La semántica en los motores de búsqueda

Sonia Ordóñez Salinas[1]

RESUMEN

Este documento presenta una revisión de los principales aportes que se han hecho en el tema del manejo de la semántica en los Motores de Búsqueda o en los Sistemas de Recuperación de Información. Como cualquier otro sistema de software los Sistemas de Recuperación de Información cuentan con una arquitectura y unas estructuras que les permiten funcionar. Es por ello que este artículo revisa no solo los aspectos generales de dichos sistemas sino que detalla como ha sido el tratamiento que se le ha dado a la semántica dentro de los mismos. Así mismo el marco del desarrollo del trabajo se inscribe dentro de lo que se considera la semántica a nivel computacional y de manera general presenta las características más relevantes, de tal forma que permita entender los temas tratados durante la revisión.

Palabras claves: Motores de Búsqueda, Sistemas de Recuperación de Información en el Web, representación de la semántica, indexación de la semántica, Arquitectura y semántica, Semántica en la Web

Semantics in the search motors

ABSTRACT

This document presents a revision of the main contributions made about handling of semantics in the Search Motors or in the Systems of Information Recovery. As any other software system, the Systems of Information Recovery have architecture and some structures that allow them to work. This paper revises the general aspects of these systems, and details the treatment that has been given to the semantics inside this software. This work is framed into the computational semantics and in a general way, it presents the most relevant characteristics, in order to understand the topics treated throughout this revision.

Key words: Search motors, Systems of Recovery of Information in the Web, representation of the semantics, indexation of the semantics, Architecture and semantics, Web Semantic

1. INTRODUCCIÓN

La acumulación de documentos y su búsqueda en la Web se ha convertido en un gran problema, En razón de que se habla de grandes volúmenes de información, pues se trata de cerca de tres billones de documentos estáticos en la Web, usados por más de 20 millones de usuarios[1]. Los usuarios establecen una consulta a través de un motor de búsqueda que retorna una lista de títulos, descripciones y enlaces de los documentos relevantes a la consulta. Generalmente, dichos motores generan sus listas con base en la presencia de las palabras digitadas por el usuario en los documentos. El que se trabaje con las palabras presenta muchas dificultades entre las que cabe destacar la aparición de documentos que nada tienen que ver con el dominio de interés del usuario, debido a la aparición de palabras cuyo significado depende del contexto. Una alternativa que se ha dado con el fin de presentar resultados más certeros, es la de incluir el significado de las palabras o más precisamente la semántica dentro los motores de búsqueda.

Sin embargo, la inclusión de la semántica es aún muy incipiente, por lo que se considera de interés revisar el manejo de las partes que se podrían constituir como la arquitectura de los sistemas de recuperación de Información.

2. SEMÁNTICA EN LA COMPUTACIÓN

Según las definiciones dadas por Rajesh [2] el estudio del Procesamiento del lenguaje natural[1] (PLN) es un área de la ciencia de la com-

[1] Directora del Grupo de Investigación GESDATOS de la Universidad Distrital Francisco José de Caldas.

Artículo recibido el 3 de Mayo de 2005 aprobado para publicación el 12 de Mayo de 2006

[1] En inglés Natural Language Processing (NLP)

putación que trata el procesamiento del lenguaje humano a través de los computadores. En PLN, se analizan las palabras, la semántica y la gramática en un lenguaje independiente del contexto. Se trata de construir sistemas computacionales que puedan reconocer y generar lenguajes entendibles al ser humano. El estudio de lenguaje humano es complejo, no solo por la ambigüedad que presentan los diferentes usos y significados de las palabras, sino por el vasto vocabulario que se maneja y el significado dependiente del contexto.

El análisis semántico es el proceso de interpretar la forma lógica de una sentencia a partir de las palabras que la conforman. La forma lógica codifica todos los posibles significados que se derivan de las palabras dentro de la sentencia.

Para el conocimiento del lenguaje natural se requiere del análisis morfológico, sintáctico, semántico y pragmático.

El análisis Morfológico trata del estudio de la estructura, variabilidad y formación de las palabras. La unidad más pequeña de una palabra es el morfema. Y una nueva palabra se crea adicionando un sufijo un prefijo o ambos. Una analizador morfológico analiza una palabra, la separa en la raíz + las flexiones y las almacena en una base de datos Léxica (ver Fig 1)

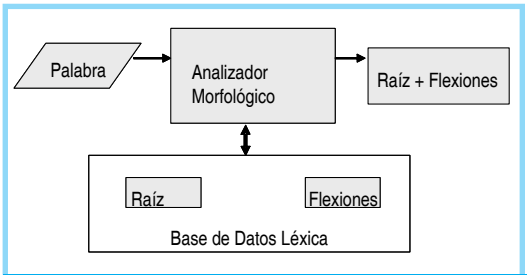

Figura 1. Analizador Morfológico adaptado de Rajesh [1]

El análisis sintáctico usa el resultado del análisis morfológico y construye una descripción estructural de la frase. Este proceso se conoce como análisis gramatical y consiste en representar la frase a través de un árbol jerárquico con las palabras que la conforman. El análisis sintáctico chequea la sintaxis a través de los nodos y las hojas del árbol.

El análisis semántico estudia el significado libre del contexto, estudia cómo las palabras se usan en diferentes situaciones y cómo el uso de estas palabras afecta la interpretación. Analiza cómo una frase se afecta por la frase que la precede o la sucede.

En PLN, las formas lógicas son las expresiones que codifican el significado de todas las frases que hacen parte de una sentencia. La representación del significado, independiente del contexto, se llama forma lógica de la frase. El proceso de representar la sentencia de acuerdo a su forma lógica se conoce como interpretación semántica. La forma lógica incluye los términos u objetos y los predicados (relaciones o propiedades), por ejemplo, María1 y María2 referencian a dos personas diferentes. La forma lógica nos da la posibilidad de construir preposiciones, combinarlas a través de operadores lógicos y resolver las posibles ambigüedades. La representación de una sentencia se puede hacer a través de estructuras sofisticadas de datos tales como redes semánticas y las ontologías.

