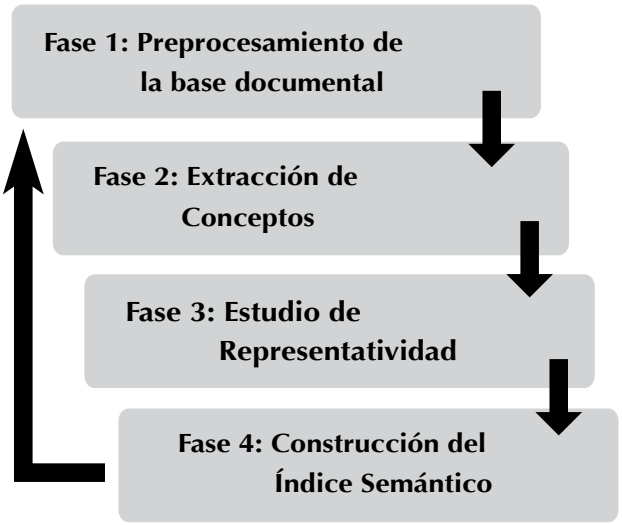


Figura 2: Fases propuestas para la indexación semántica
Fuente: Presentación propia de los autores

documental preprocesada. Para ambos casos el procesamiento es el mismo: primero, al conjunto de documentos iniciales se les ha convertido en texto plano, segundo un análisis léxico en el cual se identifican los tipos de palabras encontradas, posteriormente se eliminan las palabras vacíasy finalmente se llevan a los lexemas las palabras que concuerdan con un mismo origen gramatical, reduciendo así el conjunto de términos a tener en cuenta.

En el caso del índice plano, corresponde a una estructura de datos que será consultada de manera eficiente, para el cual dado un término se pueden obtener los documentos relacionados.

En el caso de la base documental preprocesada, se tiene un conjunto de documentos con términos relevantes en su contenido.

Como se puede apreciar en la Figura 4, la fase 2 inicia con la selección de los términos de los documentos preprocesados o el índice creado. Lo primero que se hace es definir si se obtendrán frases nominales, el hecho de crear las frases nominales da la posibilidad de utilizar los algoritmos de espacios conceptuales y la aplicación de tesauros y ontologías, para ampliar o reducir los conceptos y solucionar problemas de homonimia, sinonimia y polisemia, entre otros. La salida corresponde a un conjunto de vectores con conceptos relacionados a los documentos.

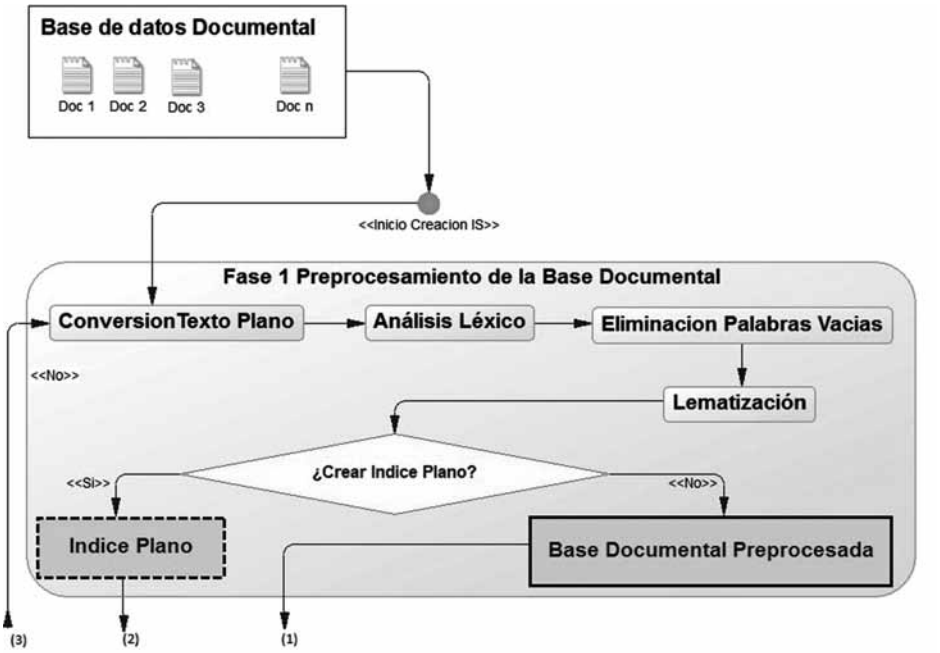


Figura 3. Diagrama de actividades: Procedimiento de creación de índices semánticos, Fase 1 **Fuente**: Presentación propia de los autores

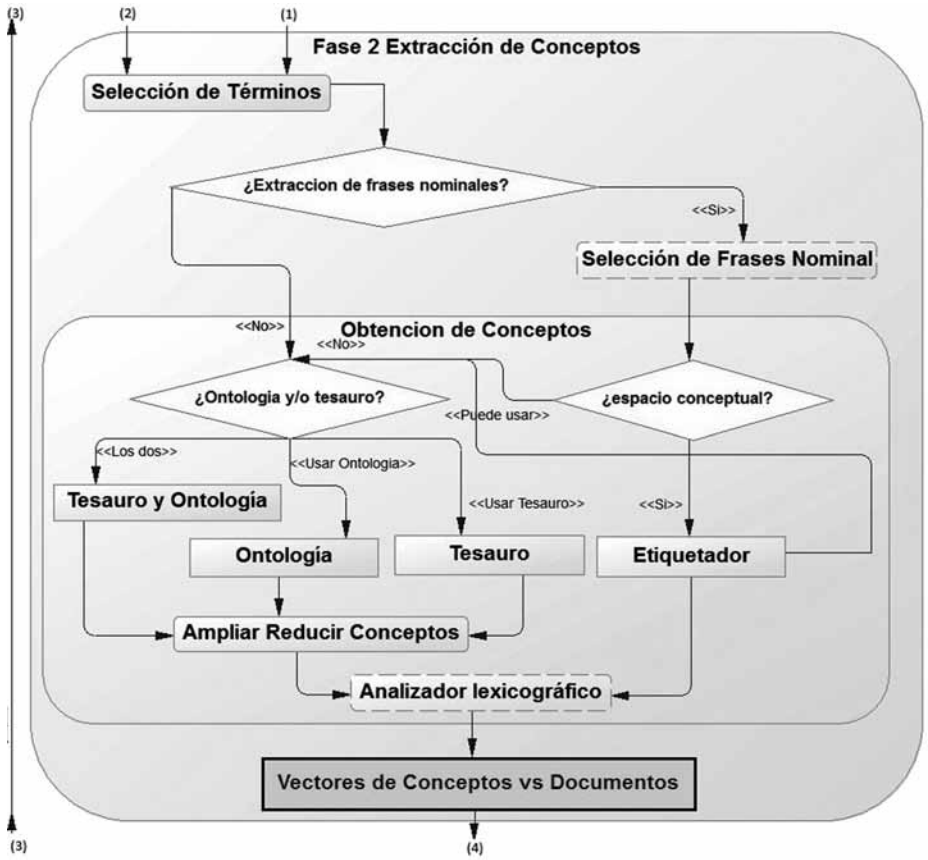


Figura 4. Diagrama de actividades: Procedimiento de creación de índices semánticos, Fase 2 **Fuente:** Presentación propia de los autores

En la Figura 5 se presenta la siguiente fase del proceso, en esta se toman cada uno de los conceptos y se verifica su coocurrencia en cada uno de los documentos, obteniendo unos pesos que dependerán de cada estudio, con el fin de hacer el análisis de relevancia correspondiente con una función de similitud semántica apropiada. Así se obtendrá un espacio conceptual de documentos vs. conceptos debidamente ponderaros y organizados por relevancia.

Finalmente, en la Figura 6 se muestra cada una de las decisiones importantes al momento de construir y evaluar el buen funcionamiento del índice semántico. Hay que destacar que en la última decisión, después de evaluar el índice, si los resultados no son satisfactorios se debe volver a revisar todo el proceso desde la primera fase, con el fin de encontrar puntos de sintonización que permitan optimizar el índice hasta obtener un funcionamiento adecuado lo que depende de su ámbito de implementación.

En la Tabla 2 se resume cada una de las Fases y los pasos relacionados, detallando con mayor precisión lo que se hace en cada uno (Ve páginas 268-271).

