I + D

Esquema metodológico para la construcción automática de ontologías

Methodological framework for the automatic construction of ontologies

Fabrizio Bolaño López*
José Nelson Pérez Castillo**

Fecha de recepción: 20 de enero de 2013
Fecha de aprobación: 30 de abril de 2013

Resumen

El presente artículo pretende mostrar los resultados de una investigación en la que se propone un modelo para la definición y formalización de un esquema metodológico para la construcción de ontologías de forma automática, en este modelo se plantea como factor innovador una fase llamada "Fase de Construcción Automática". En dicha fase se utiliza un analizador sintáctico para procesar los documentos de la temática objeto de estudio, y además, se aplica un modelo temático probabilístico para capturar las características significativas (temas importantes) de dichos documentos. El desarrollo de la investigación se planteó bajo un esquema de fases, en el cual para cada una se especifican detalladamente las entradas, los procesos y las salidas con la implementación de elementos de trabajo denominados *work ítems*.

Como principal resultado se mostrarán los pasos detallados de la metodología utilizada, las herramientas desarrolladas, tales

* Universidad Distrital Francisco José de Caldas. Correo electrónico: fabriziobolano@gmail.com

** Universidad Distrital Francisco José de Caldas. Correo electrónico: jnperezc@gmail.com

como el proceso de etiquetado, el proceso de extracción de términos y el proceso de construcción y armado de la ontología en un archivo OWL. El resultado final de la investigación permite mostrar una efectividad del 60 % de la metodología propuesta; además establece unos retos interesantes sobre la necesidad de analizadores sintácticos y bases de datos léxica sobre todo para el idioma español.

Palabras clave

Construcción automática de ontología, generación de conocimiento, ingeniería ontológica, método Latent Dirichlet Allocation.

Abstract

The present paper shows the results of a study where a model for the definition and formalization of a methodological scheme is proposed to automatically construct ontologies. A novel feature of such a model is a phase called Automatic Construction Phase. During this phase, a syntactic analyzer is used in order to process documents related to a particular subject of study; moreover, a probabilistic theme model is applied to capture the most significant features (relevant themes) of these documents.

This research was set under a phase-wise scheme, providing a detailed per-phase description of the inputs, the processes and the outputs by implementing the so called work items.

The main contribution is the presentation of the detailed steps involved in this methodology and the tools developed; namely the labeling process, the terms-extraction process and the ontology construction-and-assembly process into an OWL file.

The proposed methodology proves 60% effective; additionally it poses interesting challenges on the type of syntactic analyzers and lexical databases required, especially for the Spanish language.

Key words

Automatic Building Ontologies, Knowledge Generation, Method, Latent Dirichlet Allocation, Ontological Engineering.

1. Introducción

La gran cantidad de información digital que encontramos hoy en día en los diferentes portales de internet han llevado a las universidades y a los investigadores de las ciencias de la información a buscar mecanismos que permitan localizar información específica, procesarla, analizarla y generar respuestas confiables acordes con las preguntas y requerimientos de los diferentes usuarios.

En el presente artículo se pretende abordar el problema de la gestión de conocimiento a partir de la construcción de ontologías de forma automática apoyándose en técnicas de procesamiento de lenguaje natural y aprendizaje automático. Este interés se fundamenta en el hecho de mejorar los tiempos de construcción de una ontología, variable que es muy alta en las metodologías actuales, también se fundamenta en la carencia de ingenieros de conocimiento en las empresas en Colombia; ingenieros que puedan asistir los procesos de generación de ontologías y que finalmente conlleven a la extracción de conocimiento útil para las organizaciones. Como alcance de la investigación se propuso la realización de un experimento informático que permitiera verificar la aplicación de la nueva metodología propuesta, para ello se planteó la utilización de una ontología de referencia construida manualmente por expertos y que aborda el vocabulario de música clásica.

