PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

DISEÑO DE UNA HERRAMIENTA PARA LA ANOTACIÓN SEMÁNTICA AUTOMÁTICA DE DOCUMENTOS BASADOS EN ONTOLOGÍAS EN EL DOMINIO DE LA INGENIERÍA INFORMÁTICA

Tesis para optar el Título de Ingeniero Informático**,** que presenta el bachiller:

Rodrigo Jesús Espinoza Florez

ASESOR: Héctor Andrés Melgar Sasieta

Lima, noviembre de 2014

Resumen

Analizando la situación de la Web en la actualidad en cuanto a la gestión y búsqueda de la información que hay en ella, el siguiente documento propone una herramienta de anotación semántica automatizada como alternativa de solución al trato de la información que se genera en línea. Básicamente, una herramienta de anotación semántica puede contribuir con muchas otras aplicaciones como herramientas de búsqueda, de organización, repositorios, etc.; y al apoyarse en una ontología de un campo determinado, el desarrollo de la herramienta puede extenderse a otros campos específicos mientras se cuente con la información y los expertos respectivos en el modelado del conocimiento.

El siguiente proyecto en específico será beneficioso para la búsqueda y organización de diferentes documentos del campo de las ciencias de la computación desarrollados tanto en la universidad como fuera. Esto supondría que todos los miembros de la comunidad universitaria pudieran tener acceso a todos los contenidos del campo sin tener que gastar muchos recursos como tiempo y dinero.
Entre los principales beneficios está la reducción de tiempo en búsqueda de materiales de información del campo, así como evitar volver a generar conocimiento que ya se encuentra en la Web o ya ha sido investigado en la universidad.

Por último, además de la información recopilada en la investigación de una herramienta de esta naturaleza, se propone un diseño y un conjunto de recursos para desarrollarla, los cuales fueron probados en un conjunto de documentos pertenecientes al campo de la ingeniería informática en la universidad.

Dedicatoria

Dedico el presente trabajo a mi familia y a todos los que de alguna forma me acompañaron desde que ingrese a esta casa de estudios.

Agradecimientos

Agradezco a mi familia por su apoyo y comprensión, el cual nunca mermo desde que ingrese a la universidad hasta terminar mi carrera.

Al profesor Andrés Melgar por dedicarme su tiempo y mostrarme nuevos campos de la carrera que era muy interesantes y yo desconocía.

A Héctor y Bonnie, por su ayuda y confianza en esta aventura, a pesar de conocerme desde hace poco.

Y a todos aquellos que supieron comprenderme y alentarme durante el tiempo que estuve abocado a este trabajo.

Contenido

Resumen *2*

1 CAPÍTULO 1: PLANTEAMIENTO *10*

1.1 Problemática **10**

1.2 Objetivo general **11**

1.3 Objetivos específicos **12**

1.4 Resultados alcanzados **12**

1.5 Herramientas, métodos, metodologías y procedimientos **13**

1.5.1 Introducción **13**

1.5.2 Herramientas **14**

1.5.3 Métodos y Procedimientos **16**

1.5.4 Metodologías **16**

1.5.5 Alcance **17**

1.5.6 Limitaciones **17**

1.5.7 Riesgos **18**

2 CAPÍTULO 2: MARCO DE REFERENCIA *19*

2.1 Marco Conceptual *19*

2.1.1 Introducción **19**

2.1.2 Web Semántica **19**

2.1.3 Ontología **21**

2.1.4 Anotación **23**

2.1.5 Metadatos **24**

2.1.6 Procesamiento de Lenguaje Natural **25**

2.1.7 Conclusión **25**

2.2 Estado del arte *26*

2.2.1 Introducción **26**

2.2.2 Método **27**

2.2.3 Aplicaciones **28**

2.2.3.1 Plataforma KIM 28

2.2.3.2 Proyecto MOLTO 28

2.2.3.3 Framework de anotación semántica automatizada para artículos de Wikipedia 29

2.2.3.4 La herramienta FLERSA 30

2.2.3.5 Apache Stanbol 31

2.2.3.6 OpenCalais 31

2.2.4 Tabla de comparación **32**

2.2.5 Conclusiones sobre el estado del arte **33**

3 CAPÍTULO 3: DISEÑO DE LA HERRAMIENTA *34*

3.1 Objetivo Especifico N°1: Proponer una estructura modular para el desarrollo de una herramienta de anotación semántica automatizada. 34

3.2 Resultado Alcanzado N°1: Modelo de diseño de una herramienta de anotación semántica automatizada. 35

3.3 Consideraciones Finales del resultado N°1 38

4 CAPÍTULO 4: PROCESAMIENTO TEXTUAL DE LOS DOCUMENTOS 39

4.1 Objetivo Específico N°2: Soportar el procesamiento de la información textual de documentos 39

4.2 Resultado Alcanzado N°2: Mecanismo de procesamiento de lenguaje natural que permita obtener los diversos conceptos que se encuentran en un documento. 39

4.3 Resultado Alcanzado N°3: Mecanismo de procesamiento del lenguaje natural para la obtención de la forma canónica de un concepto determinado 42

4.4 Consideraciones Finales de resultados N°2 y N°3 43

5 CAPÍTULO 5: ONTOLOGÍA DE LA HERRAMIENTA 46

5.1 Objetivo Especifico N°3: Permitir la representación del contenido de diversos documentos cuya información se encuentra en una ontología del campo de la Ingeniería Informática. 46

5.2 Resultado Alcanzado N°4: Modelo de Ontología cuyo contenido sea el de los cursos Fundamentos de Programación, Lenguaje de Programación I y Sistemas Operativos de la especialidad de Ingeniería Informática en la Pontificia Universidad Católica del Perú. 46

5.3 Consideraciones Finales de Resultado N°4 49

6 CAPÍTULO 6: DESAMBIGUACIÓN DE TÉRMINOS 50

6.1 Objetivo Especifico N°4: Permitir la desambiguación de términos de un documento del campo de las Ciencias de la Computación usando una ontología del campo de la Ingeniería Informática. 50

6.2 Resultado Alcanzado N°5: Mecanismo de desambiguación de términos en un documento del campo de las ciencias de la computación. 50

6.3 Consideraciones Finales del resultado N°5 53

7 CAPÍTULO 7: PERSISTENCIA DE LAS ANOTACIONES 55

7.1 Objetivo Especifico N°5: Soportar la persistencia de anotaciones semánticas en documentos. 55

7.2 Resultado Alcanzado N°6: Formato de anotaciones semánticas en documentos. . 55

7.3 Resultado Alcanzado N°7: Mecanismo de persistencia de anotaciones semánticas en una base de datos relacional. 56

7.4 Consideraciones Finales de resultados N°6 y N°7 57

8 CAPÍTULO 8: CONCLUSIONES Y RECOMENDACIONES 58

8.1 Conclusiones 58

8.2 Recomendaciones para trabajos futuros 59

Referencias bibliográficas 60

Tabla de Ilustraciones

Ilustración 1: Representación de capas de la Web semántica. Imagen recuperada de M.C. Daconta, L. J. Obrst y K. T. Smith (2003) 21

Ilustración 2: Una ontología y su relación con el contenido de un documento. Imagen recuperada de A. Kiryakov (2004) 22

Ilustración 3: Representación la arquitectura de MOLTO. Imagen recuperada de M. Chechev (2012) 29

Ilustración 4: Herramientas compatibles con OpenCalais. Imagen capturada de la página oficial de OpenCalais [url= http://www.opencalais.com/about] [Último acceso: 11 Noviembre 2013] 32

Ilustración 5: Diagrama de arquitectura de la herramienta de anotación semántica automatizada propuesta. Imagen producida por el autor. 37

Ilustración 6: Conversión de Documento PDF a Texto Plano. Imagen producida por el autor. 40

Ilustración 7: Diagrama de clases lingüistas de Freeling. Imagen recuperada de Padró, Lluís and Evgeny Stanilovsky (2012) 41

Ilustración 8: Conceptos extraídos del documento. Imagen producida por el autor. 42

Ilustración 9: Conceptos extraídos del documento junto a su respectivo lema. Imagen producida por el autor 43

Ilustración 10: Ontología creada desde cero con Protégé. Imagen producida por el autor 47

Ilustración 11: Ontología del dominio de Ingeniería Informática en la Universidad. Imagen producida por el autor. 48

Ilustración 12: Estructura de un objeto concepto. Imagen producida por el autor. 53

1 CAPÍTULO 1: PLANTEAMIENTO

1.1 Problemática

La Web, como la conocemos en la actualidad, almacena gran parte del conocimiento en el mundo, aportado por todos sus usuarios de ella en el planeta. Esta crece a pasos agigantados. Según estadísticas de NetCraft[1] la cantidad de sitios que había en la Web no superaba las 100,000 en 1995; en 2013 el número de sitios en línea es de casi 1,000,000,000 [19].

Ésta fue pensada para ser entendida por los seres humanos, por lo que casi toda la información que contiene no puede ser ni entendida ni procesada por las máquinas. Esto trae como consecuencia problemas en la búsqueda, organización y mantenimiento de las páginas que aloja. Los problemas de gestión del conocimiento en ella están muy relacionados con el tamaño que tiene; mientras más cantidad de páginas existan, las búsquedas se hacen más lentas, lo mismo ocurre para su organización y mantenimiento [1, 3, 4].

Tener búsquedas y organización ineficientes, y contenido mal gestionado, provoca que cada día que pasa se suba información redundante en la Web, sobrecargándola con contenidos que ya existen y siguiendo con el círculo vicioso del deterioro en la eficiencia de la gestión del conocimiento en la Web. Las búsquedas ineficientes generan una gran cantidad de transacciones, lo que lleva a explotar la ya bastante saturada red mundial provocando enormes gastos en mantenimiento y obligando a buscar nuevas soluciones sobre cómo mejorar su infraestructura. Además de no poder analizar el contenido de las páginas, hay muchos problemas para transmitir el conocimiento entre ellas, ya que los diferentes vocabularios y formatos que se manejan en la red no pueden ser entendidos por todos los sistemas de gestión del conocimiento desarrollados en diferentes instituciones educativas y empresas [4].

Conociendo los problemas que presenta la Web, ¿qué situaciones deberían darse para que el panorama cambie, y el conocimiento que se encuentre en ella pueda ser difundido fácilmente? La respuesta sería básicamente que las máquinas puedan comprender los recursos (páginas) que se encuentran en la Web, ser capaces de procesarlos y analizarlos para trabajar mejor en las búsquedas y clasificación en base al contenido de las páginas [1].

[1] [url: http://www.netcraft.com/, Último acceso: 11 Noviembre 2013]

Una forma de hacer que las computadoras puedan entender las páginas es dotarlas de metadatos debidamente estructurados que cuenten con información coherente de los conceptos más importantes del contenido del documento según su dominio. Los metadatos son información acerca del contenido de un documento, que facilitan su procesamiento por agentes de software [5, 11, 13].

Uno de los recursos capaces de dotar a las páginas de esa información enriquecida son las herramientas de anotación semántica que hacen uso de ontologías. El objetivo de su uso es anotar metadatos en las páginas y documentos según la información que se encuentren en estos, usando una ontología del respectivo dominio de la información de los documentos en cuestión. El uso de ontologías nos permite unificar un solo concepto en diversas representaciones heterogéneas. El análisis se puede realizar de una palabra o una frase estableciendo una relación entre el contenido principal de la página y los elementos existentes en la ontología [15, 16, 17].

El presente proyecto propone una alternativa de solución al problema de la Web actual (sintáctica), pero estableciendo como dominio los documentos pertenecientes al campo de la Ingeniería Informática, y teniendo como campo de acción los documentos producidos por profesores y alumnos de la Pontificia Universidad Católica del Perú. Actualmente, la mayoría de estos documentos están alojados en bases de datos de la universidad, pero también en repositorios en la nube como dropbox, google drive, y otros lugares de almacenamiento. Esto trae como consecuencia que la búsqueda y organización de estos documentos sea una tarea prácticamente imposible.

La alternativa de solución propuesta es una herramienta de anotación semántica automática que utiliza una ontología cuyo dominio sea Ingeniería Informática. La herramienta será utilizada para anotar semánticamente documentos producidos en la universidad que se encuentren en el dominio de la ontología. Estas anotaciones permitirán que otras herramientas de búsqueda y gestión de la información fomenten su uso entre la comunidad universitaria.

1.2 Objetivo general

Desarrollar una herramienta que, usando una ontología en el dominio de la Ingeniería Informática, permita la anotación semántica automática de documentos como soporte a la recuperación de conocimiento.

1.3 Objetivos específicos

Objetivo 1.- Proponer una estructura modular para el desarrollo de una herramienta de anotación semántica automatizada.

Objetivo 2.- Soportar el procesamiento de la información textual de documentos.

Objetivo 3.- Permitir la representación del contenido de diversos documentos cuya información se encuentra en una ontología del campo de la Ingeniería Informática.

Objetivo 4.- Posibilitar la desambiguación de términos de un documento del campo de las Ciencias de la Computación usando una ontología del campo de la Ingeniería Informática.

Objetivo 5.- Soportar la persistencia de anotaciones semánticas en documentos.

1.4 Resultados alcanzados

Resultado 1 para el objetivo 1: Modelo de diseño de una herramienta de anotación semántica automatizada.

Resultado 2 para el objetivo 2: Mecanismo de procesamiento de lenguaje natural que permita obtener los diversos términos que se encuentran en un documento.

Resultado 3 para el objetivo 2: Mecanismo de procesamiento de lenguaje natural de obtención de la forma canónica de un concepto determinado.

Resultado 4 para el objetivo 3: Modelo de Ontología cuyo contenido sea el de los cursos de Ingeniería Informática en la Pontificia Universidad Católica del Perú.

Resultado 5 para el objetivo 4: Mecanismo de desambiguación de términos en un documento del campo de las ciencias de la computación.

Resultado 6 para el objetivo 5: Formato de anotaciones semánticas en documentos para ser alojadas en un repositorio.
Resultado 7 para el objetivo 5: Mecanismo de persistencia de anotaciones semánticas en una base de datos relacional.