2.1 Red Semántica

Una red semántica es la representación del conocimiento a través de nodos y arcos. Se puede usar la red semántica para representar el conocimiento léxico. Cada nodo representa un concepto y los arcos representan las relaciones entre los conceptos involucrados (ver Fig 2).

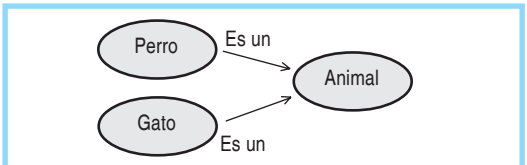

Figura 2. Red Semántica con multi-nodo

Se conocen cinco tipos de redes semánticas. 1) Red por definición, donde se establecen relaciones entre los nodos y subtipos del nodo, permitiendo la herencia y la generalización. 2) Red de implicación, usa implicaciones para la conexión de los nodos, permite representar patrones de creencia, casualidad e inferencia. 3) Red ejecutable, contiene métodos que permiten modificar y ejecutar acciones dentro de la red. Con estos métodos se pueden crear relaciones, pasar mensajes y buscar patrones. 4) Redes de aprendizaje para desarrollar información por conocimiento heredado de los ejemplos; se pueden utilizar para modificar nodos y relaciones. 5) Redes Híbridas, combinan dos o más técnicas en una o varias redes.

2.2 Manejo de la Ambigüedad de las Palabras[2]

El proceso de manejar la ambigüedad de las palabras permite determinar el contexto adecuado a partir de las palabras [3]. En este proceso se tratan de resolver algunas características lingüísticas como la homonimia, palabras con la misma ortografía y forma fonológica pero que varía su significado de acuerdo al contexto. Y la sinonimia, palabras que se escriben diferente y tienen el mismo significado. Con base en desarrollos estadísticos y de aprendizaje, a partir de la información coleccionada de los diferentes significados de cada palabra en uno o diferentes contextos, se pueden buscar relaciones que permitan determinar adecuadamente el contexto de acuerdo a una palabra. Existen varios métodos estadísticos que permiten desarrollar esta tarea, entre ellos los clasificadores Bayesianos y las cadenas de Markov o métodos de Montecarlo.

2.3 Análisis de Texto

El análisis de texto es el proceso de estructurar los datos textuales en un formato entendible para la máquina [3]. Por tanto, es el estudio del significado de las palabras y sus interrelaciones en un texto específico[4]. Para esto se requieren varias técnicas como el análisis gramatical, la identificación del lenguaje, la semejanza de textos, la clasificación de textos, el análisis de palabras y análisis de contenido. La identificación del lenguaje es necesaria para así identificar el lenguaje del documento y los caracteres usados dentro del texto[3]. Después de identificar el lenguaje en el texto, se pueden analizar la estructura y el significado del texto. Para determinar el significado del texto, se requieren conocer la estructura, los significados de las palabras y cómo las palabras se relacionan dentro del texto. El análisis de texto clasifica los documentos. Cuando se requiere recuperar información de un tópico en particular, se usan las clases predefinidas para buscar la información. Es útil en aplicaciones como minería de texto, seguridad de información, filtro de contenidos, y recuperación de información.

De acuerdo a [3], los algoritmos de semejanza de texto, permiten realizar búsquedas de una cadena o de un patrón de caracteres dentro de un documento. Existen varios algoritmos que permiten lograr esto, entre ellos los algoritmos de Knuth-Morris-Pratt, Naïve, Espacio vectorial y el de Boyer-Moore. El algoritmo de Naïve Bayes, clasifica las cadenas de caracteres de acuerdo al teorema Probabilístico de Bayes. El algortimo de Knuth-Morris-Pratt (KMP), es similar al anterior, con la diferencia de que no realiza comparaciones redundantes y almacena el conocimiento obtenido durante las comparaciones. En el algoritmo del espacio vectorial los documentos y las consultas se representan como vectores, donde cada entrada a un vector (documento) puede representar 1) la presencia o ausencia de una palabra clave; 2) la frecuencia de aparición de la palabra dentro del documento o 3) una distribución de comportamiento.

El Boyer-Moore (BM) es el algoritmo más eficiente para semejanza de cadenas de caracteres usado en los editores de texto y aplicaciones de semejanza.[3]. El algoritmo de BM compara los caracteres de una cadena dada con el texto a la derecha y a la izquierda.

El análisis de palabras permite hacer un análisis gramatical a través de técnicas como las redes neuronales. Se pueden identificar entre otros las diferentes partes de las palabras: prefijos, sílabas, morfemas y las frecuencias de aparición y pesos de las interrelaciones.

2.4 Bases Léxicas

Existen en la actualidad herramientas que permiten almacenar en una base de datos el conocimiento relacionado con el lenguaje natural. Las bases de conocimiento se utilizan para crear sistema expertos[4], que permitan procesar información para un usuario final o simplemente para experimentación. El conocimiento se adquiere entre otros de textos, de papeles técnicos, de bases de datos, documentos, informes. Para la adquisición de dicho conocimiento existen varias técnicas entre las que cabe destacar, algoritmos genéticos, redes neuronales y en general máquinas de aprendizaje, donde se busca siempre mejorar el desempeño con base en la experiencia. Ya en el mercado exis-

[2] En ingles The Word Sense Disambiguation (WSD).
[3] En la actualidad existen varias herramientas computacionales que permiten realizar esta tarea.

[4] Un Sistema experto es un programa que permite no solo depurar su propio código, sino reducir la complejidad al usar y asimilar las palabras almacenadas en una base de datos y generalmente de un domino específico.

ten varias herramientas para la adquisición de conocimiento léxico como las mencionadas en [4]: ES Shells, Repertory Grids y Text Analysis and Knowledge Mining (TAKMI), Interlinear Text (IT), Shoebox, TACT, Concordance, Conc, Text STAT, Text Analysis Tool for Object Encoding (TATOE), Intext. Algunas de ellas, no solamente pueden extractar información de grandes volúmenes de datos, sino que a través de interfaz gráfica permiten calcular frecuencias, analizar tendencias y en general generar información que permite realizar análisis sobre las palabras y las relaciones. Dentro de algunas facilidades que permiten estas herramientas están las de traducir la información a notación matemática, indexar, realizar procesos de búsqueda y armar redes semánticas.

