ANALES DE LA
Academia de Ciencias de Cuba

CIENCIAS TÉCNICAS

Artículo original de investigación

Contribuciones a la minería de textos desde una perspectiva no supervisada: nuevos enfoque y métodos

Alfredo Javier Simón Cuevas [1*] https://orcid.org/0000-0002-6776-9434
Yamel Pérez Guadarrama [2] https://orcid.org/0009-0002-7408-1248
Wenny Hojas Mazo [1] https://orcid.org/0000-0002-8298-3439
José Ángel Olivas Varela [3] https://orcid.org/0000-0003-4172-4729
Francisco Pascual Romero Chicharro [3] https://orcid.org/0000-0002-6993-2434
Manuel Barreiro Guerrero [1] https://orcid.org/0009-0004-4383-3907
Eduardo Javier Valladares-Valdés [1] https://orcid.org/0000-0003-1559-0173
Manuel de La Iglesia Campos [1] https://orcid.org/0000-0002-2866-8454
Jesús Serrano Guerrero [3] https://orcid.org/0000-0002-6177-8188
Oleyda del Camino Valle [1] https://orcid.org/0009-0007-0618-6976

[1] *Universidad Tecnológica de La Habana José Antonio Echeverría. La Habana, Cuba*
[2] *Centro de Aplicaciones de Tecnologías de Avanzada. La Habana, Cuba*
[3] *Universidad de Castilla La Mancha, Ciudad Real, España*

[*] *Autor para la correspondencia:* asimon77@gmail.com

Editor

Lisset González Navarro
Academia de Ciencias de Cuba.
La Habana, Cuba

Traductor

Darwin A. Arduengo García
Academia de Ciencias de Cuba.
La Habana, Cuba

RESUMEN

Introducción: El volumen de información textual está creciendo exponencialmente. Embebido en eso textos se encuentra una gran cantidad de información relevante y conocimientos extraordinario valor para la actividad humana, a nivel económico, político, intelectual, académico, social, y otros. Por tanto, la identificación y extracción de esa información relevante, así como el descubrimiento de esos conocimientos valiosos, constituyen intereses primordiales en muchos ámbitos. En este contexto las soluciones innovadoras de minería de texto son cada vez más necesarias. **Objetivos:** Mejorar la eficacia en la recuperación de la información a través de la personalización. Concebir y desarrollar nuevas soluciones para la síntesis automática de textos que aporten mayor eficacia. Concebir y desarrollar una solución para el análisis y descubrimiento de conocimiento en los textos a partir de su representación en forma de grafos. **Métodos:** Se proponen un conjunto de soluciones para la recuperación de información personalizada. Se proponen métodos para la generación automática de resúmenes extractivos partiendo de múltiples documentos y para la extracción de frases relevantes en textos. Se concibió un modelo para el análisis computacional de textos basado en grafos de conocimiento. **Resultados:** Las contribuciones propuestas se evaluaron y validaron a nivel experimental y varias de ellas también a través de su aplicación práctica en varios escenarios. Se demostraron los aportes y beneficios de las soluciones en el procesamiento de diferentes tipos de textos, desde noticias, artículos científicos, correos, entre otros. Se lograron impactos importantes para el análisis de información en el ámbito de la seguridad y el orden interior. **Conclusiones:** Las soluciones desarrolladas representan contribuciones que arrojan

Premio Nacional Anual de la Academia de Ciencias de Cuba, 2024

resultados prometedores en el estado del arte de la minería de texto con enfoque no supervisada, con énfasis en el tratamiento de la semántica y el uso de grafos de conocimiento.

Palabras clave: minería de texto; recuperación de información personalizada; generación de resúmenes; extracción de frases relevantes; análisis de textos basada en grafos

Contributions to text mining from an unsupervised perspective: new approaches and methods

ABSTRACT

Introduction: The volume of textual information is growing exponentially. Embedded in these texts is a wealth of relevant information and knowledge of extraordinary value to human activity, economically, politically, intellectually, academically, socially and otherwise. Therefore, the identification and extraction of this relevant information, as well as the discovery of this valuable knowledge, is of paramount interest in many fields. In this context, innovative Text Mining solutions are increasingly necessary. **Objectives:** To improve the efficiency of information retrieval through the personalization. To conceive and develop new solutions for the automatic synthesis of texts that provide greater efficiency. To conceive and develop a solution for the analysis and discovery of knowledge in texts based on their representation in the form of graphs. **Methods:** A set of solutions for personalized information retrieval are proposed. Methods for the automatic generation of extractive summaries from multiple documents and for the extraction of relevant sentences in texts are proposed. A model for computational text analysis based on knowledge graphs was developed. **Results:** The proposed contributions were evaluated and validated experimentally, and several of them also through their practical application in various scenarios. The contributions and benefits of the solutions in the processing of different types of texts, from news, scientific articles, mails, among others, were demonstrated. Significant impacts were achieved for the analysis of information in the field of security and internal order. **Conclusions:** The developed solutions represent promising contributions to the state of the art of text mining with an unsupervised approach, with emphasis on the treatment of semantics and the use of knowledge graphs.

Keywords: text mining; personalized information retrieval; text summarization; key phrase extraction; graph-based text analysis

INTRODUCCIÓN

La acelerada evolución de las tecnologías de cómputo, su despliegue en el ámbito social, la globalización de la economía, y el uso intensivo de internet, ha provocado un crecimiento exponencial del volumen de contenido textual en formato digital. Embebido en eso textos se encuentra una gran cantidad de información relevante y conocimientos de extraordinario valor para la actividad humana, a nivel económico, político, intelectual, académico, social, y otros. La identificación y extracción de esa información relevante, así como descubrimiento de conocimientos valiosos, constituyen intereses primordiales en muchos ámbitos, como el empresarial, gubernamental, y académico.