1.5 Herramientas, métodos, metodologías y procedimientos

1.5.1 Introducción

El siguiente proyecto pertenece al campo de las ciencias de la computación. Su objetivo principal es el de facilitar la extracción de conocimiento en la Web. La esencia del proyecto es la de contribuir a la consecución de la Web semántica, posibilitando que el contenido en la Web pueda ser procesable por diversas herramientas de búsqueda de información. A continuación, se presentarán las herramientas que se utilizan para materializar los resultados del proyecto.

Resultados Esperados	Herramientas a usarse
RE1: Modelo de diseño de una herramienta de anotación semántica automatizada.	CommonKADS
RE2: Mecanismo de procesamiento de lenguaje natural que permita obtener los diversos términos que se encuentran en un documento.	Freeling, Apache Tika, Lenguaje de programación Java
RE3: Mecanismo procesamiento de lenguaje natural de obtención de la forma canónica de un concepto determinado.	Freeling, Lenguaje de programación Java
RE4: Modelo de Ontología cuyo contenido sea el de los cursos Fundamentos de Programación, Lenguaje de Programación I y Sistemas Operativos de la especialidad de Ingeniería Informática en la Pontificia Universidad Católica del Perú.	Protegé, RDF, OWL
RE5: Mecanismo de desambiguación de términos en un documento del campo de las ciencias de la computación.	Lenguaje de programación Java, API Jena Ontology, RDF , SPARQL
RE6: Formato de anotaciones semánticas en documentos alojadas en un repositorio digital.	CommonKADS
RE7: Mecanismo de persistencia de anotaciones semánticas en una base de datos relacional.	Lenguaje de programación Java, MySQL

Tabla 1: Resultados y Herramientas

1.5.2 Herramientas

Las herramientas principales para obtener los resultados del proyecto son aquellas que nos ayudan a trabajar con una ontología y poder representar el conocimiento de los documentos que se anoten, en este caso los del campo de las ciencias de la computación. También las que nos permitan definir una estructura de datos fácil de procesar que contenga todas las anotaciones realizadas en los documentos analizados.

Freeling

Freeling es una librería open-source de procesamiento de múltiples lenguajes, que provee un amplio rango de funcionalidades de análisis del lenguaje natural. Esta librería fue desarrollada por un grupo de investigación de la Universidad Politécnica de Catalunya. En la actualidad, cuenta con soporte para más de ocho idiomas, como el español y el inglés con muy buenos resultados [32]. Desde un punto de vista más técnico, la herramienta está orientada a integrarse con servicios de análisis de lenguaje en aplicaciones de más alto nivel. Freeling presta diversos servicios de análisis del lenguaje, entre los que destacan el análisis morfológico de texto, la etiquetación de palabras y la identificación de sus bases morfológicas.

La librería está desarrollada en el lenguaje C++, pero cuenta con extensiones para Java y Python. Para este proyecto se usará la API Java que hace uso de los recursos de Freeling en C++ [32].

Apache Tika

Esta herramienta es un proyecto lanzado por la comunidad apache con el fin de brindar una solución de extracción de contenido y metadatos de documentos en distintos formatos. Para ello, Tika hace uso de diferentes librerías ya existentes que soportan la extracción de texto en múltiples formatos como HTML, PDF, ODT, DOC, XLS, etc. [2]

Protégé

La herramienta open-source Protégé fue desarrollada por la Universidad de Stanford. Esta plataforma libre provee un conjunto de herramientas para construir modelos de dominio y aplicaciones basadas en conocimiento por medio de ontologías. Protégé está desarrollada en Java y cuenta con una interfaz de usuario bastante amigable además de poder ampliar sus funcionalidades haciendo uso de extensiones

[2] Información extraída de: http://wiki.apache.org/tika/

desarrolladas en Java para otras tareas relacionadas con la gestión de conocimiento [31, 33].

Apache Jena

Jena es un framework open-source hecho en java utilizado en el desarrollo de aplicaciones de Web Semántica y de información enlazada. Éste es compatible con el lenguaje OWL y con otras herramientas bastante utilizadas en el campo (como RDF). Cuenta con un API de manejo de archivos RDF, además de un motor compatible con el lenguaje de consultas SPARQL [28].

RDF

EL marco de descripción de recursos RDF (Resource Description Framework) es un modelo de datos para la representación de información sobre los recursos en la World Wide Web. Con estas herramientas podemos representar los conceptos contenidos en la ontología previamente desarrollada. Este marco trabaja su contenido en elementos llamados triples, su nombre responde a que sus expresiones están conformadas por un sujeto, un predicado y un objeto. [33, 34, 27].

Lenguaje OWL

El lenguaje OWL (Ontology Web Language) es un lenguaje de marcado usado para publicar y compartir datos en la Web usando ontologías. Los archivos OWL están codificados en el lenguaje de marcado XML y en su mayoría tienen su contenido siguiendo los patrones de expresión de RDF. Su nombre fue propuesto como acrónimo de One World Language, lo que se entiende como único lenguaje mundial. Este significado está relacionado con la Web semántica y su objetivo de fomentar y unificar el conocimiento en el mundo. [31]

SPARQL

SPARQL es un lenguaje de consulta desarrollado principalmente para hacer consultas en grafos RDF. Fue publicado el año 2004 por el RDF Data Access Working Group, así mismo desde el año 2006 es el lenguaje recomendado por la W3C para la ejecución de consultas en RDF. El funcionamiento de esta herramienta consiste en la extracción de información de una o más fuentes por medio de la coincidencia de patrones entre grafos. Una consulta en SPARQL consiste de tres partes: la primera vendría a ser la coincidencia de patrones, la segunda son los modificadores de solución los cuales permiten modificar los resultados obtenidos y la tercera que sería el output de toda la consulta. [34, 35].

1.5.3 Métodos y Procedimientos

Para obtener la información respectiva del dominio del conocimiento que se va a modelar se recurrirá a material tanto de la universidad como de repositorios digitales debido a que el campo de las ciencias de la computación está directamente relacionado con la naturaleza del proyecto.
Los resultados de los análisis de distintas fuentes de conocimiento procesadas se contrastarán con análisis realizados por personas expertas en el tema. Al comparar ambos resultados se podrá medir la efectividad tanto de las herramientas usadas como de la metodología elegida.

1.5.4 Metodologías

Siendo el proyecto a desarrollar perteneciente al campo de la Ingeniería del Conocimiento se seguirán algunos principios de la metodología CommonKADS. Esta metodología es producto de múltiples investigaciones internacionales y proyectos de aplicaciones de la Ingeniería del Conocimiento desde 1983. Históricamente estos proyectos se desarrollaban en su mayoría bajo la premisa de prueba y error, pero con el paso del tiempo las técnicas y prácticas de esta metodología fueron adoptadas en su mayoría por la mayoría de organizaciones y desarrolladores en el campo de la Ingeniería del Conocimiento [27].

De todas las secciones de la metodología, se piensa seguir algunas que se consideran más importantes para el desarrollo de este proyecto:

- Sección 1: El valor del conocimiento, esta sección brinda un acercamiento metodológico a la ingeniería y administración del conocimiento.

- Sección 2: Bases de la Ingeniería del Conocimiento, esta sección es necesaria para una metodología en la Ingeniería del Conocimiento y contiene los principios fundamentales de CommonKADS.

- Sección 3: Contexto organizacional, el cual nos permite entender el impacto que traen consigo sistemas y soluciones apoyadas en el conocimiento.

- Sección 4: Administración del conocimiento, nos ilustra sobre cómo administrar el conocimiento en la Ingeniería del Conocimiento.

1.5.5 Alcance

En esta sección del documento se describirá el alcance del proyecto, mencionando las limitaciones y los riesgos que se podrán materializar a lo largo de su desarrollo.
El proyecto a realizar es un desarrollo tecnológico que busca realizar anotaciones en documentos en la Web con el fin de que puedan ser procesados y localizados en función a su contenido y no simplemente a las etiquetas o títulos que tenga. Para lograr esto, se necesita modelar el conocimiento que poseen estos documentos, pero al ser este demasiado amplio, se decidió solo considerar el dominio de las ciencias de la computación. Puesto que además de limitar por mucho el campo del conocimiento a modelar, se cuenta ya con conocimiento previo de algunos de sus conceptos al estar íntimamente relacionado con la carrera.

1.5.6 Limitaciones

Algunas limitaciones consideradas en el desarrollo del proyecto:

- La utilización de herramientas NLP estándar puede conllevar cierto porcentaje de sesgo (por debajo del 10%) en el procesamiento del contenido de los documentos.
- El desarrollo de ontologías en el idioma español es bastante limitado, por lo que se tendrán que seguir estándares generales recomendados para el idioma inglés o llegar a un punto medio que se acomode bien con el idioma español.
- El corpus para trabajar con la herramienta es limitado pues no se cuenta con ningún repositorio oficial de documentos académicos en la especialidad de Ingeniería Informática de la universidad.

1.5.7 Riesgos

El proyecto a realizar es un desarrollo tecnológico que busca realizar anotaciones en documentos en la Web con el fin de que puedan ser procesados y localizados en función a su contenido.

N°	Riesgo Identificado	Impacto en el proyecto	Medidas correctivas para mitigar
1	Dificultad para integrar las herramientas en el desarrollo del proyecto	ALTO	Investigar sobre la compatibilidad e integración de las herramientas usadas en proyectos anteriores de anotación semántica.
2	Reestructuración de los metadatos a utilizar en los documentos analizados	ALTO	Definir una estructura de metadatos que siga los patrones de modelos de representación de conocimiento descritos en CommonKADS
3	Problemas en el modelamiento de la ontología	MEDIO	Modelar los objetos que representen los conceptos de la ontología orientándose de proyectos pasados que hayan usado OWL y la API Jena.
4	Anotaciones resultantes no concuerden con el contenido de los documentos analizados	MEDIO	Realizar pruebas de efectividad de las anotaciones realizada de manera que se pueda hacer un seguimiento de la evolución del proyecto.

Tabla 2: Tabla de riesgos del proyecto

2 CAPÍTULO 2: MARCO DE REFERENCIA

2.1 Marco Conceptual

2.1.1 Introducción

En este capítulo se presentan y describen los conceptos más relevantes relacionados con una herramienta de anotación semántica que usa ontologías. Cada descripción agrupa las definiciones principales de cada uno de los conceptos, presentados de forma que puedan ser entendidos por cualquiera aún sin estar relacionado con el tema. Los conceptos presentados serán: Web Semántica, Anotaciones, Ontologías y Metadatos.

El objetivo de este capítulo es que el lector entienda claramente los conceptos relacionados con el proyecto. Términos como anotación y ontología tienen connotaciones distintas dependiendo del campo donde se estén desarrollando; en este caso los conceptos que se emplearan son los que se usan en el campo de las Ciencias de la Computación, específicamente en la Ingeniería del Conocimiento. Conocerlos es de suma importancia para comprender la utilidad que se le puede dar a una herramienta de anotación. El punto de partida para entender la motivación de este proyecto y entender la utilidad de anotaciones, ontologías y metadatos parte del concepto de Web semántica, es por ello que el presente marco se explayará primero sobre ella dándonos a conocer su significado e importancia en el manejo del conocimiento; esto también hará más sencillo entender por qué se utilizan anotaciones semánticas y ontologías para desarrollar la Web semántica haciendo uso del contenido que está presente en la Web actual.

2.1.2 Web Semántica

La mayor parte del contenido que se encuentra en la Web, se hizo pensando en que solo tuviera que entenderlo un humano, mas no una máquina. Esto es una dificultad al momento de intentar buscar información en la red ya que muchas veces el resultado de nuestras búsquedas termina en respuestas que no son las que realmente buscamos. Aquí entra a tallar la Web Semántica, ésta añade una estructura al contenido relevante de cada página, haciendo que los agentes computacionales puedan trabajar con ellas ya sea en búsqueda de información, indexación,

clasificación, consolidación, etc. Actualmente a lo que llamamos Web vendría a ser la Web sintáctica, la cual realiza búsquedas simplemente encontrando coincidencias de palabras o frases que le hayamos indicado.
La Web Semántica no tiene la misma naturaleza, sino más bien es una extensión de la sintáctica, que aporta un soporte semántico al contenido de las páginas que permite tanto a las personas como a las computadoras trabajar en conjunto con la información de la Web [1].

El añadido que se realiza a las páginas que se encuentran en la Web actual está conformado por metadatos o meta-información que permitirá que las máquinas puedan entender y procesar como si lo hiciera un ser humano, esto permitirá brindar mejores servicios en el manejo de los conocimientos haciendo que la Web deje de ser un simple repositorio en cual buscar información útil sea una tarea sumamente ardua [2,3].

Para lograr este soporte su estructura es más compleja que la de la Web actual, como se muestra en la Ilustración 1, posee varias capas donde destacan las de reglas, ontología que es donde se encuentra modelado el conocimiento que se necesita para procesar la información, RDF y XML para representarla en metadatos y una URI la cual es un identificador que busca relacionar toda la información relacionada al documento procesado en un solo enlace.

Actualmente podemos observar el manejo del conocimiento como lo requiere la Web semántica en intranets corporativos y sistemas de información de grandes trasnacionales, para estas organizaciones la información es uno de sus activos más importantes [3]. El gasto que estas sociedades hacen en gestionar su información es sumamente alto pero es una inversión que les trae muchos beneficios, porque hoy en día no hay beneficio más grande que tener la información necesaria a la mano y mejor si la tienes antes que tu competencia. Un buen ejemplo de Web Semántica aplicado a negocios sería Amazon el cual trata de proveer ofertas a sus clientes en base a inferencias que obtiene como resultado del análisis de los productos que suele comprar o suele seguir; el impacto de una aplicación de este tipo en su público se traduce en grandes ganancias para el gigante de compras por internet.

La Web Semántica va más allá de eso, su objetivo es que el conocimiento pueda llegar a todos utilizando de la mejor forma posible los recursos informáticos que utilicemos; el economizar la Web y llenarla de información útil sin caer en la

redundancia es sin duda una meta que requerirá mucho trabajo y compromiso de toda la comunidad pero los beneficios que nos traería son incalculables [3,4].

