SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Método Híbrido para Extracción del Contexto Semántico de Términos Mediante el uso de Ontologías

presentada por

Ing. Rubén Peralta Espinoza

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Director de tesis

Dra. Alicia Martínez Rebollar

Codirector de tesis

Dr. Juan Francisco Mosiño

Cuernavaca, Morelos, México. Junio de 2018.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Morelos a 25 de junio del 2018
OFICIO No. DCC/205/2018

Asunto: Aceptación de documento de tesis

DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **Ing. Rubén Peralta Espinoza**, con número de control M16CE013, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado "**Método Hibrido para la Extracción del Contexto Semántico de Términos mediante el uso de Ontologías"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTORA DE TESIS

Dra. Alicia Martínez Rebollar
Doctora en Informática
7399055

CO-DIRECTOR DE TESIS

Juan Francisco Mosiño

Dr. Juan Francisco Mosiño
Doctor en Ciencias (Óptica)

REVISOR 1

Dr. Luis Gerardo Vela Valdés
Doctor en Ciencias en Ingeniería Electrónica
7980044

REVISOR 2

Dr. Manuel Mejía Lavalle
Doctor en Ciencias Computacionales
8342472

C.p. M.T.I. María Elena Gómez Torres - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 26 de junio de 2018
OFICIO No. SAC/301/2018

Asunto: Autorización de impresión de tesis

ING. RUBÉN PERALTA ESPINOZA
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Método Hibrido para la Extracción del Contexto Semántico de Términos mediante el uso de Ontologías"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
EXCELENCIA EN EDUCACIÓN TECNOLÓGICA®
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

SEP TecNM
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

C.p. M.T.I. María Elena Gómez Torres .- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Agradecimientos

Al Consejo Nacional de Ciencias y Tecnología (CONACYT) por el apoyo económico proporcionado para mis estudios, los cual me permitieron realizar esta investigación de tesis.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (**CENIDET**) por brindarme la oportunidad de realizar los estudio de una maestría.

A mi directora de tesis la Dra. Alicia Martínez Rebollar y a mi codirector el Dr. Juan Francisco Mosiño por el apoyo, la ayuda y el tiempo proporcionado en todo momento en el transcurso del desarrollo de esta investigación.

A mis revisores el Dr. Luis Gerardo Vela Valdez y al Dr. Manuel Mejía Lavalle, por las observaciones y correcciones que han permitido el progreso de esta investigación.

A mis profesores y personal académico, por su experiencia y su conocimiento los cuales han brindado la estructura de mi formación.

Al personal administrativo por el apoyo proporcionado para la realización de los trámites, proyectos y actividades.

Dedicatoria

A Dios que me ha dado vida y salud. A mi esposa y mi hijo que me animaron con gran perseverancia para hacerme continuar.

Abstract

From the entrance of the Internet era, thousands of people began to upload pages, forming *Web* 1.0 that continued to grow, achieving the existence of millions of pages. At that time the page search was based on the comparison of keywords completely ignoring the meaning of the information and consequently obtaining little precision in the searches.

In response to this problem Tim Berners-Lee [1] proposed the *semantic Web* in the framework of *Web* 2.0. The aim of the semantic *Web* is to add meaning to the elements of a *Web* page using a set of technologies and models of knowledge representation within which the ontology stands out.

Ontology is a formal representation of knowledge in a specific domain, composed of concepts, relations, functions, instances and axioms. Querying the semantic *Web* using ontologies seeks the meaning of terms [2], rather than keywords. The terms being associated with other terms providing greater meaning and consequently the searches become more accurate. In this sense, the precision is a product of the closeness between terms that is made evaluated of the distance between the entities to which we will call semantic similarity of terms.

Semantic similarity measures the closeness of the relationship of each of the terms of ontology by establishing the semantic context of the terms. In particular, the semantic context allows us to determine which pair of terms is more related to each other.

This thesis proposes the development of hybrid method for extracting the semantic context of terms in documents through the use of ontologies by using the measure of similarity of entities, taking into account the weight of the edges and the relationships between nodes.

Resumen

Desde la entrada de la era del Internet miles de personas comenzaron a subir páginas, formando así la *Web 1.0* que siguió creciendo logrando la existencia de millones de páginas. En ese entonces la búsqueda de páginas se basada en la comparación de palabras clave ignorando completamente el significado de la información y en consecuencia obteniendo poca precisión en las búsquedas.

En respuesta a esta problemática Tim Berners-Lee [1], propuso la *Web semántica* en el marco de la *Web 2.0.* El objetivo de la Web semántica es agregar significado a los elementos de una página *Web* utilizando un conjunto de tecnologías y modelos de representación del conocimiento dentro de las cuales destaca la ontología.

La ontología es una representación formal del conocimiento en un dominio específico, formada por conceptos, relaciones, funciones, instancias y axiomas. Las consultas la *Web* semántica utilizando las ontologías busca el significado de los términos [2], en lugar de palabras clave. Los términos al estar asociados con otros términos proporcionando un mayor significado y en consecuencia las búsquedas se tornan más precisas. En este sentido, la precisión es producto de la cercanía entre términos que se realiza evaluado de la distancia entre las entidades a lo cual llamaremos similitud semántica de términos**.**

La similitud semántica mide la cercanía de la relación de cada uno de los términos de una ontología estableciendo el *contexto semántico* de los términos. En particular, el contexto semántico nos permite determinar qué par de términos está más relacionado entre sí.

En esta tesis se establece el desarrollo de un método híbrido para la extracción del contexto semántico de términos en documentos mediante el uso de ontologías a través de la medida de similitud de entidades, tomando en cuenta el peso de las aristas y las relaciones entre nodos.

Tabla de contenido

Índice de tablas

Índice de figuras

Capítulo 1. Introducción

En este capítulo, se describe el problema que motivó el desarrollo de la tesis, los objetivos y la estructura de esta tesis.

1.1. Antecedentes

Este proyecto forma parte de un trabajo de investigación que actualmente se está desarrollando en el área de sistemas distribuidos a nivel doctoral. El título del proyecto es: **"Un nuevo enfoque de anotación semántica para una Arquitectura Integral de Búsqueda de Información basada en ontologías"**. La propuesta en este trabajo de investigación es la siguiente:

Actualmente las técnicas para la descripción de contenidos y procesamiento de consultas en la búsqueda de la Información se basan en palabras clave por lo que proporciona capacidades limitadas para capturar la conceptualización asociadas con las necesidades del usuario y del contenido. Esto se debe a que gran parte de la información se encuentra en forma no estructurada.

Debido a las necesidades de las organizaciones para mejorar el proceso de búsqueda de información, se han creado nuevas estrategias para mejorar este proceso. Con el objetivo de solucionar las limitaciones que tienen los modelos basados en palabras clave en la búsqueda de información, se presenta la búsqueda semántica que realiza búsqueda mediante significado en lugar de literales; dicho enfoque ha sido ampliamente utilizado en el campo de la recuperación de información. Sin embargo, para realizar búsqueda semántica en documentos Web no estructurados, se requiere de distintos procesos entre los que destaca el proceso de anotación semántica, el cual enlaza el significado de datos no estructurados a los conceptos de una base de conocimiento.

En esta propuesta doctoral se presenta un nuevo enfoque de anotación semántica en documentos no estructurados. Para el enfoque de anotación semántica se usará un marco estadístico y la información semántica conceptual de los conceptos de la base de conocimiento para enlazarlos en el corpus de documentos no estructurados.

1.2. Planteamiento del problema

La similitud semántica es utilizada en distintas áreas de investigación tales como inteligencia artificial, gestión de la información, minería, recuperación de la información y en distintas aplicaciones biomédicas.
A pesar de que existen distintas propuestas para medir la similitud, siguen existiendo retos para mejorar este proceso. La mayoría de los métodos utilizados para calcular la similitud semántica no abarcan todos los factores relacionados, tales como la estructura ontológica, contenido de la información, tipo de relaciones, términos similares, múltiples antecesores, rutas semánticas, etc.
A continuación se enumeran las desventajas que presentan los enfoques encontrados en la literatura y mencionados más ampliamente en el marco conceptual:

- *Enfoque basado en conteo de aristas*: el problema de este enfoque es que la distancia entre términos pares con valores similares computados en este tipo de métodos son muy sensitivos a la estructura de las ontologías donde se aplica [39]**.**
- *Los enfoques basados en el contenido de la información*: El problema de este enfoque surge cuando cualquier par de términos de diferentes subcapas (capas secundarias) con el mismo menor ancestro común tienen el mismo valor de similitud lo cual no es razonable en la realidad y si dos términos están bien anotados cerca de la raíz de la ontología, su similitud será muy cercaba a 1 proporcionando así un resultado erróneo [39].
- *Los enfoques basados en Características*: el problema de aquellos que utilizan este enfoque es confiar en características no taxonómicas que son raramente encontradas en ontologías y que requieren un ajuste fino de los parámetros de ponderación para integrar evidencias semánticas heterogéneas.
- *Los enfoques basadas en comentarios*: el problema es que los comentarios de WordNet son muy cortos para ser usados como características y tener un buen resultado de similitud.
- *Los enfoques híbridos*: el problema de este enfoque es que son muy complicados y no siempre tienen un buen desempeño [40]. Es claro que por sí solos los métodos de similitud semántica tienen sus debilidades y fortalezas ya que cada uno fue concebido para un propósito en particular y por ende tendrán un mejor desempeño cuando se sigan las condiciones propuestas. Por ello se han propuesto los algoritmos híbridos que persiguen unir la mayor parte de características posibles (contenido de la información, estructura de la ontología, definiciones, etc.) y obtener un grado mayor de similitud [39].

Con lo anterior, queda claro que la extracción de la similitud es de suma importancia para medir la relación que existe entre una entidad. Bajo este contexto proponemos una estrategia que mida la similitud de las entidades en una ontología utilizando un algoritmo híbrido para determinar su contexto semántico.

1.2.1 Propuesta de solución

En esta tesis se propone el desarrollo de un método híbrido para la extracción del contexto semántico de términos en documentos mediante el uso de ontologías a través de la medida de similitud de entidades, tomando en cuenta el peso de las aristas y las relaciones entre nodos.

1.3 Objetivos

1.3.1 Objetivo general

El objetivo de esta tesis es desarrollar método híbrido para la extracción del contexto semántico de términos en documentos mediante el uso de ontologías a través de la medida de similitud de las entidades, tomando en cuenta el peso de los nodos y de las relaciones o aristas.

1.3.2 Objetivos específicos

- Extraer términos en un corpus de textos para realizar una lista de términos y poder ser buscados en la ontología.
- Desarrollar un método híbrido que calcule la similitud semántica de términos utilizando el peso de los nodos y las relaciones entre aristas.
- Desarrollar un prototipo para determinar el contexto semántico en documentos electrónicos.
- Realizar pruebas del prototipo y comparar el método propuesto con la literatura.

1.4 Estructura de la tesis

La tesis está estructurada de la siguiente manera:

- **Capítulo 2. Marco conceptual**: En este capítulo, se presentan los conceptos utilizados en la investigación.
- **Capítulo 3. Estado del arte:** En este capítulo, se presentan las herramientas que realizan anotación semántica. Las herramientas han sido divididas en anotación semántica de recursos multimedia y anotación semántica de textos o páginas Web.
- **Capítulo 4. Metodología de solución:** En este capítulo, se describe la metodología para realizar la anotación semántica de videos deportivos.
- **Capítulo 5. Herramienta:** En este capítulo, se muestra el sistema de software de anotación semántica.
- **Capítulo 6. Resultados y discusiones:** En este capítulo, se presentan los resultados obtenidos de las pruebas con el sistema realizado y se inicia una discusión.
- **Capítulo 7. Conclusiones:** En este capítulo, se muestran las conclusiones obtenidas a partir de la Investigación realizada para este proyecto.

Capítulo 2. Marco teórico

En este capítulo, se presentan los conceptos que han sido utilizados en esta tesis.

2.1 Representación del conocimiento

El conocimiento es un conjunto de *representaciones abstractas* que se almacenan mediante la experiencia, la adquisición de conocimiento o a través de la observación. En el sentido más extenso se trata de la tenencia de varios *datos* interrelacionados que al ser tomados por sí solos, poseen un menor valor cualitativo [3].

El conocimiento puede ser de tipo procedimental, declarativo o heurístico [4]. A continuación se describen cada uno de ellos:

- **Conocimiento procedimental** es aquel que es compilado, se refiere a la forma de realizar una cierta tarea (el saber cómo hacerlo). Por ejemplo, el proceso estándar para el ensamble de un vestido, una computadora, una maquina; la realización de cierta pintura, la resolución de ecuaciones algebraicas.
- **El conocimiento declarativo** es conocimiento pasivo, sentencias que expresan hechos del mundo que nos rodea (el saber que hacer). Por ejemplo, la información en una base de datos.
- **El conocimiento heurístico** es algo especial para resolver problemas complejos. Es un criterio, estrategia, método o proceso que simplifica resolver problemas.

Gruber [5] menciona que un cuerpo de conocimiento formalmente representado se basa en la conceptualización de los objetos, conceptos y otras entidades de interés, así como las relaciones que mantienen entre ellas, es decir, representar el conocimiento significa especificar las entidades que forman parte del cuerpo de conocimiento y especificar sus relaciones.

Por otro lado, Norvin y Russell [6] mencionan que existen 3 modelos distintos que clasifican la representación del conocimiento:

- **El modelo conceptual:** es una representación del conocimiento de un dominio, independientemente de cómo se implemente, utilizando estructuras no computables que modelan el problema y la solución en un dominio concreto.
- **El modelo formal:** es una representación “semi-interna” o “semi-computable” del conocimiento de un dominio. Este modelo formal se obtiene a partir del modelo conceptual.
- **El modelo computable:** hace que el modelo formal sea totalmente operativo, y está formado por una base de conocimiento, un motor de inferencias y una serie de estrategias de control.

2.2 Web Semántica

La *Web semántica* es la siguiente generación de la *Web* donde se expresa el significado de los datos mediante metadatos, es decir, crear definiciones formales de dichos datos usando lenguajes de representación especiales, permitiendo con esto la explotación computacional de sus significados. Tim Berners- y su grupo son reconocidos como pioneros de la propuesta de la Web semántica[1]; estos autores son los fundadores del consorcio W3C, *World Wide Web Consortium* que recomienda estándares y orienta a equipos de investigación y desarrollo de la Web semántica en todo el mundo.

La Web semántica prevé una Web donde la información sea accesible y procesable tanto para humanos como para computadoras. La ontología es la piedra angular para realizar esta visión de la Web semántica ya que proporciona la semántica en los datos, es decir, transforman los datos en significado [7].

En la figura 1, se ilustra la transformación en de los datos de un documento HTML en significados [8].

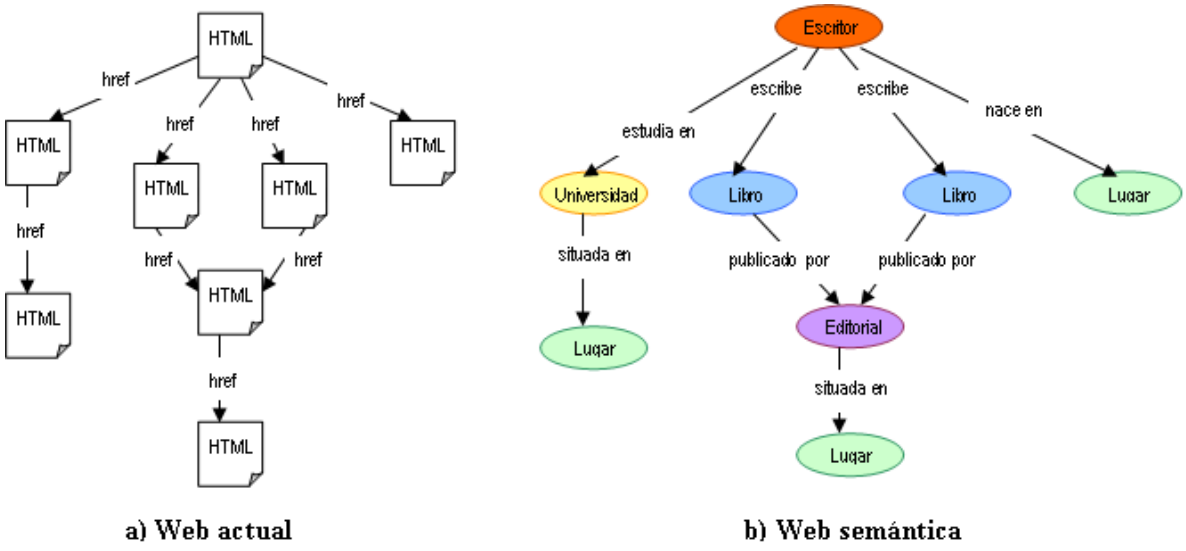

Figura 1. Comparación entre la Web actual y la Web Semántica

2.3 Ontología

Según la definición de Gruber [5] una ontología es una especificación explicita y formal de una conceptualización compartida. Para ampliar esta definición se describen los términos relevantes:

- **Una conceptualización** es una vista simplificada y abstracta del mundo que deseamos representar para algún propósito en específico, definiendo un vocabulario controlado.
- **Explicita** significa que el tipo de conceptos utilizados sean explícitamente definidos, esto es que si también pueden describir otros conceptos del mismo tipo, se definan detalladamente.
- **Formal** se refiere al hecho de que la ontología debe ser legible por la máquina, es decir, que se almacene en un formato digital.
- **Compartido** refleja la noción de que la ontología no es restringida solo para un individuo, sino que es aceptada por un grupo de personas.

A continuación en la figura 2, se muestra un fragmento esquemático de una ontología.

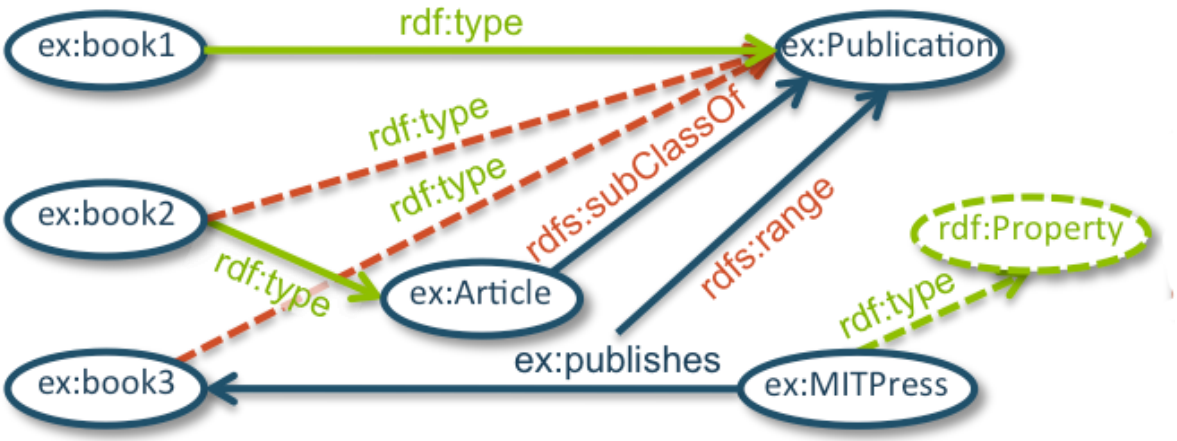

Figura 2. Clases y subclases de una ontología sobre periféricos de ordenador

En una ontología, los conceptos son las unidades fundamentales para la especificación. Proveen una base para la descripción de información. Cada concepto consta de 3 componentes básicos [9]:

- Los **términos** son los nombres utilizados para referirse a un concepto específico que puede incluir un conjunto de sinónimos que especifican los mismos conceptos.
- Los **atributos** son las características de un concepto y describen el concepto a más detalle.
- Las **relaciones** se utilizan para representar correspondencias entre diferentes conceptos y para proveer una estructura general a la ontología.

A continuación se mencionan algunas de las propiedades más importantes que deben cumplir las ontologías [9]:

- **Claridad**: para comunicar el significado intencionado de los términos definidos.
- **Coherencia**: para sancionar inferencias que son consistentes con las definiciones.
- **Extensibilidad***:* para anticipar el uso de vocabulario compartido.
- **Sesgo de codificación mínimo**: para especificar al nivel de conocimiento sin depender de una codificación particular a nivel de símbolo.
- **Mínimo compromiso ontológico:** para hacer la menor cantidad de "pretensiones" acerca del mundo modelado.

2.4 Similitud semántica

La similitud semántica es la fuerza de la interacción entre elementos semánticos (palabras o conceptos) basado en sus significados [10], esta fuerza refleja la cantidad de relación existente y generalmente se encuentra en un rango de 0 a 1.