Los ejes temáticos abordados como referentes bibliográficos para la realización de la investigación se dividen en: a) las generalidades de la ingeniería ontológica, b) los fundamentos teóricos para la construcción de ontologías [1], c) generalidades de la inteligencia artificial y sus áreas de aplicación como son: el procesamiento de lenguaje natural y el aprendizaje automático[2, 3], d) las relaciones semánticas entre palabras [4], y por último e) los elementos claves de la web semántica como XML, RDF y OWL[5].

Para el desarrollo de la investigación se priorizaron tres objetivos que se describen a continuación: a) la estructuración y conceptualización de los pasos y componentes de la nueva metodología, b) la especificación y caracterización de las técnicas de aprendizaje automático y procesamiento de lenguaje natural para el etiquetado de textos en el idioma español, y c) la identificación de un modelo probabilístico para la extracción de tópicos y temas de un conjunto de documentos de forma automática.

2. Metodología

La metodología que se empleó para el desarrollo de investigación consta de dos aspectos o fases principales, un aspecto teórico y otro estructural que incluye una definición y formalización. Para el desarrollo del primer aspecto se argumentan cada uno de los conceptos y referentes teóricos necesarios para la adquisición de la información pertinente a la construcción de ontologías. Seguidamente se realiza un análisis de las debilidades en las metodologías existentes para establecer los puntos sobres los cuales es necesario un nuevo esquema metodológico.

Culminada esta fase, se procedió con el segundo aspecto, el cual pretende proponer un esquema metodológico que permita realizar de forma automática la construcción de una ontología minimizando la asistencia de los ingenieros de conocimiento y los tiempos de respuesta en la construcción de esta; también se plantea y se desarrolla una experimento informático que permite simular y aplicar la metodología propuesta en la primera fase. En detalle la metodología seguida en este trabajo de investigación se dividió en cuatro etapas principales, agrupadas en dos fases asociadas con la parte de investigación teórica y la parte de investigación estructural (definición y formalización).

A continuación, se explica brevemente cada una de las etapas y sus respectivas actividades.

Fase 1:

Etapa 1: identificación de las principales metodologías para la construcción de ontologías. Se realizó una revisión bibliográfica

que permitió identificar cuáles son las metodologías actuales para la construcción de ontologías tanto de forma manual como de forma semiautomática. También se realizó una revisión sobre los mecanismos para validar ontologías.

Etapa 2: identificación de técnicas de procesamiento de lenguaje natural y aprendizaje automático. Durante esta etapa se realizó una revisión bibliográfica de las principales técnicas y métodos de procesamiento de lenguaje natural y aprendizaje automático para el idioma español. En esta revisión se determinó cuales métodos son los más rápidos y efectivos para el tratamiento de textos en el idioma español y se procedió a seleccionar un etiquetador (tagger) que cumpliera con las expectativas buscadas, finalmente se seleccionó el TreeTagger [6].

Fase 2:

Etapa 3: identificar un método probabilístico para hallar las características significativas (temas o conceptos) de un conjunto de textos. Se realizó una revisión bibliográfica sobre los posibles métodos que permitieran identificar características significativas en un texto de forma automática y con alto porcentaje de certeza. En esta revisión se seleccionó el método Latent Dirichlet Allocation [7, 8].

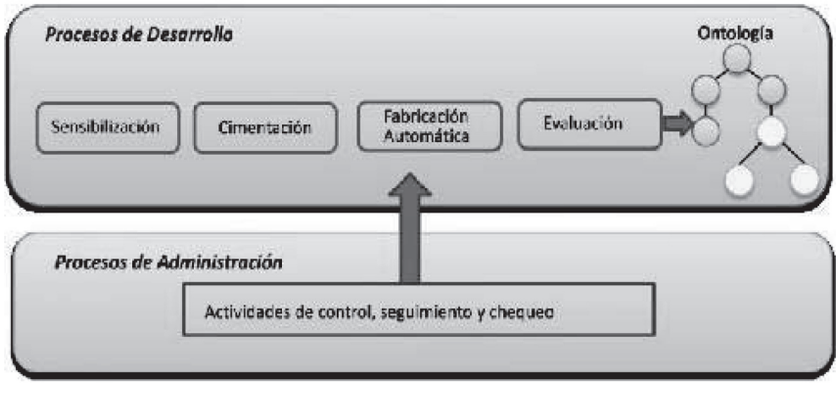
Etapa 4: definición y formalización de una metodología para la construcción automática de ontologías apoyadas con técnicas de aprendizaje automático. Se planteó un esquema metodológico que permita llegar a construir una ontología de un dominio específico de forma automática, es decir, procesando los textos e identificando los temas principales existentes en ellos de una forma automática. Durante esta etapa también se diseñó un experimento informático que permitiera simular un ambiente de prueba en el que se verifica el modelo propuesto a partir de la utilización de una ontología de referencia sobre temas de música clásica. A continuación se describe detalladamente la etapa 4.