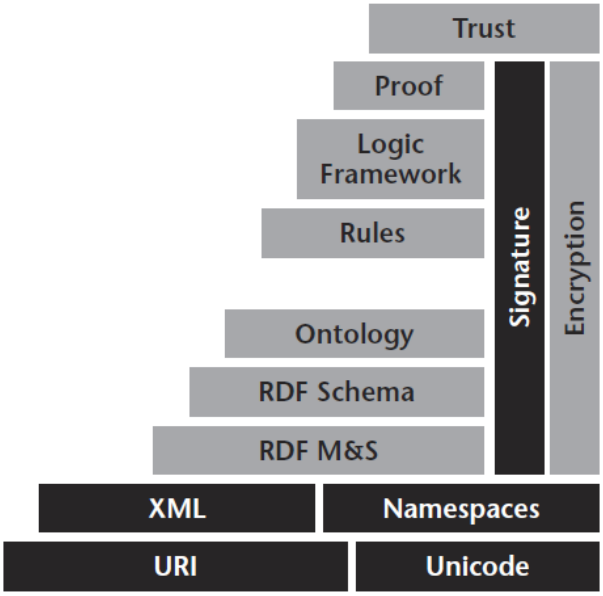


Ilustración 1: Representación de capas de la Web semántica. Imagen recuperada de M.C. Daconta, L. J. Obrst y K. T. Smith (2003)

2.1.3 Ontología

Para comprender lo que son las ontologías lo más apropiado seria revisar la definición que hizo sobre ellas el científico Thomas R. Gruber quien fue el primero en desarrollar trabajos sobre ontologías en el campo de la inteligencia artificial. Según Gruber para poder definir qué es una ontología, primero se debe entender el concepto de conceptualización. Conceptualización se puede definir como una representación abstracta de algún mundo que queramos representar, es decir, representa los objetos o conceptos existentes en ciertas áreas y las relaciones que existen entre ellos. Por lo tanto una ontología vendría a ser una especificación explícita de una conceptualización, es decir, llevada al ámbito formal. Gruber menciona también que la existencia de una ontología implica el establecimiento de axiomas que restringen la definición y el buen uso que se le dé a estos. Además de ser esta independiente al nivel de conocimiento que tenga los agentes, es decir, define un vocabulario que se puede usar para hacer aseveraciones y preguntas entre agentes [5].

Desde el punto de vista de la filosofía, la ontología es el estudio de las cosas que existen. Para la inteligencia artificial, se refiere a un vocabulario especializado para cierto dominio del conocimiento. Más allá del vocabulario es, en esencia, las conceptualizaciones que este define. Se podría cambiar de idioma la ontología sin afectar la conceptualización. Identificar el vocabulario y las conceptualizaciones requiere un análisis exhaustivo de los tipos de objetos y relaciones de un dominio [4,6].

El poder manejar una definición clara sin importar el vocabulario o quien la esté usando es uno de los motivos por los cuales las ontologías se están volviendo muy populares en la gestión del conocimiento. En el ámbito de la tecnología, las ontologías ofrecen una manera de hacer frente a las representaciones heterogéneas del contenido en la Web. En otras palabras, nos ayudan a relacionar conceptos con representaciones distintas pero con significados iguales en un determinado contexto [2]. Para la Web semántica, las ontologías tienen importancia en la medida que apoyan la búsqueda de información al delimitar los dominios de búsqueda y poder llegar a las fuentes que son realmente útiles para la consulta que se esté realizando. También ayudan en la reutilización y clasificación de la información al poder manejar de forma más clara los conceptos sin importar las fuentes o los agentes de donde vengan.

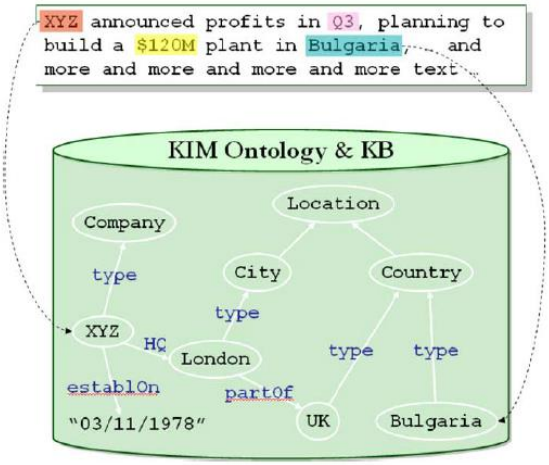


Ilustración 2: Una ontología y su relación con el contenido de un documento. Imagen recuperada de A. Kiryakov (2004)

2.1.4 Anotación

Cuando se habla de anotaciones, lo más común es referirse a poner notas en un libro, marcar una hoja con algo que necesitemos recordar. De forma más formal se puede decir que estas son un recurso de información que puede estar plasmada en comentarios, notas, explicaciones, que se añaden a un documento o a una parte de un documento. Pueden ser consideradas de tipo externas si no modifican el documento o internas si lo hacen. Conceptualmente las anotaciones son consideradas como metadatos los cuales nos brindan información sobre una porción de datos [8]. Las personas en el ámbito académico han venido usando las anotaciones en libros, papers, revistas, etc. con diversos objetivos como marcar información que requiera de su atención para una futura revisión, marcar secciones donde se necesiten referencias adicionales para comprender su contenido, resaltar lo más importante de un texto, anotar alguna idea en relación con lo leído, etc. El anotar lo que se necesite permitirá más adelante poder sacar más provecho al documento leído ya sea para analizarlo o para realizar una revisión rápida sobre él [9, 10, 11]. Las anotaciones también se pueden realizar en el campo de la computación, recibiendo el nombre de anotaciones digitales. Estas anotaciones son hechas por los usuarios en la Web y pueden contener información variada como opiniones personales, comentarios o impresiones de la página visitada así como metadatos que tengan información más formal de la página consultada. Estas anotaciones se pueden contrastar entre los usuarios, con ello la capacidad para juzgar el contenido de las páginas puede aumentar mucho así como la participación de los usuarios en la creación o modificación de las anotaciones.

Estas anotaciones pueden ser utilizadas para gestionar el contenido que se encuentra en las páginas Web, aunque no todas las anotaciones son útiles, para esto se necesita que tengan un nivel de formalidad. Siguiendo este criterio podemos clasificar las anotaciones en anotaciones formales e informales. Las formales son anotaciones que tienen un nivel de formalidad que permite asegurar su interoperabilidad entre diversos agentes. Teóricamente estas anotaciones son más aptas para ser interpretadas de la misma forma por diferentes mecanismos de consulta; un ejemplo de este tipo de anotación serían los metadatos, específicamente la que sigue estándares en su estructura y tiene asignada valores que usa nombres convencionales autorizados. Por otro lado las anotaciones informales vendrían a ser las notas o anotaciones que escribimos en un libro o artículo mientras lo vamos leyendo; estas notas pueden tener diversas utilidades como recordatorios, citas, críticas, etc. [13].

2.1.5 Metadatos

Los metadatos son información acerca de la información, parte de una información secundaria que se refiere a una información primaria estando separadas. Algunos ejemplos de metadatos son: esquemas, restricciones de integridad, comentarios acerca de los datos, ontologías, parámetros de calidad, comentarios, anotaciones, fuentes y políticas de seguridad [14].

En el manejo de la información los metadatos resultan muy útiles para clarificar su significado así como para prevenir sus malinterpretaciones y favorecer su manejo y extracción. Otro aspecto que favorece sumamente su uso es que pueden añadirse a una gran variedad de documentos en la Web, en nuestras computadoras, en libros físicos, etc., incluso aplicaciones que estemos ejecutando. Esta también puede expresarse en una gran cantidad de lenguajes y vocabularios como también estar disponible tanto en formato físico como electrónico [14, 15, 16].

También pueden ser utilizados como anotaciones digitales, ofreciendo grandes ventajas en la gestión del conocimiento en la Web como proveer formalización al contenido de los documentos anotados para facilitar su búsqueda y clasificación, también destacar que los metadatos son herramientas muy flexibles que pueden ser entendidas fácilmente por los humanos y también por las máquinas [17, 18]. Esta flexibilidad y sencillez en su rendimiento favorece el uso de la anotación con metadatos usando ontologías. Las ontologías como ya se describió antes nos ayudan a intercambiar el conocimiento basado en la suposición de que hay una sola realidad y la distribución de este conocimiento es una cuestión del alineamiento de las diferentes formas de pensar tanto de las personas o sistemas respecto a ello. Usados en conjunto son especialmente útiles en la anotación semántica porque son fáciles de entender, sencillos de construir y mantener, y es sencillo llegar a un consenso respecto a la información que nos ofrecen [16, 17].

Enfocándonos en el valor que tienen los metadatos usando ontologías en la Web Semántica, estos nos ayudan a entender mejor las entidades que existen en un documento en base al dominio al cual pertenece [15].

2.1.6 Procesamiento de Lenguaje Natural

El procesamiento de lenguaje natural (NLP, siglas en inglés) nace como la gran promesa de lograr que las máquinas pudieran comunicarse con las personas de forma sencilla. El objetivo de esta rama de las ciencias de la computación es el que las personas logren hablar en su propia lengua con las máquinas y así evitar que se tengan que aprender complejos lenguajes de computadora para cada orden que se les quisiera dar [29, 30].

La aplicación del NLP está incluida en muchos estudios como la traducción hecha por maquinas, el procesamiento y resumen textos en lenguaje natural, interfaces de usuario, extracción de información, inteligencia artificial, sistemas expertos entre otros. Pero trasladándonos al presente una de las áreas que más importancia está cobrando su uso es la proliferación de la Web y los repositorios digitales, diversas investigaciones apuntan a la necesidad facilitar la extracción de información en múltiples lenguajes para explotar a fondo los beneficios que trae la Web y los repositorios digitales. En el presente proyecto se utilizará especialmente las técnicas de separación de oraciones para identificar las oraciones en un texto, la técnica de tokenización para separar las oraciones en tokens y la técnica de etiquetado sintáctico para identificar la función de cada uno de las palabras que componen las oraciones reconocidas [31].

2.1.7 Conclusión

Repasando los conceptos se puede tener una idea de en qué consiste la Web Semántica y cuál es su objetivo. El desarrollo de una herramienta de anotación semántica que use ontologías supone un gran apoyo en la tarea de transformar la Web actual como la conocemos en una Web que pueda ser comprendida tanto por humanos como por máquinas, una donde la búsqueda de la información no sea ningún obstáculo.

Conocer estos conceptos nos permite comprender la naturaleza de la herramienta que propone este proyecto de fin de carrera, las anotaciones, los metadatos y las ontologías en conjunto pueden llevar el manejo del conocimiento en la Web a otro nivel, logrando beneficios que en la actualidad no podríamos calcular.

2.2 Estado del arte

2.2.1 Introducción

La anotación semántica como herramienta para enriquecer los documentos que se encuentran en la Web lleva ya casi dos décadas de propuesta. Pero no es hasta 2001 que Tim Berners-Lee, quien es considerado el padre de la World Wide Web, publica su artículo “The Semantic Web” [1], que realmente se toma conciencia de que la Web actual no funciona como lo hubieran deseado sus creadores. La Web actual no es capaz de transmitir el conocimiento que contiene a todos sus usuarios, muchas veces por dificultad en las búsquedas, otras porque los conceptos están plasmados en instancias heterogéneas y aunque puedan tener el mismo significado las máquinas no son capaces de comprenderlos.

Berners-Lee propone a la Web semántica como una extensión de la Web que conocemos, una donde el conocimiento pueda ser fácilmente trasmitido y todos manejemos los mismos conceptos sin importar que sistemas, lenguajes o métodos usemos. Es aquí donde son importantes las herramientas de anotación semántica, estas pueden proveer a los documentos en la Web de los metadatos y estructuras necesarias para que las máquinas puedan procesarlos y facilitar la gestión de su conocimiento.

Otro problema de la Web es su gran tamaño por lo que las anotaciones de antaño que muchas veces eran hechas por herramientas manuales y semi-manuales, están prácticamente en desuso. En la actualidad la mayoría de propuestas de este tipo de herramientas son automatizadas y utilizan ontologías para tener un soporte de conocimiento.

Los objetivos de esta revisión son conocer las alternativas de solución que se han propuesto para resolver el problema de la incapacidad de las máquinas de poder procesar los documentos que se encuentran en la Web. Nos centraremos en las propuestas más actuales y en resaltar las características principales que comparten en común.

2.2.2 Método

El método usado para la realización del estado del arte fue la revisión sistemática. Según Pino [20], "*esta revisión permite identificar, evaluar, interpretar y sintetizar todas las investigaciones existentes y relevantes de un tema en particular*".

Este tipo de revisión en el contexto de la ingeniería de software fue propuesto por Barbara Kitchenham, se plantea una revisión con tres fases independientes: la primera consistía en planificar la revisión identificando las necesidades de esta y el protocolo de revisión que luego usaríamos para buscar material de investigación relacionados a los temas que estamos indagando. La segunda fase es la del desarrollo de la revisión identificando lo que se estudia, seleccionando la literatura primaria necesaria, su evaluación, extracción y síntesis. La tercera y última fase comprende la publicación de los resultados de la revisión [20].

El protocolo de revisión utilizado se compone de las siguientes partes:

a) Pregunta Norteadora.- Esta pregunta nos dará un criterio sobre lo que queremos encontrar en la revisión que realizamos.
   ¿Cómo han sido utilizadas las ontologías en el proceso automático de anotación semántica?

b) Criterio de Exclusión.- Nos indica que literatura vamos a descartar en nuestra revisión.
   Todo estudio primario que no trate sobre anotación semántica automática.

c) Llaves de búsqueda.- Son los parámetros de búsqueda que usaremos para recopilar fuentes de información. Para nuestra revisión buscaremos en cuatro repositorios de documentos de investigación conocidos en el campo de la investigación.

