UNIVERSIDAD
Gastón Dachary

Ingeniería en Informática

TRABAJO FINAL DE CARRERA

Análisis de documentos Web utilizando métricas de relación y similitud semántica

Alumnos:

- Pfeifer, Hernán Ariel
- Rojas, Matias Gabriel

Director de T.F.C.: Dr. Karanik, Marcelo

Co-Director de T.F.C.: Ing. Favret, Fabian

POSADAS – MISIONES

2019

ÍNDICE

RESUMEN

La competitividad y subsistencia de las empresas y organizaciones, dependen fuertemente de la toma de decisiones acertadas, que permitan hacer frente a los problemas y aprovechar las oportunidades.

Un requisito esencial para la toma de decisiones es reducir la incertidumbre existente con respecto al tema a considerar. Para esto, es fundamental contar con información adecuada y oportuna.

En la actualidad, la fuente de información por excelencia es internet, debido a que contiene en un solo lugar, información de distinta índole y proveniente de distintos orígenes. Sin embargo, su rápido crecimiento provocó un fenómeno conocido como sobresaturación de información, que implica que el usuario explore una cantidad infinita de información con el fin de satisfacer una necesidad. Con el paso del tiempo, la cantidad de información se tornó inmanejable, lo que derivó en un problema.

Varios trabajos de investigación tienen como objetivo resolver este problema, proponiendo herramientas inteligentes que pudieran identificar la información relevante a la necesidad que posee el usuario. En este sentido, en el ámbito de la Universidad Gastón Dachary se propuso un sistema de recuperación de información, que evalúa la relevancia de recursos web (documentos web) mediante la utilización de técnicas basadas en correspondencia lexicográfica.

Estas técnicas buscan representar el criterio de determinación de relevancia del usuario, exigiendo coincidencias exactas de palabras de una clave de búsqueda en el contenido de los recursos web. Sin embargo, resultan en un análisis incompleto, debido a que se dejan de lado ciertos aspectos determinantes, como ser las características contextuales de la clave de búsqueda, palabras relacionadas a las que se definieron en la clave, entre otros.

Es por ello que, en este trabajo de investigación, se propone un modelo de determinación de relevancia basado en técnicas de análisis semántico, que asume que cada palabra es utilizada en un contexto determinado, que le proporciona un significado específico y un conjunto de relaciones semánticas con otras palabras.

Este modelo utiliza estas relaciones y similitudes semánticas, propias del contexto al que pertenece cada palabra de la clave de búsqueda, para determinar qué tan relevante es un recurso analizado.

Por otro lado, también integra los criterios considerados por ambas técnicas (lexicográficas y semánticas) con el fin de realizar un análisis que contemple una mayor cantidad de factores a la hora de determinar la relevancia y, por ende, mejorar los resultados obtenidos.

Sobre este modelo se ejecutaron un conjunto de pruebas, considerando dos escenarios, siendo cada uno de ellos, propuesto por un experto distinto. Los resultados obtenidos mostraron que la implementación de las técnicas semánticas mejoró a los resultados proporcionados por las técnicas basadas en correspondencia lexicográfica, lo que expone la factibilidad de su utilización.

Además, en las pruebas realizadas sobre la integración de técnicas, se observaron mejores resultados combinando los criterios que utilizando cada técnica por separado.

Palabras Clave: análisis semántico de recursos Web, métricas de relación semántica, análisis de relevancia, sistemas de recuperación de información Web, ranking de recursos Web.

ABSTRACT

The competitiveness and subsistence of companies and organizations, depend heavily on making the right decisions, to deal with problems and make the most of opportunities.

An essential requirement for decision-making is to reduce the existing uncertainty regarding the issue to be considered. For this, it is essential to have appropriate and timely information.

Currently, the main source of information is the Internet, because it contains in one place, information of different nature and from different origins. However, its rapid growth caused a phenomenon known as information over-saturation, which implies that the user explores an infinite amount of information in order to satisfy a need. With the passage of time, the amount of information became unmanageable, resulting in a problem.

Several research projects are aimed to solve this problem, proposing intelligent tools that could identify the information relevant to the need that the user has. In this regard, within the scope of Gaston Dachary University, an information retrieval system has been proposed which evaluates the relevance of web resources (web documents) by using lexicographical correspondence-based techniques.

These techniques seek to represent the criterion of relevance determination of the user, requiring exact matches of words of a search key in the web resources content. However, they result in an incomplete analysis, due to the fact that certain determining aspects are left aside, such as the contextual characteristics of the search key, words related to those defined in the key, among others.

That is why, in this research work, a relevance determination model is proposed based on semantic analysis techniques, which assumes that each word is used in a specific context, providing a specific meaning and a set of semantic relationships with other words.

This model uses these semantic relationships, specific to the context to which each word of the search key belongs, to determine how relevant an analyzed resource is.

On the other hand, it also integrates the criteria considered by both lexicographic and semantic techniques, in order to perform an analysis that includes a greater number of factors when determining relevance and, therefore, improve the obtained results.

Regarding this model, a set of tests were applied, considering two scenarios, each of them being proposed by a different expert. The results obtained showed that the implementation of semantic techniques improved the results provided by the lexicographical correspondence-based techniques, which exposes the feasibility of its use.

In addition, in the tests performed concerning the integration of techniques, better results were found by combining the criteria than using each technique separately.

Keywords: semantic analysis of Web resources, semantic relationship metrics, analysis of relevance, Web information retrieval systems, ranking of Web resources.

CAPÍTULO 1
INTRODUCCIÓN

En este capítulo, se presentan los aspectos generales del presente trabajo, introduciendo en principio la problemática que motivó su realización y los objetivos que se persiguen. Asimismo, se detalla el contexto en el que se desarrolla, se especifica la organización del documento y se exponen algunos trabajos antecedentes relacionados.

1.1 MOTIVACIÓN

Con el paso del tiempo, el paradigma competitivo y de subsistencia de las empresas y las organizaciones ha cambiado y generado el replanteo de estrategias con el fin de lograr mantenerse y posicionarse en un mercado que día a día se va tornando más complejo. Hoy en día, ya no basta solo con una estrategia basada en la disminución de precios o el aumento de calidad en el producto, para poder posicionarse como un serio competidor en el mercado, sino que entra en juego la innovación que tenga la compañía para poder introducir nuevos o mejorados productos y servicios.

A partir de esto, surge la necesidad de utilizar herramientas que provean a las empresas la información necesaria para hacer frente a fenómenos o acontecimientos que puedan llegar a ocurrir en un futuro cercano, abrirse a nuevos mercados, conocer los últimos avances tecnológicos, establecer que investigaciones realizar, etc. Todo esto apunta a intentar identificar oportunidades y amenazas, y por ende tomar decisiones estratégicas acertadas [1].

Hoy en día, la fuente principal de información es Internet debido a que contiene una gran diversidad de documentos web, artículos, trabajos de investigación realizados por otras empresas, opiniones, y toda clase de recursos que pueden ser útiles para tomar decisiones. Sin embargo, a diferencia de hace algunas décadas, el problema no es la falta sino la sobresaturación de información. Poder establecer qué es importante y relevante sobre un tema en particular no es una tarea trivial ya que se requiere de conocimiento específico del dominio [2][3].

En este contexto, se intenta dar apoyo a la búsqueda de información mediante varias herramientas inteligentes, entre las que se encuentran la Vigilancia Tecnológica y la Inteligencia Competitiva (VTeIC). Estas herramientas tienen por objetivo, la obtención de información relevante que pueda satisfacer las necesidades de información y su presentación en un formato adecuado, de manera que contribuya a la toma de decisiones [4].

Una de las principales actividades que realiza un sistema VTeIC, es la denominada Recuperación de la Información (RI). Esta actividad se centra en la utilización de distintos mecanismos para la obtención de información relevante.

Existen trabajos de investigación [5][6][7] donde se han tratado los problemas de la VTeIC y RI, implementando un sistema compuesto por dos módulos. El primero es un proceso que da soporte a los usuarios en la generación de claves de búsqueda y presenta los resultados obtenidos. El segundo, lleva a cabo un proceso de búsqueda continua y determinación de la relevancia de recursos web.

La determinación de la relevancia de los recursos recuperados se realiza, mediante Técnicas Lexicográficas (TL), que consisten en el conteo de la frecuencia de repetición de palabras de la clave de búsqueda.

Si bien es acertado considerar a la frecuencia de repetición de palabras, como medida de la relevancia de recursos, resulta un análisis incompleto, debido a que deja afuera aspectos determinantes.

Un primer aspecto no tenido en cuenta es el contexto del recurso, dado que estas técnicas solamente se ocupan de la ocurrencia explicita de palabras pertenecientes a la clave de búsqueda, independientemente de su sentido, o la categoría sintáctica a la que pertenezca (sustantivo, verbo, adjetivo, adverbio).

Tampoco son consideradas las relaciones entre palabras, como la hiperonimia (el significado de una palabra engloba a otra), la hiponimia (una palabra es contenida por otra más general), la sinonimia (igualdad existente entre el significado de dos o más palabras) o la antonimia (oposición entre los significados de dos palabras), etc.

Este tipo de relaciones permite mirar más allá de las palabras presentes en el recurso y la clave de búsqueda, logrando contemplar de una manera más amplia y precisa la correspondencia existente entre ambos [8][9].

No tener en cuenta estos aspectos, a la hora de determinar la relevancia de recursos, puede derivar en la obtención de resultados no adecuados. Puesto que un recurso puede hacer referencia a un tema buscado sin necesidad de que las palabras de la clave de búsqueda aparezcan con frecuencia dentro del mismo o, la situación contraria donde las palabras de la clave de búsqueda están contenidas en el recurso, pero, aun así, no se relaciona con lo que se busca.

Es a partir de esto que surge la necesidad de considerar técnicas que contemplen el contexto de las palabras a la hora de determinar la relevancia de un recurso.

Asimismo, como ambos enfoques poseen criterios distintos de determinación de relevancia, es posible combinarlos con el fin de realizar un análisis que contemple una mayor cantidad de factores, y así obtener mejores resultados.

1.2 CONTEXTO

Este Trabajo Final de Carrera (TFC) está radicado en el ámbito de los proyectos de investigación Modelo de Análisis de Información Desestructurada Utilizando Técnicas de Recopilación y Minería Web adjudicado por la resolución N° 07/A/17 de la Universidad Gastón Dachary (UGD) y Análisis de Información en Grandes Volúmenes de Datos Orientado al Proceso de Toma de Decisiones Estratégicas perteneciente a la Universidad Tecnológica Nacional – Facultad Regional Resistencia (UTN – FRRe).

Dada una necesidad de información presentada por una organización, el objetivo de estos proyectos, es la recuperación de información relevante almacenada en la web, mediante la aplicación de técnicas de VTeIC, que sirva de soporte a la toma de decisiones estratégicas.

Para tal fin, en la actualidad se encuentran implementados dos módulos separados físicamente, donde el primero tiene como objetivo la transformación de necesidades de información de un usuario en requerimientos y la presentación de resultados obtenidos, y el segundo lleva a cabo un proceso de búsqueda continua de recursos disponibles en la web, para satisfacer los requerimientos definidos por el primer módulo.

La disposición del Módulo de Recopilación de Requerimientos (MRR) y del Módulo de Minería Web (MMW), se puede apreciar en la Figura 1.1.

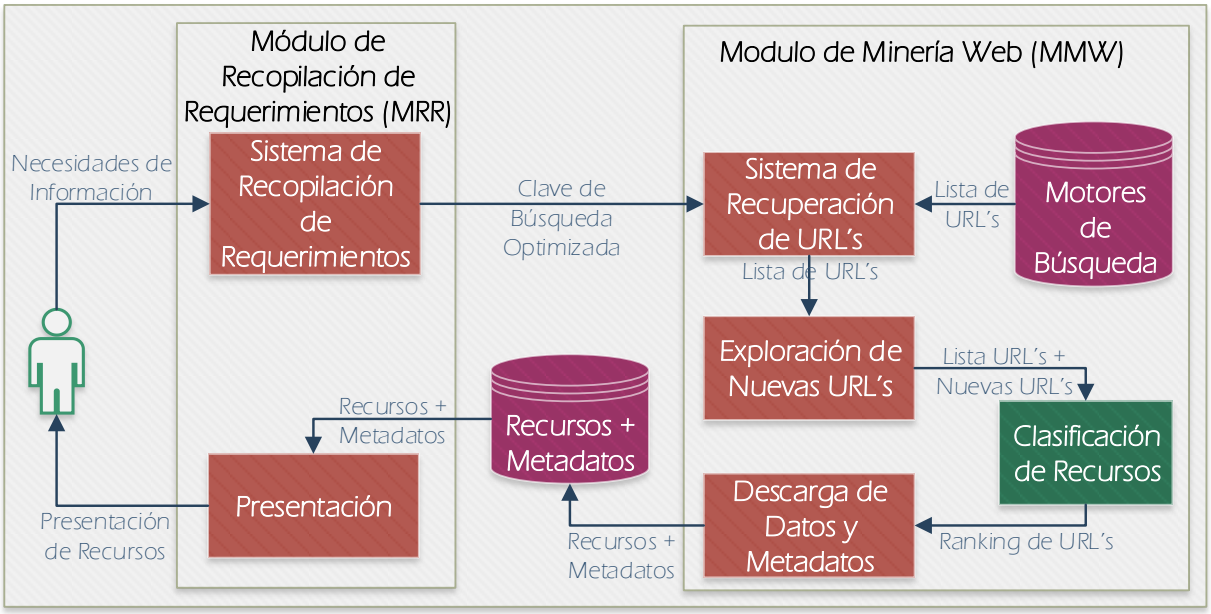

Figura 1.1 *- Modelo del proyecto VTeIC*

El proceso de búsqueda es iniciado por el usuario, que plantea una necesidad de información y la ingresa al MRR. El sistema de recopilación de requerimientos la transforma en una clave de búsqueda optimizada, con la que se obtienen las URL's raíces desde distintos motores de búsqueda en el MMW.

Como resultado, se obtiene una lista de URL's, a partir de las cuales se realiza la exploración de enlaces relacionados, que tiene como fin, descubrir nuevas URL's que puedan satisfacer las necesidades de información.

Con las URL's raíces y las descubiertas en el proceso de exploración, se desencadena el proceso de clasificación de recursos. Para ello, se determina la relevancia de los recursos de cada URL, mediante las TL.

Finalizados los procesos realizados por el MMW, se obtiene un ranking ordenado de acuerdo a la relevancia de los recursos con respecto a la clave de búsqueda optimizada.

Seguidamente, se descarga el contenido y los metadatos de las URL's, para que el MRR realice la presentación de los recursos web al usuario.

En este trabajo, se describe la implementación de las Técnicas Semánticas (TS), que, mediante la consideración de relaciones y similitudes semánticas, determinan la relevancia de recursos. Estas técnicas complementaran al análisis realizado por las TL ya implementadas.

1.3 OBJETIVOS

1.3.1 Objetivo General

Desarrollar un mecanismo de clasificación de recursos web, a partir de técnicas de determinación de relevancia basadas en métricas de relación y similitud semántica.

1.3.2 Objetivos Específicos

- Analizar el estado del arte en lo referido a técnicas de determinación de la relevancia de recursos basado en el análisis semántico.
- Diseñar un modelo de clasificación de recursos web que utilice técnicas de determinación de relevancia basadas en relación y similitud semántica.
- Implementar el modelo propuesto de clasificación de recursos web que complemente al sistema VTeIC ya implementado.
- Evaluar resultados obtenidos por técnicas basadas en frecuencias de repetición de palabras, técnicas basadas en métricas de relación y similitud semántica y la combinación de estas dos técnicas, mediante la comparación de la eficiencia en el ordenamiento de recursos.

1.4 ORGANIZACIÓN DEL DOCUMENTO

Este TFC está organizado en base los objetivos específicos manteniendo el orden en que fueron definidos, con el fin de lograr una mejor comprensión a medida que se avanza por los distintos capítulos. Para ello, la disposición es la siguiente.

En el capítulo 2, se da una introducción a la recuperación de información, repasando conceptos asociados, tales como Sistemas de Recuperación de Información (SRI), relevancia y métricas de evaluación de SRI. También, se describe al sistema actualmente puesto en producción, que implementa a las TL. Este sistema es la base sobre la cual se implementan las TS propuestas en este trabajo.

En el capítulo 3, se introduce teóricamente a las herramientas utilizadas por las TS. En primer lugar, se exponen los conceptos de relación, similitud y distancia semántica entre pares de palabras, que son la base para determinar la relación y similitud semántica existente entre recursos y clave de búsqueda. Seguidamente, se presentan las relaciones semánticas contempladas en la informática, que se dividen en dos categorías: las clásicas y las no clásicas. También se exponen herramientas tales como las taxonomías semánticas y las métricas de relación y similitud semántica entre pares de palabras, aplicables sobre estas taxonomías.

En el capítulo 4, se detalla el modelo propuesto, explicando las tareas involucradas en el proceso de evaluación semántica de recursos. Además, se especifica el Modelo de Integración Léxico - Semántico (MILS) que consiste en la utilización de una fórmula de unificación de rankings, modificada para combinar los criterios considerados tanto por las TL, como por las TS.

En el capítulo 5, se presentan las pruebas realizadas al modelo propuesto, con el objetivo de observar los efectos positivos o negativos, de la implementación de las TS y el MILS, por sobre los resultados obtenidos por las TL. Para esto, se realizan tres tipos de pruebas: Pruebas de recuperación, pruebas de ordenamiento (ambas aplicadas sobre las TL y las TS) y la evaluación de los resultados obtenidos por el MILS, con respecto a los de las TL y las TS.

Finalmente, en el capítulo 6 se presentan las conclusiones derivadas de la ejecución de las pruebas y los trabajos futuros.

1.5 ANTECEDENTES

En los últimos años, los avances en el área del procesamiento de lenguaje natural (NLP - *Natural Language Processing*), han permitido la aparición de nuevas propuestas en cuanto a la RI. Uno de los campos que fue de gran interés para distintas investigaciones, es la determinación de relevancia de recursos mediante técnicas de análisis semántico. En la presente sección se presentan algunos de ellos.

Entre los antecedentes a este trabajo, se encuentra el propuesto en [10] en el que se presenta un método que considera relaciones de herencia y la distancia semántica, para medir el grado de correspondencia entre palabras. Para ello, utiliza recursos externos como WordNet, mediante el que obtiene un fragmento de la taxonomía, que contiene a las palabras a evaluar. El proceso tiene como entrada dos palabras, que son evaluadas mediante los siguientes cuatro pasos: Primero, se determinan los pesos de las relaciones entre el nodo raíz y los nodos del fragmento de la taxonomía. Luego, se genera una tabla de enrutamiento, donde se registran todas las rutas posibles entre el nodo raíz y los nodos de las palabras a evaluar. En el tercer paso, se calcula la distancia semántica de cada ruta existente en la tabla de enrutamiento. Finalmente, se determina la similitud semántica considerando las distancias semánticas obtenidas en el tercer paso.

Otro enfoque es el trabajo realizado en [11] donde se presenta un sistema de respuesta a preguntas, que busca determinar la oración que mejor conteste a una pregunta ingresada por el usuario. Estas oraciones son extraídas de una colección de recursos, obtenidos a partir de técnicas de RI. Para determinar la mejor respuesta, utilizan métricas de relación y similitud semántica, aplicadas a cada una de las oraciones con respecto a la pregunta. Además, presentan técnicas de mejoramiento del análisis, como separación de frases nominales, etiquetado de partes de la oración, entre otros. Como conclusión especificaron, que no toda la información lingüística es útil y que ciertas características de las oraciones son más importantes que otras a la hora de determinar la relevancia.

Una variante interesante es la que se presenta en [12], en la que se propone la evaluación de una métrica de similitud semántica, con el fin de utilizarla en el futuro, con fuentes de datos reales. Para ello se solicita a distintos expertos que puntúen la similitud existente entre las sinopsis de un conjunto de pares de películas. Luego el sistema puntúa la similitud para el mismo conjunto, utilizando la taxonomía DBPedia. Finalmente, evalúa a cada una de las métricas, considerando la correlación entre los puntajes otorgados por el experto y los otorgados por el sistema. Los resultados obtenidos mostraron una correlación de Pearson de 0,69 con respecto a los puntajes de los expertos.

Otro antecedente es el presentado en [13], donde se propone un modelo de *clustering* de resultados de motores de búsqueda, que se compone por los siguientes seis pasos: primero obtiene resultados de los motores de búsqueda para una clave de búsqueda ingresada por el usuario. Luego, continúa con el preprocesamiento de los recursos obtenidos, extrayendo características de cada uno de ellos. Seguidamente, realiza el enriquecimiento de características de los recursos, mediante la taxonomía WordNet y se construye una red semántica que modela al recurso. En el paso siguiente, se aplica el algoritmo *spreading activation*, a la red semántica construida. Posteriormente, se calcula la matriz de desemejanza entre los recursos recuperados, utilizando las características más significativas que los representan y finalmente, se aplica el algoritmo de *clustering* jerárquico aglomerativo, a la matriz de desemejanza. Los experimentos confirmaron que la solución presentada por este trabajo, arrojó resultados notables en la precisión de los *clusters* obtenidos.

También es importante destacar al modelo presentado en [6], que es el antecesor directo al modelo propuesto en el presente trabajo. En el mismo se llevó a cabo la

implementación de dos módulos principales que cumplen con los siguientes objetivos: el primero, generar un proceso que de soporte a la recopilación de requerimientos de usuario y presentación de los resultados. El segundo objetivo, consiste en generar un proceso de búsqueda continua de recursos en la web, evaluando su relevancia a partir de técnicas de correspondencia lexicográfica, para posteriormente ordenarlos de acuerdo al grado de relevancia obtenido con respecto a la clave de búsqueda ingresada por el usuario. Los resultados obtenidos fueron satisfactorios para los distintos escenarios en los que fue evaluado, debido a que recupera información acorde al tipo de respuesta esperada.

En líneas actuales de investigación relacionadas al análisis semántico, se continúa avanzando sobre la correcta determinación del contexto. Como ejemplo se puede observar el enfoque presentado en [14], donde se propone el enriquecimiento del contexto de los recursos a analizar, para utilizarlo en la estimación de la relación semántica. Este enriquecimiento se hace mediante las definiciones de palabras presentes en los recursos, obtenidas mediante DBPedia y Wikipedia. En tareas de determinación de la similitud entre recursos, esta solución superó a los resultados obtenidos por los métodos tradicionales y logró una performance similar a aquellos basados en conocimiento (que dependen del aporte de expertos). En tareas de RI mejoró la calidad de los resultados con respecto a las técnicas utilizadas hasta ese momento.

Además, también existen avances con respecto a la implementación de técnicas de *machine learning* y *deep learning*, aplicables a la recomendación de recursos relevantes.

Un enfoque es el propuesto en [15], en el que se plantea la utilización de una red neuronal recurrente para medir la relevancia de *papers* académicos, utilizando técnicas de bolsas de palabras (*BOW - bag of words*) y TF-IDF (*term frecuency and inverse document frecuency*). Este enfoque ha mostrado buenos resultados obteniendo un valor Kappa de 0,869.

Una propuesta similar es la presentada en [16], donde se sugiere la utilización de redes neuronales convolucionales para la extracción de la importancia de los términos de una clave de búsqueda, la frecuencia de los términos de búsqueda y la relevancia obtenida mediante el método BM-25, para posteriormente determinar un puntaje de relevancia global de artículos de noticias. Los experimentos en grandes conjuntos de artículos de noticias demostraron la efectividad del modelo propuesto, en comparación con modelos tradicionales.

El avance de estas líneas permitirá, a largo plazo, obtener técnicas y procesos que resulten más fiables, ya que contemplaran una mayor cantidad de factores, donde no solo las características del recurso sean importantes a la hora de determinar su relevancia, sino que también se consideren las características subjetivas que puedan aportar los usuarios indirectamente, aprendidas por las técnicas de *machine learning* y *deep learning*.

CAPÍTULO 2
SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

En el presente capítulo, se da introducción a todos los conceptos relacionados a los sistemas de recuperación de información, comenzando por el de recuperación de información, que es el que establece las bases sobre las que se construyen estos sistemas. También se lo compara con respecto a la recuperación de datos y se expone el concepto de relevancia, que es fundamental en la evaluación de recursos. Asimismo, se exponen métricas de evaluación de estos sistemas y algunos modelos generales existentes en la actualidad. Para finalizar, se describe un sistema que implementa técnicas lexicográficas, que será complementado con el modelo de análisis semántico propuesto en este trabajo.

2.1 RECUPERACIÓN DE INFORMACIÓN Y CONCEPTOS RELACIONADOS

Con el paso de los años, la necesidad de contar con la información adecuada de manera oportuna, en el proceso de toma de decisiones, dejó de ser un aspecto secundario para pasar a ser primordial. Es entonces, que cobró relevancia el concepto de Recuperación de Información (RI) [2].

La RI es el área de la informática que, mediante el empleo de diversas técnicas y herramientas, permite la adquisición, representación, almacenamiento, procesamiento y presentación de la información relacionada a la necesidad que poseen los usuarios [17][18].

La adquisición de información hace referencia al empleo de distintas técnicas para obtener recursos de diversas fuentes, tratándose usualmente de contenidos presentes en la web. Esta actividad representa un filtro inicial, ya que se trata de obtener el conjunto de recursos que tentativamente responda a la necesidad del usuario.

La representación, consiste en la manera en que esta información adquirida se representa internamente, previo a su almacenamiento. Se realiza con el fin de que tanto el contenido de los recursos como la clave de búsqueda, sean uniformemente representados, para lograr una comparación efectiva.

El almacenamiento, se corresponde con la manera en que la información persistirá en el tiempo, es decir, decidir si llevar a cabo un almacenamiento temporal o permanente, y que técnicas se utilizarán para ello. Su importancia radica en permitir el rápido acceso a la representación de los recursos, realizada en la actividad anterior.

El procesamiento, incluye todas las actividades relacionadas a la determinación de la relevancia de los recursos y la generación de rankings ordenados de acuerdo a esta relevancia. Es la actividad principal de todo Sistema de Recuperación de Información (SRI), ya que, en base a ella se determina la pertinencia de los recursos a la necesidad de información del usuario.

Por último, la presentación tiene que ver con la manera en la que el usuario tendrá acceso a la información recuperada y la forma en que ésta será organizada para su correcta comprensión. Es vital, que el usuario pueda comprender correctamente los resultados proporcionados, de manera que pueda darle una utilidad.

Asimismo, la presentación puede incluir procesos de retroalimentación en los que el usuario evalué los resultados presentados y de esta manera contribuya a mejorar la estimación de relevancia.

La implementación de la RI, se lleva a cabo mediante los SRI, que pueden consistir en programas informáticos que realizan simples tareas de recuperación o involucran tareas avanzadas de análisis de contenido [3].

En la Figura 2.1 se puede observar la arquitectura básica de un SRI propuesta por Tolosa y Bordignon [2], que es una generalización de la estructura que soporta las actividades elementales que realiza cualquier SRI.

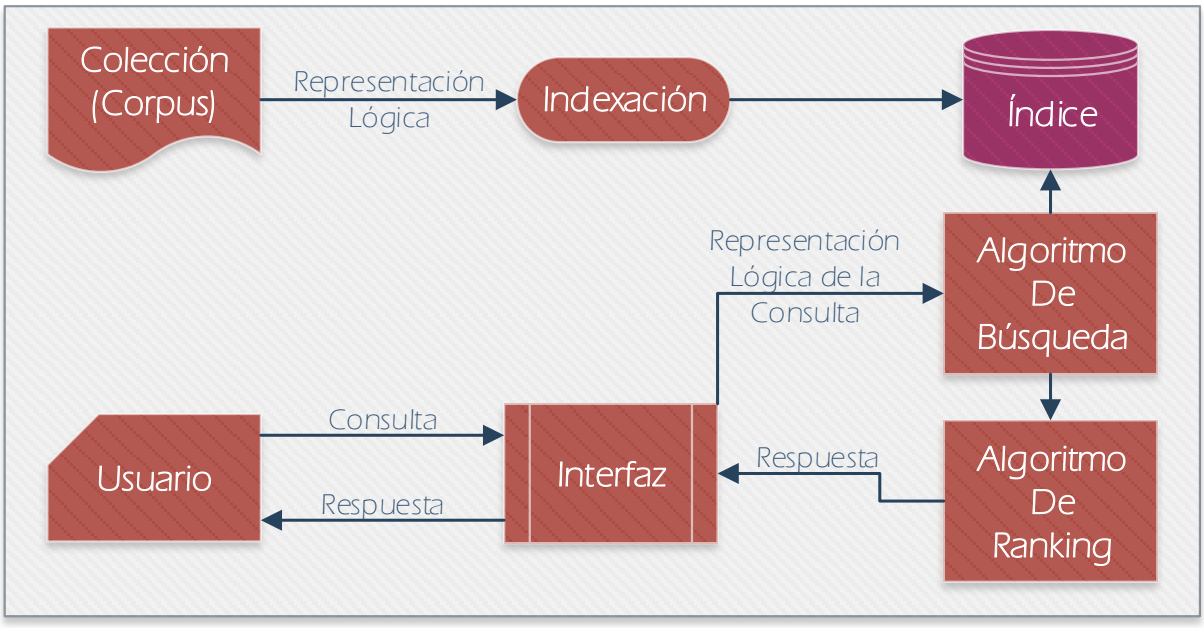

***Figura 2.1**- Arquitectura de un SRI basado en [2]*

En principio se cuenta con un conjunto de recursos, donde cada uno se compone por sucesiones de palabras que forman estructuras gramaticales, como ser oraciones y párrafos. Este conjunto generalmente se denomina corpus, colección o base de datos documental.

Para poder aplicar las operaciones de la RI sobre el corpus, es necesario que sus recursos, sean representados lógicamente mediante términos, frases u otras unidades (sintácticas o semánticas), que permitan caracterizarlos [2][17].

Luego estos recursos se almacenan en discos (generalmente denominados, repositorio central), generando sobre ellos una estructura denominada índice, que permite su rápida ubicación, recuperación y clasificación [17]. Dicho índice consiste en el estado inicial del SRI.

El funcionamiento del SRI es iniciado por un usuario que posee una necesidad de información que desea satisfacer. Este usuario transmite su necesidad de información en forma de clave de búsqueda, que se transforma en una representación lógica de la clave de búsqueda.

Luego, mediante la utilización de un Algoritmo de Búsqueda, se ubica a los recursos que posiblemente satisfagan a la clave de búsqueda, en el índice generado a partir del corpus del SRI. Finalmente, mediante un Algoritmo de Ranking, se determina la relevancia de cada recurso y en base a ella, se construye una lista ordenada.

Un resultado ideal para un SRI es una respuesta que únicamente esté compuesta por recursos relevantes. Sin embargo, en la práctica esto no es posible, debido a la dificultad de compatibilizar la expresión de la necesidad de información con el lenguaje de los recursos y la inherente subjetividad para determinar la relevancia, que varía de acuerdo al usuario. Por

lo tanto, su objetivo es recuperar la mayor cantidad posible de recursos relevantes minimizando la cantidad de los no relevantes (ruido) [2][17].

Teniendo en cuenta esto, se puede decir que el problema de la RI puede ser estudiado desde dos puntos de vistas, el computacional y el humano [17]. El primer caso se refiere a la construcción de estructuras de datos y algoritmos eficientes que mejoren la calidad de las respuestas, y el segunda caso, al estudio del comportamiento y las necesidades del usuario. En las secciones siguientes se realiza una comparación entre la RI y la RD, y se introduce al concepto de relevancia, siendo ambos concernientes a la RI.

2.1.1 Recuperación de información y recuperación de datos

En ocasiones se utilizan indistintamente los términos datos e información, sin embargo, su significado es diferente. Los datos, son valores obtenidos como producto de mediciones, resultados de cálculos, etc., lo que en simples palabras significa que por sí solos no son una fuente de ayuda para la toma de decisiones debido a su carencia de significado.

Al procesar estos datos, pasan a ser información, ya que adquieren un contexto y relaciones con otros datos, lo que lo vuelve útil para el soporte a la toma de decisiones [2].

De igual manera, existen diferencias entre la RI y la Recuperación de Datos (RD). Estas diferencias están relacionadas a los tipos de objetos con los que trata cada una, la representación de estos objetos, la especificación de las consultas y los resultados que se obtienen.

En primer lugar, la RD trata con valores y claves de búsqueda con una estructura definida, mientras que la RI debe lidiar, en ambos casos, con las dificultades del procesamiento del lenguaje natural.

Por otro lado, se puede ver a la RD como una aproximación a la RI, en las situaciones en que se busca determinar los recursos de una colección que contienen las palabras de la clave de búsqueda ingresada por el usuario.

Sin embargo, desde el punto de vista de la RI, lo más probable es que los resultados obtenidos sean irrelevantes para la necesidad de información que el usuario posee, ya que la ocurrencia de palabras de la clave de búsqueda en un recurso, no es suficiente como para afirmar si éste es relevante o no [17].

Con el fin de diferenciar de manera más precisa a la RI y la RD, en la Tabla 2.1, se exhibe una comparación entre sus características.

Tabla 2.1 *- Características de la recuperación de datos y recuperación de información [2]*

Características	***Recuperación de datos***	***Recuperación de información***
Acierto	Exacto	Parcial, el mejor
Inferencia	Algebraica	Inductiva
Lenguaje de consulta	Fuertemente Estructurado	Estructurado o Natural.
Especificación consulta	Precisa	Imprecisa
Estructura	Información estructurada.	Información semi o no estructurado
Error en la respuesta	Sensible	Insensible
Recuperación	Determinística	Probabilística

A partir de estas comparaciones, se puede afirmar que, tanto la RI como la RD representan enfoques distintos, donde uno busca encontrar recursos que se adecuan a la necesidad de información que posee el usuario y el otro busca encontrar respuestas exactas a lo que el usuario desea [17].

2.1.2 Relevancia

La relevancia es uno de los conceptos más importantes en la teoría de la RI, que surge de considerar que, si un usuario de un SRI tiene una necesidad de información, entonces algunos recursos de una colección pueden ser relevantes a esta necesidad o en otras palabras, la información que se considera relevante para la necesidad del usuario es aquella que puede ayudarlo a satisfacerla [2].

Como se mencionó anteriormente, el objetivo de un SRI es obtener la mayor cantidad de recursos relacionados a la temática de la clave de búsqueda ingresada por el usuario. El criterio a partir del cual se determina esta relación, se denomina relevancia [17].

Sin embargo, determinar si un recurso es relevante es una actividad compleja, debido a que se trata de un juicio subjetivo, lo que significa, que dos usuarios pueden determinar niveles de relevancia distintos para un mismo recurso. Por lo tanto, se hace evidente que la relevancia consiste en realidad en la consideración de varios factores, entre los que se encuentran las características del recurso, las características de la necesidad de información y la subjetividad del usuario que elabora la clave de búsqueda [2].

Con respecto a esto, Blair [19] afirma que es más fácil llevar a cabo la determinación de la relevancia de un recurso determinado, que explicar cómo fue obtenida o qué criterios se utilizaron. Esto claramente, tiene fundamento en que se trata de una actividad que los seres humanos realizan de manera automática, mediante procesos cognitivos, para lo que no existe un consenso en cuanto a cómo se lleva a cabo.

En definitiva, si bien es cierto que los métodos de determinación de relevancia se ven afectados por la subjetividad, se puede decir que los existentes hasta el momento, logran representar los criterios considerados por los usuarios de manera parcial, siendo esto suficiente como para obtener una precisión aceptable en cuanto a la cantidad de recursos relevantes recuperados. Igualmente, los avances cotidianos permiten vaticinar un acercamiento a la determinación cognitiva de la relevancia que llevan a cabo los usuarios.

2.2 MÉTRICAS DE EVALUACIÓN DE SRI

Como se sabe, la respuesta de un SRI generalmente no es exacta, debido a dos factores: la infinita cantidad de recursos disponibles para evaluar y la subjetividad inherente a la determinación de la relevancia [2]. Es por esto, que la respuesta es más acertada, cuantos más recursos relevantes la conformen.

Con el objetivo de evaluar, que tan acertada es la respuesta proporcionada al usuario y, por lo tanto, que tan efectiva es la determinación de relevancia del SRI, es necesario utilizar un conjunto de métricas que permitan llevar a cabo dicha labor.

En la siguiente sección se describen dos de las métricas más utilizadas en la evaluación de SRI. Estas son la precisión y la exhaustividad.

2.2.1 Precisión y exhaustividad

Existen varias métricas que permiten evaluar la calidad de los resultados obtenidos, de las cuales las más utilizadas son la precisión y exhaustividad, planteadas por Cleverdon [20]. Gráficamente estos conceptos pueden ser representados mediante la Figura 2.2.

***Figura 2.2**- Conjunto de recursos dada una solicitud de información [2]*

La precisión se define como la proporción de los recursos recuperados que son relevantes y permite evaluar la habilidad del sistema para generar un ranking en el que las primeras posiciones estén ocupadas por recursos relevantes [17]. La precisión se calcula mediante la Ecuación (2.1).

$$Precisión = \frac{\#RecursosRelevantesRecuperados}{\#RecursosRecuperados} \quad (2.1)$$

Donde $\#RecursosRelevantesRecuperados$ es la cantidad de recursos recuperados que son relevantes (en la Figura 2.2, es la intersección entre el conjunto azul y el conjunto rojo) y $\#RecursosRecuperados$ es la cantidad total de recursos recuperados (conjunto rojo).

La exhaustividad se define como la proporción de los recursos relevantes que han sido recuperados y permite evaluar la habilidad del sistema para encontrar todos los recursos relevantes de una colección [17]. La exhaustividad se calcula mediante la utilización de la Ecuación (2.2).

$$Exhaustividad = \frac{\#RecursosRelevantesRecurperados}{\#RecursosRelevantes} \quad (2.2)$$

Una situación ideal es que los SRI proporcionen al mismo tiempo un 100% de exhaustividad y un 100% de precisión, es decir, que se recuperen todos los recursos relevantes y tan sólo los recursos relevantes. Sin embargo, en la práctica estos dos indicadores se comportan de manera opuesta, ya que a medida que se incrementa la exhaustividad, se disminuye la precisión y viceversa [3]. Este comportamiento se puede apreciar en la Figura 2.3.

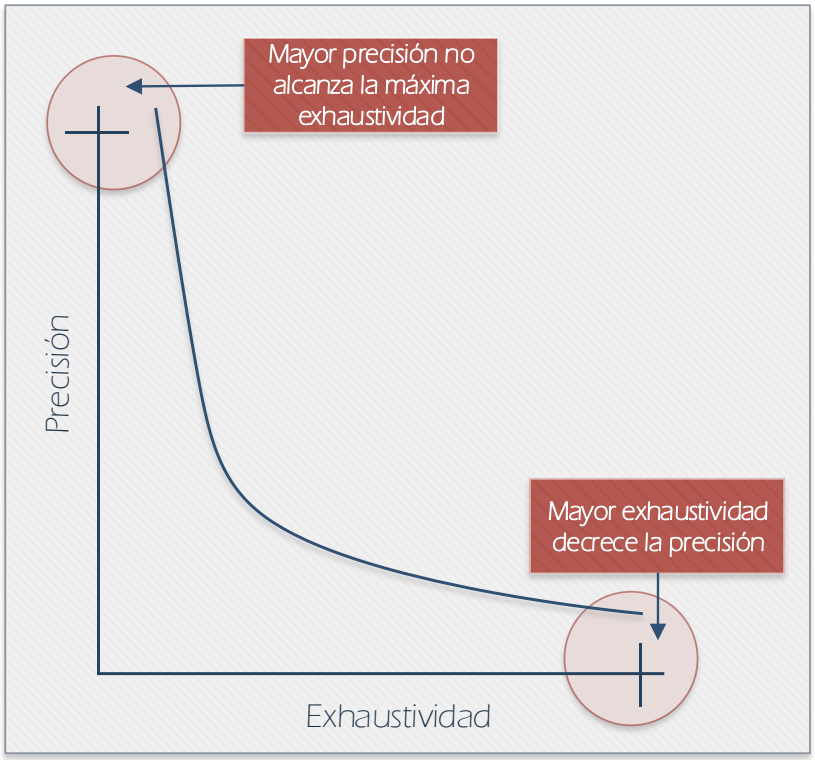

Figura 2.3 *- Relación entre Precisión y Exhaustividad [2]*

Si se recupera una pequeña cantidad de recursos, siendo todos estos relevantes, la precisión es perfecta, pero se descartarían otros recursos relevantes causando un valor bajo de exhaustividad. Por el contrario, si se recuperan todos los recursos relevantes, la exhaustividad es perfecta, pero la precisión es baja por la cantidad de recursos no relevantes recuperados.

Si bien la precisión y la exhaustividad representan un enfoque básico, resultan en un análisis completo y fiable de la respuesta del sistema con respecto a la esperada por el usuario, debido a que involucra su criterio de evaluación. Es por esto, que constantemente se lo selecciona como método de evaluación en ámbitos no solo de la RI, sino que en todos los que se requiera contrastar una respuesta del sistema con respecto a una esperada.

2.2.2 Precisión y exhaustividad para un Top k del ranking

En base a la explicación anterior, se puede extender la aplicación de las métricas de precisión y exhaustividad a los Top k de un ranking de recursos recuperados. De esta manera, la precisión para un Top k determinado, representa la cantidad de recursos relevantes presentes en las k primeras posiciones del ranking, y se calcula mediante la Ecuación (2.3).

$$Precision(k) = \frac{\#RecursosRelevantesRecuperados\ (k)}{k} \tag{2.3}$$

Donde $\#RecursosRelevantesRecuperados\ (k)$ es la cantidad de recursos relevantes presentes en el Top k y k es la cantidad de posiciones del ranking consideradas.

Asimismo, la exhaustividad para un Top k, representa la proporción de la totalidad de recursos relevantes, que se ubican en las primeras k posiciones del ranking, y se obtiene mediante la Ecuación (2.4).

$$Exhaustividad(k) = \frac{\#RecursosRelevantesRecuperados(k)}{\#TotalRecursosRelevantes} \tag{2.4}$$

Donde $\#RecursosRelevantesRecuperados(k)$ es la cantidad de recursos relevantes presentes en el Top k y $\#TotalRecursosRelevantes$ es la cantidad total de recursos relevantes recuperados.

Para una mejor comprensión, se presenta el ejemplo de la Tabla 2.2, considerando un ranking de 10 posiciones, donde 7 de los recursos recuperados son relevantes y 3 son no relevantes.

Tabla 2.2 *- Ejemplo de las métricas de precisión y exhaustividad aplicadas sobre los Top de un ranking*

TOP	Cant. Relevantes	Cant. No Relevantes	Precisión	Exhaustividad
1	1	0	1/1 = 1	1/7 = 0,14
5	3	2	3/5 = 0,60	3/7 = 0,43
10	7	3	7/10 = 0,70	7/7 =1

En primer lugar, se considera el Top 1, en el que se obtiene una precisión de 1, debido a que el primer recurso del ranking es relevante. Además, al recuperar 1 de los 7 recursos relevantes presentes en las diez posiciones del ranking, la exhaustividad es de 0,14.

En el Top 5, se tiene que 3 recursos son relevantes, lo que provoca la disminución de la precisión para el conjunto de las 5 primeras posiciones, obteniendo un valor de 0,60. Además, al recuperar 3 de los 7 recursos relevantes, la exhaustividad aumenta al 0,43.

Al analizar todas las posiciones del ranking (Top 10), se recuperan 7 recursos relevantes, lo que arroja una precisión del 0,70 y una exhaustividad de 1.

A partir de la aplicación de estas métricas, a los distintos Top k de un ranking, es posible generar un gráfico que plasme la relación existente entre la precisión y la exhaustividad, haciendo visible como responde una de ellas, a la variación en el comportamiento de la otra y viceversa.

Este gráfico se obtiene mediante la interpolación de los valores de precisión (eje Y) y exhaustividad (eje X) obtenidos para los Top k evaluados. Para el ejemplo anterior, se muestra en la Figura 2.4 el grafico generado.

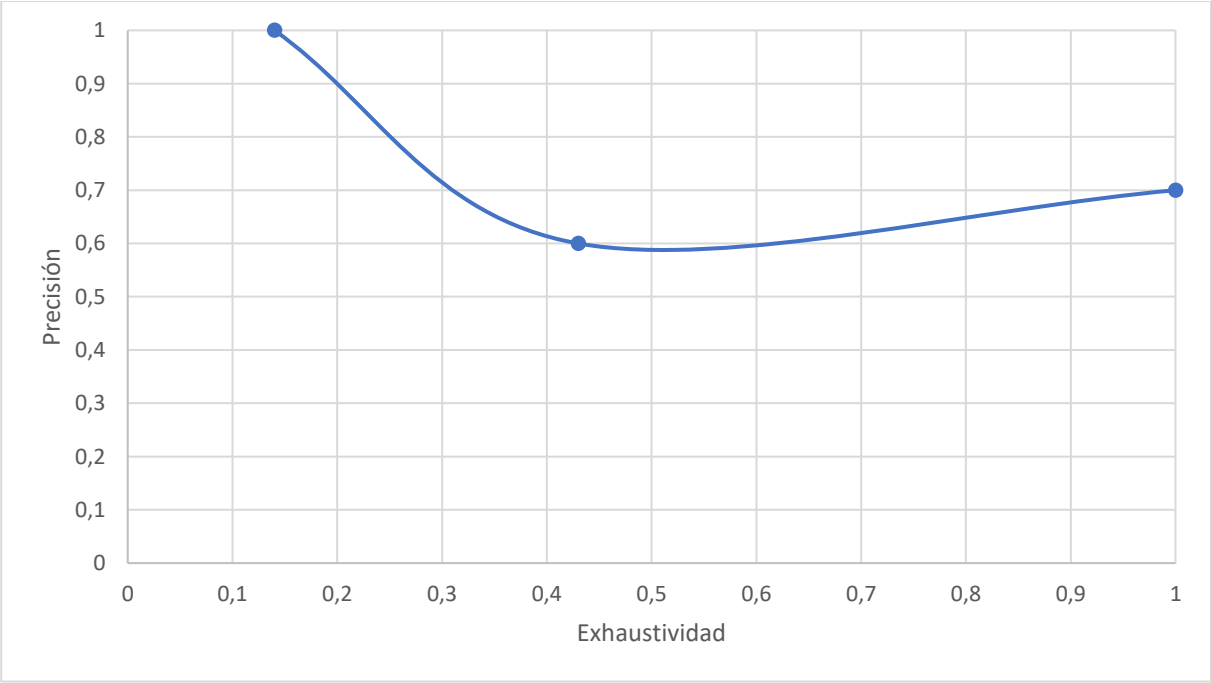

Figura 2.4 *- Gráfico de Precisión y Exhaustividad Interpolada generado a partir del ejemplo de la* ***Tabla 2.2***

Este gráfico permite, evaluar el comportamiento del sistema a medida que se recuperan recursos, comparar el comportamiento de distintos sistemas para una misma clave de búsqueda o comparar el comportamiento de un mismo sistema utilizando distintas claves de búsqueda [21][22].