La aplicación y estudio de las tecnologías computacionales para el procesamiento inteligente de textos viene ganado gran relevancia desde hace varios años. El gran volumen de contenido textual disponible y que se genera constantemente, su heterogeneidad, la no estructuración del contenido textual, y la ambigüedad inherente al lenguaje natural, provocan que el incremento de la eficacia en el procesamiento computacional de los textos constituya todavía un gran desafío. En el desarrollo de soluciones que resuelvan estos desafíos, se hace imprescindible la aplicación de tecnologías de inteligencia artificial. En este sentido, uno de los enfoques de solución es a partir de técnicas no supervisadas, las cuales tienen como ventaja no requerir procesos de entrenamiento que deman-

dan de colecciones de textos etiquetados (no suelen disponerse en escenarios prácticos), lo que permite que este tipo de soluciones sea más generalizable.

El acceso a los datos textuales es una de las tareas fundacionales de la Minería de Texto (MT), sin embargo, la sobrecarga de información a la que las personas se enfrentan cada día, dificulta llegar fácilmente a la información relevante que se necesita. A partir de este fenómeno es cada vez más creciente el interés por la personalización de la recuperación de información. Aun cuando puede que el acceso a la información de interés esté garantizado, en muchas ocasiones el volumen de contenido dificulta su análisis y la toma de decisiones basada en ella, por lo que se hacen necesarias soluciones que sinteticen de forma automática el contenido textual, por ejemplo, mediante resúmenes, tópicos o frases relevantes.

Sin embargo, las tasas de eficacia de esas soluciones aún resultan bajas. La síntesis de los textos mediante unidades de contenido a nivel de palabras o frases relevantes, también aporta mucho valor a los procesos computacionales. Este tipo de elementos pueden ser de gran utilidad para obtener una caracterización más eficaz de los documentos, aspecto muy necesario para otras tareas computacionales de MT, fundamentalmente de aquellas que dependan de la identificación de patrones característicos de los textos.

Por otra parte, para poder transformar el contenido textual en conocimiento útil no basta con tener acceso a la información, aunque esté resumida o caracterizada, es imprescindible disponer de soluciones de análisis y descubrimiento de conocimiento, ya sea de forma completamente automática o con la intervención humana. De este modo el procesamiento y análisis automático de contenidos de textos representados en forma de grafos ofrece potencialidades extraordinarias.

El uso de grafos para representar el contenido significativo de los textos permite modelar relaciones complejas entre conceptos, entidades y contextos de una manera más estructurada e intuitiva. El análisis basado en este tipo de representación facilita la extracción y descubrimiento de conocimiento profundo, mediante la identificación de patrones en las estructuras conceptuales, la obtención de relaciones jerarquías o dependencias que antes no existían, y que son difíciles de detectar en la revisión directa de los textos, sobre todo cuando se trata de múltiples documentos, o grandes volúmenes.

Además, los grafos permiten una visualización intuitiva de la información potenciando la comprensión y toma de decisiones basada en datos interconectados. En tal sentido los retos objetivos de esta investigación fueron: mejorar la eficacia en la recuperación de la información, a partir de la concepción de soluciones innovadoras que personalicen el proceso de recuperación; concebir y desarrollar nuevas soluciones para la síntesis automática de textos que aporten mayor eficacia en los resultados; concebir y desarrollar una solución para el análisis y descubrimiento de conocimiento en los textos partiendo de su representación en forma de grafos.

MÉTODOS

Modelo de recuperación de información personalizada

En la actualidad, la recuperación de información (RI) tiene una importancia significativa en la vida cotidiana de las personas debido a su integración en diversas funciones útiles como la navegación por Internet y los sistemas de búsqueda. [1] Por lo general los sistemas de búsqueda no siempre ofrecen la información que el usuario necesita, ya que generalmente no tratan la semántica del contenido, ni los intereses de los usuarios. [2] En este sentido, la personalización de la recuperación de información constituye una de las líneas fundamentales en las que se ha venido trabajando para incrementar la calidad de los resultados de los buscadores web. [3,2,4] La personalización de la recuperación de información (RIP) es un enfoque que busca adaptar los resultados de búsqueda a las necesidades, preferencias y contexto específico de cada usuario. [3] Este enfoque fue el adoptado en esta investigación, proponiéndose un nuevo modelo de recuperación de información personalizada (MRIP) orientado a un buscador Web. El MRIP se concibió en 4 procesos: procesamiento semántico de contenidos, generador de perfiles, proceso de personalización y analizador de ranking, y se enfoca en el tratamiento de la semántica, en combinación con los intereses del usuario para lograr una personalización más eficaz de la información recuperada. [5,6]

Algoritmo de *ranking* de recuperación basado en la relevancia y la calidad de los documentos

Otro de los enfoques trabajados para reducir la sobrecarga de información que generan los sistemas de búsqueda y con el objetivo de combinarlo con la personalización, es la concepción de algoritmos que mejoren la construcción del *ranking* de recuperación. [7] En este sentido, se propone un novedoso algoritmo de construcción de *ranking* que combina los intereses de los usuarios con una medición de la calidad y relevancia de los documentos, el cual fue evaluado en el ámbito de la información científica médica. En la figura 1 se muestra el flujo de trabajo general del algoritmo propuesto. A partir de una consulta de usuario enviada a un motor de búsqueda, se recupera una lista de documentos (L1) y se clasifica según criterios de relevancia. Seguidamente, los documentos son agrupados mediante algún algoritmo de *clustering* de tal