SCOPUS	TITLE-ABS-KEY("automatic semantic annotation" AND ontology)
ACM	(Title:"automatic semantic annotation" and Title:ontology) or (Abstract:"automatic semantic annotation" and Abstract:ontology)
IEEE	(("automatic semantic annotation") AND ontology)
ScienceDirect	TITLE-ABSTR-KEY("automatic semantic annotation") and TITLE-ABSTR-KEY(ontology)

Tabla 3: Tabla de riesgos de llaves de búsqueda

2.2.3 Aplicaciones

En esta sección se describirán algunos proyectos de anotación semántica automática que usan ontologías de diversos dominios, la mayoría de ellos fueron concebidos con la misma finalidad, la de poder brindar un soporte a la gestión del conocimiento en su respectiva área.

2.2.3.1 Plataforma KIM

La plataforma KIM provee servicios e infraestructura para la anotación semántica, indexación semántica y extracción semántica. Para tener un funcionamiento consistente, su rendimiento en la extracción de la información está basado en una ontología y una gran base del conocimiento [24]. La ontología contiene definiciones de entidades de clase, atributos y relaciones, tanto como de recursos léxicos. Las descripciones de las entidades y las relaciones permanecen en la Knowledge Base (KB) codificada en la ontología KIM, residiendo en el mismo repositorio semántico. En lo que respecta a su arquitectura la plataforma consiste de la KIM Ontology (KIMO), la Knowledge Base (KB), el KIM Server (que contiene el KIM API) y los front-ends que soporta como Internet Explorer, Mozilla Firefox, etc. La anotación semántica se realiza mediante un servidor de anotaciones que se comunica con el API de anotaciones y que en conjunto con el repositorio semántico realizan las anotaciones. La herramienta que usan para las anotaciones contenidas en sus repositorios es RDF. El desarrollador del proyecto es la empresa Ontotext AD, cuya información está disponible en la Web.

2.2.3.2 Proyecto MOLTO

La motivación principal de este proyecto es la de desarrollar una solución de traducción multilingüe en tiempo real para páginas Web, todo esto con una alta calidad; respondiendo a la necesidad de tener la información disponible para la Web en diferentes idiomas [26]. Soportar la sincronía de la traducción y actualización de documentos en la Web es una tarea sumamente complicada y realizarla con una traducción humana no es posible. El proyecto MOLTO intenta hacer esa sincronía posible, cubriendo actualmente 15 idiomas. La sincronía es posible porque se trabaja con un lenguaje de traducción más restringido pero mucho más escalable y práctico. Los programas que provee son open-source, siendo estos un API y un IDE de traducción. La creación de

la base gramática para que actúe el API se ampara en una base del conocimiento que contiene anotaciones en RDF. Por el lado de las ontologías desarrolladas estas se presentan en OWL. Su arquitectura se observa con más detalle en la Ilustración 3.

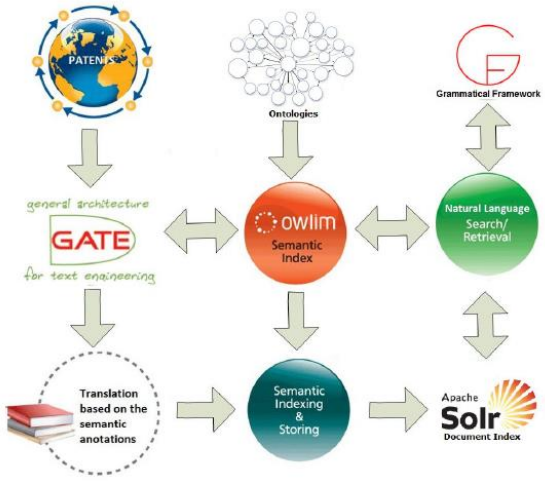


Ilustración 3: Representación la arquitectura de MOLTO. Imagen recuperada de M. Chechev (2012)

2.2.3.3 Framework de anotación semántica automatizada para artículos de Wikipedia

Este proyecto fue desarrollado por Arianna Pipitone y Roberto Pirrone del departamento de ingeniería informática de la Universidad del Estudio de Palermo. Resaltan la importancia de las wikis semánticas y como una de las más grandes wikis en el mundo como lo es la Enciclopedia Wikipedia es no semántica (aunque actualmente tiene documentos con anotaciones semánticas, la cantidad de ellos no es considerable como para llamarla wiki semántica) [23].

Las wikis semánticas son wikis que poseen modelos del conocimiento internamente en sus páginas los cuales describen su información y permiten facilitar la identificación y captura de los datos que contienen. Wikipedia está hecho con el motor Mediawiki el cual cuenta con excelentes cualidades para la gestión de la información pero aun así no estaba previsto para funcionar como una wiki semántica.

Para que la Wikipedia contara con páginas con soporte semántico, se creó la extensión Semantic MediaWiki (SMW) la cual es una extensión Open-Source del MediaWiki. El SMW enriquece la información de la wiki insertando etiquetas

de anotaciones semánticas en la página, para así no tener que modificar el texto que ya contiene. Su mecanismo de anotación está basado en el estándar de los formalismos de la Web Semántica.

El framework propuesto en este caso tiene tres enfoques:
-Analizar la estructura del documento para determinar una correcta estructura en las anotaciones.
-Análisis de los párrafos mediante reglas lingüísticas.
-Generación de anotación semántica

La arquitectura del framework consiste en mapear los conceptos principales de la página seleccionada de Wikipedia y relacionarlas con los conceptos presentes en su tabla de contenido; luego utilizando un mapa ontológico conocido como FOAM (Framework for Ontology Alingment and Mapping), extrae los conceptos respectivos de sus respectivos dominios lo que da como resultado una lista de conceptos mapeados listos para ser usados junto con el analizador de textos. El siguiente paso relacionar los conceptos según sus atributos semánticos para finalmente proceder a la anotación semántica, teniendo como salida una página de wiki semántica.
Para el desarrollo de las ontologías se usa OWL y RDF para las anotaciones.

2.2.3.4 La herramienta FLERSA

La herramienta FLERSA (Flexible Range Semantic Annotation) fue desarrollada por José Navarro-Galindo y José Samos de la universidad de Granada, está pensada para usuarios que tienen como principal objetivo realizar anotaciones semánticas en documentos en la Web.
Es una herramienta sencilla de usar en el entorno Web, pudiéndose integrar con los navegadores más populares. Sus herramientas de anotación usan estándares abiertos como XML, RDF, RDFa y OWL para promover la interoperabilidad y extensibilidad.
FLERSA usa una arquitectura cliente-servidor, la cual cuenta con un servidor Web que trabaja con la interface, una API Web y otra ontológica, servidores con la base del conocimiento y guardan las anotaciones tanto por el lado del cliente como del servidor. Estas se guardan en el formato RDF. La anotación automática se basa en las palabras más frecuentes de los documentos

apoyándose por un modelo de vectores espaciales con pesos para cada elemento del documento [22].

2.2.3.5 Apache Stanbol

Apache Stanbol es un conjunto de herramientas usada para enriquecer documentos Web con anotaciones que brinden mayor significado a su contenido. Stanbol está construida de forma modular por un conjunto de componentes, cada componente es accesible por su propia interface Web RESTful. Sus componentes no dependen el uno del otro y pueden ser usados de forma combinada según diferentes escenarios [21].

El conjunto de herramientas cuenta con un analizador y un componente de reglas que junto con un administrador de ontologías permiten el agregar anotaciones al documento Web analizado. Trabaja cada elemento como una entidad.

Se puede trabajar con sus APIs utilizando la plataforma java.

2.2.3.6 OpenCalais

OpenCalais es la versión OpenSource del conjunto de herramientas Calais desarrollados por la empresa Thomson Reuters.

El servicio Web de OpenCalais permite crear metadata con contenido semántico para el documento que se suba a su servicio usando procesamiento de lenguaje natural y de aprendizaje máquina [25]. Este servicio Web permite trabajar con múltiples herramientas como se muestra en la Ilustración 4, para promover su uso entre los desarrolladores.

La herramienta clasifica a las entidades, hechos y eventos creando mapas enlazados a documentos o sitios relacionados con ellos.

Las ontologías en Calais están desarrolladas usando OWL y las anotaciones con RDF.

El análisis de la herramienta utiliza un sistema de puntaje para cada entidad que analiza para así encontrar el dominio y los conceptos que más se asemejen a la ontología que va a utilizar.

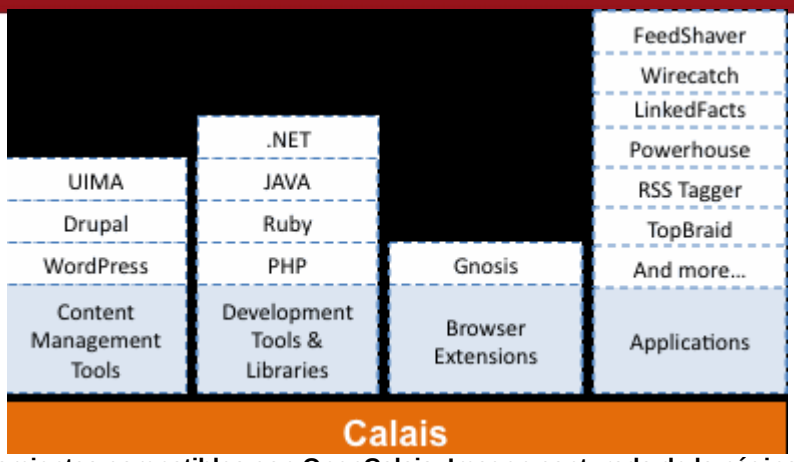


Ilustración 4: Herramientas compatibles con OpenCalais. Imagen capturada de la página oficial de OpenCalais [url= http://www.opencalais.com/about] [Último acceso: 11 Noviembre 2013]

2.2.4 Tabla de comparación

La siguiente tabla nos muestra una comparativa de las aplicaciones antes descritas, haciendo énfasis al dominio que pertenecen, las herramientas que utilizaron y que organizaciones propusieron su desarrollo.

Propuesta	Dominio	Recursos Utilizados	Tipo de organización detrás del proyecto	Extensibilidad
Plataforma KIM	Anotación semántica de documentos Web en general	RDF, OWL	Empresa del campo de gestión del conocimiento	
Proyecto MOLTO	Traducción de páginas usando anotaciones	RDF, OWL	Iniciativa Europea de programación	
Framework Anotación Semántica Wikipedia	Anotación semántica de páginas y contenido de wikipedia	Semantic Media Wiki, RDF, OWL	Proyecto universitario	SemanticMediaWiki
FLERSA	Anotación semántica de documentos Web en general	XML, RDF, RDFa, OWL	Proyecto universitario	

Apache Stanbol	Anotación semántica de documentos Web en general	OWL	Proyecto de empresa de software	Con java y algunos software de apache
Open Calais	Anotación semántica de cualquier documento Web no estructurado	OWL, RDF	Proyecto de empresa de software	Con la mayoría de lenguajes existentes y plataformas de publicación

Tabla 4: Tabla de comparación de aplicaciones de anotación semántica

2.2.5 Conclusiones sobre el estado del arte

Se puede concluir que las iniciativas para realizar este tipo de herramientas puede provenir tanto de empresas grandes de software, empresas medianas o pequeñas y personas del ámbito universitario interesadas en investigar el tema.

Conforme fueron saliendo las primeras herramientas de anotación semántica, el dominio que tenían planeando abarcar era bastante extenso. Conforme pasó el tiempo fueron optando por un dominio más limitado, dejando los proyectos más grandes a empresas con mayor respaldo.

Sobre el uso de herramientas, prácticamente todos los proyectos desarrollados los últimos 10 años usan OWL (Web Ontology Language) para desarrollar sus ontologías, las cuales son guardadas en repositorios semánticos para su respectivo uso. El que estén desarrolladas en un mismo lenguaje ayuda a que estas puedan ser reutilizadas y quien sabe en un futuro poder tener a todas las ontologías de diferentes dominios trabajando en conjunto.

También se destaca el uso del RDF (Resource Description Framework) para realizar las anotaciones; estas se suelen guardar en repositorios de anotaciones para que los buscadores los usen como apoyo para realizar las extracciones de contenido.

El tema de la extensibilidad en estas herramientas depende de quienes realizan el proyecto, las herramientas automatizadas de anotación semántica van cobrando interés pero aún no son lo suficientemente populares como para que existan muchos proyectos extensibles con otras herramientas Web. Generalmente son las empresas grandes quienes tienen frameworks o aplicaciones extensibles.

3 CAPÍTULO 3: DISEÑO DE LA HERRAMIENTA

3.1 Objetivo Especifico N°1: Proponer una estructura modular para el desarrollo de una herramienta de anotación semántica automatizada.

El diseño de una herramienta de Ingeniería del Conocimiento si bien es cierto sigue las prácticas comunes en el diseño de software, no es necesariamente parecido a cualquier sistema que busque soportar los procesos de una organización. Debido a que el proceso de negocio principal en esta rama vendría a ser la administración del conocimiento. Es por ello que se ha optado por seguir algunos lineamientos de la metodología CommonKADS, la cual es un marco de trabajo para la construcción de herramientas de Ingeniería del Conocimiento [27].

La metodología comprende 6 modelos interrelacionados que permiten capturar los principales rasgos, funcionalidades y características del proyecto de Ingeniería del Conocimiento que se va a realizar [27]. Estos modelos son:

- Modelo de Organización
- Modelo de Tareas
- Modelo de Agente
- Modelo de Conocimiento
- Modelo de Comunicación
- Modelo de Diseño

Tratándose de un proyecto de diseño el que se propone en este trabajo, se desarrollará el modelo de Diseño, pues los anteriores aunque, fueron tomados como referencia para la elección de herramientas y su tipo de utilización, responden a la fase de análisis de una aplicación de Ingeniería del Conocimiento. Lo cual requiere un trabajo exhaustivo en cuanto a los pasos que se seguirán en el desarrollo de la base del conocimiento, como también el análisis del impacto de la herramienta en la organización donde se busca aplicarlo y el análisis de cada una de las tareas a realizar, los agentes que la realizan y la forma detallada en que se comunican entre ellos [27].