Las dos herramientas de evaluación presentadas en esta sección, proporcionan elementos de análisis adicionales que permiten una mejor comprensión del comportamiento general del SRI, mostrando la evolución de las métricas a medida que se conforman las posiciones del ranking y permitiendo sectorizar el análisis a realizar. Además, el hecho de utilizar un gráfico que plasme la relación entre estas métricas, permite una representación clara y concisa de los datos obtenidos, facilitando la obtención de conclusiones.

2.3 MODELOS PARA RECUPERACIÓN DE INFORMACIÓN

Un modelo de RI es una representación formal que define como se estima la relevancia de los recursos analizados y, establece el criterio de generación de rankings. Este modelo es mejor cuanto más se aproxime al criterio de determinación relevancia que posee el usuario.

Existen varios enfoques de modelos de RI, que representan distintos puntos de vista. Gran parte de estos, se centran en el cálculo de la probabilidad de que el contenido de los recursos se relacione al contenido de la clave de búsqueda. Un ejemplo, son los algoritmos de ranking basados en el conteo de repetición de palabras.

Por otro lado, enfoques más avanzados se centran en las características lingüísticas del texto, considerando sus propiedades sintácticas y semánticas [2][21].

En las secciones siguientes, se presentan algunos de los modelos de RI más representativos, entre los que se encuentran: el modelo Booleano, el Vectorial, el Probabilístico y el enfoque Semántico.

2.3.1 Modelo Booleano

El modelo booleano es la alternativa de modelos de RI más simple, debido a su básica evaluación de la relevancia. Establece que, dado un recurso D y una clave de búsqueda Q, D es relevante si y solo si, contiene todos los términos definidos en Q [17][23].

Para determinar la relevancia, generalmente se utilizan los operadores unión, intersección y negación, propios de la teoría de conjunto [23].

La principal desventaja del modelo booleano, es que no distingue niveles de relevancia. Esto impide que pueda ser utilizado como método de construcción de rankings, ya que se asume que todos los recursos recuperados son igualmente relevantes.

Por otro lado posee varias ventajas, entre las que se encuentran, su facilidad de comprensión y explicación, y su capacidad de eliminar rápidamente los recursos que considere no relevante [21].

Es por estas ventajas y porque representa un enfoque simple e intuitivo, que se sigue utilizando este modelo en la actualidad. Además, representa un punto de partida, a partir del cual se puede obtener métricas más confiables y con mejores resultados.

2.3.2 Modelo Vectorial

El modelo vectorial, definido por Salton [18], es el más utilizado en operaciones de RI, debido a su simplicidad para el cálculo de peso de términos, armado de rankings y el plasmado

de la retroalimentación [21]. Además, estima el grado de pertinencia de un recurso a una clave de búsqueda, lo que significa, que distingue niveles de relevancia.

De manera general, este modelo representa recursos y claves de búsqueda mediante la utilización de vectores, donde cada uno de ellos se compone por un conjunto de términos y un peso que indica la importancia de cada término (cuyo método de determinación varía según el algoritmo utilizado).

Entonces, la relevancia del recurso analizado, se obtiene calculando la similitud existente entre los dos vectores. Para esto, generalmente se utiliza la distancia del coseno entre ambos, resultando en que un recurso es más relevante cuanto menor sea la distancia existente entre su vector y el de la clave de búsqueda [17].

Sea $D = \{d_1, d_2, \ldots, d_N\}$ el conjunto de recursos recuperados, Q una clave de búsqueda definida por el usuario y $T = \{t_1, t_2, \ldots, t_k\}$ el conjunto de términos pertenecientes a D y Q. Podemos representar un recurso d_i como el vector $d_i \rightarrow \sim d_i = \{w(t_1, d_i), \cdots, w(t_k, d_i)\}$, donde $w(t_r, d_i)$ es el peso del término t_r en el recurso d_i.

De igual manera, se puede representar a la clave de búsqueda como un vector $Q \rightarrow \sim Q = \{w(t_1, Q), \cdots, w(t_k, Q)\}$, donde $w(t_r, Q)$ es el peso del termino t_r presente en la clave de búsqueda Q. Entonces, para calcular la similitud entre un recurso d_i y Q, se utiliza la distancia del coseno, obtenida mediante la Ecuación (2.5) [23].

$$Coseno(d_i, Q) = \frac{\sum_{r=1}^{k} w(t_r, d_i) * w(t_r, Q)}{\sqrt{\sum_{r=1}^{k} w(t_r, d_i)^2 * \sum_{r=1}^{k} w(t_r, Q)^2}} \tag{2.5}$$

En Definitiva, este modelo consta de un grado de confianza superior al enfoque booleano, debido a la distinción de niveles de relevancia y la consideración de los pesos de términos que los discriminan de acuerdo a su importancia. Son estos aspectos, los que justifican su uso en una gran cantidad de investigaciones y fundamentan que aun en la actualidad se sigan utilizando en muchos desarrollos de SRI.

2.3.3 Modelo Probabilístico

El modelo probabilístico, propuesto en 1976 por Robertson y Sparck Jones [24], plantea una solución al problema de la RI basada en un marco probabilístico. Utiliza la representación binaria de recursos, al igual que el modelo booleano, indicando presencia o ausencia de términos mediante 0 y 1. La diferencia, radica en el uso de la teoría probabilística y en las premisas bajo las que se constituye su funcionamiento, las cuales son [17]:

- Según la clave de búsqueda planteada por el usuario, los recursos de la colección se clasifican en dos grupos: el conjunto de recursos relevantes y el conjunto de recursos no relevantes.
- Existe una respuesta ideal del sistema, constituida por el conjunto de recursos relevantes, a la que se denomina conjunto de respuesta ideal.
- Existe una clave de búsqueda ideal, que es aquella que proporciona un conjunto de respuesta ideal.
- A priori se desconocen los términos que deberían pertenecer a la clave de búsqueda para obtener el conjunto de respuesta ideal.

El objetivo del modelo probabilístico, es refinar constantemente la clave de búsqueda del usuario hasta obtener la clave de búsqueda ideal, mediante la reformulación sucesiva de sus términos y sus ponderaciones.

El proceso de ponderación de los términos consiste en el cálculo de la probabilidad de que pertenezcan al conjunto de recursos relevantes y la probabilidad de que pertenezcan al conjunto de los no relevantes, obtenido mediante el teorema de Bayes [25].

Luego se calcula la similitud entre la clave de búsqueda y los recursos de la colección, mediante alguno de los modelos antes descriptos, permitiendo ordenarlos de acuerdo a la relevancia con respecto a la clave de búsqueda del usuario [3].

Como se puede observar, el modelo probabilístico representa un enfoque confiable debido al aval de los métodos probabilísticos que son la base de su implementación. Sin embargo, todavía existen ciertos aspectos que generan duda en cuanto a cómo utilizar el teorema de Bayes para calcular las probabilidades necesarias y como estimar el puntaje de relevancia del recurso.

2.3.4 Enfoque Semántico

El enfoque semántico surge para evitar la restricción que imponen las técnicas anteriores, en cuanto a que un recurso es relevante si en su contenido posee ocurrencias explicitas de las palabras de la clave de búsqueda. Esta restricción se transforma en problema debido a que un recurso puede ser relevante sin necesariamente contener dichas palabras en su contenido.

Para dar solución a este problema, se hace foco sobre las características semánticas de las palabras que conforman a los recursos y la clave de búsqueda, como método para calcular los puntajes de relevancia.

En general, desde el punto de vista de la semántica, cada palabra posee un significado independiente del contexto, denominado denotación, un significado dependiente del contexto, denominado connotación o sentido, y un conjunto de relaciones semánticas que también dependen del contexto [26].

Teniendo como base esta definición, la aproximación planteada por el enfoque semántico es determinar el puntaje de relevancia, considerando las relaciones semánticas existentes entre el contenido de los recursos y la clave de búsqueda.

Para determinar estas relaciones, existen dos aproximaciones consideradas en la actualidad: la aproximación basada en corpus y la aproximación basada en conocimiento [27].

La aproximación basada en corpus permite la determinación de relaciones semánticas, a partir del análisis de unidades de lenguaje extraídas desde grandes corpus generados a partir de combinaciones de textos aleatorios. Mediante dicho corpus, se obtienen valores de frecuencia de repetición de palabras, frecuencias de palabras yuxtapuestas, entre otros, útiles para los cálculos a realizar.

El fundamento de esta aproximación, tiene origen en la hipótesis distribucional, que busca definir como el análisis estadístico de grandes corpus de texto, distribuciones de palabras y regularidades estadísticas ligadas a contextos lingüísticos, pueden ser utilizados para modelar la semántica de las palabras. Para ello, se basa en las siguientes suposiciones:

- El contexto asociado a una palabra está determinado por las palabras que la rodean.
- Palabras que ocurren en contextos similares, tienen más probabilidad de estar relacionadas semánticamente.

La técnica de la aproximación basada en corpus, más conocida y ampliamente utilizada, es el análisis semántico latente (LSA - *Latent Semantic Analysis*) [28], que representa el perfil de una palabra determinada, mediante matrices de contexto, que

contienen la palabra original y las que se relacionan a ella en una determinada connotación. Entonces, para obtener el valor de relación semántica entre dos palabras, utiliza el cálculo de la similitud del coseno entre sus matrices de contexto.

Otros ejemplos de esta aproximación, son el análisis semántico explicito (ESA – *Explicit Semantic Analysis*) [29], hiperespacio análogo al lenguaje (HAL – *Hyperspace Analogue to Language*) [30], entre otros [27].

La aproximación basada en conocimiento, por su parte, basa su análisis en la representación del conocimiento de expertos mediante ontologías, por lo que su eficacia depende de que tan completas y correctas sean estas.

Las ontologías son un tipo de taxonomía, que estructuran en clases a distintos elementos que poseen características similares. Por lo tanto, su uso puede ser extendido para representar los distintos contextos de las palabras y plasmar sus relaciones semánticas.

Existen diversas ontologías semánticas, que representan los distintos contextos de las palabras y sus correspondientes relaciones semánticas, entre las que se encuentran WordNet [31], ConceptNet [32], etc. Estas generalmente poseen una estructura de árbol, donde por cada nivel se tiene un nodo padre, que es la palabra que engloba a las que se encuentran por debajo (hiperónimo) y nodos hijos, que subsumen al nodo padre (hipónimos).

Además, cada nodo posee un conjunto de palabras que tienen el mismo significado y que por lo tanto pertenecen al mismo contexto (sinónimos). También, vale aclarar que dado el hecho de que una palabra puede tener distintos sentidos de acuerdo al contexto al que pertenece (fenómeno denominado polisemia), es factible que tengan varias apariciones en la ontología [27].

Estas ontologías, proporcionan un entorno simple para aplicar métricas que permitan evaluar la relación semántica existente entre las palabras. Existe un gran conjunto de métricas, entre las que se encuentran: la métrica de Wu and Palmer [33], Li [34], Lin [35], Resnik [36], entre otras [27].

En definitiva, el enfoque semántico representa un análisis de recursos más complejo, en el que se consideran más factores que en los modelos presentados en las secciones anteriores. Esto hace suponer que tendrán incidencia en la mejora de resultados.

En cuanto a las dos aproximaciones presentadas, se puede apreciar que la basada en corpus, al adquirir conocimiento de manera automática, posee menor precisión en la determinación de relaciones semánticas que la basada en conocimiento.

Sin embargo, la aproximación basada en corpus no se limita por la cantidad de palabras y relaciones existentes como si ocurre en el caso de la aproximación basada en conocimiento. Por lo tanto, la aproximación a utilizar debe ser seleccionada de acuerdo a las necesidades que se posean, es decir, optar por la precisión de la aproximación basada en conocimiento o la cantidad de palabras y sus relaciones que contempla la basada en corpus.

Como el enfoque semántico, mediante la aproximación basada en conocimiento, es el utilizado para desarrollar el modelo que se propone en este trabajo, en el capítulo 3 se presentan con mayor detalle sus conceptos relacionados.

2.4 SISTEMA BASE

En [5][6][7] se describe un modelo de RI que tiene dos objetivos bien definidos: el primero consiste en capturar de manera precisa la necesidad de información que posee el usuario y presentar los resultados obtenidos.

El segundo, consiste en la implementación de métodos y técnicas que permitan la identificación, recuperación y análisis continuo de recursos existentes en la Web.

El esquema general, que representa el funcionamiento de este sistema, se muestra en la Figura 2.5. En principio, los dos objetivos mencionados, son implementados mediante el ***Módulo de Recopilación de Requerimientos*** y el ***Módulo de Minería Web***, respectivamente.

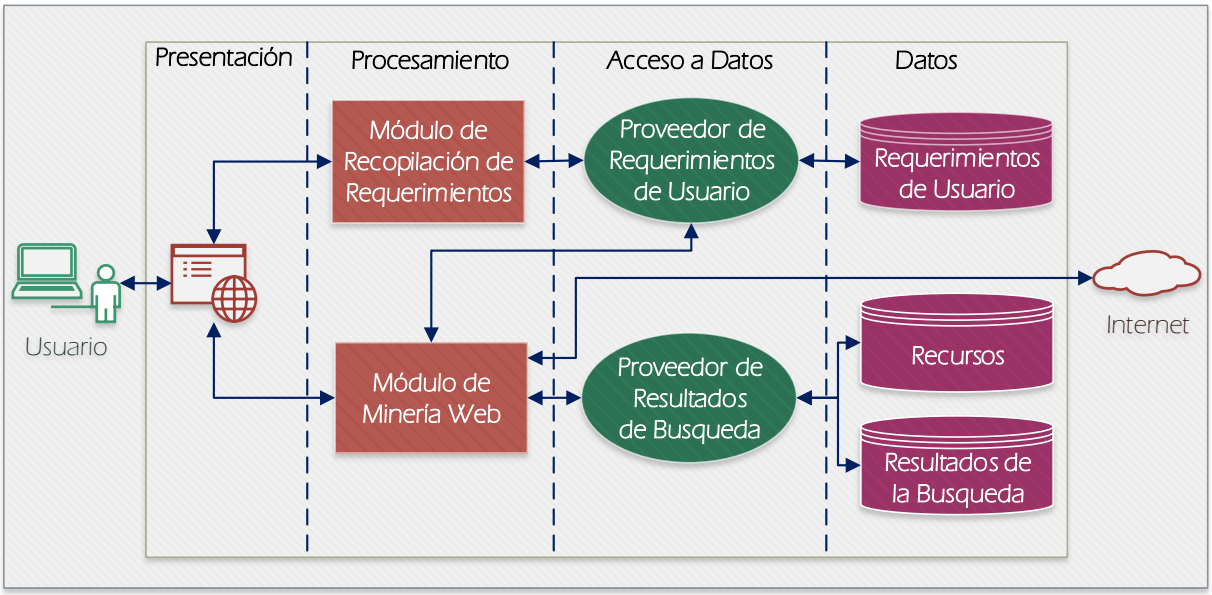

Figura 2.5 *- Esquema general del sistema implementado [6]*

El sistema inicia su funcionamiento, cuando el usuario ingresa una clave de búsqueda. Durante este proceso, el usuario debe responder una serie de preguntas de ámbito general, tales como características adicionales al tema principal de búsqueda, año o década en la que centrar la búsqueda, tipo de resultados que desea, que pueden ser científicos o no, etc.

Estas preguntas contribuyen a aumentar la precisión en la definición de la necesidad de información que el usuario posee. Como resultado, se obtiene un conjunto de claves de búsqueda, donde cada una está completamente estructurada y optimizada para ser utilizada en motores de búsqueda.

Estas claves, se envían al ***Módulo de Minería Web***, iniciando así su funcionamiento. En la Figura 2.6 se presenta el diagrama de bloques de este módulo, en el que se puede observar los módulos que lo componen y el intercambio de información entre estos.

Inicialmente se reciben las claves en el ***Webminer Controller***, que es el que controla todo el funcionamiento del ***Módulo de Minería Web***.

Las claves se envían al ***Módulo de Búsqueda***, donde se obtienen las URL de los primeros diez resultados de los buscadores Google, Bing, Intelligo (Buscador de Patentes) y Msmlx excite (metabuscador). Estas URL's se retornan al ***Webminer Controller***, donde se conforma una única lista, sin repeticiones.

La lista obtenida, se envía al módulo ***Web Crawler***, donde se la utiliza para el proceso continuo e iterativo de exploración y descubrimiento de enlaces. Como resultado, se obtiene un conjunto de grafos, donde cada URL de la lista es la raíz de un grafo, los enlaces descubiertos a partir de ella son los nodos hijos y las aristas que las unen representan las relaciones entre ellos.

Cada grafo es enviado al ***Módulo de Recuperación de Información*** para establecer una puntuación a los recursos web recuperados, con el fin de obtener una lista ordenada de

acuerdo a su nivel de relevancia. Para esto, se utilizan tres algoritmos de determinación de relevancia (presentados en la sección 2.4.1), obteniendo por cada uno de ellos, una lista de recursos ordenados de acuerdo a su criterio de relevancia.

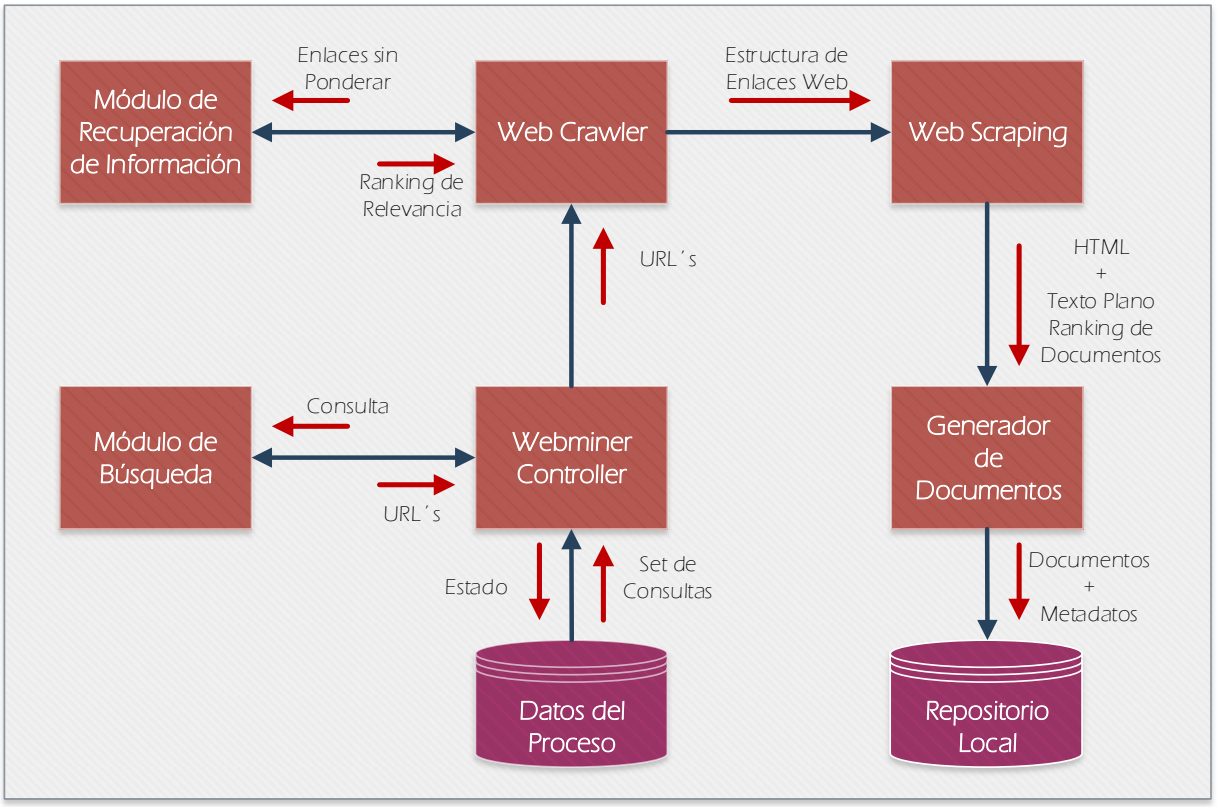

Figura 2.6 *- Diagrama de Bloques del Módulo de Minería Web [6]*

Por cada recurso se computa su posición en cada lista, y luego, mediante la utilización de un método de unificación de rankings, se obtiene un puntaje final, que es utilizado para confeccionar una única lista ordenada de manera descendente a partir del puntaje obtenido.

Esta lista, se envía al módulo ***Web Scraping***, donde se descarga el contenido de los primeros 50 recursos de la lista ordenada y se generan los metadatos necesarios para la presentación de los resultados.

Luego, en el módulo ***Generador de Documentos*** se crean los documentos dependiendo de la extensión (.pdf, .html, .asp, .php, etc.) y se los asocia con los metadatos, para almacenarlo en un directorio que está sincronizado con todas las ubicaciones en las que se requiere la información (Repositorio Local).

Finalizado este paso, se reinicia el proceso, tomando otra URL de la lista y descubriendo sus enlaces relacionados, mediante el módulo ***Web Crawler***. El proceso de este módulo finaliza cuando el usuario decide parar con la búsqueda de recursos o cuando no existen más URL's por explorar.

Como uno de los objetivos del presente trabajo, es complementar a los algoritmos de determinación de relevancia lexicográficos de este sistema base, mediante la utilización de técnicas de análisis semántico, se toma como punto de partida a la arquitectura presentada en esta sección.

2.4.1 Algoritmos de determinación de relevancia utilizados en el Sistema Base.

En el sistema base, presentado en la sección anterior, se utilizaron tres algoritmos de determinación de relevancia. Estos son, el Modelo de Espacio vectorial, el Modelo Okapi BM25 y el Modelo C-Rank, y se presentan a continuación:

- **Modelo de Espacio Vectorial (VSM):** Cada recurso de la colección está representado por un vector t-dimensional, donde t es la cardinalidad del conjunto de términos en el corpus de recursos y cada elemento del vector posee un peso del término asociado a esa dimensión. De igual manera, la clave de búsqueda también se representa mediante un vector t-dimensional. Para calcular la similitud entre el recurso y la clave de búsqueda, generalmente se utiliza la similitud del coseno, que determina la distancia existente entre los vectores generados para ambos [37].
- **Modelo Okapi BM25:** Es un modelo probabilístico que incorpora en su cálculo la frecuencia de aparición de términos de la clave de búsqueda, la longitud promedio de los recursos de toda la colección y la longitud del recurso analizado. Dada una clave de búsqueda Q, cuyos términos son $Q = \{q_1, \dots, q_n\}$, el puntaje de relevancia otorgado por el modelo Okapi BM25 para un recurso R, se determina a partir de la Ecuación (2.6) [38][39].

$$Okapi(R,Q) = \sum_{i=1}^{n} \left(\log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \right) * \frac{f(q_i,R) * (k+1)}{f(q_i,R) + k * \left(1 - b + b * \frac{|R|}{avglr}\right)} \qquad (2.6)$$

  Donde N es el número total de recursos en la colección, $n(q_i)$ es el número de recursos que poseen al término q_i de la clave de búsqueda, $f(q_i,R)$ es la frecuencia de aparición del término q_i en el recurso R, $|R|$ es la longitud del recurso R (cantidad de palabras) y $avglr$ es la longitud promedio de los recursos de la colección. k y b son constantes definidas para adecuar la función a características concretas de la colección de recursos. En este sistema base, se establecieron los valores $k = 2$ y $b = 0{,}75$, determinados como óptimos a partir de resultados experimentales.

- **Modelo C-Rank**: Para la determinación de la relevancia, este modelo considera al contenido del recurso analizado y al contenido de sus recursos relacionados. Como resultado se obtiene que un recurso es más relevante, cuanto más relacionado esté a la clave de búsqueda y cuanta más correspondencia exista entre la clave y sus recursos relacionados. Para obtener el puntaje de relevancia de un recurso R, se utiliza la Ecuación (2.7) [40].

$$Crank(R) = Rel_R * \lambda + Cont_r * (1 - \lambda) \qquad (2.7)$$

  Donde Rel_R es la relevancia del recurso, determinada mediante el método TF/IDF y, $Cont_r$ es la contribución de los recursos relacionados a R. λ es una constante de proporcionalidad, que establece el aporte al puntaje final, de Rel_R y $Cont_r$. En este sistema base, se estableció el valor de $\lambda = 0{,}8$ recomendado en [40].

CAPÍTULO 3
ANÁLISIS SEMÁNTICO

Este capítulo tiene el objetivo de brindar un marco teórico que sirva de sustento a las técnicas de análisis semántico que son implementadas en el presente trabajo. Para ello, Inicialmente se da una introducción a los conceptos de relación, similitud y distancia semántica. Luego, se presenta una clasificación de las distintas relaciones semánticas existentes, planteadas por la lingüística y otras áreas, que fueron extendidas para su uso en el procesamiento del lenguaje natural. Seguidamente, se introducen a las taxonomías semánticas, que son las herramientas utilizadas por el modelo propuesto. Finalmente se da a conocer un conjunto de métricas de relación y similitud semánticas de pares de palabras, aplicables sobre las taxonomías a utilizar.

3.1 RELACIÓN, SIMILITUD Y DISTANCIA SEMÁNTICA

Un problema que atañe al área del Procesamiento del Lenguaje Natural (NLP – *Natural Language Processing)* es la Recuperación de Información (RI), que consiste entre otras cosas, en la construcción de rankings de recursos ordenados de acuerdo a la relevancia, con respecto a una clave de búsqueda ingresada por un usuario.

El enfoque semántico, propone lidiar con este problema, determinando la relevancia de los recursos a partir de su correspondencia semántica con la clave de búsqueda. Para ello, propone tener en cuenta las relaciones semánticas entre pares de palabras, pertenecientes al recurso y la clave de búsqueda, basándose en el planteo de la lingüística de que cada palabra se compone por un significado independiente del contexto (denotación), un significado propio del contexto en el que se la utiliza (connotación o sentido) y relaciones con otras palabras, propias de ese contexto [26][41].

Desde la informática, se han diferenciado tres definiciones útiles para la evaluación de la correspondencia semántica entre pares de palabras, ellos son: la relación semántica, la similitud semántica y la distancia semántica [42].

En este sentido, Resnik [36] intentó demostrar la diferencia entre las dos primeras definiciones con la utilización del siguiente enunciado: "*car* (automóvil) y *gasoline* (gasolina) deberían estar más cercanamente relacionados, que *car* (automóvil) y *bicycles* (bicicleta), pero este último par es ciertamente más similar".

En otras palabras, la relación semántica hace referencia a todas las relaciones posibles existentes entre dos palabras y la similitud semántica es un caso especial de relación semántica, en la que solo se tienen en cuenta las relaciones de sinonimia e hiperonimia. Es evidente que la relación semántica es más general que la similitud semántica, ya que palabras no similares pueden estar semánticamente relacionadas como sucede, por ejemplo, con la relación de meronimia entre *car* (automóvil) y *wheel* (rueda).

El término distancia semántica, por otro lado, puede ser utilizado cuando se hable tanto de relación como de similitud semántica. Plantea que existe una mayor relación semántica entre palabras, cuando la distancia existente entre ellas es menor. Sin embargo, no siempre es un indicador fiable de que exista una relación entre dos palabras, como sucede en el caso de la relación de antonimia, en la que dos palabras se encuentran distantes una de la otra, pero en realidad, están relacionadas semánticamente [32][43].

En la Tabla 3.1 se presenta un resumen de los tres conceptos presentados en esta sección, plasmando las relaciones semánticas que contemplan cada uno de ellos, y cómo diferenciar en qué caso se manifiesta uno u otro.

Tabla 3.1 *- Resumen de Relación, Similitud y Distancia Semántica*

	Relaciones Contempladas	¿Cómo se manifiesta?
Relación Semántica	• Relaciones clásicas (Sinonimia, antonimia, meronimia, etc.) • Asociaciones funcionales • Relaciones no clásicas [9]	Con la existencia de algún tipo de relación semántica entre dos palabras. Ej.: Hueso es merónimo de Brazo.
Similitud Semántica	• Sinonimia • Hiperonimia	Cuando las dos palabras pertenecen a la misma clase, o poseen el mismo significado. Ej.: "Automóvil" y "Bus", pertenecen a la clase "Vehículo".
Distancia Semántica	• No tiene en cuenta relaciones.	Cercanía o lejanía entre palabras.

Los conceptos presentados en esta sección, proporcionan un marco sobre el cual basarse para poder estimar la relación y similitud semántica entre pares de palabras, que ya de por si son conceptos abstractos y, por lo tanto, complejos.

Existen diversos métodos desarrollados para estimarlos, que varían en la precisión obtenida, debido a la subjetividad implicada en la estimación de que tan relacionadas se encuentran dos palabras.

No obstante, la eficacia de estos métodos, permiten utilizarlos en modelos que evalúen la correspondencia semántica existente entre dos conjuntos de palabras, como ser entre un recurso Web y una Clave de Búsqueda, lo que los hace útiles para Sistemas de Recuperación de Información (SRI).

3.2 TIPOS DE RELACIONES SEMÁNTICAS

Una parte del campo de la semántica, consiste en el estudio de las distintas relaciones semánticas existentes entre palabras enmarcadas en un contexto específico (connotación). La importancia de comprender estas relaciones radica en su contribución a la comprensión de textos.

En síntesis, pares de palabras relacionadas pueden juntarse para formar grandes grupos de palabras que pueden extenderse incluso en la totalidad de una oración y contribuir a su vez, al significado y sentido del texto en su totalidad [9]. Asimismo, Cruse [44] afirma que el sentido de una palabra puede verse reflejado en sus relaciones contextuales, haciendo notar, que estas relaciones contribuyen a la definición de la connotación.

Las relaciones semánticas de las palabras, han sido de gran interés por áreas de la filosofía, psicología cognitiva, lingüística, científicos de la computación y otras áreas cuyo interés esté enfocado en las palabras, sus significados o cómo funciona la mente [45].

Un aporte a la distinción de tipos de relaciones semánticas surge a partir del trabajo presentado por Morris y Hirst [9], en el que se plantea una separación en dos grandes categorías de relaciones semánticas: Las relaciones semánticas clásicas y las no clásicas.

Al hablar de relaciones semánticas clásicas, se hace referencia a todas aquellas relaciones que surgen como producto de las características compartidas entre palabras, como ser las relaciones de sinonimia, hiperonimia, meronimia, etc. En cambio, las relaciones semánticas no clásicas, son relaciones entre palabras que no dependen de las características compartidas entre ambas.

A continuación, tomando la clasificación propuesta en [9], se pormenorizará cada uno de estos tipos de relaciones, dividiéndolos en dos categorías, relaciones clásicas y no – clásicas.

3.2.1 Relaciones semánticas Clásicas

El primer tipo de relaciones semánticas a estudiar son las Relaciones Semánticas Clásicas, que agrupa a todas aquellas que se dan entre dos palabras que pertenecen a la misma clase sintáctica, es decir, relaciones entre pares de verbos, pares de sustantivos, etc. En la Figura 3.1 se presentan las distintas relaciones pertenecientes a esta categoría.

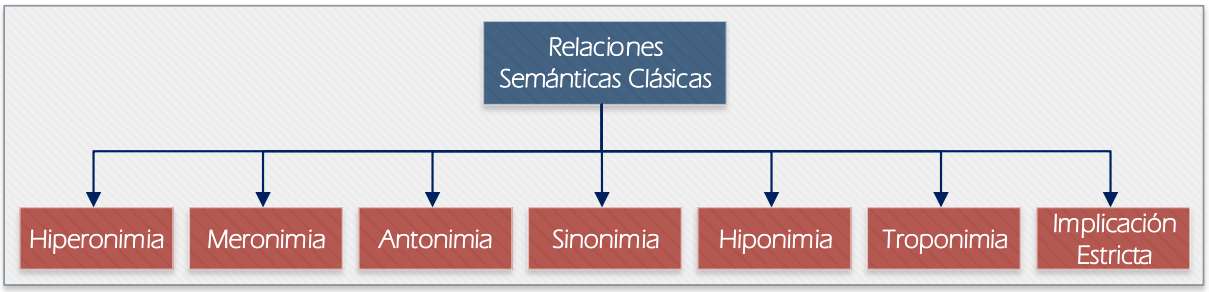

Figura 3.1 *- Tipos de Relaciones Semánticas*

Según varios autores, este tipo de relaciones se corresponden con las relaciones semánticas paradigmáticas, ya que agrupan a conjuntos de palabras que forman algún tipo de paradigma. En este caso, los paradigmas son las clases sintácticas de las palabras, por lo que se tienen conjuntos de palabras pertenecientes al paradigma sustantivo, al paradigma adjetivo, etc.

Las relaciones semánticas clásicas tienen la particularidad de que son sustituibles por otro miembro perteneciente al mismo paradigma, como por ejemplo, la frase “una silla ___” la puede completar cualquiera de los miembros pertenecientes al paradigma “adjetivos”, entre los que se encuentran azul, baja, linda, etc. [44][45].

Este tipo de relaciones son las ampliamente conocidas por la lingüística, lo que implica una vasta teoría y fundamento sobre ellas. Sin embargo, dado el hecho que pueden existir infinitas relaciones entre los significados y sentidos de las palabras, es evidente que son limitadas, abarcando solo una porción de los tipos de relaciones existentes.

Hiperonimia

La hiperonimia es la relación que se expresa por la aserción $A\ es\ un\ B$, donde A es un tipo de palabra específico y B es un tipo de palabra general. Un ejemplo de esta relación es: Gerente es un Empleado [46].

Las relaciones de hiperonimia son un tipo de relaciones jerárquicas, donde un elemento ubicado más arriba en la jerarquía, “engloba” a otro ubicado más abajo en la misma jerarquía. Un ejemplo de hiperonimia, se presenta en la Figura 3.2, donde Gerente y Operario poseen una relación, indicando que son tipos de Empleado.

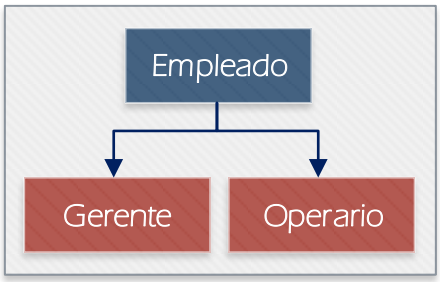

Figura 3.2 *- Estructura jerárquica en la que se refleja la relación de hiperonimia*

Con respecto a esto, Chaffin [47] ha identificado los siguientes cuatro tipos hiperonimia, presentadas en la Tabla 3.2.

Tabla 3.2 *- Tipos de relaciones de hiperonimia [46]*

Tipo	Ejemplo
Objeto Natural - Tipo	Empleado es una Persona
Artefacto - Tipo	Microcomputadora es una Computadora
Estado – Tipo	Soltero es un tipo de Estado Marital
Actividad - Tipo	Asesoría es un tipo de Trabajo

Tomando como base este ejemplo, la relación de hiperonimia puede ser vista como una categorización, donde las palabras ubicadas más abajo en la jerarquía, forman parte de la categoría determinada por la palabra que se ubica más arriba en la jerarquía.

Hiponimia

Se denomina hiponimia al tipo de relación entre palabras específicas y generales, en la que una de ellas está incluida en la otra. Por ejemplo, Gato es hipónimo de Animal.

La hiponimia es el punto de vista opuesto a la hiperonimia explicada en la sección anterior, debido a que se trata de una palabra que pertenece a una categoría determinada por otra más general. En el ejemplo presentado, Animal es un hiperónimo de Gato [48].

El conjunto de términos que son hipónimos del mismo hiperónimo, son denominados cohipónimos. Un ejemplo de esto se puede ver en la Figura 3.3, en el que se observan tres palabras, Ómnibus, Automóvil y Camión, que son hipónimos de un mismo hiperónimo, que es Vehículo.

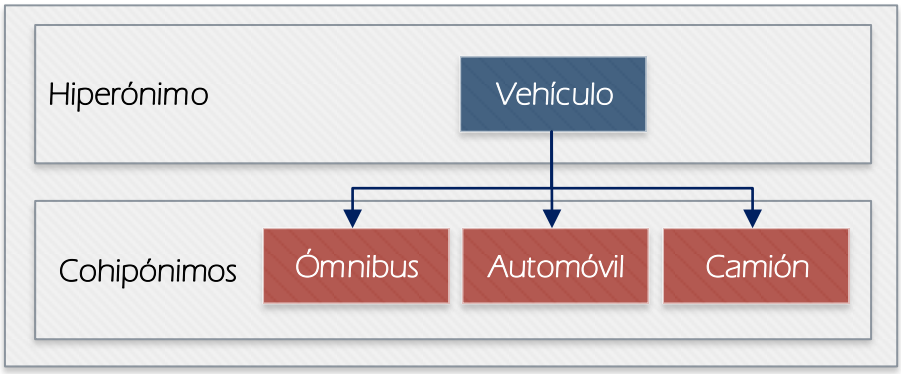

Figura 3.3 *- Ejemplo de relación entre Hiperónimo e Hipónimo*

Meronimia

La meronimia (del griego meros, que significa parte) es una relación en la que una palabra representa una parte de otra, por ejemplo, Rueda es merónimo de Automóvil. A continuación, se describen siete tipos de relaciones de meronimia, propuestos en [46]:

- **Componente - Objeto:** Relaciones entre un componente y el objeto al que forma parte. Ejemplo: Motor es componente de Automóvil.
- **Característica – Evento:** Un ejemplo de esto puede ser Acto Trapecista es parte del Circo. Los eventos pueden tener partes (características) que ocurren en diferentes momentos en el tiempo, cosa que lo diferencia del objeto, que requiere que ambas cosas ocurran en simultáneo.
- **Miembro – Colección:** Se vincula con las relaciones Pertenece - a, donde un conjunto de miembros es considerado un objeto. Un ejemplo de este tipo de relaciones podría ser Empleado es miembro de Comité. Difiere de la relación Componente – Objeto, en que no se asume que el objeto pueda cumplir una función en la colección, sino que solamente forma parte de ella.
- **Porción** – **Masa:** En una relación de Porción – Masa, la parte es similar a todas las otras partes y a la totalidad, por ejemplo, Porción de Torta es una parte de una Torta. Cada Porción de Torta es Torta y es similar a las otras porciones y a la Torta en su totalidad.
- **Fase** – **Actividad:** Relaciona una fase a una actividad o proceso, y se diferencia de la relación Característica-Evento, en que las fases no pueden ser separadas de la actividad (mientras que una característica si puede ser separada de su evento). Un ejemplo de este tipo de relación puede ser Adolescencia es parte del Crecimiento.
- **Lugar – Área:** Relaciona un área y un lugar o localización especial dentro de él. Por ejemplo, Central Park está ubicado en Nueva York. Se diferencia de la relación Porción – Masa, en que un lugar no puede ser separado del área.
- **Cosas – Objeto:** Relaciona una cosa que constituye a un objeto, pero difiere con la relación componente – objeto, en que la cosa no puede ser separada físicamente del objeto, sin alterar su identidad (como si puede ocurrir con los componentes). Ejemplo: Bicicleta es en parte Aluminio.

Antonimia

La antonimia es la relación semántica que existe entre dos o más palabras que tienen significado opuesto. Pares de palabras antónimas, son las que comparten todas las características semánticas, excepto una. Esa característica no compartida está presente, en una palabra, pero ausente en otra, un ejemplo de esto se presenta en la Figura 3.4.

Figura 3.4 *- Ejemplo de antónimos*

Basados en la literatura, existen tres tipos de antónimos, los cuales se describen a continuación [41][49]:

- **Antónimos complementarios o contradictorios:** Son pares de palabras en las que un miembro tiene una cierta propiedad semántica que otro miembro no tiene.

Por lo tanto, en el contexto en que un miembro es verdadero, el otro debería ser falso. Ejemplo: Masculino/Femenino.

- **Antónimos Relacionales:** Pares de palabras en las que la presencia de una cierta propiedad semántica en una implica la presencia de una propiedad semántica distinta en la otra. Es decir, la existencia de una palabra implica la existencia de otra. Ejemplo: Sobre/Debajo.
- **Antónimos escalares:** Pares de palabras que son contrastadas con respecto al grado de posesión de una cierta propiedad semántica. Cada palabra representa un punto final o extremo en una escala. (por ejemplo, temperatura, tamaño, altitud, etc.). Entre estos extremos existen otros puntos intermedios. Ejemplo de este tipo de antónimos son: Caliente/Frio, Grande/Pequeño, etc.

Además, los antónimos, pueden ser morfológicamente no relacionados, es decir, que los elementos del par de palabras no derivan el uno del otro (Ejemplo: Bueno/Malo); o morfológicamente relacionados, donde una de las palabras del par de antónimos es derivada del otro miembro (Ejemplo: Igual/Desigual).

Sinonimia

A la relación entre palabras, que poseen el mismo significado, se la denomina sinonimia. Para que dos palabras sean sinónimas, no es necesario que sean idénticas en significados, ni que posean connotaciones idénticas (a esto se lo denomina sinonimia total).

Se puede decir que una relación es de sinonimia, si ambas palabras están lo suficientemente cerca en cuanto a significados, como para permitir que se haga una elección entre ellas, sin que exista ninguna diferencia para el significado de la oración como un todo [48]. En otras palabras, la sinonimia es la equivalencia semántica entre ítems léxicos, donde ambos ítems comparten todas sus propiedades semánticas. Ejemplos de relaciones sinonímicas son: Esconder/Encubrir, Cesar/Frenar, etc.

Debido a que la sinonimia representa la cercanía o igualdad entre los significados de las palabras, se dice que este es el tipo principal de relaciones semánticas.

Implicación estricta

La implicación estricta (o simplemente implicación), es una propiedad de las proposiciones lógicas, que fue extendida para relacionar semánticamente a dos verbos.

Dados los verbos $V1$ y $V2$, existe una implicación estricta cuando la acción representada por $V1$ implica la realización de la acción representada por $V2$. La implicación estricta es una relación unilateral, es decir, si $V1$ implica a $V2$, no puede darse que $V2$ implique al $V1$. Si ambos verbos se implican mutuamente, entonces se trata de una relación de sinonimia [50].

Un ejemplo de implicación estricta, se muestra en la Figura 3.5, entre los verbos roncar y dormir, ya que, para roncar se debe estar dormido, pero, dormir no implica roncar.

Figura 3.5 *- Ejemplo de relación de implicación estricta entre los verbos roncar y dormir*

Troponimia

La troponimia es un tipo de relación semántica que plantea que, dado un verbo $V1$ específico y un verbo $V2$ general, la actividad que representa $V1$ implica la realización de $V2$,

siempre que sean temporalmente co-ocurrentes y estén involucradas en una relación jerárquica.

Este tipo de relación es análoga a la relación de hiponimia, descripta en secciones anteriores. Un ejemplo se puede observar en la Figura 3.6, entre los verbos balbucear y hablar, ya que al balbucear también se está hablando y hablar es un verbo general que incluye a balbucear. Lo mismo sucede con los verbos susurrar y gritar.

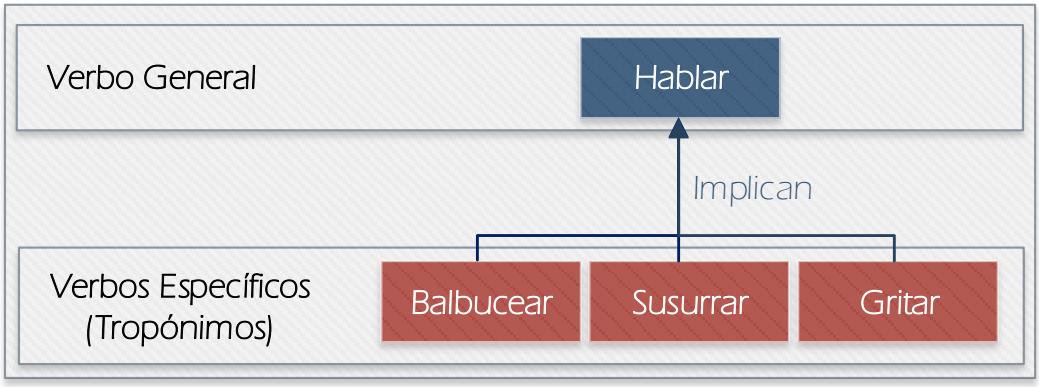

Figura 3.6 *- Ejemplo de relación de troponimia*

Esto deja en evidencia que la troponimia es un caso especial de implicación, exclusivo de pares de verbos que ocurren al mismo tiempo y en donde existe una relación jerárquica entre un verbo especifico y uno global que lo incluye [50].

3.2.2 Relaciones Semánticas No Clásicas

Las relaciones semánticas no clásicas, son relaciones que no dependen de las características compartidas entre palabras.

