CATEGORIZACIÓN DE RESÚMENES DE PUBLICACIONES CIENTÍFICAS BASADA EN SIMILITUD SEMÁNTICA

José Alejandro Reyes Ortiz

Universidad Autónoma Metropolitana, Unidad Azcapotzalco

jaro@correo.azc.uam.mx

Maricela Claudia Bravo Contreras

Universidad Autónoma Metropolitana, Unidad Azcapotzalco

mcbc@correo.azc.uam.mx

Resumen

Los resúmenes de las publicaciones científicas se encuentran disponibles de manera abierta, es decir en repositorios con acceso libre. En el área de las ciencias computacionales, estos repositorios no están organizados con temáticas de dominio específicas, esto lleva a que una tarea de localización de una publicación de interés requiere de un trabajo exhaustivo por parte de las personas interesadas. En este artículo se describe un enfoque para la categorización de resúmenes de publicaciones científicas utilizando mediciones de similitud basadas en conocimiento semántico para obtener el grado de relación entre los textos. Una experimentación ha sido presentada en términos de Precisión, Exhaustividad y medida F, la cual muestra resultados prometedores para la categorización de resúmenes en el dominio de las ciencias computacionales.

Palabra(s) Clave(s): categorización de textos, medidas basadas en conocimiento, publicaciones científicas, similitud semántica.

1. Introducción

En la actualidad, grandes cantidades de artículos científicos del área de las ciencias computacionales son publicados en repositorios, donde los resúmenes tienen un acceso libre y, en la mayoría de los casos, el resto del artículo tiene

acceso restringido. Los resúmenes de las publicaciones proporcionan la idea general, la aportación, la solución, el método, algoritmo, técnica, enfoque o herramienta para un problema específico. Existen varios sitios o repositorios sobre publicaciones, tales como:

- El sitio Google Académico[1] soportado por la empresa Google que contiene información sobre citas y perfiles especializada en literatura científica-académica.
- DBLP[2] es un sitio web, patrocinado por la Universidad de Trier, que ofrece un servicio para consultar información bibliográfica sobre revistas y memorias sobre el área de las ciencias computacionales.
- CiteSeerX[3] es una librería digital para publicaciones científicas y académicas, principalmente en los campos de la computación e informática, el cual incluye citas, documentos y estadísticas.

Estos repositorios no tienen una estructura semántica o una categorización temática de sus publicaciones. Sin embargo, ofrecen herramientas que hacen posible acceder a los resúmenes de publicaciones científicas mediante el uso de buscadores basados en palabras clave. Para que los usuarios puedan localizar un artículo de su interés, es necesario invertir mucho tiempo y esfuerzo, ya que se requiere un trabajo exhaustivo de análisis de resultados proporcionados por el buscador. Aunado a esto, se tiene que los buscadores son propensos a proporcionar resultados irrelevantes debido a la carencia de estructuras semánticas en los datos que ayuden a mejorar los resultados de las búsquedas.

El enfoque propuesto en este artículo se centra en categorizar resúmenes científicos utilizando mediciones de similitud basadas en conocimiento semántico extraído de WordNet [1]. La idea principal consiste en obtener una similitud semántica entre palabras y luego entre pares de resúmenes que sean las bases del algoritmo de categorización de resúmenes del área de las ciencias computacionales. El objetivo de este enfoque es mejorar la organización de los

[1] Google Académico, `https://scholar.google.com.mx/`

[2] DBLP Computer Science Bibliography, `http://dblp.uni-trier.de/`

[3] CiteSeerx, `http://citeseerx.ist.psu.edu/index`

repositorios de resúmenes de tal forma que se faciliten la localización de artículos de interés para el usuario.

El resto del artículo está organizado de la siguiente manera. En la Sección 2 se presentan los trabajos relacionados con la categorización de textos y resúmenes de publicaciones científicas. La Sección 3 explica seis mediciones de similitud basadas en conocimiento semántico. El enfoque propuesto en este artículo para la categorización de resúmenes de publicaciones científicas basada en métricas de similitud semántica, se muestra en la Sección 4. La experimentación y resultados de la categorización se exponen en la Sección 5. Finalmente, las conclusiones y el trabajo a futuro son presentados en la Sección 6.