Como caso especial dentro de las herramientas se encuentra WebBase, desarrollado por la universidad de Stanford[5]. El sistema implementa un gran repositorio de datos con aproximadamente 40 millones de páginas web y algunas características estructurales que lo plantean como novedoso. Uno de los subsistemas convierte los archivos lexicograficos en grupos lógicos de sinónimos. Otro caso especial es la base lexicográfica WordNet desarrollada por el laboratorio de Ciencia Cognitiva de la Universidad de Princenton, donde además de encontrarse bastante literatura [4] ha sido utilizada a nivel experimental como en [7].

2.5 Ontologías

En respuesta a la necesidad de gestionar el conocimiento, y a la gran cantidad de formatos que existen en los documentos de la Web, aunados al problema que se genera para el mantenimiento, surgen las ontologías como espacios de nombres, que permiten no solo expresar a través de meta-información los documentos, las terminologías y el conocimiento sino que proporcionan un camino fácil para el mantenimiento de las fuentes.

El término ontología, es utilizado en filosofía y se define como la teoría sobre el ser o la realidad. Una ontología provee una perspectiva sobre el mundo real o una parte de este. Sin embargo en el campo computacional, especifica cómo representar los conceptos y cómo relacionarlos. Las ontologías surgen no solo para estandartizar el problema de los cientos de formatos que se manejan en la Web, sino como una solución a la navegación basada en la semántica.

La definición más generalizada, la provee Gruber [8], como una especificación formal explícita de un concepto compartido. Donde un concepto se refiere al modelo abstracto de algún fenómeno dentro de un dominio, y lo formal se refiere al hecho que la ontología debe ser entendible por la máquina. Compartido, a su vez, refleja la noción de que una ontología captura el conocimiento consensual, es decir, que no es restringido a un individuo, es aceptado por un grupo.

La mayoría de los investigadores están de acuerdo con que la ontología debe incluir vocabulario y las definiciones correspondientes, pero existe la dificultad en que no se tiene un consenso en los detalles que la deben caracterizar [1].

Las ontologías se han desarrollado en el campo de la inteligencia artificial para facilitar compartir y reutilizar el conocimiento. A través de las ontologías se puede compartir y estandarizar el conocimiento. La noción de ontología se ha extendido entre otros campos a la integración de información inteligente, sistemas de información cooperativos, recuperación de información, comercio electrónico, y administración del conocimiento [1].

Al lado de las ontologías surge el objetivo de permitir que la Web tenga sentido y que no sea simplemente una red interconectada. Es así que el mismo inventor de la Web, Berners-Lee, ha acuñado el término Web Semántico, para describir una Web con sentido y esto solo se logra si se pueden describir explícitamente cada uno de sus recursos [9].

Las ontologías en la web surgen como una necesidad que es claramente expresada en [1] La Web crece inicialmente alrededor del HyperText Markup Language HTML. La simplicidad del HTML trajo consigo el crecimiento desmesurado de la World Wide Web WWW y la imposibilidad de crecer hacia dominios específicos o tareas más avanzadas. Para suplir la deficiencia mencionada aparece el Standard Generalized Markup Language (SGML), y particularmente el Extensible Markup Language XML, que a través de marcas y su propio

metadato brinda la posibilidad de describir la estructura de cada documento HTML. Se redefine el HTML como una capa superior al XML, permitiendo no solo el manejo de dominios arbitrarios sino la inclusión de la semántica como una aplicación de XML. Posteriormente surge el estándar o especificación Resource Description Framework RDF, desarrollada y mantenida bajo los auspicios del Consorcio de la World Wide Web, que integra una variedad de aplicaciones para librerías de catálogos, directorios de páginas, manejo de noticias, colecciones multimediales y una convención sintáctica sobre el XML. Por último, se definen los esquemas (RDF Schema) RDFS que permiten definir primitivas de modelado ontológico en una capa superior del RDF. Estos RDFS se definen con el Ontology Inference Layer (OIL) y el Agent Markup Language-Ontology (DAML_OIL) creado por Defense Advanced Research Projects Agency (DARPA)[13].

RDF[10] es una norma de metadatos generada por el grupo W3C [11] y define el modelamiento ontológico de primitivas, permitiendo así describir cualquier recurso de la Web. El RDFS es una recomendación formulada por Brickley [12] definida para una capa superior al RDF. Permite la definición de clases, es decir, conceptos, herencia jerárquica de clases, propiedades, restricciones de dominio y rangos de restricciones para propiedades.

OIL [14] unifica los aspectos más relevantes proporcionados por las diferentes comunidades. Permite modelamiento de formas a través de primitivas, soporta una semántica formal y un eficiente razonamiento para la descripción lógica y estándar para el intercambio de anotaciones propias del mundo Web.

En el mercado se encuentran una gran variedad de lenguajes que permiten trabajar con ontologías como Knowledge Interchange Format (KIF), Ontolingua, KL-ONE, KRL, LOOM, Classic, «Simple HTML Ontology Extensions» (SHOE), OWL.

Los textos escritos bajo lenguaje natural, como fue expuesto, cuentan con restricciones morfológicas, sintácticas, semánticas y conceptuales, por lo que construir ontologías manualmente no solamente es muy difícil sino que requiere de mucho tiempo. Es así que han surgido una serie de herramientas que permiten entre otras tareas diseñar ontologías. Estas herramientas combinan máquinas de aprendizaje, extracción de información y técnicas lingüísticas. Dentro de las tareas principales y que se espera que una herramienta cuente con ellas, se encuentran: 1) Representación de ontologías a través de lenguajes formales; 2) Aprendizaje automático para la creación y actualización de las ontologías. 3) Interfaces amigables y adaptables para cualquier dominio y 4) Servicios de razonamiento e inferencia que permitan realizar búsquedas avanzadas. Entre otras herramientas se encuentran Apollo, Circa, Protege[16], Coherence, Ontobroker, DOE, Ontobuilder [17].