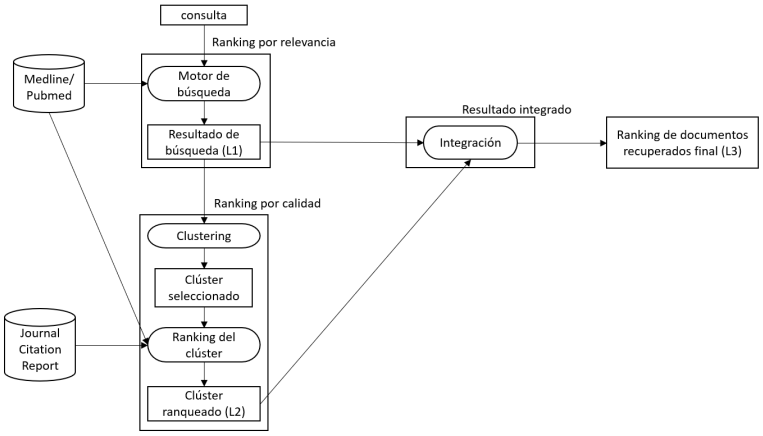

Fig. 1. Vista general del flujo de trabajo del algoritmo de ***ranking*** [7]

forma que esos grupos puedan ser examinados por el experto humano que realizó la consulta inicial. Una vez obtenidos los grupos de documentos (representan los tópicos en esa evidencia), el experto puede seleccionar el clúster LC que satisfaga sus necesidades de información. Este *clúster* debe ser evaluado según criterios de calidad, en este caso, los criterios utilizados son la relevancia de los autores, el tipo de cada documento y sus fechas de publicación. Este proceso devuelve una lista clasificada de resultados (L2). Una vez obtenidas L1 y L2, se fusionan ambas listas en una sola L3, con el objetivo de combinar ambos criterios: calidad y relevancia.

Método de generación automática de resúmenes extractivos a partir de múltiples documentos

La generación automática de resúmenes tiene el propósito de condensar en un texto corto la información más relevante y esencial contenida en uno o varios documentos de textos. [8] Esta problemática ha sido ampliamente abordada, pero sigue siendo una tarea desafiante, sobre todo cuando se quiere obtener el resumen de múltiples documentos. [9] Los resúmenes automáticos se pueden obtener mediante métodos extractivos, abstractivos, o híbridos, siendo el primero el adoptado en esta investigación. [8]

El método propuesto fue concebido en 6 fases: A) preprocesamiento, B) generación de grafos semánticos, C) integración de los grafos semánticos, D) detección de tópicos, E) evaluación de relevancia de las oraciones, y F) construcción del resumen. Este método constituye un nuevo enfoque no supervisado para la generación automática de resúmenes extractivos a partir de múltiples documentos, en el que se combina el uso de grafos semánticos, con una estrategia de evaluación de relevancia basado en el modelado de tópicos. [8,10] En ese nuevo enfoque, la conceptualización y estructura semántica subyacente del contenido de los documentos se representa mediante grafos semántico generados automáticamente, partiendo de la identificación de conceptos y relaciones semánticas entre ellos a partir de WordNet. [11] Estos grafos se integran para obtener una única representación del contenido de los múltiples documentos. Desde ese grafo se identifican los tópicos más relevantes tratados, aplicando un algoritmo de construcción de clúster (representan tópicos) y estos son utilizados para medir la relevancia de las oraciones, según su relación con esos tópicos detectados.

Método para la extracción automática de frases relevantes en textos

La identificación de información relevante dentro del gran volumen de información textual constituye una tarea desafiante, que demanda de soluciones cada vez más sofisticadas de MT. [12] En este sentido, se puede obtener una descripción de alto nivel de un documento a partir de un conjunto de palabras o frases relevantes contenidas en él. Las frases relevantes proporcionan una comprensión concisa de un texto, lo que permite captar la idea central y los tópicos principales que aborda. [13] Además, constituye una valiosa alternativa de caracterización textos, muy empleada en otras tareas de mayor alcance como la clasificación, la recuperación de información, el indexado de documentos, la generación de resúmenes, entre otras problemáticas se abordan desde la MT. [12] El método propuesto fue diseñado en cinco fases: A) preprocesamiento del texto, B) extracción de frases candidatas, C) identificación de tópicos, D) construcción de ranking de tópicos y E) selección de frases relevantes. [14,15,16,17,18]

En la primera fase se aplicaron técnicas de procesamiento de lenguaje natural (PLN) para extraer la información sintáctica del texto y el etiquetado de la parte del discurso (POS, por sus siglas en inglés) de las palabras. Este etiquetado es utilizado en la segunda fase para identificar secuencias de palabras que cumplan con un conjunto de patrones definidos, las cuales constituyen las frases candidatas. En la tercera fase se identifican los tópicos del documento mediante el agrupamiento de las frases candidatas. Cada tópico es representado por un grupo de frases candidatas. Las frases candidatas se agrupan usando como criterio el grado de relación semántica (*GRS*) entre las frases, el cual se obtiene a través de la agregación difusa de un conjunto de métricas de relación semántica. En la cuarta fase los tópicos se representan en forma de grafo para calcular un peso de relación entre ellos a partir de lo cual se crea el ranking de tópicos. En la última fase se selecciona, por cada uno de los *n* tópicos con mayor puntuación en el ranking, la frase con mayor frecuencia de aparición en el texto.