Para reforzar el concepto de similitud se plantea el ejemplo de las figuras geométricas en la figura 3:

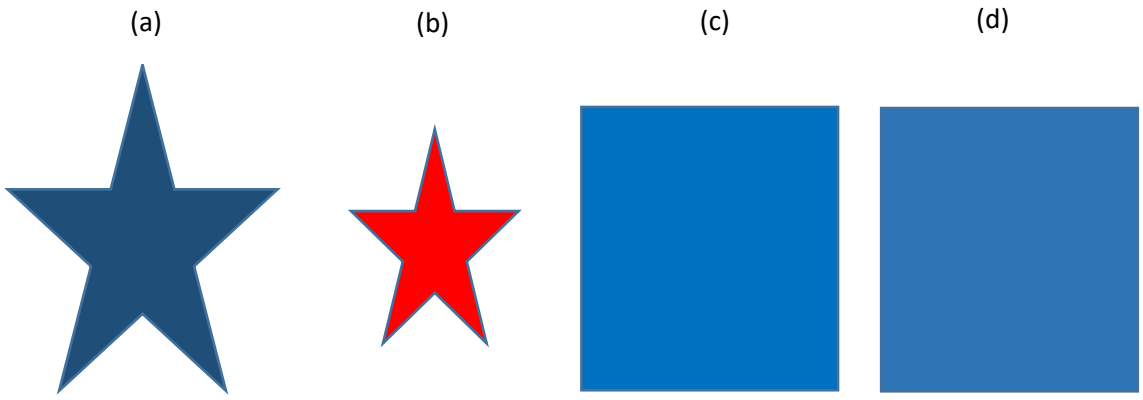

Figura 3. Ejemplo de similitud en figuras geométricas

Inmediatamente podremos percatarnos que las figuras **a** y **b** son similares porque son estrellas o diferentes porque no tienen el mismo tamaño y color; también es cierto que **a** y **d** son similares pues tienen el mismo color o diferentes porque tienen distinta forma. En ambos casos tanto la forma, color y tamaño son características que se pueden medir.

Con el ejemplo anterior la similitud puede resultar una tarea no trivial y menos si en lugar de cuerpos geométricos tenemos conceptos que pertenecen a un concepto más general, formando así, una estructura conceptual [11].

La herramienta matemática *Medida de similitud semántica* es la encargada de estimar la fuerza de las relaciones semánticas entre unidades del lenguaje, conceptos o instancia, a través de una

descripción (numérica) obtenida de la comparación de información que respalda su significado[10].

2.4.1 Enfoques para calcular la Similitud semántica entre un par de términos

A continuación se describen los enfoques para calcular la similitud semántica de conceptos, encontrados en la literatura:

Enfoque basado en conteo de aristas que evalúan la similitud basado en el número de enlaces taxonómicos y la mínima distancia de ruta entre 2 conceptos presentes en una ontología. Este enfoque lo presentan Wu and Palmer [12], Leacock and Chodorow [13], Li et al[14], Rada et al [15].

Los enfoques basados en el contenido de la información cuantifican la similitud entre conceptos como una función del contenido de la información que ambos conceptos tienen en común en una ontología dada. Este enfoque lo presentan Zhou et al [16], Tversky [17]; Sánchez et al [18], Seco et al [19].

Los enfoques basados en Características estiman la similitud de acuerdo a la suma ponderada del número de características comunes y no comunes (Sánchez et al. [20]). Como características, los autores usualmente consideran información taxonómica y no taxonómica, además de las descripciones de conceptos recuperadas de diccionarios. Tversky [17], Petrakis et al [21] son algunos de los autores que tienen el enfoque.

Los enfoques basadas en comentarios explotan las pequeñas definiciones proporcionadas por *WordNet* con el fin de cuantificar los traslapes entre comentarios de dos conceptos con sus vecinos semánticos Este enfoque lo presentan Banerjee y Pedersen [22], Lesk [23] también Patwardhan y Pedersen [24].

Los enfoques híbridos combinan las mediadas concebidas de los enfoques anteriormente mencionados con el fin unir las ventajas y disminuir sus desventajas. Algunos de los representantes son: Zhou et al [16], Jeong C. J. [25], Wu et al [26] y Zhang et al [27].

2.4.2 Contexto semántico en una ontología

En la lengua una palabra tiene múltiples aplicaciones. Para seleccionar la acepción adecuada, las palabras necesitan ubicarse en un contexto es decir, colocarse entre otras palabras. En un sentido amplio, el contexto es un marco de referencia con respecto al cual los signos adquieren un significado determinado [28].

En una ontología los términos están relacionados con otros términos y se utiliza la similitud semántica para determinar cuan similar son los términos [10], sin embargo existe un par de términos que está más relacionado entre sí al que llamaremos el *contexto semántico*.

El contexto semántico es ampliamente utilizado también en diferentes áreas tales como: geolocalización [29], redes colaborativas [30], reconocimiento de imágenes [31], sistemas de recomendación [32], etc.

Capítulo 3. Estado del arte

En este capítulo, se presentan los trabajos de investigación que son relevantes o de apoyo teórico para la propuesta de tesis que se presenta.

3.1 Una nueva medida de Similitud Semántica para la Ontología del Gen basada en descendientes y longitud de trayectoria

En este trabajo de investigación [33] se propone una medida de similitud semántica híbrida llamada por sus siglas en inglés **HSS** (***H**ybrid **S**tructural **S**imilarity*) que toma en cuenta la longitud de ruta entre dos término así como la especificidad del ***m**as **i**nformativo **a**ncestro **c**omún* (**MICA**).
Mientras que otros métodos han usado técnicas como la profundidad o el contenido de la información del MICA dentro de una base de datos de organismos específicos, este trabajo propone calcular la especificidad del MICA basado en el conteo relativo de descendiente en la ontología.
El método que propone este trabajo es considerablemente menos extensivo computacionalmente ya que no depende de la selección de un corpus. Este método se considera híbrido por que calcula la similitud semántica como una función de la longitud de ruta así como el contenido de la información estructural del ancestro menos común.
A continuación se muestra la ecuación para calcular la similitud semántica:

$$sim_{HSS} = IC_{structural}\big(LCA(t1,t2)\big) * Z(PL(t1,t2))$$

Para validar el método propuesto los autores utilizaron la base de datos de Interacción Proteica ó DIP por sus siglas en inglés. Además compararon los resultado del nuevo método con los métodos de: Resnik [34], Lin [35], Jiang [36], y Wang [37].

A diferencia del método propuesto, nosotros si definimos un corpus de textos, sin embargo su cálculo se realizara en una sección dedicada para el proceso de lenguaje natural. De esta manera se tomará ventaja al separar el análisis de texto del cálculo de similitud semántica para no afectarle en cuanto a coste computacional.

Nuestro método será evaluado a través de *WordNet* y esto porque nuestro método no está restringido para un tipo de ontología, puede utilizar cualquier ontología. Para evaluar nuestro método también utilizaremos el método de Resnik[34].

3.2 Medida híbrida para la Similitud Semántica de términos de la Ontología del Gen

En este trabajo de investigación [38] se propone un método híbrido de similitud semántica entre dos términos de una ontología del Gen tomando en cuenta los múltiples ancestros en común que tienen el par de términos, y agregando la información semántica y la información de la profundidad de los ancestros comunes no redundantes. El método propuesto busca ancestros comunes no redundantes de manera efectiva.

El método hibrido propone resolver los problemas de las técnicas que utiliza:

- Técnicas basadas en nodo: el problema que se herede desde Resnik [34] consiste en tomar en cuenta el contenido de información de un ancestro cuando un término puede porvenir de varios ancestros.
- Técnicas basadas en aristas: esta técnica tiene deficiencias para determinar la similitud semántica porque existen aristas que están en el mismo nivel y por lo tanto tienen la misma distancia semántica desde la raíz.

Para solucionar estos problemas se propone un método híbrido que se basa en las siguientes observaciones:

1. Un término en grafo de la ontología del gen hereda semántica de sus ancestros
2. Algunos ancestros comunes de dos términos en una ontología del Gen proveen semántica redundante para cuantificar su similitud.
3. La similitud entre términos cerca de la raíz de un grafo de ontología genética es menor que la de los términos más alejados de la raíz.

El trabajo de investigación propone la siguiente ecuación para calcular la similitud semántica:

$$sim_{\_fused}(c_1, c_2) = (sim_{dep}(c_1, c_2) + sim_{sv}(c_1, c_2))/2$$

A diferencia de este método propuesto, nuestro método no está restringido a un solo tipo de ontología, puede utilizar cualquier tipo de ontología de dominio. Y aunque nuestro método también utiliza la similitud semántica de Resnik [34] heredando el problema de los ancestros múltiples, nuestro método utiliza un segundo método de similitud semántica que es el de Sánchez [48], método que toma en cuenta el contenido de información de un nodo es decir las instancias que pueda tener.

3.3 Método Híbrido para calcular la similitud semántica basado en una ontología del Gen

En este trabajo de investigación [39] se propone un método híbrido de similitud semántica que integra el contenido de la información y la estructura de la ontología del Gen.

Este trabajo de investigación retoma los métodos que determina la similitud semántica y debilidades más importantes en los siguientes dos enfoques:

Métodos basados en grafos: este tipo de métodos determinan la similitud semántica contando el número de aristas a lo largo de los caminos que unen a los términos de la ontología. Como único parámetro se utiliza la distancia entre el ancestro común más y el termino raíz para cuantifica la similitud semántica entre un par de términos. La debilidad de este método es la sensibilidad a la estructura de la ontología, este método además ignora la distancia entre términos pares.

Métodos basados aristas: este tipo de métodos calculan la similitud mediante la comparación de propiedades de los términos implicados con sus ancestros o descendientes. El contenido de información es comúnmente usado para estimar las propiedades de un término dado. La debilidad de estos métodos surge cuando un par de términos en diferentes subcapas con el mismo ancestro menor común tienen el mismo valor de similitud, esto no es razonable en la realidad.

Y para mejorar estas deficiencias propone la utilización de características estructurales de las ontologías del gen. La característica que propone el método son las relaciones entre nodos y cuyas relaciones más comunes son:

- *"is a "* : indica que un hijo es subclase de un padre
- *"part-of"* : indica que un hijo es componente de un padre El método propone la siguiente ecuación para calcular la similitud semántica:

$$HSSCM(c_1, c_2) = \omega_1 \bullet \frac{IC_{max}(c_1, c_2) + IC_{average}(DCA)}{IC(c_1) + IC(c_2)} + exp\left[-\frac{\alpha}{dist(c_1, c_2)} - \beta \times dist(LCA, root)\right]$$

Nuestro método a diferencia del propuesto no está restringido a un solo tipo de ontología, el único requisito para la ontología que utiliza nuestro método es que esta tenga base de conocimiento. Ya que nuestro método también utiliza el contenido de información para calcular la similitud entre términos es propensa a tener errores en el cálculo cuando existen 2 términos en diferentes subcapas y ambos proceden de un mismo ancestro. Para este salir de este problema también se propuso un método que tome en cuenta la estructura de la ontología, en nuesto caso utilizamos en método de Resnik [34].

3.4 Nuevo método para calcular la similitud semántica basada en ontología

En este trabajo de investigación [40] presenta una revisión de los métodos existentes que dependen de una ontología para calcular la similitud semántica. Para su estudio los autores realizan la siguiente clasificación:

- Métodos basados en la estructura de una ontología donde típicamente se mide la distancia entre un par de nodos para cuantificar la similitud y después usar esta medida para evaluar la relación entre los términos correspondientes de la ontología. Sin embargo el uso de estos métodos posee una variedad de problemas. Entre lo más comunes es cuando existe un término en la ontología que tiene más de un nodo padre y esos dos términos pueden tener dos o más nodos LCA (***l**owest **c**ommon **a**ncestor*).
- Métodos basados en el contenido de información donde típicamente se determina la similitud semántica de dos términos basados en el contenido de la información de su nodo LCA. Este tipo de métodos basados en el contenido de la información pueden ser inexactos debido a la poca profundidad de las anotaciones.
- Métodos híbridos que usualmente consideran varias características tales como atributos de similitud, herencia ontológica, contenido de la información, profundidad del nodo LCA, y otros. Uno de los más representativos es el método OSS en el que una puntuación a priori se usó para calcular la distancia entre dos términos, y entonces la distancia se transformó en similitud semántica [41]. Aunque existen muchos métodos híbridos la mayoría resultan complicados y con un desempeño bajo.

Por ello los autores proponen un nuevo método llamado DOPCA, el cual combina dos técnicas:

- Grados de solapamientos en las rutas (*degrees of overlap in paths,* DOP) y
- La profundidad del menor ancestro común (*depth of the lowest common ancestor,* DLCA).

El método propuesto es flexible y puede ser aplicado en ontologías de diversos dominios, incluyendo la ontología del gen, ontología de plantas y muchas otras. El método propone la siguiente ecuación para calcular la similitud semántica:

$$SIm_{DOPCA}(A,B) = W_{DOP}Sim_{DOP} + W_{DLCA}Sim_{DLCA}$$

La similitud más relevante de este método propuesto con nuestro método es que ambos pueden utilizar más de un solo tipo de ontologías y aunque en este trabajo de investigación no se especifica si la ontología debe tener base de conocimiento, nuestro método así lo requiere. Otra similitud importante es que nuestro método pretende ser sencillo de comprender y utilizar para ello diseñar una herramienta computacional para el cálculo de similitud semántica.

3.5 Una nueva ruta basada en una medida híbrida para la similitud de la ontología del gen

En este trabajo de investigación [42] se realiza un recorrido a varios métodos existentes para calcular la similitud semántica además discute sus limitaciones. Con el fin de superar las limitaciones que poseen los métodos los autores proponen una nueva medida llamada SPBHM.
El método propuesto en el trabajo de investigación combina los valores del contenido de la información de los términos de la ontología del gen a lo largo de su estructura. Para determinar la similitud semántica entre dos términos los autores consideran los tres caminos más cortos:

- El primero desde el termino LCA (***L**owest **C**ommon **A**ncestors*) al termino raíz,
- Los otros dos de los términos anotados del LCA.

La primera ruta contribuye con el componente de similitud, mientras que la última ruta contribuye con el componente de disimilitud. Una intuición natural observando un grafo de la ontología del gen es que un los términos ancestros comunes son más específicos por lo tanto la similitud debe ser más alta cuanto más generales sean los términos. Lo inverso ocurre para los demás términos o términos poco comunes. Es decir, si los términos son más específicos en la ontología del gen, entonces la disimilitud deberá ser más baja que si los términos fueran más generales.

A través de esta observación importante los autores formularon una ecuación para calcular la similitud semántica:

$$GOsim_{SPBH}(t1,t2) = (\frac{arctan(sim(t1,t2))}{\frac{\pi}{2}} + (1 - \frac{arctan(dis(t1,t2))}{\frac{\pi}{2}}))/2$$

Nuestro trabajo de investigación toma la lógica propuesta en este trabajo de investigación la cual dice que los términos ancestros comunes son más específicos y que su similitud es más alta cuanto más general sea el término. Estos términos son los que tienen las relaciones *"is a"* (indica que un hijo es subclase de un padre) y *"part-of"* (indica que un hijo es componente de un padre). Para calcular este tipo de similitud utilizamos el método de Sánchez [48].

Capítulo 4. Metodología de solución

En este capítulo, se describe la metodología de solución para el método híbrido de extracción del Contexto semántico de términos mediante el uso de ontologías. En la primera sección se describe brevemente la metodología de solución y en las siguientes secciones se describe de manera detallada cada una de las fases que la componen.

4.1. Descripción general de la metodología de solución

En este proyecto de investigación se llevó a cabo el desarrollo de una herramienta computacional que utilice un método híbrido para la extracción del contexto semántico de términos mediante el uso de ontologías. La herramienta es capaz de analizar un texto determinando los términos relevantes y a través de una ontología de dominio extraer el contexto semántico de los términos.
La presente tesis utiliza técnicas de procesamiento del lenguaje natural para la extracción de términos relevantes. Los términos relevantes es un término que utilizamos para referirnos a las entidades nombradas que se encuentran en el texto. Una entidad nombrada puede definirse como una palabra o secuencias de palabras que se identifican como nombre de persona, organización, lugar, fecha, tiempo, porcentaje o cantidad [44]. El método propuesto en esta investigación impacta en la gran cantidad de textos no estructurados logrando extracción de información de manera correcta y precisa para su análisis y clasificación con mayor eficacia.
En una ontología los términos están relacionados con otros términos y se utiliza la similitud semántica para determinar cuan similar son los términos [10], sin embargo existe un par de términos que está más relacionado entre sí al que llamaremos el *contexto semántico*.

En el desarrollo de la herramienta se creó una interfaz humano-computadora que permite guiar al usuario desde la selección del texto deseado, la selección de la ontología hasta la obtención y visualización del contexto semántico de términos.

La verificación del funcionamiento de la herramienta es fundamental y para ello se realizaron pruebas del funcionamiento utilizando 5 ontologías de dominio, tres de ellas fueron realizadas por alumnos del **CENIDET**.
La herramienta para la extracción de contexto semántico de términos se desarrolló en cuatro fases. En la figura 4, se muestra el diagrama de la metodología de solución propuesta. A continuación, se describen brevemente cada una de las fases.

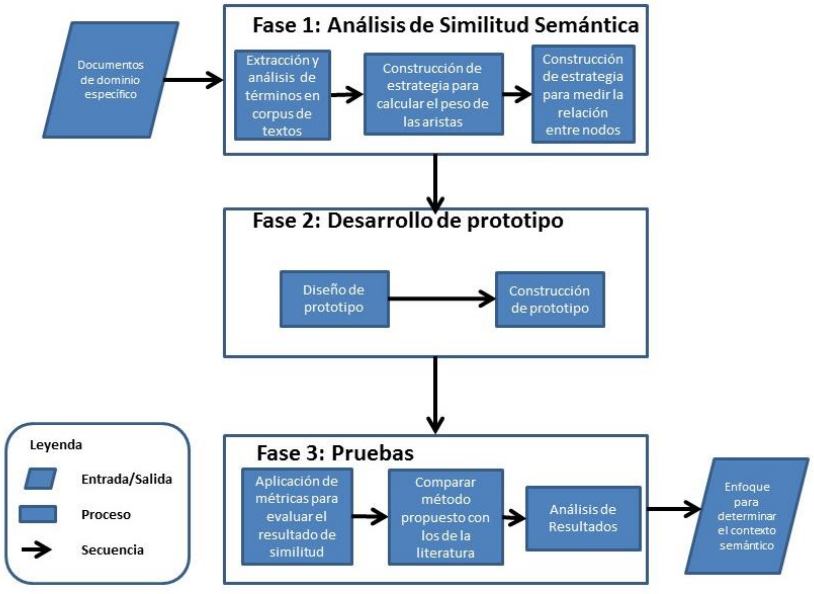

Figura 4. Metodología de solución propuesta

Fase 1: Análisis de Similitud Semántica

En esta fase se desarrolló e implemento el método para la extracción de términos relevantes en textos. Se analizaron y construyeron las estrategias de similitud semántica de los términos a utilizar en la herramienta para obtener el contexto semántico.

Fase 2: Desarrollo de prototipo

En esta fase se llevó a cabo el diseño y construcción de una herramienta para la extracción del contexto semántico de términos.

Fase 3: Pruebas

En esta fase se llevó a cabo la evaluación de las estrategias utilizadas para la obtención de similitud entre términos relevantes utilizando métricas y la comparación de nuestro método con los de la literatura.

4.2 Fase 1: Análisis de Similitud Semántica

La primera fase de nuestra metodología costa de tres procesos: extracción y análisis de términos en un corpus de textos, construcción de estrategia para calcular el peso de las aristas y construcción de estrategia para medir la relación entre nodos. A continuación se describirán cada uno de los tres procesos.

Extracción y análisis de términos en corpus de Textos. En este proceso se analizó y seleccionó el corpus de texto y la ontología a utilizar. El corpus fue realizado con textos referentes a la ontología de animales, tomados de páginas de internet. Cada fragmento del corpus tiene el URL de referencia. Una vez definido el corpus de textos se utilizaron técnicas de procesamiento del lenguaje natural para obtener los términos relevantes. Para el procesamiento de los textos se utilizó Freeling que es una librería de código abierto para el procesamiento multilingüe automático [45]. Se decidió utilizar FreeLing por que nos ofrece librerías para las aplicaciones de Procesamiento del Lenguaje Natural tales como: análisis y anotación lingüística de textos con esto se logró la reducción del coste en la construcción de nuestra herramienta.