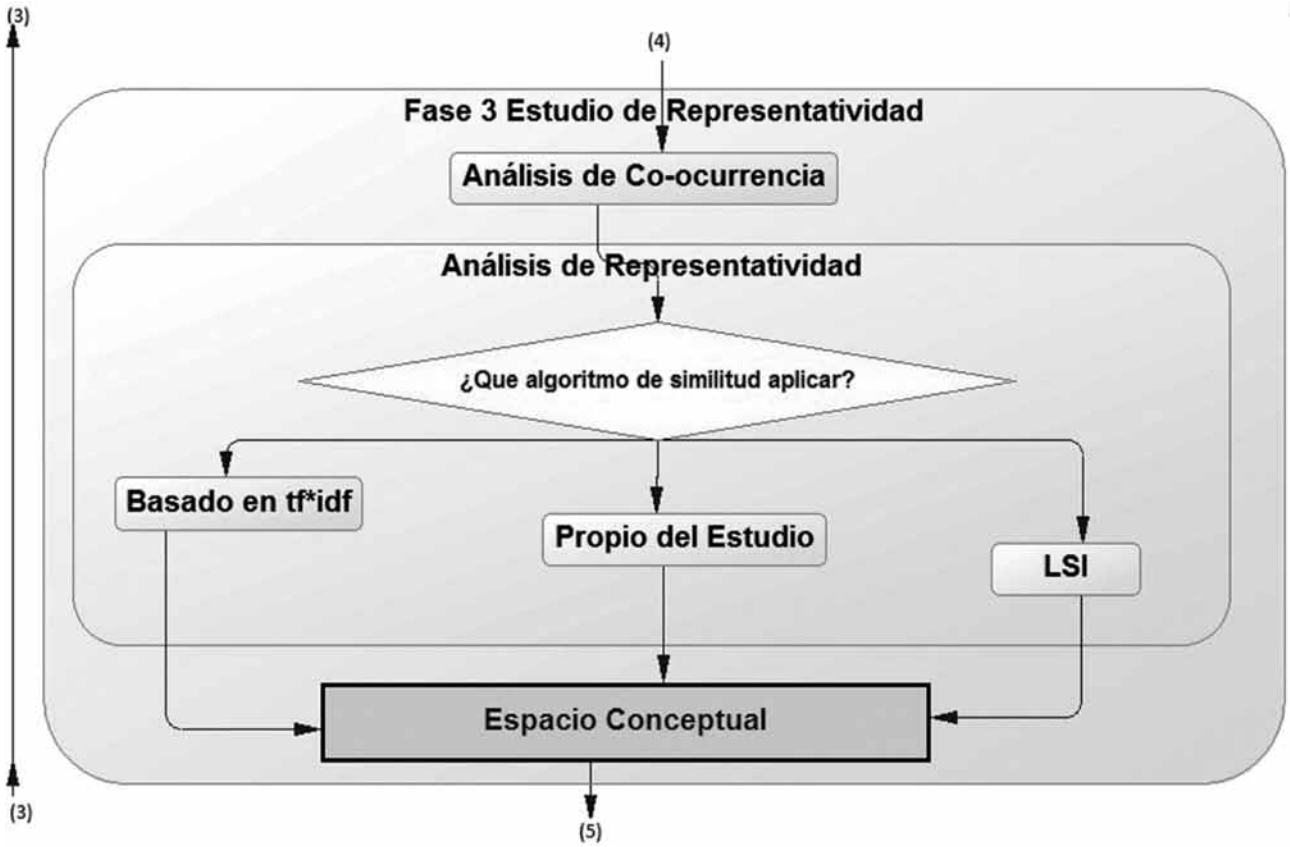
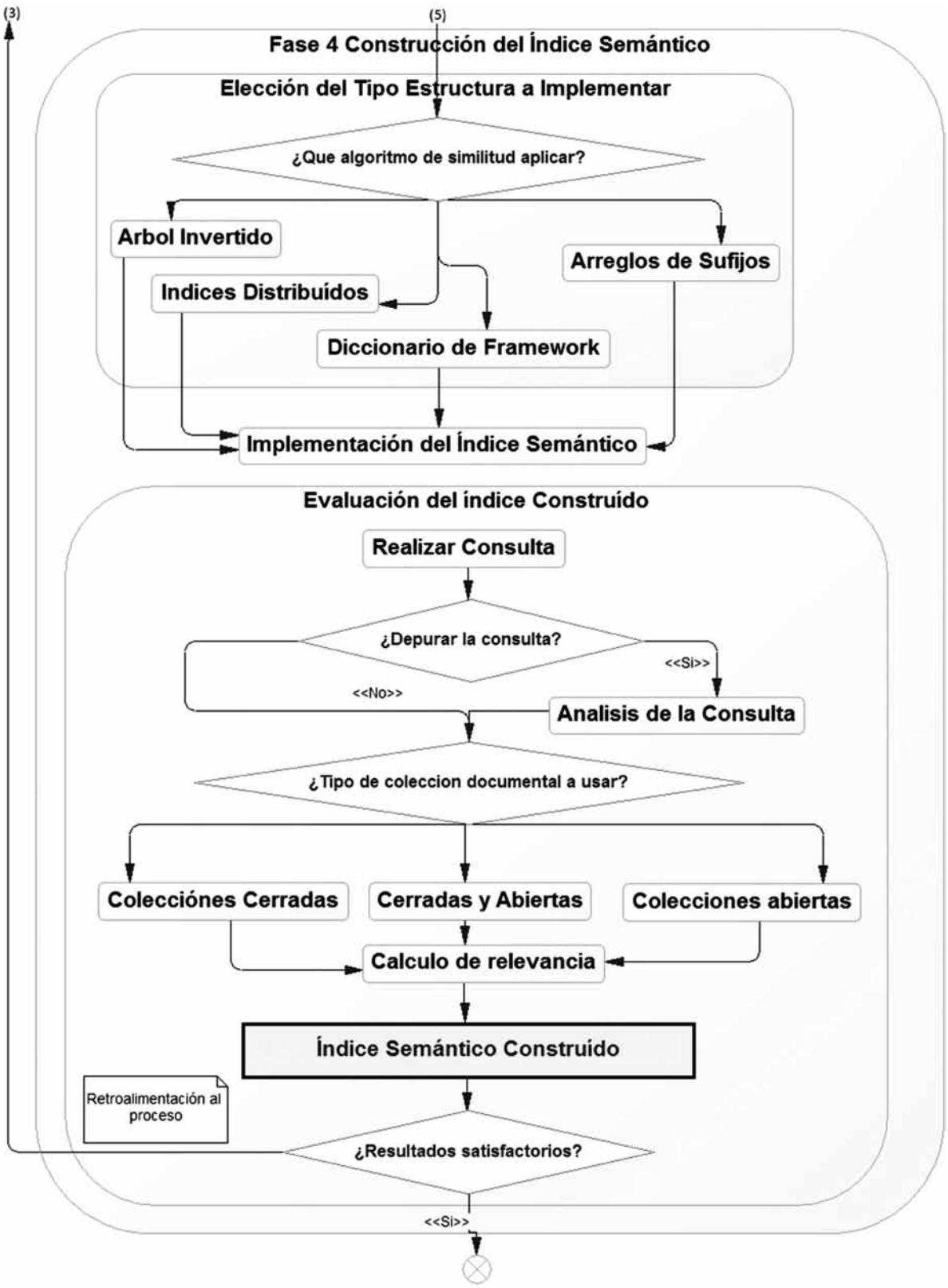


Figura 5: Diagrama de actividades: Procedimiento de creación de índices semánticos, Fase 3. **Fuente:** Presentación propia de los autores

Ver en la siguiente página:
Figura 6. Diagrama de actividades: Procedimiento de creación de índices semánticos, Fase 4. **Fuente.** Presentación propia de los autores


(3)
(5)
Fase 4 Construcción del Índice Semántico
Elección del Tipo Estructura a Implementar
¿Que algoritmo de similitud aplicar?
Arbol Invertido
Indices Distribuídos
Arreglos de Sufijos
Diccionario de Framework
Implementación del Índice Semántico
Evaluación del índice Construído
Realizar Consulta
¿Depurar la consulta?
<<Si>>
<<No>>
Analisis de la Consulta
¿Tipo de coleccion documental a usar?
Colecciónes Cerradas
Cerradas y Abiertas
Colecciones abiertas
Calculo de relevancia
Índice Semántico Construído
Retroalimentación al proceso
¿Resultados satisfactorios?
<<Si>>