3. Metodología ACO

La metodología ACO es una metodología para la construcción automática de ontologías apoyada en técnicas de procesamiento de lenguaje natural y aprendizaje automático, que fue inspirada en las metodologías manuales existentes para la construcción de ontologías.

En la figura 1 se presenta la metodología ACO, la cual está compuesta por dos grandes contenedores.

Figura 1. Metodología ACO



Fuente: elaboración propia.

3.1 Contenedor de procesos de administración

Este contenedor está integrado por las tareas de control, seguimiento y chequeo que el equipo del proyecto debe llevar a cabo para soportar todo el proceso de desarrollo.

3.2 Contenedor de procesos de desarrollo

Está integrado por todas las actividades que conforman el ciclo de vida de la construcción de la ontología. Las fases del contenedor de desarrollo son:

- Fase de sensibilización: permite establecer la justificación del porqué y para qué se quiere desarrollar una ontología, se identifican los beneficios y los retos del proyecto y se designan los roles y los beneficiarios.

- Fase de cimentación: ermite estructurar el dominio objeto de estudio de la ontología mediante esquemas formales de representación. El resultado de esta actividad es el modelo conceptual de la ontología. El primer paso para la representación formal de la ontología será construir un glosario de términos de referencia que incluye todos los términos relevantes del dominio (conceptos, instancias, atributos, relaciones entre conceptos, etc.), sus descripciones en lenguaje natural, y sus sinónimos y acrónimos.

- Fase de fabricación automática: permite extraer términos de un conjunto de documento y llevarlos a un lenguaje de ontologías pero de forma automática. Esta fase es el gran factor innovador del presente trabajo, está compuesto por varios Work Items (elementos de trabajo), los cuales se describen en detalle a continuación

Work Item: Preprocesamiento: en este elemento de trabajo se reciben como entrada los documentos seleccionados del dominio de aplicación seleccionado. A estos documentos se le aplicará de forma automática los procesos de etiquetado utilizando el analizador sintáctico TreeTagger y luego se realizará una selección de sustantivos. El primero consiste en obtener la estructura gramatical (verbos, preposición, sustantivo, etc.) de las palabras de cada sentencia del documento, y el segundo consiste en eliminar todas las palabras que no sean sustantivos.

*Work Item: Inserción de datos en Sql Server: e*n este elemento de trabajo se recibe como entrada los documentos de la fase anterior con los sustantivos y se registrarán en una base de datos denominada ACO, desarrollada por los autores en Sql Server. El modelo que soportará el proceso se presenta en la tabla 1 y se muestra a continuación:

*Work Item: matriz documentos-palabras: e*n este elemento de trabajo se ejecutarán unas consultas en la base de datos para crear dos tipos de archivos que serán la entrada al modelo LDA. Los archivos que se crearán son: archivo 1 (vocabulario.csv): este archivo contendrá un listado de todas las palabras de los documentos seleccionados en la etapa anterior. Es importante anotar que cada palabra solo aparecerá una vez en este archivo. La consulta para generar este archivo es la siguiente:

SELECT ROW_NUMBER() over (order by palabra) as Id, palabra INTO Vocabulario FROM (select distinct palabra from docpalabra) AS Palabra

Archivo 2 (Index_Doc_Palabras.csv): este archivo contendrá un listado con el número que identifica una palabra y el número del documento donde se encuentra. La consulta para generar este archivo es la siguiente:

SELECT DP.doc,V.Id

FROM DocPalabra as DP LEFT JOIN vocabulario as V on

V.palabra=DP.palabra

Los documentos creados serán la entrada para el modelo LDA.