2. Trabajos relacionados

La categorización de textos es un tema de investigación que ha sido ampliamente estudiado desde diversos enfoques estadísticos y semánticos. Sin embargo, la categorización de textos científicos ha sido abordada, en menor medida, utilizando métricas de similitud basadas en conocimiento semántico. Por lo tanto, en esta sección se presenta una revisión de trabajos relacionados, iniciando por el tema general, categorización de textos, y terminando, con la categorización de resúmenes de publicaciones científicas.

Diversos enfoques estadísticos han sido propuestos para la categorización de textos independientes del idioma. En [2] se describe un enfoque para la clasificación de documentos basada en las frecuencias de distribución de las palabras de los textos en las clases, con una técnica no supervisada por lo que no necesita ejemplos previamente etiquetados. Las Máquinas de Soporte Vectorial (MSV) han sido una técnica utilizada para la categorización automática de textos como en [3] y [4], los cuales utilizan las máquinas de soporte vectorial como un clasificador de textos con aprendizaje automático a partir de ejemplos etiquetados para cada clase, los autores argumentan que el uso de MSV logran un desempeño en la categorización de textos igual que métodos robustos. Además, las Máquinas de Soporte Vectorial pueden ser enriquecidas con mediciones de distancias entre los vectores de características textuales, como en [5] que proponen un enfoque

basado en la función de la distancia Euclidiana con MSV tanto para la fase de categorización como de entrenamiento. Finalmente, en esta categoría, se presenta el trabajo de [6], el cual consiste en un enfoque Bayesiano para la categorización automática de textos usando características específicas para cada clase.

También, las Redes Neuronales Artificiales (RNA) han sido utilizadas para la categorización de textos, como en [7] que se presenta un marco de trabajo semi-supervisado con una red neuronal convolutiva para la categorización de textos, mediante la integración de regiones de textos cortos como datos en la etapa de entrenamiento, la cual no necesitará ejemplos etiquetados. En [8] se mejora el enfoque de la red neuronal convolutiva mediante la exploración de nuevas regiones de textos para su integración haciendo uso de del método llamado memoria extensa de términos cortos.

Las mediciones de similitud cuantifican el grado de relación que existe entre dos palabras. Por su parte, estas mediciones consideran información a partir de las redes semánticas. Diversas métricas basadas en conocimiento semántico extraído de *WordNet* han sido utilizadas como es el caso del trabajo presentado en [9], el cual expone seis métricas para calcular el grado de similitud entre dos textos, mediante la combinación de la similitud palabra a palabra dentro de una métrica global para obtener le grado de relación semántica entre dos textos.

Finalmente, la categorización de resúmenes científicos ha sido abordada en los siguientes trabajos. En [10] se utiliza el, ya bien conocido, algoritmo supervisado del k-vecino más cercano aplicando una técnica de punto de transición durante el proceso de selección de términos relevantes, para la categorización de resúmenes. Esta técnica utiliza la frecuencia media de los términos para categorizar un texto debido al hecho de que ella aporta un contenido semántico alto. Un enfoque simple para el agrupamiento de resúmenes científicos es presentado en [11], el cual consiste en agrupar palabras clave y usar mediciones de similitud sintáctica entre documentos y en [12] que presentan una metodología para la construcción de sistemas de clasificación a nivel de instancias de publicaciones científicas en el área de las ciencias de la computación. Los autores

proponen un agrupamiento de las publicaciones dentro de áreas de investigación basado en las relaciones de las cita.

3. Mediciones de similitud basadas en conocimiento semántico

Las mediciones de similitud cuantifican el grado de relación que existe entre dos palabras. Por su parte, las mediciones semánticas consideran información a partir de las redes semánticas de las palabras para obtener su grado de relación.

En este artículo hemos recopilado, a partir del estado del arte, un conjunto de mediciones de similitud basadas en conocimiento (redes semánticas de las palabras): *Wu & Palmer, Lin, Leacock & Chodorow, Lesk, Resnik y Jian & Conrath*. Estas mediciones de similitud semántica utilizan la base de datos *WordNet*, una red semántica de conceptos (palabras), y utilizando las relaciones jerárquicas (es-un) y no jerárquicas con la finalidad de obtener el grado de relación. A continuación se presenta una breve descripción de estas medidas de similitud basadas en conocimiento semántico.