El último modelo, y el que se utilizará como referencia en el proyecto, es el utilizado para definir la arquitectura y diseño de las herramientas de Ingeniería del Conocimiento. Todas las especificaciones técnicas y mecanismos computacionales involucrados en este modelo servirán para realizar la implementación del producto.

3.2 Resultado Alcanzado N°1: Modelo de diseño de una herramienta de anotación semántica automatizada.

Para cumplir con el primer objetivo de diseño, se usará como guía el proceso de construcción de un modelo de diseño según la metodología CommonKADS. Este paso según la metodología se realiza posterior al análisis de todos los aspectos relacionados con el proyecto. Se pretende cubrir esto con la investigación realizada en trabajos similares a los de la herramienta a diseñar, así como la comprensión del problema y las alternativas de solución propuestas para enfrentarlo.

Según CommonKADS, las entradas necesarias para realizar este modelo son:

- Requisitos para solucionar el problema. Estos vendrían a ser el contar con expertos o personas capacitadas para la elaboración de la base del conocimiento. También es necesario contar con todas herramientas de terceros disponibles para poder diseñar la herramienta.
- Las reglas de interacción entre los componentes, que serían las reglas que determinen la comunicación entre cada uno de los módulos de la herramienta. Esto será tomado en cuenta haciendo que cada módulo funcione independientemente del otro, solo pidiendo soporte a las APIs de terceros usadas.
- Los requisitos no funcionales de la herramienta a construir, entre los que se comprende la reusabilidad de los módulos, la escalabilidad y la adaptabilidad haciendo posible que la herramienta pueda funcionar haciendo uso de otras librerías que cumplan las mismas tareas.

Los pasos para la construcción del modelo están definidos según las siguientes plantillas que aporta CommonKADS. En total se seguirán cuatros pasos que serán los siguientes:

1. Diseño de la Arquitectura

   En este paso descompondremos la herramienta a diseñar en módulos de software, y estos vendrían a ser:

   Módulo de extracción de contenido de documentos.- el cual se encargará de extraer el contenido textual de los documentos que ingresen como parámetros de entrada a la herramienta.

Módulo de identificación de términos según su categoría gramatical.- este se encargara de procesar el contenido textual extraído por medio de técnicas de NLP para conseguir términos relevantes en el documento.

Módulo de desambiguación de términos.- el cual se encargará de buscar los términos procesados en la base del conocimiento para así encontrar su concepto más idóneo en relación al contexto del documento.

Módulo de persistencia de anotaciones.- viene a ser el último módulo de la herramienta y se encargara de guardar las anotaciones semánticas en un repositorio digital.

2. Especificación de la plataforma de implementación

Debido a que la mayoría de herramientas libres de procesamiento de lenguaje natural y gestión del conocimiento están desarrolladas o tienen versiones hechas en lenguaje Java se ha elegido la plataforma Java 1.7 para la implementación del proyecto, además de estar pensado para ser usado solo en Unix por temas de compatibilidad de la librería en C++ utilizada por la herramienta Freeling. En el caso de los archivos relacionados a la base del conocimiento, se utilizará el formato de archivos OWL con el formato RDF/XML por ser de los más comunes en el desarrollo de herramientas que trabajan con ontologías según las investigaciones realizadas en el estado del arte.

Finalmente para la persistencia se eligió un manejador de base de datos MySQL por su alto grado de compatibilidad con la plataforma Java.

3. Especificación de los componentes de la arquitectura

En este paso se detallarán las especificaciones de cada uno de los módulos de la herramienta:

-Especificaciones del módulo de extracción de contenido de documentos:

*Deberá soportar el procesamiento de documentos en formato PDF.

*Deberá soportar el ingreso de por lo menos 30 documentos por vez.

-Especificaciones del módulo de identificación de términos según su categoría gramatical:

*Deberá identificar todas las oraciones del texto procesado.

*Deberá identificar todos los tokens del texto procesado.

*Deberá etiquetar gramaticalmente cada una de las palabras en el texto.

*Deberá encontrar la forma canónica de cada término seleccionado luego de las fases anteriores.

-Especificaciones del módulo de desambiguación de términos:

*Deberá trabajar con ontologías en archivos OWL y formato RDF/XML.

-Especificaciones del Módulo de persistencia de anotaciones:

*Deberá trabajar con una base de datos relacional MySQL.

4. Detalle del diseño de la aplicación

El flujo de funcionamiento de la herramienta se podrá observar en la Ilustración 5 donde cada uno de los módulos hace uso de las herramientas o librerías propuestas para apoyar su funcionamiento. En el diagrama la entrada son los documentos en formato PDF y la salida son las anotaciones ya grabadas en la base de datos relacional que toma el rol de repositorio digital. El funcionamiento de cada una de las partes mostradas en la Ilustración se irá explicando en los resultados de cada uno de los objetivos siguientes.

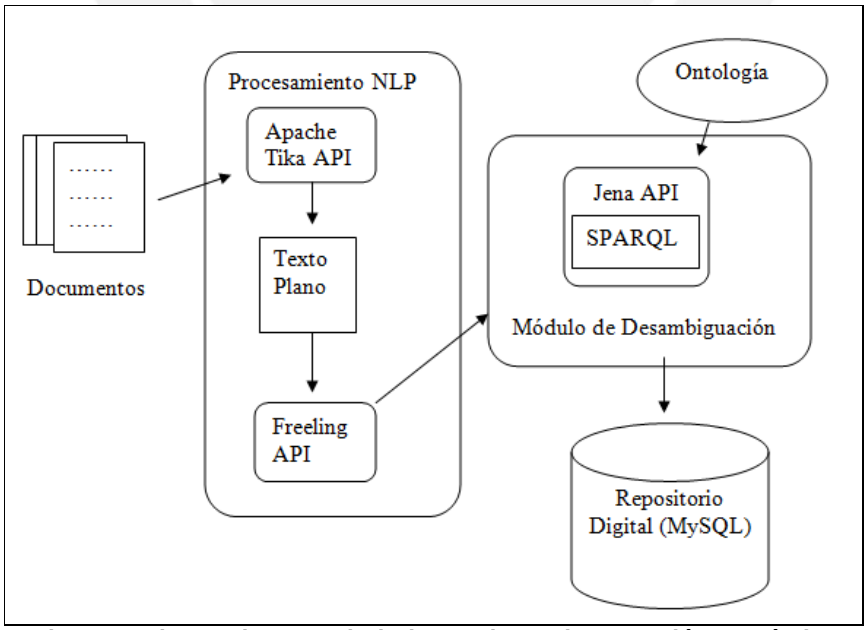


Ilustración 5: Diagrama de arquitectura de la herramienta de anotación semántica automatizada propuesta. Imagen producida por el autor.

3.3 Consideraciones Finales del resultado N°1

Se consideró solo el uso de la plantilla de diseño de CommonKADS debido a la naturaleza del proyecto que se centra en el diseño de una herramienta más que en un análisis previo de cada uno de los elementos que la componen. No obstante, el estudio de cada una de las APIs y librerías a utilizarse así como el campo específico de estudio al que pertenecen también corresponde a un análisis sobre como intentar resolver el problema que se busca solucionar con el proyecto.

Un diseño que separa claramente cada uno de los componentes de la herramienta permite que su implementación no implique el uso de las mismas librerías en caso se desee variar el tipo de solución que se brindará. Una solución modular también ayuda a que se pueda prestar atención en alguno de los módulos en específico para mejorar los resultados de la herramienta, esto quedaría para posibles trabajos futuros en caso se quiera refinar el proceso que seguirá tanto el módulo de NLP, el de desambiguación y la ontología a utilizar.

4 CAPÍTULO 4: PROCESAMIENTO TEXTUAL DE LOS DOCUMENTOS

4.1 Objetivo Específico N°2: Soportar el procesamiento de la información textual de documentos

El segundo objetivo específico de este proyecto básicamente ofrece una alternativa de soporte al procesamiento de la información textual de documentos. En este caso nos centraremos en documentos en formato PDF que contengan información sobre el campo de las ciencias de la computación que es el dominio de interés del desarrollo de este proyecto.

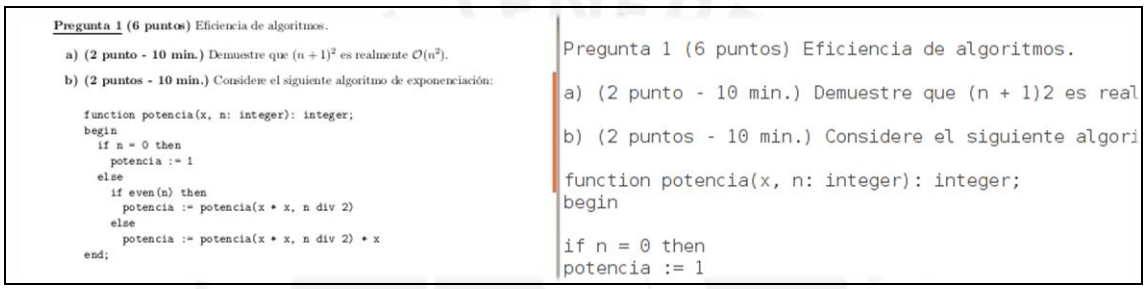
Para procesar el contenido de los documentos será necesario aplicar técnicas y herramientas de Procesamiento de Lenguaje Natural o NLP. Esta rama de las ciencias de la computación nos ayudará a que la herramienta a realizar tome como datos de entrada documentos con contenido textual y los pueda procesar. El trabajo de procesamiento no será una simple búsqueda de coincidencia de palabras, sino un análisis de cada uno de los párrafos y oraciones contenidos en los documentos, lo cual se traducirá en la ejecución de tareas dentro del NLP como son la separación de oraciones, la tokenización, la etiquetación de cada una de las partes de la oración y la reducción de las palabras encontradas a su forma más básica (Lema).

Desarrollando cada una de estas tareas es que la herramienta podrá obtener resultados que se traduzcan en el entendimiento del texto extraído de os documentos procesados, solo así se podrá realizar la anotación semántica de estos, puesto que para realizar la anotación la computadora debe comprender el contenido del recurso donde está anotando.

4.2 Resultado Alcanzado N°2: Mecanismo de procesamiento de lenguaje natural que permita obtener los diversos conceptos que se encuentran en un documento.

El primer resultado alcanzado de este objetivo consiste en usar técnicas de procesamiento de lenguaje natural para obtener los términos que se encuentran en los documentos a procesar. En este caso serían los documentos del campo de la ciencias de la computación que se van a procesar.

Como se subrayó en la descripción del objetivo N°2, el formato de los documentos a procesar será PDF por ser uno de los formatos más populares para documentos en la Web. El primer paso para comenzar su procesamiento es convertir todo el contenido textual del documento a un conjunto de caracteres que pueda ser manejado por la computadora. Esta conversión se realizará haciendo uso del proyecto Apache Tika, el cual está compuesto por un conjunto de herramientas que convierte documentos de diversos formatos como .doc, .pdf, .rtf, etc, en texto plano como se observa en la Ilustración 6. Luego de este pre procesamiento de los documentos es que se puede pasar su contenido como parámetros de entradas para las múltiples herramientas de procesamiento de lenguaje natural que existen.



Pregunta 1 (6 puntos) Eficiencia de algoritmos.

a) (2 punto - 10 min.) Demuestre que $(n+1)^2$ es realmente $\mathcal{O}(n^2)$.

b) (2 puntos - 10 min.) Considere el siguiente algoritmo de exponenciación:

```
function potencia(x, n: integer): integer;
begin
  if n = 0 then
    potencia := 1
  else
    if even(n) then
      potencia := potencia(x * x, n div 2)
    else
      potencia := potencia(x * x, n div 2) * x
end;
```

```
Pregunta 1 (6 puntos) Eficiencia de algoritmos.

a) (2 punto - 10 min.) Demuestre que (n + 1)2 es real

b) (2 puntos - 10 min.) Considere el siguiente algori

function potencia(x, n: integer): integer;
begin

if n = 0 then
potencia := 1
```

Ilustración 6: Conversión de Documento PDF a Texto Plano. Imagen producida por el autor.

Una vez obtenido el texto plano del documento, se puede proceder a su procesamiento con cualquier herramienta de NLP. En este caso, se utilizara la API en lenguaje Java de la librería Freeling, la cual puede ejecutar múltiples tareas de procesamiento del lenguaje natural. Los componentes con los que trabaja la arquitectura de la librería se pueden observar en la Ilustración 7. La librería básicamente divide el contenido del documento en párrafos, estos son divididos en oraciones y luego en palabras. Durante este proceso se realizan las tareas de etiquetación sintáctica de oraciones, lematización de palabras y reconocimiento del idioma utilizado.

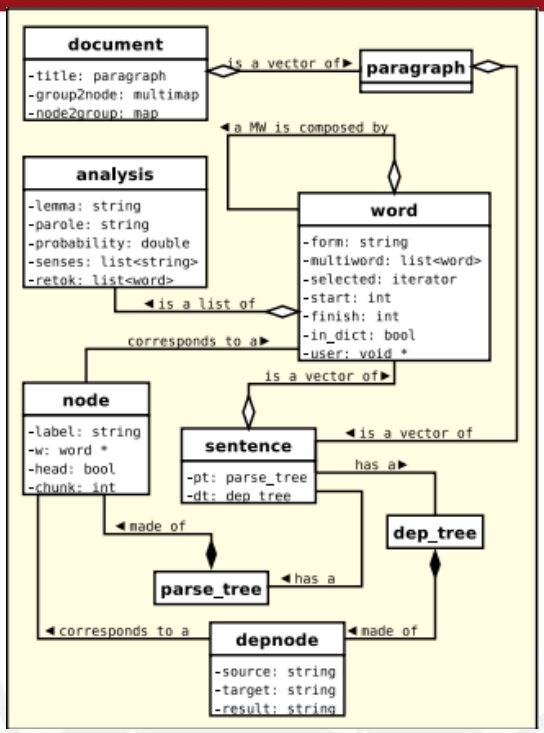


Ilustración 7: Diagrama de clases lingüistas de Freeling. Imagen recuperada de Padró, Lluís and Evgeny Stanilovsky (2012)

En el caso de los documentos que interesan al proyecto, nos centraremos en extraer todos los sustantivos y adjetivos que se encuentren en sus textos. El motivo de este filtro es el recoger la mayor cantidad de conceptos posibles en el contenido de los documentos, al margen de si estos pertenecen al dominio del conocimiento que se buscar, utilizar para la anotación semántica. Un mecanismo de esta naturaleza podría ser replicado para cualquier tipo de herramienta automatizada de anotación semántica sin importar el campo del conocimiento que se busque apoyar. Además de evitar la limitación de ceñirse a un vocabulario establecido que tenga que ser actualizado. La idea del proyecto es que la valoración de los conceptos que se encuentren en los documentos la realice la ontología pues esta es la que guarda el significado y las relaciones de estos.