FreeLing distinguen cuatro clases de entidades nombradas: Persona (etiqueta NP00SP0), Ubicación geográfica (NP00G00), Organización (NP00O00) y Otros (NP00V00). Estas entidades nombradas nosotros las consideramos como términos relevantes en el texto. De esta manera, esto nos permite identificar las palabras y/o caracteres que no tienen un significado relevante en nuestro análisis. Para esto, se desarrolló un algoritmo que lee el archivo generado por Freeling.

El algoritmo selecciona las palabras cuyas etiquetas representan las entidades nombradas (estas entidades nombradas que seleccionamos, las llamamos términos relevantes) y las almacena en arreglos. En la figura 5, se muestra un fragmento del análisis del texto: *"El etileno es una sustancia natural del crecimiento vegetal, con numerosos efectos sobre el crecimiento y desarrollo de la vida"*. El análisis se realizó de la siguiente manera: palabra la palabra original y lema, seguido del etiquetado que identifica la palabra y al final el grado de similitud que tiene con el diccionario de la herramienta Freeling.

Para nuestro caso de estudio se necesitó extraer los términos relevantes que tengan la etiqueta NCMS, NCMP, NCMFS como se muestra en la figura 5.

Figura 5. Análisis de texto no estructurado con Freeling

Para seleccionar las etiquetas fue necesaria la implementación de un algoritmo de selección implementado en el lenguaje java. El resultado de la selección siguiendo el caso de estudio fue: *etileno, sustancia, crecimiento, efecto, crecimiento, desarrollo.*

Así mismo en esta etapa se realizó la implementación de métodos y herramientas computacionales con los que se logró el acceso y consulta con la ontología. El objetivo de este proceso es obtener los términos de una ontología de dominio. Para la identificación de los términos ontológicos de la ontología de dominio, se consultó la ontología utilizando Apache Jena [46] que es un framework de la Web Semántica de código abierto para Java. La API de Jena para ontologías nos permitió extraer todas las instancias, clases y subclases; a estos les llamamos términos ontológicos. Para extraer los términos ontológicos se realizaron consultas SPAQL utilizando el framework de Jena. En la figura 6, se la consulta realizada a la ontología de Wikipedia disponible en [49] y en la figura 7, se muestra el resultado la consulta realizada a la ontología de Wikipedia mostrando todos los sujetos y predicados no repetidos.

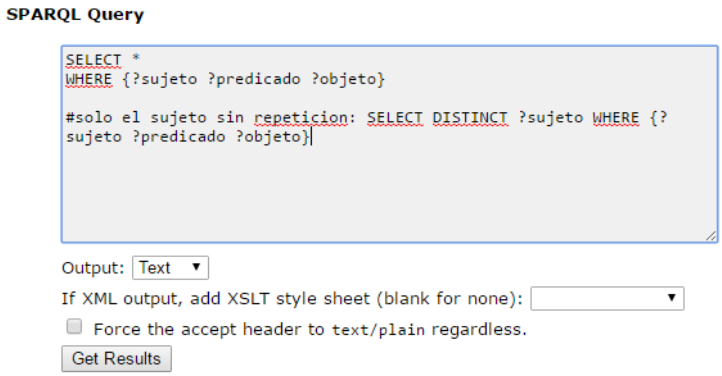

Figura 6. Consulta para obtener todos los elementos de una ontología sin repeticiones

Se realizó también un corpus de ontologías, para cada ontología que se obtuvo se verificó que tuviera Base de conocimiento es decir, necesitamos que la ontología este poblada. Si la ontología es pequeña se pueden agregar instancias manualmente con la herramienta de *Protegé* [50]. Para un mejor desempeño de nuestra herramienta, se almacenaron los términos ontológicos en arreglos.

```
-------------------------------------------------------------------------------------------
| sujeto                                                                  | predicado
|
|
| <http://dbpedia.org/ontology/reservations>                              | <http://www.w3.org/ns/prov#wasDerivedFrom>
<http://mappings.dbpedia.org/index.php/OntologyProperty:reservations>
|
| <http://dbpedia.org/ontology/reservations>                              | <http://www.w3.org/2000/01/rdf-schema#range>
<http://www.w3.org/2001/XMLSchema#boolean>
|
| <http://dbpedia.org/ontology/reservations>                              | <http://www.w3.org/2000/01/rdf-schema#domain>
<http://dbpedia.org/ontology/Restaurant>
|
| <http://dbpedia.org/ontology/reservations>                              | <http://www.w3.org/2000/01/rdf-schema#comment>
reservations required for the establishment or event?"@en
|
| <http://dbpedia.org/ontology/reservations>                              | <http://www.w3.org/2000/01/rdf-schema#label>
"Reservierungen"@de
|
| <http://dbpedia.org/ontology/reservations>                              | <http://www.w3.org/2000/01/rdf-schema#label>
"reservations"@en
```

Figura 7. Resultado de la consulta de la ontología Wikipedia realizado con el Framework de Jena

Para finalizar esta fase se realizó y utilizo un algoritmo que compara los términos ontológicos y los términos relevantes de un texto, el algoritmo se muestra en la figura 8.

```
89          String aux1;
90          String aux2;
91          for (int i = 0; i < arreglo1.size(); i++) {//aqui va el arreglo mas grande para que no mande error en la iteracion
92              aux1 = arreglo1.get(i);
93              for (int j = 0; j < arreglo2.size(); j++) {
94                  aux2 = arreglo2.get(j);
95                  if (aux1.equals(aux2)) {
96                      System.out.println("Se encontro una ocurencia en posicion: " + i + "," + j);
97                      arreglo3.add(aux1);
98                      //System.out.println("y la palabra es: " + arreglo1.get(i));
99                  }
00              }
01              for (int j = 0; j<arreglo3.size(); j++) {
02                  System.out.println(arreglo3.get(j));
03              }
04
tput - SMLparaOWL (run)
   run:
   Se encontro una ocurencia en posicion: 3,4
   planta
   Se encontro una ocurencia en posicion: 4,0
   planta
   raiz
   BUILD SUCCESSFUL (total time: 1 second)
```

Figura 8. Algoritmo de comparación de términos relevantes y términos ontológicos

En la figura 9, se muestra el resultado de comparar dos arreglos, el primero obtenidos de un texto de plantas y los términos de una ontología de plantas [51]. La descripción del resultado es el siguiente:

- arreglo de términos ontológico (línea 1),
- arreglo de términos relevantes (línea 2),
- y el arreglo resultante que contiene los términos coincidentes (línea 3).

El arreglo de coincidencias será utilizado para determinar el contexto semántico y la visualización.

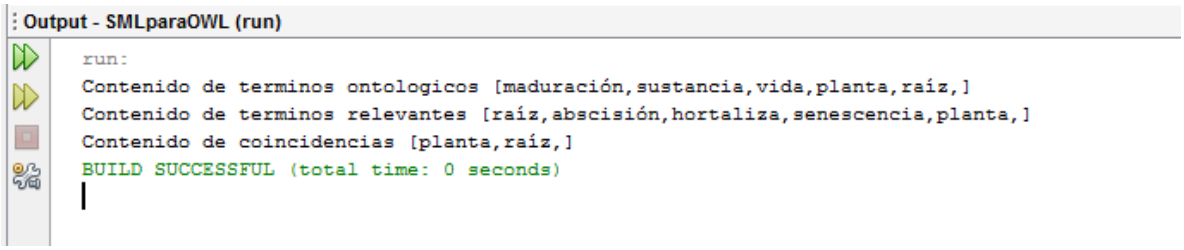

Figura 9. Resultado del algoritmo de comparación de términos

Construcción de estrategia para calcular el peso de las aristas. Para entrar a este proceso fue necesaria la obtención de coincidencias, las coincidencias delimitan los términos ontológicos y sus relaciones semánticas a utilizar para realizar cálculos de similitud semántica. De esta manera se construyó e implemento la estrategia de cálculo semántico entre términos utilizando el peso de las aristas representadas en la taxonomía de la ontología. En la figura 10, se muestra gráficamente un ejemplo de la relación taxonómica de la clase "publicaciones".

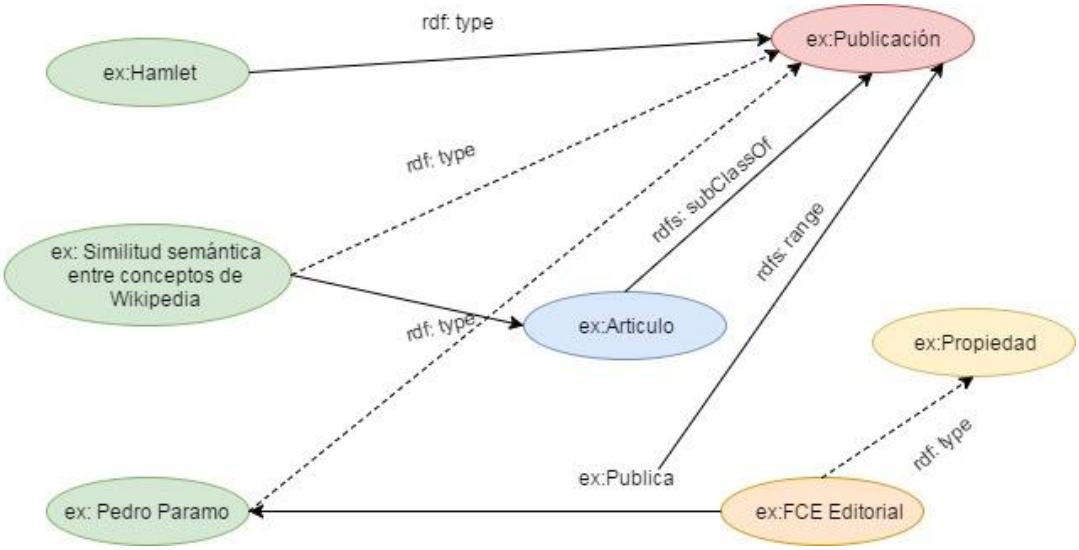

Figura 10. Relación taxonómica entre la clase Publicaciones

La estrategia consistió en analizar todas las relaciones existentes de aquellos términos ontológicos que tienen coincidencia con los términos relevantes del texto. Primero se deben explorar las instancias ontológicas y sus relaciones a través de su URI. Se analizan cada una de las relaciones con una métrica de similitud semántica que calcula el peso de las aristas. Para calcular el peso de las aristas se utiliza la cantidad de información que propuso Resnik [34] para medir la fuerza de identificación de una relación entre dos conceptos c1 y c2 como lo muestra la Formula 1.

$$Sim\big(p(c_1,c_2)\big) = \frac{IC(MSCA(c_1,c_2))}{IC(c_1)+IC(c_2)} \quad (1)$$

Para calcular la similitud semántica de la relación del ejemplo propuesto en figura 10, de una ontología de libros [52], se realizarán a continuacion dos substituciones en la ecuacion pero con dos pares de terminos propuestos:

Concepto uno (c_1)	*Concepto uno (c_2)*
ex: Similitud Semántica entre conceptos de Wikipedia	ex: publicación

$$Sim(p(c_1, c_2)) = \frac{IC(MSCA(0.43,0.20))}{IC(0.43) + IC(0.20)} \quad (2)$$

$$Sim(p(c_1, c_2)) = \frac{(0.57)}{(0.43) + (0.20)} \quad (3)$$

$$Sim(p(c_1, c_2)) = .9047 \quad (4)$$

Y para calcular la similitud de c1 y c2:

Concepto uno (c_1)	*Concepto uno (c_2)*
ex: Similitud Semántica entre conceptos de Wikipedia	ex: articulo

$$Sim(p(c_1, c_2)) = \frac{IC(MSCA(0.45,0.22))}{IC(0.45) + IC(0.22)} \quad (5)$$

$$Sim(p(c_1, c_2)) = \frac{(0.65)}{(0.45) + (0.22)} \quad (6)$$

$$Sim(p(c_1, c_2)) = 0.9701 \quad (7)$$

MSCA(c1, c2) denota el ancestro común de c1 y c2 con el contenido de información alto, IC es el contenido de información que se calcula para cada nodo c en la ontología. El propósito del IC es que cuanto más específico es el nodo en la ontología, mayor es su contenido de información. Por último se guardara el peso de cada una de las relaciones que posee el término ontológico.

Para el calculo de similitud semantica entre los terminos ontológicos se utilizaron las librerias SML [54]. Para realizar el cálculo iterativo del peso de las aristas para cada una de los terminos obtenido previamente se realizó un metodo en lenguaje java. En la figura 11, se muestra la instaciacion del método para configurar el IC y la similitud semantica entre par de nodos.

```
// Definimos el calculo del peso de las aristas con Resnil (IC) y
ICconf icConf = new IC_Conf_Topo("Sanchez", SMConstants.FLAG_ICI_RESNIK_1995);
//ICconf icConfRes = new IC_Conf_Corpus(SMConstants.FLAG_IC_ANNOT_RESNIK_1995);

// Despues definimos la configuracion de la similitud semantica
SMconf smConf = new SMconf("Lin", SMConstants.FLAG_IC_ANNOT_RESNIK_1995_NORMALIZED);
smConf.setICconf(icConf);
```

Figura 11. Instanciación del método para obtener el contenido de la información (IC)

Para la implementación iterativa de nuestro método que calcula la similitud semántica se utilizó la ontología de viajes [53]. En la figura 12, se muestra la ejecución y prueba aleatoria de similitud semántica entre pares de términos. Nuestro método realiza un análisis estructural de la ontología (padres, descendientes) para después ejecutar el motor de análisis de similitud semántica y calcular el peso de las aristas. Nuestro método realiza pruebas aleatorias con los diferentes términos que contiene la ontología y estas pruebas pueden ser ajustadas según sea necesario.

```
================================================================
Loading Semantic Measures Engine for graph http://graph/
================================================================
Graph Info:
http://graph/
Vertices
        Total   : 951  {e.g. https://sites.google.com/site/portdial2/downloads-area/Travel-Domain.owl#Departure_4_BETWEEN_MILWAUKEE}
Edges     : 3557


----------------------------------------------------------------
Pre-processing
----------------------------------------------------------------
Classes  : 37
Instances Accessor loaded: false
----------------------------------------------------------------
Inferences
----------------------------------------------------------------
Inferring ancestors...
Inferring descendants...
Inferring Conceptual Leaves...
----------------------------------------------------------------
Engine initialized
================================================================
Similarity: 0.0
Sim 1/100       Stay/Plane: 0.0
Sim 2/100       City/CarType: 0.0
Sim 3/100       Hotel/Stay: 0.0
Sim 4/100       Ontology/Thing: 0.0
Sim 5/100       Leg/Arrival: 0.0
Sim 6/100       Stay2/State: 0.0
Sim 7/100       Meal/Ontology: 0.0
Sim 8/100       Rental/Stay1: 0.0
Sim 9/100       Leg3/State: 0.0
```

Figura 12. Ejecución y pruebas aleatorias del peso entre términos

Construcción de estrategia para medir la relación entre nodos. Para aumentar la precisión y el mejoramiento del desempeño del análisis de similitud semántica se construyó una segunda estrategia para calcular la relación entre nodos utilizando el contenido de información contenido en la ontología cuyos términos son iguales a los términos relevantes del texto.

El objetivo de este proceso es analizar todas las relaciones existentes de aquellos términos ontológicos que tienen coincidencia con los términos relevantes del texto, se realiza una exploración de las instancias que tienen concordancia con los términos relevantes del texto. Para medir la fuerza de relación entre cada par de conceptos ontológicos realizaremos el cálculo de la similitud propuesto por Sánchez [21] en función al contenido de información. Su fórmula es la siguiente (2):

$$IC(c) - log\, p(c) \cong - log \left(\frac{\frac{|leaves(c)|}{|subsumer(c)|} + 1}{max\_leaves + 1} \right) \quad (1)$$

Donde las leaves(c) son el conjunto de conceptos encontrados al final del árbol taxonómico bajo el concepto c y los subsume(c) es el conjunto completo de ancestros taxonómicos de c que incluye a sí mismo. Es importante señalar que en caso de herencia múltiple todos los ancestros son considerados. La razón se normaliza por el concepto menos informativo (es decir, la raíz de la taxonomía), para el cual el número de hojas es la cantidad total de hojas en la taxonomía (max_leaves) y el número de subsumers incluyendo sí mismo es 1. Para producir valores en el rango 0...1 (es decir, en el mismo rango que la probabilidad original) y evitar los valores de log (0), se agrega 1 a ambas expresiones.

El cálculo de la relación entre nodos se realizó con la librería SML [54], en la figura 13, se muestra la instaciacion del método para obtener la relacion entre par de nodos.

```
// First we configure an intrincic IC
ICconf icConf = new IC_Conf_Topo(SMConstants.FLAG_ICI_SANCHEZ_2011_b_adapted);
// Then we configure the pairwise measure to use, we here choose to use Lin formula
SMconf smConf = new SMconf(SMConstants.FLAG_SIM_PAIRWISE_DAG_EDGE_LI_2003, icConf);

// We define the engine used to compute the similarity
SM_Engine engine = new SM_Engine(g);

double sim = engine.compare(smConf, countryURI, cityURI);
System.out.println("Similarity: " + sim);
```

Figura 13. Instanciación del método para obtener la relación entre nodos

De la misma manera, para realizar cálculos iterativos con el arreglo de coincidencias se utilizó la ontología de viajes [53]. En la figura 14, se muestra la ejecución y prueba aleatoria de similitud semántica entre pares de términos. Nuestro método realiza un análisis estructural de la ontología (padres, descendientes) para después ejecutar el motor de análisis de similitud semántica y calcular el peso de las aristas. Nuestro método realiza pruebas aleatorias con los diferentes términos que contiene la ontología y estas pruebas pueden ser ajustadas según sea necesario.

```
================================================================
Loading Semantic Measures Engine for graph http://graph/
================================================================
Graph Info:
http://graph/
Vertices
        Total   : 951  {e.g. https://sites.google.com/site/portdial2/downloads-area/Travel-Domain.owl#Departure_4_BETWEEN_MILWAUKEE}
Edges     : 3557


----------------------------------------------------------------
Pre-processing
----------------------------------------------------------------
Classes  : 37
Instances Accessor loaded: false
----------------------------------------------------------------
Inferences
----------------------------------------------------------------
Inferring ancestors...
Inferring descendants...
Inferring Conceptual Leaves...
----------------------------------------------------------------
Engine initialized
================================================================
Similarity: 0.0
Sim 1/100       Stay/Plane: 0.0
Sim 2/100       City/CarType: 0.0
Sim 3/100       Hotel/Stay: 0.0
Sim 4/100       Ontology/Thing: 0.0
Sim 5/100       Leg/Arrival: 0.0
Sim 6/100       Stay2/State: 0.0
Sim 7/100       Meal/Ontology: 0.0
Sim 8/100       Rental/Stay1: 0.0
Sim 9/100       Leg3/State: 0.0
```

Figura 14. Ejecución y pruebas aleatorias del peso entre términos

Por finalizar se realizó una suma ponderada tomando como criterios el peso de las aristas y la relación entre nodos, elegimos este último para ajustar la influencia de cada factor en el peso total como lo muestra la figura 15. Las alternativas serán cada una de las relaciones que posee el término ontológico y el peso serán los valores de similitud semántica obtenidos con cada uno de los métodos.

	C1	Cn
A1	W_{A1C1}	W_{A1C2}
A2	W_{A2C1}	W_{A2C2}

Figura. 15 Suma ponderada

Donde:

A= términos vecinos
C=similitud semántica

W= el peso mayor

La fórmula de la suma ponderada que se utilizó fue la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1z}^{n} w_i} = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_n w_n}{w_1 + w_2 + \cdots + w_n} \qquad 1)$$

El peso propuesto en nuestro método hibrido se asignó en base a juicios de expertos Jay y Jiang [59], en la tabla 1 se muestra los resultados del análisis que se realizó con 30 pares de palabras. Se utilizó la taxonomía de *WordNet* con dos medidas de similitud y la medida obtenida por humanos publicada por Miller y Charles [60]. La ponderación del peso más alto se determinó de acuerdo a Sánchez [48] que plantea el descubrimiento de terminos relevantes a partir de una base de conocimientos; por esta razón se propuso el peso mas alto (0.8) al metodo de similitud semantica que utiliza la base de conocimientos de una ontologia, metodo que maneja el contenido de la información. La ulima ponderacion de peso sera de 0.2 para ajustar la ponderacion a cualquier par de terminos. Cabe mencionar que entre mas amplia sea la base de conocimiento mayor será el grado de similitud entre terminos.