Tabla 1. Descripción de las tablas de persistencia en la metodología ACO

Nombre de la tabla	Atributos	Descripción
Documento	Id Int Descripcion varchar(100)	Contiene el código y la descripción de los documentos objeto de estudio.
Doc. palabra	Id Int IdDoc Int Palabra varchar(100)	Contiene las palabras (sustantivos) de cada documento.
Vocabulario	Id BigInt Palabra varchar(100)	Contiene todas las palabras de forma única, seleccionadas a partir de la tabla Doc. Palabra
Tópicos	Id Int Nombre varchar(30)	Contiene los tópicos extraídos de los documentos
Valores	Id Int Topico int Cadena varchar(100) Distribución doublé	Contiene las palabras de cada tópico con la distribución de probabilidad con referencia a la aparición en el documento
Medias	Topico Int Media decimal	Contiene la media de la distribución de todas las palabras contenidos en cada tópico
Taxonomía	Código int Id Varchar(100) Label Varchar(100) Etiquetapadre Varchar(100)	Contiene los términos con sus respectivos hiperónimos para el armado de la ontología

Fuente: elaboración propia.

Work Item: extracción de temas: en este elemento de trabajo se infiere un conjunto K de temas de los textos del dominio objeto de estudio mediante la aplicación del proceso generativo inverso de los modelos probabilísticos. Es decir, a partir del conjunto de documentos ya conocidos se identifican los temas abordados por estos aplicándoles el modelo LDA (Latent Dirichlet Allocation). Para esta tarea se utilizó el Matlab Topic Modelling ToolBox 1.4 disponible en la siguiente URL:

http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

Se realizaron ajustes sobre las funciones del ToolBox y se creó una función llamada Fun_Lda(int n) que recibe un parámetro de tipo numérico que representa el número de temas que se quieren identificar al realizar las corridas del modelo.

Work Item: Identificación de Temas con Latent Dirichlet

Allocation:

En este elemento de trabajo se ejecutará la función Fun_Lda()

modificando los siguientes parámetros:

- Número de temas.
- Número de iteraciones.
- Número de elementos por temas.

Se deberán realizar varias corridas con el fin de obtener resultados que permitan realizar comparaciones para determinar el mejor agrupamiento de términos en un tema. El archivo resultante de esta fase se denomina: Topic_Ejemplo.txt, el cual tendrá los siguientes campos:

- Título del tema.
- Distribución total de probabilidad del tema.
- Términos de cada tema.
- Distribución de probabilidad de cada elemento con respecto a su tema contenedor.

Work Item: extracción de términos: en este elemento de trabajo se tomará el mejor resultado de la fase anterior que se encuentra en el archivo Topic_Ejemplo.txt y se insertará en la base de datos, específicamente en las tablas: temas y valores. Luego se procederá a encontrar la media de la distribución de probabilidad de los elementos de cada tema y se guardará en la tabla: medias.

Como paso siguiente se seleccionarán los elementos de cada tema que superen el valor de la media de su grupo. La consulta para esta operación se describe a continuación:

SELECT Valores.id, Valores.topico, Valores.cadena, Valores.distribucion, Medias.media FROM Valores INNER JOIN Medias ON Valores.topico = Medias.topico WHERE Valores.distribucion>Medias.media

Al ejecutar la consulta anterior se podrán obtener los términos que se utilizarán para armar la ontología.