Al final, la herramienta Freeling brindará como parámetros de salida una lista de términos encontrados en el documento y estos serán nuestro punto de partida para comenzar a trabajar con la ontología y el conocimiento que se encuentra en ella. En la Ilustración 8 se pueden observar algunos términos extraídos de un documento relacionado a temas de programación. El análisis se realizó a un documento del idioma español por lo que el analizador tomó 5 palabras en inglés como una sola. Si bien es cierto que el análisis que realiza la herramienta obtiene sus mejores resultados

cuando trabaja en español, el margen de error aún es existente pudiendo arrojar en ciertos casos algunos verbos o adjetivos que no conozca como si fueran nombres propios; la limitación en ese aspecto se debe al diccionario en español que utiliza la herramienta Freeling [32].

```
puntos
TAD
Enriquecer
especificaciones
puntos
min
lenguaje
BASIC
BASIC
iniciales
Beginner
s
All-purpose_Symbolic_Instruction_Code
lenguaje
programación
propósito

lenguaje
funciones
cadenas

funciones
LEFT
```

Ilustración 8: Conceptos extraídos del documento. Imagen producida por el autor.

4.3 Resultado Alcanzado N°3: Mecanismo de procesamiento del lenguaje natural para la obtención de la forma canónica de un concepto determinado

El resultado alcanzado de este objetivo consiste de menos pasos que el anterior, pero es sumamente crucial para asegurar una correcta cooperación entre el procesamiento de contenido textual del documento y el análisis semántico que se realizará de los conceptos con la ontología.

Para detallar más el problema que plantea resolver este objetivo, tomaremos como ejemplo al idioma español. En este idioma existen múltiples conjugaciones tanto para sustantivos, verbos y adjetivos. Esta es sin duda una de las características más resaltantes en los idiomas latinos y que hacen que el estudio del idioma no sea relativamente sencillo para quien lo intenta adoptar como una nueva lengua.

En el plano de la gestión del conocimiento y específicamente de las ontologías, se suelen manejar diversas representaciones para conceptos con significados idénticos o similares, por lo cual se opta por representar a esos objetos en entidades cuya representación más usada sea su forma canónica o lema. En otras palabras, la forma

más simple de representar la palabra. Ejemplos de esto sería el contar con las palabras *perrito*, *perrazo*, *perrucho*, *perrita* cuyo lema de todas ellas sería *perro*. El mapear todas esas palabras en una ontología sería una tarea sumamente engorrosa pues estamos hablando de cientos de conceptos solo en ontologías de tamaño muy reducido.

La solución que se plantea con este mecanismo es el de obtener el lema de cada uno de los conceptos recuperados en el resultado alcanzado anterior para así asegurar que se puedan encontrar la mayor cantidad de conceptos posibles en la ontología y realizar su respectivo enlace para poder comenzar con las anotaciones.

No obstante, hay conceptos o sustantivos que carecen de lema, por lo que los parámetros de entrada que se usarán con la ontología serán tanto el concepto encontrado en el documento como su lema obtenido por la herramienta Freeling. Esto asegura que la búsqueda de conceptos en la ontología se realice tanto con el lema como con las palabras sin modificar.
En la Ilustración 9 podemos observar los parámetros de salida del resultado alcanzado N°2 que serían los conceptos ya extraídos con su respectivo lema.

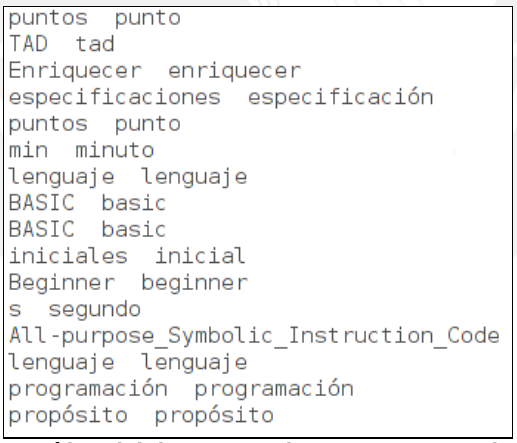


Ilustración 9: Conceptos extraídos del documento junto a su respectivo lema. Imagen producida por el autor.

4.4 Consideraciones Finales de resultados N°2 y N°3

La elección de documentos en formato PDF estuvo motivada por la gran cantidad de documentos en aquel formato producidos en la especialidad de Ingeniería Informática, además de ser uno de los formatos idóneos para la publicación de documentos científicos en el campo de las ciencias de la computación. Es debido a esto, que se

eligió como primera librería en uso al conjunto de herramientas Apache Tika. La experiencia en su uso fue sumamente notable al momento de extraer el texto plano de cada uno de los documentos que se utilizaron como datos de entrada. Además de contar con otras opciones como la extracción de metadata de los archivos la cual puede ser utilizada en análisis que impliquen como orientar el procesamiento en base a la información contenida en la metadata.

Posteriormente, la elección de la herramienta Freeling se tomó luego de experimentar con algunas herramientas de NLP del campo que tuvieran soporte para el idioma español. Su elección consideró entre otras cosas que buena parte de su equipo de desarrollo era hispanohablante, además de que publicaban resultados positivos de sus pruebas realizadas en diversos análisis del idioma español [32].

Sin embargo, Freeling no garantizó una eficiencia total al momento de procesar evaluaciones producidas por la especialidad de Ingeniería Informática de la universidad, esto debido a factores como el uso de palabras del idioma inglés en los documentos, y que la funcionalidad de reconocimiento de nombres de entidades de la herramienta era poco efectiva al momento de analizar el texto de los documentos, sobre todo cuando se trataba de entidades multi-palabra como *Fundamentos de Programación* o *Lista Adyacente*. Es por ello que se tomó la decisión de no solo rescatar del texto los sustantivos sino también los adjetivos encontrados en el documento ya que estos ayudarán al momento de desambiguar el significado de los sustantivos encontrados.

El rendimiento de la extracción de términos y nombres en general, se podría mejorar haciendo uso de una herramienta de reconocimiento de nombres en español que tenga un alto grado de eficiencia, ésta sería una buena opción de trabajo futuro que contribuiría no solo con una herramienta como la propuesta en este proyecto sino también con cualquier iniciativa o herramienta de NLP para el idioma español.

Otra limitación importante del uso de la librería Freeling es que su equipo de desarrollo no brindaba soporte para el uso de la herramienta en plataformas Windows, por lo que se tomó la decisión de desarrollar esta herramienta para plataformas Unix donde Freeling funciona perfectamente. Si bien es cierto esto supone una limitación, el carácter modular de la herramienta permite poder utilizar otra herramienta de NLP o en el mejor de los casos intentar que las librerías nativas de C++ del Freeling puedan trabajar en la plataforma Windows.

Finalmente la utilización del lema además de la palabra sin modificar corresponde a facilitar la tarea de las funcionalidades que relacionen los términos encontrados en el texto con la ontología o base de conocimiento a utilizar. Ya que en una ontología sería demasiado trabajo mapear todas las conjugaciones de cada palabra, en cambio, si se podría contar por lo menos con su forma básica o canónica.

5 CAPÍTULO 5: ONTOLOGÍA DE LA HERRAMIENTA

5.1 Objetivo Especifico N°3: Permitir la representación del contenido de diversos documentos cuya información se encuentra en una ontología del campo de la Ingeniería Informática.

Este objetivo consiste básicamente en la elaboración de la ontología que será utilizada como base del conocimiento del proyecto, y surge como alternativa de solución a los problemas con los que actualmente choca la comunidad académica y usuaria en general para poder llegar a un acuerdo sobre los conceptos que pertenecen a un determinado dominio del conocimiento. El dominio en específico serán los temas de ciencias de la computación que son impartidos en la especialidad de Ingeniería Informática de la Pontificia Universidad Católica del Perú. Específicamente de los cursos de Fundamentos de Programación, Lenguaje de Programación I y Sistemas Operativos.

Una ontología es usada para capturar el conocimiento de un dominio específico que en este caso serían los temas de los cursos ya mencionados de la especialidad de Ingeniería Informática de la universidad. Dicho conocimiento se expresara por medio de múltiples conceptos y las relaciones que existen entre estos.

5.2 Resultado Alcanzado N°4: Modelo de Ontología cuyo contenido sea el de los cursos Fundamentos de Programación, Lenguaje de Programación I y Sistemas Operativos de la especialidad de Ingeniería Informática en la Pontificia Universidad Católica del Perú.

Para la construcción de esta ontología se utilizará la herramienta open-source Protégé para la edición y gestión de ontologías. Existen diversas versiones de esta herramienta disponibles siendo algunas de estas: protegé 2000, protégé 3.1, protégé 3.4 y protégé 4.0 [33]. Además de contar con una versión Web llamada Web-Protégé, en el caso de este proyecto utilizaremos la versión de escritorio de la herramienta específicamente la versión 4.3 (Build 304).

Una vez creada una nueva Ontología con la herramienta Protégé, el punto de partida de su estructura será una clase llamada Thing la cual es la raíz de toda la ontología. Como se puede observar en la Ilustración 10, cada ontología que creemos con la herramienta sin importar el dominio que queramos representar, tendrá como inicio el nodo Thing.

Luego de creada la ontología y teniendo clara la jerarquía de los conceptos que compondrán el dominio del conocimiento que se quiera representar, se irán agregando conceptos como subclases de la clase Thing. En nuestro caso se crearan las clases Facultad, Especialidad, Programa Analítico, Unidad de Aprendizaje y Concepto, puesto que todas estas clases representan de forma bastante sencilla la jerarquía en que se organizan los cursos impartidos en la universidad. Siendo la clase de mayor nivel la Facultad y la de menor jerarquía los conceptos impartidos en cada una de los cursos dictados en ellas.

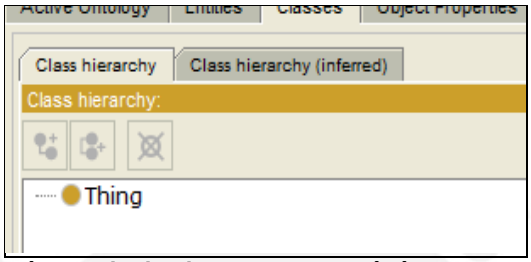


Ilustración 10: Ontología creada desde cero con Protégé. Imagen producida por el autor

Cada uno de los conceptos trasladados a la ontología heredan las propiedades de la clase Thing, y estas clases tienen relaciones determinadas que pueden ser tanto de una clase a otra como de una clase a un tipo de dato determinado. Protégé clasifica estas relaciones en Object Property cuando se refiere a una clase y Data Property cuando se refiere a un tipo de dato determinado como String, Double, etc.

En la Ilustración 11 podemos observar una estructura jerárquica preliminar de la ontología que se desarrollada en el proyecto. En esta se muestran las diferentes clases de la ontología y sus subclases. Las clases que heredan directamente de la clase Thing son clases disjuntas entre sí, osea que no pueden existir entidades que pertenezcan a ambas clases. También se encuentran subclases como Algoritmia, CienciasIngenieria, INF220, etc. Siguiendo estándares recomendados en la creación de ontologías se optó por eliminar los artículos y adverbios de cada uno de los nombres de las clases.

Una vez terminada la ontología se podrá proceder a realizar el enlace entre los conceptos mapeados en ella y los conceptos extraídos de los documentos procesados en los resultados anteriores.

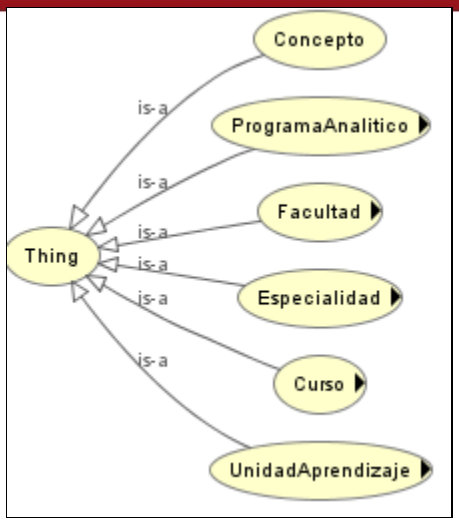


Ilustración 11: Ontología del dominio de Ingeniería Informática en la Universidad. Imagen producida por el autor.

La clase en el nivel más bajo de la ontología es la de Concepto. Esta clase está compuesta por todos los conceptos que se han mapeado en la ontología y es con ella que se realizarán la mayor parte de las tareas de procesamiento semántico del contenido de los documentos. En cuanto a los niveles de la ontología aunque todas las clases son subclases de Thing, todas tienen por lo menos una relación de pertenencia con otra clase y estas se muestran en la tabla 5.

Clase	**Relación**	**Clase**
Facultad	tieneEspecialidad	Especialidad
Especialidad	tieneCurso	Curso
Curso	tienePrograma Analitico	ProgramaAnalitico
ProgramaAnalitico	tieneUnidadAprendizaje	UnidadAprendizaje
UnidadAprendizaje	tieneConcepto	Concepto

Tabla 5: Tabla de clases y relaciones en la ontología

Las relaciones mostradas en la tabla establecen el tipo relación entre las clases que componen la ontología, al ser clases disjuntas no existen entidades que puedan ser de más de una clase al mismo tiempo así que para trabajar con ellas en conjunto se utilizarán estas relaciones de pertenencia.