Tabla 1. Índices de correlación realizada por expertos, utilizados para la asignación de pesos

Método de similitud	**Correlación (r)**
Juicio humano (replicación)	0.8848
Basado en nodos(Contenido de Información)	0.7941
Basado en aristas (Conteo de aristas)	0.6004

El contexto semántico se aquel par de términos está más relacionado entre sí. La cercanía se calculó a través de la similitud semántica y si sabemos que utilizar un solo método para determinar la similitud nos traería deficiencias en el cálculo, es entonces que nosotros utilizamos dos métodos para calcular la similitud semántica. Una vez lograda la similitud de ambos métodos es necesario la suma ponderada para unir los resultados y obtener un solo valor.

De esta manera se seleccionó el valor más óptimo de la suma ponderada como se muestra en la figura 16. Este es a lo que llamamos contexto semántico.

Termino principal	Términos Similares	Similitud Semántica Sánchez	Similitud Semántica Resnik	Ponderación de términos
Ciudad	Country	0.359995	0.5542856	0.3102837
	Aeropuerto	0.359995	0.5542856	0.3102837
	Estado Federal	0.359995	0.5542856	0.3102837
	Lugar	0.439698	0.5542856	0.3421652
Primera estancia	segunda estancia	0.359995	0.6957526	0.6298116
	estancia	0.439698	0.6957526	0.6743508
Primera Etapa de vuelo	Segunda Etapa de vuelo	0.558815	0.6160825	0.4083507
	Tercera Etapa de vuelo	0.558815	0.6160825	0.4083507
	Etapa de vuelo	0.682538	0.6160825	0.4578399
	Vuelo	0.359995	0.5588153	0.1676446

Figura 16. Tabla que visualiza los resultados para determinar de contexto semántico

4.3 Fase 2: Desarrollo de prototipo

En esta fase se llevó a cabo el diseño y desarrollo de una herramienta que sirvió para la manipulación de textos no estructurados y la obtención de términos relevantes que a su vez fueron empatados con los términos ontológicos, obteniendo de esta manera el contexto semántico. Se decidió el desarrollo de esta herramienta en el lenguaje java ya que este lenguaje permitiría la implementación multiplataforma de nuestra herramienta. Esta fase costa de 2 procesos que se describirán a continuación:

1) **Diseño del prototipo**. En este proceso se realizó el diseño conceptual y lógico de nuestra herramienta. El diseño conceptual consta de los siguientes elementos: requisitos funcionales, arquitectura, casos de uso, descripción de casos de uso, diagrama de procesos, diagrama de secuencias y diagrama de clases. En la figura 17, se muestra el diseño de nuestro método cuyas entradas son dos: el texto elegido y la ontología de dominio. También puede percibirse la salida de nuestro método híbrido: el contexto semántico.

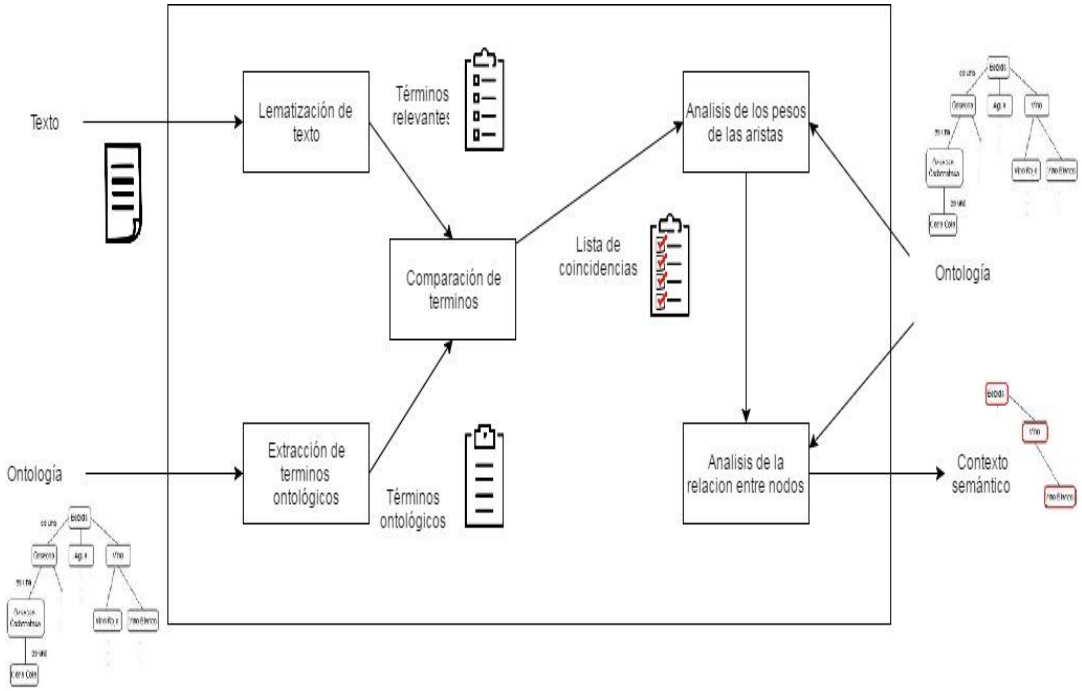

Figura 17. Proceso para la obtención del contexto semántico de términos

2) **Construcción del prototipo**. En este proceso se construyó el prototipo de nuestra herramienta tomando en cuenta el proceso para la obtención del contexto semántico de términos como se muestra en la figura 18.

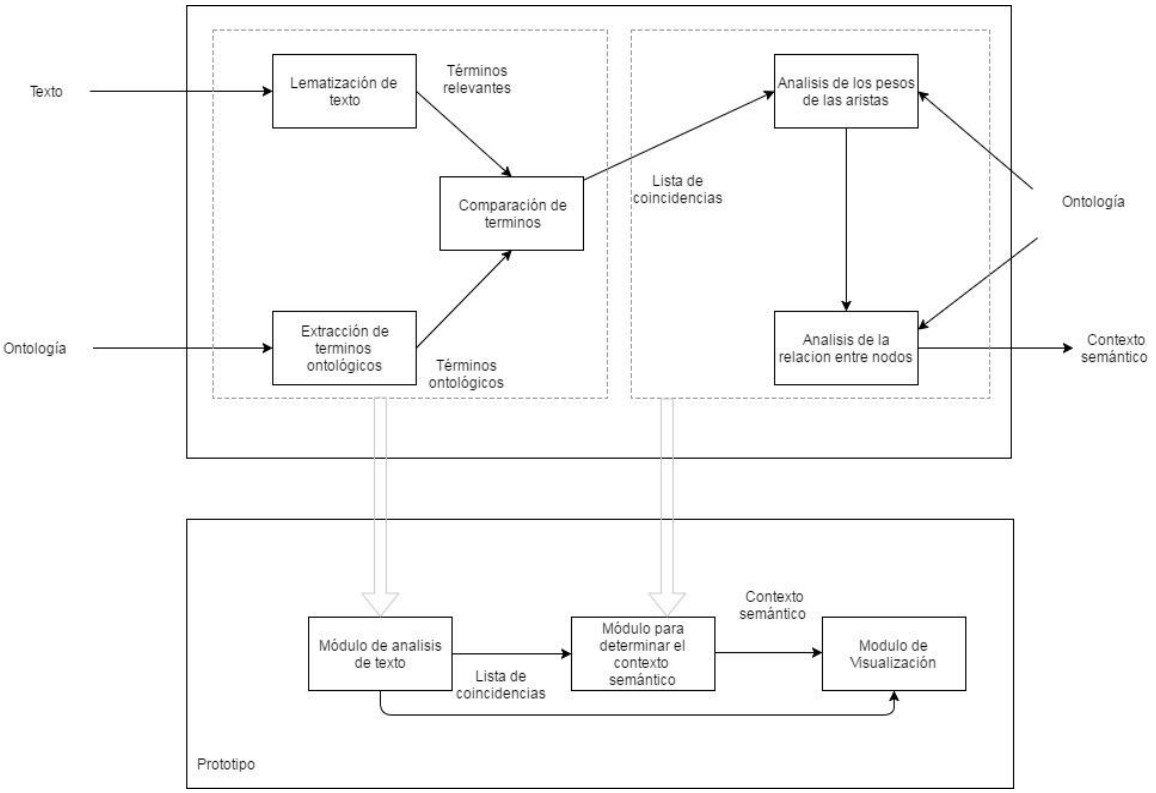

Figura 18. Diseño de la interfaz a partir del proceso de extracción de contexto semántico

Para efectos de diseño se separó el proceso en dos bloques, con el primer bloque se realizó la lematización de textos para la obtención de términos relevantes, después se realizó la extracción de términos ontológicos seguido de una comparación con la que se obtuvo una lista de coincidencias. De la división y análisis del primer bloque se obtuvo un Módulo de análisis de textos cuya entrada sería una ontología y un archivo de texto obteniendo una lista de coincidencias que sería enviada al siguiente módulo.

El siguiente bloque obtenido agrupa el análisis del peso de las aristas y el análisis de la relación entre nodos, esto logrado gracias a la lista de coincidencias. De este bloque de procesos se realizó el segundo módulo cuyas entradas serían: una lista de coincidencias (entre términos ontológicos y términos relevantes) y una ontología. La salida de este módulo fue el contexto semántico que fue enviado al último módulo.

El tercer y último módulo del prototipo permitió al usuario visualizar cada uno de los pasos que se realizaron en el proceso permitiendo un análisis trasparente. El módulo de visualización tiene interacción con los módulos de: análisis de texto y determinación del contexto semántico de manera que estas fueron sus dos entradas.

4.4 Fase 3: Pruebas

En esta fase se realizó un conjunto de pruebas que evaluaron el desempeño de nuestro método híbrido para la obtención de contexto semántico. Esta fase constó de 3 etapas:

1) **Aplicación de métricas para evaluar el resultado de la similitud.**
   Para realizar esta etapa se utilizaron 4 ontologías y utilizando un nuestro método se calculó la similitud semántica. Como se planteó desde el inicio nuestro método utiliza dos métodos para la obtención del contexto semántico, lo que permite un complemento en las deficiencias de ambos métodos. Después se comparó la similitud semántica utilizando la aplicación en línea de Wordnet llamada WS4J [55].
2) **Comparar los resultados obtenidos con otros métodos de la literatura.** En esta etapa se comparó nuestro método con 4 métodos de similitud semántica que se encontraron en la literatura:

   - Similitud semántica de Resnick
   - Similitud semántica de Sánchez
   - Similitud semántica de Seco
   - Similitud semántica de Zhou

3) **Análisis de Resultados.** Para concluir con la tercera fase se realizó un análisis de las etapas anteriores y se realizó un reporte en conjunto que concluyó los resultados del prototipo y des esta manera realizar los ajustes necesarios.

Capítulo 5. Herramienta

En este capítulo, se muestra la herramienta que implementa nuestro método híbrido para determinar el contexto semántico de términos mediante el uso de ontologías.

El desarrollo de la herramienta se realizó utilizando la herramienta de desarrollo NetBeans [34], Mientras que la manipulación de las ontologías se realizó con el framework de Jena [46]. La aplicación fue diseñada para ejecutarse en los equipos de cómputo de los usuarios y ayudarles pasa a paso en la obtención del contexto semántico.

5.1. Arquitectura de la herramienta para la extracción del contexto semántico

En esta sección se presenta la descripción de la arquitectura del funcionamiento de nuestra herramienta para la extracción de contexto semántico de términos. Nuestra herramienta es del tipo MVC, lo cual significa que se ha separado la interfaz del usuario, también se ha separado los datos de la aplicación y la lógica de control. El envío de las peticiones y respuestas se realiza por invocación a través del lenguaje JAVA.

En la figura 19, se muestra el diagrama de la arquitectura de nuestra herramienta.

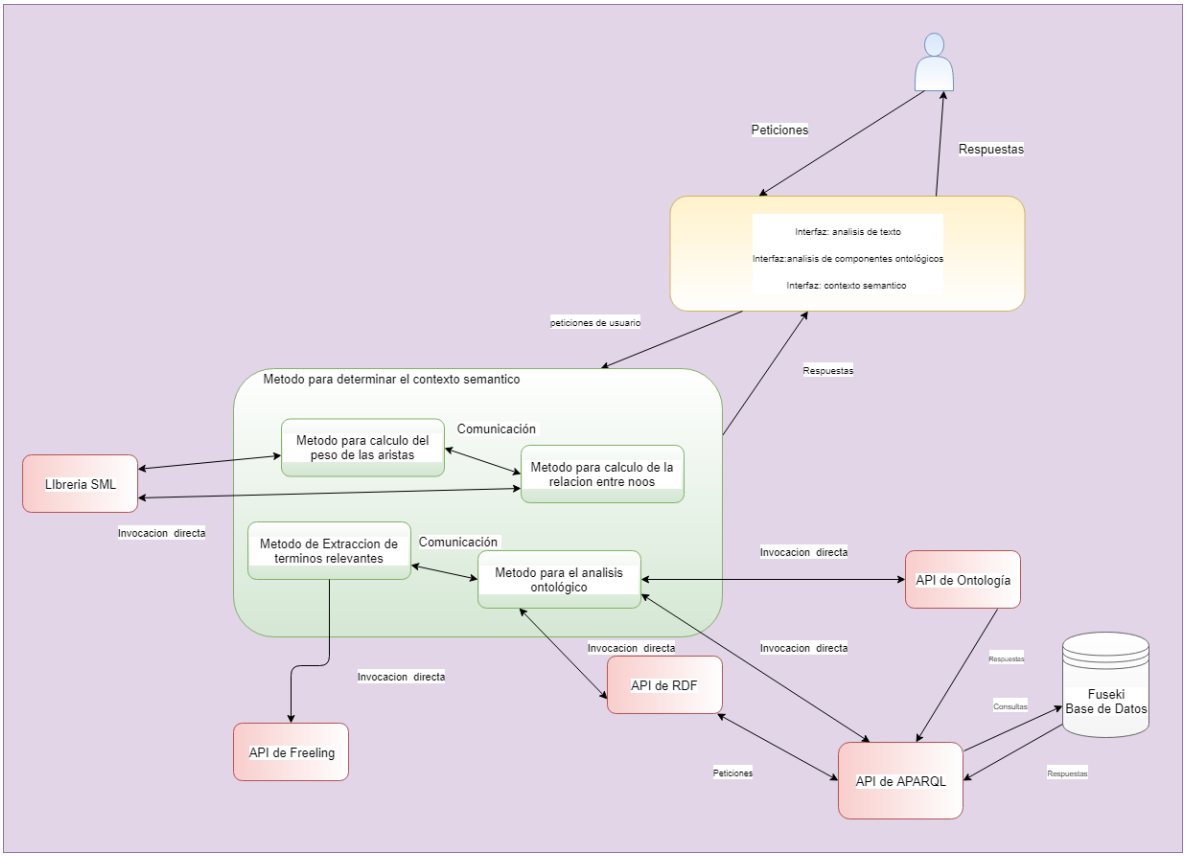

Figura 19. Arquitectura de la herramienta para la extracción de contexto semántico

5.2 Módulo de análisis de texto

La figura 20, muestra la vista general de la herramienta donde el usuario podrá:

- Cerrar la herramienta
- Minimizar la herramienta
- Maximizar la herramienta

Así mismo la ventana principal cuenta con dos módulos:

- Módulo de análisis de texto
- Módulo de análisis de componentes en la ontología
- Módulo de contexto semántico

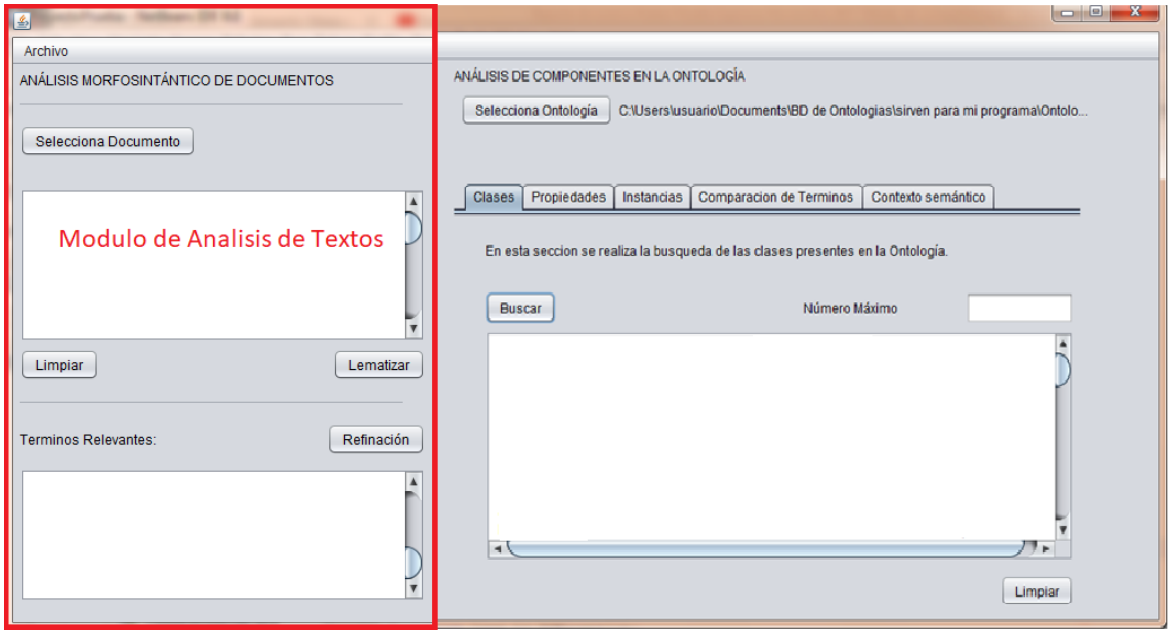

Figura 20. Vista general y Módulo de análisis de texto

En las siguientes páginas se describirán las secciones que componen cada módulo.

Sección de Análisis Morfosintáctico de Documentos

La figura 21, muestra la sección de "*análisis morfosintáctico de documentos*" donde el usuario podrá:

- Seleccionar el documento
- Limpiar área de trabajo (cuando se realiza otro análisis)
- Ejecutar el lematizado

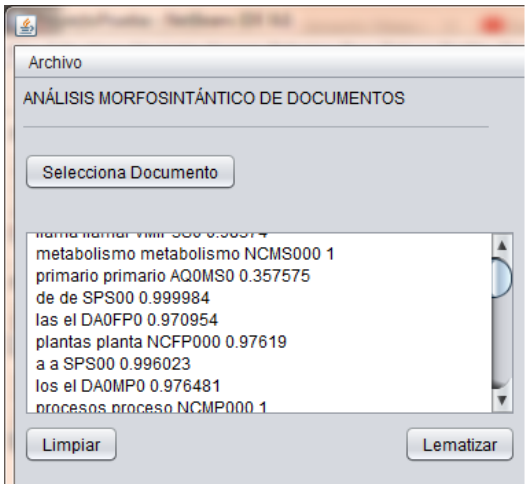

Figura 21. Lematización del texto

Sección de Términos relevantes

La figura 22, muestra la sección de "*Términos relevantes*" donde el usuario podrá:

- Visualizar la eliminación de etiquetas que genera Freeling
- Navegar entre los resultados

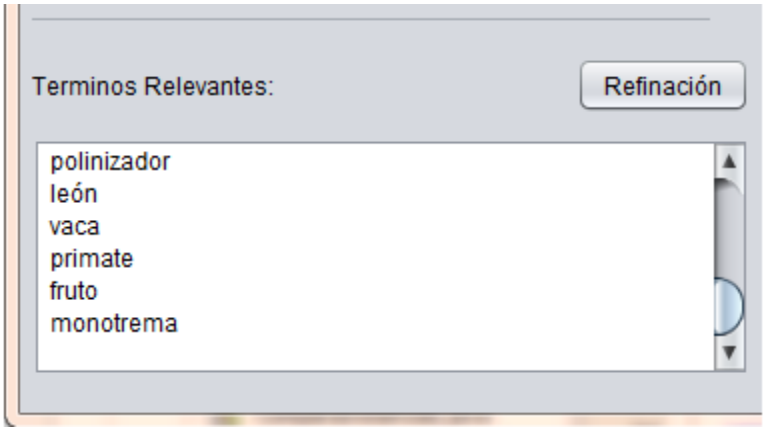

Figura 22. Términos relevantes

5.3 Módulo para el análisis de componentes en la ontología

La figura 23, muestra el "*Módulo para el análisis de componentes de la ontología*" donde el usuario podrá:

- Seleccionar la ontología a utilizar
- Realizar la búsqueda de clases
- Realizar la búsqueda de propiedades

- Realizar la búsqueda de instancias
- Realizar la comparación de términos relevantes y los términos ontológicos
- Limpiar el área de trabajo.