Work Item: identificación de hiperónimos: en este elemento de trabajo se tomarán los términos del *work item* anterior y se procederá a buscar los hiperónimos de cada uno de ellos con el fin de buscar la raíz y elementos superiores de la ontología. Para la identificación de hiperónimos se contará con la consulta de dos bases de datos léxicas: WordNet [9] y Spanish Word Net 3.0. La decisión de utilizar estos dos recursos léxicos se debe a que la presente investigación está enfocada en el idioma español y como WordNet solo cubre el idioma inglés, se necesita la segunda herramienta para poder encontrar la traducción de términos. La salida de este elemento de trabajo serán los hiperónimos de cada término encontrado en la fase anterior y se construirá una estructura a manera de árbol en una tabla denominada taxonomía. Para determinar el elemento raíz se encontrará el hiperónimo con mayor frecuencia entre los términos seleccionados.

Work Item: armado de la ontología: en este elemento de trabajo se contará con una aplicación desarrollada por los autores, en Visual Studio .NET, para recorrerla tabla taxonomía y construir de forma automática la ontología. La salida de esta fase será un archivo OWL llamado OntologiaACO.owl.

Fase de evaluación: para la realización de esta fase se utilizarán dos aspectos: el primero será el validador de la W3C que se encuentra en la siguiente URL: http://www.w3.org/RDF/Validator/, y el segundo aspecto será la utilización la Medida F [10] para identificar el porcentaje de similitud

4. Experimento informático

Para el desarrollo del experimento informático se tomó como referencia una ontología existente que estuviese construida por expertos y debidamente validada; se optó por una ontología desarrollada sobre música clásica y espectáculos de dicho tipo de música. La ontología seleccionada está disponible por medio de la URL: http://www.kanzaki.com/ns/music#

El objetivo del experimento consistió en seleccionar un conjunto de documentos web, cuyo tema central fuese la música clásica, para luego aplicar la metodología propuesta en este trabajo de investigación y así poder comprobar los resultados de su consistencia. Para la selección de los documentos se contó con la colaboración de una licenciada en Música de la Universidad del Atlántico, quien realizó el agrupamiento de los documentos web necesarios para dicho experimento. Se

aplicó cada fase descrita en la sección anterior: *fase de sensibilización, cimentación, fabricación automática y evaluación*. A continuación, solo se describirá en detalle la fase de fabricación automática realizada durante el experimento informático ya que fue el eje de mayor aporte de la presente investigación.

5. Fabricación automática para la ontología de temas de música clásica

Para el desarrollo de esta fase se procedió a procesar veintidós documentos seleccionados por la asesora en bellas artes, luego se procedió a seguir los siguientes *work items*:

- Etiquetar palabras con TreeTager: para el presente caso de estudio se procedió a utilizar la aplicación MetodologiaACO desarrollada por los autores en Visual Studio .NET.

A continuación se describe el funcionamiento del programa: Se procedió a iniciar con el paso 1 que consiste en el análisis sintáctico. En este paso se busca en el directorio donde están los archivos, se seleccionan los archivos *.txt que corresponden a los links de los sitios web seleccionados por la asesora en bellas artes, luego se pulsa un botón que ejecuta el procedimiento que aplica a cada archivo el etiquetado de cada frase.

El paso No 2, consistió en depurar los archivos procesados para que solo se mantengan los sustantivos.

En este paso el usuario busca el directorio donde están los archivos que se procesaron en la etapa anterior, luego se presiona el botón Procesar Sustantivos para eliminar todas aquellas estructuras gramaticales que no sean sustantivos.

- Insertar sustantivos en la base de datos Sql Server: en esta fase se procedió a insertar cada documento con sus respectivos sustantivos en la base de datos Sql Server.