Fase	Descripción
Fase 1: Preprocesamiento de la base documental	Su objetivo es convertir la base documental de entrada en un conjunto de palabras (tokens) adecuadas para crear el índice, reduciendo así el número de datos a analizar. Para esta fase se puede apoyar en herramientas software que contienen librerías para este procesamiento como LUCENE (Hernández & Hernández, 2008). Los pasos son: ***Paso 1. Conversión de los documentos a texto plano:*** Conversión de los textos a un formato plano (.txt) de tipo ASCII. ***Paso 2. Análisis Léxico del Texto:*** Basado en una herramienta de procesamiento 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. ***Paso 3. Eliminación de palabras Vacías (stopword removal):*** Busca eliminar los términos con poco valor en la recuperación de información como: pronombres, partículas interrogativas y ciertas preposiciones. Entre los artículos están por ejemplo: "un, la, los, el, ellos", las partículas interrogativas son: "que, cuál, quién, cómo, dónde", entre otros. Algunas preposiciones son: "con, desde, entre, hasta, por, según", etc. Dependen los términos del idioma que se escoja. ***Paso 4. Aplicación de Lematización (stemming):*** 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. A través de Lucene y otros proyectos se pueden usar diversos algoritmos de stemming, como por ejemplo el algoritmo de Porter para inglés. ***Paso 5.*** (Opcional)Creación de índice plano (índice de términos): Opcionalmente en esta fase se puede generar un índice plano, con base en uno de los modelos clásicos de indexación. Algunos proyectos lo hacen así con el fin de calcular una ponderación inicial de los términos y posteriormente en la fase de análisis conceptual realizar ajustes o reponderaciones. **Producto:** El producto de esta fase es una base documental pre-procesada, la cual será tomada para realizar los análisis semánticos correspondientes. Opcionalmente, se puede obtener también un índice plano clásico, el cual puede ser utilizado para re-ponderar sus pesos con los análisis posteriores.

Fase 2: Extracción de conceptos	Como en la indexación clásica, se debe aplicar una selección de términos candidatos, es decir, los términos relevantes en los documentos opcionados para extraer sus conceptos. Estos conceptos aparecerán a partir de un proceso de consulta en estructuras de diseñadas para almacenar conocimiento, como los tesauros y las ontologías u otras aproximaciones de la teoría cognitiva. En esta fase se identifican dos pasos: ***Paso 1. Selección de términos:*** Si en la fase anterior se creó un índice plano, se consultan los términos directamente del índice y se continúa con el siguiente paso, de lo contrario se toma cada documento pre-procesado en la fase 1 y se extraen sus términos en una estructura de datos que permita representar cada documento como un conjunto de términos, entre ellos las frases nominales. La selección de frases nominales es muy utilizada en los proyectos de indexación semántica, sin embargo es opcional y dependerá más de los objetivos del proyecto. Es muy útil en la extracción de conceptos y relaciones entre los términos consecutivos. Una frase nominal es un conjunto de términos y constituye en sí misma un nuevo término (concepto) de indexación. Este paso permite no solo extraer los conceptos de los términos aislados, sino también diferencias en sus relaciones semánticas entre sí. Para obtener la frases nominales generalmente se utiliza el algoritmo propuesto por Baziz (2005), relacionándolos con Tesauros y Ontologías para comparar las relaciones jerárquicas de estas con las frases extraídas. ***Paso 2. Obtención de los conceptos:*** En este punto existen dos variantes que los proyectos de indexación semántica han explorado: En la primera (la más usada), es la obtención de conceptos basado en ontologías y/o tesauros, los términos obtenidos (paso 1) son consultados con respecto a una estructura de representación de conocimiento como los tesauros o las ontologías, para obtener la semántica[3] de los mismos, la cual se expresa a su vez como otro conjunto de términos (conceptos) relacionados en la jerarquía del tesauro o la ontología. El proceso de obtener los conceptos tiene dos objetivos principales: a. El primer objetivo consiste en ampliar-reducir los conceptos encontrados en las estructuras de conocimiento consultadas, principalmente abstrayendo o especializando los conceptos. En este punto, cada proyecto de indexación semántica define el conjunto de conceptos relacionados de acuerdo con sus necesidades. Una de las reglas más fáciles para ampliar la representación es obteniendo de las estructuras de conocimiento con respecto a cada concepto el conjunto de sinónimos, hiperónimos, hipónimos y merónimos, entre otras posibles relaciones de conceptos que existan en las estructuras seleccionadas. b. El segundo objetivo consiste en reducir los problemas de homonimia, sinonimia y polisemia: Para este caso se pueden aplicar algoritmos que permitan hacer desambiguación de las palabras. Uno de los más utilizados es el Word Sense Disambiguation – WSD (Rada & Moldovan, 2000; Sánchez & Moreno, 2008; Sanderson, 1996) y sus variaciones. La segunda variante es la utilización de un espacio conceptual (Chung *et al.*, 1999; Schatz, 1997), el cual parte fundamentalmente de la premisa de que los conceptos presentes en un documento tienen interrelaciones que se pueden obtener a partir de técnicas de correlación estadística, adicionalmente se pueden obtener diferentes índices de dominios de conocimiento de la base documental de manera genérica, sin variaciones especiales para un dominio en particular. En la construcción de espacios conceptuales es necesario extraer frases nominales y normalmente los proyectos agregan anotaciones sobre las mismas en los documentos preprocesados identificando sus características para el siguiente análisis, para ello utilizan etiquetadores semánticos, los cuales son módulos de programa que se encargan de asignar etiquetas a cada concepto. Finalmente, cualquiera que haya sido la opción de obtención de los conceptos se lleva a un analizador lexicográfico con el fin de hacer el proceso de desambiguación descrito en el literal b. **Producto:** El producto final de ésta fase es un conjunto de conceptos que están relacionados a cada documento de la base de datos documental. Se tienen vectores de conceptos relacionados a los documentos, los conceptos pueden estar presentes en el documento mismo o no, de acuerdo a la expansión de conceptos realizada.

Fase 3: Estudio de Representatividad	Esta fase toma los vectores de conceptos y documentos obtenidos en la fase anterior con el fin de establecer pesos y/o prioridades de los conceptos con respecto a cada uno de los documentos a los cueles está relacionado. Esto se logra de dos formas: Aumentando la importancia a conceptos muy relacionados y/o eliminando conceptos que alejan el documento, eso depende de lo que se busca. Esto último se logra realizando un estudio o evaluación de representatividad de los conceptos y sus relaciones con los documentos y sus relaciones entre ellos mismos, para esto se utilizan los algoritmos de similitud semántica, los cuales pretenden encontrar, clasificar y ordenar por mayor cercanía o similitud los conceptos y por ende los documentos a los cuales están relacionados. Es aquí dónde se encuentra la principal diferencia con los indizadores clásicos. Los pasos de esta fase se resumen a continuación: ***Paso 1.Análisis de coocurrencia:*** Cualquiera que sea el modelo a seguir normalmente se inicia con un análisis de coocurrencia, con el cual se obtienen las frecuencias de aparición de los conceptos y/o frases nominales en los documentos. Si se realizó una ponderación de frecuencia previa, este estudio puede basarse en ella para re-ponderar los resultados, sin embargo a veces es mejor no ponderar los términos hasta que no se llegue hasta este paso, con el fin de no realizar reprocesos innecesarios. Dependiendo del algoritmo o estrategia de obtención de conceptos de la fase anterior, éste análisis puede ser un poco más complejo que simplemente hacer un conteo de las apariciones, por ejemplo si se decidió hacer un espacio conceptual, el algoritmo de coocurrencia debe tener encuentra las anotaciones hechas a las frases nominales encontradas y usar estructuras de datos más complejas como las matrices de coocurrencia (Chang & Schatz, 1999). ***Paso 2.Análisis de representatividad***: Para el análisis de representatividad, se puede usar cualquier algoritmo de similitud semántica existente (Cord, Lombardi, Martelli, & Mascardi, 2005; Chung *et al.*, 1999; Desmontils & Jacquin, 2002; Lin, 1998; Lv *et al.*, 2009; Mazuel & Sabouret, 2008; Resnik, 1999; Samaneh Chagheri, 2009; Song *et al.*, 2009), sin embargo, la medida de similitud y ponderación de frecuencia con TF-IDF (Salton & McGill, 1986) y las variaciones personalizadas por cada investigación (Avello, 2005; Herrera, 2006; Samaneh Chagheri, 2009), sigue siendo uno de los cálculos más simples y usados para representar los conceptos más relevantes en los textos. Por otra parte, uno de los algoritmos que ha tomado mayor fuerza es la Indexación Semántica Latente (Novoa & Ballen, 2007) o LSI (Latent Semantic Indexing), es una teoría matemática usada como modelo de RI, que permite determinar el uso y las relaciones de un término con su contexto, esto lo logra con la creación de vectores multidimensionales, los cuales permiten encontrar las relaciones existentes entre palabras, palabras y párrafos y entre párrafos, obteniendo lo que se conoce como un espacio conceptual. **Producto:** El producto final de ésta fase es un conjunto de vectores con conceptos más relevantes que representan a cada documento de la base de datos documental. A este conjunto de vectores algunos autores lo llaman Espacio Conceptual, independientemente de que se hayan usado tesauros y ontologías o por el uso de espacios conceptuales, ya que está basado en los conceptos y no en los términos. Por tal razón en este trabajo se define el producto de esta fase como el espacio conceptual.