Modelo para el análisis computacional de textos basado en grafos de conocimiento

Los modelos de análisis computacional de textos desempeñan un papel fundamental en la extracción y organización de información compleja. El análisis de información textual desde la perspectiva de la TM consta de 2 etapas principales: la estructuración del contenido textual en representaciones intermedias y el análisis de las representaciones obtenidas. Uno de los enfoques más prometedores para abordar este tipo de soluciones es el uso de grafos como forma de representación, dado que estas estructuras que permiten representar conceptos, entidades y sus relaciones de manera explícita y conectada. En la concepción del modelo que se propone se adoptaron los mapas conceptuales como grafo de conocimiento para representar el contenido de los textos. [19] Los mapas conceptuales (MC) constituyen un modelo basado en grafos para representar y organizar un conjunto de significados conceptuales en una estructura de proposiciones, expresadas mediante sentencias en lenguaje natural, lo cual los hace muy intuitivos para las personas.

En la figura 2 se muestra el modelo propuesto que consta de 3 componentes principales: A) Construcción automática de mapas conceptuales, inspirado en lo reportado en Rodríguez A., Simón A y Rodríguez A *et al.*, B) CMSemQL, que extiende el lenguaje de consultas CMQL al incorporar el tratamiento de la ambigüedad presente en los procesos de búsqueda y unificación de conceptos); y C) Recuperación de pasajes, soportado por Lucene. [20,21,22,19,23,24,25,26,27] Además, se incluyen varios tipos de recursos de conocimiento externos para facilitar la integración de los grafos de conocimiento.

RESULTADOS Y DISCUSIÓN

Evaluación del modelo de recuperación de información personalizada

El modelo fue evaluado a través de su integración al buscador web denominado Red Cuba (URL: https://www.redcuba.cu/), buscador de ámbito general concebido para ser la principal fuente de acceso a información cubana (de dominio .cu) en internet. En este proceso de evaluación se simularon las interacciones de los usuarios a través de un escenario de recuperación controlado, en el que se consideraron 154 usuarios reales, el 52 % registrados. Cada usuario definió una consulta asociada a su tema de experiencia, seleccionó el conjunto de documentos más relevantes para la consulta (según lo indexado en Red Cuba), y con esta información se

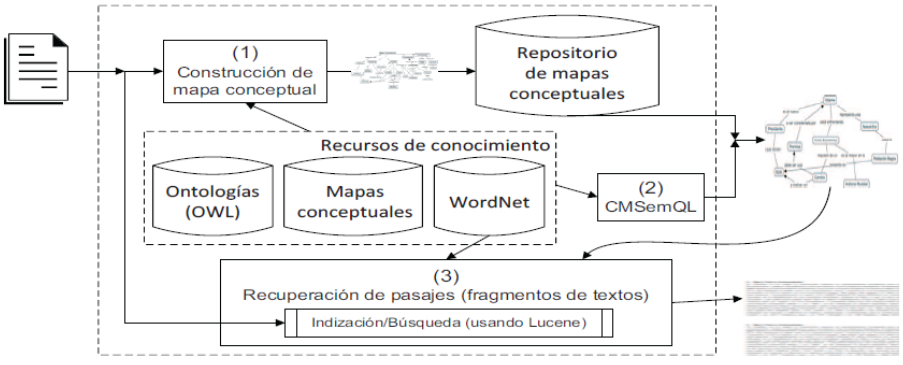

Fig. 2. Modelo para el análisis computacional de textos basado en grafos de conocimiento propuesto

construyó una colección de pruebas formada por: consultas, documentos a recuperar y perfiles de usuario. A partir de esta colección se realizaron 2 experimentos. En el primero, se evaluaron 3 escenarios: (E1) Red Cuba sin usar el MRIP, (E2) RedCuba + MRIP (considerando los perfiles registrados), y (E3) RedCuba + MRIP (sin considerar los perfiles registrados). En el segundo experimento, cada una de fases definidas en el proceso de personalización se evaluaron de forma independiente, con el objetivo de percibir con más detalle el impacto de cada una en la calidad de los resultados. Los resultados obtenidos se muestran en la tabla 1, fueron medidos considerando los 50 documentos más relevantes recuperados; similar a lo reportado en Vicente-López E *et al*. [28]

Evaluación del método de generación automática de resúmenes extractivos a partir de múltiples documentos

Esta solución propuesta se evaluó con los *corpus* de textos en español e inglés ofrecidos en el *fórum* de evaluación de tareas de resúmenes *MultiLing* 2015, específicamente en la tarea MMS (*multilingual multi-document summarization*). Los *corpus* seleccionados están formados por artículos de noticias provenientes de *WikiNews* asociados a 15 tópicos, y los mismos se han agrupado en 2 colecciones: ENCol y SPCol, respectivamente. Los resúmenes obtenidos son evaluados usando las métricas de precisión (P), *recall* (R) y medida-F (F), en el contexto de *ROUGE* (*ROUGE*-1 y *ROUGE*-2). [29] La tabla 2 muestra los resultados obtenidos y su comparación con las soluciones participantes en la tarea *MMS* de *MultiLing 2015* y otras de la literatura.