En los 3 casos de búsqueda se podrá acotar la búsqueda a un número determinado que el usuario considere.

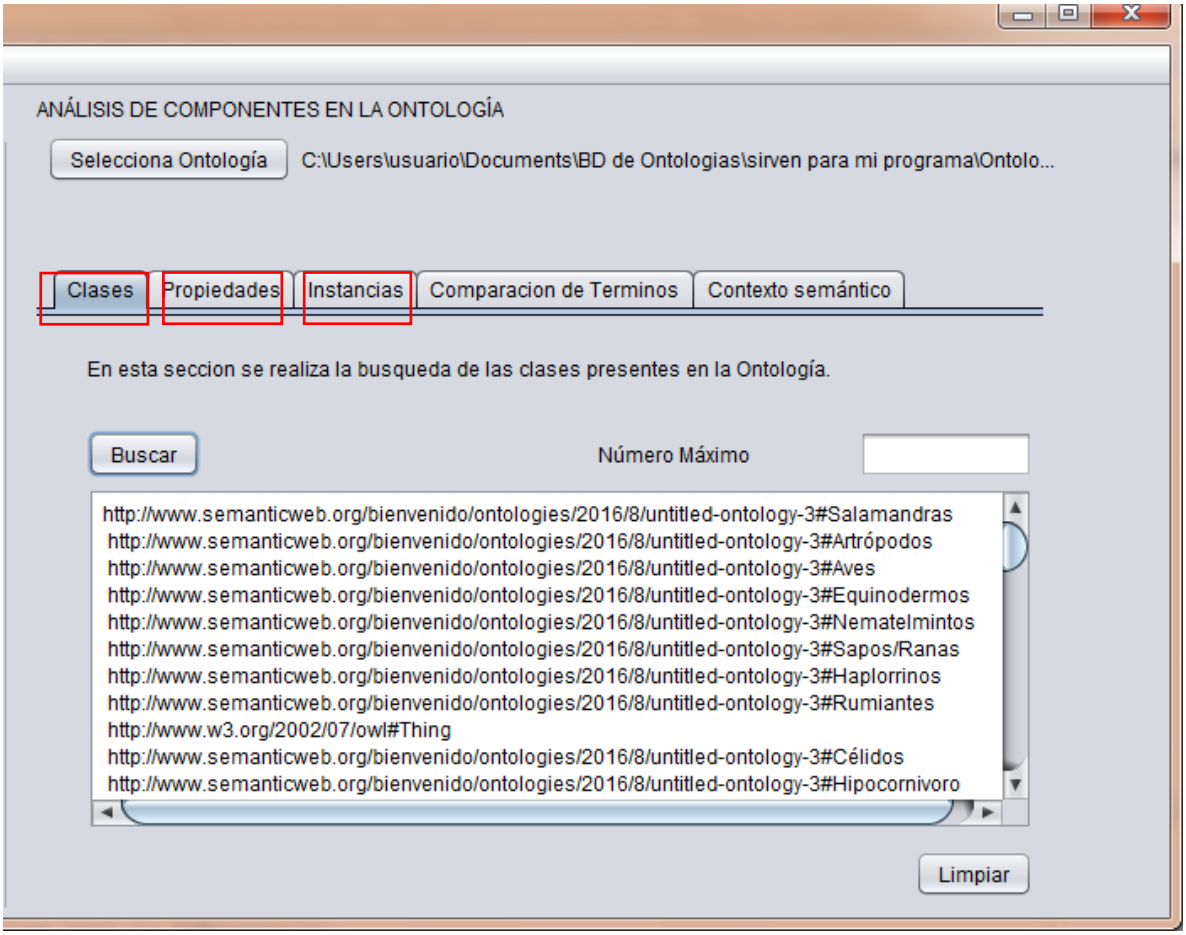

Figura 23. Módulo de análisis ontológico

5.4 Módulo para obtener el contexto semántico y su visualización

La figura 24, muestra el "Módulo extracción de contexto semántico y su visualización" donde el usuario podrá:

- visualizar los términos coincidentes
- agregar un elemento coincidente y determinar su contexto semántico,
- visualizar el contexto semántico del término seleccionado.

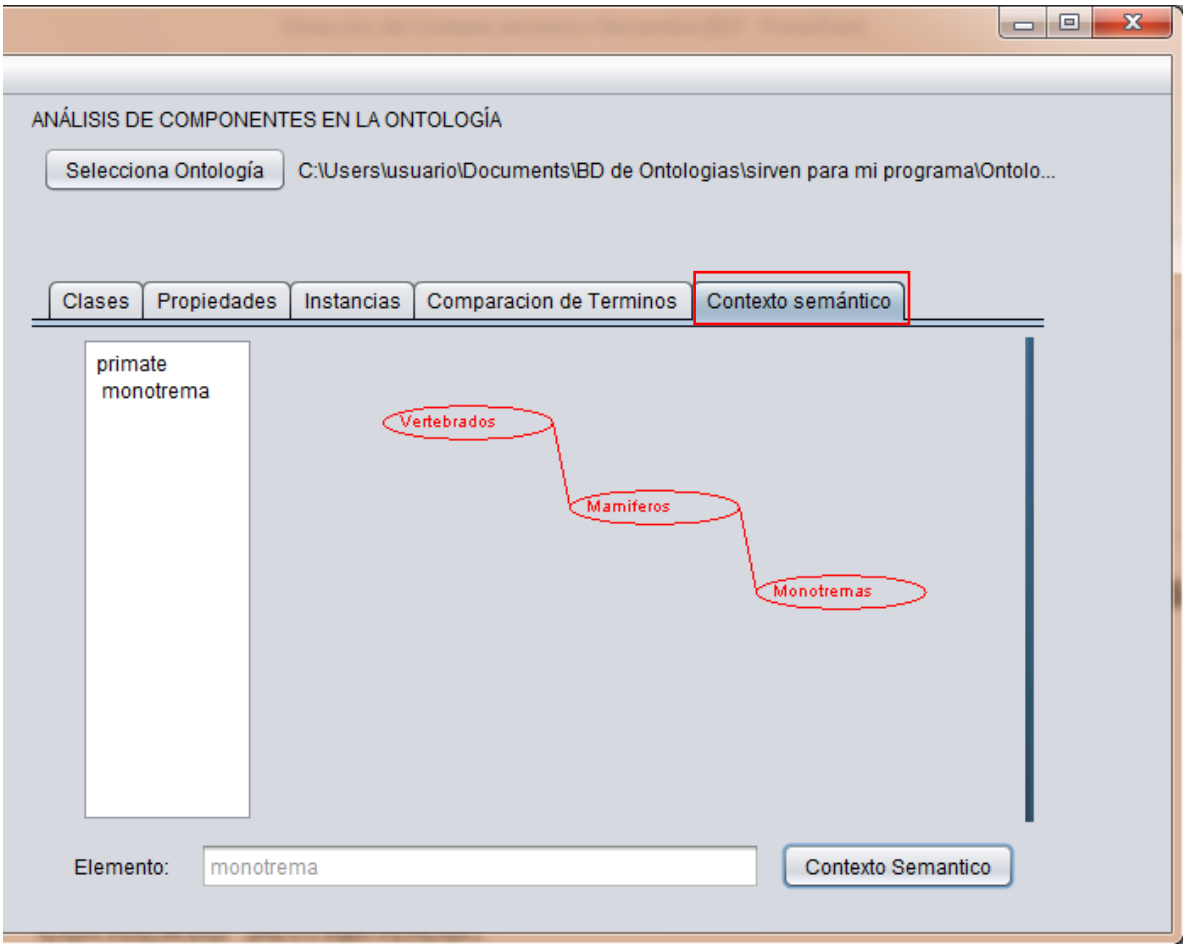

Figura 24. Módulo para obtener el contexto semántico y su visualización

Capítulo 6. Pruebas y discusiones

En este capítulo, se presentan el plan utilizado para la evaluación de los resultados de similitud semántica y la comparación entre los métodos encontrados en el estado del arte y nuestro método híbrido para la extracción de contexto semántico. Para finalizar se realizan discusiones de nuestro método híbrido.

6.1 Evaluación de los resultados de similitud semántica de nuestro método

Para evaluar los resultados de similitud semántica se utilizaron 4 ontologías de dominio.

- Ontología de animales
- Ontología de aguas residuales[56]
- Ontología de Libros[56]
- Ontología de viajes [53]

Tres de las cuatro ontologías se encuentran en línea y están disponibles para su libre uso. La ontología de animales es una ontología diseñada y desarrollada en el **CENIDET**. Las ontologías que utiliza nuestra herramienta deberán tener base de conocimiento es decir deberán tener instancias. Cuando las ontologías son pequeñas pueden ser manipuladas y agregarse instancias con la herramienta de Protégé [50].

El valor de similitud semántica fue evaluado con los valores bien definidos que proporciona WordNet, se utilizó la aplicación en línea llamada WS4J [55]. En la figura 25, se muestra la interfaz de la herramienta en línea para obtener la similitud semántica entre pares de términos.

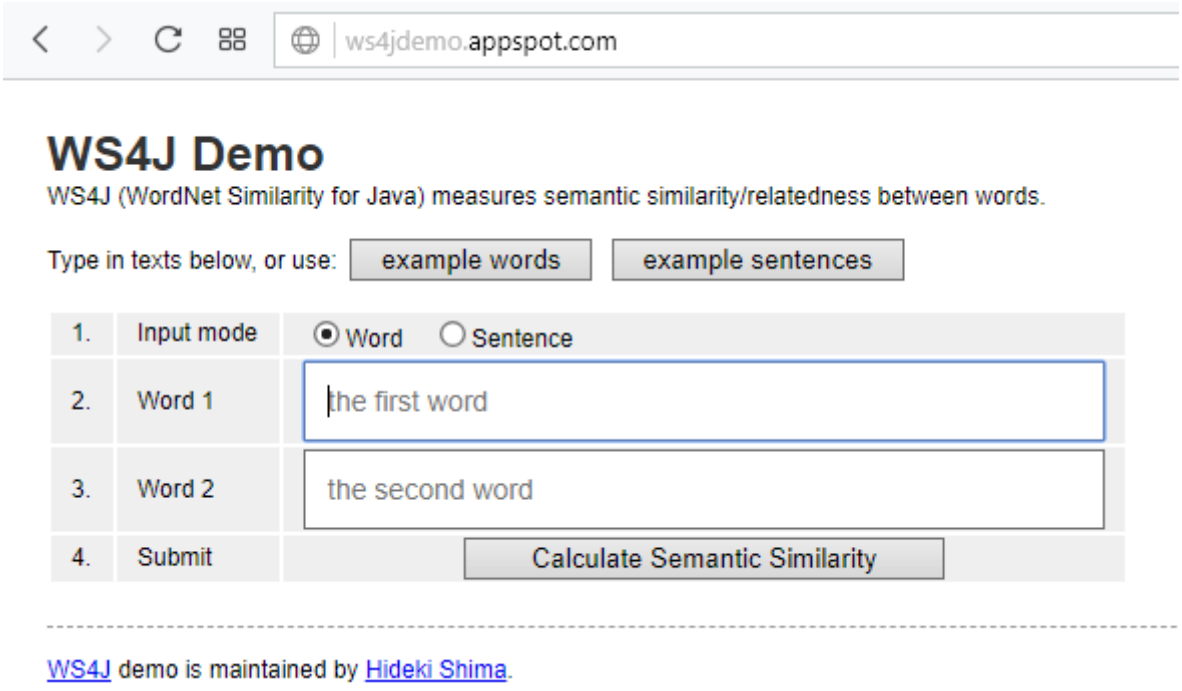

Figura 25. Interfaz de la herramienta para obtener la similitud semántica en WordNet

Para realizar la evaluación se contrasto el resultado de la similitud semántica de nuestro método, los términos utilizados individualmente para elaborar nuestro metodo: Resnik [36], Sánchez [48] y la similitud semántica obtenida de WordNet. Con esta información se formó una tabla comparativa y se realizó un análisis del porcentaje de términos en nuestro método que haya superado significativamente a los demás, los términos de nuestro método que están por debajo del grado de similitud en un grado menor y el porcentaje de términos de nuestro método que

están por debajo del grado de similitud en un grado alto. Además en la tabla se utiliza el siguiente código de colores para su fácil interpretación.

- Verde: términos en nuestro método que haya superado significativamente a los demás.
- Amarillo: los términos de nuestro método que están por debajo del grado de similitud en un grado menor.
- Rojo: el porcentaje de términos de nuestro método que están por debajo del grado de similitud en un grado alto.

Se realizó un cuadro comparativo con una muestra de términos de cada ontología por cada ontología y un análisis final.

En la Tabla 2 se muestra el análisis realizado a una muestra de términos de la ontología de animales.

Tabla 2. Análisis de ontología de animales

Ontología de animales				
Concepto1- Concepto2	**Similitud WordNet**	**sim_Sanchez**	**Sim_Resnik**	**Método Propuesto**
Anélidos-Platelmintos	0.8	0.63466305	0.65083314	0.83314701
Anélidos-Nematelmintos	0.8	0.63466305	0.65083314	0.83314701
Anélidos-Gusanos	0.85	0.7751792	0.65083314	0.94555993
Gusanos-Celentéreos	0.8	0.558815395	0.44658362	0.670344126
Gusanos-Moluscos	0.75	0.558815395	0.44658362	0.670344126
Gusanos-Equinodermos	0.8	0.558815395	0.44658362	0.670344126
Gusanos-Poríferos	0.8	0.558815395	0.44658362	0.670344126
Gusanos-Invertebrados	0.85	0.682538664	0.44658362	0.769322741
Monotremas-Terrestres	0.7	0.63466305	0.31792164	0.66669126
Monotremas-Marinos	0.7	0.63466305	0.31792164	0.66669126
Monotremas-Voladores	0.5	0.63466305	0.31792164	0.66669126
Monotremas-Mamíferos	0.7	0.7751792	0.31792164	0.779104179
Vertebrados-Invertebrados	0.85	0.35999509	0.0197647	0.297878421
Vertebrados-Reino-animal	0.85	0.439698996	0.0197647	0.361641545
Primates-Carnívoros	0.7	0.659376976	0.39604056	0.725521862
Primates-Herbívoros	0.8	0.659376976	0.39604056	0.725521862
Primates-Terrestres	0.8	0.805364857	0.39604056	0.842312167
Peces-Anfibios	0.85	0.63466305	0.21026104	0.612860959
Peces-Aves	0.5	0.63466305	0.21026104	0.612860959
Peces-Reptiles	0.5	0.63466305	0.21026104	0.612860959
Peces-Artrópodos	0.5	0.63466305	0.21026104	0.612860959

En la Tabla 3, se muestra el análisis realizado a una muestra de términos de la ontología de aguas residuales [56].

Tabla 3. Análisis de la ontología de Residuos

Ontología de Residuos				
Concepto1- Concepto2	**Valor humano**	**sim_Sanchez**	**Sim_Resnik**	**Método Propuesto**
EnviromentalValue-SocialValue	0.75	0.35999509	0.745279684	0.660635914
EnviromentalValue-EconomicValue	0.75	0.35999509	0.745279684	0.660635914
EnviromentalValue-Value	0.75	0.439698996	0.745279684	0.724399039
OPenRange-ClosedRange	0.75	0.35999509	0.798138925	0.687065535
OPenRange-Range	0.75	0.439698996	0.798138925	0.75082866
Not-Or	0.57	0.35999509	0.617919525	0.596955835
Not-And	0.57	0.35999509	0.617919525	0.596955835
Not-Quantifier	0.57	0.35999509	0.617919525	0.596955835
Not-Implication	0.57	0.35999509	0.617919525	0.596955835
ChemicalFeature-Radioactivity	0.52	0.35999509	0.308910062	0.442451104
ChemicalFeature-IndicatorParameters	0.52	0.35999509	0.308910062	0.442451104
ChemicalFeature-Microbiological	0.52	0.35999509	0.308910062	0.442451104
ChemicalFeature-WaterFeatures	0.52	0.439698996	0.308910062	0.506214228
DeonticComponent-ConstitutiveComponent	0.52	0.35999509	0.559406242	0.567699193
DeonticComponent-NormComponent	0.52	0.439698996	0.559406242	0.631462318

En la Tabla 4, se muestra el análisis realizado a una muestra de términos de la ontología libros [56].

Tabla 4. Análisis de la ontología de Residuos

Ontología de libros				
Concepto1- Concepto2	**Valor humano**	**sim_Sanchez**	**Sim_Resnik**	**Método Propuesto**
AsstLibrarian-Librarian	0.47	0.35999509	0.563705474	0.569848809
AsstLibrarian-Technician	0.47	0.35999509	0.563705474	0.569848809
AsstLibrarian-JrAssistant	0.47	0.35999509	0.563705474	0.569848809
AsstLibrarian-LibraryPersonnel	0.47	0.439698996	0.563705474	0.633611934
CD-Book	0.6	0.35999509	0.326379005	0.451185575
CD-Journal	0.6	0.35999509	0.326379005	0.451185575
CD-NewsPaper	0.6	0.35999509	0.326379005	0.451185575
CD-Thesis	0.6	0.35999509	0.326379005	0.451185575
CD-OnlineJournal	0.6	0.35999509	0.326379005	0.451185575
CD-LibraryResource	0.6	0.439698996	0.326379005	0.514948699
ReprographicService-CurretAwarenessService	0.5	0.35999509	0.472492888	0.524242516
ReprographicService-ReferenceService	0.5	0.35999509	0.472492888	0.524242516
ReprographicService-InternetAndWiFiService	0.5	0.35999509	0.472492888	0.524242516
ReprographicService-NewsPaperService	0.5	0.35999509	0.472492888	0.524242516
ReprographicService-LendingService	0.5	0.35999509	0.472492888	0.524242516
ReprographicService-LibraryService	0.5	0.439698996	0.472492888	0.588005641
Faculty-ResearchScholar	0.58	0.35999509	0.436294526	0.506143335
Faculty-GuestUser	0.3	0.35999509	0.436294526	0.506143335
Faculty-AdminStaff	0.8	0.35999509	0.436294526	0.506143335
Faculty-Student	0.28	0.35999509	0.436294526	0.506143335
Faculty-LibraryMember	0.61	0.439698996	0.436294526	0.56990646

En la Tabla 5, se muestra el análisis realizado a una muestra de términos de la ontología de viaje [53].

Tabla 5. Análisis de ontología de viaje

Ontología de viaje				
Concepto1- Concepto2	**Valor humano**	**sim_Sanchez**	**Sim_Resnik**	**Método Propuesto**
Primera etapa-Segunda etapa	0.8	0.558815395	0.61608256	0.816701852
Primera etapa-Tercera etapa	0.8	0.558815395	0.61608256	0.816701852
Primera etapa-Etapa de vuelo	0.85	0.682538664	0.61608256	0.915680468
Primera etapa-Vuelo	0.5	0.35999509	0.55881539	0.335289237
Primera estancia-Estancia	0.85	0.682538664	0.61608256	0.915680468
Primera estancia-Segunda estancia	0.8	0.439698996	0.69575263	0.769210773
Segundo alquile-Primer alquiler	0.85	0.35999509	0.69575263	0.705447649
Segundo alquile-Alquiler de coche	8	0.439698996	0.69575263	0.769210773
Ciudad-Pais	0.8	0.35999509	0.55428565	0.620567465
Ciudad-Aeropuerto	0.5	0.35999509	0.55428565	0.620567465
Ciudad-Estado federal	0.7	0.35999509	0.55428565	0.620567465
Ciudad-Lugar	0.7	0.439698996	0.55428565	0.68433059

Resultados de los análisis anteriores

Las pruebas realizadas en 4 ontologías con una muestra que el 58% supera el valor de similitud semántica de los método de Resnik[36] y Sánchez[48] y el valor obtenido con WordNet; 33% está por debajo del valor obtenido en WordNet pero en un grado pequeño y 9% está por debajo del valor humano en un grado alto. Cabe mencionar que el 33 % como se muestra en la figura 26 aunque se encuentre debajo de la similitud de WordNet no quiere decir que lo esté en comparación con los métodos: Resnik [36] y Sánchez [48].