- Creación de la matriz Documentos-Palabras: en esta fase se procedió a ejecutar las consultas para crear la matriz documentos-palabras y generar los dos archivos necesarios para alimentar el modelo LDA. Se generaron los siguientes archivos: vocabulario.csv e index_doc_palabras.csv

- Identificación de Temas con Latent Dirichlet Allocation (LDA): en esta fase se procedió a ejecutar una función en Matlab denominada Fun_Lda() con los archivos generados en la etapa anterior. La salida de este paso fue la generación del archivo TopicsEjemplo.txt

- Extracción de términos: en esta fase se procedió a tomar los términos identificados en la fase anterior y que están en el archivo TopicsEjemplo.TXT para insertarlos en la base de datos de Sql Server en las tablas temas y valores. Luego se procedió a ejecutar unas consultas para extraer solo aquellos términos cuya distribución de probabilidad supera la media de la distribución de probabilidad del tema al que pertenecen. El resultado final de esta fase fue la detección de 67 términos.

- Identificación de hiperónimos: en esta fase se procedió a utilizar las bases de datos léxica WordNet y Sapnish Wordnet para obtener los hiperónimos de cada uno de los términos obtenidos de la fase anterior. La raíz de la ontología se obtuvo hallando el hiperónimo de mayor frecuencia entre los términos objetos de estudio. Luego de hallar los hiperónimos se procedió a insertar los términos encontrados en una tabla de la base de datos llamada taxonomía con la respectiva estructura anidada para facilitar el armado de la ontología.

- Armado de la ontología: en esta fase se procedió a ejecutar una aplicación en Visual Studio .NET desarrollada por los autores, la cual permite recorrer la tabla y generar una ontología bajo el estándar

de la W3C. El resultado de éste proceso es la creación de un archivo OWL llamado OntologiaACO.owl. El archivo está disponible en la siguiente URL: http://www.beewitsoft.com/ontologias/OntologiaACO.o wl

6. Resultados

Los resultados obtenidos al finalizar el experimento informático fueron los siguientes: Como primer aspecto se utilizó el validador de la W3C (http://www.w3.org/RDF/Validator/) que permitió la verificación de la correcta estructura del archivo OWL.

Para la realización de la validación se procedió a ingresar al link descrito anteriormente y luego buscar la sección nombrada como Check By URI (Validar mediante URI), en esta sección se digitó la dirección donde se encuentra el archivo owl obtenido como resultado del experimento informático y que se encuentra disponible en: http://www.beewitsoft.com/ontologias/OntologiaACO.owl.

La anterior validación nos permitió comprobar que la estructura del archivo OWL generado por la metodología propuesta en la presente investigación cumple con el estándar definido por la W3C.

Como segundo aspecto de validación se realizó una consulta en el lenguaje Sparql mediante el software Twinkle Sparql Tools, a la ontología construida con la metodología propuesta en ésta investigación. La consulta realizada fue la siguiente: "Mostrar los nombres de los instrumentos de cuerda". La consulta Sparql que responde a la pregunta es la siguiente:

SELECT ?nombre

WHERE {

{?qw<http://www.w3.org/2000/01/rdf-schema#subClassOf>

<http://www.beewitsoft.com/ontologias/OntologiaACO.owl#Instr umentos_Cuerda>}

{?qw <http://www.w3.org/2000/01/rdf-schema#label> ?nombre}

}

La anterior actividad nos permitió concluir que la estructura del archivo OWL generado por la metodología propuesta en la presente investigación soporta consultas en el lenguaje SPARQL.

Como tercer aspecto se utilizó la medida F para identificar el porcentaje de similitud con la ontología de referencia, dicha medida utiliza las métricas de cobertura y precisión.

Precisión: proporción de conceptos correctamente identificados con respecto al total de conceptos identificados en la ontología construida durante el experimento informático.

Cobertura: proporción de conceptos correctamente identificados con respecto al total de conceptos de la ontología de referencia sobre música clásica.

En la tabla 2 se describen los datos para el cálculo de la Medida F.

Tabla 2. Variables para el cálculo de la medida

Descripción	Total
Términos de la ontología de referencia (Vocabulario Música Clásica) – TOR	49
Términos de la ontología construida durante el experimento informática – TOEI	67
Términos correctamente identificados en la ontología construida con el experimento – TCIOEI	35

Fuente: elaboración propia.