Evaluación del método para la extracción automática de frases relevantes en textos

La validación del método de extracción de frases relevantes (FTM-KPE) se diseñó en 2 escenarios: A) evaluación sobre *corpus* de referencia, donde se emplearon métricas reconocidas para medir y comparar los resultados, y B) evaluar el aporte del nuevo enfoque como parte de otras soluciones, tales como: la construcción de resúmenes extractivos y la clasificación de textos. En la evaluación sobre *corpus* se utilizaron 5 *corpus* de textos de referencia: Nus, Semeval2010, Semeval2017, Inspec, y 500N-KPCrowd, y sobre esos corpus se usaron diferentes configuraciones del método. Luego de varios experimentos se comprobó que, de todas las combinaciones evaluadas, la de mayor eficacia fue constituida por la combinación de 5 medidas con OWA, el agrupamiento con FCM y la mayor frecuencia como criterio de selección de frases candidatas. En la tabla 3 se muestra la comparación de los resultados obtenidos por la mejor configuración para el método propuesto con otros trabajos del estado del arte que mejor resultado reportan en cada uno de los corpus de evaluación seleccionados, según la medida-F.

En el segundo escenario de evaluación se aplicó FTM-KPE en la concepción de una nueva solución de generación automática de resúmenes extractivos, y en el procero de extracción de características de correos para la detección automática de *phishing*. [17] FTM-KPE permite identificar los principales tópicos de un texto representados en forma de frases relevantes, las cuales caracterizan de forma efectiva el texto. Para la evaluación de esta aplicación del método FTM-KPE Se realizaron experimentos con 5 clasificadores: Naive Bayes (NB), Ramdom Forest (RF), K-nearest neighbours (KNN), Support Vector Machine (SVM) y PART, y la base de datos pública de correos electrónicos IWSPA 2.0, con resultados muy satisfactorios. [34] Mediante este experimento se pudo comprobar los aportes del método en el aumento de la eficacia en la selección de características para ese proceso de clasificación. La aplicación de FTM-KPE en una nueva solución de resúmenes también se evaluó con MultiLing 2015 para poder establecer una línea de comparación con las soluciones en la tarea MMS (multilingual multi-document summarization), y los resultados fueron muy satisfactorios, en cuanto al aumento de la eficacia respecto a esas soluciones. [17]

Validación del modelo para el análisis computacional de textos basado en grafos de conocimiento

La evaluación del modelo propuesto se llevó a cabo a partir de su aplicación sobre diversos casos de estudios de dominios diversos, tales como, análisis de noticias, informa-

Tabla 1. Resultados experimentales

Medidas/Resultados	Primer experimento			Segundo experimento		
	E1	E2	E3	$u_i q_j$	$u_i d_j$	$u_i t_j$
Precisión	0,52	**0,80**	0,79	0,79	**0,80**	0,77
Exhaustividad	0,64	**0,93**	0,92	0,94	**0,95**	0,9
F1	0,57	**0,86**	0,85	0,85	**0,86**	0,82

Tabla 2. Resultados experimentales de la evaluación de la generación de resúmenes

Sistemas	ENCol						SPCol					
	ROUGE-1			ROUGE-2			ROUGE-1			ROUGE-2		
	P	R	F	P	R	F	P	R	F	P	R	F
SCE-Poly	0,21	0,20	0,20	0,13	0,12	0,12						
BUPT-CIST	0,12	0,11	0,12	0,02	0,01	0,02	0,13	0,12	0,12	0,03	0,03	0,03
BGU-MUSE	0,29	0,28	0,28	0,19	0,17	0,18						
NCSR/SCIFY	0,15	0,13	0,14	0,05	0,04	0,05	0,23	0,20	0,21	0,07	0,06	0,06
UJF-Grenoble	0,14	0,12	0,13	0,04	0,03	0,04						
UWB	0,34	0,33	0,33	0,19	0,18	0,18	0,41	0,39	0,40	0,26	0,25	0,25
ExB	0,23	0,23	0,23	0,08	0,09	0,08	0,25	0,24	0,24	0,10	0,10	0,10
ESIAllSummr	0,16	0,15	0,15	0,04	0,03	0,04	0,21	0,20	0,20	0,05	0,05	0,05
IDAOCCAMS	0,23	0,23	0,23	0,07	0,07	0,07	0,25	0,24	0,25	0,08	0,08	0,08
GiauUngVan	0,15	0,12	0,13	0,04	0,03	0,03						
Lead (baseline)	0,39	0,41	0,40	0,11	0,13	0,12	0,45	0,46	0,45	0,16	0,17	0,16
Solución propuesta	0,43	0,46	0,44	0,19	0,18	0,18	0,48	0,56	0,52	0,26	0,28	0,27
Características + Fuzzy [30]	0,42	0,46	0,43	0,16	0,16	0,16	0,48	0,53	0,50	0,23	0,24	0,24

ción científica, actas de reuniones, etc. Es importante destacar que para este tipo de soluciones no se han reportado esquemas de evaluación que permitan medir la calidad de los resultados. Por tanto, la aplicabilidad del modelo propuesto se llevó a cabo a través de varios casos de estudio sobre el análisis de las actas de reuniones, y en el procesamiento de artículos científicos. [19]

Conclusiones

Los métodos desarrollados representan contribuciones que arrojan resultados prometedores en el estado del arte para la minería de texto, específicamente, en el área de la recuperación de información, la generación automática de resúmenes, extracción automática de frases relevantes, y en el análisis computacional de textos. Sobre este último aspecto se logró cambien importantes contribuciones en cuanto a la construcción automáticas de grafos de conocimiento a partir de los textos y el mecanismo de consulta para extraer información y descubrir conocimiento en ellos. Desde el punto de vista práctico, se brindan herramientas que potencia el procesamiento inteligente de la información textual y facilitan el proceso de transformación de ese contenido no estructurado en conocimientos valiosos para la toma de decisiones. Se logran aplicaciones con alto valor agregado y muy beneficiosas para todos aquellos entornos donde el análisis de información textual forma parte de las tareas fundamentales.