Wu & Palmer [13] presenta una métrica de similitud que considera la profundidad de los dos conceptos dados dentro de la taxonomía de WordNet, y la profundidad del concepto común más específico (*LCS*) entre el par de conceptos (c_1, c_2), y combina estos dos valores dentro de una medición de similitud, ecuación 1.

$$Sim_{WuP}(c_1, c_2) = \frac{2 \times profundidad\,(LCS)}{profundidad\,(c_1) + profundidad\,(c_2)} \quad (1)$$

Una medida que llamaremos Lin, presentada en [14], retorna el contenido de la información (*IC*) de los dos conceptos de entrada, considerando un factor de normalización. El contenido de la información (IC) describe la cantidad de información necesaria para establecer la concordancia entre los dos conceptos de entrada, ecuación 2.

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times IC(LCS)}{IC(c_1) + IC(c_2)} \quad (2)$$

Por su parte, Leacock & Chodorow [15] presentan una métrica de similitud semántica que utiliza la longitud (*L*) del camino más corto entre los dos conceptos

usando un conteo de nodos y la profundidad máxima (*D*) de la taxonomía, ecuación 3.

$$Sim_{LCH}(c_1, c_2) = -log\frac{L}{2*D} \quad (3)$$

La similitud semántica entre dos conceptos de Lesk [16] es definida como la función de traslape entre las definiciones de los conceptos correspondientes, las cuales son extraídas de un diccionario. Esta medición es propuesta como una solución para la desambiguación de sentidos de palabras.

La medida de similitud propuesta por Resnik [17] retorna al concepto común más específico (*LCS*) entre el par de conceptos, añadiendo la probabilidad de encontrar una instancia del concepto común más específico *P(LCS)*, ecuación 4.

$$Sim_{Res}(c_1, c_2) = -\log P\,(LCS) \quad (4)$$

Finalmente, la similitud semántica presentada por Jiang & Conrath [18] que utiliza el contenido de la información (IC) del par de conceptos correspondientes y del concepto común más específico (*LCS*) entre ellos, la cual regresa un grado de similitud determinado mediante ecuación 5.

$$Sim_{JiC}(c_1, c_2) = \frac{1}{IC(c_1)+IC(c_2)-2\times IC(LCS)} \quad (5)$$

Todas las mediciones anteriores consideran un par de palabras como entrada y generan un valor que indica el grado de relación semántica existente entre ellas.

4. Categorización de resúmenes de publicaciones científicas

En esta sección se presenta el enfoque utilizado para la categorización de publicaciones científicas utilizando la similitud entre resúmenes. El proceso completo de categorización de resúmenes de publicaciones científicas incluye las siguientes etapas: pre-procesamiento de los resúmenes; segmentación de los textos y etiquetado morfológico; cálculo de la similitud entre textos y, finalmente, el agrupamiento de los resúmenes. En la figura 1, se puede observar este proceso.

Pre-procesado de los resúmenes

El texto de los resúmenes es segmentado en palabras y un etiquetado de partes de la oración es aplicado mediante *TreeTagger* [19] con el apoyo del

software *GATE- Arquitectura General para la Ingeniería de Textos* [20]. La segmentación en palabras se lleva a cabo utilizando el delimitador espacio en blanco o salto de línea.

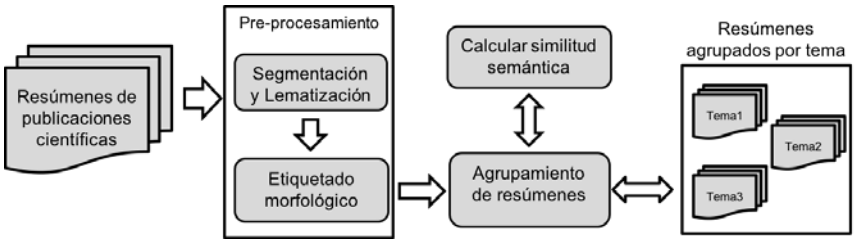

Figura 1 Proceso de categorización de resúmenes.

La lematización consiste en obtener la forma normalizada de cada palabra, es decir, convertir la forma flexionada de una palabra a su forma normal, con las siguientes reglas: singular para sustantivos, masculino y singular para adjetivos e infinitivo para verbos.