Fase 4: Construcción del Índice Semántico	Finalmente, en esta fase se procede a construir el índice semántico, el cual se puede entender como una estructura de datos sobre el texto de los documentos, para acelerar la búsqueda (Yates & Neto, 1999). Existen varios tipos de estructuras como los arboles invertidos (Zobel, 2006), arreglos de sufijos, archivos firmados y los índices distribuidos (Benavides, 2011). Los pasos son los siguientes: ***Paso 1. Elección del Tipo de Estructura a Implementar:*** Para la elección de la estructura en la cual almacenar el índice es preciso evaluar en general los siguientes aspectos: Cantidad de información a indexar, la frecuencia de actualización y consulta del índice. Normalmente cuando los valores de las variables anteriores son bajos se usan los árboles invertidos, de lo contrario se debe pensar en índices distribuidos. Por otro lado los arreglos de sufijos son eficientes cuando las búsquedas se basan en patrones. Sin embargo, cuando el índice no tiene por objetivo el manejo de gran cantidad de información se puede usar una estructura existente en los framework de programación, cómo los diccionarios. ***Paso 2. Implementación del Índice Semántico:*** La mínima información que se debe almacenar en la estructura definida son los vectores de conceptos vs. el documento que representan. Así, por cada conjunto de conceptos, se debe poder obtener los documentos con los que están relacionados. Sin embargo, los objetivos y las particularidades de cada proyecto pueden definir información adicional, como almacenar posiciones en el texto, relaciones con otros conceptos, entre otras junto a los conceptos y a los documentos representados. Esto hace una buena diferencia con respecto de los índices normales, ya que se podría obtener más información de los documentos buscados, consultando sólo el índice. ***Paso 3. Evaluación del índice construido:*** Este paso se realiza utilizando el índice con experiencias de campo, de tal forma que se pueda definir un conjunto de indicadores, generalmente de relevancia en la información recuperada, los cuales permitan establecer si el funcionamiento y el propósito del índice semántico cumple con las expectativas para las cuales fue construido. En este punto se añaden pasos para la manipulación de la consulta, los cuales están por fuera del presente estudio. Se sugieren las siguientes actividades de evaluación: a. Realización de consultas para verificar la relevancia de resultados obtenidos con el índice semántico. b. 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 de 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 realiza un mapeo semántico con un tesauro, un vocabulario controlado o una base de datos léxica de dominio general. Al realizar esta comparación se extraen los conceptos que se asemejen (sinónimos) a los términos de la consulta, y así se comparan con los documentos relevantes de acuerdo con el índice construido. c. Para observar los resultados que se obtienen con la utilización del índice se realizan evaluaciones exhaustivas, con uno de los siguientes enfoques: Las primeras, orientadas a evaluaciones automatizadas sobre colecciones de información existentes (cerradas) y previamente valoradas (documentos, consultas y documentos relevantes para cada consulta), para establecer la Precisión (Frakes, 1992; Kent, Berry, Luehrs, & Perry, 1955; Salton & McGill, 1986), Recuerdo o Exhaustividad (Kent *et al.*, 1955; Salton & McGill, 1986; Swets, 1969) y el índice MAP, entre otras; las segundas están relacionadas con evaluaciones realizadas por usuarios (colecciones abiertas) donde se pueden evaluar medidas derivadas de la precisión y el recuerdo y medidas como el estadística Kappa (Manning, Raghavan, & Schütze, 2008) para evaluar la confianza de los resultados. ***Paso 1.*** Realimentación del Índice Semántico: Una vez que se tienen los datos de las evaluaciones, si los resultados son satisfactorios, se puede dar por terminada la construcción del índice, de lo contrario, se deben definir un conjunto de hipótesis sobre los cambios que deben sufrir los algoritmos que construyen los índices. Lo anterior implica volver y revisar las fases anteriores haciendo cambios y evaluando repetidamente los resultados hasta cumplir con los objetivos previamente definidos. **Producto:** El producto final de esta fase es una estructura de datos llamada índice semántico, del cual se consultarán los conceptos y sus respectivos documentos relacionados.

Tabla 2. Fases para la construcción de índices semánticos **Fuente**. Presentación propia de los autores

Cuando se construye un índice semántico, una de las preguntas más importantes es establecer si se debe usar tesauros, ontologías o las dos. Aquí se presentan sugerencias para ayudar a tomar dicha decisión:

- **Usar sólo tesauros:** La funcionalidad primordial de un tesauro es agrupar un conjunto de palabras de un idioma particular, las cuales representan un concepto del conocimiento humano, así los tesauros en el contexto de la recuperación de la información tienen dos objetivos fundamentales, los cuales son: Controlar el vocabulario, lo cual significa identificar dentro de un campo semántico todos los conceptos representados por un término (Jiménez, 2004) y los términos relacionados con un concepto determinado.

- **Usar sólo ontologías** (Carrascal, 2004): Las ontologías representan un nivel más alto de concepción y descripción de los vocabularios, presentan un desarrollo semántico más profundo para las relaciones del tipo clase/subclase, y para las relaciones cruzadas (Jiménez, 2004), que los tesauros. Estas se pueden reutilizar y 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 profunda, así como unas relaciones enriquecidas entre clases y conceptos.

- **Usar tesauros y ontologías:** Es importante mencionar que se puede optar por la utilización de las dos herramientas para asegurar un índice semántico mejor adaptado para ciertos dominios, en el orden que se requiera. Sin embargo, la mayoría de investigaciones que utilizan las dos herramientas hacen el mapeo en el siguiente orden: en primer lugar se usa 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 conceptos, luego de realizar el mapeo de términos y conceptos con el tesauro, se procede a efectuar un mapeo de conceptos candidatos con los conceptos de la ontología de dominio. En este paso se estudia la representatividad de los conceptos con el acuerdo al dominio y se hace un cálculo de frecuencia de estos conceptos mediante ponderación de frecuencias (TF-IDF), funciones de similitud acumulativa y/o distancia semántica entre conceptos (Desmontils & Jacquin, 2002). Con estos pasos se obtienen los conceptos más representativos de los documentos y se genera el índice semántico de acuerdo con ello.

3. Plantilla de instanciación del procedimiento propuesto

Con el fin de facilitar el uso del procedimiento propuesto, se creó una plantilla en la cual se consigna paso a paso cada una de las decisiones y pasos que se personalizan dependiendo de las características y objetivos del proyecto que use este procedimiento.

La plantilla propuesta se presenta ya instanciada en el caso de estudio en la Tabla 3 (Ver página siguiente).

Debe tenerse en cuenta que la plantilla presenta unos pasos que son opcionales y otros obligatorios. Sin embargo, se deja a discreción de los diseñadores si definitivamente lo hacen o no, de tal forma que sea más una recomendación por parte del procedimiento. En el campo de observaciones se debe colocar el detalle de lo que se hizo en este paso o actividad.

4. Creación de un índice semántico para el entorno de las plantas

Teniendo el procedimiento definido se realizó un caso de estudio, relacionado con la creación de un índice semántico basado en una ontología de dominio especifico relacionado con las plantas (botánica) y para un entorno particular de la educación básica primaria. A continuación se presenta la plantilla de la instanciación del índice que se construyó teniendo en cuenta las herramientas utilizadas para este proyecto.

Posteriormente, se procedió a la implementación del índice semántico, utilizando el Framework de .NET y creando un servicio Web para el acceso al índice. La arquitectura de la aplicación se presenta en la Figura 7. La arquitectura planteada en la Figura 7, permite ver una clara separación de las capas de interfaz (AppWeb Semantic Search y Metabuscador), lógica del negocio (LuceneManager, HTMLParseManager, ScottWaterManager y OntologyManager) y acceso a los datos (servicios delicious, WordNet, la ontología de dominio plantOntology (Consortium, 2010) y el índice semántico construido).

La lógica del negocio se encuentra encapsulada en un servicio Web wsSemanticSearch (http://www.prometeo.unicauca.edu.co/wsSemanticSearch/wsSemanticSearch.asmx), el cual utiliza API´s de desarrollo (Lucene, Delicious, WordNet) para ejecutar cada una de las fases en la construcción del índice semántico.

