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.

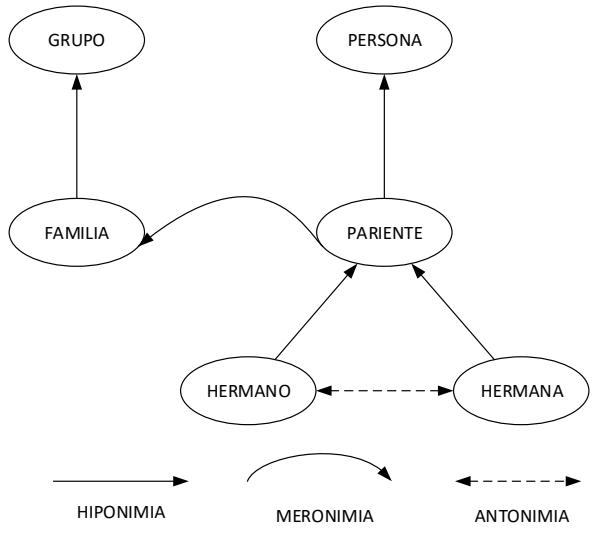

Figura 1 - Ejemplo de árbol de relaciones semánticas en *WORDNET* **[22]**.

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.

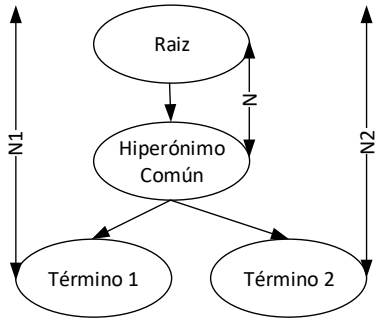

Figura 2 - Ejemplo Estructura Jerárquica **[26]**

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.

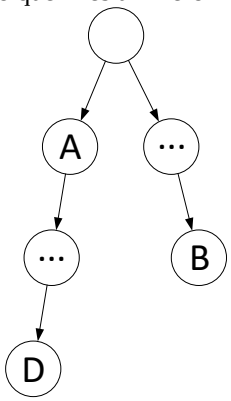

Figura 3 - Ejemplo de jerarquía **[25]**

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.

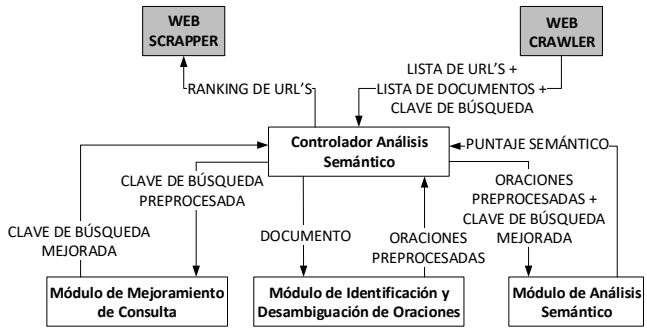

Figura 4 - Módulos del modelo propuesto

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

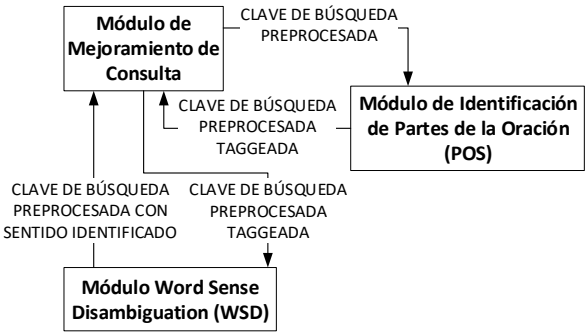

Figura 5 - Interacciones del módulo de mejoramiento de Consulta

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.

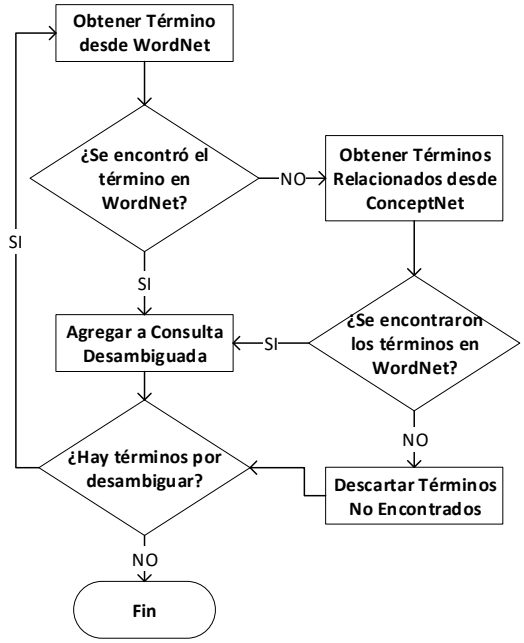

Figura 6 - Proceso de búsqueda en *WORDNET* y *CONCEPTNET*

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.

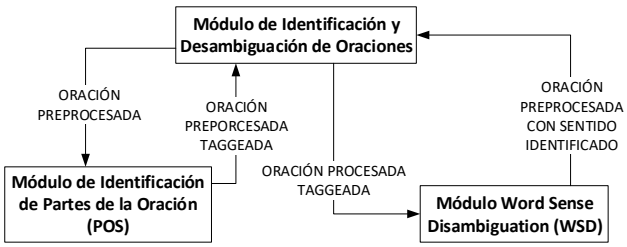

Figura 7 - Interacciones del módulo de identificación y desambiguación de oraciones

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.

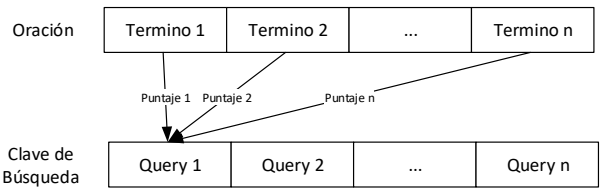

Figura 8 - Cálculo de satisfacción de términos de la clave de búsqueda por oración

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.