Por su parte, el etiquetado de partes de la oración se encarga de asignar una categoría gramatical a cada palabra de una oración. En la figura 2 se muestra un ejemplo de la salida, lema y categoría gramatical, del etiquetador *TreeTagger* para un extracto de texto de un resumen.

Texto: *This paper focuses on mining social networks using the information about event logs*
Salida: *This/this/DT paper/paper/NN focuses/focus/VB on/on/IN mining/mine/VB social/social/JJ networks/network/NN using/use/VB the/the/DT information/information/NN about/about/IN event/event/NN logs/log/NN*
donde: DT = Determinante, NN = Sustantivo, VB = Verbo, IN = Preposición y JJ = Adjetivo.

Figura 2 Texto etiquetado y lematizado con TreeTagger.

La información generada en esta etapa se utiliza para calcular la similitud semántica considerando métricas basadas en el conocimiento semántico, y mediante agrupar las palabras en categorías gramáticas (verbos, sustantivos, adverbios).

Calcular la similitud semántica entre resúmenes

El cálculo de la similitud semántica entre resúmenes se basa en las métricas de similitud entre palabras. Primero se obtiene una similitud entre pares de palabras

de la misma categoría y, después, una similitud global del resumen. Se utiliza una medida de similitud direccional presentada en [9], la cual indica la similitud semántica de un resumen con respecto al otro. Esta similitud nos proporciona un conocimiento direccional, el cual es combinado entre las dos medidas unidireccionales para obtener una similitud bidireccional.

Para un par de resúmenes de publicaciones dadas, se lleva a cabo un proceso de segmentación, lematización y etiquetado morfológico. Con estas etiquetas se crean conjuntos de palabras normalizadas para verbos, sustantivos y adjetivos. A partir de estos conjuntos se obtiene la similitud, basada en WordNet, entre pares de palabras de la misma categoría. La finalidad de la lematización es poder obtener una similitud alta, aun cuando, las palabras estén en distintas formas flexionadas, además, los conjuntos por categorías evitan comparar semánticamente sustantivos con verbos o sustantivos con adjetivos. Un ejemplo de esta creación de conjuntos por categorías gramaticales para dos extractos de resúmenes se puede observar en la figura 3.

Resumen 1: This paper focuses on mining social networks using the information about event logs.

$ConjCat_{verbo}$ = {focus, mine, use}

$ConjCat_{sust}$ = {paper, network, information, event, log}

$ConjCat_{adj}$ = {social}

Resumen 2: We study the social networks based on event types.

$ConjCat_{verbo}$ = {study, base}

$ConjCat_{sust}$ = {network, event, type}

$ConjCat_{adj}$ = {social}

Figura 3 Conjunto de categorías para dos resúmenes de artículos científicos.

La similitud entre pares de palabras se realiza de acuerdo a las mediciones de similitud presentadas en la Sección 3. La idea principal es obtener la similitud semántica más alta entre cada par de los conjuntos de palabras. Por lo tanto, se toma la similitud semántica entre dos segmentos presentada en [9] y se adapta para un par de resúmenes R_i y R_j , usando la función de similitud de palabra a palabra. Esta adaptación queda en ecuación 6.

$$Sim(R_i, R_j)_{R_i} = \frac{\sum_{i=1}^{n}(\sum_{w_k \in [ConjCat_i]} maxSim(w_k))}{\sum_{i=1}^{n} |ConjCat_i|} \quad (6)$$

La ecuación 6 se puede interpretar como sigue: w_k representa un par de palabras, $ConjCat_i$ expresa un conjunto de palabras de la misma categoría gramatical (verbo, sustantivo o adjetivo), y $maxSim(w_k)$ representa la similitud semántica más alta entre un par de palabras de la misma categoría. La tabla 1 muestra la similitud semántica unidireccional, con respecto al resumen 1, utilizando la medida de Wu & Palmer, entre los textos mostrados en la figura 3.