Es válido aclarar que los lenguajes ontológicos permiten diseñar el modelo lógico que al implementarlo sobre un repositorio de datos (sea relacional, un esquema XML o de otro tipo) permiten almacenar las ontologías de una manera organizada y eficiente [18]. De igual forma vale la pena precisar que tanto para el diseño de los repositorios, como para la depuración y escogencia de la información que alimentará dichos repositorios se requiere la ayuda del experto del dominio. La mayoría de las herramientas permiten almacenar sus ontologías no solo para realizar búsquedas sino para poder reutilizar lo aprendido en la creación y actualización de las ontologías. Un ejemplo de una gran ontología, es la definida en WordNet [5] que provee un tesauro con más de 100,000 términos explicados en el idioma natural y la CYC[19] que proporciona las teorías formales axiomáticas para muchos aspectos comunes del conocimiento. La mayoría de los investigadores están de acuerdo que si bien una ontología puede ser almacenada en una base de datos relacional, esta se puede quedar corta ya que la sola definición de la ontología es sintáctica y semánticamente mucho más rica.

Para construir las ontologías han sido propuestas una serie de metodologías que permiten guiar su desarrollo. Dentro de ellas se pueden encontrar las revisadas por Fernádez[20], como Uschold & King, Grüninger and Fox, Bernara et alia, Methodology y Sensus. Otras como la propuesta por[21], quien plantea una metodología incremental para construir ontologías. En esta última se plantean las fases de inicio, refinamiento, y evaluación.

Se han propuesto metodologías de diversos tipos, entre las cuales la formulada en[22], donde se hace una analogía con el modelo TCP/IP en cuanto a sus niveles. En esta metodología se proponen cuatro niveles: 1) El nivel de sintaxis encargado de proveer una forma de serializar la información en una secuencia de caracteres, preservando las marcas de estructura. 2) El nivel de objeto se pasa al estándar UML. Y 3) El nivel semántico que le da interpretación al modelo de objetos.

Por último, se presenta una metodología que no solo involucra la construcción de ontologías sino el sistema que contiene las ontologías[5], es decir el sistema de conocimiento, y que se asemeja a una especificación. Dicho estándar se cita en[23] y se conoce como «The Language for Knowledge Component Markup» UPML. Se presenta como una arquitectura de software diseñada para describir los sistemas de conocimiento y está concebida para que se pueda compartir y rehusar el conocimiento. Define tareas que involucra el problema a resolver, el método de razonamiento y las ontologías que proveen la terminología usada en las tareas; métodos de resolver los problemas y definición de dominios. UPML define dos tipos de adaptadores: puentes y refinadores. Los puentes modelan las relaciones entre dos partes de la arquitectura, es decir entre el dominio y la tarea y la tarea y el método de resolver el problema. Los refinadores son usados para adaptar otros elementos como los movimientos en el espacio.

3. ARQUITECTURA DE LOS BUSCADORES Y LA SEMÁNTICA

Los Sistemas de Recuperación de Información se pueden clasificar en los llamados sistemas de recuperación clásicos y los sistemas de recuperación Web o buscadores. Los segundos, a diferencia de los primeros, se enfrentan al crecimiento exponencial de información, al dinamismo o cambio que se presenta en lapsos de tiempo muy cortos, heterogeneidad en el tipo de información que se almacena, variedad de lenguajes y formatos, duplicación de información, multiplicidad de enlaces, entre otras[24].

[5] En la mayoría de la literatura cuando se habla de ontologías aparecen los sistemas de conocimiento como la realización en software de la implementación de las ontologías.

Por otro lado en los sistemas de recuperación clásicos el desempeño de los sistemas se mide tan solo en términos del recálculo y la precisión mientras que en Sistemas de Recuperación Web, también se tiene en cuenta la calidad de las páginas retornadas[24].

Dentro de los Sistemas de Recuperación Web se encuentran los motores de búsqueda de propósito general y los de dominio especifico.

Para describir lo que es la arquitectura de un motor de búsqueda se pueden mencionar cuatro elementos que están presentes en la mayoría los motores. Uno o varios rastreadores, un indexador, un servidor de Consulta y los repositorios necesarios (ver Fig 3) . El rastreador colecciona las páginas de la Web, el indexador procesa los documentos recuperados y los representa en una estructura eficiente de búsqueda. El servidor de consultas acepta los requerimientos de los usuarios y retorna las páginas. Dentro de las estructuras generalmente se cuenta entre otros sistemas o módulos los repositorio de páginas, uno o varios índices, un léxico o repositorio de palabras [24][25].

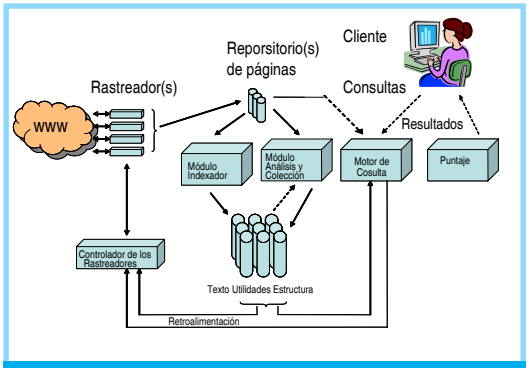

Figura 3. Arquitectura General de un Motor de Búsqueda adaptada de [25]

La arquitectura puede variar en algunos aspectos de tal suerte que le da más fortaleza a alguno de los elementos. Por ejemplo, Google cuenta con varios rastreadores distribuidos y tiene un sistema de clasificación optimizado basado en índices invertidos, las direcciones URLs y las palabras claves [24][26]. A nivel de rastreadores no solamente se ha trabajado con rastreadores distribuidos sino con rastreadores paralelos [27], rastreadores incrementales [28], donde se cambian las páginas más viejas por las más nuevas y rastreadores para la llamada información oculta o restringida[29], que cuentan con características especiales como formas de acceso y gran volumen de información estática.