Tabla 3. Comparación con los trabajos que mejores resultados reportan de medida-F

Soluciones	NUS	Semeval2010	Semeval2017	Inspec	500N-KPC
Ying *et al.* (2017) [32]	-	-	-	0,39	0,47
sCAKE (2019) [33]	-	0,41	-	0,51	-
HyperMatch (2023) [13]	0,41	0,36	-	0,32	-
Zhu *et al.* (2023) [34]	0,19	0,21	0,38	0,39	-
FTM-KPE (propuesta)	0,45	0,48	0,69	0,56	0,45

REFERENCIAS BIBLIOGRÁFICAS

1. Hambarde K, Proença H. Information Retrieval: Recent Advances and Beyond. IEEE Access. 2023 [Consultado jul 2024];99:1-1. Disponible en: https://ieeexplore.ieee.org/document/10184013
2. Singh A, Dey N, Ashour A, Santhi V. Web Semantics for Personalized Information Retrieval. En A. Singh N, Dey AS, Ashour V. Santhi (Eds.). Web Semantics for Textual and Visual Information Retrieval, IGI Global. 2017 [Consultado may 2018];166-86. Disponible en: https://www.igi-global.com/chapter/web-semantics-for-personalized-information-retrieval/198576
3. Liu J, Liu Ch, Belkin NJ. Personalization in Text Information Retrieval: A Survey, Journal of the Association for Information Science and Technology. 2020 [Consultado abr 2022];71(3):349-69. Disponible en: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24234
4. Utrera EB, Simón-Cuevas A, Olivas JA. Análisis de tendencias en la personalización de los resultados en buscadores web. Revista Cubana de Ciencias Informáticas. 2018 [Consultado jun 2024];12(2):111-28. Disponible en: http://scielo.sld.cu/pdf/rcci/v12n2/rcci09218.pdf
5. Utrera EB, Simón-Cuevas A, Olivas JA, Romero FP. Aproximación a un modelo de recuperación de información personalizada basado en el análisis semántico del contenido. Procesamiento del Lenguaje Natural. 2018 [Consultado jun 2024];61:31-8. Disponible en: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2018-61-3/3376
6. Utrera EB, Simón-Cuevas A, Olivas JA, Romero FP. A Personalized Information Retrieval Approach using Semantic Processing of Text Documents. Proceedings of the International Conference on Artificial Intelligence (ICAI'18). CSREA Press. 2018 [Consultado jun 2024];414-9. Disponible en: https://www.researchgate.net/publication/326522466_A_Personalized_Information_Retrieval_Approach_using_Semantic_Processing_of_Text_Documents
7. Serrano J, Romero FP, Olivas JA. A relevance and quality-based ranking algorithm applied to evidence-based medicine. Computer Methods and Programs in Biomedicine. 2020 [Consultado sep 2024];191:105415. Disponible en: https://www.sciencedirect.com/science/article/pii/S0169260719303785
8. El-Kassas WS, Salama ChR, Rafea AA, Mohamed HK. Automatic text summarization: A comprehensive survey, Expert Systems with Applications. 2021 [Consultado jul 2024];165:113679. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0957417420305030
9. Gambhir, M., Gupta V.: Recent automatic text summarization techniques: a survey. Artificial Intelligence Review. 2017 [Consultado mar 2019];47(1):1-66. Disponible en: https://dl.acm.org/doi/10.1007/s10462-016-9475-9
10. del Camino Valle O, Simón-Cuevas A, Valladares-Valdés E, Olivas JA, Romero FP. Generación de resúmenes extractivos de múltiples documentos usando grafos semánticos. Procesamiento del Lenguaje Natural. 2019 [Consultado jul 2024];63:103-10. Disponible en: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6100
11. Miller G, Fellbaum C. WordNet: An Electronic Lexical Database, The MIT Press: Cambridge, MA. 1998.
12. Rao SX, Piriyatamwong P, Ghoshal P, Nasirian S, de Salis E, Mitrović S, Wechner M, Brucker V, Egger P, Zhang C. Keyword extraction in scientific documents, arXiv preprint arXiv:2207.01888. 2022 [Consultado feb 2024]. Disponible en: https://arxiv.org/abs/2207.01888
13. Song M, Liu H, Hyperrank JL. Hyperbolic ranking model for unsupervised keyphrase extraction. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023 [Consultado feb 2024];16070-80. Disponible en: https://aclanthology.org/2023.emnlp-main.997/
14. Barreiro-Guerrero M, Simón-Cuevas A, Pérez-Guadarrama Y, Romero FP, Olivas JA. Applying OWA Operator in the Semantic Processing for Automatic Keyphrase Extraction. Lecture Notes in Computer Science. 2019 [Consultado jul 2024];11896:62-71. Disponible en: https://link.springer.com/chapter/10.1007/978-3-030-33904-3_6
15. Pérez-Guadarramas Y, Simón-Cuevas A, Hojas Mazo W, Romero FP, Olivas JA. A Fuzzy Approach to Improve an Unsupervised Automatic Keyphrase Extraction Process. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2018 [Consultado sep 2024];70-5. Disponible en: https://ieeexplore.ieee.org/document/8491487
16. Pérez Y, Rodríguez A, Simón-Cuevas A, Hojas W, Olivas JA. Combinando patrones léxico-sintácticos y análisis de tópicos para la extracción automática de frases relevantes en textos. Procesamiento del Lenguaje Natural. 2017 [Consultado sep 2024];59:39-46. Disponible en: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5491
17. Pérez-Guadarramas Y, Simón-Cuevas A, Romero FP, Olivas JA. Topic Modeling Based on OWA Aggregation to Improve the Semantic Focusing on Relevant Information Extraction Problems. G. Rivera ***et al***. (eds.), Data Analytics and Computational Intelligence: Novel Models, Algorithms and Applications, Studies in Big Data. 2023 [Consultado sep 2024];17-42. Disponible en: https://link.springer.com/chapter/10.1007/978-3-031-38325-0_2
18. Pérez-Guadarramas Y, Barreiro-Guerrero M, Simón-Cuevas A, Romero FP, Olivas, JA. Analysis of OWA operators for automatic keyphrase extraction in a semantic context. Intelligent Data Analysis. 2020 [Consultado sep 2024];24:43-62. Disponible en: https://journals.sagepub.com/doi/10.3233/IDA-200008
19. Hojas-Maz W, Simón-Cuevas A, Iglesia M, Romero FP, Olivas JA. A Concept-Based Text Analysis Approach Using Knowledge Graph. Communications in Computer and Information Science (CCIS). 2018 [septiembre 2024];854:696-708. Disponible en: https://link.springer.com/chapter/10.1007/978-3-319-91476-3_57
20. 20. Rodríguez A, Simón A. Método para la extracción de información estructurada desde textos. Revista Cubana de Ciencias Informáticas. 2013 [marzo 2024];7(1):55-67. Disponible en: http://www.scielo.sld.cu/pdf/rcci/v7n1/rcci07113.pdf
21. Rodríguez A, Simón A, Hojas W, Perea JM. Extracción de Datos Enlazados desde Información No Estructurada Aplicando Técnicas de PLN y Ontologías. CEUR-WS Proceedings Series. 2016 [septiembre 2021];1797. Disponible en: https://ceur-ws.org/Vol-1797/paper8.pdf
22. Simón A, Ceccaroni L, Rosete A, Suárez-Rodríguez A, Victoria R. A support to formalize a conceptualization from a concept map repository. En Cañas AJ, Reiska P, Ahlberg MK, Novak JD. (Eds.). Proceedings of the 3rd International Conference on Concept Mapping. 2008 [mayo 2021];1:68-75. Disponible en: https://cmc.ihmc.us/cmc2008papers/Backup/cmc2008-p291.pdf
23. Hojas-Mazo W, Simón-Cuevas A, de la Iglesia Campos M, Ruíz-Carrera JC. Semantic Processing Method to Improve a