5.3 Consideraciones Finales de Resultado N°4

La ontología como una de las partes más importantes de este proyecto tomó buena parte del tiempo dedicado al diseño de toda la herramienta. El alcance de solo tres cursos de la especialidad de Ingeniería Informática se justifica en que mapear la totalidad de cursos de la especialidad o más suponía un esfuerzo que escapaba de los recursos y las capacidades establecidas para el proyecto. La creación de una ontología en el común de los casos es un trabajo que suele tomar años, mediante un proceso iterativo de creación y mejora por parte de grupos expertos en el dominio a modelar. Es por ello que aunque la ontología desarrollada en este proyecto solo contempla tres materias de la especialidad; la estructura de clases que compone la ontología permite que esta se expanda hasta poder contener cualquiera de las materias impartidas en la universidad. Ya que cada una de estas se encuentra dentro de una Unidad de Aprendizaje, la cual forma parte del programa analítico de un curso, siendo cada curso parte del programa de estudios de una especialidad que compone una determinada Facultad. La clase de mayor jerarquía propuesta es la Facultad puesto que un conjunto de estas conforma lo que conocemos por universidad si tomamos en cuenta solo el ámbito académico de la organización.

La elección de Protégé como herramienta de creación para la ontología estuvo motivada por las evidencias encontradas en la revisión literaria sobre ontologías y sus posibles usos. Esta herramienta además de tener una interfaz gráfica que facilita el trabajo con las ontologías, también muestra compatibilidad con múltiples formatos capaces de contener ontologías como XML/RDF, OWL/RDF, Turtle, entre otros.

El uso de una ontología como base de conocimiento para realizar anotaciones semánticas no solo resulta conveniente al momento de identificar los conceptos encontrados en los contenidos a procesar, sino también, porque las ontologías a diferencia de algunas alternativas como las soluciones basadas en reglas son más beneficiosas en cuestiones de escalabilidad y de trabajo en conjunto. Lo cual se traduce en el objetivo principal de la herramienta que es gestionar de forma eficiente el conocimiento generado.

6 CAPÍTULO 6: DESAMBIGUACIÓN DE TÉRMINOS

6.1 Objetivo Especifico N°4: Permitir la desambiguación de términos de un documento del campo de las Ciencias de la Computación usando una ontología del campo de la Ingeniería Informática.

Este objetivo busca brindar una alternativa de solución al problema de la identificación del significado de una palabra basándose en el contexto en el que se encuentra. Existen diversas estrategias de desambiguación de términos para resolver problemas de gestión de conocimiento y de mejor de consultas de información; en el caso específico de la desambiguación de términos en un documento, las estrategias pueden estar basadas en el análisis de las oraciones o párrafos en los cuales está contenido el término, así como también en la contabilización de coincidencias de palabras clave del concepto que implica el término en el documento. Otras estrategias van más allá y determinan el significado idóneo de un término realizando búsquedas en grandes bases de conocimiento alojadas en la Web, la mayoría de estas opciones elige el mejor concepto según la mayor cantidad de uso que se le dé en esos repositorios de conocimiento [37].

Para este proyecto la alternativa de solución propuesta es una estrategia de desambiguación de términos en función al contexto del documento que contenga al término. El soporte de conocimiento para esta estrategia será la ontología obtenida en el capítulo anterior. En otras palabras, el grado de importancia de los conceptos que sean candidatos para ser relacionados con el término dependerá de la relación que estos tengan entre ellos y con el resto de conceptos existentes en el documento.

6.2 Resultado Alcanzado N°5: Mecanismo de desambiguación de términos en un documento del campo de las ciencias de la computación.

Para poder soportar la desambiguación de términos en un documento, se utilizará una estrategia de desambiguación de términos en función al contexto del documento en el que se encuentran. La estrategia en función al contexto se reflejará en la función número 1 cuyos parámetros de entrada serán el documento y el término [36,37].

$$desambiguacion(doc, term) = primer\{concepto\ e\ conc(term) \mid \max(funcMer(doc, vecin(concepto)))\}$$

Función N°1: Función de desambiguación de términos adaptada de Hotto, Staab y Stumme[36]

Glosario de la función:
$conc(term)$: está referido al conjunto de conceptos que se pueden identificar de un término
$funcMer$: está referido a la función de mérito en un conjunto de elementos
$vecin(concepto)$: está referido a la vecindad de un concepto
$primer$: El primer elemento del conjunto
doc : está referido al documento al cual pertenece al término
$term$: está referido al término del documento

El resultado de esta función es el concepto del término que guarda mayor relación con el contexto del documento, esto quiere decir que los elementos del conjunto dentro de la función serán todos los conceptos que puedan ser representados explícitamente con el término ingresado a la función, por ejemplo la palabra *cola* que podría estar relacionada a más de un concepto en la ontología. La siguiente parte de la función será aplicar el criterio de relación que guarda cada uno de los conceptos del conjunto con el contexto del documento, en otras palabras verificar a que campo de estudio en específico de las ciencias de la computación se refiere. Esta relación se evaluará calculando la frecuencia absoluta de la coincidencia de los conceptos pertenecientes a la vecindad del concepto que se evalúa en el criterio. Esta vecindad está compuesta por los conceptos más cercanos al concepto evaluado según la base del conocimiento que se esté utilizando como soporte para la desambiguación, en el caso de este proyecto la base será la ontología obtenida previamente en el objetivo N°3.

Un ejemplo del funcionamiento de esta función sería tomar parámetros de entrada de un documento cuyo contexto sea el tema *Tipo de Datos Abstractos* y en el documento se tenga que desambiguar la palabra *cadena*, la cual tiene diversas acepciones en el idioma castellano; además de utilizarse en diferentes temas de las ciencias de la computación como algoritmos, lenguajes de programación en general, etc. Mediante el uso de la ontología podremos encontrar conceptos que se puedan reflejar en el término cadena, pero es con esta función que podremos comparar todos estos conceptos respecto al contexto del documento y finalmente obtener como respuesta un concepto relacionado a cadenas de datos en el ámbito de los tipos de datos abstractos que es a lo que probablemente se refiera el término *cadena*.

En cuanto a la frecuencia, no se puede evaluar de la misma manera la coincidencia de conceptos que se encuentren en niveles lejanos al concepto evaluado, debido a que una ontología es un grafo, el grado de cercanía en los conceptos se medirá en cuanto a la distancia en aristas que exista entre ellos.

Una vez que se cuenta con una estrategia de desambiguación esta se aplicara al conjunto de términos obtenidos por el procesamiento de lenguaje natural realizado en el Objetivo N°2. El resultado de correr los algoritmos que ejecuten la función antes descrita resultará en un conjunto de conceptos, siendo estos los pertenecientes al conjunto de términos trabajados respectivamente. También puede ocurrir que no se encuentren conceptos relacionados con uno de los términos utilizados debido a que en la base de conocimiento no se llegó a ningún tipo de coincidencia con el término. Esto sucederá en el caso que los términos no tengan ningún tipo de relación con el campo de las ciencias de la computación; el otro caso que puede suceder es que en la ontología no tenga mapeada el término procesado ya que el límite de la ontología es el dominio de las ciencias de la computación en la especialidad de informática en la universidad y la ontología fue realizada en un lapso corto de tiempo con motivos académicos.

La búsqueda de conceptos en la ontología se realizará con la herramienta de gestión de ontologías Apache Jena y su motor compatible con el lenguaje de consultas SPARQL.

El motor de consultas nos permite recorrer la ontología de forma eficiente y rápida aprovechando que ésta es un grafo y no simplemente un repositorio digital de conceptos. Las consultas realizadas consistirán en la búsqueda de conceptos en cuyos atributos se encuentre el término con el que se está trabajando además de otro tipo de consultas que busquen conceptos relacionados con los conceptos anteriormente encontrados, el resultado de estas consultas vendría a conformar la vecindad de un concepto. Claro está que el peso de estas relaciones no será el mismo, como se señaló en el párrafo previo, hay conceptos que se encuentran más relacionados a otros y esto se refleja en la distancia entre aristas de un elemento a otro del grafo que vendría a ser la ontología. Según el orden de mérito que se obtenga de todos los conceptos analizados es que se decidirá cuál es el concepto correcto para el término según el contexto del documento.

El conjunto de conceptos obtenidos tendrá tres componentes básicos como se observa en la Ilustración 12 los cuales son el término que se utilizó para su obtención y las URI (Universal Resource Identifier) del concepto y de la unidad de aprendizaje a la que pertenece según se plantea en la ontología, estas direcciones serán las que figure en la base del conocimiento que de soporte a la desambiguación que en este caso sería la ontología.

Término	URI Concepto	URI Unidad Aprendizaje

Ilustración 12: Estructura de un objeto concepto. Imagen producida por el autor.

Según la ontología que se está utilizando la Unidad de Aprendizaje agrupa una determinada cantidad de conceptos relacionados con un mismo tema como Programación Dinámica, Complejidad de Algoritmos, etc.

6.3 Consideraciones Finales del resultado N°5

El proceso de desambiguación de términos según el contexto del documento procesado es una de las tareas más relevantes que realizará la herramienta. Debido a que en este módulo es donde los resultados del procesamiento de NLP y la ontología trabajan en conjunto. Todo esto se realiza mediante la función de desambiguación planteada en la función N°1 ya descrita anteriormente, la cual es una adaptación de una de las estrategias de desambiguación de palabras de la publicación *Ontologies Improve Text Document Clustering* publicada por Hotto, Staab y Stumme [36].

Los resultados de usar este algoritmo están limitados por el producto obtenido en la fase de NLP del texto de los documentos y también por la comparación sintáctica que se realiza entre los términos obtenidos del documento y los términos que representan cada uno de los conceptos de la ontología. A pesar de que al hacer uso de la ontología se adquiere un valor semántico en la identificación de conceptos, también existe un componente sintáctico que como se ha mencionado se encuentra limitado por la eficacia de la fase de NLP en la herramienta.

Para evaluar la eficacia de la función y la merma en su rendimiento causada por las dificultades para extraer términos en la etapa de procesamiento de texto. Se realizó pruebas con un corpus de 12 documentos cuyos conceptos más relevantes según el campo de estudio de la Ingeniería Informática fueron mapeados en la ontología a utilizar. Las pruebas que se realizaron fueron las de Precision y Recall, las cuales son métricas empleadas para medir el rendimiento de aplicaciones de búsqueda de información y reconocimiento de patrones. Específicamente, se entiende por Precision a la fracción de elementos recuperados que son relevantes y se conoce a Recall como

la fracción de elementos relevantes que han sido recuperados de entre todos los elementos relevantes que se encuentran en el conjunto analizado [38].

En la tabla 6 se observan los resultados de las pruebas realizadas; se entiende por relevante a todo aquel concepto cuyo significado y unidad de aprendizaje tengan una relación directa con los temas de cada documento y se entiende por recuperado todo aquel concepto que fue encontrado en los documentos.

Dato	Resultado
Conceptos recuperados relevantes	133
Conceptos recuperados no relevantes	31
Conceptos relevantes no recuperados	22
Precision	(133)/(133+31) = 0.81
Recall	(133)/(133+22) = 0.86

Tabla 6: Atributos de la tabla anotación

Los resultados finales arrojaron un valor de Precision de 0.81 y un valor de Recall de 0.86, los que según el análisis de la funcionalidades de la herramienta nos permiten concluir que a pesar de que la función de desambiguación contempla la evaluación semántica de los términos extraídos de los documentos, son las comparaciones sintácticas y la falta de una función de reconocimiento de nombres de entidades lo que disminuye la eficiencia de la función; esto se refleja en la Precision cuyo valor está más lejano a 1 debido a la considerable cantidad de conceptos recuperados que no son relevantes para el contenido general de cada uno de los documentos. Los resultados pueden mejorar aumentando el detalle de los elementos de la ontología pero la idea de tener a la ontología como base del conocimiento es que maneje términos y conceptos de forma estándar para no estar limitada por el tipo de herramienta que la utiliza como soporte. Es por eso que como trabajos futuros una mejora de las herramientas de NLP para el idioma español también ayudarían a darnos una evaluación mucho más certera de la eficiencia del algoritmo de desambiguación; que si bien es cierto arroja buenos resultados, estos pueden variar tomando ontologías mucho más extensas y documentos de múltiples campos.

7 CAPÍTULO 7: PERSISTENCIA DE LAS ANOTACIONES

7.1 Objetivo Especifico N°5: Soportar la persistencia de anotaciones semánticas en documentos.

El objetivo final que se desea alcanzar con este proyecto es el de realizar la persistencia de anotaciones semánticas luego de haber procesado el contenido de los documentos del campo de las ciencias de la computación que se introdujeron como output desde los resultados del objetivo N°2. Gestionar la información en la Web, repositorios digitales y bases del conocimiento privadas es una de las tareas más arduas de cualquier organización que haga uso intensivo del conocimiento que aloja en sus repositorios. Una alternativa de apoyo para la gestión es contar una base de datos con anotaciones semánticas de cada uno de estos documentos que permita facilitar las tareas de organización de contenido así como también mejorar las búsquedas para extracción de información. Un soporte semántico permite que los mecanismos de gestión del conocimiento se asemejen más al accionar de una persona al momento buscar información, pues ésta no solo realiza sus búsquedas en base a coincidencias entre lo que ve sino también en cuanto a lo que infiere sobre las entidades concretas que observe.

En el diseño de la herramienta propuesta se ofrece una alternativa de anotaciones en función al contexto del documento, esto permite enriquecer el contenido del documento procesado lo que facilitará su clasificación y extracción al momento de que los usuarios realicen búsquedas de documentos con ese contenido sin necesidad de que sus consultas contengan el mismo contenido en forma sintáctica; en otras palabras, las anotaciones permitirán el uso de consultas que infieran lo que el usuario está buscando haciendo que la máquina que procese la consulta realice un proceso de identificación de conceptos similar al que realizan las personas.