El índice semántico se crea en un proceso previo, ya que el análisis que se hace a los documentos, consume un tiempo considerable, se toman los primeros 40 documentos devueltos por delicious (esta cantidad normalmente obtiene los más relevantes y le pone un límite al procesamiento) y se examinan en profundidad obteniendo a su vez los enlaces, un conjunto de subdocumentos, pre-procesándolos e indexándolos de acuerdo al procedimiento desarrollado y tomando como fuente de conocimiento para extraer los conceptos la ontología PlantOntology. En esta capa se genera un archivo plano (.txt) del Índice Semántico obtenido al llevarse a cabo el proceso de indexación, este archivo es cargado a la estructura de datos en memoria cuando se ejecuta la aplicación se búsqueda Web.

Una vez construido el índice, cada vez que se hace una consulta la cual recibe ya sea de la interfaz propia que se desarrolló para las evaluaciones con los usuarios, denominada "Semantic Search" se llama el método SemanticSearch() del servicio Web, enviando la consulta, ésta se procesa con la ontología y WordNet si es necesario, para finalmente buscar sus coincidencias en el índice creado. El servicio web retorna un dataset en xml en el cual se tiene el título, el resumen (snippet) y la url de cada documento relevante encontrado, este dataset lo recibe la interfaz y lo presenta a los usuarios.

En la Figura 8 se muestra la interfaz desarrollada para presentar los resultados de la búsqueda en el índice semántico de plantas.

Fases – Pasos - Actividades		¿Se realiza?		Observaciones y/o Comentarios
Fase 1	Pasos - Actividades	Si	No	
Preprocesamiento de la Base Documental	Paso 1: Conversión de documentos a texto plano (obligatorio)	X		El repositorio de documentos es abierto, documentos Web. Sin embargo se decidió utilizar una herramienta de marcado colaborativo de la Web 2.0 llamada "Delicious" (http://www.delicious.com). Esta herramienta ofrece un API de desarrollo que permite extraer todos los documentos que la comunidad ha etiquetado. Para la conversión en texto plano se usó la clase HTMLParseManager.
	Paso 2: Análisis léxico del texto (obligatorio)	X		Usa la librería Lunece
	Paso 3: Eliminación de Stopwords (obligatorio)	X		Usa la librería Lunece
	Paso 4: Lematización (obligatorio)	X		Usa la librería Lunece
	Paso 5: Creación de un Índice Plano (opcional)	X		Usa la librería Lunece

Fase 2					
Extracción de conceptos	Paso 1: Selección de términos (obligatorio)	Actividad 1: Extraer los términos simples de cada documento (obligatorio).	X		Los términos obtenidos de cada documento se almacenaron en una matriz de términos por documentos, ya que se decidió construir un índice plano inicial.
		Actividad 2: Extraer Frases Nominales – FN (opcional)		X	No se obtuvieron FN, ya que la mayoría de los términos relacionados a las plantas son simples.
		Actividad 3: Construir espacio conceptual etiquetando las FN (opcional)		X	Se decide implementar métodos personalizados con la Ontología seleccionada y no los métodos estadísticos aplicados a los espacios conceptuales.
	Paso 2: Obtención de Conceptos	Actividad 1: Elegir entre usar tesauros, ontologías o los dos. (obligatorio)	X		La ontología escogida es PlantOntology http://www.plantontology.org/. (Consortium, 2010): Describe un vocabulario controlado con las estructuras de las plantas así como su crecimiento y etapas de desarrollo, proporcionando un marco de trabajo semántico de las consultas a través de especies significativas. Es importante mencionar que la ontología es manejada en lenguaje OWL y se manejo con la API de Protege 3.1.1.
		Actividad 2: Ampliar – Reducir los conceptos (obligatorio)	X		Después de haber seleccionado la ontología, se utilizan los diferentes métodos para hacer el tratamiento respectivo de la misma, para tal caso se utilizó la librería de ScottWaterManager la cual permitió obtener las súper clases, subclases y los respectivos axiomas de la ontología.
		Actividad 3: Análisis léxico para desambiguar las palabras. (opcional)		X	No se aplicó ningún algoritmo de desambiguación de palabras, con el fin de hacer el algoritmo del índice semántico caso de estudio lo más simple y básico posible.
Fase 3					
Estudio de Representatividad	Paso 1: Análisis de coocurrencia		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 el framework de .NET, 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 2: Análisis de Representatividad	Actividad 1: Elegir el algoritmo de similitud semántica a aplicar. (obligatorio)	X		Para nuestro proyecto se utilizó el cálculo de TF – IDF (frecuencia de términos – frecuencia inversa del documento), donde se mide 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 a los padres. Realizando este proceso se consiguió mejorar los resultados obtenidos por el índice semántico creado.
		Actividad 2: Crear el espacio conceptual con el algoritmo seleccionado (obligatorio)			Se construye la matriz de ponderación de conceptos. Esta estructura almacena el índice semántico como producto de la indexación.

Fase 4					
Construcción del Índice Semántico	Paso 1: Elección del tipo de estructura a implementar (obligatorio). Analizar tamaño de la base documental, frecuencia de actualización y consulta.		X		Teniendo en cuenta que el índice trabajará sólo con los documentos provistos por delicious, el tamaño es relativamente pequeño. El servicio devuelve 95.654 (consulta 12/05/2011) enlaces por la consulta de la palabra "plant" el índice toma sólo los primeros 40 enlaces, reduciendo los mismos cuando se consulta por partes de la planta que se encuentran en la ontología seleccionada. Por otro lado no es necesario actualizar constantemente el índice ya que en esta experiencia se añaden nuevos documentos acerca del tema con baja frecuencia. Finalmente, se espera que el número de usuarios que consultan sea bajo, ya que se pretende crear un meta buscador para el uso de estudiantes y profesores cuando estén impartiendo este tema en ciencias básicas en educación básica primaria de ciertas regiones de Colombia. Por lo anterior se escogió una estructura de Diccionario provista por el framework de .NET.
	Paso 2: Implementación del índice semántico (obligatorio)		X		Se almacena el índice en un archivo en disco el cual es subido a la memoria en el diccionario, la primera vez que se ejecuta la aplicación que lo usa.
	Paso 2: Evaluación del índice construido (obligatorio)	Actividad 1: Depurar la consulta (0pcional)	X		Para probar la aplicación se realizaron pruebas con dos colegios de la ciudad (como se especifican más adelante). Adicionalmente, para este caso se realizó 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 se estaba utilizando, se procedía a compararlo con un tesauro de domino general. Para este caso se tomo el tesauro WordNet.
		Actividad 2: Elegir el tipo de colección documental a usar. ¿Cerrada o abierta? (obligatorio)	X		Se escogió hacer las pruebas en los dos tipos de colecciones, para ello se creó una colección cerrada. Para el caso se obtuvo los archivos el repositorio de marcado social llamado Delicious (http://www.delicious.com) y se evaluaron como se detalla en la siguiente actividad.
		Actividad 3: De acuerdo a la decisión anterior evaluar la relevancia con los indicadores adecuados. (obligatorio)	X		Aunque la Base Documental por naturaleza es abierta, se decidió obtener manualmente un conjunto de documentos vs. ciertas consultas y evaluarlas, con el fin de poder calcular las curvas de indicadores de precisión, precisión – recuerdo y el índice MAP e indicadores asociados a los usuarios: estadísticas KAPPA.
	Paso 3: Realimentación del índice semántico (obligatorio)	Actividad 1: Evaluar si los resultados son satisfactorios. (obligatorio)	X		Los resultados de la primera iteración no fueron tan relevantes, por lo cual se tuvo que hacer una segunda revisión de lo que se implementó.
		Actividad 2: Realimentar el índice construido volviendo a los pasos anteriores. (obligatorio)	X		Al observar que los resultados no fueron tan satisfactorios. Se revisó de nuevo el índice creado, encontrando en la Fase 2, paso 2, que se le había dado altos pesos a los conceptos hiperónimos, lo cual hacía que los mismos trajeran documentos de dominios más amplios, la solución fue reorganizar los pesos, obteniendo con esto una gran mejoría en los resultados que se presentaban a los usuarios.