Query-Based Approach for Mining Concept Maps. Advances in Intelligent Systems and Computing. 2019 [junio 2024];1078:22-35. Disponible en URL: https://link.springer.com/chapter/10.1007/978-3-030-33614-1_2

24. Hojas W, Simón A, Rodríguez A. Aplicación de técnicas de minería de grafos para el análisis de textos. III Congreso Internacional de Ingeniería Informática y Sistemas de Información (CIIISI 2016). La Habana, Cuba, 2016 [junio 2024]. Disponible en: https://www.researchgate.net/publication/312727056_Aplicacion_de_tecnicas_de_mineria_de_grafo_para_el_analisis_de_texto
25. Hojas W, Simón A, de la Iglesia M. Método de análisis semántico basado en WordNet para la extracción de información en mapas conceptuales. Research in Computing Science. 2016 [junio 2024];124:81-92. Disponible en: https://rcs.cic.ipn.mx/2016_124/Metodo%20de%20analisis%20semantico%20basado%20en%20WordNet%20para%20la%20extraccion%20de%20informacion.pdf
26. Hojas-Mazo W, Simón-Cuevas A, Romero FP, Olivas JA. Procesamiento Semántico Difuso Aplicado a un Modelo de Análisis de Textos basado en Grafos. Actas de XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018). 2018 [Consultado jun 2024];279-84. Disponible en: https://www.researchgate.net/publication/328463183_Procesamiento_Semantico_Difuso_Aplicado_a_un_Modelo_de_Analisis_de_Textos_basado_en_Grafos
27. Suárez LM, Hojas W, Simón-Cuevas A. Un método para la recuperación de pasajes de texto a partir de mapas conceptuales usando Lucene. XVI Congreso Internacional de Informática en la Educación (InforEdu'16). Habana, Cuba, 2016 [Consultado jun 2024]. Disponible en: https://www.researchgate.net/publication/299398019_Un_metodo_para_la_recuperacion_de_pasajes_de_texto_a_partir_de_mapas_conceptuales_usando_Lucene
28. Vicente-López E, M de Campos L, Fernández-Luna JM, Huete JF, Tagua-Jiménez A. Tur-Vigil C. An automatic methodology to evaluate personalized information retrieval systems, User Modeling and User-Adapted Interaction. 2014 [Consultado ene 2019];25(1):1-37. Disponible en: https://link.springer.com/article/10.1007/s11257-014-9148-9
29. Lin C.-Y. ROUGE: a package for automatic evaluation of summaries. En Proceedings of the ACL-04 workshop. 2004 [Consultado feb 2019];74-81. Disponible en: https://aclanthology.org/W04-1013/
30. Valladares-Valdés E, Simón-Cuevas A, Romero FP, Olivas JA. A Fuzzy Approach for Sentences Relevance Assessment in Multi-document Summarization. Advances in Intelligent Systems and Computing. 2019 [Consultado feb 2019];950:57-67. Disponible en: https://link.springer.com/chapter/10.1007/978-3-030-20055-8_6
31. Ying Y, Qingping T, Qinzheng X, Ping Z, Panpan L. A graph-based approach of automatic keyphrase extraction. Procedia Computer Science. 2017 [Consultado ene 2024];107:248-55. Disponible en: https://www.sciencedirect.com/science/article/pii/S1877050917303629
32. Duari S., Bhatnagar V. sCAKE: semantic connectivity aware keyword extraction. Information Sciences. 2019 [Consultado ene 2024];477:100-17. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0020025518308521
33. Zhu X, Lou Y, Zhao J, Gao W, Deng H. Generative non-autoregressive unsupervised keyphrase extraction with neural topic modeling. Engineering Applications of Artificial Intelligence. 2023 [Consultado feb 2024];120:105934. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0952197623001185
34. Verma RM, Zeng V, Faridi H. Data quality for security challenges: Case studies of phishing, malware and intrusion detection datasets. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019 [Consultado mar 2024]; 2605-07. Disponible en: https://dl.acm.org/doi/10.1145/3319535.3363267