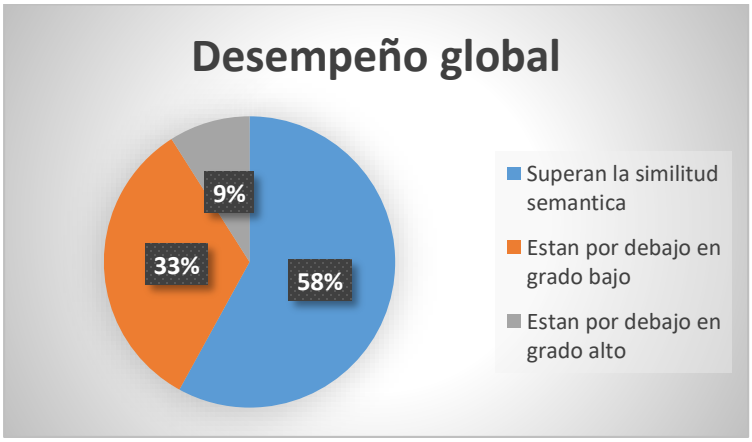

Figura 26. Desempeño global de nuestro método híbrido frente a la similitud semántica de *WordNet*

6.2 Comparar los resultados obtenidos con otros métodos de la literatura

Para evaluar los resultados de similitud semántica y compararlos con los otros métodos de la literatura se utilizaron 4 ontologías de dominio.

- Ontología de animales
- Ontología de aguas residuales[56]
- Ontología de Libros[56]
- Ontología de viajes [53]

Tres de las cuatro ontologías se encuentran en línea y están disponibles para su libre uso. La ontología de animales es una ontología diseñada y desarrollada en el **CENIDET**. Las ontologías que utiliza nuestra herramienta deberán tener base de conocimiento es decir deberán tener instancias. Cuando las ontologías son pequeñas pueden ser manipuladas y agregarse instancias con la herramienta de *Protégé* [50].

Para realizar la comparación de nuestro método con los de la literatura se utilizaron 4 métodos que se encuentran en la literatura:

- Similitud semántica de Resnick [36]
- Similitud semántica de Sánchez [48]
- Similitud semántica de Seco [19]
- Similitud semántica de Zhou [57]

Además en las tablas se utiliza el siguiente código de colores para su fácil interpretación.

- Verde: términos en nuestro método que haya superado significativamente a los demás.
- Amarillo: los términos de nuestro método que están por debajo del grado de similitud en un grado menor.
- Rojo: el porcentaje de términos de nuestro método que están por debajo del grado de similitud en un grado alto.

En la siguiente sección se presentara el cuadro comparativo con una muestra de términos de cada ontología y un gráfico que representa el análisis de desempeño de nuestro método frente a los métodos hallados en la literatura.

La Tabla 6, muestra la comparación de los resultados de nuestro método frente a cuatro métodos existentes en la literatura utilizando los términos de la ontología de animales.

Tabla 6. Tabla comparativa de términos de la ontología de animales y los métodos en la literatura

Ontología de animales					
Termino1-Termino2	Resnik	Sánchez	Seco	Zhou	Método Propuesto
Anélidos-Platelmintos	0.65083314	0.63466305	0.662773803	0.687594088	0.837998037
Anélidos-Nematelmintos	0.65083314	0.63466305	0.662773803	0.687594088	0.837998037
Anélidos-Gusanos	0.65083314	0.7751792	0.662773803	0.687594088	0.908256112
Gusanos-Celentéreos	0.44658362	0.558815395	0.465509123	0.515042078	0.636674593
Gusanos-Moluscos	0.44658362	0.558815395	0.465509123	0.515042078	0.636674593
Gusanos-Equinodermos	0.44658362	0.558815395	0.465509123	0.515042078	0.636674593
Gusanos-Poríferos	0.44658362	0.558815395	0.465509123	0.515042078	0.636674593
Gusanos-Invertebrados	0.44658362	0.682538664	0.465509123	0.515042078	0.698536228
Monotremas-Terrestres	0.317921639	0.63466305	0.34124707	0.526830722	0.571668836
Monotremas-Marinos	0.317921639	0.63466305	0.34124707	0.526830722	0.571668836
Monotremas-Voladores	0.317921639	0.63466305	0.34124707	0.526830722	0.571668836
Monotremas-Mamíferos	0.317921639	0.7751792	0.34124707	0.526830722	0.641926911
Vertebrados-Invertebrados	0.019764697	0.35999509	0.053286375	0.204746781	0.195809303
Vertebrados-Reino_animal	0.019764697	0.439698996	0.053286375	0.204746781	0.235661256
Primates-Carnivoros	0.396040562	0.659376976	0.416694515	0.621890995	0.646520938
Primates-Hervívoros	0.396040562	0.659376976	0.416694515	0.621890995	0.646520938
Primates-Terrestres	0.396040562	0.805364857	0.416694515	0.621890995	0.719514878
Peces-Anfibios	0.210261038	0.63466305	0.237268201	0.474841288	0.485540355
Peces-Aves	0.210261038	0.63466305	0.237268201	0.474841288	0.485540355
Peces-Reptiles	0.210261038	0.63466305	0.237268201	0.474841288	0.485540355

En la figura 27, se muestra el desempeño de nuestro método aplicado a la ontología de animales y comparado con 4 métodos encontrados en la literatura.

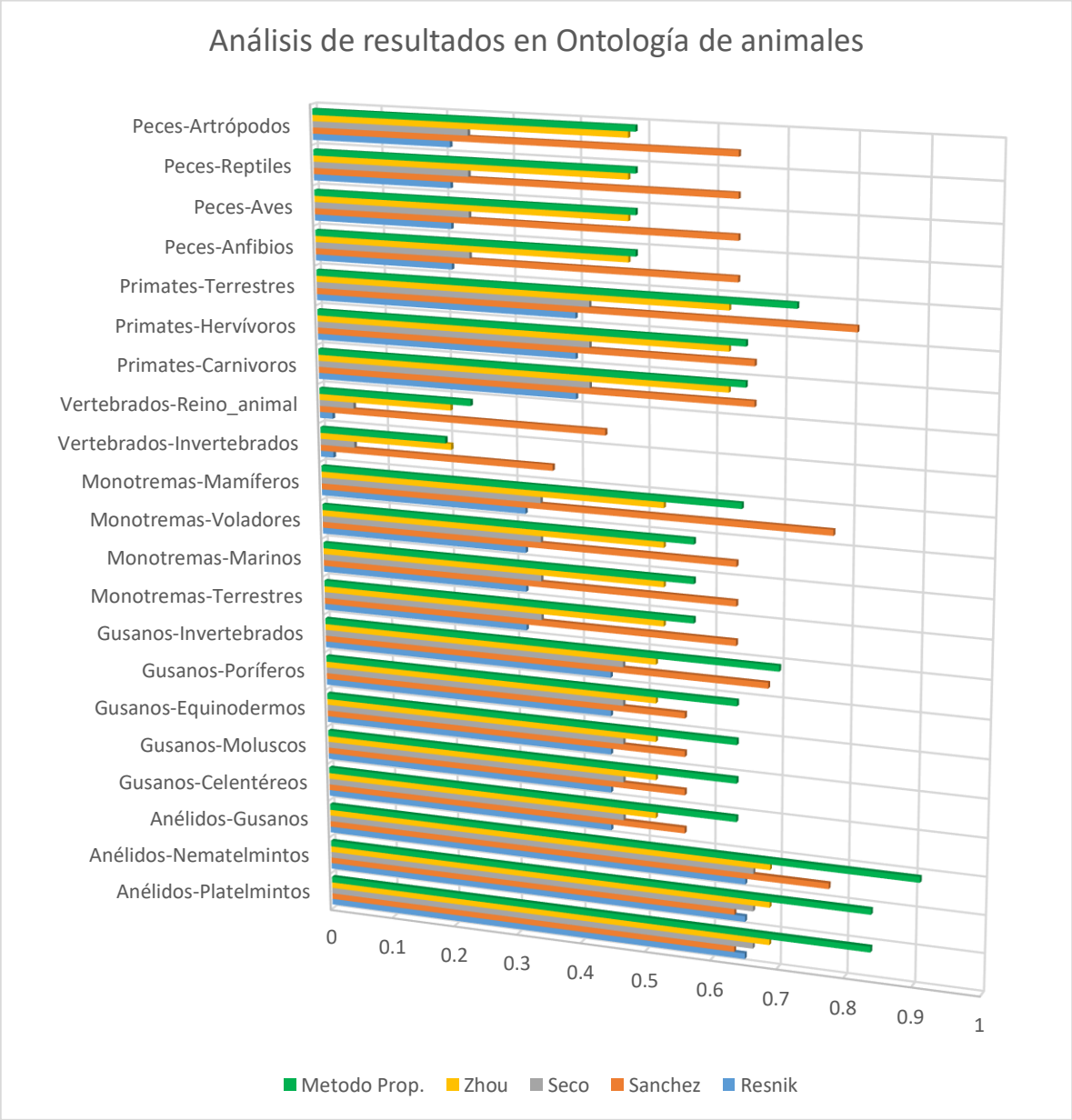

Figura 27. Desempeño de nuestro método con la ontología de animales frente a los métodos en la literatura

La Tabla 7, muestra la comparación de los resultados de nuestro método frente a cuatro métodos existentes en la literatura utilizando los términos de la ontología de aguas residuales.

Tabla 7. Tabla comparativa de términos de la ontología de aguas residuales y los métodos en la literatura

Ontología de aguas residuales					
Termino1-Termino2	Resnik	Sánchez	Seco	Zhou	Método Propuesto
EnviromentalValue-SocialValue	0.745279684	0.35999509	0.753186602	0.554696894	0.776221292
EnviromentalValue-EconomicValue	0.745279684	0.35999509	0.753186602	0.554696894	0.776221292
EnviromentalValue-Value	0.745279684	0.439698996	0.753186602	0.554696894	0.816073245
OPenRange-ClosedRange	0.798138925	0.35999509	0.80440501	0.580306098	0.818508685
OPenRange-Range	0.798138925	0.439698996	0.80440501	0.580306098	0.858360638
Not-Or	0.617919525	0.35999509	0.629779903	0.492993545	0.674333166
Not-And	0.617919525	0.35999509	0.629779903	0.492993545	0.674333166
Not-Quantifier	0.617919525	0.35999509	0.629779903	0.492993545	0.674333166
Not-Implication	0.617919525	0.35999509	0.629779903	0.492993545	0.674333166
ChemicalFeature-Radioactivity	0.308910062	0.35999509	0.330362578	0.343284883	0.427125595
ChemicalFeature-IndicatorParameters	0.308910062	0.35999509	0.330362578	0.343284883	0.427125595
ChemicalFeature-Microbiological	0.308910062	0.35999509	0.330362578	0.343284883	0.427125595
ChemicalFeature-WaterFeatures	0.308910062	0.439698996	0.330362578	0.343284883	0.466977548
DeonticComponent-ConstitutiveComponent	0.559406242	0.35999509	0.573082963	0.464645075	0.627522539
DeonticComponent-NormComponent	0.559406242	0.439698996	0.573082963	0.464645075	0.667374492

En la figura 28, se muestra el desempeño de nuestro método aplicado a la ontología de aguas residuales y comparadas con 4 métodos encontrados en la literatura.

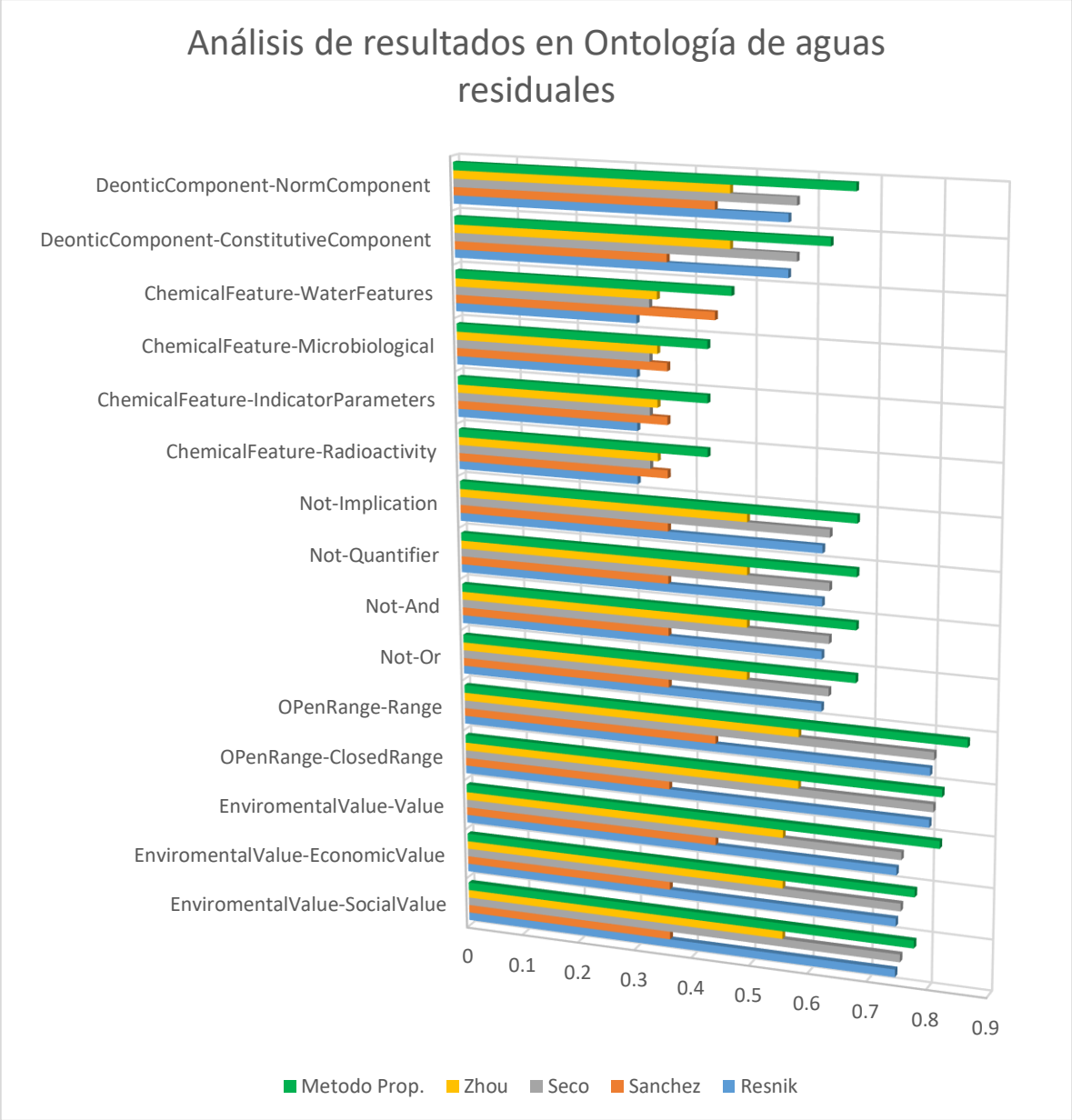

Figura 28. Desempeño de nuestro método con la ontología de aguas residuales frente a los métodos en la literatura.

La Tabla 8, muestra la comparación de los resultados de nuestro método frente a cuatro métodos existentes en la literatura utilizando los términos de la ontología de libros.

Tabla 8. Tabla comparativa de términos de la ontología de aguas residuales y los métodos en la literatura

Ontología de libros					
Termino1-Termino2	Resnik	Sánchez	Seco	Zhou	Método Propuesto
AsstLibrarian-Librarian	0.56370547	0.35999509	0.663824399	0.547250479	0.630961925
AsstLibrarian-Technician	0.56370547	0.35999509	0.663824399	0.547250479	0.630961925
AsstLibrarian-JrAssistant	0.56370547	0.35999509	0.663824399	0.547250479	0.630961925
AsstLibrarian-LibraryPersonnel	0.56370547	0.439699	0.663824399	0.547250479	0.670813878
CD-Book	0.326379	0.35999509	0.480958551	0.455817555	0.441100749
CD-Journal	0.326379	0.35999509	0.480958551	0.455817555	0.441100749
CD-NewsPaper	0.326379	0.35999509	0.480958551	0.455817555	0.441100749
CD-Thesis	0.326379	0.35999509	0.480958551	0.455817555	0.441100749
CD-OnlineJournal	0.326379	0.35999509	0.480958551	0.455817555	0.441100749
CD-LibraryResource	0.326379	0.439699	0.480958551	0.455817555	0.480952702
ReprographicService-CurretAwarenessService	0.47249289	0.35999509	0.593542871	0.512109715	0.557991855
ReprographicService-ReferenceService	0.47249289	0.35999509	0.593542871	0.512109715	0.557991855
ReprographicService-InternetAndWiFiService	0.47249289	0.35999509	0.593542871	0.512109715	0.557991855
ReprographicService-NewsPaperService	0.47249289	0.35999509	0.593542871	0.512109715	0.557991855
ReprographicService-LendingService	0.47249289	0.35999509	0.593542871	0.512109715	0.557991855
ReprographicService-LibraryService	0.47249289	0.439699	0.593542871	0.512109715	0.597843808
Faculty-ResearchScholar	0.43629453	0.35999509	0.565651148	0.498163853	0.529033166
Faculty-GuestUser	0.43629453	0.35999509	0.565651148	0.498163853	0.529033166
Faculty-AdminStaff	0.43629453	0.35999509	0.565651148	0.498163853	0.529033166
Faculty-Student	0.43629453	0.35999509	0.565651148	0.498163853	0.529033166
Faculty-LibraryMember	0.43629453	0.439699	0.565651148	0.498163853	0.568885119

En la figura 29, se muestra el desempeño de nuestro método aplicado a la ontología de libros y comparado con 4 métodos encontrados en la literatura.

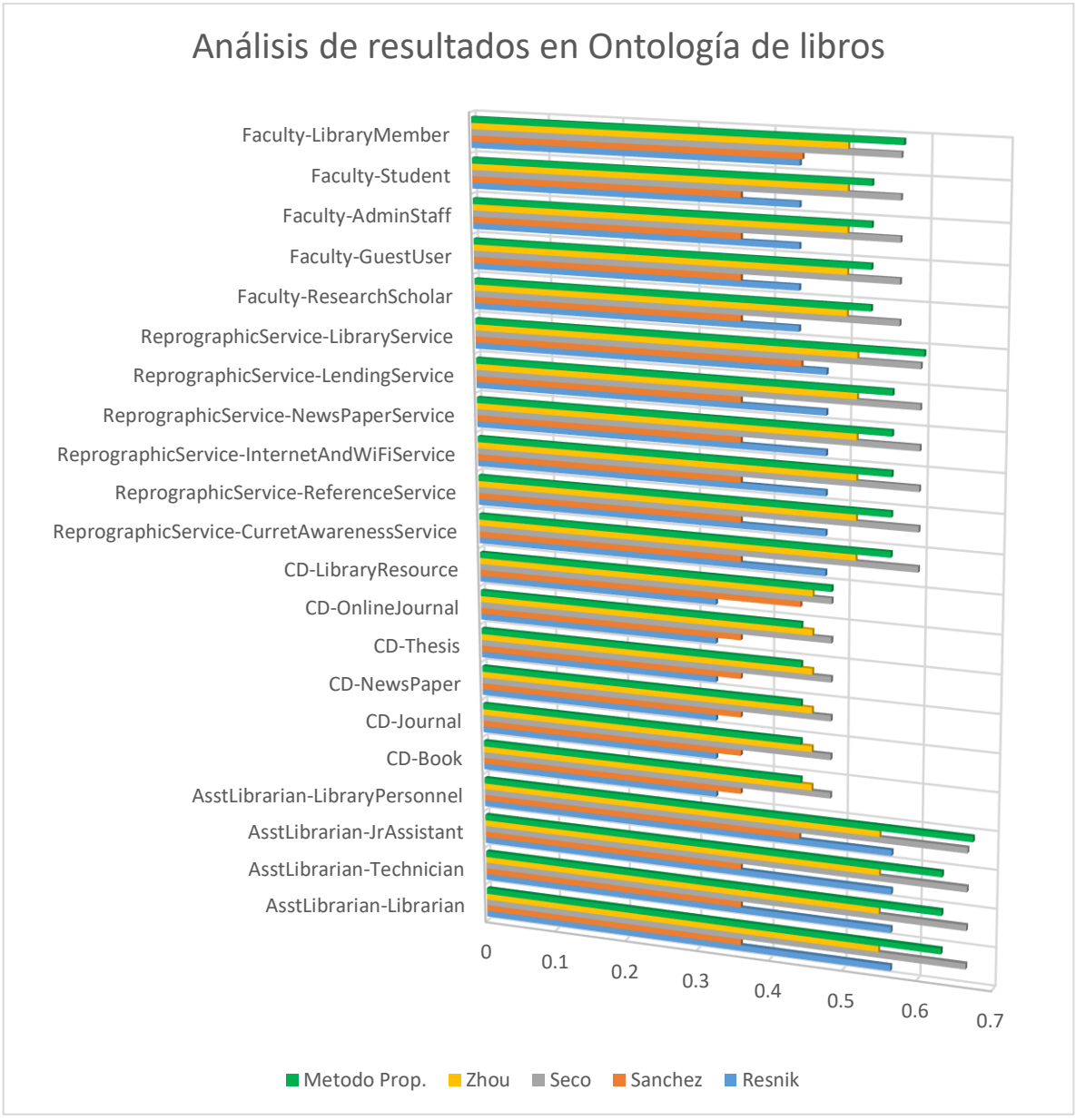

Figura 29. Desempeño de nuestro método con la ontología de libros frente a los métodos en la literatura

La Tabla 9, muestra la comparación de los resultados de nuestro método frente a cuatro métodos existentes en la literatura utilizando los términos de la ontología de viajes.