Una primera aproximación a las relaciones no clásicas, surge a partir del planteamiento de Barsalou [51], que propuso la creación de categorías independientes, que fueran “inventadas sobre la marcha para algún propósito inmediato”, lo que implica considerar la interacción de palabras dentro de un texto especifico, en lugar de suponer que estas pertenecen a una categoría especifica. Un ejemplo puede ser: cosas que llevar a un campamento.

En general, al hablar de relaciones no clásicas, se hace referencia a las distintas formas de relacionar objetos, acciones y actividades, sin la restricción de que estas relaciones sean entre palabras de una misma clase sintáctica.

Este tipo de relaciones, generalmente no posee un nombre que identifique como se relacionan las palabras, sino que está representada y fundamentada por el nombre de la categoría a la que pertenece. Un ejemplo de esto puede verse en la relación: Pelota/Campo o Pelota/Umpire, que pertenecen a la categoría Baseball.

Los siguientes, son los tipos de relaciones no clásicas más conocidas, encontradas en la literatura, que no son mutuamente exclusivas [9]:

- **Relaciones entre miembros de categorías no clásicas de Lakoff**: Pelota, Campo y Umpire son parte de una actividad estructurada Baseball [52].
- **Relaciones de Caso:** Describen cosas que son típicamente verdaderas en el mundo real, por ejemplo, cosas que los agentes usan, o actividades que realizan. Generalmente refieren a relaciones de uso. A continuación, se describen tres tipos de relaciones de caso [46][53]:

  - **Relaciones que involucran agentes:** Relaciones agente-acción (Ej.: Programador realiza Programación), agente-instrumento (Ej.: Programador usa Computadora) y agente-objeto (Ej.: Carpintero usa Madera).
  - **Relaciones que involucran acciones:** Los tipos de relaciones existentes son acción – recipiente (relaciones entre la acción y el recipiente que recibe esa acción, Ej.: Facturar - Cliente) y acción – instrumento (Ej.: Producción utiliza Maquinas).
  - **Relaciones que involucran atributos:** Relaciones entre cosas (una entidad) y atributos que probablemente estén asociados a él. (Ej.: Empleado es Experto).
- **Relaciones instrumentales:** Entre las que se encuentran las relaciones Cavar/Pala o Barrer/Escoba. Es este caso, se observa, que la definición del sustantivo probablemente deba contener al verbo [44].
- **Relaciones objetivas:** Por ejemplo, Conducir/Vehículo o Paseo/Bicicleta. Este tipo de relación, probablemente implique que la definición del verbo contenga al sustantivo [44].
- **Relación de endonimia:** Es una relación que involucra la incorporación del significado de una palabra en el significado de otra, por ejemplo: Universidad/Lector/Estudiante [44].
- **Relación de origen:** Por ejemplo, la relación entre Agua/Bienestar. Esta pone de manifiesto que una palabra es origen de otra [54].

Es evidente que las relaciones semánticas no clásicas rompen con las limitaciones de las clásicas, lo que se traduce a un conjunto de ventajas.

En principio, permite el solapamiento de categorías sintácticas, lo que significa que un mismo tipo de relación puede contener a verbos, adjetivos, sustantivos, etc. siempre y cuando estén relacionados por el contexto. Otra ventaja es que contemplan relaciones semánticas de acciones y eventos cotidianos, lo que lo vuelve un enfoque más realista.

Sin embargo, puede resultar complejo representar todas las relaciones derivadas de acciones y eventos cotidianos, por lo que se vuelve necesario utilizar una estandarización, como las presentadas en los párrafos anteriores.

En definitiva, se puede afirmar que el avance sobre este tipo de relaciones puede resultar en un mayor acercamiento a la correcta determinación de relación y similitud semántica existente entre dos palabras.

3.3 DESAMBIGUACIÓN DEL SENTIDO DE LA PALABRA

Desde el punto de vista de la semántica, existe un fenómeno que afecta a las palabras, que es conocido como polisemia.

La polisemia (contrario a la univocidad) refiere al hecho de que una palabra puede tener relacionado más de un significado, ya sea conceptualmente o históricamente. Estos significados no son intercambiables y son específicos del contexto (connotación). Por ejemplo, la palabra Diamante, puede hacer referencia a una forma geométrica y también a un campo de baseball (que posee esa forma) [49].

A la hora de realizar el análisis semántico de recursos, una actividad que generalmente demanda la mayor parte del esfuerzo en los sistemas de NLP es la de determinar el correcto sentido de las palabras.

Esto contribuye a eliminar la ambigüedad en cuanto al significado de un recurso y, por lo tanto, realizar una correcta interpretación del mismo. Los humanos, realizan esta tarea de

manera casi inconsciente y automática, sin embargo, para las computadoras resulta en una actividad compleja.

En el campo del NLP, estas actividades son denominadas desambiguación del sentido de la palabra *(WSD – Word Sense Disambiguation*). Dada una palabra perteneciente a una oración y un inventario de posibles etiquetas para esa palabra, donde cada una de esas etiquetas representan a un sentido distinto, el objetivo del WSD es responder a la pregunta ¿Cuál es la etiqueta adecuada para esa palabra en ese contexto? Por lo que, puede ser visto como un problema de clasificación [55].

A continuación, se presentarán las distintas clasificaciones de métodos de WSD, comenzando por una de las primeras aproximaciones, denominada Métodos Basados en Diccionario, para luego pasar a los Métodos Basados en Corpus, entre los cuales se encuentran las aproximaciones supervisadas y no supervisadas.

3.3.1 Métodos basados en Diccionario

El origen de esta aproximación, se da por la necesidad de evitar utilizar una gran cantidad de datos de entrenamiento, debido a la dificultad inherente de representar a la inmensa cantidad de palabras junto a sus distintos sentidos.

Es así que surgen los métodos basado en diccionario, que proponen explotar el conocimiento almacenado en los recursos léxicos disponibles, tales como diccionarios, tesauros y corpus de texto [55].

Existe una gran cantidad de trabajos realizados bajo esta aproximación. El principal exponente y el que desató el interés de distintos investigadores, fue el algoritmo de Lesk [56], que se describe en la siguiente sección.

Algoritmo original LESK

El algoritmo original Lesk [56], consiste en la desambiguación de sentidos, siendo particularmente adecuado para palabras que conformen a frases cortas.

Dada una palabra a desambiguar, este algoritmo realiza la comparación de cada una de sus definiciones de diccionario (sentidos) con respecto a cada una de las definiciones de las demás palabras pertenecientes a la frase. El sentido más adecuado, es el que cuya definición posea la mayor cantidad de palabras en común con respecto a las definiciones de las otras palabras de la frase [57].

Lesk [56] demuestra el funcionamiento de su algoritmo, considerando a la frase en inglés “PINE CONE”, para lo cual obtuvo las definiciones del diccionario de aprendizaje avanzado de Oxford. Las definiciones de estas palabras son presentadas en la Figura 3.7. En este ejemplo, y en otros presentados en el capítulo, se utilizará al idioma inglés, con el fin de lograr una mejor comprensión de los conceptos presentados.

Se puede observar, que la primera definición de Pine y la tercera definición de Cone, poseen la mayor cantidad de palabras coincidentes (indicadas con color rojo), por lo que son los sentidos más apropiados cuando Pine y Cone se utilizan en conjunto.

El algoritmo propuesto por Lesk, representa una de las aproximaciones ampliamente utilizadas en la actualidad, por su simplicidad y fácil compresión. Sin embargo, para situaciones en las que la desambiguación resulta un aspecto crucial, su baja precisión puede significar un problema, por lo que se debe tener en cuenta este detalle a la hora de decidir si utilizar o no este algoritmo.

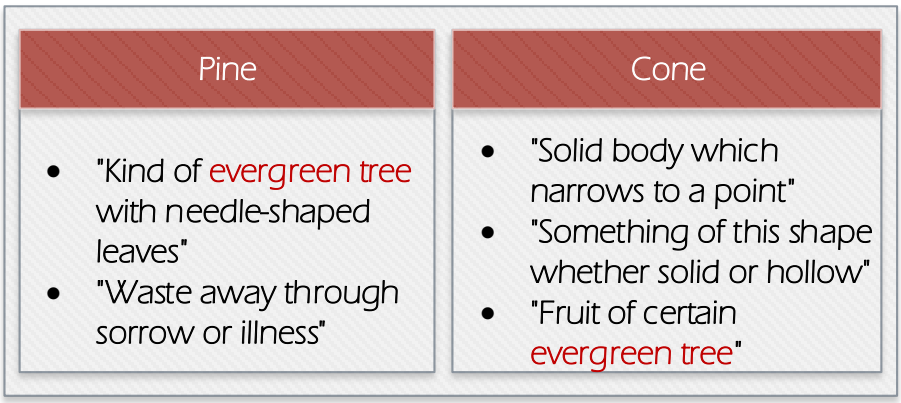

Figura 3.7 *- Definiciones de las palabras conformantes de la frase "PINE CONE" [56]*

3.3.2 Métodos basados en Corpus

Como se expuso en secciones anteriores, los procesos de WSD pueden ser vistos como problemas de clasificación, ya que se debe determinar cuál es el sentido más adecuado de una palabra determinada, de acuerdo a las características del contexto en el que se la utiliza. Es por esto, que es posible su implementación mediante técnicas de *machine learning* (ML), que son especialmente aplicables a este tipo de problemas [58].

Los procesos de WSD llevados a cabo mediante la utilización de técnicas de ML, son conocidos como métodos basados en corpus y pueden ser divididos en dos categorías principales: supervisados y no supervisados.

Los algoritmos WSD supervisados son aquellos que necesitan de ejemplos de entrenamiento para poder llevar a cabo su proceso, ya que los necesitan para el aprendizaje de sus modelos.

Estos ejemplos, generalmente están compuestos por las características del contexto de cada sentido de la palabra, junto a etiquetas que identifican al sentido implicado.

Por ejemplo, para la palabra planta, las características del contexto, con sus respectivas etiquetas, se presentan en la Tabla 3.3. En el primer y segundo caso, al tratarse de una planta de ensamblaje y una planta nuclear, la etiqueta asignada es fábrica. El último dato de entrenamiento, corresponde a una planta tropical, por lo que la etiqueta es ser vivo.

Tabla 3.3 *- Ejemplos de entrenamiento para la palabra Planta*

Características del Contexto	Etiqueta
Plante de ensamblaje	Fábrica
Planta nuclear	Fábrica
Planta tropical	Ser vivo

Entre las distintas técnicas de WSD supervisado se puede nombrar al basado en modelos Naive-Bayes presentado en [59], o el basado en máquinas de vectores de soporte (SVM) planteado en [60], entre otros.

Estos algoritmos cuentan con la limitación de que no todas las palabras poseen una distinción exacta del sentido, debido a los escasos datos de entrenamiento relacionados a ellas o la no existencia de datos de entrenamiento.

Esto motivó el surgimiento de los algoritmos WSD levemente supervisados, que buscan lograr conocimiento con pocos datos existentes. Entre estos se encuentran, los algoritmos de WSD basados en desambiguación de clases de palabras, algoritmos basados en gráficos, etc.

Por otro lado, la otra categoría perteneciente a los métodos basados en corpus son los algoritmos WSD no supervisados. Esta categoría, tiene el objetivo de agrupar en clusters las instancias de las palabras polisémicas, utilizando alguna forma de representación de sus características o del contexto [55]. Para esto, generalmente se hace uso de algoritmos de clustering jerárquicos, como es el caso del trabajo presentado en [61].

En todas las técnicas presentadas en esta sección, hay un aspecto que resalta, que es la necesidad de datos de entrenamiento. Esto representa la principal dificultad de esta aproximación debido a la gran cantidad de palabras que posee un lenguaje completo y la dificultad de establecer etiquetas para cada uno de los sentidos de estas palabras. Sin embargo, los continuos avances en la obtención automática de sentidos a partir de corpus de longitud infinita, permiten suponer que este problema será solventado.

3.4 TAXONOMÍAS SEMÁNTICAS

Taxonomía es un término procedente del griego “taxis” (ordenamiento) y “nomos” (regla), utilizado en principio por la biología para categorizar organismos que poseen características en común. Esta idea fue extendida a la semántica, con el fin de estructurar palabras por medio de clases ordenadas.

Una taxonomía se expresa formalmente, mediante la utilización de propiedades o axiomas particulares. Es por esto, que su precisión en la clasificación es absoluta. Al representar a la semántica mediante la utilización de taxonomías, se logra, además de exactitud, evitar recaer en la ambigüedad que poseen los distintos sentidos de las palabras.

Formalmente, una taxonomía semántica puede ser expresada de la siguiente manera: dado un conjunto de elementos P (palabras), una taxonomía es un conjunto parcialmente ordenado no estricto de P. Por lo tanto, esta puede ser definida mediante $\preccurlyeq p$ (relación binaria $\preccurlyeq$ definida sobre P), que cumple con las propiedades descriptas en la Tabla 3.4 [27].

Tabla 3.4 *- Propiedades de un conjunto no estricto parcialmente ordenado*

Propiedad	***Notación de conjuntos***	***Descripción textual***
Reflexiva	$\forall\, p \in P : p \preccurlyeq p$	Relación binaria de una palabra consigo misma.
Asimétrica	$\forall\, u, v \in P :\ (u \preccurlyeq v \wedge v \preccurlyeq u) => u = v$	No existe relación bidireccional, salvo una relación de una palabra consigo misma.
Transitiva	$\forall\, u, v, w \in P (u \preccurlyeq v \wedge v \preccurlyeq w) => u \preccurlyeq w.$	Al relacionarse una palabra con otra y esta última con una tercera, entonces existe una relación entre la primera palabra y la tercera palabra.

Dado que generalmente se tiene un elemento raíz denotado por T, que subsume a todos los otros elementos, esta taxonomía se puede representar como un grafo acíclico dirigido con raíz. Un ejemplo puede ser visto en la Figura 3.8.

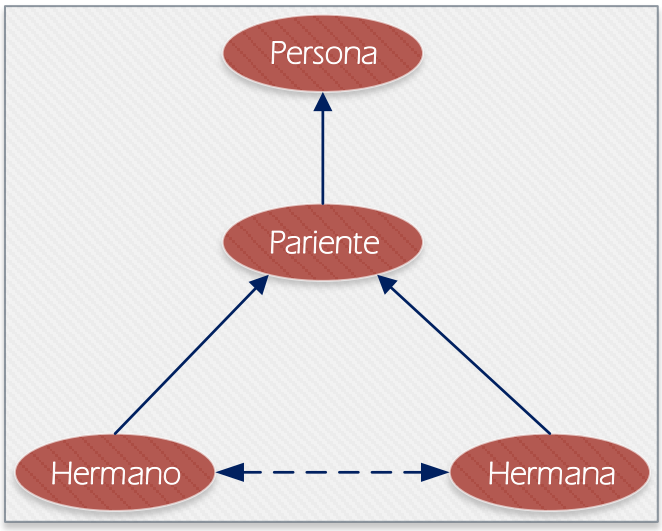

Figura 3.8 *- Ejemplo de taxonomía semántica [50]*

Una taxonomía de palabras $\preccurlyeq p$, puede ser formalmente definida como un grafo semántico O: $\langle P, R, E, A^O \rangle$, donde P es un conjunto de palabras, R define a los predicados que pueden ser utilizados para ordenar a las palabras, como por ejemplo, $R = \{SubClaseDe\}$ y $E \subseteq P \; x \; R \; x \; P$, es el conjunto de relaciones orientadas (aristas) que definen el orden de P. A^O representa el conjunto de axiomas o propiedades, que establece que O es una taxonomía y no una simple estructura de datos representada en forma de grafo [27].

En definitiva, las taxonomías son una aproximación adecuada para representar las palabras con sus distintos sentidos y relaciones, debido a la robustez de las expresiones formales. Esto permite obtener un modelo confiable y preciso.

Sin embargo, es evidente que el contenido de estas taxonomías debe provenir de alguna fuente, ya sea a partir de expertos o mediante la obtención automática desde corpus de texto, lo que representa una dificultad.

Este aspecto en general, resulta decisivo a la hora de optar por usar o no una taxonomía semántica, ya que por naturaleza están limitadas a representar una parcialidad del conjunto total de palabras de un lenguaje, con sus sentidos y relaciones.

En las secciones siguientes, se presentan dos de las taxonomías ampliamente utilizadas en trabajos de investigación, las cuales son WordNet y ConceptNet.

3.4.1 WordNet

WordNet [31] es una taxonomía semántica, en la que sustantivos, verbos, adjetivos y adverbios son organizados en conjuntos de sinónimos, representados mediante nodos.

Fue desarrollado en 1986 en la universidad de Princeton, estando actualmente en continua evolución. Surgió a partir del interés de George A. Miller por los experimentos de Inteligencia Artificial aplicados específicamente a la comprensión de la memoria semántica de los humanos. Para esto, buscó representar el conocimiento de miles de palabras y sus sentidos, de manera que sean simples de acceder y puedan ser almacenados en forma eficiente y económica [62].

A continuación, se explica cómo se representa internamente en WordNet al idioma inglés, resaltando los aspectos más importantes de este lenguaje y la representación equivalente de estos aspectos en la taxonomía.

El vocabulario de un lenguaje determinado, puede ser representado mediante un conjunto de pares (f, s), donde una forma f es un string definido sobre un alfabeto finito y s es un sentido perteneciente a un conjunto de connotaciones. Cada forma f en conjunto con un sentido s, es una palabra de ese lenguaje.

En WordNet, una forma f se representa mediante un string de caracteres ASCII, y un sentido s se representa mediante un conjunto de sinónimos, denominado *Synsets*. En su versión más reciente, contiene 155.327 formas de palabras f distintas y 207.016 pares (f, s), todas estas, pertenecientes al lenguaje inglés.

Los *Synsets* son conjuntos de pares (f, s) que tienen el mismo sentido s y se representan mediante una estructura que se compone por el conjunto de palabras que son sinónimos en un sentido determinado, la categoría sintáctica, el índice que identifica al sentido dentro de la taxonomía, una definición de diccionario y una frase de ejemplo. Esta estructura se presenta en la Tabla 3.5.

Tabla 3.5 *- Estructura de un Synset en WordNet*

Sinónimos	Categoría sintáctica	Índice del sentido en WordNet	Definición	Frase de ejemplo
"*Car, Auto, Automovile, Machine, Motorcar"*	N (Sustantivo)	1	"*A motor vehicle with four wheels, usually propelled by an internal conbustion engine*"	"*He needs a car to get to work*"

Por otro lado, en un idioma, cada palabra tiene un conjunto C de contextos lingüísticos en los que puede ser utilizada. La sintaxis del lenguaje, parte a C en distintos subconjuntos para cada categoría sintáctica. Es decir, se obtiene un subconjunto N de sustantivos, un subconjunto V de verbos y así sucesivamente. Asimismo, dentro de estos subconjuntos, pueden existir distintos sentidos para una palabra.

WordNet posee cuatro taxonomías distintas, donde cada una se corresponde con una categoría sintáctica diferente (sustantivo, adjetivo, verbo y adverbio) y son independientes entre sí. Un ejemplo de esto se da con las palabras *back* (atrás), *right* (derecho) y *well* (bien), que pueden ser sustantivos en un contexto lingüístico, verbos en otros contextos y adjetivos o adverbios en otros contextos, donde para cada uno de ellos existe una instancia distinta en WordNet.

Otra cuestión relacionada a un lenguaje, es su morfología, que se define en términos de un conjunto M de relaciones entre formas de palabras. La morfología del inglés está particionada en relaciones morfológicas flexivas[1], derivacionales[2] y composiciones[3] [66].

La morfología flexiva tiene representación en WordNet. Por ejemplo, si se requiere información para *went* (fuimos), el sistema retorna la palabra transformada, que en este caso es *go* (ir).

Por otro lado, las morfologías compuestas y derivacionales, se representan mediante ocurrencias independientes. Por ejemplo, *interpret* (interpretar), *interpreter* (interprete), *misinterpret* (malinterpretado), son ingresadas por separado en WordNet.

[1] **Morfología flexiva:** Transformación de la palabra para indicar número, posesión, tiempo verbal, etc. Por ejemplo: *write* (escribir) y *writes* (escribe) [63].

[2] **Morfología derivacional:** Tipo de formación de palabras que crea nuevos lexemas, mediante cambios en la categoría sintáctica y/o añadiendo nuevos significados sustanciales a una base libre o consolidada. Por ejemplo: *employment* (empleo), *employer* (empleador), *employee* (empleado) [64].

[3] **Composiciones en morfología**: Es un proceso de formación basados en la combinación de elementos léxicos (palabras). Por ejemplo flower (flor) y pot (lata), forman flowerpot (meseta) [65].

Finalmente, los aspectos léxicos semánticos de un lenguaje, son definidos en términos de un conjunto S de relaciones entre los sentidos de las palabras, que se conocen como relaciones semánticas.

En WordNet, se define un conjunto de relaciones semánticas, que fueron seleccionadas debido a que son ampliamente utilizadas en el idioma inglés y no es necesario tener un entrenamiento avanzado para comprenderlos.

Estas se implementan mediante aristas que interconectan dos *Synsets*, plasmando así los distintos tipos de relaciones que pueden existir entre los sentidos de dos palabras distintas, exceptuando a la sinonimia que se contempla en cada uno de los *Synsets*. Las relaciones consideradas en WordNet se presentan en la Tabla 3.6 [66].

Tabla 3.6 *- Relaciones semánticas contempladas en WordNet [66].*

Relación Semántica	***Categoría sintáctica***	***Ejemplo***
Sinonimia	N,V,Aj,Av	*Sad* (triste), *unhappy* (infeliz)
Antonimia	N, V, Aj, Av	*Wet* (mojado), *dry* (seco)
Hiperonimia	N	*Plant* (planta), *tree* (árbol)
Hiponimia	N	*Sugar maple* (arce azucarero), *maple* (arce)
Meronimia	N	*Wheel* (rueda), *car* (automóvil)
Troponimia	V	*March* (caminata), *walk* (caminar)
Implicación	V	*Drive* (conducir), *ride* (pasear)
Referencias	**N** = Sustantivo; **Aj** = Adjetivo; **V** = Verbo; **Av** = Adverbio	

Lo expuesto, permite afirmar que WordNet provee un entorno idóneo para su aplicación en la determinación de relevancia semántica, debido a su confiabilidad y robustez propias de una taxonomía, la gran cantidad de características del lenguaje representadas y la cantidad de palabras que la componen.

3.4.2 ConceptNet

ConceptNet [32], es una taxonomía semántica compuesta por una gran cantidad ítems de conocimiento de sentido común y un conjunto de herramientas que permiten realizar tareas de razonamiento sobre textos.

El conocimiento del sentido común, se obtiene desde sucesos de la vida cotidiana, lo que abarca aspectos espaciales, físicos, sociales, temporales y psicológicos.

ConceptNet, surge con el fin de expandir el alcance de WordNet, buscando incorporar una mayor cantidad de conocimiento práctico y mantener su facilidad de uso. Esta expansión contempla tres puntos principales.

En primer lugar, se amplió la noción de nodo, que además de contener palabras y frases simples con significado atómico, también comprenden conceptos de orden superior, tales como frases compuestas (por ejemplo: *Adventure Books*).

Esto brindó la posibilidad de representar el conocimiento a partir de un gran rango de palabras encontradas en la vida cotidiana.

La segunda extensión, fue aplicada sobre las relaciones semánticas, agregando un conjunto extra de relaciones a las ya contempladas por WordNet, obtenidas a partir de actividades de la vida cotidiana.

Como resultado, en esta taxonomía se contemplan tanto a las relaciones semánticas clásicas, como a las no clásicas, expuestas en la sección 3.2. Una porción de estas es plasmada en la Tabla 3.7 [67].

Tabla 3.7 *- Ejemplo de Relaciones Semánticas contempladas en ConceptNet [68]*

Relación	Ejemplo	Relación	Ejemplo
IsA	*"Car" is a type of "Vehicle"*	***LocatedNear***	*"Chair" is near to "Table"*
UsedFor	*"Drum" is used for "Make Music"*	***DefinedAs***	*"Energy" is defined as "capacity to perfom work"*
HasA	*"Guitar" has "Strings"*	***AtLocation***	"Apple" Located in "Apple tree"
CapableOf	*"Airplane" is capable of "Arraive at the airport"*	***ReceivesAction***	*"room" can be "entered through a doorway"*
Desires	*"Dog" wants "a Bone"*	***HasPrerequisite***	*"Bed" requires "Going to sleep"*
CreatedBy	*"Cake" is created by "Baking it"*	***MotivatedByGoal***	*"Memorize" is motivated to "be able to repeat information"*
PartOf	*"Accelerator" is part of "Car"*	***CausesDesire***	*"Dire" could make you want to "take shower"*
Causes	*Causes of "Disease" is a "Virus"*	***MadeOf***	*"Tire" is made of "Rubber"*

Estas relaciones se representan internamente en ConceptNet mediante aserciones, que poseen la siguiente estructura: $Relación(Palabra_1, Palabra_2)$. Un ejemplo de esto puede ser: $IsA(Jazz, Genre\ of\ music)$.

Cada una de estas aserciones son obtenidas a partir de textos en lenguajes natural, extraídos por mecanismos automáticos. Actualmente ConceptNet cuenta con más de 8.7 millones de aserciones, que conectan 3.9 millones de palabras [68].

El tercer cambio realizado con respecto a WordNet, refleja que el conocimiento almacenado en ConceptNet es de naturaleza más informal, descartable y valorada desde el punto de vista de la práctica.

Que el conocimiento sea valorado desde la práctica, significa que vincula objetos de acuerdo a su participación en una actividad. Por ejemplo, en WordNet no se relaciona a Perro con Mascota, mientras que sí a Perro con Canino, Carnívoro y Mamífero Placentario. Por el contrario, ConceptNet si tiene en cuenta a este tipo de relaciones.

Por otro lado, el conocimiento descartable, es aquel que tiene probabilidades de ocurrir, pero que no siempre ocurre de la misma manera, por ejemplo, Caerse De La Bicicleta provoca el efecto de Lastimarse. Este tipo de conocimiento es útil, teniendo en cuenta que gran parte de los sucesos cotidianos son descartables [67].

Es evidente que ConceptNet representa una alternativa superadora a WordNet, lo que viene relacionado a que esa fue su motivación desde su concepción.

El hecho de contemplar relaciones semánticas no clásicas lo convierte en un marco adecuado para realizar análisis de mayor complejidad, pero también implica contemplar una menor cantidad de detalles, por ejemplo, no identifica los distintos sentidos de la palabra.

Esto se debe a que el enfoque de ConceptNet está más orientado a realizar inferencias sobre un conjunto de palabras, útiles en tareas de NLP como la generación de texto en lenguaje natural.

Sin embargo, es una alternativa confiable para obtener relaciones semánticas entre pares de palabras sin considerar el sentido, lo que resulta útil en las tareas de evaluación de la relación semántica entre clave de búsqueda y recursos web.

3.5 MÉTRICAS DE RELACIÓN Y SIMILITUD SEMÁNTICA

Las métricas de relación y similitud semántica, son las que, en base a distintos métodos, determinan que tan relacionadas se encuentran dos palabras. En general, se distinguen dos aproximaciones: la basada en Corpus y la basada en Conocimientos.

Las métricas basadas en corpus, estima la relación existente entre dos palabras, teniendo como base el análisis estadístico de grandes corpus de texto, las distribuciones de las palabras y las regularidades estadísticas observadas. Por ejemplo, determina la relación y similitud semántica considerando la frecuencia en que dos palabras aparecen juntas en un corpus de texto.

Por otro lado, las métricas basadas en conocimiento realizan su evaluación teniendo como base la distancia taxonómica entre pares de palabras definidas en una taxonomía semántica [27].

En esta sección se hará foco sobre las métricas basadas en conocimiento, debido a su aplicabilidad sobre taxonomías semánticas, tales como WordNet y ConceptNet, que son utilizadas en el marco de este trabajo de investigación.

En la actualidad existen cuatro enfoques distintos de métricas basadas en conocimiento, utilizadas para comparar palabras definidas sobre una taxonomía, que son [27]:

- Métricas basadas en el análisis de la estructura del grafo.
- Métricas basadas en el análisis de propiedades de las palabras.
- Métricas basadas en la teoría de información.
- Métricas hibridas.

En las secciones siguientes, se presentan a cada uno de estos enfoques, explorando sus características y las diversas maneras en las que proponen calcular la relación y similitud semántica entre pares de palabras.

3.5.1 Métricas basadas en el análisis de la estructura del grafo

Las métricas basadas en estructura, o también conocidas como métricas de conteo de aristas, estiman la similitud semántica en función al grado de interconexión existente entre las palabras.

Son generalmente reconocidas como métricas enmarcadas en el modelo espacial, ya que la similitud entre dos palabras se estima mediante la distancia que las separa dentro del grafo [27]. Esta distancia, generalmente se calcula contando la cantidad de aristas que las separan [69].

Las propuestas desarrolladas siguiendo este enfoque, son las que se describen a continuación:

- *Shortest Path* o camino más corto
- La métrica de Wu and Palmer
- La métrica de Slimani
- La métrica de Li
- La métrica de Leacock and Chodorow

Shortest Path (Camino más corto)

Esta métrica surge a partir de la propuesta realizada por Rada y otros [70], donde se planteó que la distancia conceptual entre dos palabras ubicadas en un espacio multidimensional puede ser utilizada como métrica de similitud.

Este planteo fue extendido a las taxonomías, mediante el siguiente enunciado “Dadas dos palabras A y B, representadas en una red semántica mediante los nodos a y b respectivamente, la distancia conceptual entre A y B está determinada por el mínimo número de aristas que las separan”. El planteo de esta métrica derivó en la formula presentada en la Ecuación (3.1).

$$Sim(P_1, P_2) = 2 * Max(P_1, P_2) - SP \quad (3.1)$$

Donde P_1 y P_2 son las dos palabras, sobre las que se desea conocer la similitud semántica, $Max(P_1, P_2)$ es la máxima longitud de camino entre P_1 y P_2, y SP es la menor longitud de camino entre ambas palabras [27][69].

Métrica de Wu And Palmer

Wu y Palmer [33], proponen que la relación entre dos palabras se obtiene mediante una división que considera al camino más corto que enlace a las palabras evaluadas y la profundidad de su ancestro común más cercano [27].

Dadas dos palabras, P_1 y P_2 pertenecientes a una taxonomía. El cálculo de relación semántica está basado en la distancia $N1$ y $N2$ que separan a P_1 y P_2 del nodo raíz y la distancia N que separa al ancestro común más cercano del nodo raíz (ver Figura 3.9).

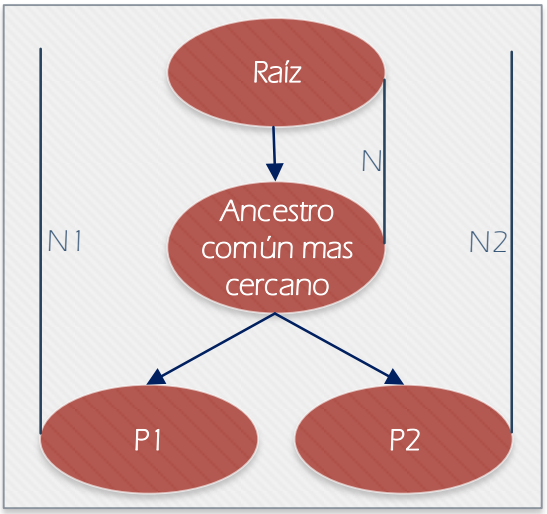

Figura 3.9 *- Parámetros utilizados en la métrica de Wu and Palmer en un extracto de una taxonomía [71]*

El valor que indica que tan relacionados semánticamente están P_1 y P_2, se obtiene mediante la fórmula presentada en la Ecuación (3.2) [71].

$$Sim_{wp} = \frac{2 * N}{N1 + N2} \quad (3.2)$$

Métrica de Slimani

Se trata de una extensión a la métrica de Wu And Palmer (WUP), en la que se busca mejorar sus resultados debido a que no siempre son satisfactorios.

Específicamente, se centra en que WUP valora mejor a la relación entre palabras ubicadas en distintas ramas de la jerarquía, que a la existente entre palabras ubicadas en una misma rama. Un ejemplo se presenta en la Figura 3.10.

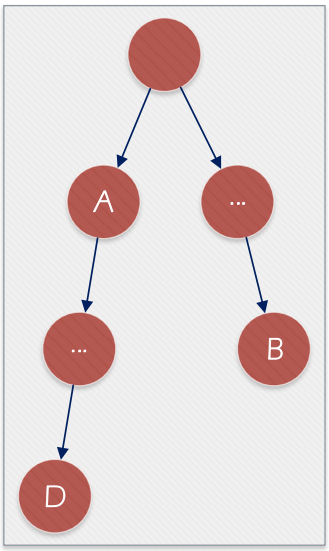

Figura 3.10 *- Ejemplo de problema en el que incurre la métrica de Wu and Palmer [69]*

Dado este ejemplo, WUP determina que el valor de relación entre la palabra A y la palabra B es mayor que el valor de relación entre la palabra A y la palabra D.

Esto es un problema porque, la palabra D pertenece a la categoría encabezada por la palabra A, lo que implica una relación más directa que la existente entre la palabra A y B [69].

Para paliar esta situación, Slimani y otros [71], propusieron la formula presentada en la Ecuación (3.3).

$$Sim_{tbk} = Sim_{wp}(P1, P2) * PF(P1, P2) \tag{3.3}$$

Donde $Sim_{wp}(P1, P2)$ es la similitud semántica entre las palabras $P1$ y $P2$, determinada por WUP y $PF(P1, P2)$ es una penalización para palabras que no se encuentren en la misma jerarquía, cuyo valor es obtenido a partir de la Ecuación (3.4).

$$PF(P1, P2) = (1 - \lambda) * (\min(N1, N2) - N) + \lambda * (|N1 - N2| + 1)^{-1} \tag{3.4}$$

Donde $\min(N1, N2)$ es el valor mínimo de distancia de las palabras $P1$ y $P2$ al nodo raíz, N es la distancia a la raíz del ancestro común más cercano (ver Figura 3.9) y λ es un coeficiente booleano, que retorna 0 si las dos palabras están en una misma jerarquía, o 1 si están en jerarquías diferentes [71].

Métrica de Li

La métrica de similitud de Li [34], es en una función paramétrica que considera la longitud del camino más corto y la profundidad del ancestro común más cercano, para determinar la relación y similitud semántica entre dos palabras [27]. El puntaje se obtiene mediante la fórmula presentada en la Ecuación (3.5).

$$Sim_{li}(P1, P2) = e^{-\alpha * SP} * df(P1, P2) \tag{3.5}$$

Donde SP es el camino mas corto que separa a las dos palabras $P1$ y $P2$, determinado a partir del conteo de aristas, $\alpha > 0$ es un coeficiente que determina la contribución del SP al valor final de similitud semántica y $df(P1, P2)$ es un valor denominado factor de profundidad, que se obtiene aplicando la Ecuación (3.6).

$$df(P1, P2) = \frac{e^{\beta * N} - e^{-\beta * N}}{e^{\beta * N} + e^{-\beta * N}} \tag{3.6}$$

Donde $\beta > 0$ es un coeficiente que determina la importancia de N, es decir, especifica la contribución al valor final, de la distancia que separa al ancestro común más cercano y la

raíz. $df(P1, P2)$ consiste en una función tangente hiperbólica que es normalizada en un intervalo $[0,1]$ y define el grado de no linealidad asociado a la profundidad del ancestro común más cercano [27][69].

Métrica Leacock and Chodorow

En la investigación realizada por Leacock y Chodorow [72], se propuso una métrica de relación y similitud semántica, que considera al camino más corto SP y la máxima profundidad de la taxonomía D. Esta métrica calcula el puntaje de relación y similitud semántica mediante la Ecuación (3.7).

$$Sim_{lc}(P1, P2) = -\log\left(\frac{SP}{2 * D}\right) \tag{3.7}$$

Donde SP es la longitud del camino más corto entre $P1$ y $P2$, obtenido mediante el conteo de aristas que las separan, y D es la profundidad máxima de la taxonomía [27][69].

3.5.2 Métricas basadas en el análisis de propiedades de las palabras

Las métricas pertenecientes a este enfoque calculan la relación y similitud semántica, existente entre pares de palabras, considerando sus características compartidas y no compartidas.

El puntaje obtenido será únicamente influenciado por la estrategia adoptada para caracterizar las propiedades de la palabra, y la adoptada para la comparación [27][69].

A continuación, se presentarán dos de las métricas pertenecientes a este enfoque, las cuales son: La métrica de similitud semántica de Tversky y la métrica X-Similarity.

Métrica de similitud semántica de Tversky

La métrica de Tversky [73], tiene en cuenta las propiedades de las palabras para calcular la relación y similitud semántica entre ellas.

Cada palabra posee un conjunto de términos que indican sus propiedades (generalmente su definición de diccionario). Las propiedades en común (términos en común) tienden a incrementar el puntaje de relación, mientras que las no comunes, tienden a disminuirlo. Esta métrica está definida por la Ecuación (3.8).

$$Sim_{tvsk}(P1, P2) = \frac{|P1 \cap P2|}{|P1 \cap P2| + \alpha|P1 - P2| + \beta|P2 - P1|} \tag{3.8}$$

Donde $|P1 \cap P2|$ es la cantidad de propiedades compartidas por las palabras $P1$ y $P2$, $|P1 - P2|$ son todas las propiedades de $P1$, que no comparte $P2$, $|P2 - P1|$ son todas las propiedades de $P2$ no compartidas por $P1$ y $\alpha, \beta \in [0,1]$ es la importancia relativa de las características no compartidas por ambas palabras. El valor de α y β se incrementa con las características en común y decrementa con las características no compartidas [69].

X-Similarity

Petrakis y otros [74], proponen una métrica basada en el planteo de que dos palabras son más similares, cuanto más similares léxicamente sean sus definiciones y las de sus vecinos (obtenidas por relaciones semánticas).

Debido a que no todas las palabras vecinas poseen las mismas relaciones, el cálculo se realiza por tipo de relación semántica SR. Dadas dos palabras A y B, la métrica de relación y similitud semántica propuesta, se expresa en la Ecuación (3.9):

$$Sim_{xsim} = \begin{cases} 1, si\ S_{synsets}(A,B) > 0 \\ \max\{S_{vecindario}(A,B), S_{descripcion}(A,B)\}, si\ S_{synsets}(A,B) = 0 \end{cases} \quad (3.9)$$

Siendo i, un tipo de relación semántica (SR), la similitud de los vecinos $S_{vecindario}$ se obtiene mediante la siguiente formula:

$$S_{vecindario} = \max_{i \subset SR} \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (3.10)$$

Para implementar la Ecuación (3.10) es necesario armar por cada palabra A y B, un corpus conformado por las definiciones de cada palabra relacionada por medio del tipo de relación semántica i, hasta llegar a la raíz.

Luego, mediante estos corpus, se obtiene la cantidad de términos coincidentes y se los divide por la cantidad total de términos sin repetición (que se obtiene combinando los dos corpus generados). Es importante tener en cuenta que, esto se realiza por cada por cada relación semántica i.

Finalmente, se selecciona el máximo de estos valores como el valor final de relación semántica, que se corresponde con alguno de los tipos de relaciones semánticas i a partir de las que se obtienen los corpus de comparación.

Un ejemplo, puede ser la siguiente situación: Inicialmente se conforma el corpus A_{hip}, que consiste en la concatenación de todas las definiciones de las palabras relacionadas a A por la relación de hiperonimia. De igual manera se obtiene el corpus B_{hip}.

Sobre estos corpus, se aplica la Ecuación (3.10) y se procede a repetir el proceso con el resto de las relaciones semánticas (por ejemplo, con la relación de meronimia). El valor final de relación semántica, será el máximo valor obtenido a partir de la Ecuación (3.10).

Por otro lado, para obtener los valores de $S_{descripcion}$ y $S_{synsets}$, se aplica la formula presentada en la ecuación (3.11).

$$S_{description} = S_{synsets} = \frac{|A \cap B|}{|A \cup B|} \quad (3.11)$$

Estos dos valores se obtienen mediante la división entre la cantidad de palabras coincidentes en las definiciones de las palabras A y B, y la cantidad de palabras distintas en la unión de ambas definiciones [69][74].

3.5.3 Métricas basadas en la teoría de la información

Las métricas de este enfoque se basan en una estimación de la cantidad de información transportada por cada palabra, es decir, su contenido de información (IC)[4].

Se centran en evaluar la similitud semántica de acuerdo a la cantidad de información compartida y la no compartida entre las palabras comparadas. Pero a diferencia del enfoque basado en propiedades, considera el grado de información de las palabras en lugar de hacer un análisis booleano de presencias de términos [27].

En las secciones siguientes, se presentan tres métricas basadas en la teoría de la información: La métrica propuesta por Resnik, la métrica de similitud de Lin y la métrica de Jiang and Conrath.

[4] **Contenido de información:** Cantidad de información que transmite una palabra. Puede estimarse en función al tamaño del universo de interpretaciones asociadas a él (es decir, sus instancias) [27].

Todas estas métricas utilizan el contenido de información de los padres (ancestro común) que comparten las palabras $P1$ y $P2$ analizadas.

Es importante destacar que estas palabras pueden compartir ancestros comunes a través de distintos caminos. Cuando sucede esto, se utiliza el mínimo $p(P)$, siendo P_{mis} el ancestro común más informativo. Esto se determina mediante la Ecuación (3.12).

$$P_{mis}(P1, P2) = \min_{P \in S(P1,P2)} \{p(P)\} \tag{3.12}$$

Donde $p(P)$ es la probabilidad de aparición del ancestro común P en un corpus de comparación construido para tal fin (ejemplos de estos pueden ser el corpus de Resnik y el corpus de Brown) y $S(P1, P2)$ es el conjunto de palabras que son ancestros comunes de $P1$ y $P2$ [69]. Por lo tanto, el ancestro común más informativo, será aquel que menos apariciones tenga en el corpus de comparación.

Métrica de Resnik

La métrica de Resnik [75] plantea que dos palabras son más similares, si comparten más información.

La información compartida por dos palabras $P1$ y $P2$, está determinada por el contenido de información del ancestro común más informativo. Esta métrica se define formalmente mediante la Ecuación (3.13):

$$Sim_{Resnik}(P1, P2) = -\ln(p_{mis}(P1, P2)) \tag{3.13}$$

Siendo $p_{mis}(P1, P2)$ definido en la Ecuación (3.12) [69]. La métrica de Resnik no captura explícitamente las especificidades de las palabras comparadas, lo que significa que, pares de palabras que comparten el mismo ancestro común más informativo, tendrán la misma similitud semántica [27].

Métrica de Lin

En el enfoque de Lin y otros [35], se propone una métrica basada en la restricción de la taxonomía a enlaces jerárquicos (relaciones de meronimia e hiperonimia) y un corpus de comparación. Esta métrica se define en la Ecuación (3.14).

$$Sim_{lin}(P1, P2) = \frac{2 * \ln(p_{mis}(P1, P2))}{\ln(p(P1)) + \ln(p(P2))} \tag{3.14}$$

Donde $p_{mis}(P1, P2)$ se obtiene mediante la Ecuación (3.12), y $p(P1)$ y $p(P2)$, son las probabilidades de ocurrencia de las palabras $P1$ y $P2$ en un corpus determinado [69]. Esto permite contemplar las especificidades de las palabras comparadas [27].

Métrica de Jiang and Conrath

La propuesta de Jiang y Conrath [76], estima la distancia semántica de dos palabras, utilizando un corpus de comparación en conjunto con una taxonomía, de manera similar a lo realizado por las aproximaciones de Resnik y Lin.

Plantea que la distancia semántica existente entre dos palabras $P1$ y $P2$, es la diferencia entre la suma del contenido de información de las dos palabras y el contenido de información del ancestro común más informativo. Formalmente, esta métrica se presenta en la Ecuación (3.15).

$$Sim_{jc}(P1, P2) = -2 * \ln(p_{mis}(P1, P2)) - (\ln(p(P1)) + \ln(p(P2))) \tag{3.15}$$

Donde $p_{mis}\ (P1, P2)$ se obtiene mediante la Ecuación (3.12), y $p(P1)$ y $p(P2)$, son las probabilidades de ocurrencia de las palabras $P1\ y\ P2$ en un corpus determinado [69].

3.5.4 Métricas híbridas

Las métricas hibridas son aquellas que surgen con bases en múltiples paradigmas, combinando por ejemplo factores como contenido de información (IC), densidad, profundidad, propiedades de palabras, entre otros factores [27][69]. Existen diversas propuestas basadas en estas categorías, entre las que se encuentran:

- **Métrica de Knappe** [77]**:** Define una métrica de relación y similitud semántica que tiene en cuenta la coincidencia en las definiciones de sus ancestros, sean comunes o no.
- **Métrica de Zhou y otros** [78]**:** Proponen una métrica que tiene como parámetro a las métricas de contenido de información y las métricas basadas en estructuras.
- **Métrica de Álvarez y Yan** [79]**:** Proponen una métrica que explota tres componentes de evaluación de palabras: el camino más corto, el ancestro común más cercano y sus definiciones literales.

3.5.5 Consideraciones sobre las métricas

Las métricas analizadas poseen varios aspectos destacables que deben ser mencionados. Por ejemplo, un aspecto se relaciona a las métricas basadas en la estructura del grafo, que dependen de las relaciones plasmadas en la taxonomía y que tan bien fueron representadas.

En el caso de WordNet, las relaciones son creadas por expertos, lo que proporciona confianza a la hora de ejecutar métricas de relación y similitud semántica sobre las palabras. No obstante, esto provoca que sus métricas sean susceptibles a la subjetividad de los expertos que definieron las relaciones, lo que representa una desventaja.

Las métricas basadas en propiedades, consideran a las definiciones de diccionario para caracterizar a las palabras evaluadas. El aspecto a resaltar, es que estas definiciones, deben ser escritas mediante un conjunto de términos estándar, con el fin de que la evaluación de términos coincidentes arroje resultados confiables. El problema es que generalmente, las definiciones no utilizan términos estándar, por lo que los resultados de estas métricas no suelen ser satisfactorios.