Tabla 3. Plantilla del IS en las plantas **Fuente**: Presentación propia de los autores

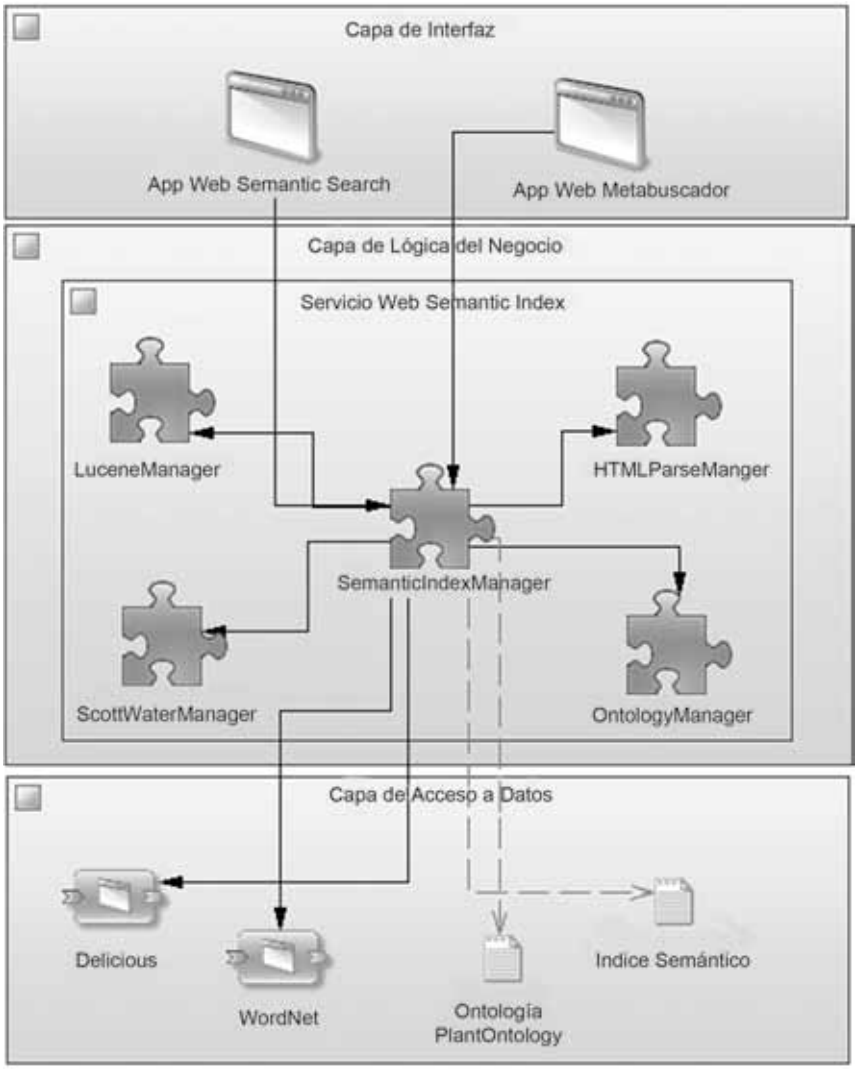


Figura 7. Arquitectura del Servicio Web, Índice Semántico en Plantas **Fuente:** Presentación propia de los autores

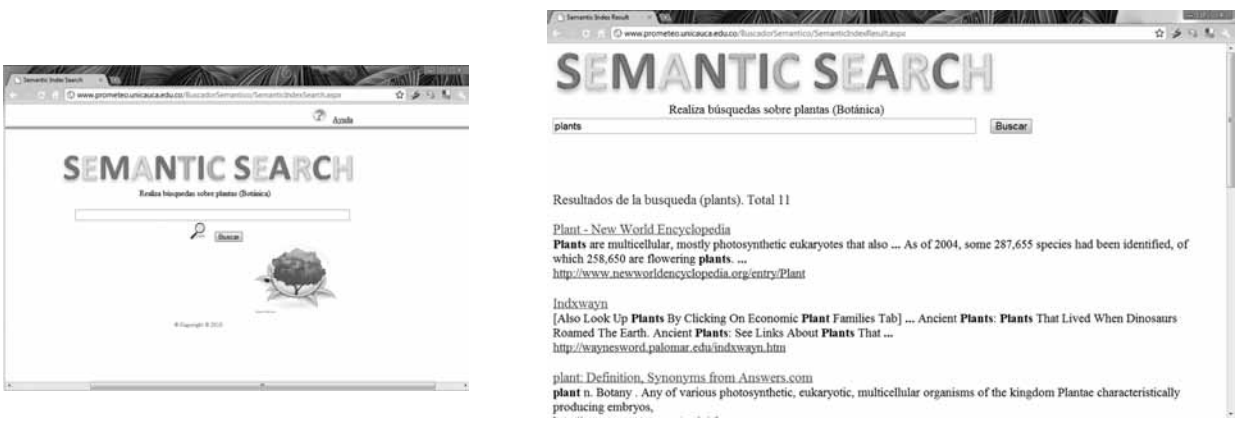


Figura 8. Interfaz del buscador Semantic Search **Fuente:** Presentación propia de los autores

5. Resultados

En la evaluación del índice creado para el presente proyecto se llevaron a cabo mediciones específicas en la recuperación de información, las cuales se describen a continuación.

5.1. Curva de precisión-recuerdo

Para realizar la evaluación del índice, se decidió tomar los primeros cuarenta documentos de la base documental con respecto a cinco consultas. Cada uno de los documentos se evaluó manualmente y se ordenaron por relevancia, de tal forma que se constituyeran en una base documental cerrada.

Para las pruebas se tomaron cinco conceptos y se midieron, para cada uno, la precisión y el recuerdo en las URLs retornadas. Luego, se realizaron consultas y se calculó la curva de precisión-recuerdo. En la Figura 9 se presenta la curva obtenida para uno de los conceptos consultados. En dicha curva se aprecia una precisión del 100% en el primer nivel de recuerdo y una precisión que varía entre el 70% y el 80% para los valores de recuerdo restantes.

Los resultados promedios para cada una de las consultas se presentan en la Tabla 4. Estos resultados muestran que la precisión es buena, a pesar del incremento en el número de documentos.

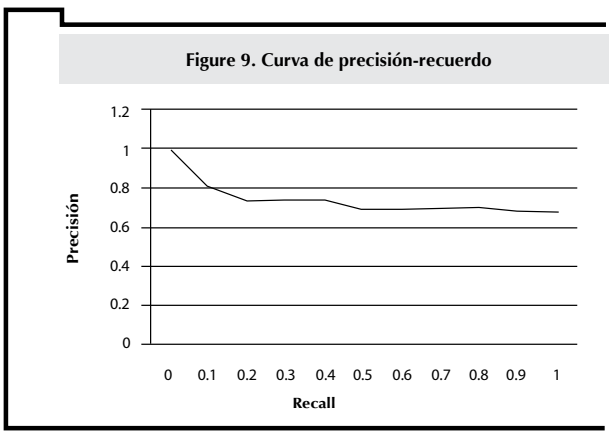


Figura 9. Curva de Precisión - Recuerdo para "flower"
Fuente: Presentación propia de los autores

Concepto buscado	Documentos relevantes	Promedio precisión
Flower	34	0.686
Seedling	7	0.201
plant structure	5	0.205
Seed	20	0.5822
Leaf	21	0.712

Tabla 4. Resumen de precisión en las consultas hechas al índice semántico de Plantas
Fuente: Presentación propia de los autores

La precisión disminuyó en algunas consultas por la falta de documentos recuperados ya que solamente se estaban analizando los primeros cuarenta documentos que devolvía delicious y sus enlaces internos, quedando temas de las plantas con muy pocos documentos para su exploración.

5.2. Índice MAP

El índice MAP (Mean Average Precision), 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 las cinco (5) consultas.

Al realizar el cálculo del Índice MAP se observa un promedio aceptable (0.48), teniendo en cuenta los resultados promedio en un sistema de recuperación de información, los cuales varían entre 0.1 y 0.7 (Manning *et al.*, 2008). Esto significa una aceptable precisión en general, de acuerdo con las consultas realizadas. Sin embargo, es necesario ampliar la base de documentos a analizar con el fin de verificar si el índice mejora.

5.3. Estadístico Kappa

El estadístico Kappa permite medir el nivel de acuerdo 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.

Para realizar las pruebas se contó con dos colegios de la ciudad, en los cuales se escogió un grupo de estudiantes entre cuarto, quinto (básica primaria)

y sexto (secundaria), dos al azar por cada una de las cinco consultas probadas. A cada uno se le asignó unas consultas para que ellos decidieran la relevancia de los resultados obtenidos en cada Url retornada y así poder establecer el acuerdo que hay entre los jueces.

El primer experimento se realizó en el colegio Campestre Americano de la ciudad de Popayán obteniendo un valor Kappa de 0.53, lo cual indicó un valor moderadamente confiable ya que el máximo valor es de 1.

Al realizar el análisis de las pruebas y las anotaciones de los comportamientos de los jueces, se detectaron varios elementos que influían en los resultados finales, inicialmente, la mayoría de los estudiantes buscaban la palabra clave dentro de las páginas retornadas sin interesarles el contenido y su relación con la búsqueda, si no la encontraban en los párrafos iniciales abandonaban la lectura. Por otro lado, los documentos retornados se relacionaban con la consulta, pero eran más generales a lo buscado.

Otro inconveniente tuvo que ver con el manejo del idioma (inglés), 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.