Con relación a los de propósito general se encuentran Gigablast[30], WiseNut[31], iWon[32], HotBot[33], Yahoo[34], Teoma[35], MSN[36], Google[37], y al realizar una revisión en el «Search Engine Showdown Reviews, The users´ Guide to Web Search»[38] publicado en el 2004, se puede fácilmente concluir que ninguno de estos incluye la semántica, como tal, sea a través de redes u ontologías.

Otro tipo de Sistemas de Recuperación para Web o motores de búsqueda son los dedicados a un dominio específico. Dichos motores se pueden lograr de tres diferentes formas:

1) La primera se presenta a través del uso de rastreadores que coleccionen solamente las páginas del dominio en particular, como ejemplo Cora[39], especializado en ciencias de la Computación y que utiliza un rastreador basado en máquinas de aprendizaje, SPIRAL[40] y WebKB[41].

2) Otra forma es a través del uso de filtros. Dichos filtros pueden ser implementados de tres formas diferentes: por palabras claves, por la lista retornada o por medio de un «gateway». Dentro de estos buscadores se encuentran FileWacher [42], que contiene cuatro grupos de búsquedas específicas y Ahoy [43], un motor especializado en localizar páginas personales.

3) Por último, se tiene la opción del uso del refinamiento de la consulta a través de las palabras pertenecientes a la especie del dominio. Para ello se parte inicialmente de un motor de búsqueda de propósito general y a través de algoritmos se redefine la consulta. Dentro de esta categoría está el motor propuesto por Oyama [44], donde se utilizan árboles de decisión tanto para descubrir las palabras de la especie como para clasificar los documentos.

GETESS es un motor que existe comercialmente y que se podría catalogar de propósito general dado que recupera cualquier tipo de información, pero que sus creadores lo catalogan de domino especifico al generar resultados en lengua inglesa y alemana, sin que el usuario deba filtrar en qué idioma quiere el resultado. El motor GETESS se basa en el trabajo desarrollado por [45] que permite encontrar resultados de un mismo contexto sin importar la lengua (ingles, o alemán) en que esté escrito el texto. La consulta se plantea en lenguaje natural y se soporta en ontologías. En este sistema se presentan varios elementos sobresalientes, una base de datos completa, una ontología que provee un meta-conocimiento sobre las consultas y la posibilidad de refinar la consulta por parte del usuario. El lenguaje se reconoce por la fonética y por deletreo. La identificación también puede ser encontrada por el origen de la lengua. Así el uso de ciertas combinaciones de letras, o palabras claves particulares de un lenguaje pueden proveer la heurística para reconocer y clasificar el lenguaje del documento.

A nivel experimental se cuenta con desarrollos de sistemas de recuperación de Información que de una u otra forma tienen en cuenta la semántica. Dentro de los trabajos que vale la pena resaltar está el presentado por la Universidad de Bowie [47], la que desarrolla un sistema que permite construir ontologías para un dominio específico. A partir de una colección de documentos de texto relacionados, extracta la información estadística ontológica, que resulta de una serie de subsistemas: Pre-procesamiento, Normalización, Indexación por semántica Latente ISL a través del teorema de la descomposición singular, construcción del grafo y construcción de una interfaz gráfica. El Preprocesamiento se encarga de extraer los términos significativos, calcular la frecuencia de aparición dentro del texto y así obtener la matriz *documento-término.* La normalización permite calcular los pesos normalizados de cada una de las palabras obtenidas. La relación entre los documentos y los términos (ontología) se obtiene a través de la técnica de semántica Latente (ISL). Dicha técnica se basa en la descomposición espectral de la matriz *documento-término* en la forma cuadrática USV, donde la matriz U corresponde a los nodos de conceptos, la matriz V a los nodos de términos y S a los pesos correspondientes. Este sistema presenta una interfaz gráfica que permite visualizar el grafo bipartido y muestra las relaciones entre los términos y los conceptos.

Otro sistema que se merece mencionar es el WordNet, visto ya en la sección de Bases Léxicas. Este sistema incluye el manejo de la semántica a través de características de desempeño e indexación. Permite distribuir el repositorio transparentemente a través de cluster de

computadores y discos. Usa una red de discos que permite ampliar automáticamente el repositorio (grind). Apoya el acceso de cadenas de caracteres, acceso aleatorio y acceso basado en consultas. Permite manejar las actualizaciones de las páginas almacenando únicamente la última versión y eliminando las páginas obsoletas. Y como ya se había mencionado, crea una base lexicográfica que almacena las palabras con sus sinónimos y posibles significados, armando grupos lógicos de contexto [5].

La semántica se puede extractar de diferentes partes del documento como del título y de los vínculos. Dentro de sistemas de recuperación o motores de búsqueda que incluye la semántica a partir del título y principalmente de los vínculos que contenga el documento, se encuentra el desarrollado por Iraklis y demás [7] llamado THESUS. Este sistema recoge investigaciones anteriores sobre la utilización de los vínculos como el propuesto por Phelps [48] e introduce la idea de los hipervínculos para extracción de información semántica y los motores de búsqueda Kartoo [49] y Vivísimo [50] donde se utilizan estructuras para la información extractada de los hipervínculos. La técnica que permite extractar la información semántica de los hipervínculos y en sus fronteras es lo que en THESUS han definido como semántica del vínculo. La técnica consiste en extractar las palabras claves que se encuentran dentro de la frontera delimitada por <A REF.=> y </A> y a los alrededores, tanto de los vínculos que enlazan al documento como los que este enlaza. Las palabras claves se asocian a un mapa de palabras a través de WordNet obteniendo así lo que los autores llaman la semántica del vínculo. THESUS aporta, entre otras cosas, un modelo y un lenguaje para manejar la extracción de los enlaces y, un mecanismo de agrupamiento que explota una medida de similitud para organizar un conjunto de documentos web dentro de grupos de páginas similares. Dicha medida de similitud distribuye un conjunto de términos ponderados dentro de una jerarquía, teniendo en cuenta que la similitud entre dos conjuntos no se basa en las semejanzas de los términos, sino que combina todos los términos de los dos conjuntos. Permite a los usuarios a través de una interfaz gráfica ejecutar consultas SQL sobre los datos almacenados y habilitar búsquedas sobre documentos de interés basadas en semántica y enlaces.