Cálculo de la precisión:

$$\text{Cálculo de la Precisión} : \frac{TCIOEI}{TOEI} = \frac{35}{67} = 52.238806$$

$$\text{Cálculo de la Cobertura} : \frac{TCIOEI}{TOR} = \frac{35}{49} = 71.4285714$$

Cálculo de Medida F :

$$\frac{2 * Precisión * Cobertura}{Precisión + Cobertura} = \frac{2 * 52.238806 * 71.4285714}{52.238806 + 71.4285714}$$

$$\frac{2 * Precisión * Cobertura}{Precisión + Cobertura} = 60.34482759$$

En la tabla 2 se presenta un resumen de la cantidad de elementos necesarios para calcular la medida F, la cual nos permitió comprobar que existe un 60,34 % de semejanza entre la ontología de referencia y la construida en el experimento informático aplicando la metodología propuesta en la presente investigación.

7. Conclusiones

La presente investigación se planteó para responder a la pregunta: ¿qué grado de confiabilidad tendría una ontología creada de forma automática apoyándose en técnicas de aprendizaje automático? Su desarrollo y la realización del experimento informático permitieron responder que el grado de confiabilidad de una ontología creada bajo un esquema de fabricación automática sería del 60 % aproximadamente. También se pudo identificar al TreeTagger como una herramienta muy precisa y eficaz de etiquetado para el idioma español; de igual forma, nos permitió probar el método Latent Dirichlet Allocation para la identificación del tema a partir de un conjunto de contenidos sobre un tema específico, con resultados sobresalientes.

De lo anteriormente expuesto se puede concluir que la aplicación de técnicas de aprendizaje automático y procesamiento de lenguaje natural como las incluidas en la presente investigación son de gran utilidad para la generación y procesamiento de conocimiento útil para las organizaciones. La metodología ACO propuesta en esta investigación es un gran punto de partida para automatización de fases en la construcción de ontologías y en la consolidación del aprovechamiento de la web semántica en las universidades y organizaciones de Colombia.

8. Referencias

[1] A. Gómez-Pérez *et al.*, "Methodologies and methods for building ontologies" in Ontological Engineering: with examples from the areas of Knowledge Management, e- Commerce and the Semantic Web. Springer Verlag, Inglaterra 2004, ch 3, Sec 3.3, pp 113-153.

[2] M. Vallez y R. Pedraza. "El procesamiento del lenguaje natural en la recuperación de información textual y áreas afines". [En línea]. Hipertext.net, núm. 5, 2007. Disponible en http://www.hipertext.net.

[3] L. de León, S.Schawb y E. Wehrli, "Análisis sintáctico profundo del español: un ejemplo del procesamiento de secuencias idiomáticas. Spanish deep parsing: the example of idiomatic sequences processing". Procesamiento del lenguaje natural, Ginebra Univ, vol. 41, pp. 37-44, 2008.

[4] N. Campoy Garrido, "Relaciones semánticas entre las palabras: hiponimia, sinonimia, polisemia, homonimia y antonimia. los cambios de sentido", Contribuciones a las Ciencias Sociales. 2010. [En línea] disponible en www.eumed.net/rev/cccss/08/ncg.htm.

[5] E. Méndez. "RDF: un modelo de metadatos flexible para las bibliotecas digitales del próximo milenio", E-prints in Library and Information Science, Barcelona. 2009.

[6] H. Schmid, "Probabilistic part-of-speech tagging using decision trees". 1994. [En línea] disponible en: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf.

[7] D. Blei *et al.*, "Latent dirichlet allocation", The Journal of Machine Learning Research, Massachusetts, vol. 3, pp. 993-1022. 2003.

[8] D. Blei, "Introduction to probabilistic topic models", Communications of the ACM, Princeton Univ, 2011.

[9] C. Fellbaum, WordNet: An electronic lexical database, Massachusetts, MA: MIT Press, 1998.

[10] A. Rodrigo *et al.*, "Comparación de enfoques para evaluar la validación de respuestas", Procesamiento del lenguaje natural, Madrid, vol. 43, pp. 277-285. 2009.