Tabla 9. Tabla comparativa de términos de la ontología de viajes y los métodos en la literatura

Ontología de viajes					
Termino1-Termino2	Resnik	Sánchez	Seco	Zhou	Método Propuesto
Primera etapa-Segunda etapa	0.5588154	0.61608256	0.702731325	0.747606288	0.816701852
Primera etapa-Tercera etapa	0.5588154	0.61608256	0.702731325	0.747606288	0.816701852
Primera etapa-Etapa de vuelo	0.68253866	0.61608256	0.702731325	0.747606288	0.915680468
Primera etapa-Vuelo	0.35999509	0.55881539	0.702731325	0.747606288	0.335289237
Primera estancia-Estancia	0.68253866	0.61608256	0.764420149	0.632210074	0.915680468
Primera estancia-Segunda estancia	0.439699	0.69575263	0.764420149	0.632210074	0.769210773
Segundo alquile-Primer alquiler	0.35999509	0.69575263	0.764420149	0.632210074	0.705447649
Segundo alquile-Alquiler de coche	0.439699	0.69575263	0.764420149	0.632210074	0.769210773
Ciudad-Pais	0.35999509	0.55428565	0.654881756	0.577440878	0.620567465
Ciudad-Aeropuerto	0.35999509	0.55428565	0.654881756	0.577440878	0.620567465
Ciudad-Estado federal	0.35999509	0.55428565	0.654881756	0.577440878	0.620567465
Ciudad-Lugar	0.439699	0.55428565	0.654881756	0.577440878	0.68433059

En la figura 30, se muestra el desempeño de nuestro método aplicado a la ontología de viajes y comparado con 4 métodos encontrados en la literatura.

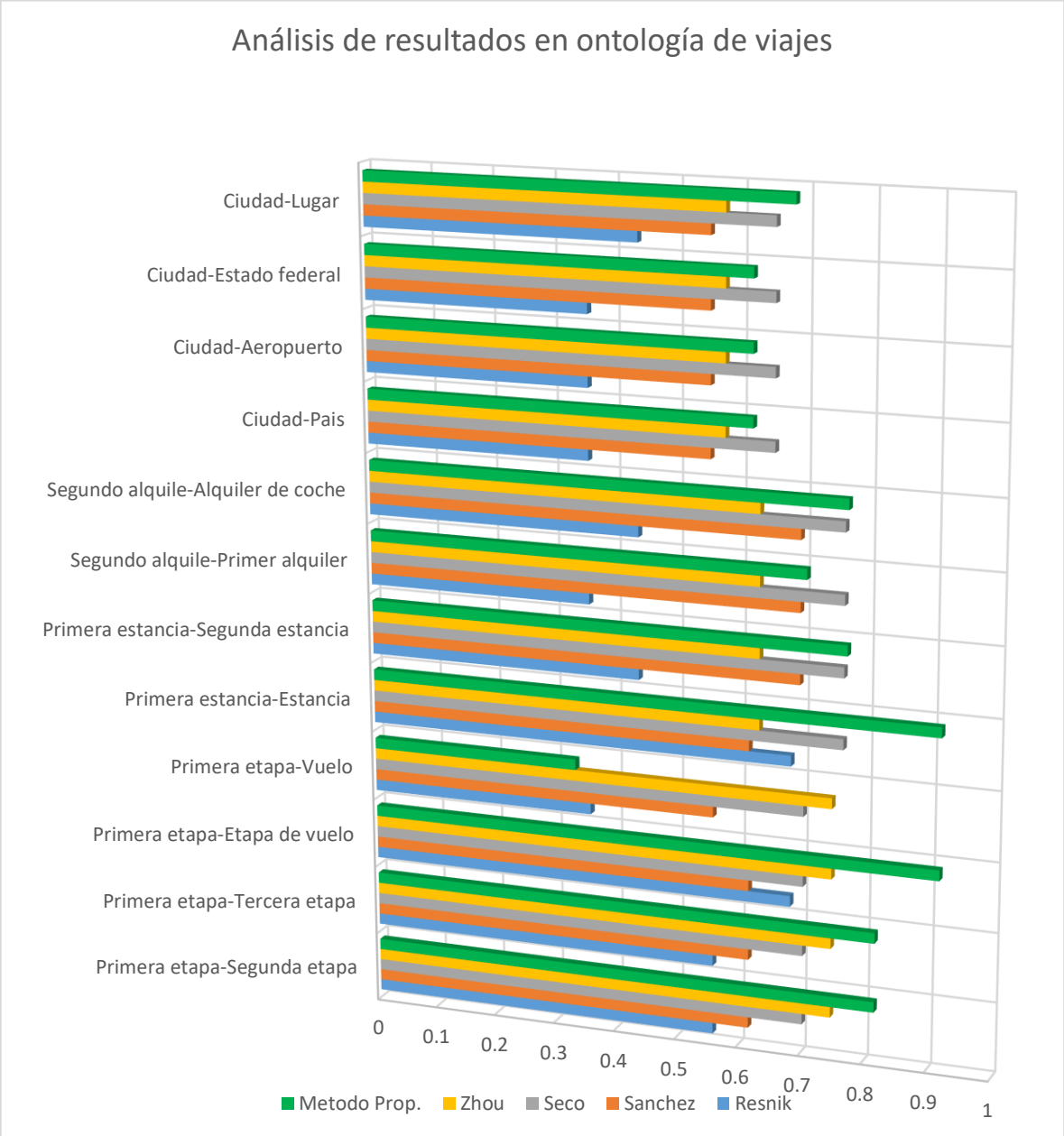

Figura 30. Desempeño de nuestro método con la ontología de viajes frente a los métodos en la literatura

Resultados de los análisis anteriores

La figura 31 muestra que nuestro método logró superar el total de los métodos propuestos en la literatura en un 38% de términos analizados, un 52% obtuvo un grado de similitud inferior pero de bajo grado en relación a alguno de los métodos y un 10% obtuvo un grado de similitud inferior en un grado alto respecto a los métodos de la literatura.

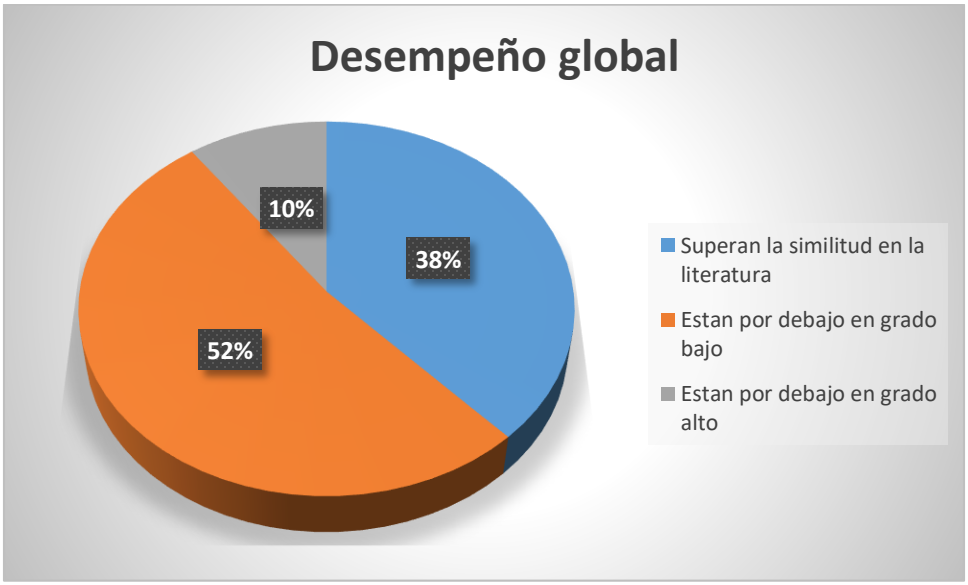

Figura 31. Desempeño global de nuestro método híbrido respecto a la similitud semántica en la literatura

Para calcular el porcentaje de similitud utilizado en cada tabla, primero se determinó la media de todos los valores, después de se calculó la desviación estándar de la muestra. Por último se calcularon el índice superior e índice superior. De esta manera se obtuvieron los rangos para poder seleccionar aquellos que encuentra dentro del promedio del valor de similitud. De lo anterior se obtuvieron los índices:

- Índice superior: 0.694111352
- Índice inferior: 0.558944978

Los índices tienen un grado de confianza del 95% y un α= 5%

6.3 Discusiones

En la presente tesis se muestra que es factible utilizar más de un método híbrido de similitud semántica de manera sencilla pero eficaz para recuperar el contexto semántico de términos relevantes en textos no estructurados. Y aunque el desempeño global de frente a los cuatro método de similitud encontrados en la literatura es del 52 %, cuando se contrasta uno por uno puede notarse la diferencia de desempeño como se muestra en la figura 32.
Tomando como ejemplo la ontología de animales, nuestro método hibrido:

- Supera en el 95% de términos la similitud propuesta por Zhou.
- Supera en el 100% de términos la similitud propuesta por Seco.
- Supera en el 90% de términos la similitud propuesta por Resnik.
- Supera en el 38% de términos la similitud propuesta por Sánchez.

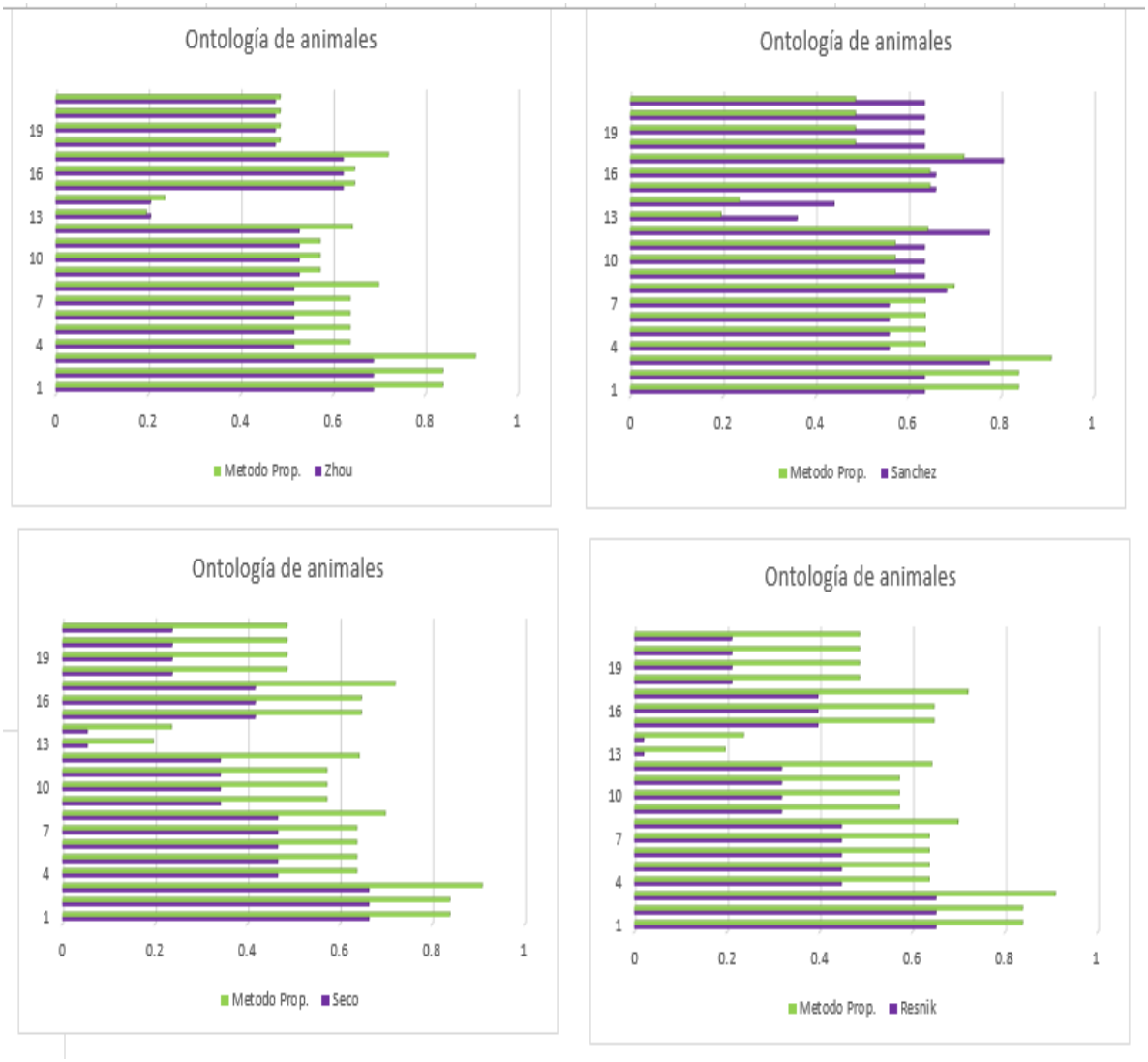

Figura 32. Desempeño de nuestro método hibrido utilizando la ontología de animales

El siguiente contraste individual entre nuestro método hibrido y los método de la literatura se encuentra en la figura 33. Tomando como ejemplo la ontología de viajes, nuestro método hibrido:

- Supera en el 95% de términos la similitud propuesta por Zhou.
- Supera en el 63% de términos la similitud propuesta por Seco.
- Supera en el 95% de términos la similitud propuesta por Resnik.
- Supera en el 95% de términos la similitud propuesta por Sánchez.

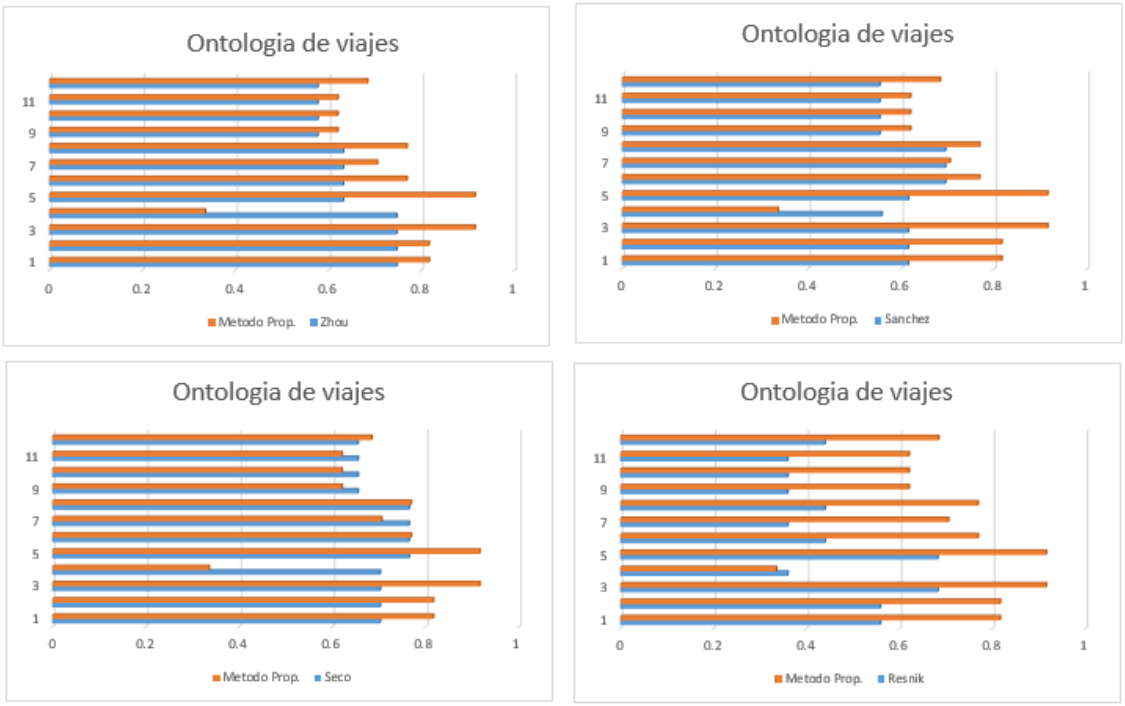

Figura 33. Desempeño de nuestro método hibrido utilizando la ontología de viajes

Otro contraste individual entre nuestro método hibrido y los método de la literatura puede hallarse en la figura 34. Tomando como ejemplo la ontología de residuos, nuestro método hibrido:

- Supera en el 100% de los términos en todos los métodos encontrados en la literatura.

El ultimo contraste individual entre nuestro método hibrido y los método de la literatura que compara uno por uno para apreciar la diferencia de desempeño como se muestra en la figura 36.

- Supera en el 81% de términos la similitud propuesta por Zhou.
- Supera en el 100% de términos la similitud propuesta por Seco.
- Supera en el 100% de términos la similitud propuesta por Resnik.
- Supera en el 95% de términos la similitud propuesta por Sánchez.

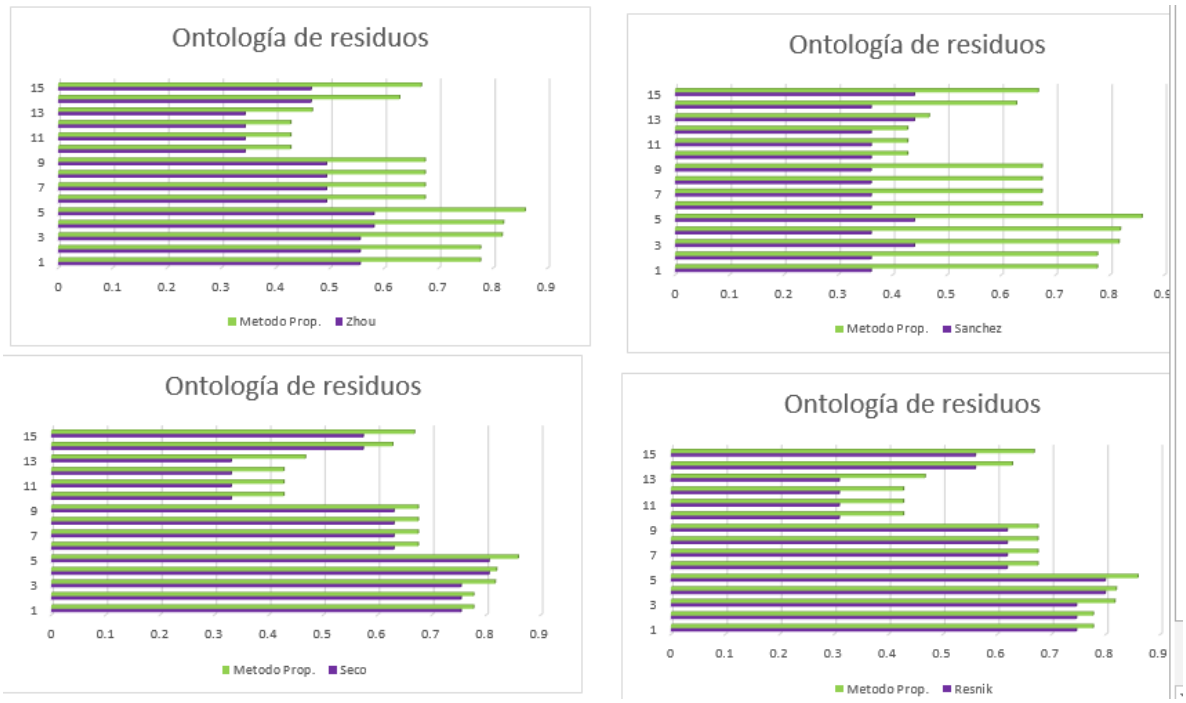

Figura 34. Desempeño de nuestro método hibrido utilizando la ontología de residuos

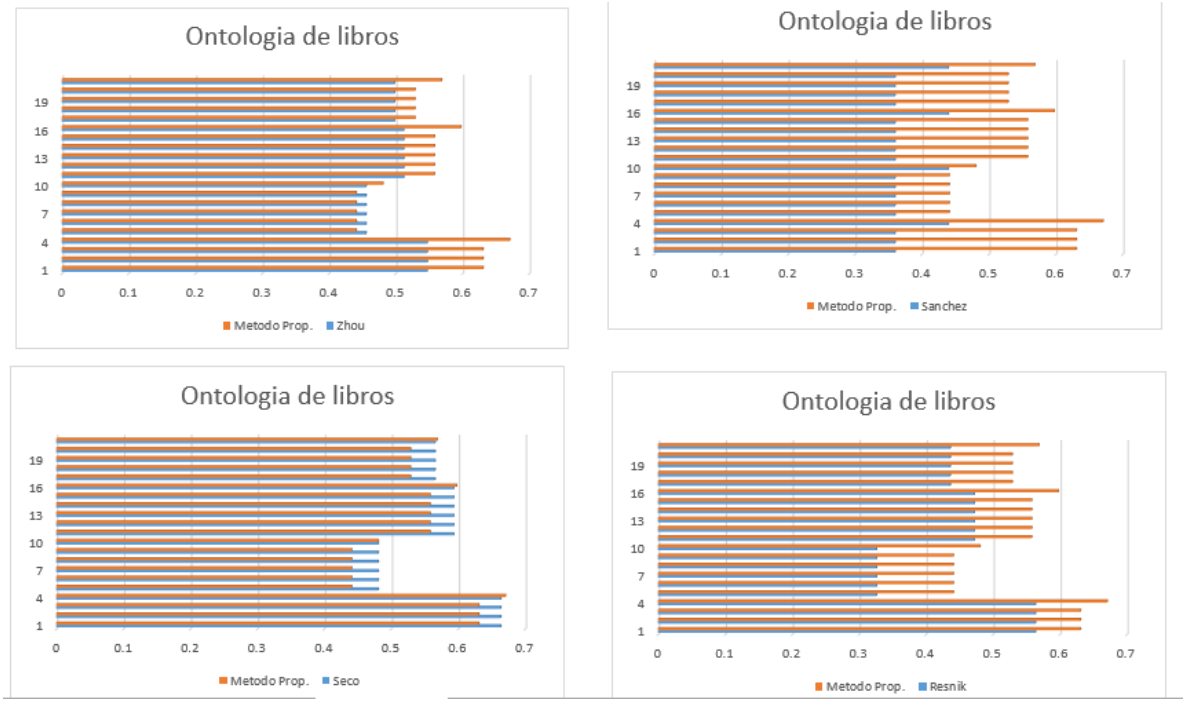

Figura 35. Desempeño de nuestro método hibrido utilizando la ontología de libros

Podemos discutir que cuando medimos el desempeño global contabilizado los términos efectico solo cuando el índice se similitud semántica de nuestro método hibrido superaba al total de los métodos en la literatura, es cuando el desempeño global es bajo y no necesariamente sea así. También en algunas ocasiones se observa que el desempeño de nuestro método hibrido frente al método de similitud propuesto por Sánchez [48] es bajo y esto debido a que Sánchez toma en cuenta la base de conocimiento de la ontología y cuanto más grande esté la base de conocimiento mayor será el incremento en su índice de similitud.