4. REPRESENTACIÓN DE DOCUMENTOS A TRAVÉS DE ONTOLOGÍAS

En esta sección se tratará el tema de cómo representar y cómo extraer la información de un documento por medio de ontologías y algoritmos que permitan extractar la semántica.

Se tratarán solamente extractores o representaciones basadas en información semi-estructurada o no estructurada, es decir aquellas que no tienen una estructura regular como en las bases de datos.

Dentro de los extractores merece mención el prototipo desarrollado por la universidad de Stanford [51] e instalado en el TSIMMIS (The Stanford IBM Manager of Multiple Information Sources) como parte de DARPA. Presenta una herramienta para extractar datos semi-estructurados de un conjunto de paginas HTML y luego convertir la información extractada en objetos. La salida del extractor se provee bajo la especificación Object Exchange Model, OEM. Dicha especificación es de la forma (variables, fuente, patrón), donde la fuente especifica la entrada del texto que será considerada, el patrón indica como encontrar el texto dentro de la fuente y las variables corresponden a una o más variables extractoras; una variable puede ser una secuencia de comandos.

Se han presentado muchas formas de estructurar la información que hace parte de un documento, dentro de ellas algunos investigadores han presentado la opción de trabajar con plantillas como en [52]. En el trabajo de la Universidad de Stanford y reseñado en [52] automáticamente se extracta la información de las páginas Web sin ejemplos de aprendizaje u otras ayudas humanas. Se define la noción de una plantilla, y se propone un modelo que describe cómo los valores se codifican dentro de las páginas usando la plantilla. Se presentan algoritmos de extracción que usan conjuntos de palabras que tiene patrones con similares ocurrencias en las páginas de entrada para construir la plantilla. A partir de una página, por ejemplo en Amazon, se construye la plantilla, es decir se deduce todo lo que aparece común en una página, como precio de venta, título, autor, editorial, etc... Una vez construida la plantilla, el texto que sobra se constituye como los datos.

Generalmente las ontologías representan modelos basados en texto como en [16] [53].

Los modelos basados en texto son fáciles de construir, pero sus relaciones estructurales son difíciles de visualizar. Esto es importante si la ontología tiene que ser representada por un experto humano o un usuario final[54]. En el trabajo presentado por [54][55], a partir de trabajos previos de modelamiento de bases de datos orientadas a objetos se representa la ontología a través de grafos dirigidos. En este trabajo, el sistema ONION, aborda el problema de representar la semántica y la gran diversidad de fuentes que se encuentran en la Web con una construcción ontológica basada en grafos dirigidos, donde cada fuente se modela como una ontología individual, y con base en reglas de articulación relaciona los términos con la fuente. Cada ontología que se presenta como un grafo, refleja parte de una fuente de conocimiento externo. La ontología se basa en la especificación IDL, documentos basados en XML y listas simples adyacentes. ONION es un sistema diseñado para usuarios expertos en el dominio. A través de formas gráficas puede visualizar ontologías existentes, importar o borrar ontologías, especificar las reglas de articulación, visualizar posibles enlaces semánticos, formular consultas, entre otros.

Existe una tendencia a utilizar nuevos elementos del lenguaje de marcado para relacionar las secciones con los términos de la ontología, conocida comúnmente como anotación. En esta línea se encuentran bastantes trabajos que permiten representar los documentos a través de ontologías, partiendo de nuevas marcas insertadas dentro de los documentos. Vale la pena resaltar el desarrollado en [56], donde a los documentos se les hacen algunas modificaciones insertando marcas que permiten manejar las anotaciones. Dichas anotaciones determinan los conceptos y las relaciones a través de reglas establecidas. El modelo se presenta en el prototipo InfoSleuth, que cuenta con dos agentes; el Extractor de Texto que accede a recursos textuales y extrae el texto de acuerdo a una consulta; y el clasificador de texto que automáticamente construye ontologías de dominio específico y recupera conceptos de una colección de documentos. A cada documento se le colocan marcas de anotación que identifican los sustantivos, los verbos, etc. Se clasifican algunas frases como posibles conceptos (clases de la ontología) y a partir de éstos se generan las relaciones.

Al igual que en conjuntos, se han definido operaciones unarias y binarias para ontologías. Por ejemplo en ONION [55], la unión entre dos ontologías también se representa como un grafo ontológico y corresponde a los enlaces semánticos que además pueden verse como una serie de reglas algebraicas ontológicas.

Las reglas de articulación propuestas por SKAT (Semantic Knowlodge Articulation Tool), sistema desarrollado por la universidad de Stanford [57], permiten que a través de un proceso semi-automático se creen articulaciones ontológicas o uniones entre fuentes. La articulación entre dos ontologías involucra los siguientes pasos: 1) El experto proporciona a la herramienta las reglas necesarias para crear la relación, 2) SKAT sugiere la relación y la articulación basada en las reglas proporcionadas por el experto; 3) El experto a) acepta, b) rechaza o c) marca como irrelevante; 4) SKAT crea las reglas correctas y calcula la ganancia del rechazo o aceptación y 5) almacena las reglas encontradas para ser utilizadas en un futuro como retroalimentación.

El álgebra ontológica para articulaciones mencionada en [55], establece un operador unario y tres operadores binarios. El operador unario se refiere a la selección igual que en el álgebra relacional, y permite extractar toda o una parte de una ontología. Es decir que selecciona el subárbol rotado en el nodo objeto y las aristas que conectan los nodos correspondientes. Las operaciones binarias tienen como entrada dos ontologías, y retorna una sola. Dentro de estas encontramos la intersección, la unión y la diferencia. Estas operaciones se resuelven con base en funciones que incluyen las dos ontologías de entrada y sus reglas de articulación. Se resuelven sobre la base de que la intersección de dos ontologías es única; siempre puede ser determinada y las demás operaciones se generan a partir de la intersección.