Tabla 1 Similitud unidireccional de Wu & Palmer para un par de resúmenes.

		verbos		**sustantivos**			**adjetivos**
		study	**base**	**network**	**event**	**type**	**social**
Verbos	**focus**	0.40	**0.41**	-	-	-	-
	mine	0.33	**0.33**	-	-	-	-
	use	0.50	**0.50**	-	-	-	-
Sustentivos	**paper**	-	-	0.43	**0.46**	0.32	-
	network	-	-	**1.00**	0.55	0.35	-
	information	-	-	0.50	**0.55**	0.35	-
	event	-	-	0.55	**1.00**	0.50	-
	log	-	-	**0.43**	0.40	0.29	
adjetivos	**social**	-	-	-	-	-	**1.00**

Se utilizan los valores de máximas similitudes entre pares de palabras, marcadas en la tabla 1 y, con ellos, se aplica la fórmula (6), para obtener una similitud unidireccional entre el resumen 1 y el resumen 2 de 0.631.

Esta ponderación, la cual se expresa con un valor entre 0 y 1, es el grado de similitud unidireccional basado en el significado de las palabras con respecto a un resumen (R_i) de esta manera se obtiene: $Sim(R_i, R_j)_{R_i}$. Con la finalidad de obtener una similitud bidireccional se utiliza una combinación de las similitudes unidireccionales, obtenidas del trabajo presentado en [9], usando una simple función de promedio, esta similitud se expresa mediante ecuación 7.

$$Sim(R_i, R_j) = \frac{sim(R_i,R_j)_{R_i} + sim(R_i,R_j)_{R_j}}{2} \quad (7)$$

El cálculo de la similitud semántica bidireccional entre un par de resúmenes se utiliza para estimar el grado de similitud de cada resumen con respecto al resto, con la finalidad de encontrar sus *k*-vecinos más cercanos, y de esta manera determinar la categoría (tópico) de pertenencia.

Agrupamiento de resúmenes de publicaciones científicas

El agrupamiento de resúmenes de publicaciones científicas fue realizado con el algoritmo de agrupamiento *k*-NN (*k*-vecinos más cercanos) utilizando el cálculo de similitud semántica entre un par de resúmenes descrito anteriormente. Además de utilizar similitudes de palabra a palabra basadas en *WordNet* y descritas en la Sección 3.

El algoritmo *k*-vecinos más cercanos [21] es un método de agrupamiento supervisado, cuya idea es encontrar la categoría de pertenencia de un objeto a partir de los ejemplos (vecinos) más cercanos en función de una similitud en el espacio de características. Este algoritmo necesita un conjunto de ejemplos de entrenamiento, es decir, ejemplos previamente etiquetados con sus categorías, con ellos el algoritmo será capaz de determinar, mediante un valor de *k* y la función de similitud, la clase de pertenencia de los ejemplos que se desean categorizar (no etiquetados).

En nuestro caso, los objetos corresponden a los resúmenes de las publicaciones científicas y la función de similitud está determinada por la similitud semántica entre resúmenes. De esta manera, el algoritmo *k*-vecinos más cercanos será utilizado para asignar la categoría a un conjunto de resúmenes que se desean clasificar. Este algoritmo necesita un conjunto de resúmenes de entrenamiento, es decir, resúmenes etiquetados con sus tópicos para el número (N) de categorías, con ellos el algoritmo, mediante un valor de k y la función de similitud semántica basada en conocimiento, determina la categoría de pertenencia de nuevos resúmenes.

Para la categorización de resúmenes de publicaciones científicas, el algoritmo *k*-vecinos más cercanos queda descrito de la siguiente manera:

- Se tiene como entrada un conjunto de resúmenes clasificados D = {(R_1,C_1), … , (R_N,C_N)} y un R_x a ser clasificado.
- Para cada resumen clasificado $R_i \in D$, calcular su grado de similitud semántica con el resumen a clasificar (R_x) mediante la fórmula descrita en la ecuación 7, de esta manera se obtiene: $Sim(R_i, R_x)$.
- Ordenar las similitudes $Sim(R_i, R_x)$ de mayor a menor.

- Seleccionar los *k* resúmenes más cercanos a R_x, es decir, cuya similitud sea mayor.
- Asignar al resumen R_x la etiqueta de la categoría con mayor frecuencia (mayoría de votos) de sus *k* resúmenes más cercanos.