Si bien, se pueden realizar índices de confianza para saber si el promedio de similitud está dentro de un par de índices de confianza, el número de error sería mayor ya que las muestras con las que se realizaron pruebas son pequeñas. Por ello se recomienda que se amplié el tamaño de muestra para realizar este tipo de análisis.

Con los resultados obtenidos podemos concluir que nuestro método hibrido supera el valor de similitud para los métodos propuestos en la literatura.

Otro aspecto importante que hay que discutir es que en las ontologías de aguas residuales y de libros, están contenidos términos compuestos por dos o términos, y como menciona Savary [58] el significado de este término no sería el mismo si solo se tomara uno de los componentes de este término. Por ello se recomienda el análisis y estudio más profundo de términos multipalabra para poder ser incluidos en el método hibrido.

También podemos discutir la manera en que nuestro método hibrido ataca la ambigüedad de términos, en la figura 36 se muestra como nuestro método puede distinguir la diferencia de términos aunque el grado de similitud sea la misma. Logra distinguir una diferencia cuando se anexa un segundo método de similitud y es a este método al que se le asigna un peso mayor.

Termino principal	Terminos Similares	Similitud Semantica Sanchez	Similitud Semantica Resnik	Ponderación de terminos
Ciudad	Country	0.35999509	0.554285655	0.5651389
	Aeropuerto	0.35999509	0.554285655	0.5651389
	Estado Federal	0.35999509	0.554285655	0.5651389
	Lugar	0.439698996	0.554285655	0.628902025
	Ponderación de criterios	0.8	0.5	

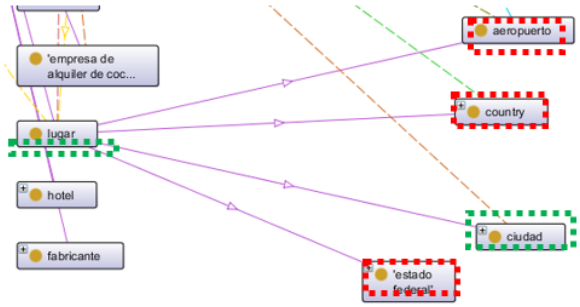

Figura 36. Problemas de la ambigüedad de términos

Capítulo 7.Conclusiones y trabajos futuros

En esta sección se muestran las conclusiones obtenidas a partir de la investigación realizada para esta tesis y los trabajos futuros.

7.1 Conclusiones

En la presente tesis se propuso utilizar un método híbrido de similitud semántica que toma en cuenta el peso de las aristas y las relaciones entre nodos para recuperar el contexto semántico de términos relevantes en textos no estructurados. Dicho método utiliza la ontología como herramienta que proporciona una estructura semántica excelente y la similitud entre términos para obtener aquel par de términos que guarde una similitud mayor.

Nuestro trabajo de tesis está enmarcado en el proyecto doctoral *"Un nuevo enfoque de anotación semántica para una Arquitectura Integral de Búsqueda de Información basada en ontologías"* y aprovechó el reto aun existente en la búsqueda de información: evitar la limitación de resultados no pertinentes en la búsqueda por palabras clave, por ello, afrontamos el reto utilizar la semántica de las palabras para la recuperación de información importante en documentos electrónicos no estructurados.

En el capítulo tres se mostró un pequeño resumen sobre la diferencia que guardan otros métodos híbridos para obtener la similitud semántica entre palabras y se muestran algunas de sus características para hacer notar que los trabajos desarrollados, si bien converge en muchas de los métodos para determinar la similitud semántica y el uso de ontologías, guardan cierta distancia que nos permite posicionar nuestro trabajo de tesis y desarrollar una herramienta computacional útil para obtener el contexto semántico de términos en documentos no estructurados.

Del resumen obtenido podemos mencionar que existe un enorme reto en la aportación que facilite la implementación de varios métodos de similitud semántica, y estos, no estén acotados a un solo tipo de ontología ni a un solo dominio de información.

Es cierto que individualmente, cualquier método de similitud semántica tiene desventajas pero se ven disminuidas con la implementación de otro método de similitud. Por ello la necesidad de utilizar un método híbrido en nuestro trabajo de tesis. Un método hídrico no es más que la capacidad de lograr que dos o más métodos trabajen y den lugar a otro método de similitud semántica más robusto. Un método de similitud semántica se traduce en la obtención de un grado de similitud más alto y atacar el reto que existe en la similitud entre palabras y este es: la ambigüedad de términos.

Para lograr nuestro objetivo se diseñó e implemento una metodología que se siguió a lo largo de nuestro trabajo de investigación. Nuestra metodología consta de 3 fases, la primera fase engloba aquellas actividades de procesamiento del lenguaje natural en textos y la similitud semántica de términos en ontologías; la segunda fase engloba todo el diseño e implementación de nuestro método de extracción de contexto semántico en una herramienta computacional; en la etapa tres se engloban los esfuerzos para probar el desempeño de nuestro método para la extracción del contexto semántico utilizando 4 ontologías de dominio y 4 métodos de similitud semántica obtenidos en la literatura.

Nuestro método hibrido se probó cuatro ontologías y cuatro métodos hallados en la literatura, con los resultados obtenidos se superó la similitud semántica de términos en un 90% frente a los propuestos en la literatura. También se realizó un la comparación de similitud semántica frente a WordNet y se obtuvieron los índices de confiabilidad al 95% con una desviación estándar de 0.17243, los índices de confiabilidad son:

- Índice de confianza superior 0.67244474
- Índice de confianza inferior 0.524946588

Del análisis a las pruebas realizadas con nuestro método concluimos lo siguiente:

- Los resultados muestran una mejoría en el 90% de los términos analizados cuando se utiliza más de un métodos de similitud semántica. Estos terminos se obtuvieron con dos índices cuyo grado de confianza es del 95% y un margen de error del 5%.
- Cuando en un método de similitud se obtiene una ambigüedad de términos el otro método de similitud ayuda a la desambiguación.
- Puede realizarse una mejoría en la extracción de contexto semántico si en lugar de utilizar palabras utilizamos frases, pero requiere modificar la herramienta.
- En varias ocasiones el método de similitud semántica propuesto por Sánchez presenta mejores resultados que nuestro algoritmo debido a que este método usa la Base de conocimiento (instancias) de la ontología. Por eso nuestro método hibrido aprovecha esta característica y ocupa este método también para calcular la similitud.

7.2 Trabajos futuros

Los trabajos futuros que se proponen para ampliar y/o mejorar este proyecto de investigación se listan a continuación:

- Considerar el análisis de términos multipalabra en los textos no estructurados. Esto debido que existen palabras compuestas por 2 o más palabras que ligadas originan un significado y denotan conceptos que semánticamente no es posible deducir por las palabras presentes de manera individual [58]. De esta manera se complementaría el análisis que se realiza con Freeling.
- Programar directamente de la herramienta de similitud semántica WS4J [55] para realizar para implementar un método de evaluación directa en nuestra herramienta de extracción de contexto semántico.
- Programar el *plugin* de *Protege* [50] en nuestra herramienta para poder manipular las ontologías de manera visual y especializada.

Referencias

[1] T. Berners-Lee, "Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor", *Paw Prints*, 2008.

[2] A. Andrejev and T. Risch, "Scientific SPARQL: Semantic Web queries over scientific data", *Proc. - 2012 IEEE 28th Int. Conf. Data Eng. Work. ICDEW 2012*, pp. 5–10, 2012.

[3] Venemedia, "CONCEPTODEFINICION.DE", articulo online disponible en: http://conceptodefinicion.de/conocimiento/. Ultima vez accesado: 2018-01-12.

[4] C. Torres, *Inteligencia artificial Conceptos básicos*. Bogotá D. C., Colombia, ISBN 958-33-7213-7, 2007.

[5] T. R. Gruber, "A traslation approach to portable ontologies," *Knolowledge Acquis.*, vol. 5, no. 2, pp. 199–220.

[6] S. Norvig, P., & Russell, "Inteligencia artificial," in *Elsevier Brasil*, vol. 1, no. 3, 2014, p. 1021.

[7] J. Ashraf, E. Chang, O. K. Hussain, and F. K. Hussain, "Ontology usage analysis in the ontology lifecycle: A state-of-the-art review", *Knowledge-Based Syst.*, vol. 80, pp. 34–47, 2015.

[8] N.G. Nazareth, "Web Semantica, HTLM, Ontologia", 2015. articulo online disponible en: http://Websemanticahtlmontologia.blogspot.mx/2015/04/preguntas-de-autoevaluacion.html. Ultima ves accesado: 2018-01-08.

[9] S. Esther, Tesis "Modelo de indexación de formas en sistemas VIR basado en ontologías", Universidad de las Américas Puebla, 2007.

[10] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic Similarity from Natural Language and Ontology Analysis", *Synth. Lect. Hum. Lang. Technol.*, vol. 8, no. 1, pp. 1–254, 2015.

[11] M. Iribe, "Uso de medidas de similitud semántica para procesamiento de informacion geoespacial", Instituto Politecnico Nacional, México, 2010.

[12] M. Wu, Z. and Palmer, "Verb semantics and lexical selection", in *In 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, 1994, pp. 133–138.

[13] M. Leacock, C. and Chodorow, "Filling in a sparse training space for word sense identification", 1994.

[14] Y. J. Lee, J. H., Kim, M. H., and Lee, "Information retrieval based on conceptual distance in is-a hierarchies", in *Artif. Intell. Stat.*, pp. 65–72, 1993.

[15] M. Rada, R., Mili, H., Bicknell, E., and Blettner, "evelopment and application of a metric on semantic nets", *IEEE Trans. Syst. Man Cybern.*, vol. 19, pp. 17–30, 1989.

[16] Y. Xu, T., Du, L., and Zhou, "Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data.," *BMC Bioinformatics*, vol. 9, p. 472, 2008.

[17] A. Tversky, "Features of similarity", *Psychol. Review*, vol. 84, pp. 327–352, 1977.

[18] D. Sánchez, D., Batet, M., and Isern, "Ontology-based information content computation", *Knowledge-Based Syst*, vol. 24, pp. 297–303, 2011.

[19] J. Seco, N., Veale, T., and Hayes, "An intrinsic information content metric for semantic similarity in WordNet", in *In 16th European Conference on Artificial Intelligence (ECAI 2004)*, 2004, pp. 1–5.

[20] A. Sánchez, D., Batet, M., Isern, D., and Valls, "Ontology-based semantic similarity: a new feature-based approach.," *Expert Syst. Appl.*, vol. 39, pp. 7718–7728, 2012.

[21] P. Petrakis, E., Varelas, G., Hliaoutakis, A., and Raftopoulou, "X-Similarity: computing semantic similarity between concepts from different ontologies", *J. Digit. Inf. Manag.*, vol. 4, pp. 233–237, 2006.

[22] T. Patwardhan, S., Banerjee, S., and Pedersen, "Using measures of semantic relatedness for word sense disambiguation", *In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational*, 2003, pp. 241–257.

[23] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries", in *In Proceedings of the 5th annual international conference on Systems documentation (SIGDOC '86)*, 1986, pp. 24–26.

[24] T. Patwardhan, S. and Pedersen, "Using WordNet-based context vectors to estimate the semantic relatedness of concepts", in *EACL Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, 2006, pp. 1–8.

[25] J. C. Jeong and X. Chen, "over Gene Ontology", in IEEE/ACM Trans. Comput. Biol. Bioinformatics vol. 12, no. 2, pp. 322–334, doi 10.1109/TCBB.2014.2343963,2015.

[26] X. Wu, E. Pang, K. Lin, and Z. M. Pei, "Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method", *PLoS One*, vol. 8, no. 5, 2013.

[27] P. Zhang, Z. Zhang, and W. Zhang, "An approach of semantic similarity by combining HowNet and Cilin", *Proc. - 2013 IEEE Int. Conf. Green Comput. Commun. IEEE Internet Things IEEE Cyber, Phys. Soc. Comput. GreenCom-iThings-CPSCom 2013*, no. 13, pp. 1638–1643, 2013.

[28] E. R. Trives, libro "Aspectos de semántica linguistica-textual", 1a. ed., S. A. Alcala, Ed. Madrid, 1979, p. 43.

[29] J. Mansouri, B. Seddik, S. Gazzah, T. Chateau, and R. U. Sage, "Coarse Localization Using Space-Time and Semantic-Context Representations of Geo-Referenced Video Sequences", pp. 1006–1010, 2014.

[30] V. Franzoni and a Milani, "A Pheromone-Like Model for Semantic Context Extraction from Collaborative Networks", *2015 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, vol. 1, pp. 540–547, 2015.

[31] M. a R. Pieck, F. Van Der Sommen, S. Zinger, and P. H. N. De With, "Real-time semantic context labeling for image understanding", *Proc. - Int. Conf. Image Process. ICIP*, vol. 2015-Decem, pp. 3180–3184, 2015.

[32] V. Maran, J. P. M. De Oliveira, R. Pietrobon, and I. Augustin, "Ontology network definition for motivational interviewing learning driven by semantic context-awareness", in *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2015-July, pp. 264–269, 2015.

[33] A. Nagar and Hisham Al-Mubaid, "A Hybrid Semantic Similarity Measure for Gene Ontology Based On Offspring and Path Length", *Comput. Intell. Bioinforma. Comput. Biol. (CIBCB), 2015 IEEE Conf.*, 2015.

[34] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *arXavi Prepr.*, no. cmp-lg/9511007, 1995.

[35] D. Lin, "An information-theoretic definition of similarity in ICML", vol. 98, pp. 296–304, 1998.

[36] J. jiang and D. W. Conrath, "semantic similarity based on corpus statistics and lexical taxonomy", *arXiv Prepr. C.*, 1997.

[37] Wang, Z. Du, R. Payattakool, S. Y. Philip, "A new method to measure the semantic similarity of GO terms", *Bioinformatics*, vol. 23, pp. 1274–1281, 2007.

[38] S. Zhang and J. H. Lai, "A hybrid measure for the semantic similarity of gene ontology terms", *2014 2nd Int. Conf. Syst. Informatics, ICSAI 2014*, no. Icsai, pp. 911–916, 2015.

[39] L. Liu, X. Dai, C. Du, H. Wang, and J. Lu, "A New Hybrid Semantic Similarity Computation Method Based on Gene Ontology", pp. 739–744, 2014.

[40] M. Gan, X. Dou, D. Wang, and R. Jiang, "DOPCA: A new method for calculating ontology-based semantic similarity", *Proc. - 2011 10th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2011*, pp. 110–115, 2011.

[41] F. Schickel-Zuber V, "OSS:A Semantic Similarity Function based on Hierarchical Ontologies", in *Proceedings of International Joint Conference on Artificial Intelligence*, 2007, pp. 551–555.

[42] S. Bandyopadhyay and K. Mallick, "A new Path Based Hibrid Measure for Gene Ontology Similarity", vol. 11, no. 1, pp. 116–127, 2014.

[43] Y. Ni, Q. K. Xu, F. Cao, Y. Mass, D. Sheinwald, H. J. Zhu, and S. S. Cao, "Semantic Documents Relatedness Using Concept Graph Representation", *Proc. Ninth ACM Int. Conf. Web Search Data Min.*, pp. 635–644, 2016.

[44] Hoffart, Johannes and Yosef, "Robust Disambiguation of Named Entities in Text",*Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 782-792, 2011.

[45] Freeling, http://nlp.lsi.upc.edu/freeling/index.php/node/1, ultima vez accesado: 2018-02-04.

[46] Apache Jena, https://jena.apache.org/download/index.cgi, ultima vez accesado: 2018-02-03.

[47] Framework de Netbeans, https://netbeans.org, ultima vez accesado: 2018-14-01.

[48] Sánchez, D. Batet, M. and Isern, "Ontology-based information content computation", *Knowledge-Based Syst.*, vol. 24, pp. 297–303, 2011.

[49] DBpedia.org, https://wiki.dbpedia.org/services-resources/ontology, ultima vez accesado: 2018-22-01.

[50] Protégé, https://protege.stanford.edu, ultima vez accesado: 2018-02-03.

[51] Ontologia de plantas, http://bhort.bh.cornell.edu/histology/ontologia.html, ultima vez accesado: 2018-02-01.

[52] La ontologia bibliografica, http://bibliontology.com/#, ultima vez accesado: 2018-02-01.

[53] Ontologia de viajes, http://metashare.elda.org/repository/browse/travel-domain-ontology/417e7cb0ea3411e380b5842b2b6a04d7368c0e7c2c9b49caa07e1eb71444006e/, ultima vez accesado: 2018-01-04.

[54] Biblioteca de medidas semantica & *Kit de herramientas, disponible en:* http://www.semantic-measures-library.org/sml/, ultima vez accesado: 2018-14-01.

[55] Similitud de WordNet para Java, http://ws4jdemo.appspot.com, ultima vez accesado: 2018-15-01.

[56] W3C, Ontology repositories, https://www.w3.org/wiki/Ontology_repositories, ultima vez accesado: 2018-02-03.

[57] Z. Zhou, Y. Wang and J. Gu, “A New Model of Information Content for Semantic Similarity in WordNet”, 2008 Second International Conference on Future Generation Communication and Networking Symposia, Sanya, 2008, pp. 85-89.

[58] Savary A., Sailer M., Parmetier Y., Rosner M., et al. *"PARSEME – PARSing and Multiword Expressions within a European multilingual network"*. Submitted to Language & Technology Conference (LTC’15), Poznan , November 27-29, 2015.

[59] Jay J. Jiang and . "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,. CoRR, cmp-lg/9709008, 1997.

[60] Miller, G., 1990, “Nouns in WordNet: A Lexical Inheritance System”, International Journal ofLexicography, Vol. 3, No. 4, 245-264.