5. CONCLUSIONES

El tema de la semántica, tanto en la computación como en los motores de búsqueda, es aún muy incipiente en la búsqueda de alternativas que permitan satisfacer las necesidades de la web, de las bases del conocimiento y las bases documentales cada vez más específicas.

Encontrar algoritmos que permitan entender exactamente y contextualmente lo que la mente humana quiere transmitir todavía se enfrenta a varios problemas, dentro de ellos el que el orden de las palabras no ha sido un factor suficientemente tomado en cuenta en la semántica computacional, y dado que en el lenguaje humano este factor es definitivo, tal limitante ha hecho del tema de búsquedas, relaciones, índices y por ende traducción del significado, algo altamente ineficiente.

De otro lado, el relativo gran número de palabras que constituye cualquier vocabulario y su casi innumerable número de combinaciones con sentido, a la hora del análisis implica un gran desafío para los procesos computacionales, pues requiere del uso de algoritmos estadísticos multidimensionales y por ende de la resolución de grandes ecuaciones matriciales.

Igualmente, el aspecto del contexto en el que se inscriben las frases y las palabras, es altamente significativo, pues aún en el caso de identidad morfológica, éstas pueden asumir significados diferentes de acuerdo con las temáticas en las que puedan estar inscritas (el sentido de la palabra matriz, por ejemplo, es sustancialmente diferente en matemática, medicina o mecánica). A lo anterior se suma el problema de las morfologías homólogas pero con sentido diferente en idiomas diferentes; además de los modismos y nuevos sentidos que los lenguajes dinámicos van señalando a los diferentes términos.

Lo anterior tiene como consecuencia que las arquitecturas de los buscadores o Sistemas de Recuperación de Información, pese a la utilización de grandes plataformas de hardware y software, aún se vean enfrentadas al crecimiento ilimitado de la Web y por ende al crecimiento desmesurado de sus repositorios con todo lo que ello implica. Así mismo, las deficiencias señaladas se traducen en que en los resultados de la búsqueda se genere una buena parte de producto «basura» que no tiene ningún significado para el usuario y que se constituye para éste en perdida significativa de tiempo e incluso en amenaza de desorientación si se trata de un usuario inexperto.

A pesar de que el uso de ontologías se mostraba como una potencial solución, se ha encontrado que si bien tal uso permite almacenar documentos o páginas de manera estandarizada, la gran diversidad de formatos y lenguajes con los que éstos se encuentran en la Web, en realidad a lo que conducen es a incluir un elemento adicional a los buscadores: la traducción y el almacenamiento de los documentos en los repositorios de las ontologías; haciendo aún más complejo el procesamiento. Lo anterior sin tener en cuenta que la alimentación de dicha ontología incluye la mano de un experto que difícilmente se puede automatizar ciento por ciento, por lo que en el fondo se esta encareciendo el procedimiento.

Como se puede ver, la exploración del campo de la semántica en la computación está en sus inicios, pues enfrenta el problema de entender, no solo la complejidad estructural del lenguaje sino las variabilidades del contexto que enriquecen el mundo del significado. Esto implica nada menos que ahondar en las dinámicas de la mente humana y del fenómeno sociológico.

6. REFERENCIAS BIBLIOGRÁFICAS

[1] Fensel D., Hendler J., Lieberman H., Wahlster W. Introduction, Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, The MIT Press Cambridge, Massachusetts London, England Massachusetts Institute of Technology, Copyright © 2003

[2] Rajesh K., Khurana R., Buiding and Using Semantic Representation, Natural Language Processing ReferencePoint Suite, Published by SkillSoft Corporation, 2004

[3] Saraswat P. Introduction to Text Analysis, Natural Language Processing ReferencePoint Suite, Published by SkillSoft Corporation, 2004

[4] Nawalgaria R.. Acquiring, Analyzing, and Storing Natural Words, Natural Language Processing ReferencePoint Suite, Published by SkillSoft Corporation, 2004

[5] Hirai J., Raghavan S., Garcia-Molina H. Paepcke Andreas WebBase : A repository of web pages, System Integration Technology Center, Toshiba Corp., 3-22 Katamachi, Fuchu, Tokyo 183-8512, Japan, Computer Science Department, Stanford University, Stanford, CA 94305, USA

[6] http://www.cogsci.princeton.edu/~wn/

[7] Varlamis I., Vazirgiannis M., Halkidi M., Nguyen B.. Thesus, A Closer View on Web Content Management Enhanced with Link Semantics, IEEE Transactions on Knowledge and Data Engineering (vol 16) pages: 601-611, 2004

[8] Gruber, T. R.. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5: 199–220. 1993

[9] Heflin J., Hendler J., Luke S.. SHOE—A Blueprint for the Semantic Web, Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, The MIT Press Cambridge, Massachusetts London, England Copyright © 2003 Massachusetts Institute of Technology , 2003

[10] http://www.w3.org/RDF/

[11] http://www.w3.org

[12] Brickley, D., R. Guha.. Resource Description Framework (RDF) Schema Specification 1.0 World Wide Web Consortium. 2000 Available from http://www.w3.org/TR/2000/CR-rdf-schema-20000327

[13] http://www.daml.org/

[14] http://www.ontoknowledge.org/oil

[15] Patel-Schneider P., Bechhofer, Broekstra, Decker, Erdmann, Fensel, C. Goble, F. van Harmelen, I. Horrocks, M. Klein, D. McGuinness, E. Motta, S. Staab, and R. Studer.. An Informal Description of Standard Oil and Instance OIL. 2000 http://www.ontoknowledge.org/oil/downl/oil-whitepaper.pdf.