5. Experimentación y resultados

Con la finalidad de evaluar el enfoque, se llevan a cabo una serie de experimentos, para los cuales utilizamos una colección de resúmenes de publicaciones científicas compiladas y disponibles por la herramienta ArnetMiner [22], las cuales fueron recopiladas a partir de DBLP, ACM y Citeseer. Este conjunto de datos está dividido en 10 categorías. A partir de éste, se extrae un subconjunto de resúmenes con tres categorías: a) Inteligencia Artificial, Ingeniería de Software e Interacción Humano-Computadora. Como resultado, nuestro conjunto está compuesto de 4467 resúmenes y 277 454 palabras (aproximadamente 62 palabras por resumen). Al estar usando un algoritmo supervisado para la categorización, nuestra colección de resúmenes es dividida en 2976 resúmenes para el entrenamiento y 1491 para las pruebas. La distribución de los resúmenes en las tres categorías se puede observar en la tabla 2.

Tabla 2 Distribución de nuestro conjunto de resúmenes.

Categoría	**Número total de resúmenes**	**Número de resúmenes de entrenamiento**	**Número de resúmenes de pruebas**
Inteligencia Artificial	1826	1217	609
Ingeniería de Software	1463	974	489
Interacción Humano-Computadora	1178	785	393

Los experimentos para evaluar la categorización de resúmenes se llevan a cabo con el algoritmo de *k*-vecinos más cercanos, y utilizando un valor de *k=5*. Para el cálculo de similitud entre resúmenes, se llevan a cabo experimentos con las 6 mediciones de similitud entre palabras basada en conocimiento semántico, y utilizando las ecuaciones 6 y 7.

La evaluación de todos los experimentos se realizó utilizando las métricas de Precisión (P), Exhaustividad (R) y medida F ampliamente utilizadas en la tarea

categorización y recuperación de información, en nuestro caso, categorización de resúmenes de artículos científicos. Estas métricas comparan los resultados del algoritmo de categorización a ser evaluado con los valores externos de confianza (resúmenes previamente clasificados, proporcionados en la etapa de entrenamiento), utilizando los siguientes valores: a) Verdadero Positivo (VP) es el número de predicciones correctas del algoritmo de categorización de resúmenes que corresponden al juicio externo de confianza (resúmenes preclasificados); Verdadero Negativo (VN) es el número de predicciones correctas del algoritmo de categorización de resúmenes que no corresponden al juicio externo de confianza; Falso Positivo (FP) corresponde al número predicciones incorrectas del algoritmo de categorización de resúmenes que corresponden al juicio externo de confianza; y, finalmente Falso Negativo (FN) es el número de predicciones incorrectas del algoritmo de categorización de resúmenes que no corresponden al juicio externo de confianza.

Bajo estos criterios, se emplea la *Precisión (P)* para evaluar los algoritmos en términos de los valores de predicciones positivas, la cual se define en ecuación 8.

$$P = \frac{VP}{VP+FP} \tag{8}$$

También, se utiliza el *Exhaustividad (R)* para expresar la tasa de correspondencias correctas con los resúmenes preclasificados de manera externa con una confianza alta, el cual se define en ecuación 9.

$$R = \frac{VP}{VP+FN} \tag{9}$$

Finalmente, la *medida F* que representa la media armónica entre Precisión y Exhaustividad, la cual tiene como fundamento obtener un valor único ponderado entre ellas y se define en ecuación 10.

$$medida\ F = 2 \times \frac{P \times R}{P+R} \tag{10}$$

Los experimentos se han organizado para medir el impacto de las seis mediciones de similitud basadas en conocimiento semántico para la categorización de resúmenes científicos. Bajo esta consideración, la tabla 3 muestra los resultados, por categoría, de los experimentos para tres mediciones de similitud semánticas (Wu & Palmer, Lin y Leacock & Chodorow) en términos de Precisión y

Exhaustividad, utilizando las ecuaciones 8 y 9 respectivamente. También, se proporciona un peso promedio, resultado de considerar la *medida F,* ecuación 10 y un factor que refleja la importancia (número de resúmenes) de cada categoría.

Tabla 3 Resultados de tres mediciones para la categorización de resúmenes científicos.