En la segunda versión del índice semántico se hizo un cambio, tomando los conceptos relacionados con la búsqueda hacia niveles inferiores en la jerarquía. Antes de esto se realizaba tomando los conceptos generales, es decir, los niveles superiores (padres). Con este cambio se buscaba retornar búsquedas más específicas sobre las consultas y así, encontrar los documentos relacionados con los conceptos buscados.

Para resolver 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 previamente estudiado.

Hecho lo anterior, se decidió hacer el experimento nuevamente con otra evaluación del índice Kappa en el segundo colegio, el Alejandro de Humboldt, ya que los grupos en los que se hizo la prueba anteriormente se consideraba que ya estaban sesgados por historia, los resultados del acuerdo en este caso fue mayor: 87.2%, que en el total calculado en el primer colegio evaluado y mayor que en varios casos específicos, lo cual brinda una mayor confiabilidad en los resultados de la búsqueda. Con los cambios hechos, los estudiantes encontraron más rápidamente los conceptos en los documentos y entendieron mejor el tema de que hablaban los mismos.

6. Conclusiones y trabajo futuro

Se procedió a la construcción de un procedimiento para generar índices semánticos, lo cual permite a nuevos investigadores decidir los pasos a seguir para una indexación semántica de documentos en la Web o intranet corporativa. En la definición del procedimiento se establecieron las fases, pasos, actividades, requisitos y elementos necesarios e importantes que deben tenerse en cuenta al momento de crear un índice semántico.

La aplicación de ontologías en la creación de dichos índices es de gran ayuda para mejorar los procesos de recuperación de información por sus relaciones semánticas. En el caso de estudio se utilizó una ontología de dominio particular llamada PlantOntology (creada en idioma inglés), para la extracción de las relaciones semánticas necesarias en la recuperación de información sobre el tema de botánica. Esto proporcionó una mejora en la relevancia de los resultados de las búsquedas realizadas por los diferentes usuarios de la aplicación.

El uso de un servicio Web de marcado colaborativo como Delicious, proporciona una ayuda importante en la extracción de información relevante para las búsquedas sobre un dominio específico. La ventaja de utilizar este servicio es que la base de datos documental ya ha sido clasificada y anotada colaborativamente por los usuarios del servicio, generando una ventaja en la obtención de documentos cuya relación con los temas de búsqueda con seguridad son mayores que los que se obtienen de buscadores de propósito general.

La validación del prototipo mostró una precisión media MAP de 0.48, con un confiabilidad final en los resultados dados por el índice Kappa de 87.2%. Esto indica que aunque la precisión está por encima del promedio es necesario mejorar el índice semántico creado para el caso de estudio, teniendo en cuenta las variables adicionales detectadas al momento de su utilización, así se observó que se usó un reducido conjunto de documentos

analizados y también la falta de preparación de los evaluadores (estudiantes de básica primaria) en cuanto a la forma de analizar los documentos y el manejo del idioma inglés.

En la mayoría de los casos, el nivel de precisión se mantuvo constante a pesar del incremento de los documentos recuperados, lo cual es un buen indicador para el usuario que espera relevancia en sus resultados.

Finalmente, la relevancia de los resultados del índice semántico que se construya, dependerá de varios factores que se presentan en la construcción del mismo, las decisiones que tomen sus implementadores, como el algoritmo de similitud semántica o el algoritmo de obtención de conceptos y los objetivos de uso del índice, entre otros. Por ello este trabajo permite dar una visión en cómo orientar a los investigadores en la construcción de índices semánticos y no en definir unas reglas estáticas al respecto.

Como trabajo futuro se espera probar el procedimiento propuesto en el desarrollo de índices semánticos en otros campos y aplicaciones que lo necesiten, dentro del grupo de investigación.

Agradecimientos

Se hace un especial agradecimiento a la Universidad del Cauca y al convenio de Computadores para Educar – CPE de la Universidad del Cauca, por su apoyo logístico y financiero.

NOTAS

1. Este artículo se deriva de un proyecto de investigación denominado (Procedimiento Para La Creación De Índices Semánticos Basados En Ontologías De Dominio). Desarrollado por el grupo de investigación en tecnologías de la información GTI de la Universidad del Cauca, Popayán, Colombia. Inició en Marzo del 2010 y finalizó en Diciembre del 2010. Fue cofinanciado por el Convenio Computadores para Educar Universidad del Cauca – Ministerio de Educación Nacional.

2. Análisis del Lenguaje Natural

3. Según el diccionario de la Real Academia de la Lengua Española es "Estudio del significado de los signos lingüísticos y de sus combinaciones, desde un punto de vista sincrónico o diacrónico" http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=semántica

REFERENCIAS BIBLIOGRÁFICAS

1. **AUFAURE, M. A., SOUSSI, R., & BAAZAOUI, H. SIRO: On-line semantic information retrieval using ontologies. En: IEEEXplore: Digital Information Management (2007).**
2. **AVELLO, D. G. BlindLight- 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. Oviedo, 2005, 241. Doctoral Tesis Universidad de Oviedo. Departamento de Informática.**
3. **BARITE, M. Diccionario de Organización y representación del Conocimiento: Clasificación, Indización, terminología. En: Página Web versión HTML. Montevideo (2000) [citado 13 de Abril de 2011], Disponible en Internet: <http://www.eubca.edu.uy/diccionario/index.htm>**
4. **BAZIZ, M., BOUGHANEM, M., & AUSSENAC-GILLES, N. Evaluating a Conceptual Indexing Method by Utilizing WordNet. En: (2005); 8.**
5. **BENAVIDES, K. D. R. Índices de RI. En: Página Web versión HTML. (2011) [citado 14 Abril de 2011], Disponible en Internet: <http://www.kramirez.net/RI_Maestria/Material/Presentaciones/Indices%20de%20RI.pdf>**
6. **CARRASCAL, C. Tesauros y Ontologías. En: Página Web versión HTML. (2004) [citado 14 de Abril del 2011], Disponible en Internet: <http://personales.upv.es/ccarrasc/doc/2003-2004/TesaurosOnto/principal.html>**
7. **CONSORTIUM, P. O. Plant Ontology. En: Plant Ontology™ Consortium Página Web versión HTML. (2010) [citado 10 de agosto de 2010], Disponible en Internet: <http://www.plantontology.org/]>**
8. **CORD, V., LOMBARDI, P., MARTELLI, M., & MASCARDI, V. An Ontology-Based Similarity between Sets of Concepts. En: (2005).**
9. **CHANG, C., & SCHATZ, B. Performance and Implications of Semantic Indexing in a Distributed Environment. En: Proceedings of the eighth international conference on Information and knowledge management Kansas City, Missouri, United States. (1999)**
10. **CHEN, H., SCHATZ, B., NG, D., MARTINEZ, J., KIRCHHOFF, A., & LIN., C. A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project. En: IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8) (1996); 39.**
11. **CHUNG, Y.-M., HE, Q., POWELL, K., & SCHATZ, B. Semantic Indexing for a Complete Subject Discipline. En: Proceedings of the fourth ACM conference on Digital libraries (1999); 39-48.**
12. **DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., & HARSHMAN, R. Indexing by latent semantic analysis. En: Society for information systems (1990).**
13. **DESMONTILS, E., & JACQUIN, C. Indexing a Web Site with a Terminology Oriented Ontology. En: CiteSeerX (2002); 181-198.**
14. **DESMONTILS, E., JACQUIN, C., & SIMON, L. Ontology enrichment and indexing process. En: Institut de Recherche en Informatique de Nantes 2, rue de la Houssinire Página Web versión HTML. (2003) [citado, Disponible en Internet: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.7308&rep=rep1&type=pdf>**