Además, es posible que pares de palabras cuya relación semántica sea mínima, tengan una cierta cantidad de términos coincidentes en sus definiciones, provocando un puntaje de relación semántica alto. Esto, representa otro punto de falla.

Las métricas basadas en la teoría de la información, proponen una mejora al enfoque basado en propiedades, debido a que consideran el grado de información de las palabras analizadas. Puede ser visto como un enfoque probabilístico, ya que, para determinar el grado de información, utiliza la probabilidad de que las palabras consideradas y sus ancestros comunes aparezcan en un corpus construido para tal fin.

Sin embargo, la utilización de un corpus específico, se convierte en el talón de Aquiles de este enfoque, ya que posee una cantidad de palabras limitada y la efectividad de las métricas dependen de que tan bien estén representadas.

Las métricas hibridas, al ser una combinación de las anteriores, apuntan a mejorar problemas y aprovechar virtudes, por lo que no se pueden distinguir aspectos específicos de este enfoque.

En definitiva, cada métrica posee virtudes y defectos, por lo que deben ser evaluadas de acuerdo al contexto en la que serán utilizadas. Por ejemplo, si se busca precisión en la determinación de relación y similitud semántica, posiblemente el enfoque basado en la

estructura del grafo o el basado en la teoría de la información sean los más adecuados, mientras que, si se busca obtener rapidez en la respuesta, el enfoque basado en propiedades puede ofrecer buenos resultados.

CAPÍTULO 4
MODELO PROPUESTO

En este capítulo, se propone un modelo para la evaluación de la relevancia de recursos web que utiliza técnicas semánticas. Inicialmente se define el modelo general que emplea técnicas lexicográficas y técnicas semánticas. Posteriormente, se explica detalladamente el modelo propuesto y se presentan algunos ejemplos de aplicación. Finalmente, se define el modelo de integración léxico – semántico que incluye los criterios considerados tanto por las técnicas léxicas, como por las técnicas semánticas.

4.1 MODELO GENERAL

Con el fin de implementar un sistema que lleve a cabo la determinación de la relevancia de recursos, mediante las Técnicas Semánticas (TS) propuestas en este trabajo y las Técnicas Lexicográficas (TL) ya implementas en [5][6][7], se define la arquitectura del sistema, que se presenta en la Figura 4.1.

En los párrafos posteriores, se procede a explicar esta arquitectura, detallando el funcionamiento general del sistema.

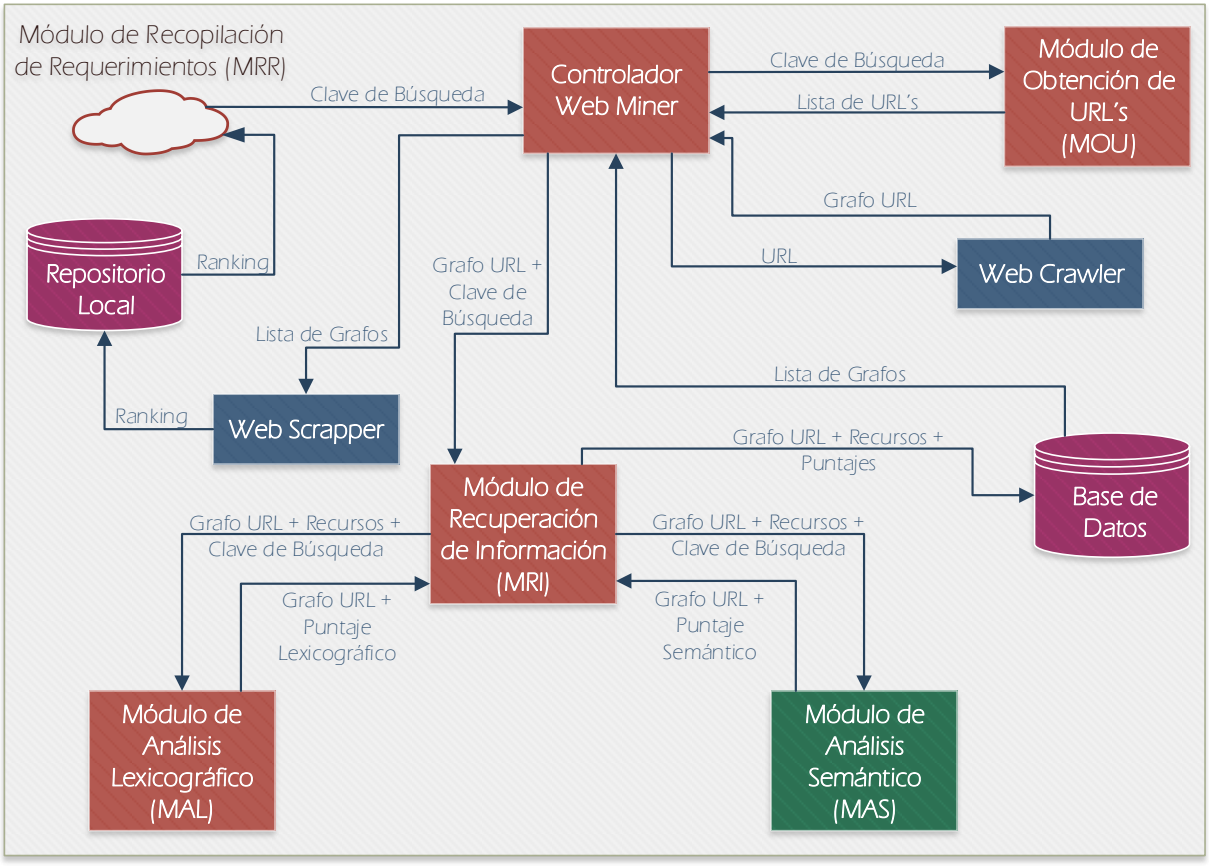

Figura 4.1 *- Arquitectura general del sistema propuesto*

El proceso de análisis de recursos y generación de rankings comienza cuando el ***Módulo de Recopilación de Requerimientos (MRR)***, envía la clave de búsqueda ingresada por el usuario al ***Controlador Web Miner***, que se encarga de coordinar el funcionamiento de todo el sistema.

Esta clave, se pasa al ***Módulo de Obtención de URL's (MOU),*** donde se genera una lista única de URL's conformada por los primeros diez resultados de los siguientes

buscadores: *Google*, *Bing*, *MSXML Excite* e *Intelligo.* Confeccionada dicha lista, se la retorna al ***Controlador Web Miner***, donde se descartan las repetidas e inicia el proceso de análisis (Figura 4.2).

Cada una de las URL resultantes, son representadas mediante nodos, con estado de procesamiento “No Explorado”. La representación mediante nodos permite mantener una estructura de datos por cada URL, compuesta por los puntajes procedentes de los análisis de las TL, las TS y el MILS, el contenido de dicha URL (recurso Web) y el estado de procesamiento, que puede ser “Explorado” o “No Explorado”.

La siguiente operación consiste en enviar uno de los nodos marcado como “No Explorado” al módulo ***Web Crawler***, donde se descubren los enlaces directamente relacionados a su URL y se construye un grafo acíclico dirigido que represente estas relaciones. En este grafo, el nodo recibido es la raíz y los enlaces descubiertos se representan como sus nodos hijo. Como resultado de esto, se cambia el estado de procesamiento de la raíz a “Explorado” y se retorna el grafo completo al ***Controlador Web Miner***.

Seguidamente, se envía el grafo generado y la clave de búsqueda, al ***Módulo de Recuperación de Información (MRI)***, donde por cada nodo se obtiene el contenido del mismo (HTML o PDF) y se ejecuta el análisis de relevancia, que determina su grado de correlación con respecto a la clave de búsqueda.

Para el análisis de relevancia se utilizan dos técnicas: las TL implementadas mediante el ***Módulo de Análisis Lexicográfico (MAL)*** y las TS implementadas mediante el ***Módulo de Análisis Semántico (MAS)***.

Como resultado de la ejecución de estos módulos, por nodo, se obtienen cuatro puntajes. Los tres primeros correspondientes a las TL, que representan a las técnicas *CRank*, Okapi y VSM respectivamente, y el cuarto, que representa al puntaje de relevancia obtenido mediante las TS.

Al finalizar estos análisis, se almacena en la base de datos los nodos del grafo analizado, es decir, los puntajes, el contenido de la URL y el estado de procesamiento de cada nodo.

Posteriormente, en el ***Controlador Web Miner***, se comienza con la presentación de los resultados. Para ello, inicialmente, se genera una lista compuesta por los grafos obtenidos hasta el momento.

Esta lista se envía al módulo ***Web Scrapper***, donde se calcula el puntaje final de cada nodo mediante el Modelo de Integración Léxico – Semántico (MILS) (descripto en la sección 4.3) y se confecciona el ranking de recursos a ser presentado al usuario.

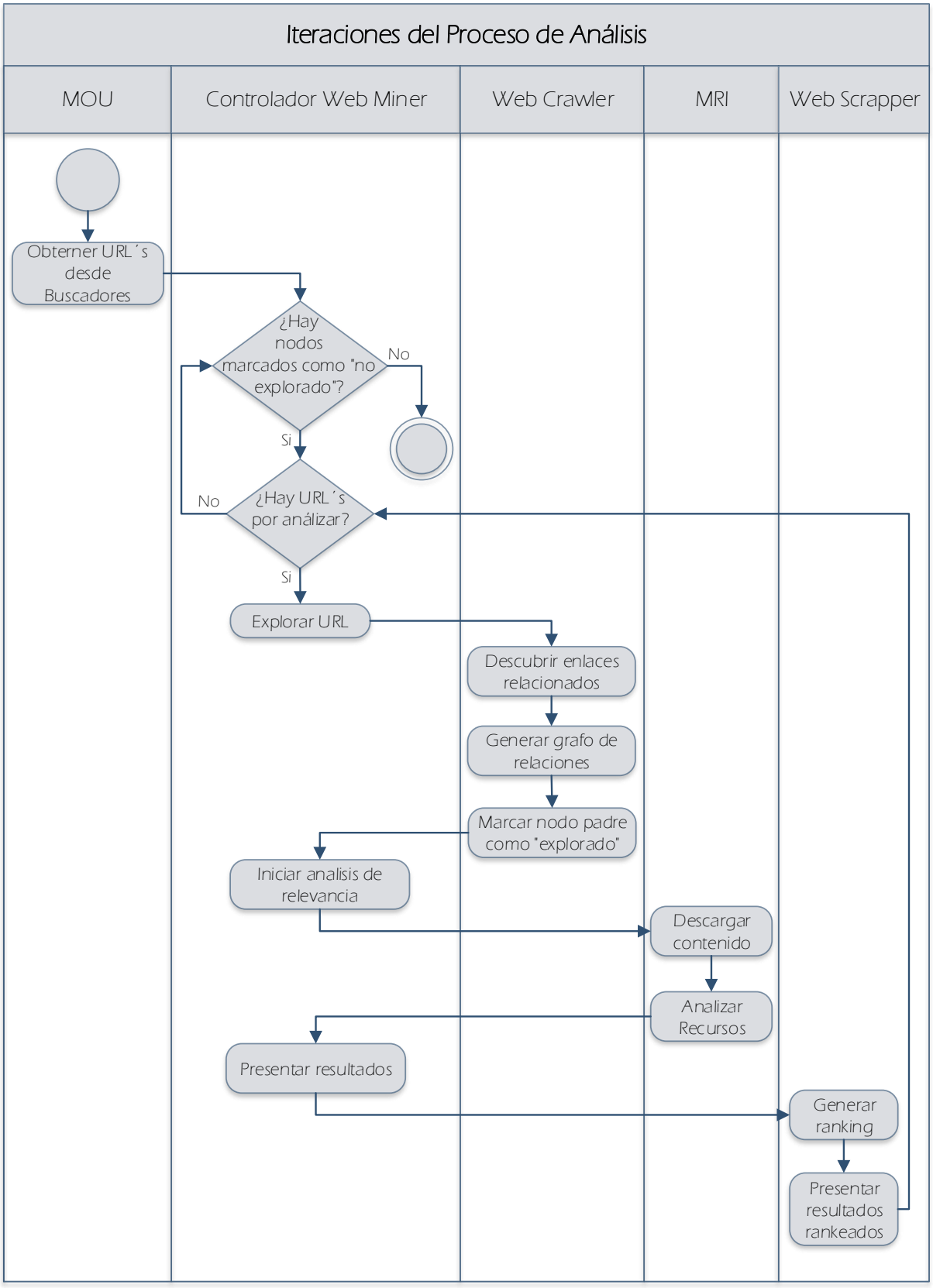

Figura 4.2 *- Diagrama de actividad del proceso de análisis*

Este ranking consta de cincuenta posiciones, con el fin de limitar la cantidad de información presentada al usuario. Para generarlo, en principio, se agrupan los recursos de acuerdo su dominio. Como se observa en la Figura 4.3, por cada grupo se posee un recurso

principal, cuyo puntaje otorgado por el MILS es el mayor y un conjunto de recursos relacionados, pertenecientes al mismo dominio.

Ya con los grupos conformados, se determinan las cincuenta posiciones del ranking, considerando los puntajes arrojados por el MILS de los recursos principales de cada dominio.

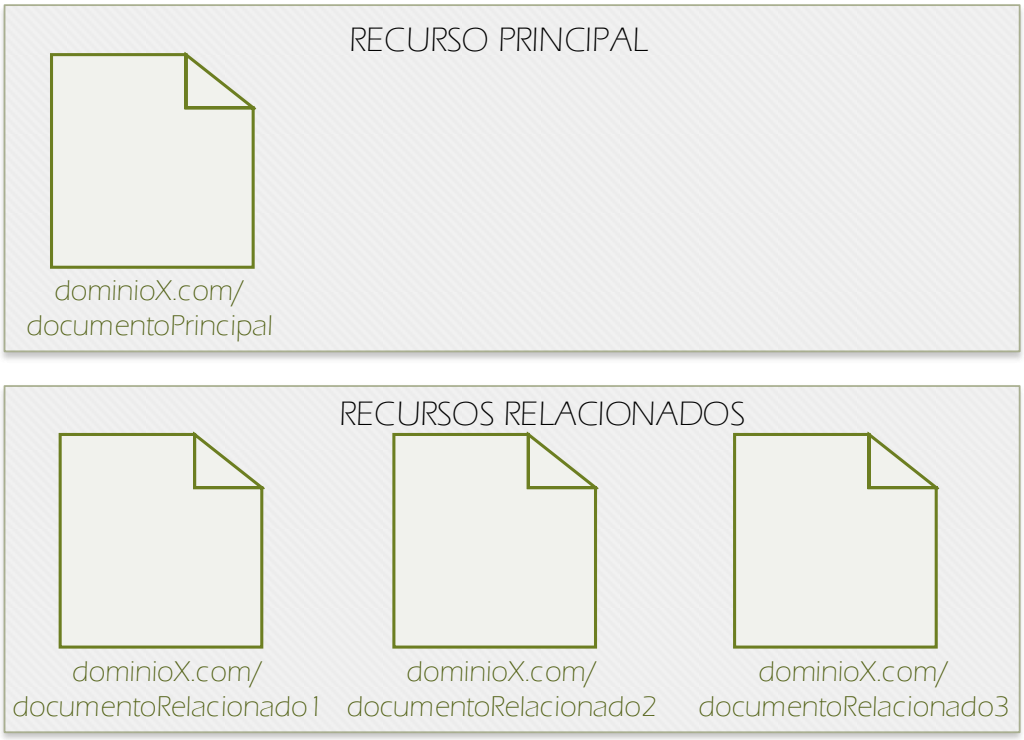

Figura 4.3 *- Ejemplo de una posición del ranking generado en el módulo* ***Web Scrapper***

Por cada posición del ranking, se genera un archivo JSON que contiene el puntaje de relevancia correspondiente al recurso principal, la posición que ocupa en el ranking y las URL's de los recursos relacionados al mismo. Dichos archivos son almacenados en un repositorio local, permitiendo que sean recuperados por el MRR, para presentarlos al usuario.

El proceso continúa con la expansión de los demás nodos marcados como "No Explorado" (indicado con color rojo en la Figura 4.4). Finalizada la primera iteración, se procede a verificar si existen nodos obtenidos a partir del descubrimiento de enlaces relacionados (indicado con color azul en la Figura 4.4).

De ser así, se da inicio a una nueva iteración, lo que implica el descubrimiento de nuevos nodos y la realización del análisis de relevancia por cada uno de ellos, resultando en una reestructuración del ranking a ser presentado al usuario. En caso contrario, se da por finalizado el proceso de análisis de recursos y generación de rankings. Además, cabe destacar que el usuario puede finalizar dicho proceso cuando desee, lo que supone otro punto de corte para la ejecución del sistema.

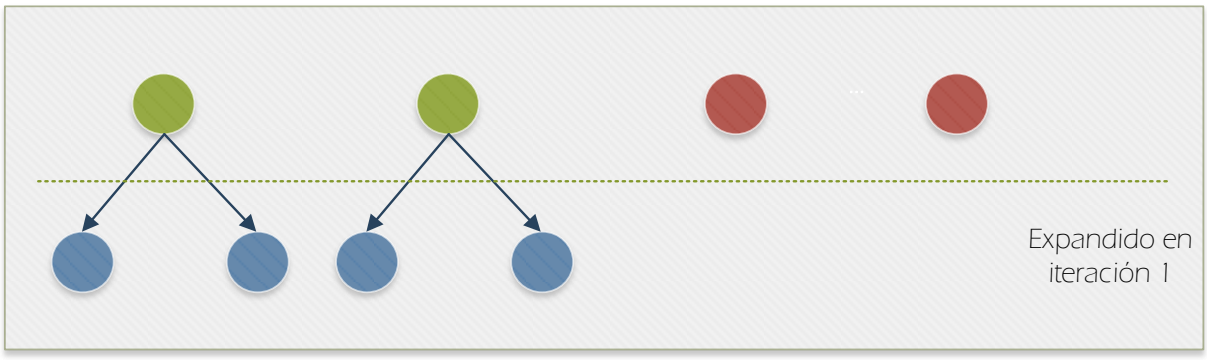

Figura 4.4 *- Ejemplo de exploración de nodos*

En las siguientes secciones, se presentan detalles de la implementación del ***MAS*** y el MILS, que representan los objetivos del presente trabajo.

4.2 MÓDULO DE ANÁLISIS SEMÁNTICO

Para determinar la relevancia de recursos se tienen dos técnicas: las TL, que actualmente se encuentran puestas en producción, y las TS, que son las planteadas e implementadas en este trabajo mediante el ***MAS***, cuyo funcionamiento se explica en esta sección.

El objetivo de este módulo es implementar métricas de relación y similitud semántica, que contribuyan a determinar la relevancia de los recursos analizados. El esquema general de este módulo se presenta en la Figura 4.5.

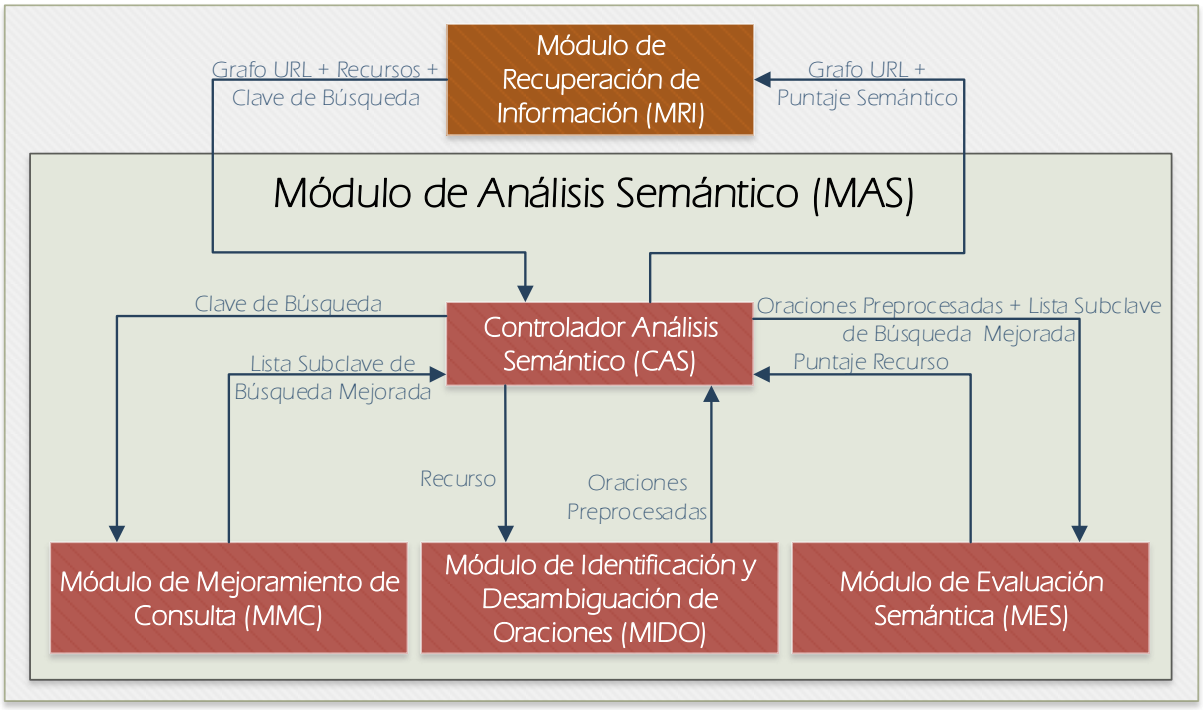

Figura 4.5 *- Esquema del* ***MAS***

Como se puede apreciar, el ***MAS***, está compuesto a su vez por un conjunto de módulos que interactúan entre sí, donde el ***Controlador Análisis Semántico (CAS)*** es el encargado de coordinar toda su operatoria.

El proceso de análisis semántico comienza al recibir del ***MRI*** el Grafo URL a analizar y la clave de búsqueda, donde cada nodo del grafo está compuesto por su correspondiente recurso. Las actividades que se llevan a cabo en este módulo, se plasman en la Figura 4.6.

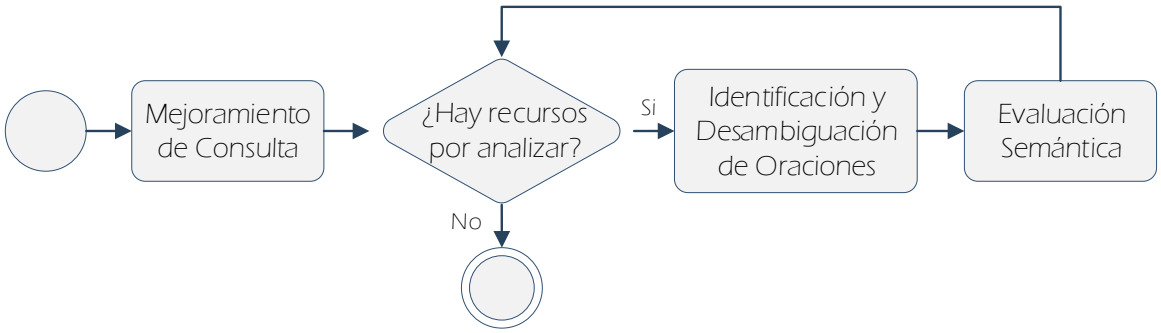

Figura 4.6 *- Diagrama de actividad del* ***MAS***

En primera instancia, el ***CAS*** envía la clave de búsqueda al ***Módulo de Mejoramiento de Consulta (MMC)***, donde se eliminan errores ortográficos, los stopwords[5] y se identifica el sentido de los términos que la componen.

Además, se la segmenta en subclaves, lo que permite que se tenga en cuenta la importancia de las distintas partes de la clave de búsqueda. Como resultado, se retorna la Lista de Subclave de Búsqueda Mejorada al ***CAS***.

Seguidamente, se envía un recurso correspondiente a un nodo, al ***Módulo de Identificación y Desambiguación de Oraciones (MIDO)***, donde se segmenta a su contenido en las oraciones que lo conforman. Por cada oración, se desambigua el sentido de las palabras que la componen. Como resultado, se obtiene una lista de Oraciones Preprocesadas, que se retorna al ***CAS***.

A continuación, se envía esta lista de Oraciones Preprocesadas junto a la Lista Subclave de Búsqueda Mejorada al ***Módulo de Evaluación Semántica (MES)***, donde se aplica la métrica de relación y similitud semántica para determinar el puntaje de relevancia correspondiente al recurso analizado. Finalmente se retorna el Puntaje Recurso al ***CAS***.

Hecho esto, se comprueba si existen nodos por analizar en el grafo, de ser así, se vuelve a realizar el mismo procedimiento. En caso contrario, se finaliza el análisis semántico, retornando al ***MRI*** el Grafo URL junto a los puntajes semánticos correspondientes.

En las secciones siguientes, se amplía en detalle el funcionamiento de los módulos que forman parte del ***MAS***.

4.2.1 Módulo de Mejoramiento de Consulta

El ***MMC*** lleva a cabo el preprocesamiento de la clave de búsqueda, necesario para poder evaluar la correspondencia semántica existente entre el contenido de los recursos y la clave de búsqueda. Los módulos que lo componen, junto a sus interacciones, se presentan en la Figura 4.7.

[5] **Stopword:** Palabras vacías, o comúnmente utilizadas y que no contribuyen al significado.

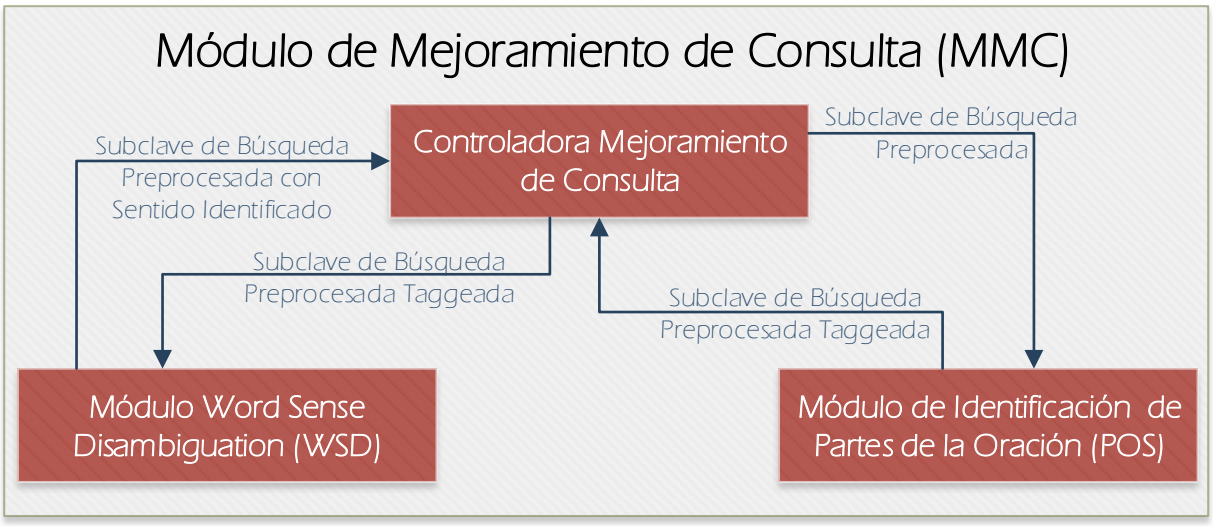

Figura 4.7 *- Esquema del* ***MMC***

El proceso comienza segmentando la clave de búsqueda recibida desde el ***CAS***, teniendo en cuenta los conectores "AND" y "OR" definidos por el usuario en el ***MRR*** (ver Figura 4.8). Esto permite que en pasos posteriores se establezca una ponderación de acuerdo a la importancia de las distintas partes de la clave de búsqueda, teniendo como base el orden en el que fueron ingresadas. De esta segmentación, se obtiene un conjunto de subclaves de búsqueda.

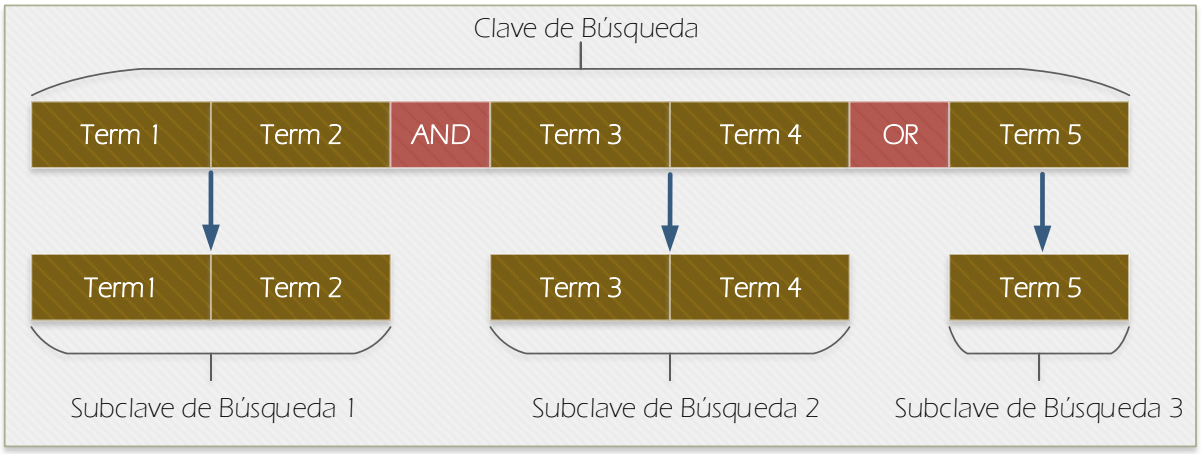

Figura 4.8 *- Segmentación de la Clave de Búsqueda*

Por cada subclave, inicialmente se corrigen palabras con errores ortográficos, actividad que se conoce como "*Spelling*". En este trabajo, se utiliza el método propuesto por Peter Norvig [80], que consiste en la corrección de palabras mediante los siguientes cuatro pasos:

- **Generación de candidatos de distancia uno:** Se genera una lista de palabras candidatas a ser correcciones de la palabra que presuntamente posee errores ortográficos, mediante la aplicación de cuatro tipos de operaciones:
    - **Borrado:** Eliminar una letra de la palabra. Ejemplo: a *'spellling'* se le elimina la letra '*l',* obteniendo '*spelling'*.
    - **Transposición:** Cambiar la posición de dos letras adyacentes. Ejemplo: a partir de '***ps****elling*' se obtiene '***sp****elling'*.

- **Reemplazo:** Consiste en cambiar una letra por otra. Ejemplo: a partir de '*sp**a**lling'* se obtiene '*sp**e**lling'*.
- **Inserción:** Agregar una letra a la palabra. Ejemplo: a partir de '*splling*' se obtiene '*sp**e**lling'.*

- **Generación de candidatos de distancia dos:** Se genera una lista de correcciones candidatas, aplicando las operaciones descriptas en el paso anterior, a la lista de candidatos de distancia uno, lo cual permite tener mayor cantidad de posibilidades de comparación.
- **Cálculo de probabilidades de candidatos:** Se calcula la probabilidad de ocurrencia de cada candidato generado en los pasos anteriores, dentro de un corpus de un millón de palabras, eliminando aquellos candidatos cuya probabilidad de ocurrencia sea 0. Esta probabilidad se obtiene mediante la Ecuación (4.1):

$$P(c) = \frac{\#\ ocurrencias\ de\ c}{\#\ total\ de\ palabras} \tag{4.1}$$

  Donde c es una de las correcciones candidatas para una palabra determinada.

- **Selección del candidato más adecuado:** Consiste en la selección de un candidato, como corrección de una palabra determinada, teniendo en cuenta al siguiente orden de prioridad:
  - La palabra original, si no posee errores ortográficos.
  - El candidato de la lista de palabras generadas con distancia uno, cuya probabilidad sea máxima, si es que existe.
  - El candidato de la lista de palabras generadas con distancia dos, cuya probabilidad sea máxima, si es que existe.
  - La palabra original, aunque posea errores ortográficos.

De esta subclave libre de errores ortográficos, se eliminan los *stopwords* (palabras ampliamente utilizadas, que no contribuyen a determinar el contexto, como 'the', 'in', etc.) y las palabras resultantes se convierten a su raíz, agrupando las morfológicamente relacionadas (ejemplo: '*Argued* (argumentó) es llevado a su raiz '*Argue*' (argumentar)). Estos dos procesos son llevados a cabo mediante las funciones provistas por la librería *NLTK* [81].

Posteriormente, se envía la subclave de búsqueda preprocesada al ***Módulo de Identificación de Partes de la Oración (POS)***, donde se identifica la categoría sintáctica de cada palabra que la conforma, teniendo en cuenta el contexto. Esta actividad, se realiza mediante el método *Parse*, perteneciente a la librería *Pattern* [82].

Las posibles etiquetas para una palabra son: Sustantivo, Adjetivo, Adverbio y Verbo. Para una mejor comprensión, se proporciona un ejemplo de esta actividad en la Figura 4.9.

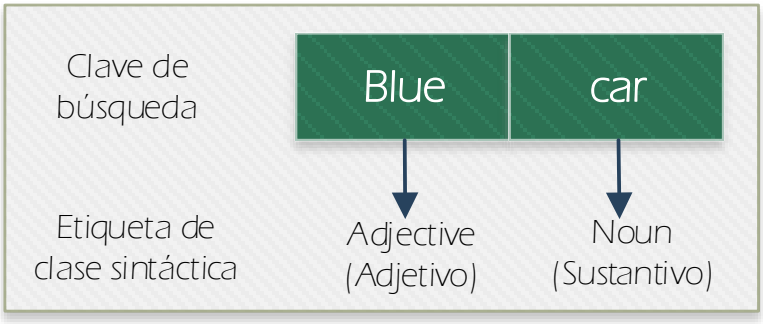

Figura 4.9 *- Ejemplo de etiquetado de partes de la oración*

Como resultado de esta actividad, se retorna la Subclave de Búsqueda Preprocesada Taggeada a la ***Controladora Mejoramiento de Consulta***.

Posteriormente, esta subclave se envía al ***Módulo Word Sense Disabiguation (WSD)***, donde por cada palabra que la conforma, se busca identificar el sentido más aproximado al contexto al que pertenece, teniendo en cuenta su categoría sintáctica. El funcionamiento de este módulo, se explica en la sección siguiente.

Como resultado de la ejecución del ***WSD***, se retorna esta Subclave de Búsqueda con Sentido Identificado a la ***Controladora Mejoramiento de Consulta***, donde se genera una lista de subclaves que contendrá a todas aquellas que resultan de los procesos antes descriptos, junto a la posición en la que el usuario la ingresó dentro de la clave de búsqueda.

A continuación, se aplica el mismo procedimiento a las siguientes subclaves, si es que existen. En caso contrario, se retorna la Lista de Subclave de Búsqueda Mejorada al ***CAS***, con el fin de continuar con los demás pasos del análisis semántico.

4.2.1.1 Módulo Word Sense Disambiguation

Como se explicó en la sección 3.1, desde el punto de vista de la semántica, cada palabra posee un significado independiente del contexto, conocido como denotación y un significado dependiente del contexto, conocido como connotación (sentido). Este contexto es determinado por la oración o la subclave a la cual pertenece la palabra.

Teniendo en cuenta esto, el objetivo de este módulo es determinar dicha connotación, para lo cual se presenta su diagrama de actividad en la Figura 4.10.

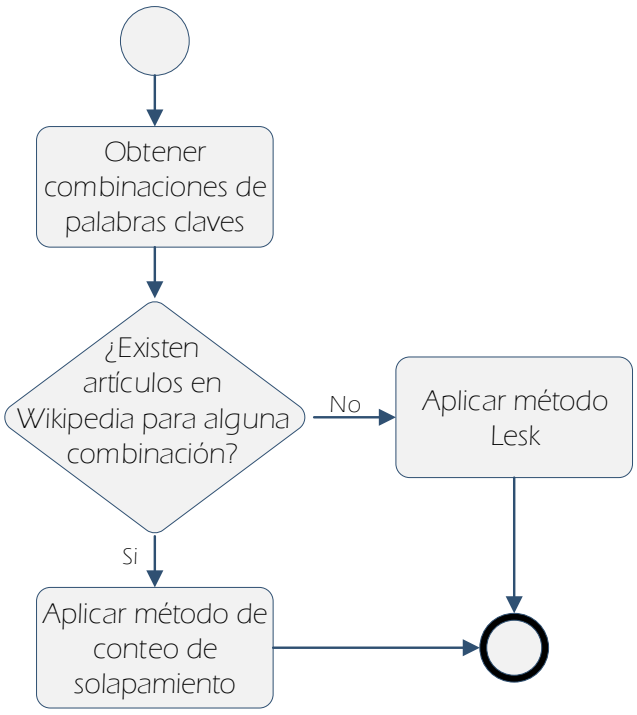

Figura 4.10 *- Diagrama de Actividad del* ***WSD***

Una vez que se recibe de la ***Controladora Mejoramiento de Consulta*** la Subclave de Búsqueda Preprocesada Taggeada, se generan combinaciones entre palabras yuxtapuestas de la subclave, respetando el orden en el que se ingresan. Un ejemplo de esto se puede observar en la Figura 4.11.

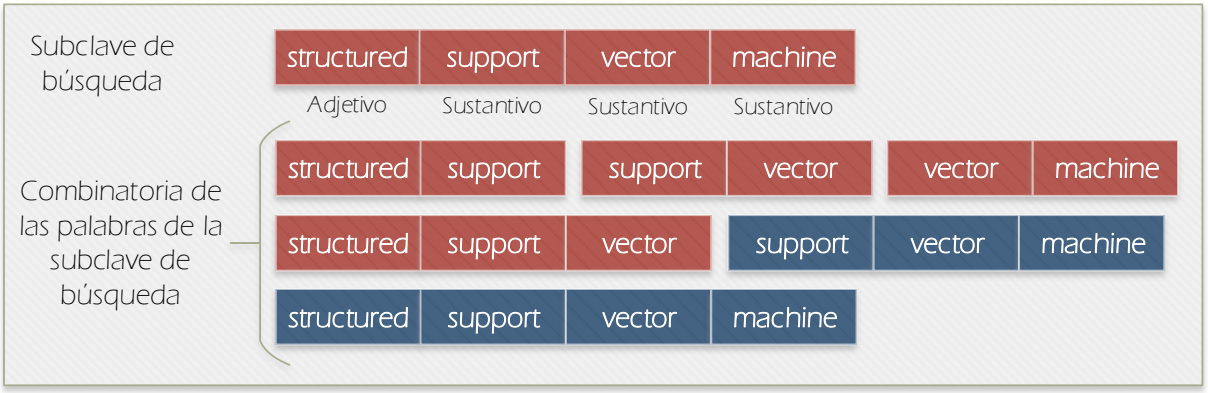

Figura 4.11 *- Ejemplo de combinación de Clave de Búsqueda*

Con estas combinaciones, se verifica en Wikipedia si existen artículos cuyo título se corresponda con alguna de ellas. Tomando el ejemplo que se presenta en la Figura 4.11, inicialmente se busca artículos con el título "*structured support*", luego se intenta con "*support vector*" y así sucesivamente.

El paso siguiente, va a depender de si se encontraron artículos en Wikipedia para alguna de las combinaciones generadas. Si se encontró algún artículo, se comienza con el método de conteo de solapamiento, cuya actividad se representa mediante la Figura 4.12 y se describe en los párrafos posteriores.

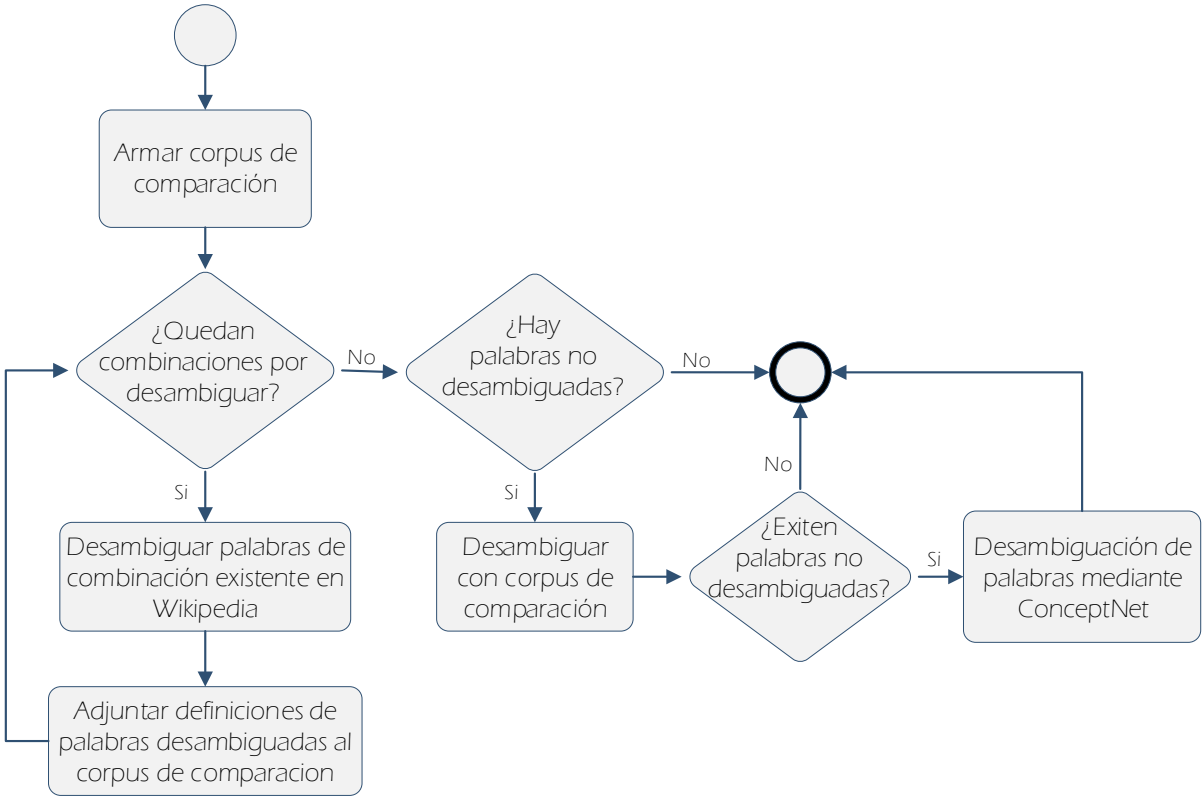

Figura 4.12 *- Diagrama de actividad del método de conteo de solapamiento*

En principio, se crea un corpus de comparación mediante la combinación de todos los artículos encontrados (ver Figura 4.13). Este corpus surge con el fin de obtener una mayor cantidad de palabras que permitan definir el contexto más adecuado al que se enmarca la subclave de búsqueda.

Al tratarse de artículos cuyo título es una combinación de palabras yuxtapuestas de la subclave de búsqueda, se asume que definen a tales palabras en un sentido determinado, descartando todos los otros sentidos posibles.

Para el ejemplo de la Figura 4.11, se tienen dos combinaciones que coinciden con artículos en Wikipedia (indicados con color azul), que son "*support vector machine*" y "*structured support vector machine*". Estos artículos se combinan con el objetivo de obtener el corpus de comparación a ser utilizado por el ***WSD*** (ver Figura 4.13).

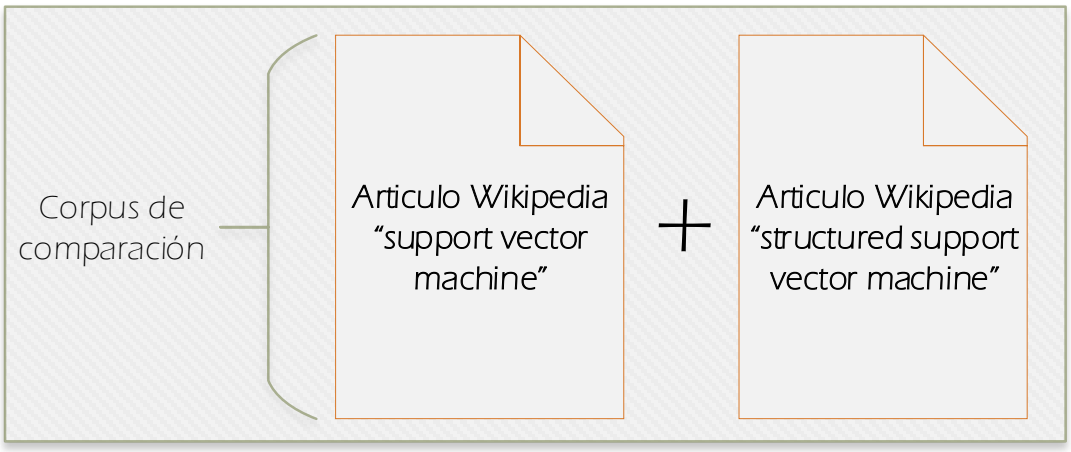

Figura 4.13 *- Corpus generado para la búsqueda de la Figura 4.11*

Ya con el corpus generado, se desambiguan las palabras de las combinaciones que tienen asociadas artículos en Wikipedia.

Para esto, por cada palabra se obtiene, mediante la taxonomía WordNet (presentada en la sección 3.4.1), el conjunto de definiciones correspondiente a los distintos sentidos que posee, considerando su categoría sintáctica.

Por ejemplo, para la palabra "*support*", que es un sustantivo dentro de la subclave de búsqueda, se obtienen las definiciones presentadas en la Figura 4.14, donde cada una de ellas se corresponde con un sentido distinto.

A partir de estas definiciones, se selecciona como el sentido más adecuado a aquel cuya definición posea mayor cantidad de solapamientos con respecto al corpus de comparación, es decir, el que más palabras de su definición aparezcan en el corpus de comparación.