Categoría	**Wu & Palmer**			**Lin**			**Leacock & Chodorow**		
	P	R	F	P	R	F	P	R	F
Inteligencia Artificial	0.55	0.70	0.61	0.65	0.80	0.71	0.40	0.70	0.50
Ingeniería de Software	0.70	0.60	0.65	0.85	0.70	0.77	0.36	0.22	0.28
Interacción Humano-Computadora	0.79	0.68	0.73	0.81	0.76	0.78	0.7	0.42	0.53
Peso promedio	0.68	0.66	0.66	0.77	0.75	0.76	0.48	0.45	0.43

Por su parte, la tabla 4 muestra los resultados de los experimentos para tres mediciones de similitud semánticas restantes (Lesk, Resnik y Jiang & Conrath).

Tabla 4 Resultados de tres mediciones para la categorización de resúmenes científicos.

Categoría	**Lesk**			**Resnik**			**Jiang & Conrath**		
	P	R	F	P	R	F	P	R	F
Inteligencia Artificial	0.66	0.74	0.70	0.38	0.4	0.39	0.76	0.94	0.84
Ingeniería de Software	0.79	0.74	0.76	0.4	0.36	0.38	0.93	0.76	0.84
Interacción Humano-Computadora	0.85	0.80	0.83	0.46	0.48	0.47	0.92	0.86	0.88
Peso promedio	**0.77**	**0.76**	**0.76**	**0.41**	**0.41**	**0.41**	**0.87**	**0.85**	**0.86**

Los resultados mostrados en las tablas 3 y 4 hacen notar que la medición de similitud semántica de *Jiang & Conrath* como ponderación para la categorización de resúmenes científicos es la mejor alternativa considerando las tres categorías. Con esta métrica se ha logrado un 86 % de resúmenes clasificados correctamente.

Los resultados de nuestra experimentación demuestran la efectividad de nuestro enfoque para la categorización de resúmenes científicos. A pesar de que los resultados son más alentadores para la categoría “Interacción Humano-Computador”, el enfoque puede ayudar a los usuarios del ámbito científico a localizar artículos existentes en un repositorio desorganizado de resúmenes, ya

que pueden enfocar su búsqueda en una categoría y no en todo el repositorio. Además, se pueden hacer búsquedas por similitud semántica, proporcionando un resumen como ejemplo.

6. Conclusiones

Este artículo ha presentado un enfoque para la categorización de resúmenes de artículos científicos del área de las ciencias computacionales, considerando las siguientes categorías: artículos científicos sobre Interacción Humano-Computadora, Ingeniería de Software e Inteligencia Artificial. Este enfoque utiliza mediciones de similitud, primero entre palabras y después entre resúmenes, basadas en conocimiento semántico extraído de la base de datos *WordNet*. Estas métricas obtienen una similitud entre palabras, luego con ellas se calcula una similitud global entre un par de resúmenes separando las clases gramaticales (verbos, sustantivos y adjetivos). La similitud global es utilizada, por el algoritmo de categorización, para encontrar los k-vecinos más cercanos de cada resumen con la finalidad de asignar la categoría de pertenencia.

Como una evaluación de nuestro enfoque, se han presentado una serie de experimentos con seis mediciones de similitud semántica. Los experimentos han mostrado resultados prometedores para la categorización de resúmenes científicos, demostrado que la métrica de *Jiang & Conrath* resulta ser la mejor alternativa como medida de ponderación de similitud entre palabras, la cual se utiliza como base del algoritmo de categorización que presenta un desempeño promedio de 86% de resúmenes categorizados correctamente para los tres tipos de textos.

La categorización presentada en este artículo puede ayudar a los usuarios a localizar un artículo científico de manera más rápida, ya que no tendrán que buscar en todo un repositorio, por el contrario se enfocarían en una categoría específica. Como trabajo futuro, se pretende desarrollar un sistema de búsqueda de artículos científicos basada en la categorización de resúmenes o bien, proporcionando un resumen de ejemplo.

7. Bibliografía y Referencias

[1] G. A. Miller, “WordNet: a lexical database for English”. Communications of the ACM. Vol. 38. No. 11. 1995. Pp. 39-41.

[2] L. D. Baker, A. K. McCallum, “Distributional clustering of words for text classification”. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998. Pp. 96-103.

[3] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features”. European conference on machine learning. Springer Berlin Heidelberg. 1998. Pp. 137-142.

[4] F. Sebastiani, “Machine learning in automated text categorization”. ACM computing surveys. Vol. 34. No. 1. 2002. Pp. 1-47.