7.2 Resultado Alcanzado N°6: Formato de anotaciones semánticas en documentos.

Las anotaciones a realizarse en los documentos estarán estrechamente relacionadas con la base de conocimiento a utilizar en su procesamiento, la que en este caso sería la ontología. Por lo que los valores a guardarse dentro de cada anotación serán los siguientes:

- Término, correspondiente a la palabra anotada que pertenece al documento analizado.
- URI Concepto, correspondiente al identificador del concepto desambiguado que se encuentra en la ontología, un ejemplo de una URI sería el siguiente: *http://dominio#elemento* donde lo que va antes del # sería el dominio del identificador y lo que va luego el nombre de espacio.
- URLDOC, correspondiente al URL del documento procesado, ya que las anotaciones serán hechas en base a un documento específico. Las especificaciones se realizarán por documento y no por palabra debido que la herramienta busca sobre todo agilizar las consultas y gestión de los documentos en masa, permitiendo la anotación de muchos documentos sin necesidad de intervención humana en el proceso.
- URI Unidad de Aprendizaje, correspondiente al URI de la Unidad de Aprendizaje a la que pertenece el concepto.

Según las resultados que debemos obtener luego de realizadas todas las tareas previas, los componentes de cada una de las anotaciones son los mínimos requeridos para que cualquier herramienta de gestión de conocimiento o extracción de información pueda administrar los documentos procesados usando el apoyo de la ontología con la que se realizaron las anotaciones. La URI no solo brinda la identificación del concepto en la ontología sino también la identificación de la ontología que correspondería al dominio del identificador.

7.3 Resultado Alcanzado N°7: Mecanismo de persistencia de anotaciones semánticas en una base de datos relacional.

La persistencia de las anotaciones se realizará por medio de un JavaBean cuyas partes serían los componentes del formato de anotaciones anteriormente propuesto. Estos objetos serán guardados en una base de datos relacional MySQL por medio de una clase en hecha en Java con métodos que realicen inserciones en la base de datos haciendo uso de las librerías controladoras que permiten la comunicación entre la máquina virtual de Java y el manejador de base de datos MySQL. La base de datos tendrá una sola tabla cuyos atributos corresponderán a los objetos dentro del Bean como se observa en la tabla 7.

Término
URI_Concepto
URL_Documento
URI_UnidadAprendizaje

Tabla 7: Atributos de la tabla anotación

Debido a que la herramienta trabaja estrechamente con una ontología determinada, solo se guardará el URI del concepto anotado pues con esta dirección se podrían acceder a todos los atributos del concepto desde la ontología además de poderse acceder a otros conceptos que sean cercanos semánticamente hablando del concepto anotado. Además se incluirá el URI de la unidad de aprendizaje a la que pertenece el concepto, puesto que hay casos donde un concepto está relacionado con más de una Unidad de Aprendizaje.

7.4 Consideraciones Finales de resultados N°6 y N°7

La fase de persistencia de las anotaciones es una de las más sencillas de plantear aunque para escoger su estructura se debe tomar en cuenta que información relevante es necesaria que este enlazada con el concepto, en este caso, además de la URI del concepto se añade la URI de la Unidad de Aprendizaje ya que esta permite notar con mayor claridad el contexto al que pertenece el término anotado.

Para prevenir además posibles casos de duplicación de anotación los campos de URI y dirección del documento serán llaves primarias de la tabla de la base de datos, con lo que se garantiza tomar en cuenta conceptos similares pertenecientes a diferentes Unidades de Aprendizaje o de contextos diferentes.

8 CAPÍTULO 8: CONCLUSIONES Y RECOMENDACIONES

En este capítulo se mencionarán las principales conclusiones rescatadas del desarrollo del proyecto, así mismo se harán recomendaciones para trabajos futuros relacionados con el proyecto, los que pueden ser tanto mejoras como alternativas para ampliar el alcance de la herramienta.

8.1 Conclusiones

- Los resultados obtenidos en la fase de NLP dependen no solo de las funcionalidades con las que se cuente para analizar los documentos, sino también del tipo de documentos que se analiza, en el corpus analizado existían formulas, ecuaciones, y muestras de lenguajes de programación los cuales supusieron una merma en los resultados y acrecentaron la dificultad para el análisis sintáctico de los términos. Es por ello que se tomó la decisión de extraer también adjetivos además de sustantivos lo que supuso una mejora en la identificación de conceptos de la ontología.

- Fue sumamente importante el uso de un corpus de evaluaciones de cursos de ciencias de la computación dictados en la universidad en la creación de la ontología, ya que estos además de ser realizados por profesionales y estudiosos del campo, están ordenados según un currículo predeterminado el cual toca un campo de conocimiento específico por documento.

- Utilizar una técnica de desambiguación en función al contexto del documento, implico procesar todo el contenido del documento para realizar el análisis de los términos extraídos de él. Los resultados mostrados en las cifras de Precision y Recall demuestran que se puede desambiguar términos sin necesidad de llegar a utilizar una gran cantidad de reglas que definan los significados así como el uso de diccionarios técnicos para conocer los significados exactos de los términos encontrados. Usando una ontología debidamente estructurada se puede lograr encontrar los conceptos deseados y esto puede mejorarse aún más si se refina la funcionalidad de NLP en la herramienta y si se extiende aún más la ontología.

8.2 Recomendaciones para trabajos futuros

- Lo ideal para una herramienta de NLP que procese documentos académicos del campo de las Ciencias de la Computación sería que cuente con funcionalidades de reconocimiento de nombres de entidades expresados en más de una palabra, así como un diccionario que permita detectar términos técnicos que son utilizados en un determinado tipo de documentos académicos como las evaluaciones realizadas en la especialidad de Ingeniería Informática de la Universidad. La realización de una herramienta que pueda reconocer nombres de entidades para el idioma español supone un trabajo bastante extenso por si solo y que incluso puede especializarse en los temas que contienen los documentos a trabajar.

- La utilización de una plantilla de CommonKADS para modelar el diseño de una herramienta de este tipo no es de carácter obligatorio pero si supone una ayuda significativa no solo para manejar de forma más ordenada las funcionalidades de la herramienta sino también para poder demostrar de forma más sencilla el flujo que sigue para realizar la anotación semántica así como para tener una visión de todas librerías y funcionalidades de terceros usadas en general.

- La estructura de la ontología modelada permite ser llenada por el resto de conceptos y cursos impartidos en todas las especialidades de la universidad. En caso se desarrollará una ontología de esa magnitud sería recomendable contar con especialistas de cada uno de los contenidos que se modelaran debido a que la ontología debe plasma el conocimiento de un experto en el campo de forma que pueda ser entendido por quienes no tienen un absoluto dominio de sus conceptos.

- La utilización de más de una técnica de desambiguación podría mejorar la eficiencia de esta y otras herramientas relacionadas, en especial si se hace uso no solo de una ontología sino de repositorios semánticos de mayor envergadura como WordNet.

Referencias bibliográficas

[1] Berners-Lee, Tim, James Hendler and Ora Lassila. "The Semantic Web." *Scientific american* 284, no. 5 (2001): 28-37.

[2] Davies, John, Dieter Fensel and Frank. Van Harmelen. *Towards the Semantic Web*: Wiley Online Library, 2003.

[3] Daconta, Michael C, Leo J Obrst and Kevin T Smith. *The Semantic Web: A Guide to the Future of Xml, Web Services, and Knowledge Management*: Wiley. com, 2003.

[4] Studer, Rudi, Stefan Decker, Dieter Fensel and Steffen Staab. "Situation and Perspective of Knowledge Engineering." *Knowledge Engineering and Agent Technology. IOS Press, Amsterdam*, (2000).

[5] Gruber, Thomas R. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing?" *International journal of human-computer studies* 43, no. 5 (1995): 907-928.

[6] Chandrasekaran, Balakrishnan, John R Josephson and V Richard Benjamins. "What Are Ontologies, and Why Do We Need Them?" *Intelligent Systems and Their Applications, IEEE* 14, no. 1 (1999): 20-26.

[7] O'Leary, Daniel E. "Using Ai in Knowledge Management: Knowledge Bases and Ontologies." *Intelligent Systems and Their Applications, IEEE* 13, no. 3 (1998): 34-39.

[8] Meena, Erica, Ashwani Kumar and Laurent Romary. "An Extensible Framework for Efficient Document Management Using Rdf and Owl." In *Proceeedings of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology*, 51-58: Association for Computational Linguistics, 2004.

[9] Marshall, Catherine C. "Annotation: From Paper Books to the Digital Library." In *Proceedings of the second ACM international conference on Digital libraries*, 131-140: ACM, 1997.

[10] O'Hara, Kenton and Abigail Sellen. "A Comparison of Reading Paper and on-Line Documents." In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, 335-342. Atlanta, Georgia, USA: ACM, 1997.

[11] Wolfe, Joanna L. "Effects of Annotations on Student Readers and Writers." In *Proceedings of the fifth ACM conference on Digital libraries*, 19-26. San Antonio, Texas, USA: ACM, 2000.

[12] Sannomiya, Takeshi, Toshiyuki Amagasa, Masatoshi Yoshikawa and Shunsuke Uemura. "A Framework for Sharing Personal Annotations on Web Resources Using Xml." In *Information Technology for Virtual Enterprises, 2001. ITVE 2001. Proceedings. Workshop on*, 40-48: IEEE, 2001.

[13] Marshall, Catherine C. "Toward an Ecology of Hypertext Annotation." In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems*, 40-49. Pittsburgh, Pennsylvania, USA: ACM, 1998.

[14] Srivastava, Divesh and Yannis Velegrakis. "Intensional Associations between Data and Metadata." In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 401-412: ACM, 2007.

[15] Corcho, Oscar. "Ontology Based Document Annotation: Trends and Open Research Problems." *International Journal of Metadata, Semantics and Ontologies* 1, no. 1 (2006): 47-57.

[16] Kiryakov, Atanas, Borislav Popov, Ivan Terziev, Dimitar Manov and Damyan Ognyanoff. "Semantic Annotation, Indexing, and Retrieval." *Web Semantics: Science, Services and Agents on the World Wide Web* 2, no. 1 (2004): 49-79.

[17] Uren, Victoria, Philipp Cimiano, José Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta and Fabio Ciravegna. "Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art." *Web Semantics: science, services and agents on the World Wide Web* 4, no. 1 (2006): 14-28.

[18] Agosti, Maristella and Nicola Ferro. "A Formal Model of Annotations of Digital Content." ACM Transactions on Information Systems (TOIS) 26, no. 1 (2007): 3.

[19] Netcraft, September 2013 Web Server Survey, 2013, http://news.netcraft.com/archives/2013/09/05/september-2013-Web-server-survey.html [Consulta: 18 de Septiembre del 2013]

[20] Pino, F, Félix García and Mario Piattini. "Revisión Sistemática De Mejora De Procesos Software En Micro, Pequeñas Y Medianas Empresas." *Revista Española de Innovación, Calidad e Ingeniería del Software* 2, no. 1 (2006): 6-23.

[21] Joksimovic, Srecko, Jelena Jovanovic, Dragan Gasevic, Amal Zouaq and Zoran Jeremic. "An Empirical Evaluation of Ontology-Based Semantic Annotators." In *Proceedings of the seventh international conference on Knowledge capture*, 109-112: ACM, 2013.

[22] Navarro-Galindo, José L and José Samos. "Manual and Automatic Semantic Annotation of Web Documents: The Flersa Tool." In *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services*, 542-549: ACM, 2010.

[23] Pipitone, Arianna and Roberto Pirrone. "A Framework for Automatic Semantic Annotation of Wikipedia Articles." In *6th Workshop on Semantic Web Applications and Perspectives, Bressanone, Italy*, 2010.

[24] Popov, Borislav, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff and Miroslav Goranov. "Kim–Semantic Annotation Platform." In *The Semantic Web-Iswc 2003*, 834-849: Springer, 2003.

[25] De Maio, C, G Fenza, M Gallo, V Loia and S Senatore. "Formal and Relational Concept Analysis for Fuzzy-Based Automatic Semantic Annotation." *Applied Intelligence*, (2013): 1-24.

[26] Chechev, Milen, Meritxell Gonzàlez, Lluís Màrquez and Cristina España-Bonet. "The Patents Retrieval Prototype in the Molto Project." In *Proceedings of the 21st international conference companion on World Wide Web*, 231-234: ACM, 2012.

[27] Schreiber, Guus. *Knowledge Engineering and Management: The Commonkads Methodology*: the MIT Press, 2000.

[28] Apache Jena, URL: http://jena.apache.org/documentation/ontology/ , [Último acceso: 15/11/2013]

[29] Grishman, Ralph. "Natural Language Processing." *Journal of the American Society for Information Science* 35, no. 5 (1984): 291-296.

[30] Mihalcea, Rada, Hugo Liu and Henry Lieberman. "Nlp (Natural Language Processing) for Nlp (Natural Language Programming)." In *Computational Linguistics and Intelligent Text Processing*, 319-330: Springer, 2006.

[31] Chowdhury, Gobinda G. "Natural Language Processing." *Annual review of information science and technology* 37, no. 1 (2003): 51-89.

[32] Padró, Lluís and Evgeny Stanilovsky. "Freeling 3.0: Towards Wider Multilinguality." (2012).

[33] Jain, Vishal, and Mayank Singh. "*Ontology Development and Query Retrieval using Protégé Tool.*" International Journal of Intelligent Systems & Applications5.9 (2013).

[34] Pérez, Jorge, Marcelo Arenas, and Claudio Gutierrez. "*Semantics and Complexity of SPARQL.*" The Semantic Web-ISWC 2006 (2006): 30-43.

[35] Hartig, Olaf, Christian Bizer, and Johann-Christoph Freytag. "*Executing SPARQL queries over the Web of linked data.*" The Semantic Web-ISWC (2009): 293-309.

[36] Hotho, Andreas, Steffen Staab, and Gerd Stumme. "*Ontologies improve text document clustering.*" Third IEEE International Conference on. IEEE (2003).

[37] Plaza, Laura, Mark Stevenson, and Alberto Díaz. "*Improving summarization of biomedical documents using Word Sense Disambiguation.*" Association for Computational Linguistics, 2010.

[38] Goutte, Cyril, and Eric Gaussier. "*A probabilistic interpretation of precision, recall and F-score, with implication for evaluation.*" Advances in information retrieval. Springer Berlin Heidelberg, 2005.