Por cada palabra desambiguada, se agrega al corpus de comparación, la definición del sentido más adecuado, junto a las definiciones de sus hipónimos inmediatos y su hiperónimo inmediato, obtenidos mediante WordNet. Esto permite que el corpus de comparación crezca continuamente y agregue palabras que contribuyan a la univocidad del contexto de la subclave de búsqueda.

Seguidamente, se verifica si existen palabras no desambiguadas. En caso de no existir se finaliza el funcionamiento del ***WSD***, retornando a la ***Controladora Mejoramiento de Consulta***, la Subclave de Búsqueda Preprocesada con Sentido Identificado.

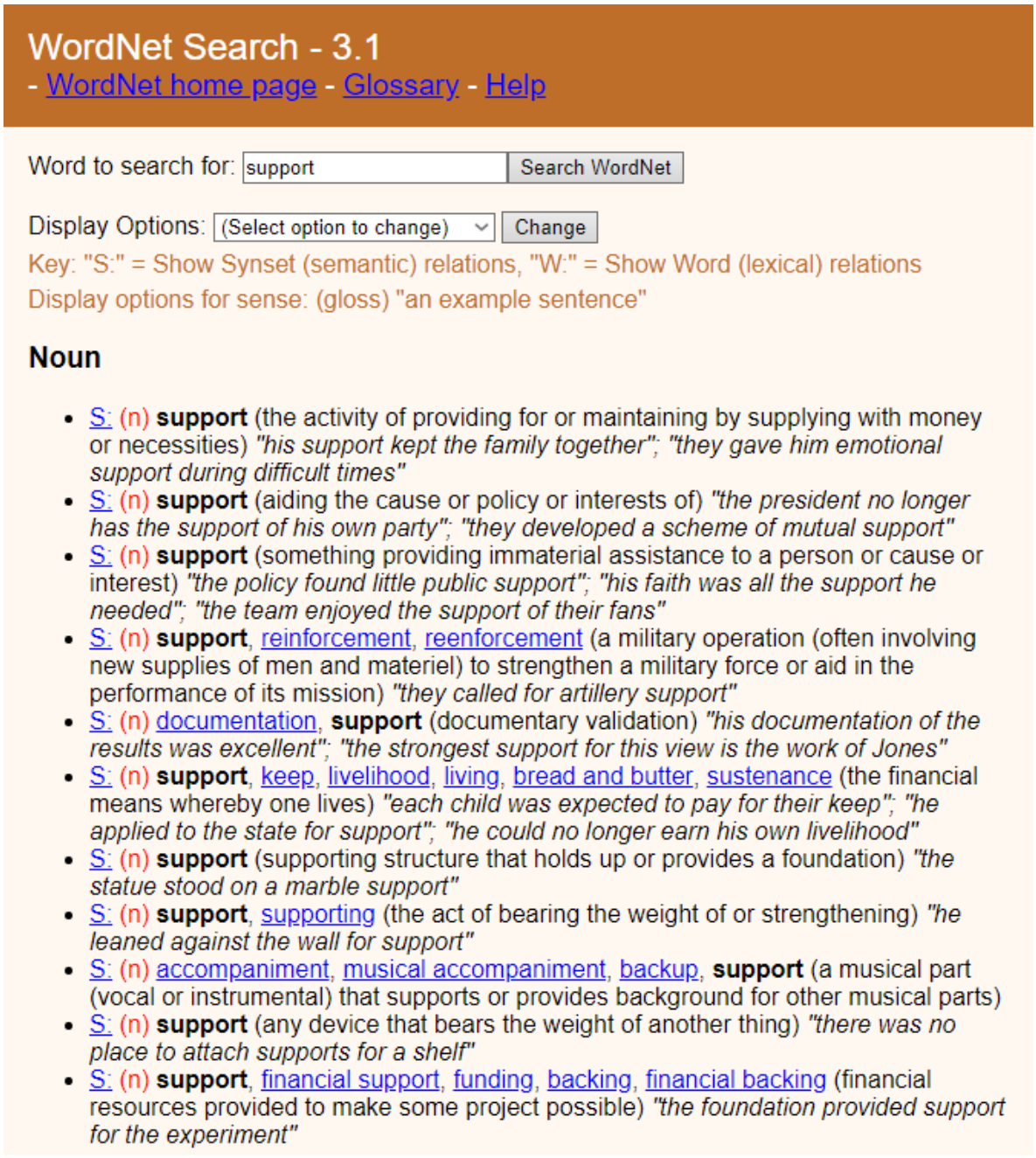

Figura 4.14 *- Definiciones para "support" obtenido a partir de Wordnet [31]*

En cambio, si existen palabras por desambiguar, se procede a desambiguarlas utilizando el corpus de comparación generado en los pasos anteriores, mediante la técnica de conteo de solapamientos. Luego, se comprueba nuevamente si existen palabras no desambiguadas, esto atendiendo al caso de que es posible que no se encuentren definiciones en WordNet para una palabra determinada.

De no existir, se da por finalizado el proceso de desambiguación, retornando a la ***Controladora Mejoramiento de Consulta***, la Subclave de Búsqueda Preprocesada con Sentido Identificado.

Caso contrario, se ejecuta el proceso de desambiguación, que consiste en obtener las palabras relacionadas a las no existentes en WordNet, mediante la taxonomía de ConceptNet (presentada en la sección 3.4.2).

Por cada una de estas palabras relacionadas, se identifica el sentido más adecuado en WordNet, mediante el algoritmo Lesk (presentada en la sección 3.3.1). Esto, debido a que las métricas utilizadas para determinar la relevancia, se aplican únicamente sobre dicha taxonomía.

Seguidamente, se confecciona la subclave de búsqueda final, que está compuesta por las palabras cuyo sentido fue identificado en WordNet.

También, se agregan a esta subclave, aquellas palabras relacionadas descubiertas mediante ConceptNet con sentido identificado en WordNet y, las palabras originales sin desambiguar a partir de las cuales se obtuvieron estas palabras relacionadas.

Esto último evita que se descarten términos que no existan en la taxonomía de WordNet, y permite realizar un análisis de correspondencia lexicográfica que complemente al puntaje de sus palabras relacionadas semánticamente.

Cabe destacar que, si no se encuentran relaciones en ConceptNet para una palabra determinada o ninguna de ellas pudo ser desambiguada, esta se elimina de la subclave.

Finalizado este proceso, se retorna a la ***Controladora Mejoramiento de Consulta***, la Subclave de Búsqueda Preprocesada con Sentido Identificado.

En la Figura 4.15 se presenta un ejemplo de la obtención de palabras relacionadas descripta en los párrafos anteriores. Los Term. 1 y Term. 3, fueron desambiguados mediante WordNet. En cambio, el Term. 2 (indicado con color rojo) no pudo ser desambiguado, por lo que fue necesario la obtención de sus palabras relacionadas mediante ConceptNet.

Como resultado se obtienen tres palabras relacionadas. Para los TermC.1 y TermC.2, se pudo identificar un sentido en WordNet, mientras que el TermC.3 no pudo ser desambiguado (indicado con color rojo) y por lo tanto fue descartado.

La Subclave de Búsqueda Preprocesada con Sentido Identificado, queda compuesta por los términos que fueron desambiguados (Term. 1 y Term. 3) y el Term.2 de la subclave de búsqueda sin sentido identificado, junto a sus palabras relacionadas con sentido identificado (TermC.1 y TermC.2).

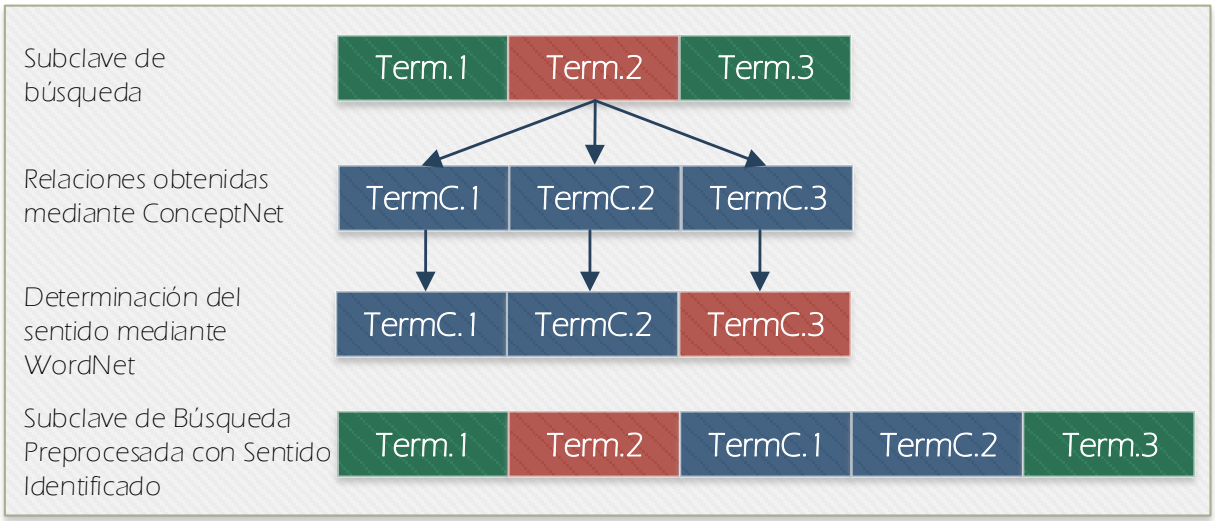

Figura 4.15 *- Ejemplo de obtención de palabras relacionadas en ConceptNet*

Retomando la Figura 4.10, en el caso de que no se encuentren artículos de Wikipedia para ninguna de las combinaciones, las palabras de la subclave se desambiguan mediante el algoritmo Lesk (presentado en la sección 3.3.1). Este proceso se explica mediante el diagrama de actividad presentado en la Figura 4.16.

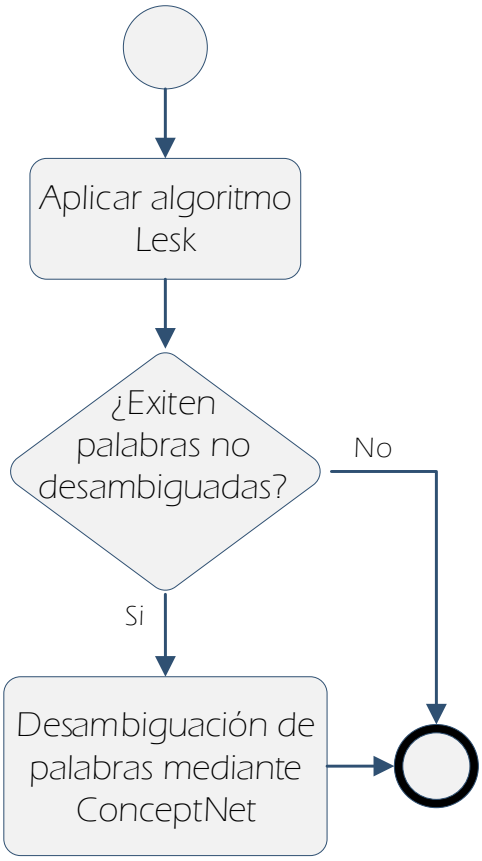

Figura 4.16 *- Diagrama de actividad del método Lesk*

Si alguna de estas palabras no pudo ser desambiguada debido a que no existen en WordNet, se ejecuta para cada una de ellas, el proceso de desambiguación de palabras mediante ConceptNet, descripto en los párrafos anteriores.

Finalmente, se retorna a la ***Controladora Mejoramiento de Consulta***, la Subclave de Búsqueda Preprocesada con Sentido Identificado.

4.2.2 Módulo de Identificación y Desambiguación de Oraciones

Debido a que un texto está compuesto por distintas unidades gramaticales, yendo desde párrafos hasta oraciones y palabras, es posible dividirlo en partes manejables para hacer eficiente su procesamiento.

Teniendo en cuenta esto, una oración, es la unidad mínima en la que se puede dividir un texto, manteniendo presente al contexto, ya que expresa un juicio con sentido completo y autonomía sintáctica, lo que, dicho de otra forma, significa que posee un conjunto de palabras enlazadas, pertenecientes a un mismo contexto.

El objetivo de este módulo es llevar a cabo la división del texto en oraciones, además de realizar todos los procesamientos necesarios que permitan evaluar la correspondencia semántica existente con respecto a la clave de búsqueda. Para esto se dispuso el esquema que se presenta en la Figura 4.17.

El proceso comienza al recibir un recurso desde el ***CAS***, donde se divide su contenido en las oraciones que lo conforman. El principal desafió de esta tarea es delimitar el comienzo y fin de la oración, para la cual existen diversos métodos.

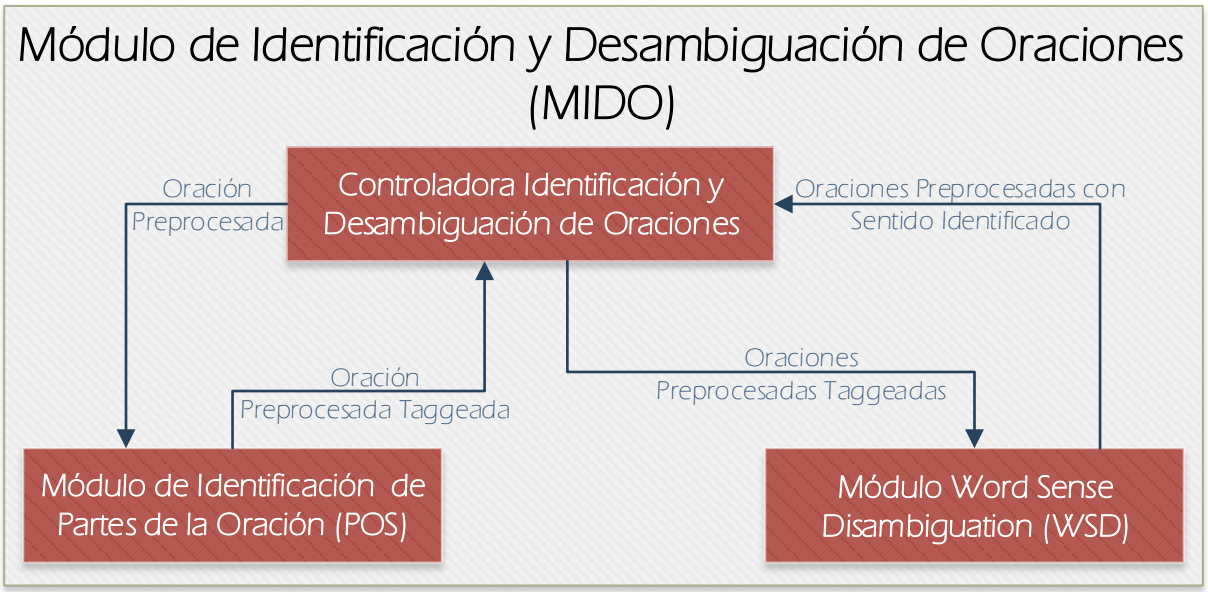

Figura 4.17 *- Esquema del* ***MIDO***

En este módulo, en primera instancia, se segmenta al recurso mediante la utilización de una expresión regular que se propone en [83], la cual se presenta en la Ecuación (4.2).

$$(?<!\w\.\w.)(?<![A-Z][a-z]\.)(?<=\.|\?)(\s|[A-Z].*) \quad (4.2)$$

Esta expresión regular busca representar la estructura de una oración, donde se agrupa a todos aquellos conjuntos de palabras que cumplan con la misma o, dicho de otra manera, las oraciones del recurso serán todas aquellas que satisfagan esta expresión regular. Se puede observar que esta expresión aplica una serie de filtros divididos en cuatro bloques, que se explican a continuación:

- $(?<!\w\.\w.)$: Permite que en una oración determinada aparezca un carácter alfanumérico seguido de un punto y seguido de un carácter alfanumérico, sin necesidad de considerarlos como oraciones distintas. Esto valida casos como “Google.com” o “1.5”.
- $(?<![A-Z][a-z]\.)$: Permite que dentro de una oración aparezca una letra en mayúsculas, seguida de una letra en minúsculas y seguido de un punto, sin considerarlo necesariamente como la finalización de una oración. Este bloque valida casos como “Mr.”, “Jr.”.
- $(?<=\.|\?)$: Este bloque reconoce como indicador de fin de oración cuando se produce la aparición de un punto o un signo de interrogación.
- $(\s|[A-Z].*)$: Este bloque indica que el comienzo de una oración está dado por un espacio en blanco o una letra en mayúsculas seguida de un conjunto infinito de caracteres (establecido infinito por el desconocimiento de la cantidad de caracteres que pueden ocurrir).

Luego se divide a cada oración, mediante los saltos de línea, debido a que no se contemplan en la expresión regular. Como resultado de esto, se obtiene una lista de oraciones que forman parte del recurso recibido.

Por cada una de estas oraciones, se eliminan aquellas palabras consideradas *stopwords* y las resultantes se llevan a su correspondiente raíz de manera que se agrupen las palabras morfológicamente relacionadas. Estos dos procesos son llevados a cabo mediante las funciones provistas por la librería NLTK [81].

A medida que se procesan las oraciones, son enviadas al ***Módulo de Identificación de Partes de la Oración (POS)***, donde mediante la utilización del método *Parse* de la librería *Pattern* [82], se identifica la categoría sintáctica a la que pertenece cada palabra. Luego de realizar este proceso, se retorna la Oración Preprocesada Taggeada a la ***Controladora Identificación y Desambiguación de Oraciones***.

Esta oración, se envía al módulo ***WSD***, donde por cada palabra se identifica el sentido más adecuado mediante el algoritmo Lesk, que realiza la desambiguación de sentidos mediante la taxonomía de WordNet. Si para alguna de las palabras no se encuentra un sentido en WordNet, esta se descarta.

Cabe destacar que, para el proceso de desambiguación aplicado sobre las oraciones, se utiliza únicamente el algoritmo Lesk, por cuestiones de rendimiento y eficiencia.

Como resultado, se retorna al módulo ***Controladora Identificación y Desambiguación de Oraciones*** las Oraciones Preprocesadas con Sentido Identificado, lo que a su vez es retornado al ***CAS***.

4.2.3 Módulo de Evaluación Semántica

El objetivo principal del ***MES*** es determinar la correspondencia existente entre el contenido de los recursos y la clave de búsqueda ingresada por el usuario, teniendo como criterio a la relación y similitud semántica. El funcionamiento de este módulo se representa en la Figura 4.18.

Inicialmente se recibe el conjunto de oraciones del recurso a analizar y la lista de subclave de búsqueda desde el ***CAS***.

En primera instancia, se calculan las ponderaciones asociadas a cada término conformante de la clave de búsqueda, con el fin de diferenciar el aporte al puntaje final de los términos que representan el tema principal de búsqueda (que poseen mayor importancia), de aquellos que son orientativos de la búsqueda (y por ende son menos importantes).

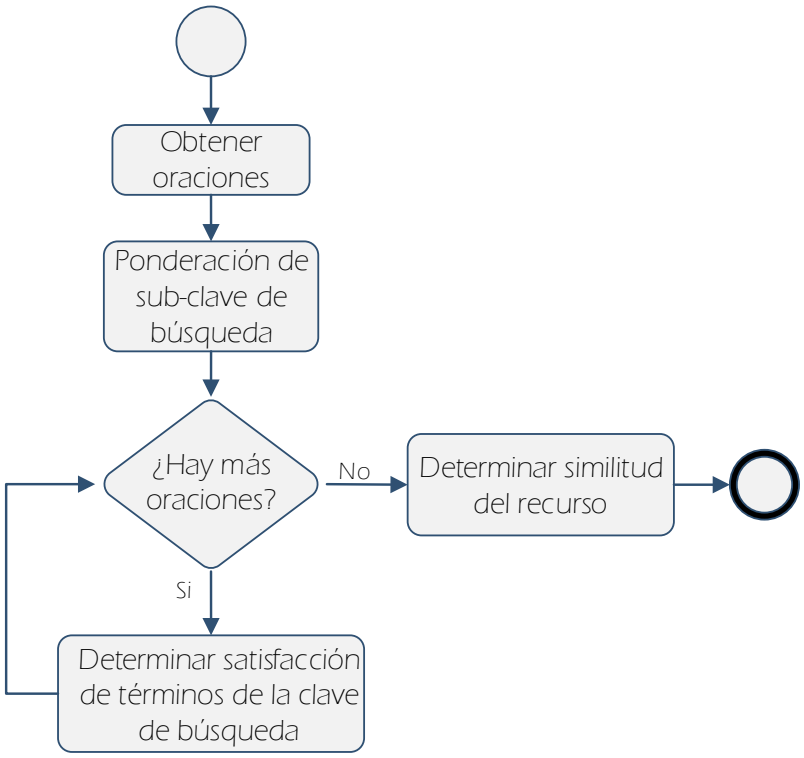

Figura 4.18 - *Funcionamiento del* ***MES***

Para calcular dicha ponderación, se propone utilizar orden en que fue ingresada la subclave de búsqueda a la que pertenece cada término, como indicativo de la importancia de este último. Este valor se obtiene mediante la aplicación de la Ecuación (4.3) [84].

$$Ponderación_{ji} = 1 - (j-1) * \left(\frac{1}{m}\right) \quad (4.3)$$

Donde $Ponderación_{ji}$ es la ponderación del $i-ésimo$ término perteneciente a la subclave de búsqueda ingresada en la posición j, y m es la cantidad de subclaves existentes en la clave de búsqueda.

Por otro lado, dado que el puntaje de relevancia para un recurso determinado, pertenece a un intervalo cerrado [0,1], es necesario normalizar cada ponderación, de manera que no se sobrepasen dichos límites. Para conseguir esto se propone la Ecuación (4.4) [84].

$$PonderaciónNormalizada_{ji} = \frac{Ponderación_{ji}}{\sum_{j,i=0}^{n} Ponderación_{ji}} \quad (4.4)$$

Después, se genera una única clave de búsqueda compuesta por los términos $Query_{ji}$, correspondientes a cada subclave recibida y sus respectivas $PonderaciónNormalizada_{ji}$, con el fin de agilizar el análisis de correspondencia a ejecutar en pasos posteriores.

Luego, por cada oración del recurso a analizar, se determina la satisfacción de las mismas a cada término $Query_{ji}$ de la clave de búsqueda.

Estas satisfacciones se obtienen acumulando por cada término $Query_{ji}$, las relaciones semánticas existentes con términos $Termino_k$ de la oración analizada, cuyo valor es determinado por la métrica desarrollada por Slimani [71] (explicada en la sección 3.5.1). Esto se puede observar en la Figura 4.19.

Dicha acumulación debe satisfacer dos limitaciones. La primera es que $Query_{ji}$ y $Termino_k$ deben pertenecer a la misma categoría sintáctica. Y la segunda es que la métrica de Slimani, debe arrojar un valor de relación semántica de al menos 0.5.

La primera limitación está fundada en que los tipos de relaciones contempladas (relaciones semánticas clásicas) se dan únicamente entre palabras pertenecientes a la misma categoría sintáctica (ver sección 3.2.1), mientras que la segunda limitación tiene asidero en que existen relaciones débiles entre palabras, que pueden provocar la introducción de ruido al análisis semántico.

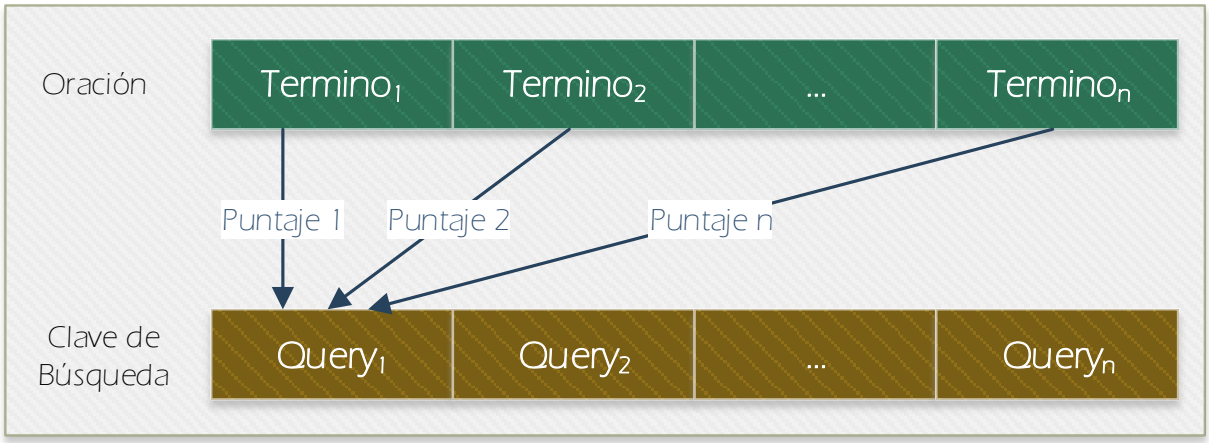

Figura 4.19 *- Determinación de satisfacción de términos de la Clave de Búsqueda*

La acumulación, se lleva a cabo por cada oración del recurso a analizar, por lo que, al procesar todas las oraciones, se obtiene la satisfacción acumulada de cada término $Query_{ji}$. Posteriormente, se calcula la satisfacción promedio de cada $Query_{ji}$, con el fin de obtener un puntaje entre 0 y 1. Para esto, se propone la Ecuación (4.5) [84].

$$PuntajePromedioQuery_{ji} = \frac{PuntajeQuery_{ji}}{TermRelacionados_{ji}} \quad (4.5)$$

Donde $PuntajeQuery_{ji}$ es la sumatoria de todas las relaciones (mayores o iguales a 0.5) que posee $Query_{ji}$, y $TermRelacionados_{ji}$ es el conteo de los términos que obtuvieron una valor de relación semántica de al menos 0.5 con $Query_{ji}$.

A continuación, a cada $PuntajePromedioQuery_{ji}$ se lo multiplica por la $PonderaciónNormalizada_{ji}$ calculada en la Ecuación (4.4), de manera que se tenga en cuenta su importancia. Para esto, se propone la Ecuación (4.6) [84].

$$SatisfacciónQuery_{ji} = PuntajePromedioQuery_{ji} * PonderacionNormalizada_{ji} \quad (4.6)$$

Alcanzado este punto, se calcula el puntaje semántico del recurso, para lo que se propone la Ecuación (4.7) [84].

$$PuntajeSemántico_x = \sum_{j=0}^{m} \left(\frac{\sum_{i=0}^{n} SatisfacciónQuery_{ji}}{n} \right) \quad (4.7)$$

Donde $\sum_{i=0}^{n} SatisfacciónQuery_{ji}$ es la sumatoria de las $SatisfaccionQuery_{ji}$ de la subclave de búsqueda j y n es la cantidad de términos $Query_{ji}$ de la subclave j.

De esta manera, se conserva la satisfacción promedio de cada subclave, lo cual es necesario teniendo en cuenta que cada una de estas representa un aspecto particular de la clave de búsqueda.

Entonces, $PuntajeSemántico_x$ es la sumatoria de la satisfacción promedio de cada subclave j perteneciente a la clave de búsqueda original introducida por el usuario.

Este puntaje representa, a su vez, la correlación semántica existente entre el recurso x analizado y la clave de búsqueda, lo que se traduce en la relevancia semántica de dicho recurso.

Para una mejor comprensión del proceso llevado a cabo en el ***MES***, se plantea el siguiente ejemplo: Dado una clave de búsqueda, conformada por dos subclaves, se procede a determinar el $PuntajeSemántico_x$ para cuatro recursos. Estas subclaves están compuestas a su vez por dos términos cada una (ver Figura 4.20).

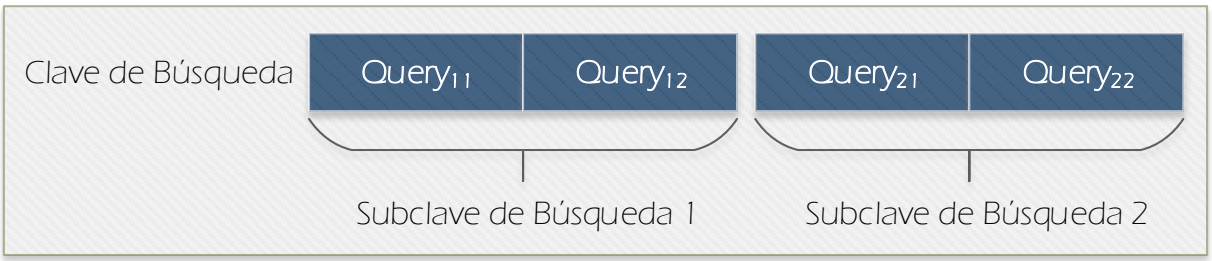

Figura 4.20 *- Clave de Búsqueda de ejemplo*

Inicialmente, se calculan las $Ponderación_{ji}$ y $PonderaciónNormalizada_{ji}$, para cada uno de los términos $Query_{ji}$, mediante la utilización de las Ecuaciones (4.3) y (4.4) respectivamente. Los valores obtenidos, se presentan en la Tabla 4.1.

Tabla 4.1 *- Ponderación para cada* $Query_{ji}$

	$Ponderación_{ji}$	$PonderaciónNormalizada_{ji}$
$Query_{11}$	1	0.66
$Query_{12}$	1	0.66
$Query_{21}$	0.5	0.33
$Query_{22}$	0.5	0.33

Luego se calculan los $PuntajePromedioQuery_{ji}$, considerando en este ejemplo, los valores presentados en la Tabla 4.2, elegidos con el fin de observar como la distribución de estos puntajes afecta a la satisfacción de cada subclave.

Tabla 4.2 *-Valores considerados para cada* $PuntajePromedioQuery_{ji}$

Nro. Recurso	$PuntajePromedioQuery_{ji}$			
	$Query_{11}$	$Query_{12}$	$Query_{21}$	$Query_{22}$
1	0.9	0.8	0.5	0.98
2	0.9	0.7	0.9	0.9
3	0.9	0.8	0.65	0.95
4	0.9	0.8	0.78	0.9

Posteriormente, se calculan las $SatisfacciónQuery_{ji}$ correspondientes a cada uno de los cuatro recursos analizados, utilizando la Ecuación (4.6). Los resultados se presentan en la Tabla 4.3.

Tabla 4.3 *- Calculo de* $SatisfacciónQuery_{ji}$

Rec.	$SatisfacciónQuery_{11}$	$SatisfacciónQuery_{12}$	$SatisfacciónQuery_{21}$	$SatisfacciónQuery_{22}$
1	0.9 * 0.66 = **0.594**	0.9 * 0.66 = **0.594**	0.9 * 0.33 = **0.297**	0.9 * 0.33 = **0.297**
2	0.8 * 0.66 = **0.528**	0.7 * 0.66 = **0.462**	0.8 * 0.33 = **0.264**	0.8 * 0.33 = **0.264**
3	0.5 * 0.66 = **0.33**	0.9 * 0.66 = **0.594**	0.65 * 0.33 = **0.215**	0.78 * 0.33 = **0.257**
4	0.98 * 0.66 = **0.647**	0.9 * 0.66 = **0.594**	0.95 * 0.33 = **0.314**	0.9 * 0.33 = **0.297**

Seguidamente, se calculan los $PuntajeSemántico_x$ mediante la Ecuación (4.7). Para facilitar la comprensión de los valores obtenidos, se divide el cálculo en dos pasos. Inicialmente en la Tabla 4.4 se presenta, la satisfacción promedio de cada subclave de búsqueda.

Tabla 4.4 *- Cálculo de satisfacción promedio por cada Subclave*

Rec.	Satisfacción Subclave 1	Satisfacción Subclave 2
1	(0.594 + 0.594) / 2 = **0.594**	(0.297 + 0.297) / 2 = **0.297**
2	(0.528 + 0.462) / 2 = **0.495**	(0.264 + 0.264) / 2 = **0.264**
3	(0.33 + 0.594) / 2 = **0.462**	(0.215 + 0.257) / 2 = **0.236**
4	(0.647 + 0.594) / 2 = **0.621**	(0.314 + 0.297) / 2 = **0.306**

Luego, se determina el $PuntajeSemántico_x$ correspondiente a cada recurso, mediante la sumatoria de los valores presentados en la tabla anterior. Esto se puede observar en la Tabla 4.5.

Tabla 4.5 *- Cálculo de* $PuntajeSemático_x$

Rec.	$PuntajeSemántico_x$
1	0.594 + 0.297 = **0.891**
2	0.495 + 0.264 = **0.759**
3	0.462 + 0.236 = **0.698**
4	0.621 + 0.306 = **0.927**

En el ejemplo presentado, se observa que debido a la utilización de la Ecuación (4.7), emerge una dificultad significativa para lograr alcanzar valores máximos o al menos acercarse a ellos, requiriendo para esto que los $PuntajePromedioQuery_{ji}$ sean muy próximos a 1.

Esta dificultad, se ve incrementada debido a que las relaciones existentes para cada termino $Query_{ji}$ no siempre son las de sinonimia o relaciones semánticamente cercanas, lo cual provoca que se disminuya el $PuntajeSemántico_x$ a obtener.

Para flexibilizar los resultados obtenidos a partir de la ecuación (4.7), se propone una fórmula que consiste en una función sigmoidal o también conocida como logística y permite alcanzar los extremos sin necesidad de que los valores de $PuntajePromedioQuery_{ji}$ sean cercanos a 1. Esta fórmula se presenta en la Ecuación (4.8) [84].

$$PuntajeRecurso_x = \frac{1}{1 + e^{-[10*(PuntajeSemántico_x - 0.5)]}} \quad (4.8)$$

Donde $PuntajeSemántico_x$ se obtiene mediante la Ecuación (4.7). El comportamiento de la Ecuación (4.8) puede ser observado en la Figura 4.21, donde se aprecia que se logran valores cercanos a 1, cuanto mayor sea el valor del $PuntajeSemantico_x$. En el caso contrario, para valores de $PuntajeSemántico_x$ cercanos a 0, el divisor tenderá a infinito, por lo que el resultado de la división se acercará a 0.

Una particularidad de esta función se da cuando $PuntajeSemántico_x$ se aproxima a 0.5, provocando que $PuntajeRecurso_x$ también se aproxime a dicho valor, es decir, la función posee características lineales para valores cercanos a 0.5, mientras que suaviza la llegada a los extremos para valores inferiores y superiores a 0.5.

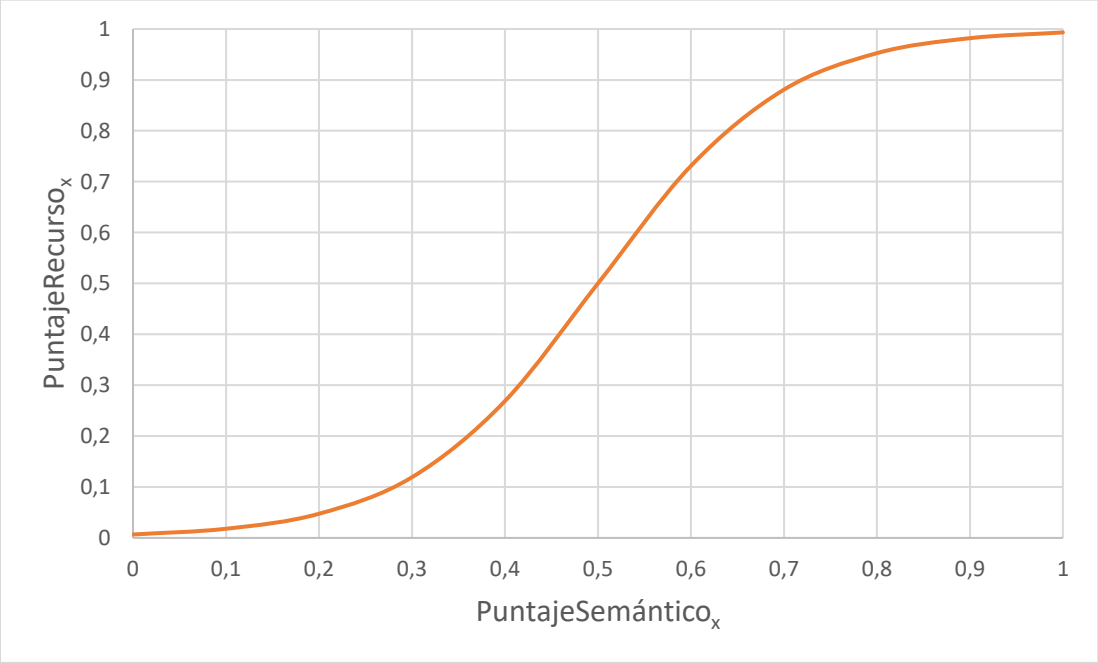

Figura 4.21 *- Comportamiento de la Ecuación (4.8)*

Presentada la Ecuación (4.8), se procede a calcular el $PuntajeRecurso_x$, para cada recurso del ejemplo anterior. Los resultados obtenidos se plasman en la Tabla 4.6.

Tabla 4.6 *- Calculo de $PuntajeRecurso_x$ a partir del $PuntajeSemántico_x$*

Rec.	$PuntajeSemántico_x$	$PuntajeRecurso_x$
1	0.891	0.98
2	0.759	0.93
3	0.698	0.88
4	0.927	0.986

A partir de este ejemplo, se observa una mejora significativa en los puntajes obtenidos por la Ecuación (4.8), con respecto a los obtenidos por la Ecuación (4.7), dado que permite

alcanzar valores máximos para recursos que posean características destacables desde el punto de vista de la semántica.

Finalizando con la descripción de la funcionalidad del ***MES***, una vez que se obtuvo el $PuntajeRecurso_x$ del recurso analizado, se lo retorna al ***CAS***.

4.2.4 Elección de Métrica de Relación y Similitud Semántica de Pares de Palabras

El modelo propuesto en la sección anterior, representa una aproximación para la determinación de la relevancia de recursos con respecto a una clave de búsqueda, mediante la estimación de la relación y similitud semántica existente entre ambos.

Para realizar esta estimación, el modelo utiliza una métrica que determina la relación y similitud semántica entre pares de palabras, con el fin de evaluar cada término de las oraciones de los recursos con respecto a cada término de la clave de búsqueda. La métrica utilizada es la propuesta por Slimani et al. [71] descripta en la sección 3.5.1.

Esta métrica se seleccionó mediante la comparación con respecto a diez consideradas. Para esto, se toma un conjunto de pares de palabras y se determina, por cada una, su puntaje de relación y similitud semántica mediante cada métrica.

El conjunto de pares de palabras a utilizar se obtuvo mediante la colección de prueba *WordSimilarity – 353* [85], que consta de 353 pares de palabras para las que trece expertos evalúan la relación y similitud semántica existente entre ellos, siendo el puntaje final para cada par de palabra, el promedio de los trece puntajes otorgados.

Luego, se evalúa la coincidencia existente entre los puntajes otorgados por cada métrica, con respecto a los otorgados por los expertos.

Para medir dicha coincidencia, se utiliza el coeficiente correlación de Spearman [86], que otorga valores en un intervalo cerrado $[-1,1]$, para indicar la correlación existente entre dos conjuntos de datos. Un valor cercano a -1, indica una correlación negativa fuerte o que a medida que los valores de un conjunto aumentan, los del otro conjunto disminuyen. Un valor cercano a 1, indica una correlación positiva fuerte, lo que significa que ambos conjuntos tienen comportamientos similares. Un valor de 0, indica que ambos conjuntos no tienen relación.

Los resultados de correlación obtenidos por cada métrica, son presentados en la Tabla 4.7, donde se puede observar que la métrica propuesta por Slimani, es la más certera con relación a la evaluación realizada por los distintos expertos.

Tabla 4.7 *- Correlación de Spearman para las Métricas de Relación y Similitud Semántica*

Métrica	*Correlación*
Slimani	0,3541
Wu and Palmer	0,3538
Resnik	0,3467
Li	0,3293
Shortest Path	0,3144
Leacock and Chodorow	0,3144
Lin	0,31
Jiang and Conrath	0,2974
Tversky	0,2765
X-Similarity	0,2596

4.3 MODELO DE INTEGRACIÓN LÉXICO-SEMÁNTICO

Teniendo en cuenta que las TL y las TS poseen criterios de análisis distintos, donde uno busca ocurrencias explicitas de las palabras de la clave de búsqueda en el contenido de los recursos y el otro contempla relaciones semánticas de esta clave, una aproximación posible es combinar ambos criterios de análisis con el fin de obtener un único puntaje que permita estimar de manera más precisa la relevancia de los recursos.

Para ello, mediante la utilización de una fórmula de unificación de rankings, se busca obtener un puntaje final que integre tanto la posición promedio otorgada por las TL como la posición otorgada por las TS.

Al considerar las posiciones otorgadas por cada técnica, se tiene en cuenta el criterio de determinación de relevancia de cada una, ya que generan sus rankings basadas en ese criterio.

Esta fórmula de unificación de rankings se denomina Modelo de Integración Léxico – Semántico (MILS). A continuación, se explica la manera en que se obtiene dicho puntaje para un recurso determinado.

Una vez que se obtienen los puntajes a partir de las TL y las TS para cada recurso analizado en una iteración determinada, el ***Controlador Web Miner*** envía al módulo ***Web Scrapper***, la lista de grafos analizados, que contienen a los nodos con sus recursos y sus puntajes.

Por nodo, se poseen cuatro puntajes distintos, donde tres de ellos corresponden a las TL (*CRank*, *Okapi* y *VSM*) y el restante pertenece a las TS. Entonces, el siguiente paso consiste en generar cuatro rankings de nodos (o recursos), uno por cada técnica, es decir, se genera un ranking para *CRank*, uno para *Okapi*, y así sucesivamente.

Posteriormente, por cada recurso x, se obtiene el puntaje final mediante la utilización del MILS, comenzando con el cálculo de la $RelevanciaLexicográfica_x$ (Ecuación (4.9)), el cual considera la posición de dicho recurso, dentro de cada ranking correspondiente a las TL.

$$RelevanciaLexicográfica_x = \left(\frac{\left(\frac{1}{Pos_{Crank_x}} + \frac{1}{Pos_{Okapi_x}} + \frac{1}{Pos_{VSM_x}} \right)}{3} \right) \quad (4.9)$$

Donde la $RelevanciaLexicográfica_x$, es el valor total de relevancia para el recurso x obtenido mediante la aplicación de las TL, Pos_Crank_x se corresponde con la posición del recurso x en el ranking generado por la técnica *CRank*, Pos_Okapi_x hace referencia a la posición del recurso x en el ranking generado por la técnica *Okapi*, Pos_VSM_x se corresponde con la posición del recurso x en el ranking generado por la técnica *VSM.* En simples palabras, este puntaje de relevancia, consiste en el promedio de las inversas de las posiciones de los rankings lexicográficos correspondientes al recurso x.

Seguidamente, se obtiene la $RelevanciaSemántica_x$ mediante la utilización de la Ecuación (4.10), la cual se define como la inversa de la posición del recurso x, en el ranking generado por las TS.

$$RelevanciaSemántica_x = \left(\frac{1}{Pos_{Semántico_x}} \right) \quad (4.10)$$

Donde $RelevanciaSemántica_x$, es el valor total de relevancia para el recurso x, obtenido mediante la aplicación de las TS y $Pos_{Semántico_x}$, es la posición de dicho recurso en el ranking generado por las TS.

Una vez que se obtienen estos valores, se calcula el puntaje de relevancia final correspondiente a cada nodo, mediante la utilización del MILS, que se expresa en la Ecuación (4.11).

$$RelevanciaRecurso_x = \frac{1}{(RelevanciaLexicográfica_x * k) + \left(RelevanciaSemántica_x * (1-k)\right)} \quad (4.11)$$

Donde la $RelevanciaRecurso_X$, es el valor total de relevancia para el recurso x que se obtiene mediante la integración de ambas técnicas (lexicográfica y semántica), $RelevanciaLexicográfica_x$ se obtiene mediante la Ecuación (4.9), $RelevanciaSemántica_x$ se obtiene mediante la Ecuación (4.10) y k es la constante de proporcionalidad de las TL, que indica el aporte al puntaje final efectuado por cada técnica. Mediante resultados experimentales, se determina el valor $k = 0.25$, lo que significa que las TL representan un 25% del puntaje obtenido por el MILS y las TS un 75%.

Un aspecto a tener en cuenta, es que debido a la naturaleza de la Ecuación (4.11), los recursos mejores posicionados por cada técnica, que por ende también serán los más relevantes, obtendrán un puntaje de $RelevanciaRecurso_x$ mínimo, mientras que los puntajes de los recursos que estén ubicados en peores posiciones, tenderán a infinito, lo que conlleva a ordenar de manera ascendente al ranking generado por el MILS.

Una vez obtenidos los resultados del MILS, se arma el ranking final de los 50 recursos más relevantes a ser presentados al usuario como se explicó en la sección 4.1. Luego este ranking se almacena en el “Repositorio Local”, para que el ***MRR*** tenga acceso a él.

CAPÍTULO 5
PRUEBAS Y RESULTADOS

En este capítulo se presentan las pruebas realizadas para evaluar si la utilización, tanto de las técnicas semánticas, como el modelo de integración léxico – semántico, representa alguna mejora, respecto a los resultados obtenidos por las técnicas lexicográficas ya puestas en producción. A tal fin, se presentan los aspectos y características de diseño de las pruebas realizadas (ajustes paramétricos efectuados, escenarios de pruebas considerados y métricas de evaluación utilizadas) y los resultados obtenidos de las simulaciones llevadas a cabo.

5.1 DISEÑO DE LAS PRUEBAS

El objetivo del presente capítulo es evaluar las técnicas consideradas, mediante la colaboración de expertos, sobre dos escenarios diferentes. De esta manera, se obtiene una directa validación y verificación de los resultados.

Para realizar estas evaluaciones sobre las Técnicas Semánticas (TS), el Modelo de Integración Léxico – Semántico (MILS) (descriptos en las secciones 4.2 y 4.3 respectivamente) y las Técnicas Lexicográficas (TL) ya implementadas (descriptas en la sección 2.4), se define el proceso de ejecución de las simulaciones mostrado en la Figura 5.1.