Recibido: 2/10/2025
Aprobado: 2/11/2025

Agradecimientos

Aramis Rodríguez Blanco, Raudel Hernández León, José Manuel Perea Ortega, Juan Carlos Ruíz-Carrera, José Roberto Gordillo Rodríguez, Marlon García Galloso, Keny Noy Gómez, Amed Torres-Rondón, Dayana Pedroso Alfonso, Leslie Rodríguez Morejón, Luis Manuel Suárez González, Heysi Despaigne Alcántara, Eric Bárbaro Utrera Sust.

Conflictos de intereses

Puede estar dado que las figuras ya han sido publicadas en los artículos que avalan el trabajo.

Contribuciones de los autores

- Conceptualización: Alfredo Javier Simón Cuevas, Yamel Pérez Guadarramas, Wenny Hojas Mazo, José Ángel Olivas Varela, Francisco Pascual Romero Chicharro
- Curación de datos: Yamel Pérez Guadarramas, Wenny Hojas Mazo, Manuel Barreiro Guerrero, Eduardo Javier Valladares Valdés, Oleyda del Camino Valle
- Análisis formal: Alfredo Javier Simón Cuevas, Francisco Pascual Romero Chicharro
- Adquisición de fondos: José Ángel Olivas Varela, Francisco Pascual Romero Chicharro, Jesús Serrano Guerrero
- Investigación: Alfredo Javier Simón Cuevas, Yamel Pérez Guadarramas, Wenny Hojas Mazo, Manuel Barreiro Guerrero, Eduardo Javier Valladares Valdés, Manuel de La Iglesia Campos, Jesús Serrano Guerrero, Oleyda del Camino Valle
- Metodología: Alfredo Javier Simón Cuevas, José Ángel Olivas Varela, Francisco Pascual Romero Chicharro
- Administración del proyecto: Alfredo Javier Simón Cuevas
- Recursos: Alfredo Javier Simón Cuevas
- *Software:* Manuel de La Iglesia Campos, Yamel Pérez Guadarramas, Wenny Hojas Mazo, Manuel Barreiro Guerrero, Eduardo Javier Valladares Valdés, Manuel de La Iglesia Campos, Oleyda del Camino Valle
- Supervisión: Alfredo Javier Simón Cuevas, Francisco Pascual Romero Chicharro, Yamel Pérez Guadarramas, Wenny Hojas Mazo, Manuel de La Iglesia Campos
- Validación: Yamel Pérez Guadarramas, Wenny Hojas Mazo, Manuel Barreiro Guerrero, Eduardo Javier Valladares Valdés, Oleyda del Camino Valle
- Visualización: Alfredo Javier Simón Cuevas, Yamel Pérez Guadarramas, Wenny Hojas Mazo, Manuel Barreiro Guerrero, Eduardo Javier Valladares Valdés
- Redacción original: Alfredo Javier Simón Cuevas, Yamel Pérez

Guadarramas, Wenny Hojas Mazo, Manuel Barreiro Guerrero, Eduardo Javier Valladares Valdés, Oleyda del Camino Valle

- Redacción-revisión y edición: Alfredo Javier Simón Cuevas, José Ángel Olivas Varela, Francisco Pascual Romero Chicharro, Jesús Serrano Guerrero

Financiamientos

Esta investigación ha sido apoyada por el Fondo Europeo de Desarrollo Regional (FEDER) y la Agencia Estatal de Investigación (AEI) el Ministerio de Ciencia, Innovación y Universidades de España, a través del proyecto SAFER: PID2019-104735RB-C42 (AEI/FEDER, UE).

Cómo citar este artículo

Simón Cuevas AJ, Pérez Guadarrama Y, Hojas Mazo W, Olivas Varela JA, Romero Chicharro FP, Barreiro Guerrero M et al. Contribuciones a la minería de textos desde una perspectiva no supervisada: nuevos enfoque y métodos An Acad Cienc Cuba [Internet] 2025 [citado en día, mes y año];15(3):e3189. Disponible en: http://www.revistaccuba.cu/index.php/revacc/article/view/3189