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

[17] http://www.xml.com/2002/11/06/Ontology_Editor_Survey.html

[18] Klein M., Broekstra J., Fensel D., van Harmelen F., Horrocks I.. Ontologies and Schema Languages on the Web, Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, The MIT Press Cambridge, Massachusetts London, England Copyright © 2003 Massachusetts Institute of Technology , 2003

[19] http://www.cyc.com

[20] Fernández López M., Overview of methodologies for building ontologies, Laboratorio de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid Campus de Montegancedo, sn. Boadilla del Monte, 28660. Madrid. Spain

[21] Maedche A., Staab S., Stojanovic N., Stude R., Sure Y.. SEmantic portAL—The SEAL Approach, Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, The MIT Press Cambridge, Massachusetts London, England Copyright © 2003 Massachusetts Institute of Technology , 2003

[22] Melnik S., Decker S. A Layered Approach to Information Modeling and Interoperability on the Web, Database Group, Stanford University, fmelnik, stefangdb.stanford.edu, Revised version: Sep 4, 2000

[23] Omelayenko B., Crubézy M., Fensel Dieter, Benjamins Richard, Wielinga Bob, Motta Enrico, Mark Musen, Ying Ding, UPML-The Language and Tool Support for Making the Semantic Web Alive, Wide Web to Its Full Potential, The MIT Press Cambridge, Massachusetts London, England Copyright © 2003 Massachusetts Institute of Technology , 2003

[24] Huang, L. A Survey on Web Information Retrieval Technologies, Computer Science Department, State University of New York at Stony Brook, NY 11794-4400 , 2000

[25] Arasu A., Junghoo C., Garcia-Molina H., Paepcke A., Raghavan S. Searching the Web, Computer Science Department, Stanford University, 2000

[26] Brin S., Page L..The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, 2002 , 30 , 107-117

[27] Junghoo C., Garcia-Molina H.. Parallel Crawlers Stanford University, 2000

[28] Junghoo C., Garcia-Molina H.. The Evolution of the Web and Implications for an Incremental Crawler Department of Computer Science Stanford, CA 94305, 1999

[29] Raghavan S., Garcia-Molina H. Crawling the HiddenWeb Computer Science Department, Stanford University Stanford, CA 94305, USA

[30] http://www.gigablast.com/

[31] http://www.WiseNut.com/

[32] http://www.iWon .com/

[33] http://www.HotBot.com/

[34] http://www.Yahoo.com/

[35] http://www.Teoma.com/

[36] http://www.msm.com/

[37] http://www.Google.com/

[38] Search Engine Showdown Reviews, The users´ Guide to Web Search http://searchengineshowdown.com/reviews/

[39] McCallum A., Nigam K., RenniJason, Seymore K. Building Domain-Specific Search Engines with Machine Learning Techniques, AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace. A related paper will also appear in IJCAI'99., 1999

[40] Cohen W.W., A Web-Based Information System that Reasons with Structured Collections of Text, Proc. Second Int'l Conf. Autonomous Agents (Agents '98), pp. 116-123, 1998.

[41] Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K., and Slattery S., Learning to Extract Symbolic Knowledge from the World Wide Web, Proc. 15th Nat'l Conf. Artificial Intelligence (AAAI-98), pp. 509-516, 1998

[42] http://www.filewatcher.com

[43] Shakes J., Langheinrich M., and Etzioni O., Dynamic Reference, Sifting: A Case Study in the Homepage Domain, Proc. Sixth Int'l World Wide Web Conf. (WWW6), pp. 189-200 1997.

[44] Satoshi O., Takashi K., Toru I., Domain-Specific Web Search with Keyword Spices, IEEE Transactions on Knowledge and Data Engineering, vol 16. No1, January 2004, pág 17-27

[45] Gröticke S., Düsterhöft A., A Heuristic Approach for Recognizing a Document's Language Used for the Internet Search Engine GETESS, Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'00), IEEE, 2000

[46] DFKI Saarbrücken, Gecko mbH Rostock, and the University of Rostock

[47] Srivastava S., Gil de Lamadrid J., Karakashyan Y., Document Ontology Extractor, Department of Computer Science Bowie State University, 2000

[48] Phelps T. and Wilensky R., Robust Hyperlinks Cost Just Five Words Each, uc Berkeley TecHnical Report UCB//CSD-00-1091, Berkeley, Calif., 2000

[49] The Kartoo System, http://www.kartoo.fr, 2004

[50] Vivisimo search engine: http://www.vivisimo.com 2004.

[51] Hammer, J.; Garcia-Molina, H.; Cho, J.; Aranha, R.; Crespo, A., Extracting Semistructured Information from the Web, In Proceedings of the Workshop on Management of Semistructured Data. Tucson, Arizona, , 1997

[52] Arasu, Arvind; Garcia-Molina, Hector Extracting Structured Data from Web Pages, Project Stanford InfoLab; Database Group, Techreport, Stanford University, 2002

[53] Ontolingua, htttp://www-ksl-svc.stanford.edu:5915/doc/project-papers.html

[54] Mitra P., Wiederhold G., Kersten M., A Graph-Oriented Model for Articulation of Ontology Interdependencies. Stanford University, Stanford, CA, 94305, U.S.A. INS, CWI, Kruislaan 413, 109 GB Amsterdam, The Netherlands

[55] Mitra P., Wiederhold G An Algebra for Semantic Interoperability of Information Sources Stanford University, CA, USA 94305, U.S.A. INS, CWI, Kruislaan 413, 109 GB Amsterdam, The Netherlands.

[56] Czejdo B., Dinsmore J., Hwang C.H., Miller R., Rusinkiewicz M., Automatic Generation of Ontology Based Annotations in XML and Their Use in Retrieval Systems, First International Conference on Web Information Systems Engineering (WISE'00)-Volume1, 0296, 2000

[57] Mitra P., Wiederhold G., Jannink J. Semi-automatic Integration of Knowledge Sources Proceedings of Fusion, 1999

Sonia Ordóñez Salinas

Ingeniera de Sistemas y Especialista en Teleinformática de la Universidad Distrital Francisco José de Caldas. También es estadísta de la Universidad Nacional de Colombia. Obtuvo su Maestría en Ingeniería de Sistemas y de Computación en la Universidad Nacional de Colombia. Actualmente, se desempeña como profesora e investigadora en la Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas y dirige el Grupo de Investigación GESDATOS.