Una vez determinados los escenarios, se solicita a los expertos que definan sus requerimientos de información. Este proceso tiene la finalidad de establecer las características específicas sobre la información a recuperar y restricciones que se tendrán en cuenta.

Por cada requerimiento de información el experto define, con la ayuda del equipo de proyecto, la clave de búsqueda a ser utilizada, tanto para la recuperación de recursos como para la realización de los distintos análisis.

Esta clave de búsqueda, se construyen de la siguiente manera: inicialmente se establece el tema principal de la misma, para luego, adjuntar las características específicas expresadas por el experto, mediante los conectores “AND” (en el caso de que se trate de una característica que complemente al tema principal) y “OR” (en caso de que se trate de una característica alternativa). Finalmente, se concatenan las restricciones (si fue definida alguna) mediante el conector “NOT”.

Utilizando cada clave de búsqueda generada, se realiza la recuperación de los conjuntos de los 50 mejores recursos obtenidos tanto por las TL, como por las TS (ambos ordenados de acuerdo a la relevancia). Estos conjuntos son evaluados por el experto y sometidos a las pruebas de recuperación.

Además, utilizando las TS se ordenan los recursos obtenidos por las TL y viceversa. Esto permite realizar la comparación de la eficiencia del ordenamiento efectuado por cada técnica sobre un mismo conjunto de recursos, proporcionando así, otro punto de evaluación entre ambas.

Una vez realizadas estas evaluaciones, se genera por cada escenario, un conjunto único de recursos, producto de la combinación sin repetición de los resultados retornados por las TL y las TS. Este conjunto, se utiliza para evaluar el ordenamiento llevado a cabo por el MILS (presentado en la sección 4.3) mediante la comparación con el realizado tanto por las TL como por las TS. Esto permite, entre otras cosas, observar la contribución de cada técnica a la clasificación final y establecer los pros y contras de la utilización del MILS.

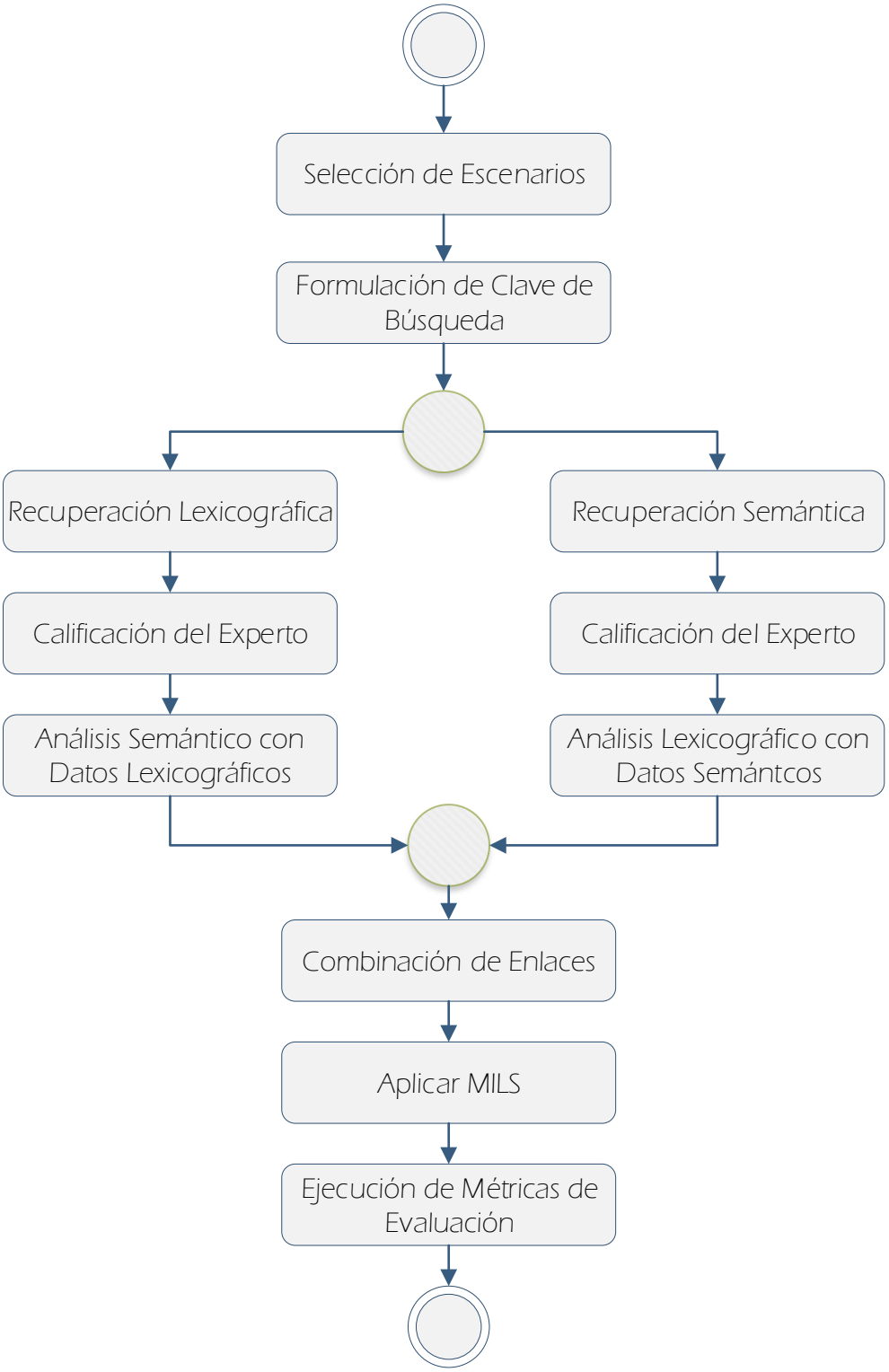

Figura 5.1 *- Diagrama de Actividad Para el Proceso de Prueba*

En todas estas pruebas, se emplean las métricas de precisión y exhaustividad de los primeros 1, 5, 10, 15, 20, 50 y de la totalidad de recursos recuperados, con el fin de observar la proporción de recursos relevantes que ocupan las primeras posiciones del ranking presentado al usuario y visualizar comportamientos destacables.

La precisión muestra la proporción de recursos relevantes existentes en la cantidad total de recursos recuperados o en un Top k determinado, lo que permite evaluar a una lista

de recursos de acuerdo a la utilidad de los mismos con respecto al criterio del experto. El valor de precisión para un Top k determinado, se obtiene mediante la Ecuación (5.1).

$$Precisión\ (k) = \frac{\#RecursosRelevantesRecuperados\ (k)}{k} \quad (5.1)$$

Donde $\#RecursosRelevantesRecuperados\ (k)$ es la cantidad de recursos relevantes presentes en el Top k y k es la cantidad de recursos recuperados considerados.

La exhaustividad muestra la proporción del conjunto total de recursos relevantes, recuperado en un momento determinado. Esta métrica permite estimar la rapidez con la que se recuperan todos los recursos relevantes e implícitamente el grado de dispersión de tales recursos a través de las distintas posiciones del ranking. Para calcular la exhaustividad en un Top k determinado, se utiliza la Ecuación (5.2).

$$Exhaustividad(k) = \frac{\#RecursosRelevantesRecuperados(k)}{\#TotalRecursosRelevantes} \quad (5.2)$$

Donde $\#RecursosRelevantesRecuperados(k)$ es la cantidad de recursos relevantes presentes en el Top k y $\#TotalRecursosRelevantes$ es la cantidad total de recursos relevantes recuperados.

A partir de los valores obtenidos por dichas métricas, se generan los gráficos de precisión y exhaustividad interpolada, que permiten observar la relación existente entre ambas, plasmando la variación en la precisión a medida que la exhaustividad aumenta.

5.1.1 Escenarios Considerados

A continuación, se presentan los escenarios considerados para la realización de las distintas pruebas. Como se expuso en la sección anterior, se cuenta con dos escenarios, siendo cada uno de ellos propuesto por un experto de un área temática determinada.

En las secciones siguientes, se exhiben detalles de la necesidad de información planteada por el experto y la clave de búsqueda resultante correspondiente a cada escenario.

Escenario 1: "Digital Storytelling"

El primer escenario considerado, está relacionado al ámbito de la educación digital, mediante la utilización de la técnica "*Digital Storytelling"*. Esta técnica consiste en el uso de herramientas audiovisuales que den soporte a la forma de enseñanza tradicional. Con respecto a esto, el experto pretende obtener información acerca de cómo articular esta técnica con las aulas digitales y la enseñanza tecnológica mediante el arte.

Al momento de definir la clave de búsqueda correspondiente a la necesidad de información, el experto especificó como tema principal, la técnica "*Storytelling"*.

En cuanto a los aspectos complementarios, expresó que se debían recuperar recursos relacionados a las aulas digitales ("*Digital Classroom")* y a las artes utilizadas en la enseñanza tecnológica ("*Art in Technology Education"*).

Además, aclaró que se deben evitar recursos relacionados a las enseñanzas del arte ("*Art Education"*). A partir de esto, se procede a definir la clave de búsqueda, la cual fue estructurada de la siguiente manera: "*Storytelling AND Digital Classroom AND Art in Technology Education NOT Art Education".*

Escenario 2: "Cookie Poisoning"

El segundo escenario considerado, responde a una necesidad de información correspondiente al área de la seguridad informática. En este caso, el experto expresó como

tema principal de búsqueda, las técnicas de ataques por envenenamiento de *cookie* (*Cookie Poisoning*).

Además, destacó como aspecto específico, la aplicación de estas técnicas a la vulneración de aplicaciones web (*Hacking Web Applications).* Más detalladamente, la búsqueda debe estar relacionada a las distintas técnicas existentes y en desarrollo, de *cookie poisoning* utilizables en la vulneración de aplicaciones web, con el fin de tomar medidas que permitan la protección ante este tipo de ataques maliciosos. A partir de lo especificado, se genera la clave de búsqueda a utilizar, la cual es: "*Cookie Poisoning AND Hacking Web Applications"*.

Como se puede observar, esta clave de búsqueda resulta menos restrictiva que la determinada en el escenario anterior, ya que se establece una sola característica que aparenta ser insuficiente para acotar el dominio de búsqueda y no se proporcionan restricciones que descarten posibles resultados no deseados. De todas maneras, se decide mantener esta estructura, con el fin de observar la incidencia de la definición de la clave en los resultados obtenidos.

5.1.2 Parámetros y preparación de los datos

Como se mencionó en la sección 2.4.1, las tres técnicas utilizadas para el análisis lexicográfico son: *Okapi BM-25,* Ranking de contribución (*C-Rank)* y *Vector Space Model (VSM)*. Para realizar las pruebas, se configuraron los parámetros presentados en la Tabla 5.1.

Tabla 5.1 *- Resumen de valores de parámetros para las TL*

Técnica	Parámetro	Valor
Okapi BM-25	b	0,75
	k	2
C-Rank	λ	0,8
VSM	No se definen parámetros	

En primer lugar, para la técnica Okapi BM-25 se definen dos constantes, $b = 0{,}75$ y $k = 2$, determinadas como óptimas a partir de resultados experimentales.

En cuanto a la técnica *C-Rank*, se establece el valor de la constante λ, que determina la contribución al puntaje final, del puntaje lexicográfico correspondiente al recurso analizado (λ) y la de los recursos relacionados a este ultimo $(1 - \lambda)$. Para ello se determina $\lambda = 0{,}8$, cuyo valor fue recomendado por Kim y colaboradores [40]. En cambio, para la técnica VSM no fue necesaria la definición de parámetros.

Por otro lado, para el MILS, se determina el valor de la constante k, que establece el porcentaje de aporte de las TL (k) y las TS $(1 - k)$ al puntaje final (ver Tabla 5.2). A partir de resultados experimentales, se considera como óptimo para las pruebas antes descriptas el valor de $k = 0{,}25$. Además, como se indica en la Tabla 5.2, para las TS no fue necesaria la definición de parámetros.

Tabla 5.2 *- Resumen de valores de parámetros para las TS y el MILS*

Técnica	Parámetro	Valor
TS	No se definen parámetros	
MILS	k	0,25

5.2 PRUEBA 1: ANÁLISIS DE RECUPERACIÓN

En esta sección, se presentan los resultados obtenidos mediante la evaluación de la eficacia en la recuperación de recursos relevantes realizada tanto a las TL, como a las TS.

Para ello, inicialmente se muestran los resultados obtenidos por cada técnica, resaltando aspectos destacables y posteriormente se presenta la comparación del comportamiento observado en ambos.

5.2.1 Escenario "*Digital Storytelling*"

Para la realización de estas pruebas, se llevó a cabo la recuperación de los cincuenta recursos más relevantes según el criterio cada técnica, mediante la utilización de la clave de búsqueda definida para el escenario "*Digital Storytelling*". Estos fueron evaluados por el experto correspondiente, el cual los clasificó como relevante o no relevante.

Con los recursos ya evaluados, se analizó la recuperación de las TL, cuyos resultados se resumen en la Tabla 5.3, presentando los valores de precisión y exhaustividad para los Top 1, 5, 10, 15, 20 y 50.

Tabla 5.3 *- Valores de Precisión (P) y Exhaustividad (E) para las TL*

	P	E
TOP 1	1	0,05
TOP 5	0,80	0,19
TOP 10	0,60	0,29
TOP 15	0,60	0,43
TOP 20	0,60	0,57
TOP 50	0,42	1

La Figura 5.2 muestra la relación entre la precisión y la exhaustividad, a partir de los valores expuestos en la Tabla 5.3.

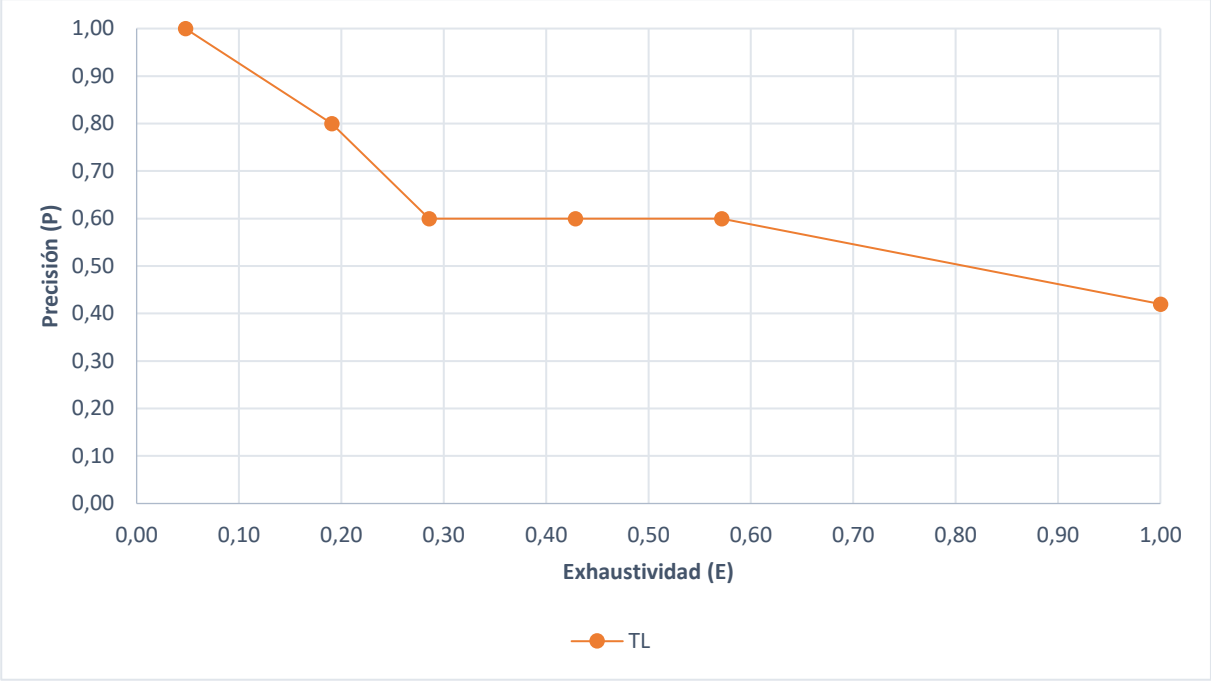

Figura 5.2 *- Gráfico de precisión y exhaustividad interpolada para las TL*

Como se observa, al recuperar el quinto recurso, se obtiene una precisión del 80%, indicando que, de los 5 primeros recursos presentados al usuario, 4 son relevantes. A partir de este punto, la precisión disminuye para estabilizarse en un valor de 60% en los Tops 10, 15 y 20, mostrando una mayor proporción de recursos relevantes en tales posiciones. Finalmente, este valor desciende al 42% al recuperar el quincuagésimo recurso, debido a un aumento en la proporción de los no relevantes, resultando en que 21 recursos de los 50 presentados, son relevantes.

En cuanto a los valores de exhaustividad, un aspecto a resaltar es que la mayor parte del conjunto de recursos relevantes, fue recuperada en las primeras 20 posiciones del ranking, alcanzando un valor de exhaustividad del 57%.

Para el mismo escenario, se analizó la recuperación de recursos efectuada por las TS, obteniendo los valores de precisión y exhaustividad que se presentan en la Tabla 5.4.

Tabla 5.4 *- Valores de Precisión (P) y Exhaustividad (E) para las TS.*

	P	E
TOP 1	1	0,04
TOP 5	0,80	0,17
TOP 10	0,90	0,39
TOP 15	0,80	0,52
TOP 20	0,70	0,61
TOP 50	0,46	1,00

A partir de estos resultados, se genera el gráfico de precisión y exhaustividad interpolada que se muestra en la Figura 5.3.

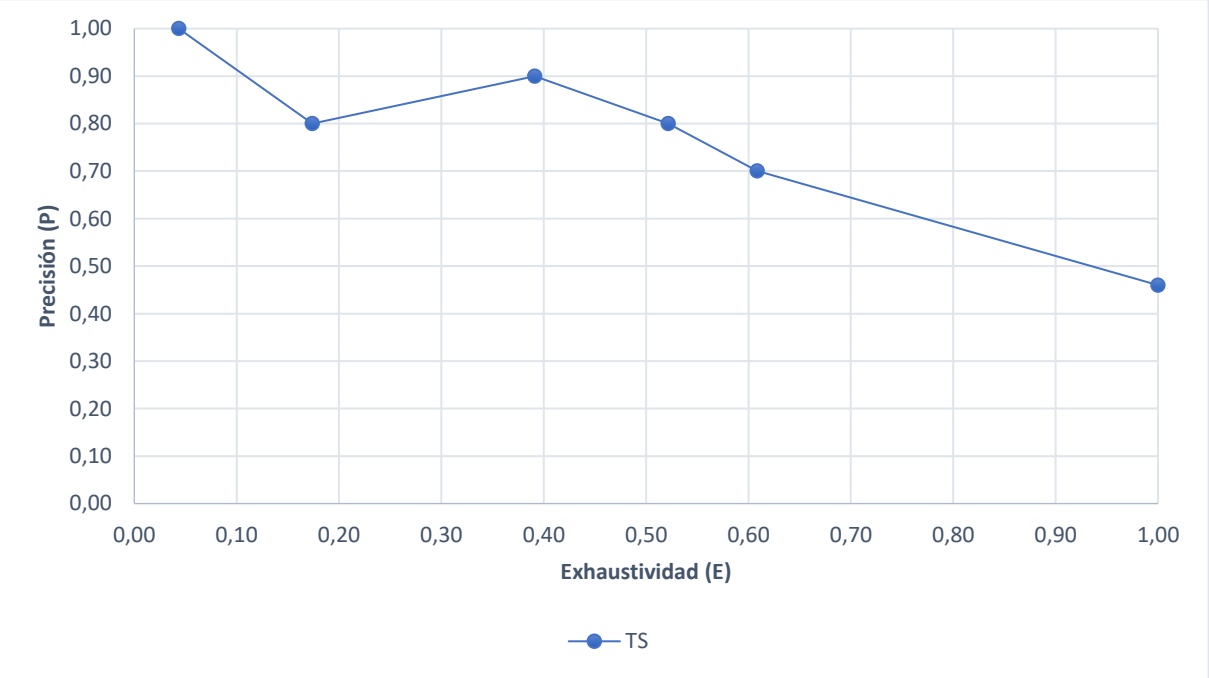

Figura 5.3 *- Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TS.*

Observando los resultados de las TS, es posible destacar varios puntos. En primera instancia, al recuperar el quinto recurso, se logra una precisión del 80%, indicando que solo 1 de los 5 recursos presentados al usuario es no relevante. Este valor se incrementa al 90% en el momento en que se recupera el décimo recurso, lo que plasma una mejora en la proporción de recursos relevantes recuperados.

Posterior a este punto, la precisión comienza a disminuir, logrando un 80% en el Top 15 y un 70% en el Top 20. Finalmente, al recuperar el último recurso, se obtiene una precisión 46% plasmando que de los 50 recursos recuperados 23 son relevantes.

Por parte de la exhaustividad, se aprecia que la mayor porción del conjunto de recursos relevantes se alcanza al recuperar el vigésimo recurso, logrando un valor del 61%.

Finalizadas las pruebas anteriores, se comparan los gráficos de precisión y exhaustividad interpolada obtenidos en ambos casos, con el fin de evidenciar la eficacia en la recuperación de recursos. Esto se presenta en la Figura 5.4.

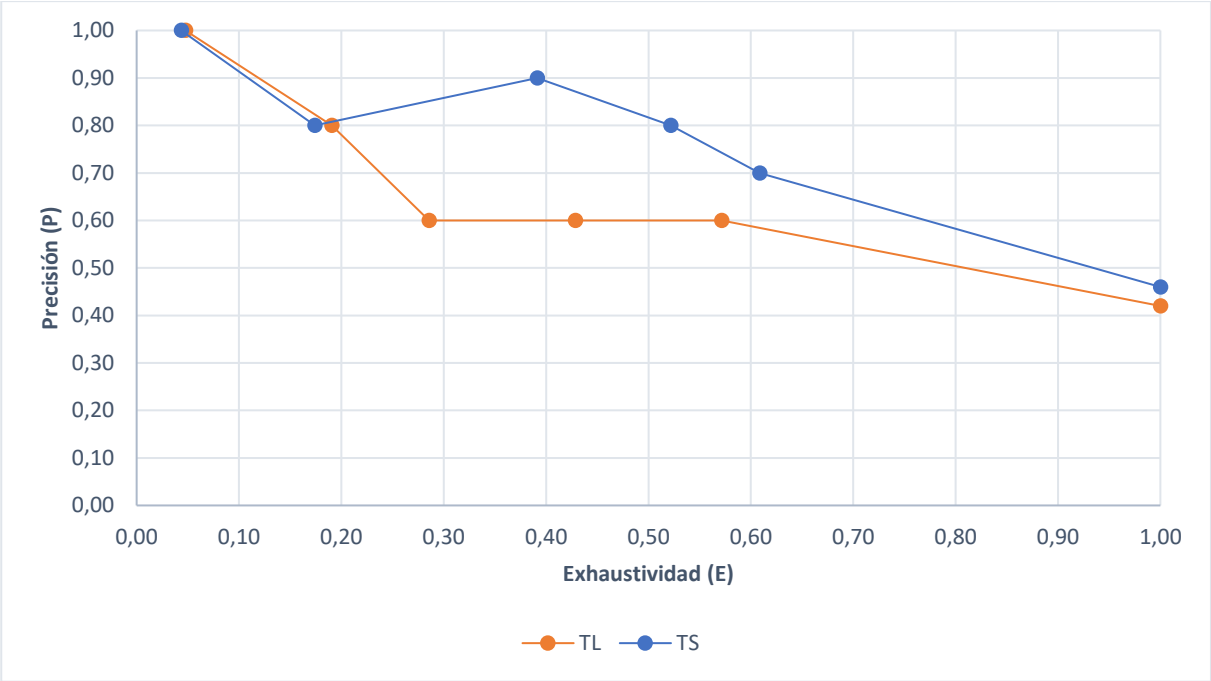

Figura 5.4 *- Comparación entre resultados obtenidos por las TL y las TS.*

A priori se observa una mejora en términos de precisión para las TS, donde se aprecia una diferencia a partir del quinto recurso recuperado.

La máxima diferencia se logra al alcanzar el Top 10, donde las TS obtienen una precisión del 90% con respecto al 60% de las TL.

Si se tiene en cuenta la totalidad de recursos presentados al usuario, las TS muestran una leve mejora obteniendo una precisión total del 46%, en comparación al 42% de las TL.

Con respecto a la exhaustividad, las TS muestran una significativa mejora en la mayor parte de las instancias analizadas (Tops 10, 15 y 20). La máxima diferencia se produjo al recuperar el décimo recurso, donde se alcanza un valor del 40%, con respecto al 29% de las TL. Esta diferencia se repite en menor medida, en los Tops 15 y 20.

5.2.2 Escenario "*Cookie Poisoning*"

De igual manera, para las pruebas que se realizaron sobre este escenario, se recuperaron los cincuenta mejores recursos por cada técnica, los cuales fueron posteriormente evaluados por el experto que presentó esta necesidad de información.

Los resultados obtenidos por las TL se presentan en la Tabla 5.5, en la que se resumen los valores de precisión y exhaustividad para los Tops 1, 5, 10, 15, 20 y 50.

Tabla 5.5 *- Valores de Precisión (P) y Exhaustividad (E) para las TL.*

	P	E
TOP 1	1	0.20
TOP 5	0.40	0.40
TOP 10	0.30	0.60
TOP 15	0.27	0.80
TOP 20	0.25	1
TOP 50	0.1	1

Con estos resultados, se confecciona el gráfico de precisión y exhaustividad interpolada que se presenta en la Figura 5.5.

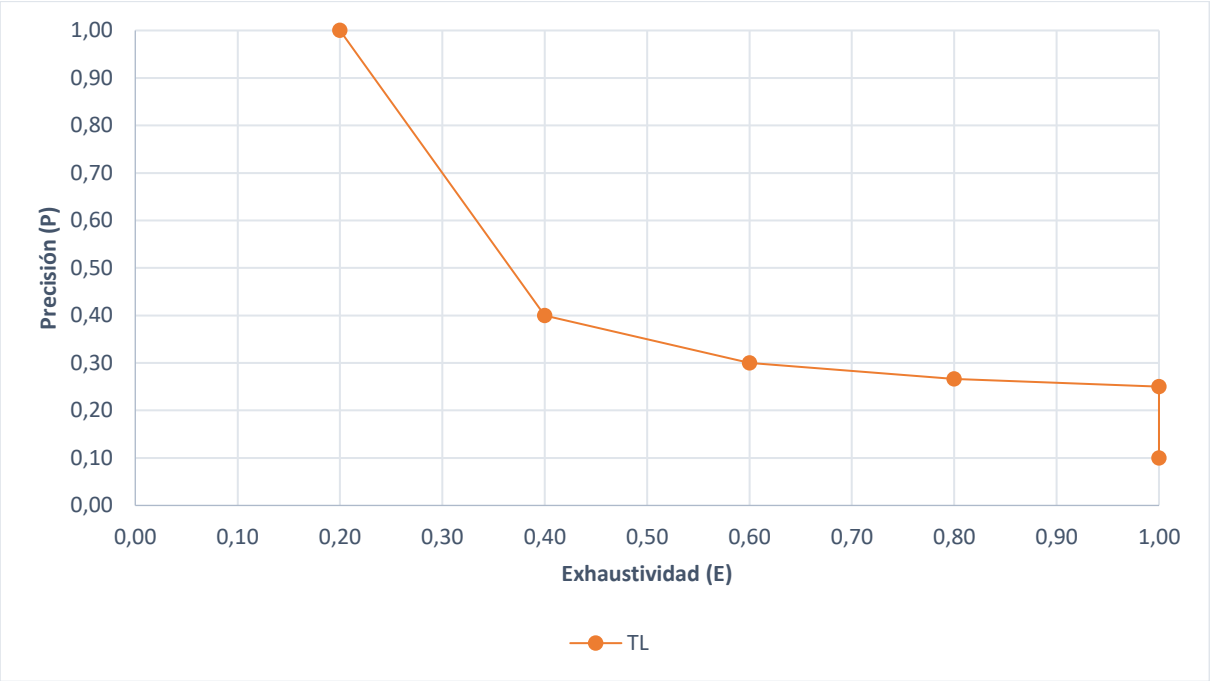

Figura 5.5 *- Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TL.*

En el grafico anterior, se puede observar un marcado descenso de la precisión al recuperar el quinto recurso, indicando que la cantidad de recursos relevantes en tales posiciones, es mínima. Esto se debe a la dispersión de los recursos relevantes a través de las distintas posiciones del ranking generado por las TL, lo cual depende de la eficacia en el ordenamiento de estas técnicas.

Por otro lado, el máximo valor de exhaustividad, se alcanza al momento de recuperar el vigésimo recurso, indicando que, en este punto, todos los recursos relevantes fueron recuperados. En dicho punto, la precisión es del 25%, lo que plasma que, de 20 recursos recuperados, solo 5 son relevantes. Finalmente, al recuperar el quincuagésimo recurso, la precisión disminuye al 10%, indicando que solo 5 recursos de 50, son relevantes.

Del mismo modo, se realizaron las pruebas para las TS, evaluando precisión y exhaustividad para los primeros 1, 5, 10, 15, 20 y 50 recursos recuperados. Estos valores se presentan en la Tabla 5.6.

Tabla 5.6 *- Valores de Precisión (P) y Exhaustividad (E) para las TS.*

	P	E
TOP 1	1	0,13
TOP 5	0,80	0,50
TOP 10	0,50	0,63
TOP 15	0,40	0,75
TOP 20	0,35	0,88
TOP 50	0,16	1

A partir de estos resultados, se muestra en la Figura 5.6, el gráfico de precisión y exhaustividad interpolada.

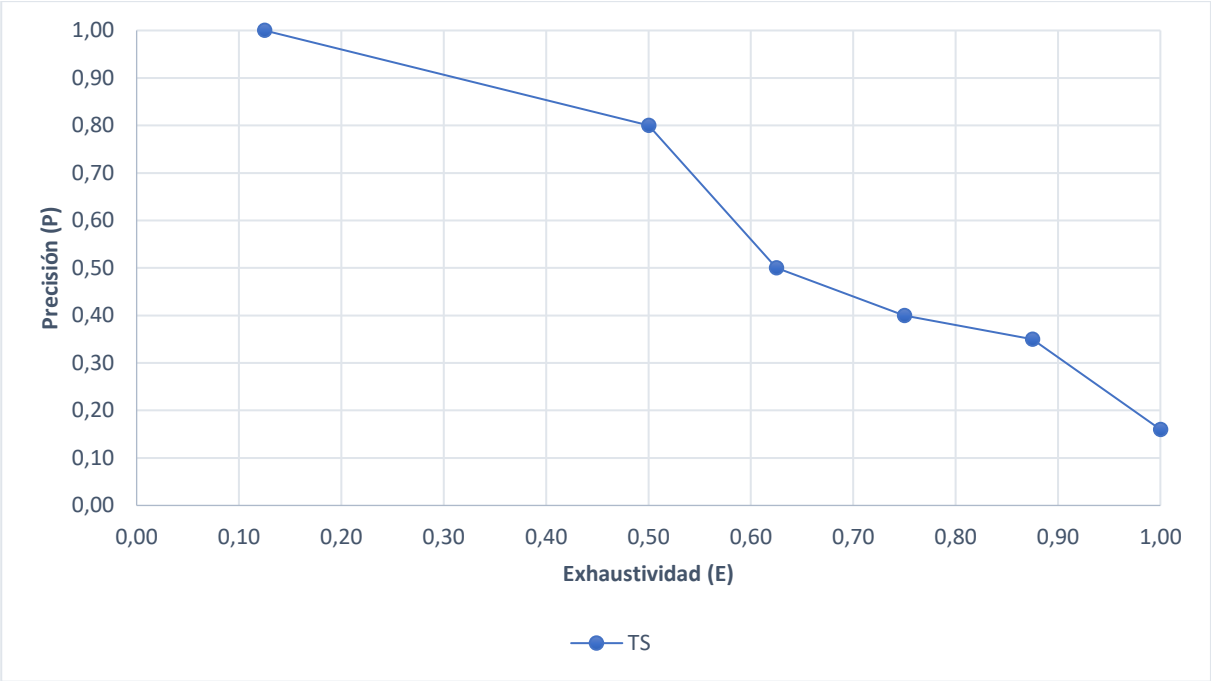

Figura 5.6 *- Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TS*

En este caso, al recuperar el quinto recurso se obtiene una precisión del 80%, lo que indica que 4 de 5 recursos, son relevantes. Otro aspecto a destacar es que, en este punto, se alcanza una exhaustividad del 50%, lo que significa, que se logra recuperar la mitad del conjunto de recursos relevantes.

El descenso observado en términos de precisión se debe a que la segunda mitad de recursos relevantes, se encuentra dispersa en las posiciones inferiores al Top 5, lo cual cobra sentido, al observar que el máximo valor de exhaustividad se alcanza al recuperar el quincuagésimo recurso.

Además, se observa que la precisión final es del 16%, reflejando la baja recuperación de recursos relevantes en este escenario.

Finalmente, mediante los resultados de las ejecuciones anteriores, se lleva a cabo la comparación entre los gráficos de precisión y exhaustividad interpolada obtenidos. Esta comparación se presenta en la Figura 5.7.

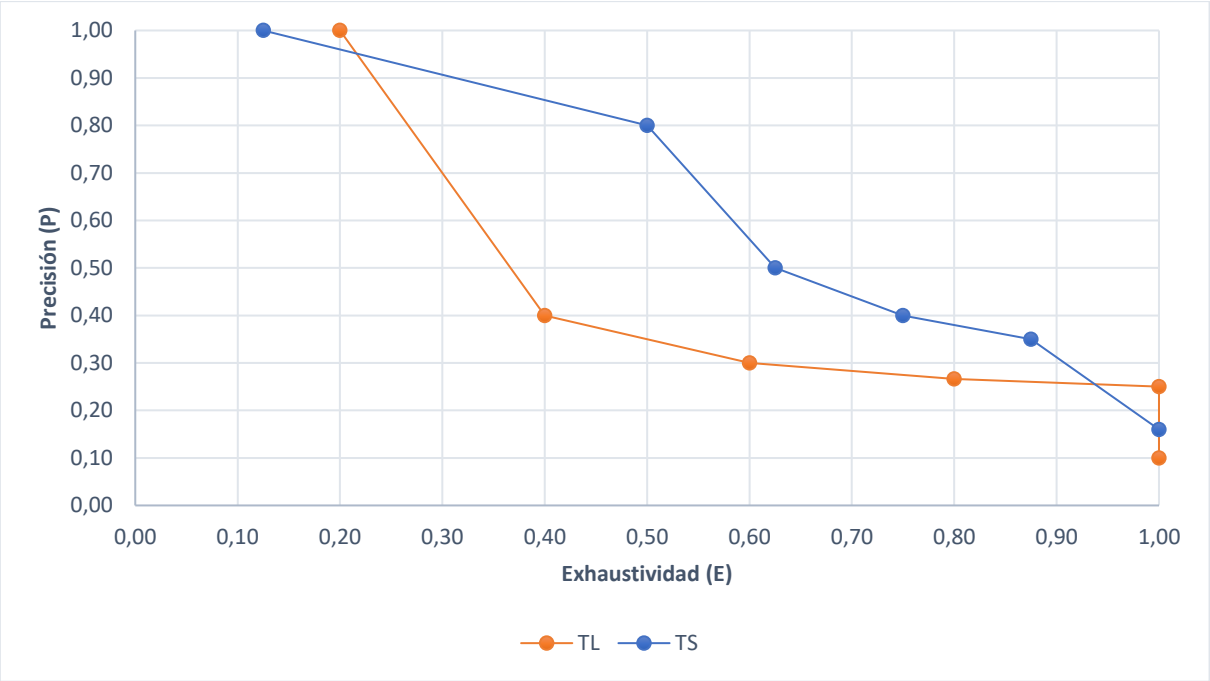

Figura 5.7 *- Comparación entre resultados obtenidos por las TL y las TS.*

En principio, se puede observar una mejora por parte de las TS, la cual se hace visible al momento de recuperar el quinto recurso y se extiende a lo largo de toda la recuperación.

El punto en el que se produce la mayor diferencia de precisión entre ambas es el Top 5, arrojando una precisión del 80% para las TS, con respecto al 40% de las TL.

Si bien, la cantidad total de recursos relevantes recuperados por ambas técnicas es mínima, en el caso de las TS se observa un mayor agrupamiento de estos en las primeras posiciones, comparado al caso de las TL, en el que se aprecia una mayor dispersión.

Además, cabe destacar, que la proporción total de recursos relevantes, es mayor para las TS logrando un 16% de precisión, comparado con las TL, que obtienen el 10%.

5.3 PRUEBA 2: ANÁLISIS DE ORDENAMIENTO

En la presente sección, se lleva a cabo la ejecución de las pruebas de ordenamiento, que permiten evaluar la eficiencia con la que ambas técnicas elaboran sus rankings. Además, teniendo en cuenta que la relevancia es el criterio a partir del cual se generan dichos rankings, implícitamente se evalúa que tan acertada es la determinación de la misma por cada técnica.

Para ello, por cada escenario, se comparó el ordenamiento realizado por ambas técnicas sobre un mismo conjunto de recursos, es decir, el ordenamiento de ambas sobre el conjunto de recursos recuperados por las TL y por las TS.

5.3.1 Escenario "*Digital Storytelling*"

Para realizar las pruebas de ordenamiento sobre el escenario "*Digital Storytelling*", inicialmente se analizó el ordenamiento de ambas técnicas sobre el conjunto de recursos recuperados por las TL. Luego, esto se realizó sobre el conjunto recuperado por las TS, permitiendo en ambos casos, determinar la precisión y exhaustividad de los rankings generados considerando los Tops 1, 5, 10, 15, 20 y 50.

Los valores de precisión y exhaustividad, para el ordenamiento realizado por las TL y las TS sobre el conjunto de recursos recuperados por las TL, se presentan en la Tabla 5.7.

Tabla 5.7 *– Precisión (P) y Exhaustividad (E) para TL y TS, utilizando la lista de recursos recuperada por las TL.*

	TL		TS	
	P	**E**	**P**	**E**
TOP 1	1	0,05	0,0	0,0
TOP 5	0,80	0,19	0,80	0,19
TOP 10	0,60	0,29	0,70	0,33
TOP 15	0,60	0,43	0,60	0,43
TOP 20	0,60	0,57	0,60	0,57
TOP 50	0,42	1	0,42	1

A partir de estos resultados, se genera el gráfico de precisión y exhaustividad interpolada para las dos técnicas, el cual se presenta en la Figura 5.8.

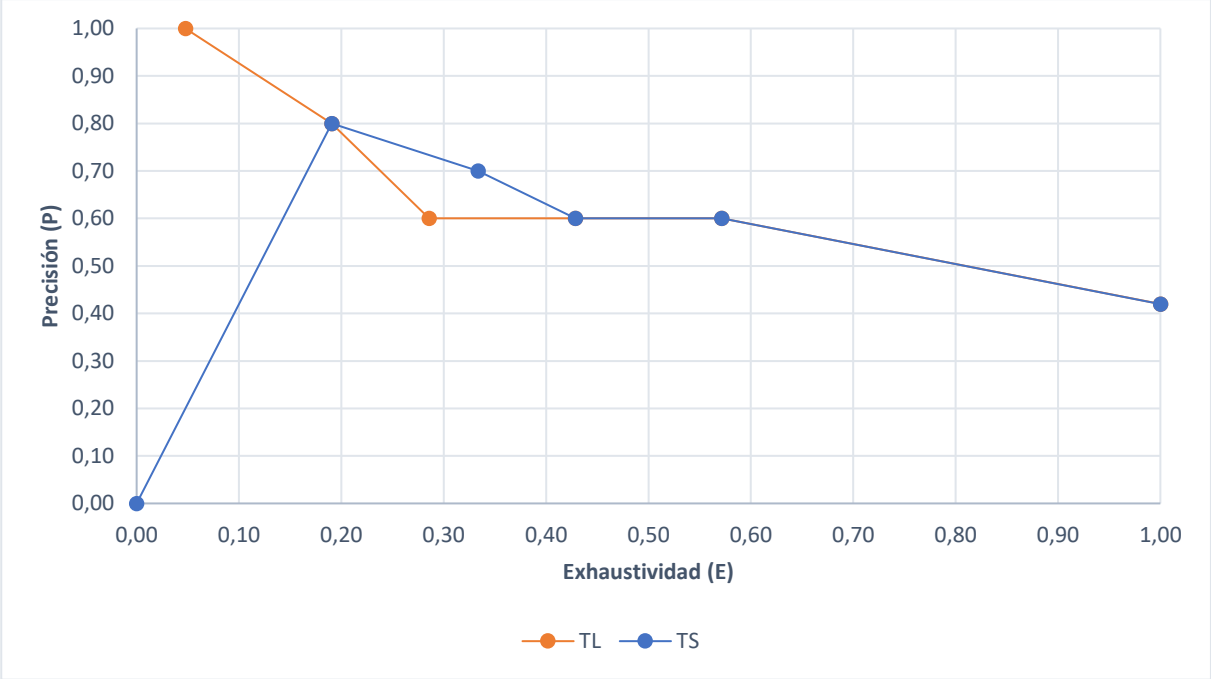

Figura 5.8 *- Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TL y las TS, utilizando la lista de recursos recuperada por las TL.*

Como se observa, la precisión y exhaustividad de las TS en el Top 1 es de 0, lo que indica que la primera posición del ranking generado por esta técnica, la ocupa un recurso no relevante.

Por otro lado, existe una diferencia al momento de obtener el décimo recurso, donde la precisión de las TS aumenta al 70% comparado al 60% obtenido por las TL. Sin embargo, en términos generales, el comportamiento fue similar en ambas técnicas.

Finalizadas estas pruebas, se procede a la evaluación de ambas técnicas, con respecto al conjunto recuperado por las TS, donde los valores de precisión y exhaustividad obtenidos se presentan en la Tabla 5.8.

Tabla 5.8 *– Precisión (P) y Exhaustividad (E) para TL y TS, utilizando la lista de recursos recuperados por las TS.*

	TL		TS	
	P	**E**	**P**	**E**
TOP 1	1	0,04	1	0,04
TOP 5	0,80	0,17	0,80	0,17
TOP 10	0,70	0,30	0,90	0,39
TOP 15	0,80	0,52	0,80	0,52
TOP 20	0,70	0,61	0,70	0,61
TOP 50	0,46	1	0,46	1

A partir de estos valores, se presenta en la Figura 5.9, el gráfico de precisión y exhaustividad interpolada, con el fin de apreciar que técnica es más efectiva en la generación de rankings.

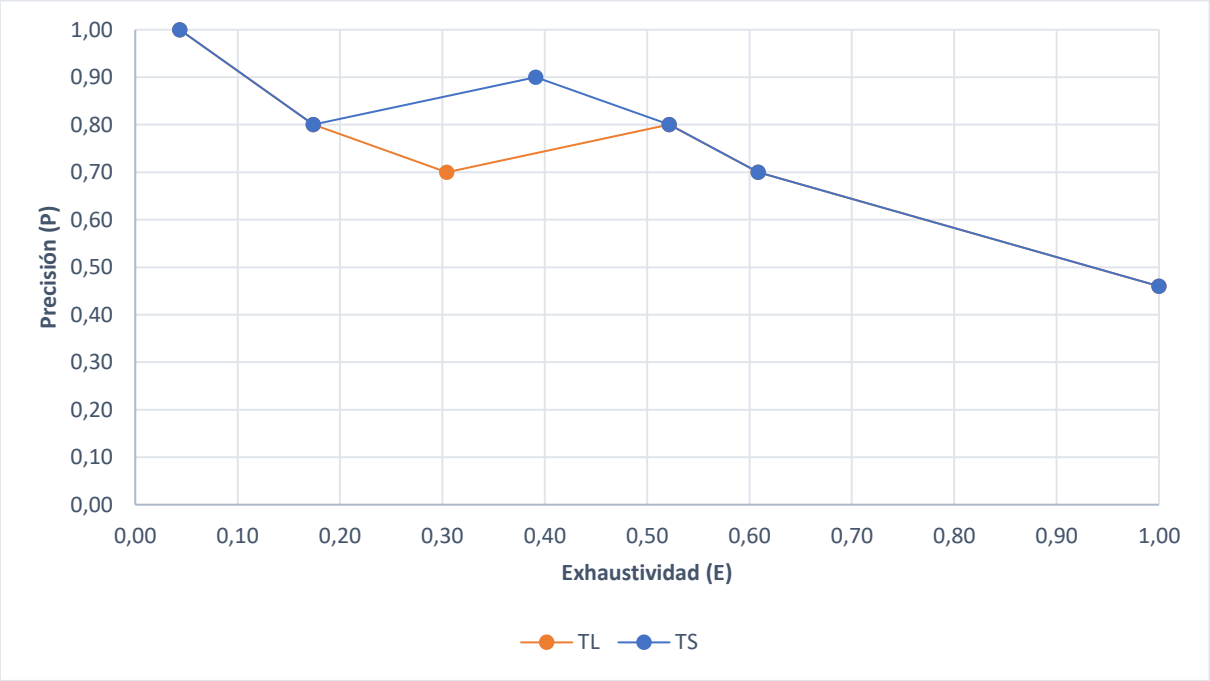

Figura 5.9 *- Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TL y las TS, utilizando la lista de recursos recuperados por las TS.*

En este caso, el comportamiento de ambas técnicas es similar, exceptuando la décima posición, donde las TS obtienen una precisión del 90% con respecto al 70% de las TL.

Además, en ambas pruebas las diferencias se producen en las primeras diez posiciones, lo que plasma una mejora en cuanto a los primeros resultados con los que el usuario tendrá contacto.

Por otro lado, el comportamiento similar responde a que, para estas pruebas, la cantidad de recursos relevantes no varía de acuerdo a las técnicas utilizadas para analizarlos, lo que explica la presencia de pequeñas diferencias en el ordenamiento de recursos y por ende pequeñas variaciones en la precisión y exhaustividad.