15. DOMINICH, S. A unified mathematical definition of classical information retrieval. En: Journal of the American Society for Information Science, 51(7) (2000); 10. 0002-8231.
16. FRAKES, W. Information retrieval: data structures and algorithms: 1992. Series,
17. GAO, M., LIU, C., & CHEN, F. An Ontology Search Engine Based on Semantic Analysis. En: Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05) Volume 2 - Volume 02.(2005)
18. GÄRDERFORS, P. Concetual Spaces as a Framework for Knowledge Representation. En: Mind and Matter, 2(2) (2004); 18.
19. HERNÁNDEZ, J. P. R., & HERNÁNDEZ, G. A. Indización y Búsqueda a través de Lucene. Veracruz, Sinaloa, 2008, Universidad Autónoma de Sinaloa, Instituto Tecnológico de Orizaba.
20. HERRERA, A. G. L. Modelos de Sistemas de Recuperación de Información Documental Basados en Información Linguística Difusa. 2006, 255. Tesis (Doctor en Informática). Universidad de Granada. Departamento de Ciencias de la Computación e Inteligencia Artificial.
21. ISO. Métodos Para el Análisis de Documentos, determinación de su Contenido y Selección de los Términos de Indización NC- ISO 5963: 2000. En: Página Web versión HTML. (2000) [citado, 1, Disponible en Internet: <http://www.energia.inf.cu/PAEC/conten/normal/CAT%C1LOGO%20DE%20NORMAS%20CUBANAS.pdf>
22. JAVIER. Ven a ver a Javier. Lista de Buscadores Semánticos. En: Página Web versión HTML. (2011) [citado, Disponible en Internet: <http://www.javi.it/semantic.html>
23. JIMÉNEZ, A. G. Instrumentos de representación del conocimiento: tesauros versus ontologías. En: Anales de documentación, Revista de Bibliotecomania y Documentación (2004). 1697-7904.
24. JONES, K. S., WALKER, S., & ROBERTSON, S. A probabilistic model of information retrieval: development and comparative experiments: Part 1. En: Information Processing and Management (2000).
25. KANG, B. Y. A Novel Approach to Semantic Indexing Based on Concept. En: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, Japan.(2003)
26. KENT, A., BERRY, M. M., LUEHRS, F. U., & PERRY, J. W. Machine literature searching VIII. Operational criteria for designing information retrieval systems. En: American Documentation, 6(2) (1955); 93-101. 1936-6108.
27. LEAL, E. T. La Desambiguación del Sentido de las Palabras: revisión metodológica. Revista multidisciplinar sobre diseño, personas y tecnología En: Página Web versión HTML. (2009) [citado 10 de marzo de 2010], Disponible en Internet: <http://www.nosolousabilidad.com/articulos/desambiguacion.htm>
28. LIN, D. An Information-Theoretic Definition of Similarity. En: Proc 15th International Conference on Machine Learning (1998); 296-304.
29. LV, G., ZHENG, C., & ZHANG, L. Text Information Retrieval Based on Concept Semantic Similarity. En: 2009 Fifth International Conference on Semantics, Knowledge and Grid (2009); 356-360.
30. MANNING, C. D., RAGHAVAN, P., & SCHÜTZE, H. Introduction to Information Retrieval: Cambridge: 2008. Series, 0521865719
31. MARCO, S. B., & KATHLEEN, S. V. An Approach to Semantic Indexing and Information Retrieval, Extraido 10 de diciembre de 2009. En: Revista Facultad de Ingeniería Universidad de Antioquia, 48 (2009); 14. 0120-6230.
32. MAZUEL, L., & SABOURET, N. Semantic Relatedness Measure Using Object Properties in an Ontology. En: Proceedings of the 7th International Conference on The Semantic Web, Karlsruhe, Germany.(2008)
33. MIHALCEA RADA, M. D. Semantic Indexing using WordNet Senses En: Department of Computer Science and Engineering (2000); 11.
34. MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D., & MILLER, K. Introduction to WordNet: An On-line Lexical Database. En: International Journal of lexicography, 3 (1993); 235-244.
35. MOLINA, M. P. Búsqueda y Recuperación de Información. En: E-COMS: Electronic Content Management Skills Página Web versión HTML. (2009) [citado 27 de abril de 2010], Disponible en Internet: <http://www.mariapinto.es/e-coms/recu_infor.htm>
36. N., J. M. D., SALTO, F., & PÉREZ, M. Recuperación de Información. . En: Página Web versión HTML. (2009) [citado 14 de julio de 2010], Disponible en Internet: <http://sites.google.com/site/glosariobitrum/Home/recuperacion-de-informacion>
37. NGUYEN, T., & PHAN, T. The effect of Semantic Index in Information Retrieval development. En: International Conference on Information Integration and web-based Applications and Services, Austria.(2008)
38. NOAH, S. A., ZAKARIA, L., ALHADI, A. C., MOHD, T., SEMBOK, T., & SAAD, S. Towards Building Semantic Rich Model for Web Documents Using Domain Ontology. En: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. (2004)
39. NOVOA, D., & BALLEN, L. La Indexación Semántica Latente en la recuperación de información. En: Preprint (2007).
40. PAGE, L., BRIN, S., MOTWANI, R., & WINOGRAD, T. The PageRank Citation Ranking: Bringing Order to the Web: 1999. Series,
41. PARDO, M. A., & FERRO, J. V. Introducción a la Recuperación de Información. En: Grupo: Lengua Y Sociedad de la Información Página Web versión HTML. Galicia (2010) [citado 09/06/2010], Disponible en Internet: <http://www.grupolys.org/docencia/ln/biblioteca/ir.pdf>
42. RADA, M., & DAN, M. Semantic Indexing using WordNet Senses En: Department of Computer Science and Engineering, In Proceedings Of Acl Workshop On Ir & Nlp, Hongkong.(2000)
43. RADA, M., & MOLDOVAN, D. I. An Iterative Approach to Word Sense Disambiguation. En: Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference Orlando, FL.(2000) http://www.cse.unt.edu/~rada/papers/mihalcea.flairs00.pdf
44. RAE. Sitio Web del Diccionario de la Real Academia Española, Segunda Edición. En: RAE Página Web versión HTML. (2011) [citado 11 de Abril de 2011], Disponible en Internet: <http://www.rae.com/>
45. RESNIK, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. En: arXiv preprint cmp-lg/9511007(02 de febrero de 2010) (1995).
46. RESNIK, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. En: Journal of Articial Intelligence Research 11 (1999); 36.

47. **ROMA-FERRI, M.T., & PALOMAR, M. Interoperabilidad Semántica de Ontologías Basada en Técnicas de Procesamiento del Lenguaje Natural. En: ISKO. CAPÍTULO ESPAÑOL. CONGRESO 7° (2005); 534-548.**

48. **S CAZALENS, E. D., C JACQUIN, AND P LAMARRE. A Web Site Indexing Process for an Internet Information Retrieval Agent System En: Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00), 1 (2000); 254-258.**

49. **SALTON, G., & MCGILL, M. J. (1986). Introduction to Modern Information Retrieval (pp. paginas 400). Retrieved from http://lyle.smu.edu/~mhd/8337sp07/salton.pdf**

50. **SAMANEH CHAGHERI, C. R., SYLVIE CALABRETTO, CYRIL DUMOULIN. Semantic Indexing of Technical Documentation. En: Laboratoire d'InfoRmatique en Image et Systèmes d'information (2009); 12.**

51. **SÁNCHEZ, D., & MORENO, A. Learning non-taxonomic relationships from web documents for domain ontology construction. En: Data & Knowledge Engineering, 64(3) (2008); 600-623. 0169-023X.**

52. **SANDERSON, M. Word sense disambiguation and information retrieval. Glasgow, 1996, 136. Tesis (PhD). University of Galsgow. Department of Computing Science.**

53. **SCHATZ, B. R. Information Retrieval in Digital Libraries: Bringing Search to the Net. En: Science - Bioinformática, 275 (1997); 327 - 334.**

54. **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. En: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (2004); 769 - 770.**

55. **SONG JUN-FENG, Z. W., XIAO W., LI G., XU Z. Ontology-Based Information Retrieval Model for the Semantic Web. 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 En: IEEE Computer Society Página Web versión HTML. Washington, DC, USA (2005) [citado, Disponible en**

56. **SONG, W., LI, C. H., & PARK, S. C. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. En: Expert Systems with Applications, 36 (2009); 9095–9104.**

57. **SWETS, J. Effectiveness of information retrieval methods. En: American Documentation (1969).**

58. **THANH NGUYEN, T. P. The effect of Semantic Index in Information Retrieval development. En: International Conference on Information Integration and web-based Applications and Services (2008); 438-441.**

59. **TUMER, D. S., M.A. BITIRIM, Y. DEPT. OF COMPUT. ENG., EASTERN MEDITERRANEAN UNIV., FAMAGUSTA. An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia. En: IEEEXplore (2009); 51-55.**

60. **UNIVERSITY, P. WordNet 3.0 Princenton University En: Página Web versión HTML. (2009) [citado 02 de febrero de 2010], Disponible en Internet: <http://wordnet.princeton.edu/wordnet>**

61. **VALLET, D., FERNANDEZ, M., & CASTELLS, P. An Ontology-Based Information Retrieval Model. En: IEEE Mendeley (2005).**

62. **YATES, R. B., & NETO, B. R. Modern Information Retrieval: New York: 1999. Series, 0-201- 39829- X**

63. **YU, D. C., L., D. J., CUADRADO, & COBURN, A. The Semantic Indexing Project knowledgesearch En: Página Web versión HTML. (2003) [citado 12/12/2009], Disponible en Internet: <http://www.knowledgesearch.org/>**

64. **ZOBEL, J. Inverted files for text search engines. En: ACM Computing Surveys (CSUR) (2006).**