[5] L. H. Lee, C. H. Wan, R. Rajkumar, D. Isa, “An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization”. Applied Intelligence. Vol. 37. No. 1. 2012. Pp. 80-99.

[6] B. Tang, H. He, P. M. Baggenstoss, S. Kay, “A Bayesian classification approach using class-specific features for text categorization”. IEEE Transactions on Knowledge and Data Engineering. Vol. 28. No. 6, 2016. Pp. 1602-1606.

[7] R. Johnson, T. Zhang, “Semi-supervised convolutional neural networks for text categorization via region embedding”. Advances in neural information processing systems. 2015. Pp. 919-927.

[8] R. Johnson, T. Zhang, “Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings”. Proceedings of The 33rd International Conference on Machine Learning. 2016. Pp. 526-534.

[9] C. Corley, R. Mihalcea, “Measuring the Semantic Similarity of Texts”. Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. 2005. Pp. 1318.

[10] D. Pinto, H. Jiménez-Salazar, P. Rosso, “Clustering abstracts of scientific texts using the transition point technique”. International Conference on

Intelligent Text Processing and Computational Linguistics. Springer. 2006. Pp. 536-546.

[11] M. Alexandrov, A. Gelbukh, P. Rosso, “An approach to clustering abstracts”. International Conference on Application of Natural Language to Information Systems. Springer. 2005. Pp. 275-285.

[12] L. Waltman, N. J. Eck, “A new methodology for constructing a publication-level classification system of science”. Journal of the American Society for Information Science and Technology. Vol. 63. No. 12. 2012. Pp. 2378-2392.

[13] Z. Wu, M. Palmer, "Verbs semantics and lexical selection". Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics. 1994. Pp. 133-138.

[14] D. Lin, "An information-theoretic definition of similarity." ICML. Vol. 98. 1998. Pp. 296- 304.

[15] C. Leacock, M. Chodorow, "Combining local context and WordNet similarity for word sense identification". WordNet: An electronic lexical database. Vol 49. No. 2. 1998. Pp. 265-283.

[16] S. Banerjee, T. Pedersen, "An adapted Lesk algorithm for word sense disambiguation using WordNet". Computational linguistics and intelligent text processing. Springer Berlin Heidelberg. 2002. Pp. 136-145.

[17] P. Resnik, “Using information content to evaluate semantic similarity”. Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, Canada. 1995.

[18] J. Jiang and D. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy”. Proceedings of the International Conference on Research in Computational Linguistics. Taiwan. 1997.

[19] S. Helmut “Improvements in Part-of-Speech Tagging with an Application to German”. Proceedings of the ACL SIGDAT-Workshop, Dublin. 1995. Pp. 47-50.

[20] H. Cunningham, “GATE, a general architecture for text engineering”. Computers and the Humanities. Vol. 36. No. 2. 2002. Pp. 223-254.

[21] D. W. Aha, D. Kibler, M. K. Albert, “Instance-based learning algorithms”. Machine learning. Vol. 6. No. 1. 1991. Pp. 37-66.

[22] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, “Arnetminer: extraction and mining of academic social networks”. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 2008. Pp. 990-998.

8. Autores

El Dr. José Alejandro Reyes Ortiz obtuvo el grado de doctor en Ciencias de la Computación en el 2013 por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). Desde julio de 2013 se encuentra trabajando como Profesor Investigador en el Departamento de Sistemas de la UAM Azcapotzalco, pertenece al Área de Investigación en Sistemas de Información Inteligentes. Sus principales áreas de interés son: procesamiento de lenguaje natural, diseño y desarrollo de ontologías; aprendizaje de ontologías a partir de textos y diseño e implementación de aplicaciones de cómputo móvil.

La Dra. Maricela Claudia Bravo Contreras obtuvo el grado de doctor en Ciencias de la Computación en el 2006 por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). Desde mayo de 2011 se encuentra trabajando como Profesora Investigadora en el Departamento de Sistemas de la UAM Azcapotzalco, pertenece al Área de Investigación en Sistemas de Información Inteligentes. Sus principales áreas de interés son: composición automatizada y optimización de servicios Web públicos, diseño y desarrollo de ontologías; diseño e implementación de aplicaciones de cómputo móvil, cómputo ubicuo y sistemas sensibles al contexto.