5.3.2 Escenario "*Cookie Poisoning*"

En esta sección se presentan las pruebas de ordenamiento, realizadas sobre el escenario "*Cookie Poisoning"*. En primer lugar, se comienza por evaluar el ordenamiento realizado por ambas técnicas, sobre el conjunto de recursos recuperado por las TL. Los valores de precisión y exhaustividad obtenidos, se resumen en la Tabla 5.9.

Tabla 5.9 *– Precisión (P) y Exhaustividad (E) para TL y TS, utilizando la lista de recursos recuperada por las TL.*

	TL		TS	
	P	E	P	E
TOP 1	1	0,2	1	0,2
TOP 5	0,40	0,40	0,60	0,60
TOP 10	0,30	0,60	0,40	0,80
TOP 15	0,27	0,80	0,33	1
TOP 20	0,25	1	0,25	1
TOP 50	0,10	1	0,10	1

A partir de estos resultados, se confecciona el gráfico de precisión y exhaustividad interpolada, que se presenta en la Figura 5.10.

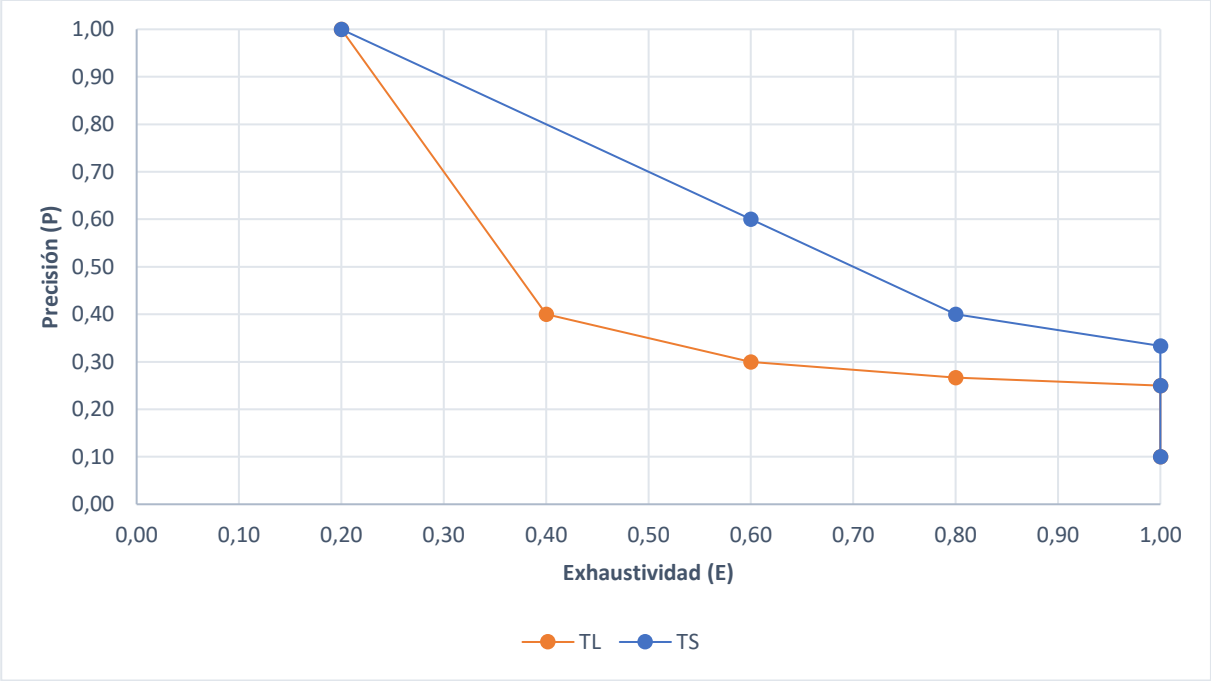

Figura 5.10 *- Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TL y las TS, utilizando la lista de recursos recuperada por las TL.*

Un aspecto destacable es la mejora de las TS por sobre las TL, lo que indica un ordenamiento de recursos más certero, por parte de esta técnica.

Esto se ratifica al observar los valores de exhaustividad, donde las TS logran recuperar todos los recursos relevantes en el Top 15 y las TL en el Top 20.

De igual manera, se realizan las pruebas de ordenamiento, utilizando el conjunto de recursos recuperado por las TS, donde los valores de precisión y exhaustividad obtenidos se presentan en la Tabla 5.10.

Tabla 5.10 *– Precisión (P) y Exhaustividad (E) para TL y TS, utilizando la lista de recursos recuperada por las TS.*

	TL		TS	
	P	**E**	**P**	**E**
TOP 1	1	0,13	1	0,13
TOP 5	0,60	0,38	0,80	0,50
TOP 10	0,40	0,50	0,50	0,63
TOP 15	0,33	0,63	0,40	0,75
TOP 20	0,25	0,63	0,35	0,88
TOP 50	0,16	1	0,16	1

Seguidamente, se presenta en la Figura 5.11, el gráfico de precisión y exhaustividad interpolada generado a partir de los resultados obtenidos.

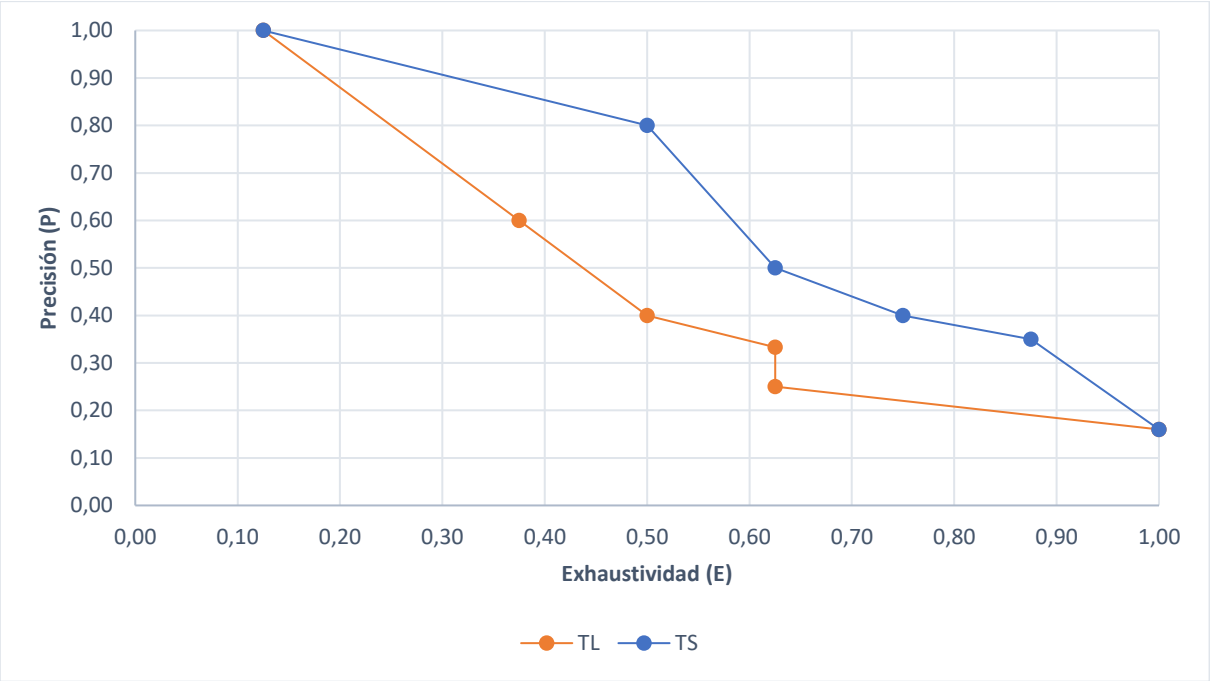

Figura 5.11 *- Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TL y las TS, utilizando la lista de recursos recuperada por las TS*

En estas pruebas se hace evidente la mejora por parte de las TS, la cual en todo momento es superior, tanto en términos de precisión como en exhaustividad. Esto indica un mejor ordenamiento de los recursos por parte de las TS y, por ende, una apreciación acertada del criterio de relevancia.

5.4 EVALUACIÓN DEL MODELO DE INTEGRACIÓN LÉXICO – SEMÁNTICO

Con el objetivo de poder obtener un puntaje final, que contemple los factores considerados por las TL y las TS, se definió un modelo que permite obtener un valor de relevancia para los recursos analizados, unificando los puntajes obtenidos por cada técnica. Para esto, se plantea el Modelo de Integración Léxico - Semántico (MILS), presentado en la sección 4.3, el cual propone calcular dicho puntaje mediante la utilización de la Ecuación (5.3).

$$RelevanciaRecurso_x = \frac{1}{(RelevanciaLexicográfica_x * k) + \left(RelevanciaSemántica_x * (1-k)\right)} \quad (5.3)$$

Para evidenciar la eficiencia de ordenamiento del MILS y, por lo tanto, las ventajas y desventajas de su utilización, en las secciones siguientes se realizan pruebas de ordenamiento, en las cuales se contrastan, por cada escenario considerado, los rankings generados por las TL, las TS y el MILS, mediante la utilización de un conjunto de recursos obtenido a partir de la combinación sin repetición de las recuperaciones llevadas a cabo por las TL y las TS.

5.4.1 Escenario "*Digital Storytelling*"

Mediante la combinación de las recuperaciones de las TL y las TS, se obtuvo para el primer escenario, un conjunto único compuesto por 92 recursos que no se repiten. A partir de dicho conjunto, se calculan los valores de precisión y exhaustividad para los Top 1, 5, 10, 15, 20, 50 y 92, correspondientes a ambas técnicas y al MILS. Estos valores se presentan en la Tabla 5.11.

***Tabla 5.11** – Precisión (P) y Exhaustividad (E) para TL, TS y el MILS, utilizando la lista de recursos recuperados por TL y las TS.*

	TL		TS		MILS	
	P	E	P	E	P	E
TOP 1	1	0,03	1	0,03	1	0,03
TOP 5	0.80	0.11	0.80	0.11	0.80	0.11
TOP 10	0.70	0.18	0.90	0.24	0.90	0.24
TOP 15	0.60	0.24	0.80	0.32	0.87	0.34
TOP 20	0.55	0.29	0.65	0.34	0.75	0.39
TOP 50	0.54	0.71	0.50	0.66	0.50	0.66
TOP 92	0.41	1	0.41	1	0.41	1

A partir de estos resultados, se confecciona el gráfico de precisión y exhaustividad interpolada, el cual se muestra en la Figura 5.12.

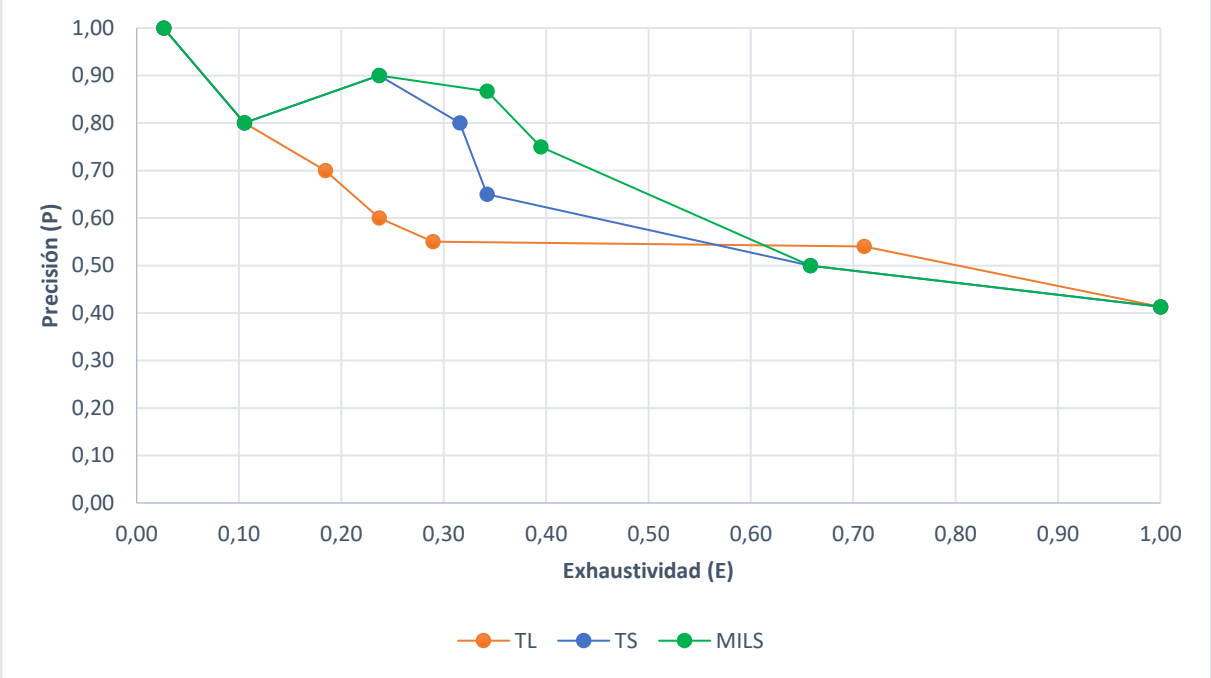

***Figura 5.12** - Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TL, las TS y el MILS*

En este caso, es evidente la superioridad tanto del MILS, como de las TS, en términos de precisión y exhaustividad. El comportamiento de estos dos, es similar hasta la décima posición, lo que indica una influencia directa de las TS sobre la precisión del MILS.

A partir del Top 10, hasta el vigésimo recurso, la clasificación realizada por el MILS demuestra una mayor precisión y por ende una mayor exhaustividad, que las otras dos técnicas, lo que permite aseverar que se produjo una mejora en la clasificación, producto de la combinación de factores considerados.

En el Top 50, las TL muestran una mejor precisión, indicando que posee una mayor cantidad de recursos relevantes distribuidas entre la posición 20 y la 50. Además, estas técnicas obtienen un mejor valor de exhaustividad en el Top 50, lo que sugiere que las TS y el MILS tienen una mayor cantidad de recursos relevantes ubicados en posiciones inferiores a la 50.

5.4.2 Escenario "*Cookie Poisoning*"

Finalmente, se realiza la comparación de los rankings generados por las TL, las TS y el MILS, para el escenario "*Cookie Poisoning"*, presentando los valores de precisión y exhaustividad obtenidos en cada caso, en la Tabla 5.12.

***Tabla 5.12** – Precisión (P) y Exhaustividad (E) para TL, TS y el MILS, utilizando la lista de recursos recuperados por TL y las TS.*

	TL		TS		MILS	
	P	E	P	E	P	E
TOP 1	1	0,08	1	0,08	1	0,08
TOP 5	0.40	0.17	0.80	0.33	0.80	0.33
TOP 10	0.30	0.25	0.60	0.50	0.60	0.50
TOP 15	0.20	0.25	0.47	0.58	0.40	0.50
TOP 20	0.25	0.42	0.40	0.67	0.40	0.67
TOP 50	0.18	0.75	0.20	0.83	0.20	0.83
TOP 93	0.13	1	0.13	1	0.13	1

A partir de estos resultados, se presenta en la Figura 5.13, el gráfico de precisión y exhaustividad interpolada, que muestra el comportamiento de estas técnicas en el Top 1, 5, 10, 15, 20, 50 y 93.

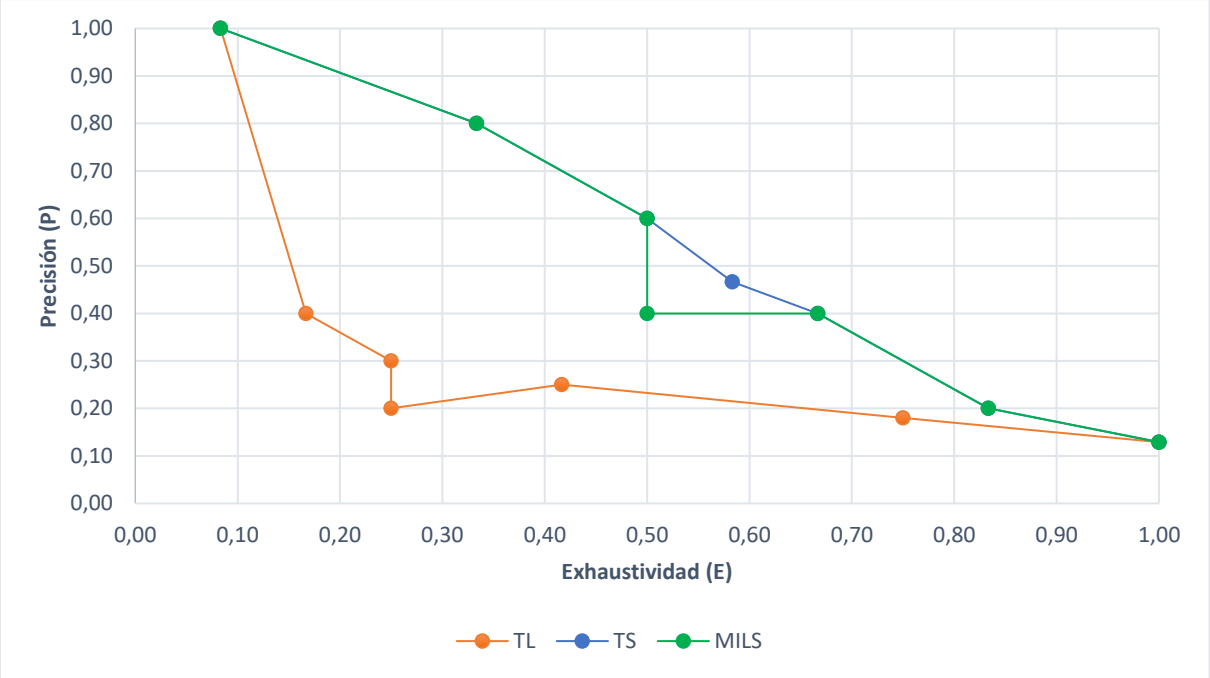

***Figura 5.13** - Gráfico de Precisión (P) y Exhaustividad (E) interpolada para las TL, las TS y el MILS*

En general, los resultados de las TS y el MILS (las cuales tienen un comportamiento similar), son superiores a los obtenidos por las TL.

El aspecto destacable se produce en la decimoquinta posición, donde el MILS disminuye en precisión, de manera similar a las TL, lo que indica que existe una influencia por parte de estas técnicas sobre los resultados del MILS. Finalmente, el comportamiento a partir del Top 20, vuelve a ser coincidente con el de las TS.

5.4.3 Consideraciones Destacadas

Durante el desarrollo de estas pruebas, se persiguió el objetivo de analizar las técnicas propuestas desde distintos puntos de vista. Esto permitió determinar aspectos relacionados a la recuperación llevada a cabo por cada técnica, la generación de rankings y como se combinan los criterios de las TL y las TS para obtener el puntaje a partir del MILS.

Las pruebas de recuperación muestran la efectividad con la que tanto las TL como las TS, obtienen recursos relevantes desde internet, descartando los no relevantes.

El objetivo de evaluar este aspecto es verificar que tan efectiva es la determinación de la relevancia sobre el conjunto infinito de recursos existentes en la web, lo cual incide en la cantidad de recursos relevantes presentados al usuario.

Las pruebas de ordenamiento, por otro lado, se centran en la capacidad de cada técnica de ubicar mejor a los recursos relevantes en los rankings que construyen. En una situación ideal, los recursos más relevantes se ubican en las primeras posiciones del ranking. Sin embargo, esto no siempre es así en la práctica, debido a la complejidad inherente a la determinación de la relevancia.

Por último, se llevaron a cabo las pruebas realizadas sobre el MILS, que permitieron observar los resultados del ordenamiento de recursos a partir de la combinación de los criterios considerados por las TL y las TS, y comprobar como la consideración de distintos enfoques, puede contribuir a la correcta estimación de la relevancia.

En definitiva, las pruebas realizadas proveyeron de elementos de análisis comparativos entre todas las técnicas, considerando distintas perspectivas que aparentan ser suficientes como para observar y destacar aspectos relacionados al comportamiento de cada una de ellas.

CAPÍTULO 6
CONCLUSIONES Y TRABAJOS FUTUROS

En este capítulo, se presentan las conclusiones obtenidas a partir de la implementación y las pruebas realizadas sobre las técnicas semánticas y el modelo de integración léxico semántico. Asimismo, se proponen líneas de investigación sobre las que se podría avanzar en trabajos futuros.

6.1 CONCLUSIONES

En este Trabajo Final de Carrera, se describe el Modelo de Integración Léxico – Semántico (MILS) para la clasificación de recursos Web utilizando Técnicas Lexicográficas (TL) y Técnicas Semánticas (TS) como medio de determinación de la relevancia. El modelo desarrollado combina los criterios evaluados por cada técnica, para establecer la correspondencia existente entre el contenido Web y las necesidades de información planteada por el usuario y expresadas en forma de claves de búsqueda.

Durante el desarrollo del MILS se ha trabajado en cuatro etapas:

Etapa 1 – Definición de las herramientas a utilizar. Luego del análisis de varias alternativas, se seleccionaron las ontologías WordNet y ConceptNet para contextualizar palabras. Estas representan de manera simple e intuitiva los distintos sentidos de tales palabras, junto a sus relaciones, facilitando la aplicación de distintas operaciones sobre ellas. Además, se estableció la métrica de relación y similitud semántica de pares de palabras a emplear, en base a la comparación entre diez métricas distintas con respecto al criterio de distintos expertos. Esto permitió tener una perspectiva clara del estado del arte respecto de los algoritmos más utilizados para el análisis semántico.

Etapa 2 – Implementación de las TS. En base a la métrica seleccionada, se diseñaron las estrategias utilizadas por las TS para generar los valores que permiten la evaluación de la relevancia de los recursos. Es importante destacar que, de manera complementaria con la evaluación de las TL, se obtienen dos puntajes por cada recurso, lo que permite tener un punto de vista más completo en el proceso de evaluación.

Etapa 3 – Construcción del modelo. Teniendo en cuenta el concepto de integración de las TL y las TS se diseñó e implementó el MILS. Para ello, se consideraron los criterios provistos por ambas técnicas, resumiendo la relevancia de los recursos en un solo valor. De esta manera se logró desarrollar un proceso para establecer la importancia de cada recurso de manera más precisa y completa.

Etapa 4 - Validación del modelo. Mediante la realización de diferentes tipos de pruebas se evaluó el rendimiento del MILS. Para ello, se comparó la eficiencia en la recuperación de recursos contra las TL y TS en dos escenarios distintos. Además, se analizaron las diferencias entre los rankings generados por las TL, TS y el MILS, tomando como base las medidas de precisión y exhaustividad.

A partir de los resultados obtenidos, se observan ciertos aspectos destacables del MILS que indican la conveniencia de su utilización. En primer lugar, los resultados muestran una mejora con respecto a la utilización de las TL y la TS, producto de una acertada combinación de criterios considerados por cada técnica. En segundo lugar, se puede apreciar que, si bien en algunos casos las TL puede disminuir el rendimiento del MILS, la clasificación correcta de las TS corrige la desviación, mejorando de esta manera la clasificación realizada por las TL. En definitiva, la combinación de ambas técnicas mediante el MILS representa un enfoque apropiado para la clasificación de recursos.

En cuanto a los valores de precisión obtenidos por el MILS, se observa que se logran resultados mejores a los obtenidos por las TL y las TS. Esto ocurre cuando las primeras posiciones del ranking generado por las TL (que poseen menor aporte al puntaje final), están ocupadas en su mayoría por recursos relevantes, lo que provoca, a partir de la combinación con el ranking de las TS, un aumento en la proporción de recursos relevantes ubicados en las primeras posiciones del ranking generado por el MILS. El aumento de la precisión, se relaciona con la mayor rapidez de recuperación de todos los recursos relevantes presentes en el conjunto total, lo que implica un aumento en la tasa de crecimiento de la exhaustividad del MILS.

Adicionalmente, al comparar los resultados de los análisis léxico y semántico, se aprecia una mejora de las TS respecto a las TL tanto en la recuperación de recursos como así también en la evaluación de la capacidad de ordenamiento. Esto se debe a que las TS consideran una mayor cantidad de aspectos a la hora de realizar el análisis de los recursos (relaciones y similitudes semánticas existentes), mientras que las TL llevan a cabo un análisis más restrictivo, exigiendo exclusivamente, la aparición de términos pertenecientes a la clave de búsqueda, lo que provoca la penalización de los recursos cuyo contenido no esté compuesto por tales términos. Esto indica que las TS son métodos adecuados para estimar con mayor precisión la relevancia de los recursos.

Una consideración importante a tener en cuenta es que, más allá de las técnicas a utilizar, se deben definir las claves de búsqueda de manera tal que precisen de forma correcta el dominio de trabajo. Esta consideración se ve reflejada en los resultados obtenidos para los dos escenarios utilizados, ya que, para el escenario con mayor definición de las claves, *Digital Storytelling*, se recuperó una mayor cantidad de recursos relevantes que para el escenario *Cookie Poisoning*. Claramente la definición de características complementarias al tema principal y restricciones que permiten descartar resultados no deseados, focaliza el ámbito de búsqueda y tiene incidencia directa en la proporción de recursos relevantes recuperados.

Otra cuestión importante, es la incidencia de los métodos denominados “Crawlers Web”, en el correcto análisis de los recursos. Estos métodos, tienen por objetivo, la recuperación del contenido principal de los recursos, lo cual representa una tarea compleja, debido a la gran diversidad de tecnologías y estructuras utilizadas en la construcción de páginas Web. Una mala recuperación por parte del “Crawler Web” se traduce en el análisis de contenido no deseado y, por ende, en la introducción de ruido en el puntaje final de relevancia.

Finalmente, cabe destacar que los resultados de este TFC, fueron publicados y expuestos en el 6to Congreso Nacional de Ingeniería Informática - Sistemas de Información (CoNaIISI 2018), bajo el título “Modelo de Análisis Semántico de Documentos Web” [84] (Anexo I).

6.2 TRABAJOS FUTUROS

Es evidente que los sistemas de recuperación de información, están atravesando su momento de máximo apogeo. Las nuevas técnicas afloran constantemente, y se observa un gran esfuerzo por obtener mejores resultados. En este sentido, durante el desarrollo de este Trabajo Final de Carrera, se ha identificado un conjunto de temas o líneas interesantes de abordar que convienen ser mencionadas.

El primer tema está relacionado a la necesidad de identificar técnicas más eficientes de desambiguación del sentido de la palabra (WSD – *Word Sense Disambiguation*), debido a la baja precisión de las utilizadas actualmente. En este contexto, algunas alternativas factibles

de ser utilizadas para mejorar la calidad de la desambiguación son las propuestas por Chaplot y Salakhutdinov [87] y por Dongsuk y colaboradores [88], entre otras.

Otro tema de investigación interesante es mejorar la dinámica del Modelo de Integración Léxico – Semántico (MILS), de manera que pueda discernir en qué situaciones es conveniente considerar un mayor aporte de las TL al puntaje final, y en que situaciones convendría un mayor aporte de las TS. Teniendo en cuenta que las TL generan tres rankings, una aproximación posible, es otorgar una mayor ponderación a estas técnicas, cuanto menor sea la dispersión existente en la posición de un recurso determinado en los tres rankings generados.

Finalmente, considerando que el usuario determina si un recurso es relevante o no de acuerdo a sus necesidades, se vuelve deseable poder establecer mecanismos de interacción con el sistema a medida que se va produciendo la recuperación. De esta manera el usuario puede aportar información que mejore los resultados que se le van a proporcionar en el futuro.

GLOSARIO DE TÉRMINOS

***Crawler*:** Programa que inspecciona las páginas del World Wide Web (WWW) de forma metódica y automatizada. Utilizado frecuentemente por los motores de búsqueda para la exploración e identificación de recursos nuevos.

ConcepNet: Es una red semántica, que contiene conocimiento de sentido común extraído de manera automática desde diversos corpus de textos.

Concepto: En el ámbito de WordNet y ConceptNet, un concepto hace referencia a una palabra, en conjunto con una connotación determinada.

Conjunto parcialmente ordenado no estricto: Un orden parcial es una relación binaria reflexiva, asimétrica y transitiva. Un conjunto, en combinación con un orden parcial, se denomina conjunto parcialmente ordenado no restrictivo.

Connotación o sentido: Se refiere al sentido asociado, expresivo o adicional que posee una palabra o frase según el contexto. La connotación de una palabra o frase indica su significado secundario en determinado contexto, lo que generalmente muestra un sentido más amplio de lo textual.

Corpus: Colección de recursos recuperados, que serán evaluados por las técnicas de determinación de relevancia de los sistemas de recuperación de información.

Denotación: Es el significado básico, formal y objetivo que posee una palabra o frase, el significado que es reconocido y entendido, en términos generales, por todas aquellas personas que hablan un mismo idioma.

Diccionario: Conjunto de palabras o términos que se encuentran ordenados alfabéticamente, junto a sus significados, definiciones, etimologías, ortografía, pronunciación, separación silábica y forma gramatical.

HTML: Lenguaje de marcado de hipertexto, que establece una estructura básica estándar y un código, para definir el contenido de un sitio web.

Machine Lerning: Es un subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan a realizar actividades sin necesidad de que sean explícitamente programadas para ello.

***Natural Languaje Processing* (NLP):** Es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano.

Ranking: Lista de recursos ordenados de acuerdo a su relevancia con respecto a una clave de búsqueda. Son el resultado del proceso de ordenamiento llevado a cabo por un sistema de recuperación de información.

Recursos o documentos web: Es información electrónica capaz de contener texto, sonido, vídeo, programas, enlaces, imágenes y otras cosas, adaptada para la World Wide Web (WWW) y que puede ser accedida mediante un navegador web.

Relevancia: La relevancia es la característica de aquello que tiene importancia y pertinencia con respecto a un tema o tópico en particular. En el caso de los sistemas de recuperación de información, permite determinar qué tan pertinente es un recurso a una clave de búsqueda definida por el usuario.

Scraper o Web Scraping: Actividad que consiste en la utilización de técnicas y herramientas para la extracción de información de los recursos web.

Synset: En WordNet, un synset es un conjunto de una o más palabras que son sinónimas en una connotación determinada y que, por lo tanto, pueden ser intercambiadas en una frase sin alterar su significado.

Taxonomía: Clasificación y organización de objetos que poseen características en común y que se encuentran relacionados entre sí.

Tesauro: Es una lista de palabras empleadas para representar conceptos. Proporciona una organización semántica considerando las relaciones establecidas entre dichos conceptos y el significado de los términos que los representan.

URL: Localizador de recursos uniforme, mediante el que se accede a recursos de información disponible en internet.

WordNet: Es una base de datos léxica que agrupa palabras en inglés, en conjuntos de sinónimos llamados synsets, proporcionando definiciones cortas y generales y almacenando sus relaciones semánticas.

Word Sense Disambiguation* (WSD)*: Es una actividad del NLP, que consiste en la identificación de la correcta connotación de la palabra de acuerdo al contexto en el que se encuentra.

REFERENCIAS

[1] B. Arango Alzate, L. Tamayo Giraldo, and A. Fadul Barbosa, “Vigilancia tecnológica: metodologías y aplicaciones,” *Rev. Gest. Las Pers. Tecnol.*, no. 13, pp. 154–158, 2012.

[2] G. H. Tolosa and F. R. A. Bordignon, *Introducción a la Recuperación de Información*. Tolosa y Bordignon, 2008.

[3] E. Abadal and L. Codina, *Bases de datos documentales: características, funciones y método*. Madrid: Síntesis, 2008.

[4] M. I. Ramírez, D. E. Rua, and B. A. Alzate, “Vigilancia Tecnológica e Inteligencia Competitiva,” *Rev. Gest. Las Pers. Tecnol.*, vol. 5, no. 13, p. 12, May 2012.

[5] K. Eckert, F. Favret, M. Barboza, L. M. Witzke, and V. M. Alvarenga, “Modelos de análisis de información para la toma de decisiones estratégicas del sector tealero,” presented at the XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina), 2016.

[6] K. Eckert, V. M. Alvarenga, M. Barboza, L. M. Witzke, and L. Airaldi, “Vigilancia tecnológica e inteligencia competitiva basada en técnicas de minería de la web,” presented at the XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)., 2016.

[7] F. Favret, R. Montiel, V. Alvarenga, M. Barboza, and L. Witzke, “Recuperación de información basada en técnicas de minería Web,” *2016*, p. 7.

[8] S. Johlic, “Understanding Semantic Analysis (and why this title is totally meta),” 11-Jul-2015. [Online]. Available: https://www.linkedin.com/pulse/understanding-semantic-analysis-why-title-totally-meta-shannon-johlic. [Accessed: 17-Sep-2018].

[9] J. Morris and G. Hirst, “Non-classical lexical semantic relations,” in *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics - CLS ’04*, Boston, Massachusetts, 2004, pp. 46–51.

[10] J. Ge and Y. Qiu, “Concept Similarity Matching Based on Semantic Distance,” in *2008 Fourth International Conference on Semantics, Knowledge and Grid*, Beijing, China, 2008, pp. 380–383.

[11] M. D. Boni and S. Manandhar, “The Use of Sentence Similarity as a Semantic Relevance Metric for Question Answering,” in *New Directions in Question Answering*, 2003.

[12] V. Groues, Y. Naudet, and O. Kao, “Adaptation and Evaluation of a Semantic Similarity Measure for DBPedia: A First Experiment,” in *2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization*, Luxembourg City, Luxembourg, 2012, pp. 87–91.

[13] S. S. Soliman, M. F. El-Sayed, and Y. F. Hassan, “Semantic Clustering of Search Engine Results,” *Sci. World J.*, vol. 2015, pp. 1–9, 2015.

[14] F. Benedetti, D. Beneventano, S. Bergamaschi, and G. Simonini, “Computing inter-document similarity with Context Semantic Analysis,” *Inf. Syst.*, Feb. 2018.

[15] K. M. Ravi, J. Mori, and I. Sakata, “Cross-Domain Academic Paper Recommendation by Semantic Linkage Approach Using Text Analysis and Recurrent Neural Networks,” in *2017 Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, OR, 2017, pp. 1–10.

[16] X. Kong, Q. Kong, W. Mao, and S. Tang, “Deep news event ranker based on user relevant query,” in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, 2018, pp. 363–367.

[17] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. New York : Harlow, England: ACM Press ; Addison-Wesley, c1999.

[18] G. Salton and M. J. McGill, *Introduction to modern information retrieval*. New York: McGraw-Hill, 1983.

[19] D. C. Blair, *Language and representation in information retrieval*. Amsterdam ; New York : New York, N.Y., U.S.A: Elsevier Science Publishers ; Distributors for the U.S. and Canada, Elsevier Science Pub. Co, 1990.

[20] C. Cleverdon, J. Mills, and M. Keen, *ASLIB Cranfield Research Project: factors determining the performance of indexing systems*. 1966.

[21] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: information retrieval in practice*, Internat. ed. Boston, Mass.: Pearson, 2010.

[22] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, “Evaluating collaborative filtering recommender systems,” *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, Jan. 2004.

[23] L. G. Jaimes and F. V. Riveros, “Modelos clásicos de recuperación de la información,” *Rev. Integr.*, vol. 23, no. 1, p. 10.

[24] S. E. Robertson and K. S. Jones, “Relevance weighting of search terms,” *J. Am. Soc. Inf. Sci.*, vol. 27, no. 3, pp. 129–146, May 1976.

[25] P. Bennet, S. T. Dummais, and E. Horvitz, “Probabilistic combination of text classifiers using reliability indicators,” in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp. 207–214.

[26] W. Phillips, “Introduction to Natural Language Processing - The Mind Project,” *Introduction to Natural Language Processing*. [Online]. Available: http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php. [Accessed: 23-Aug-2018].

[27] 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, May 2015.

[28] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse Process.*, vol. 25, no. 2–3, pp. 259–284, 1998.

[29] E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis,” in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA, 2007, pp. 1606–1611.

[30] K. Lund and C. Burgess, “Producing high-dimensional semantic spaces from lexical co-occurrence,” *Behav. Res. Methods Instrum. Comput.*, vol. 28, no. 2, pp. 203–208, Jun. 1996.

[31] “WordNet Search - 3.1.” [Online]. Available: http://wordnetweb.princeton.edu/perl/webwn. [Accessed: 24-Sep-2018].

[32] “ConceptNet.” [Online]. Available: http://conceptnet.io/. [Accessed: 25-Sep-2018].

[33] Z. Wu and M. Palmer, “VERB SEMANTICS AND LEXICAL SELECTION,” in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, USA, 1994, pp. 133–138.

[34] Y. Li, Z. A. Bandar, and D. McLean, “An Approach for Measuring Semantic Similarity Between Words Using Multiple Information Sources,” *IEEE Trans Knowl Data Eng*, vol. 15, no. 4, pp. 871–882, Jul. 2003.

[35] D. Lin, “Principle-based Parsing Without Overgeneration,” in *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 1993, pp. 112–120.

[36] P. Resnik, “Using Information Content to Evaluate Semantic Similarity in a Taxonomy,” *ArXivcmp-Lg9511007*, Nov. 1995.

[37] G. Salton, A. Wong, and C. S. Yang, “A Vector Space Model for Automatic Indexing,” *Commun ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[38] K. Sparck Jones, S. Walker, and S. E. Robertson, “A probabilistic model of information retrieval: development and comparative experiments: Part 1,” *Inf. Process. Manag.*, vol. 36, no. 6, pp. 779–808, Nov. 2000.

[39] K. Sparck Jones, S. Walker, and S. E. Robertson, “A probabilistic model of information retrieval: development and comparative experiments: Part 2,” *Inf. Process. Manag.*, vol. 36, no. 6, pp. 809–840, Nov. 2000.

[40] D.-J. Kim, S.-C. Lee, H.-Y. Son, S.-W. Kim, and J. B. Lee, “C-Rank and its variants: A contribution-based ranking approach exploiting links and content,” *J. Inf. Sci.*, vol. 40, no. 6, pp. 761–778, Dec. 2014.

[41] J. Lyons, “SEMANTICS, Volume I,” *Camb. Univ. Press*, p. 12.

[42] A. Budanitsky and G. Hirst, “Evaluating WordNet-based Measures of Lexical Semantic Relatedness,” *Comput Linguist*, vol. 32, no. 1, pp. 13–47, Mar. 2006.

[43] J. Gracia and E. Mena, “Web-Based Measure of Semantic Relatedness,” p. 15.

[44] D. A. Cruse and D. A. Cruse, *Lexical Semantics*. Cambridge University Press, 1986.

[45] M. L. Murphy, *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge University Press, 2003.

[46] V. C. Storey, “Understanding semantic relationships,” *VLDB J.*, vol. 2, no. 4, pp. 455–488, Oct. 1993.

[47] R. Chaffin, D. J. Herrmann, and M. Winston, “An empirical taxonomy of part-whole relations: Effects of part-whole relation type on relation identification,” *Lang. Cogn. Process.*, vol. 3, no. 1, pp. 17–48, Jan. 1988.

[48] D. Crystal and D. Crystal, *A dictionary of linguistics and phonetics*, 6th ed. Malden, MA ; Oxford: Blackwell Pub, 2008.

[49] V. Fromkin, R. Rodman, and N. Hyams, *An introduction to language*, 9. ed., international student ed. South Melbourne, Victoria: Wadsworth, Cengage Learning, 2011.

[50] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, “Introduction to WordNet: An On-line Lexical Database,” *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, Dec. 1990.

[51] L. W. Barsalou, “Intraconcept similarity and its implications for interconcept similarity,” in *Similarity and Analogical Reasoning*, S. Vosniadou and A. Ortony, Eds. Cambridge: Cambridge University Press, 1989, pp. 76–121.

[52] G. Lakoff, *Women, Fire, and Dangerous Things*. University of Chicago Press, 1990.

[53] R. Chaffin and D. J. Herrmann, “The similarity and diversity of semantic relations,” *Mem. Cognit.*, vol. 12, no. 2, pp. 134–141, Mar. 1984.

[54] M. W. Evens, *Lexical-semantic relations: a comparative survey*. Linguistic Research, 1980.

[55] F. J. Damerau and N. Indurkhya, *Handbook of natural language processing*. Boca Raton, FL: Chapman et Hall/CRC, 2010.

[56] M. Lesk, “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone,” in *Proceedings of the 5th Annual International Conference on Systems Documentation*, New York, NY, USA, 1986, pp. 24–26.

[57] S. Banerjee, “Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet,” p. 98.

[58] A. Montoyo, M. Palomar, G. Rigau, and A. Suarez, “Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods,” *J. Artif. Intell. Res.*, vol. 23, pp. 299–330, Mar. 2005.

[59] W. A. Gale, K. Church, and D. Yarowsky, “A method for disambiguating word senses in a corpus,” *Comput. Humanit.*, vol. 26, pp. 415–439, Dec. 1992.

[60] Y. K. Lee and H. T. Ng, “An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation,” in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 41–48.

[61] U. Zernik, “Train1 vs. Train2: Tagging Word Senses in Corpus,” in *Intelligent Text and Image Handling - Volume 2*, Paris, France, France, 1991, pp. 567–585.

[62] R. Poli, M. Healy, and A. Kameas, Eds., *Theory and Applications of Ontology: Computer Applications*. Dordrecht: Springer Netherlands, 2010.

[63] D. Geeraerts, H. Cuyckens, and L. Janda, *Inflectional Morphology*. Oxford University Press, 2012.

[64] R. Lieber, *Derivational Morphology*. Oxford University Press, 2017.

[65] P. ten Hacken, *Compounding in Morphology*. Oxford University Press, 2017.

[66] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun. Acm*, vol. 38, pp. 39–41, 1995.

[67] H. Liu and P. Singh, “ConceptNet — A Practical Commonsense Reasoning Tool-Kit,” *BT Technol. J.*, vol. 22, no. 4, pp. 211–226, Oct. 2004.

[68] R. Speer and C. Havasi, “Representing General Relational Knowledge in ConceptNet 5,” p. 8.

[69] T. Slimani, “Description and Evaluation of Semantic Similarity Measures Approaches,” *Int. J. Comput. Appl.*, vol. 80, no. 10, pp. 25–33, Oct. 2013.

[70] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and application of a metric on semantic nets,” *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 1, pp. 17–30, Feb. 1989.

[71] T. Slimani, B. B. Yaghlane, and K. Mellouli, *A New Similarity Measure based on Edge Counting*. .

[72] C. Leacock and M. Chodorow, “Filling in a sparse training space for word sense identification,” March, 1994.

[73] A. Tversky, “Features of similarity,” *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.

[74] E. G. M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, “X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies,” *J. Digit. Inf. Manag. JDIM*, vol. 4, 2006.

[75] P. Resnik, “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language,” p. 36.

[76] J. J. Jiang and D. W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” *ArXivcmp-Lg9709008*, Sep. 1997.

[77] R. Knappe, H. Bulskov, and T. Andreasen, “On Similarity Measures for Concept-based Querying,” Aug. 2008.

[78] Z. Zhou, Y. Wang, and J. Gu, “New model of semantic similarity measuring in wordnet,” in *2008 3rd International Conference on Intelligent System and Knowledge Engineering*, Xiamen, China, 2008, pp. 256–261.

[79] M. A. Alvarez and C. Yan, “A graph-based semantic similarity measure for the gene ontology,” *J. Bioinform. Comput. Biol.*, vol. 09, no. 06, pp. 681–695, Dec. 2011.

[80] “How to Write a Spelling Corrector.” [Online]. Available: http://norvig.com/spell-correct.html. [Accessed: 05-Oct-2018].

[81] “Natural Language Toolkit — NLTK 3.3 documentation.” [Online]. Available: https://www.nltk.org/. [Accessed: 05-Oct-2018].

[82] “Pattern | CLiPS,” 26-Nov-2010. [Online]. Available: http://www.clips.ua.ac.be/pattern. [Accessed: 09-Sep-2018].

[83] “nlp Python - RegEx para dividir texto en oraciones (oración-tokenización) - CODE Q&A Resuelto.” [Online]. Available: https://code.i-harness.com/es/q/188b1dc. [Accessed: 01-Feb-2019].

[84] F. E. Favret, H. A. Pfeifer, and M. G. Rojas, “Modelo de análisis semántico de contenido Web,” in *6to Congreso Nacional de Ingeniería en Informática/Sistemas de Información*, Mar Del Plata, Buenos Aires, Argentina, 2018.

[85] “The WordSimilarity-353 Test Collection.” [Online]. Available: http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/. [Accessed: 16-Oct-2018].

[86] “Comparing Variables of Ordinal or Dichotomous Scales: Spearman Rank-Order, Point-Biserial, and Biserial Correlations,” in *Nonparametric Statistics for Non-Statisticians*, Wiley-Blackwell, 2011, pp. 122–154.

[87] D. S. Chaplot and R. Salakhutdinov, “Knowledge-based Word Sense Disambiguation using Topic Models,” *ArXiv180101900 Cs*, Jan. 2018.

[88] O. Dongsuk, S. Kwon, K. Kim, and Y. Ko, “Word Sense Disambiguation Based on Word Similarity Calculation Using Word Vector Representation from a Knowledge-based Graph,” in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, 2018, pp. 2704–2714.

ANEXO I
MODELO DE ANÁLISIS SEMÁNTICO DE DOCUMENTOS WEB

Modelo de análisis semántico de contenido Web

Fabián E. Favret[1], Matías G. Rojas[2], Hernán A. Pfeifer[3]
Universidad Gastón Dachary
Av. López y Planes 6519, Posadas, Misiones
efabianfavret@citic.ugd.edu.ar[1], {rojasmatias994[2],hernan.a14[3]}@gmail.com

Resumen

El avance de la tecnología trajo consigo cambios estructurales en el funcionamiento de las organizaciones, afectando incluso a los niveles directivos donde la necesidad de información se vuelve un aspecto fundamental para la toma de decisiones. En la actualidad, la cantidad de información disponible en internet, es infinita, por lo que encontrar recursos relevantes para una necesidad de información determinada, resulta una tarea compleja. En los últimos años se produjeron avances en mecanismos que permitan la recuperación de recursos relevantes a partir de técnicas inteligentes de procesamiento de información desestructurada, para lo que se destacan dos enfoques principales: el sintáctico y el semántico. En el presente artículo, se presenta un modelo de determinación de relevancia de documentos WEB basado en técnicas de análisis semántico, que permiten la realización de una evaluación más exhaustiva al contemplar términos relacionados a la clave de búsqueda y el contexto al que se enmarca la búsqueda. Para evaluar la efectividad del mismo se lleva a cabo una comparación con técnicas de análisis de documentos basados en correspondencia lexicográfica, donde ambas técnicas fueron evaluadas de acuerdo a la coincidencia presentada con los criterios de los expertos. A partir de estas evaluaciones se determinó como resultado la factibilidad e idoneidad de utilizar estas técnicas, debido a la concordancia observada con respecto a la evaluación de los expertos.

1. Introducción

Durante las últimas décadas el formidable avance tecnológico ha generado cambios significativos en el funcionamiento de las organizaciones. Estos cambios han afectado a todas las actividades en general y, en particular, al proceso de toma de decisiones en los niveles directivos que se ve afectado por el gran volumen de información del contexto y la generación de datos internos de funcionamiento. Evidentemente, este crecimiento exponencial de la cantidad de información requiere mejoras de las técnicas de recolección y análisis [1] [2] [3].

Hoy en día el análisis de grandes volúmenes de datos se ha transformado en una prioridad para el proceso de toma de decisiones, generándose de esta manera la necesidad de utilizar técnicas inteligentes de procesamiento de información desestructurada [4] [5] [6].

Si bien, mediante la utilización de los motores de búsqueda tradicionales, obtener información de la Web es relativamente sencillo, surge una cuestión no tan simple que se debe analizar con cuidado: ¿Los resultados obtenidos son los más adecuados? Claramente, para responder este interrogante hay examinar varios factores.

El primer factor es determinar si realmente se ha hecho correctamente la solicitud de información. Este punto está relacionado a la formulación de las claves de búsqueda que utiliza el usuario y que por lo general no explota todo el potencial que los buscadores proveen. Es decir, por lo general las claves contienen el tema general de búsqueda y alguna que otra característica que se intenta satisfacer, pero no las restricciones posibles que pueden ser colocadas mediante las herramientas de la búsqueda avanzada.

Ahora bien, suponiendo que el problema se ha definido con precisión, el segundo factor clave que afecta a los resultados es si se han revisado todas las fuentes de información que se disponen en la Web. Claramente la respuesta es no, ya que es imposible hacer un análisis exhaustivo de toda la Web. Es aquí donde el usuario se encuentra a merced de los algoritmos que generan los rankings [7] [8] [9] de recursos asociados a la búsqueda. Ese es el tercer factor de evaluación a la hora de analizar resultados, es decir, qué tan bien fueron construidos los rankings de los recursos encontrados. En principio hay dos enfoques bien definidos para verificar que los requerimientos de búsqueda y los recursos encontrados coinciden: el enfoque sintáctico y el enfoque semántico [10]. En el primer caso se intenta obtener una correspondencia literal, mientras que en el segundo la idea es que se pueda contextualizar el análisis del recurso encontrado. Evidentemente el análisis semántico requiere de mayor complejidad que el sintáctico, pero tiene mayor probabilidad de retornar resultados útiles para el usuario y por ello existe una cantidad elevada de trabajos que intentan implementar este tipo de técnicas [11] [12] [13] [14].

En el contexto de este tema de investigación se ha desarrollado un sistema de búsqueda lexicográfico [15] [16] que comienza con los primeros resultados devueltos por los motores de búsqueda tradicionales y hace una exploración por niveles. Al mismo tiempo analiza los recursos obtenidos y genera el ranking correspondiente. En adición a ese trabajo se presenta aquí un enfoque semántico que tiene como objetivo evaluar los recursos utilizando conceptos de relación y similitud semántica sobre bases de conceptos y taxonomías disponibles.

Este enfoque, permite realizar una evaluación que, además de considerar la ocurrencia de términos que pertenezcan a la clave de búsqueda en los documentos analizados, también se valore la ocurrencia de términos relacionados, teniendo en cuenta el contexto en el que se encuentran enmarcados. Tales términos, surgen de relaciones semánticas existentes, como la sinonimia[1], antonimia[2], hiperonimia[3], meronimia[4], etc.

Entonces, dada una serie de requerimientos de usuarios, la idea principal es establecer la pertinencia y adecuación de recursos Web analizando la estructura semántica de los mismos. Para ello, se propone un modelo que lleve a cabo tres procesos fundamentales: el preprocesamiento y desambiguación del sentido de la clave de búsqueda, la identificación de oraciones conformantes de los documentos a analizar y desambiguación del sentido de los mismos; y la evaluación de similitud semántica existente entre la clave de búsqueda y los documentos.

Este artículo está estructurado de la siguiente manera: En la sección 2 se presenta qué es el análisis semántico, en la sección 3 se describe el modelo propuesto. En la sección 4 se muestra las pruebas y resultados. En la sección 5 se plantea una discusión con respecto a los resultados obtenidos y finalmente en la sección 6 se presentan algunas conclusiones y trabajos futuros.

2. Análisis semántico

Al momento de establecer la relevancia de un documento determinado basada en las coincidencias que este posee con respecto a la clave de búsqueda ingresada por el usuario, existen dos enfoques bien definidos: el enfoque sintáctico y el enfoque semántico.

Mientras que el enfoque sintáctico se centra en la correspondencia lexicográfica de términos de la clave de búsqueda dentro del documento analizado; el enfoque semántico hace énfasis en el significado de las palabras y las oraciones, teniendo en consideración el contexto en el que estas están enmarcadas.

Desde el punto de vista de la lingüística, se plantea que la semántica refiere a la existencia de "Significados" de la palabra, donde cada una de estas, posee un significado independiente del contexto denominado "Denotación", un significado propio del contexto al que está enmarcado denominado "Connotación" y relaciones con otras palabras también propias del contexto en el que está enmarcado [17].

A partir de esto surgen tres conceptos que contribuyen a la determinación del grado de coincidencia semántica existente entre documento y clave de búsqueda, los cuales son: relación semántica, similitud semántica y distancia semántica. A continuación, se introducen tales conceptos y se presentan algunas herramientas a ser utilizadas [18].

2.1. Relación, similitud y distancia semántica

Usualmente existe la confusión entre los conceptos de relación y similitud semántica; si bien a menudo se los utiliza de manera indiferente, no son idénticos. Para esclarecer esta diferencia, Resnik [19] plantea el siguiente ejemplo: "Automóviles y gasolinas" parecen estar más estrechamente relacionados que automóvil y bicicleta, pero evidentemente estos últimos son más parecidos. La similitud es un caso especial de relación semántica, la cual se limita a solamente relaciones del tipo "es – un" y las relaciones de sinonimia, mientras que la relación semántica contempla todas las relaciones posibles existentes entre dos términos.

Un concepto que aparece para causar aún más confusión, es el de distancia semántica, el cual puede usarse cuando se habla tanto de similitud como de relación semántica en general. Este concepto plantea que, a mayor cercanía entre dos términos en una determinada ontología, mas es la relación entre ambos [20].

Teniendo en cuenta esto, dados dos términos, $T1$ y $T2$, pertenecientes a diferentes nodos ($n1$ y $n2$) conformantes de una ontología determinada, la distancia semántica determina la relación semántica existente entre los términos $T1$ y $T2$.

2.2. WordNet

WORDNET [21][5] es una base de datos léxica, que modela el conocimiento léxico del idioma inglés, desarrollada por la Universidad de Princeton, en la que sustantivos, verbos, adjetivos y adverbios se organizan en conjuntos de sinónimos, donde cada uno de ellos representa un concepto léxico [22].

[1] **Sinonimia**: Relación de igualdad existente entre el significado de dos o más palabras.

[2] **Antonimia**: Relación de oposición entre los significados de dos palabras.

[3] **Hiperonimia**: Relación en la que el significado de una palabra engloba a otra.

[4] **Meronimia:** es una relación semántica no simétrica entre los significados de dos palabras dentro del mismo campo semántico.

[5] **WordNet** - https://wordnet.princeton.edu/ - © 2018 The Trustees of Princeton University

El vocabulario de un lenguaje es definido como un conjunto de pares (f, s), donde una forma f, es un $string$ sobre un alfabeto finito y un sentido s es un elemento de un significado determinado. Cada forma, en conjunto con un sentido, es una palabra en ese vocabulario.

En *WORDNET*, una forma y un sentido, es representado mediante un conjunto de uno o más sinónimos que posee en ese sentido, denominado Synset. En su última versión, *WORDNET* contiene alrededor de 117.659 Synsets y 209.941 pares (f, s) [23].

En *WORDNET* existe un conjunto de relaciones semánticas entre los Synsets, seleccionadas a partir de su alto uso en el idioma inglés, algunas de las cuales son: sinonimia, antonimia, hiperonimia, etc.

Cada una de estas relaciones semánticas son representadas a partir de interconexiones entre los synsets (mediante la utilización de punteros, en una estructura de árbol); un ejemplo de ello puede ser visto en la Figura 1, donde se pueden apreciar ejemplos de interconexiones a partir de las relaciones semánticas existentes entre términos.

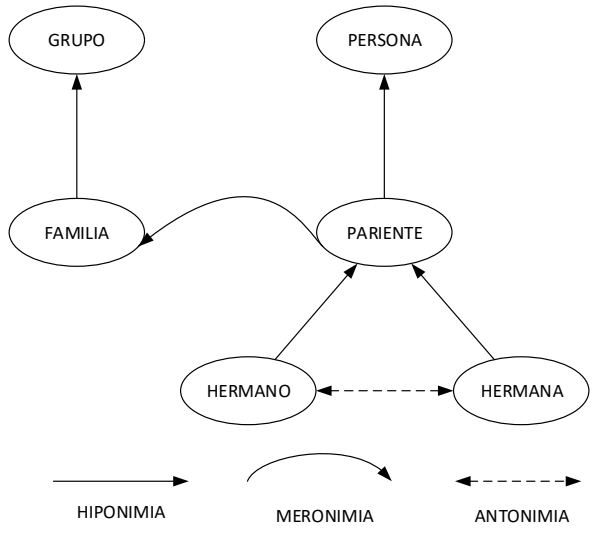

Figura 1 - Ejemplo de árbol de relaciones semánticas en *WORDNET* **[22]**.

2.3. Métrica de relación y similitud semántica de Slimani

La métrica Slimani de relación y similitud semántica es una métrica basada en estructura, lo que significa que necesita de una estructura ontológica jerárquica para poder estimar la relación semántica entre dos términos. Surge como una extensión a la métrica de Wu and Palmer [24], la cual plantea que la relación semántica entre dos términos puede ser obtenida mediante la siguiente fórmula:

$$Sim_{wp}(T1, T2) = \frac{2 * N}{N1 + N2} \quad (1)$$

Donde $T1$ y $T2$ son dos términos en una taxonomía, N es la distancia a la raíz, del hiperónimo común a los dos términos analizados. $N1$ *y* $N2$, son las distancias a la raíz, de los nodos correspondientes a los términos $T1$ y $T2$ respectivamente [25]. Ver Figura 2.

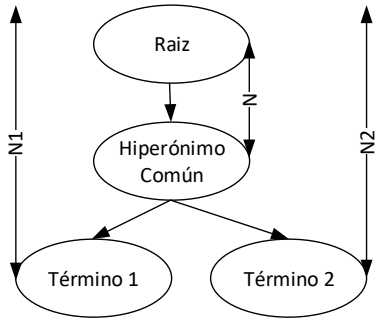

Figura 2 - Ejemplo Estructura Jerárquica **[26]**

Esta métrica tiene la desventaja de que no siempre genera resultados satisfactorios; específicamente en la situación en que otorga un valor de similitud alto a relaciones entre términos con sus vecinos, comparados con los valores obtenidos para términos pertenecientes a una misma jerarquía. Dada la jerarquía presentada en la Figura 3, la métrica de Wu and Palmer, otorga un mayor puntaje a la relación entra A y B, que, a la relación entre A y D, aun considerando que D es un merónimo de A.

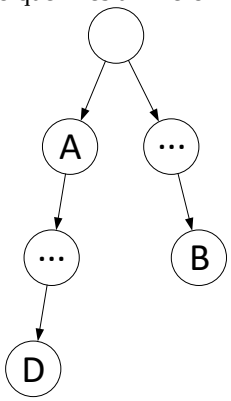

Figura 3 - Ejemplo de jerarquía **[25]**

Para dar solución a esto, Slimani [26], planteo la siguiente fórmula:

$$Sim_{sli}(T1, T2) = \frac{2 * N}{N1 + N2} * PF(T1, T2) \quad (2)$$

Donde:

- $PF(T1, T2)$ es un factor penalización para términos que sean vecinos. Y está dada por la siguiente fórmula:

$$PF(T1, T2) = (1 - \lambda) * (Min(N1, N2) - N) + \lambda * (|N1 - N2| + 1)^{-1} \quad (3)$$

Donde:

- λ es el coeficiente y es un valor booleano, que indica con un valor de 0 que dos términos están en una misma jerarquía y 1 que dos términos son vecinos [26].

2.4. ConceptNet

CONCEPTNET [27][6] es una base de conocimiento de sentido común de gran escala, con un conjunto de herramientas de procesamiento de lenguaje natural que da soporte a muchas tareas prácticas de razonamiento textual sobre documentos del mundo real. El alcance de *CONCEPTNET* es comparable a la base de datos léxica *WORDNET*. Sin embargo, existen algunas diferencias, como, por ejemplo, mientras que *WORDNET* fue optimizado para la categorización léxica y la determinación de similitud de palabras, *CONCEPTNET* fue optimizado para realizar inferencias basadas en el contexto, sobre textos del mundo real.

CONCEPTNET representa las distintas relaciones semánticas entre términos mediante aserciones de la forma $\{Término_1, Relación, Término_2\}$, conteniendo aproximadamente 1,6 millones de aserciones en su base de conocimiento, conectando más de 300.000 nodos. Estos nodos son fragmentos en el idioma inglés, semiestructurados, interrelacionados por una ontología de veinte relaciones semánticas (entre las que se encuentran "UsedFor", "LocationOf", "PartOf", etc.) [28].

3. Modelo propuesto

Para llevar a cabo la determinación de la relevancia de documentos, mediante la realización del análisis de relaciones y similitudes semántica se plantearon los módulos de la Figura 4.

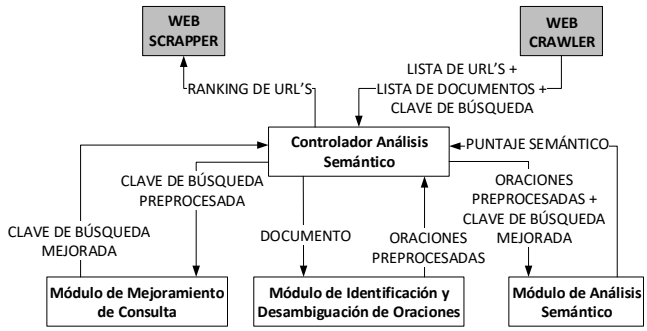

Figura 4 - Módulos del modelo propuesto

En el módulo *Web Crawler*, al recibir la clave de búsqueda proporcionada por el usuario, se desencadena un proceso de obtención de URL's a partir de cuatro buscadores distintos (Google, Bing, Intelligo, Msxml Excite), donde cada uno de ellos retorna diez URLs, las que actúan como semillas del proceso de exploración de enlaces. De este módulo, se obtiene una lista de URLs, producto de la exploración, la cual junto a la clave de búsqueda y una lista de documentos correspondiente a cada URL es enviada al *Controlador Análisis Semántico*, que es el encargado de coordinar toda la operatoria requerida para determinar la relevancia de una lista de documentos.

Una vez recibida la lista de URLs junto a sus documentos y la clave de búsqueda ingresada por el usuario, se envía esta última al *Módulo de Mejoramiento de Consulta*, el cual lleva a cabo el procesamiento de la clave de búsqueda para realizar posteriormente el análisis semántico de documentos.

Luego por cada documento, se realiza la separación del mismo en oraciones, para posteriormente desambiguar el sentido de cada palabra perteneciente a cada una de estas, mediante el *Módulo de Identificación y Desambiguación de Oraciones*.

A continuación, se envía el conjunto de oraciones pertenecientes al documento, junto a la clave de búsqueda mejorada al *Módulo de Análisis Semántico*, donde se ejecutan las métricas de similitud y relación semántica, retornando un puntaje que indica la relación semántica del documento, con respecto a la clave de búsqueda, que a su vez indica la relevancia del mismo.

Finalmente, en el *Controlador de Análisis Semántico*, se procede a realizar un ranking de URL's, de acuerdo al puntaje de relevancia obtenida a partir del *Módulo de Análisis Semántico*; el cual es enviado al módulo *Web Scrapper*, para ser presentada al usuario.

3.1. Módulo de mejoramiento de consulta

El módulo de mejoramiento de consulta es el encargado de realizar el procesamiento de la clave de búsqueda. Este proceso es necesario para ejecutar las métricas de relación y similitud semántica que serán aplicadas para determinar la relevancia de los documentos. Las interacciones de este módulo se observan en la Figura 5.

[6] **ConceoptNet** -Is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License - http://conceptnet.io/c/en

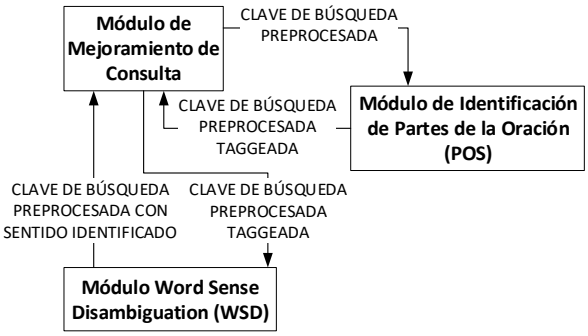

Figura 5 - Interacciones del módulo de mejoramiento de Consulta

Una vez recibida la clave de búsqueda, se busca identificar aquellas palabras que contengan errores ortográficos, con el fin de corregirlas (proceso denominado Spelling).

Posteriormente se envía la clave de búsqueda al *Módulo de Identificación de Partes de la Oración*, En el cual se identifica para cada término perteneciente a la clave, el rol que cumple en la misma, es decir, determina si un término es sustantivo, adjetivo, verbo, entre otros.

Para la ejecución de este módulo, se utiliza el método *Parse* perteneciente a la librería *Pattern* [29]. Como resultado de este módulo, se retorna la clave de búsqueda con cada uno de sus términos etiquetados de acuerdo al rol que cumple.

Luego, se envía la clave de búsqueda etiquetada, al *Módulo Word Sense Disambiguation (WSD)*, donde para cada término se busca identificar el sentido más aproximado al contexto al que la clave de búsqueda pertenece, teniendo en consideración el rol que el término cumple dentro de la misma. Para esto, se implementan dos métodos.

El primero de ellos es el *Algoritmo Original Lesk* [30] que consiste en elegir, como el sentido más adecuado de un término, a aquel cuya definición posea la mayor cantidad de palabras superpuestas con respecto a un corpus de comparación (términos de la oración a la cual pertenece el término a desambiguar (clave de búsqueda) y sus definiciones). El segundo consiste en armar un corpus de comparación, combinando artículos obtenidos desde wikipedia, cuyo título esté conformado por una frase compuesta por dos o más términos yuxtapuestos de la clave de búsqueda ingresada por el usuario.

Posteriormente, utilizando este corpus, se aplica por cada término de la clave de búsqueda el *Algoritmo Original Lesk* explicado anteriormente. Esto permite la ejecución del algoritmo, utilizando un corpus de comparación más extenso, lo que posibilita discernir el sentido de cada término de manera más precisa. Si no se obtienen artículos de wikipedia, se ejecuta el *Algoritmo Original Lesk* sin modificaciones.

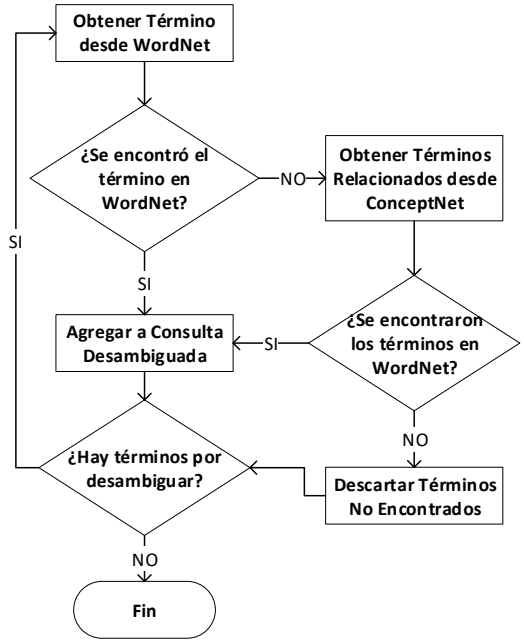

Figura 6 - Proceso de búsqueda en *WORDNET* y *CONCEPTNET*

Esta búsqueda del sentido más adecuado, se realiza sobre la base de datos léxica *WORDNET* [21]. Con el fin de evitar descartar términos que no se encuentren en ella, se realiza la búsqueda de las relaciones de estos términos en *CONCEPTNET* [27], el cual otorga una lista de términos relacionados, que al ser ubicados en la taxonomía de *WORDNET* son agregados a la clave de búsqueda, permitiendo de esta manera estimar más aproximadamente la relación semántica de tal termino no presente en la taxonomía. En el caso, de que no se encuentre en *WORDNET* los términos relacionados obtenidos a partir de *CONCEPTNET*, estos términos no encontrados son descartados y por lo tanto no son agregados a la clave de búsqueda final (Figura 6).

Finalmente se retorna la clave de búsqueda desambiguada al "Controlador Análisis Semántico".

3.2. Módulo de identificación y desambiguación de oraciones

En este módulo se lleva a cabo la extracción de las distintas oraciones que conforman a un documento. Esto se justifica, teniendo en cuenta a la definición de oración, que plantea que "la oración es conjunto de palabras que expresa un juicio con sentido completo y autonomía sintáctica", es decir, es la unidad mínima de texto que mantiene presente al contexto. La interacción de este módulo es presentada en la Figura 7.

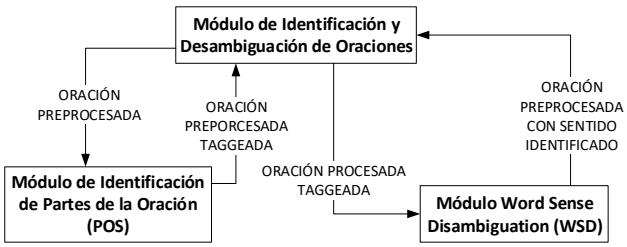

Figura 7 - Interacciones del módulo de identificación y desambiguación de oraciones

En el *Módulo de Identificación y Ponderación de Oraciones* se realiza la división del documento en las oraciones que la conforman. Posteriormente, se envían cada una de estas oraciones, al *Módulo de Identificación de Partes de la Oración*, donde para cada término se identifica el rol que cumple dentro de la oración (identificar si es sustantivo, verbo, adjetivo, etc.). Esto se lleva a cabo mediante la utilización del método *Parse* perteneciente a la librería *Pattern* [29].

Luego, cada oración es enviada al *Módulo Word Sense Disambiguation (WSD)*, donde para cada término, se determina el sentido más aproximado de acuerdo al contexto al que está enmarcado la oración a la que pertenece y también teniendo en cuenta el rol que cumple dentro de la oración. Para ello, se implementa el *Algoritmo Original Lesk* [30] explicado anteriormente.

Finalmente, se retorna al módulo *Controlador Análisis Semántico*, el conjunto de oraciones desambiguadas, pertenecientes al documento.

3.3. Módulo de análisis semántico

En este módulo se recibe un documento, segmentado en oraciones (con cada uno de sus términos desambiguados) junto a la clave de búsqueda.

Aquí se pone en ejecución la métrica basada en estructura desarrollada por "Slimani", la cual se utiliza para medir la similitud y relación semántica de las oraciones de un documento en particular con respecto a la clave de búsqueda mejorada.

Durante el proceso de determinación de la relación semántica de un documento con respecto a la clave de búsqueda, se procesa cada oración, acumulando por cada término perteneciente a la clave de búsqueda, aquellas relaciones con términos conformantes de la oración analizada, si y solo si, el puntaje de esta relación es mayor o igual al valor de 0,5. De esta manera se logra determinar la satisfacción de cada término de la clave de búsqueda por parte de cada oración (esto se ve reflejado en la Figura 8). Al procesar todas las oraciones, se obtendrá la satisfacción de cada término de la clave de búsqueda por parte del documento en su totalidad.

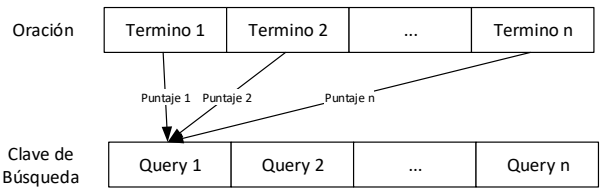

Figura 8 - Cálculo de satisfacción de términos de la clave de búsqueda por oración

Además, para cada término de la clave de búsqueda se realiza el conteo de la cantidad de términos del documento relacionados con cada uno de ellos. Esto permite obtener el puntaje de relación promedio por cada término de la clave de búsqueda.

Para determinar el puntaje de relación semántica del documento con respecto a la clave de búsqueda, se implementa la siguiente fórmula:

$$puntDocumento_x = \frac{1}{1 + e^{-[10x(puntSemantico_x - 0.5)]}} \qquad (4)$$

Tratándose en este caso de una *función logística o sigmoidal*, la cual tiene la particularidad de facilitar que los valores de $puntSemántico_x$ lleguen a los extremos, posibilitando que se alcance el puntaje máximo durante la ejecución. Donde $puntSemántico_x$ representa el puntaje de relación semántica del $documento_x$ en su totalidad, considerando la satisfacción de cada término de la clave de búsqueda por parte de cada término de la oración. Esto se obtiene mediante la siguiente fórmula:

$$puntSemántico_x = \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{n} \left(\frac{Query_{ji}}{termRelacionados_{ji}} * Ponderación_j \right)}{n} \right) \qquad (5)$$

Donde $Query_{ji}$ es la sumatoria de los puntajes de relación, de los términos del $documento_x$ relacionados al $i-ésimo$ término de la clave de búsqueda, perteneciente al concepto j. $termRelacionados_{ji}$ es la cantidad de términos del $documento_x$ que poseen un puntaje de relación y similitud semántica con respecto a la clave de búsqueda de al menos $0{,}5$. n es la cantidad de términos de la clave de búsqueda, pertenecientes al concepto j. m es la cantidad de conceptos existentes en la clave de búsqueda y $Ponderación_j$ es la ponderación de cada conjunto de términos de la clave de búsqueda ($j-ésimo$ Concepto) separados por $AND, OR\ o\ NOT$. La utilización de la ponderación se fundamenta en que el orden en el que los usuarios ingresan los conceptos pertenecientes a la clave de búsqueda está relacionado a la importancia de la presencia de los mismos en los documentos que se recuperen. Esta es obtenida mediante la siguiente fórmula:

$$Ponderación_j = 1 - (j-1) * \left(\frac{1}{m}\right) \qquad (6)$$

Una vez obtenido el $puntDocumento_x$, se aplica una fórmula de normalización, de manera que se obtengan las calificaciones pertenecientes a un intervalo cerrado [0;5], para lo cual se utiliza la siguiente fórmula:

$$puntNormalizado_x = puntDocumento_x * 5 \quad (7)$$

Finalmente, se retorna $puntNormalizado_x$ al módulo *Controlador Análisis Semántico*.

4. Pruebas y resultados

Con el objetivo de comparar las técnicas basadas en correspondencia lexicográfica y las técnicas basadas en distancia semántica, se plantean dos escenarios pertenecientes a áreas temáticas diferentes propuestos por expertos donde cada uno de ellos realiza una búsqueda relacionada con su campo de especialidad definiendo las características derivadas de sus dominios de experticia.

Una vez definidas las claves de búsqueda se ejecutan los procesos de búsqueda, utilizando los módulos de análisis lexicográfico y semántico, donde cada uno otorga una calificación para cada documento, perteneciente a un intervalo cerrado $[0; 5]$ (0, indica la no relevancia del documento y 5 se corresponde con la relevancia absoluta del documento). A la vez, los resultados obtenidos son calificados por los expertos de acuerdo al intervalo definido anteriormente.

Con el fin de evaluar las dos técnicas, se emplean el índice de correlación de ranking de Spearman [31], que otorga un valor perteneciente al rango $[-1; 1]$, para indicar el grado de similitud existente entre dos rankings. En este caso se compara la correlación entre el ranking evaluado por el experto y el ranking generado por cada técnica.

Además, se comparan los aciertos y errores en la calificación otorgada por cada técnica para cada documento, con respecto a las otorgadas por el experto; clasificando como *Coincidencias* a los aciertos, *Errores Leves* a las diferencias en las calificaciones en una unidad, *Errores Moderados* a las diferencias en las calificaciones en dos unidades y *Errores graves* a las diferencias en tres o más unidades entre ambas calificaciones.

Finalmente, se realiza el conteo de las situaciones en las que el sistema otorgó una calificación superior a la del experto (sobreestimó) y las situaciones en las que el sistema otorgó una calificación inferior a la del experto (subestimó).

4.1. Prueba 1: "Digital Storytelling"

La primera prueba realizada, pertenece al ámbito de educación digital, haciendo énfasis sobre la técnica de Digital Storytelling.

La clave de búsqueda proporcionada por el experto, es la siguiente: *Storytelling AND Digital Classroom AND Art in Technology Education NOT Art Education.*

Básicamente, se busca obtener todos los documentos WEB relacionados al *Storytelling*, que estén vinculados a las aulas digitales (*Digital Classroom*) y a las artes en la enseñanza de tecnologías (*Art in Technology Education*), exceptuando los artículos que refieran a las enseñanzas del arte (*Art Education*).

En la Tabla 1, se observan los valores correspondientes al coeficiente de correlación de Spearman, para la comparación entre el ranking generado por el experto y los rankings generados por cada una de las técnicas.

Tabla 1 - Comparación de los rankings mediante el coeficiente de Spearman

		COEFICIENTE DE SPEARMAN
Técnicas	Ranking Lexicográfico	0.180024
	Ranking Semántico	0.659447779

Se puede observar que las técnicas lexicográficas obtuvieron un valor de correlación de Spearman clasificado por [31], como una *correlación positiva débil*, lo que refleja poca coincidencia con respecto al ranking generado por el experto. En cambio, las técnicas semánticas obtuvieron un valor del coeficiente de correlación de Spearman clasificado como una *correlación positiva fuerte* [31]; es decir que los rankings generados por el experto y los generados por esta técnica, tienen un alto grado de coincidencia.

La cantidad de aciertos y errores cometidos por cada técnica, se muestran en la Tabla 2. Para el caso del análisis lexicográfico, se calificó en un 34% de manera coincidente con el criterio del experto contra un 28% obtenido por el análisis semántico. Por otra parte, los errores en la calificación se acumularon en la categoría *Errores Graves* para el caso lexicográfico mientras que para el caso semántico la mayor cantidad de ellos se acumularon en *Errores Leves.* Esto implica que el análisis semántico calificó de manera menos distante al criterio del experto, que las técnicas de análisis lexicográficas. Un aspecto a destacar aquí es que se obtuvo un porcentaje bajo de *Errores Graves* (10% de los documentos), significando que en pocas ocasiones las calificaciones obtenidas por el análisis semántico, estuvieron alejadas a las otorgadas por el experto.

La Tabla 3 muestra que en los casos en que se cometieron errores en la calificación por parte del análisis lexicográfico, el 60% de los documentos obtuvieron calificaciones inferiores a las otorgadas por el experto mientras que para el análisis semántico se trataron de sobreestimaciones, es decir, que para el 48% de los documentos, se otorgó una calificación superior a la entre-

gada por el experto, mientras que el 24% obtuvo calificaciones inferiores.

Tabla 2 - Cantidad de coincidencias y errores cometidos por ambas técnicas.

	ANÁLISIS LEXICOGRÁFICO		ANÁLISIS SEMÁNTICO	
	CANT.	%	CANT.	%
Coincidencias (diferencia 0)	17	34%	14	28%
Errores Leves (diferencia 1)	9	18%	21	42%
Errores Moderados (diferencia 2)	5	10%	10	20%
Errores Graves (diferencia 3 o más)	19	38%	5	10%
TOTAL	50	100%	50	100%

Tabla 3 - Cantidad de coincidencias, subestimaciones y sobreestimaciones producidas por ambas técnicas.

	ANÁLISIS LEXICOGRÁFICO		ANÁLISIS SEMÁNTICO	
	CANT.	%	CANT.	%
Coincidencias	17	34%	14	28%
Subestimaciones	30	60%	12	24%
Sobreestimaciones	3	6%	24	48%
TOTAL	50	100%	50	100%

4.2. Prueba 2: "Cookie Poisoning"

La segunda prueba, se corresponde al área de la seguridad informática, donde la clave de búsqueda presentada por el experto fue la siguiente: *Cookie Poisoning AND Hacking Web Applications*.

Más específicamente, la clave de búsqueda hace referencia a las distintas técnicas existentes y en desarrollo relacionadas a *cookie poisoning* utilizables en la vulneración de aplicaciones web; con el fin de tomar medidas que permitan la protección ante este tipo de ataques maliciosos. Se puede observar que esta clave es mucho menos restrictiva que la de la prueba anterior y se decidió dejarla así para evaluar situaciones de flexibilidad.

En la Tabla 4, se pueden observar que los resultados muestran que ambos Rankings obtuvieron un coeficiente de correlación que pertenece a la categoría *correlación positiva fuerte*, obteniendo en este caso, una mayor correlación con el criterio del experto, las técnicas de análisis lexicográfico.

Tabla 4 - Comparación de los rankings mediante el coeficiente de Spearman

		COEFICIENTE DE SPEARMAN
Técnicas	Ranking Lexicográfico	0.74823529
	Ranking Semántico	0.65656663

Las coincidencias y los errores obtenidos para cada técnica, se presentan en la tabla 5. En esta prueba, el análisis lexicográfico tuvo mayor porcentaje de coincidencia con las calificaciones otorgadas por el experto (72%) comparado con el análisis semántico que obtuvo el 30% de coincidencias. En cuanto a los errores, en el caso del análisis lexicográfico la mayoría de ellos se acumularon en las categorías *Errores Leves* y *Errores Moderados* mientras que para el semántico la mayoría de los errores (46%), fueron categorizados como *Errores Leves*, lo que muestra cercanía con los criterios del experto.

A partir de los errores cometidos por ambos análisis, se resumen en la tabla 6, las cantidades de sobreestimaciones y subestimaciones producidas.

Tabla 5 - Cantidad de coincidencias y errores cometidos por ambas técnicas

	ANÁLISIS LEXICOGRÁFICO		ANÁLISIS SEMÁNTICO	
	CANT.	%	CANT.	%
Coincidencias (diferencia 0)	36	72%	15	30%
Errores Leves (diferencia 1)	6	12%	23	46%
Errores Moderados (diferencia 2)	7	14%	1	2%
Errores Graves (diferencia 3 o más)	1	2%	11	22%
TOTAL	50	100%	50	100%

Tabla 6 - Cantidad de coincidencias, subestimaciones y sobreestimaciones producidas por ambas técnicas.

	ANÁLISIS LEXICOGRÁFICO		ANÁLISIS SEMÁNTICO	
	CANT.	%	CANT.	%
Coincidencias	36	72%	15	30%
Subestimaciones	10	20%	4	8%
Sobreestimaciones	4	8%	31	62%
TOTAL	50	100%	50	100%

5. Discusión

Los resultados presentados en la sección anterior, resaltan un aspecto interesante de considerar. A partir de los errores cometidos por el sistema, se contemplaron las situaciones en las que esté sobreestima o subestima a la calificación otorgada por el experto. De esto surge la necesidad de establecer cuál situación es más deseable, considerando el efecto que cada uno de ellos tenga sobre los resultados finales obtenidos.

En el caso en que el sistema subestima a la calificación otorgada por el experto, documentos con alto grado de relevancia, son calificados como poco relevante, lo que puede provocar que queden excluidos del ranking y que el experto no pueda tener acceso a ellos. En contraposición, cuando se subestima a documentos malos, se

estaría realizando una acción concordante con el criterio del experto, lo que generaría que estos se ubiquen en posiciones bajas del ranking, o sean excluidos del mismo.

En el caso en que el sistema sobrestima a la calificación otorgada por el experto, documentos poco relevantes, obtendrán calificaciones altas, aumentando la cantidad de documentos poco relevantes en el ranking. Esto posee la ventaja de que si bien, la precisión en la obtención de documentos relevantes disminuye, debido a la cantidad de documentos no relevantes recuperados, no se estaría privando al usuario de resultados que potencialmente sean buenos, otorgándole la posibilidad tener acceso a todos ellos y descartar aquellos documentos no relevantes.

6. Conclusiones y trabajos futuros

En este artículo se describe un modelo de determinación de relevancia de documentos WEB, que evalúa la relación semántica del contenido de los mismos con respecto a la clave de búsqueda ingresada por el usuario.

En las simulaciones realizadas se puede apreciar que las técnicas lexicográficas obtuvieron mejores resultados por la ocurrencia de los términos de la clave de búsqueda. Sin embargo, teniendo en cuenta a los errores cometidos por cada técnica, se puede observar que la consideración de términos relacionados y el contexto de la clave de búsqueda, producen que el criterio de determinación de relevancia de documentos sea cercano al criterio del experto. Esto se ve reflejado también en que los coeficientes de correlación de Spearman señalan una correlación positiva fuerte para el ranking semántico con respecto al ranking generado por el experto.

También se observa que solo contemplar la aparición de términos de la clave de búsqueda hace que las calificaciones tiendan a los extremos, es decir, si hay una alta aparición de términos de la clave en un documento, la calificación es alta. Esto se puede ver reflejado en los valores obtenidos para coeficiente de correlación de ranking de Spearman, donde en la primera prueba se obtuvo una correlación casi nula, del ranking obtenido por esta técnica con respecto al realizado por el experto.

Otro aspecto interesante es que en ambas pruebas, el análisis lexicográfico produjo mayor cantidad de subestimaciones y el análisis semántico produjo mayor cantidad de sobreestimaciones, haciendo evidente la influencia de contemplar términos relacionados a la clave de búsqueda, lo que provoca que se incremente la calificación de relevancia a documentos que tengan mayor cantidad de términos relacionados, contrario a lo que sucede si solo se contempla la aparición explícita de términos de la clave de búsqueda en los documentos.

Los resultados demuestran la factibilidad de utilizar técnicas semánticas como medio de determinación de relevancia de documentos, esto teniendo en cuenta que es un área que se encuentra en pleno desarrollo, y que por ende tiene grandes desafíos asociados. Uno de ellos es la dificultad que representa determinar correctamente el contexto y realizar una taxonomía que contemple la mayor cantidad de relaciones semánticas y términos posibles. Por ello, estas técnicas podrían generar buenos resultados si son utilizadas como complemento a las técnicas lexicográficas, lo que permitiría ampliar el campo de exploración a la hora de determinar la relevancia.

A partir de lo expuesto, se pretende avanzar en el sentido de identificar técnicas de desambiguación de contexto más precisas, por lo que se están explorando alternativas como: las técnicas de desambiguación del sentido de la palabra basados en modelado de tópicos [32], técnicas de desambiguación del sentido de la palabra basado en el cálculo de la similitud de palabras utilizando representación de la palabras en vectores, a partir de gráficos basados en conocimientos [33], entre otras.

Agradecimientos

Este trabajo es parte del Proyecto “Modelo de Análisis de Información Desestructurada Utilizando Técnicas de Recopilación y Minería Web”, código A07002, desarrollado en la Universidad Gastón Dachary – Posadas, Misiones, Argentina.

Referencias

[1] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval: The Concepts and Technology behind Search.*, 2ed edition. Addison-Wesley Educational Publishers Inc, 2011.

[2] Madankar, M., Chandak, M., and Chavhan, N., “Information Retrieval System and Machine Translation: A Review. Procedia Computer Science,” vol. 78, pp. 845–850, 2016.

[3] Ren, F. and Bracewell, D. B., “Advanced Information Retrieval. Electronic Notes in Theoretical Computer Science,” vol. 225, pp. 303–317, 2009.

[4] Al-Jarrah, O., Muhaidat, S., Karagiannidis, G. K., and Taha, K., “Efficient Machine Learning for Big Data: A Review. Big Data Research,” *ELSEVIER*, vol. 2, pp. 87–93, 2015.

[5] Mueller, E. T., “Commonsense Reasoning Using Unstructured Information. In Commonsense Reasoning,” *ELSEVIER*, pp. 315–335, 2015.

[6] Portugal, I., Alencar, P., and Cowan, D., “The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications,” *ELSEVIER*, vol. 97, pp. 205–227, 2018.

[7] Bozkir, A. S. and Akcapinar Sezer, E., “Layout-based computation of web page similarity ranks.

International Journal of Human Computer Studies," *ELSEVIER*, vol. 110, pp. 95–114, 2018.

[8] Yan, E. and Ding, Y., "Discovering author impact: A PageRank perspective. Information Processing & Management," *ELSEVIER*, vol. 47, pp. 125–134, 2011.

[9] Zareh Bidoki, A. M. and Yazdani, N., "DistanceRank: An intelligent ranking algorithm for web pages. Information Processing & Management," *ELSEVIER*, vol. 44, pp. 877–892, 2008.

[10] Ferreira, R., Lins, R. D., Simske, S. J., Freitas, F., and Riss, M., "Assessing sentence similarity through lexical, syntactic and semantic analysis," vol. 39, pp. 1–28, 2016.

[11] Augier, M., Shariq, S., and Thanning Vendelo, M., "Understanding context: its emergence, transformation and role in tacit knowledge sharing. Journal of Knowledge Management," *MCB UP Ltd*, vol. 5, pp. 125–137, 2001.

[12] Benedetti, F., Beneventano, D., Bergamas, S., and Simonini, G., "Computing interdocument similarity with Context Semantic Analysis," *ELSEVIER*, 2018.

[13] Hsu, P.-L., Hsieh, H. S., Liang, J. H., and Chen, Y. S., "Mining various semantic relationships from unstructured user-generated web data. Web Semantics: Science, Services and Agents on the World Wide Web," vol. 31, pp. 27–38, 2015.

[14] Oliva, J., Serrano, J., del Castillo, M. D., and Iglesias, Á., "A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering," *ELSEVIER*, vol. 70, pp. 390–405, 2011.

[15] Montiel, R., Lezcano Airaldi, L., Favret, F., and Eckert, K., "Web Information Retrieval System for Technological Forecasting," *Journal of Computer Science & Technology. UNLP*, vol. 17, pp. 49–58, 2017.

[16] Eckert, K., Favret, F., BARBOZA, M., WITZKI, A., and ALVARENGA, V., "Modelos de análisis de información para la toma de decisiones estratégicas del sector tealero," *WICC*, pp. 117–121, 2016.

[17] W. Phillips, "Introduction to Natural Language Processing - The Mind Project," *Introduction to Natural Language Processing*. [Online]. Available: http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php. [Accessed: 23-Aug-2018].

[18] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Comput Linguist*, vol. 32, no. 1, pp. 13–47, 2006.

[19] Resnik, P., "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *IJCAI-95*, vol. 1, p. 448, 1995.

[20] J. Gracia and E. Mena, "Web-Based Measure of Semantic Relatedness," *SPRINGER-VERLAG*, vol. 5175, pp. 136–150, 2008.

[21] "WordNet | A Lexical Database for English." [Online]. Available: https://wordnet.princeton.edu/. [Accessed: 10-Sep-2018].

[22] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An On-line Lexical Database*," *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, 1990.

[23] G. A. Miller, "WordNet: A Lexical Database for English," *Commun. Acm*, vol. 38, pp. 39–41, 1995.

[24] WU, Z. and Palmer, M., "Verbs semantics and lexical selection," pp. 133–138, 1994.

[25] Slimani, T., "Description and Evaluation of Semantic Similarity Measures Approaches," *Int. J. Comput. Appl.*, vol. 80, no. 10, pp. 25–33, 2013.

[26] Slimani, T., Yaghlane, B., and Mellouli, K., "A New Similarity Measure based on Edge Counting," *IJWesT*, vol. 3, no. 4, 2012.

[27] "ConceptNet." [Online]. Available: http://conceptnet.io/. [Accessed: 10-Sep-2018].

[28] H. Liu and P. Singh, "ConceptNet — A Practical Commonsense Reasoning Tool-Kit," *BT Technol. J.*, vol. 22, no. 4, pp. 211–226, Oct. 2004.

[29] "Pattern | CLiPS," 26-Nov-2010. [Online]. Available: http://www.clips.ua.ac.be/pattern. [Accessed: 09-Sep-2018].

[30] M. Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, New York, NY, USA, 1986, pp. 24–26.

[31] "Comparing Variables of Ordinal or Dichotomous Scales: Spearman Rank-Order, Point-Biserial, and Biserial Correlations," in *Nonparametric Statistics for Non-Statisticians*, Wiley-Blackwell, 2011, pp. 122–154.

[32] D. S. Chaplot and R. Salakhutdinov, "Knowledge-based Word Sense Disambiguation using Topic Models," *ArXiv180101900 Cs*, Jan. 2018.

[33] Dongsuk, S. Kwon, K. Kim, and Y. Ko, "Word Sense Disambiguation Based on Word Similarity Calculation Using Word Vector Representation from a Knowledge-based Graph," 2018.