

# An Indicator-based Multi-Objective Optimization Approach Applied to Extractive Multi-Document Text Summarization

J. Sanchez-Gomez, M. Vega-Rodríguez, and C. Pérez

**Abstract**—The massive amount of textual information on the Internet makes that automatic text summarization methods are becoming very important nowadays. Particularly, the purpose of extractive multi-document text summarization methods is to generate summaries from a document collection by, simultaneously, covering the main content and reducing the redundant information. In the scientific literature, these summarization methods have been addressed through optimization techniques, being almost all of them single-objective optimization approaches. Nevertheless, multi-objective approaches have gained importance because their results have improved the single-objective ones. On the other hand, in the multi-objective optimization field, indicator-based approaches have obtained good results in other applications. For this reason, an Indicator-based Multi-Objective Artificial Bee Colony (IMOABC) algorithm has been developed and applied to the extractive multi-document text summarization problem. Experiments have been carried out based on Document Understanding Conferences (DUC) datasets, and the obtained results have been evaluated and compared with Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics. The results have improved to the ones in the scientific literature between 7.37% and 40.76% and 2.59% and 11.24% for ROUGE-2 and ROUGE-L, respectively.

**Index Terms**—Extractive summary, Multi-document summarization, Multi-objective optimization, Indicator-based optimization, Hypervolume indicator.

## I. INTRODUCCIÓN

**H**OY en día, existe una gran cantidad de información en Internet que continúa creciendo cada vez más, y lo mismo sucede con el número de documentos digitales. Todo este gran volumen de datos dificulta poder obtener la información más importante de un tema específico, mientras que los usuarios de Internet demandan obtener esta información lo más rápidamente posible. Según [1], las herramientas de minería de texto tienen la capacidad de extraer la información más relevante a partir de un gran conjunto de documentos. Además, a partir de la información textual existente en los documentos, estas herramientas pueden generar, de forma automática, un resumen que satisfaga las necesidades de los usuarios en cuanto a la cobertura del contenido principal, a la vez que se reduce el número de oraciones redundantes [2].

Jesus M. Sanchez-Gomez, Departamento de Tecnología de Computadores y Comunicaciones, Universidad de Extremadura, Avda. de la Universidad s/n, 10003 Cáceres, España.E-mail: jmsanchezgomez@unex.es.

Miguel A. Vega-Rodríguez, Departamento de Tecnología de Computadores y Comunicaciones, Universidad de Extremadura, Avda. de la Universidad s/n, 10003 Cáceres, España.E-mail: mavega@unex.es.

Carlos J. Pérez, Departamento de Matemáticas, Universidad de Extremadura, Avda. de la Universidad s/n, 10003 Cáceres, España.E-mail: carper@unex.es.

Los resúmenes generados por estas herramientas pueden ser de varios tipos. En primer lugar, un resumen puede ser abstractivo o extractivo: los métodos abstractivos generan resúmenes formados por palabras u oraciones que pueden no existir en el conjunto de documentos, mientras que los extractivos seleccionan oraciones que existen en los documentos originales [3]. Además, los resúmenes pueden ser monodocumento o multidocumento en función del origen de la información: los métodos monodocumento se limitan a reducir la información de un único documento hasta una longitud establecida, y los multidocumento recopilan la información textual de todo el conjunto de documentos [4]. En definitiva, el resumen de texto extractivo multidocumento es uno de los métodos más usados para la generación de resúmenes automáticos, y su objetivo es el de resumir la información más importante de una colección de documentos, cubriendo el contenido de la misma y evitando la información redundante.

El problema de resumen de texto extractivo multidocumento ha sido estudiado desde los puntos de vista de la optimización monoobjetivo y de la optimización multiobjetivo. La optimización monoobjetivo se caracteriza por optimizar una única función objetivo, que incluye todos los criterios ponderados [5]. Sin embargo, esta asignación de pesos es subjetiva e influye en el resultado final. Por el contrario, la optimización multiobjetivo optimiza de forma simultánea y sin ponderación todas las funciones objetivo [6]. De hecho, la aplicación de la optimización multiobjetivo en el problema de resumen de texto extractivo multidocumento ha conseguido los mejores resultados en la literatura científica [7].

Las técnicas de optimización multiobjetivo pueden ser de diferentes tipos. Además de la tradicional basada en dominancia, también existe la basada en indicadores [8]. Una de las ventajas que tiene esta última es que se adapta mejor a los espacios objetivo de grandes dimensiones [9]. Además, la forma en que los indicadores de selección escogen las soluciones puede obtener mejores resultados que mediante la dominancia de Pareto [10]. Esta técnica ha sido aplicada con éxito en otros problemas del mundo real, como el de planificación del flujo de tiendas, el de ubicación y asignación de sistemas de telecomunicaciones en redes, o el de programación de tareas de enfermería [11]. Asimismo, uno de los indicadores más usados es el hipervolumen, ya que su aplicación ha conseguido mejorar los resultados en diversos casos [12].

En este trabajo, se ha desarrollado e implementado un algoritmo de colonia de abejas artificial multiobjetivo basado en indicador (IMOABC, *Indicator-based Multi-Objective Ar-*

*tificial Bee Colony*) para abordar el problema de resumen de texto extractivo multidocumento. El algoritmo IMOABC maximiza los objetivos de la cobertura del contenido y la reducción de la redundancia de manera simultánea. Los experimentos han sido llevados a cabo con los conjuntos de datos de DUC (*Document Understanding Conferences*) [13]. Para la evaluación de la calidad de los resúmenes, se han utilizado las métricas ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [14]. Esta aproximación ha conseguido mejorar los resultados obtenidos en los mejores trabajos de la literatura científica.

El resto de este trabajo está organizado de la siguiente manera. La Sección II incluye el estado del arte. En la Sección III se define el problema de resumen de texto extractivo multidocumento como un problema de optimización multiobjetivo. La Sección IV describe el algoritmo IMOABC. En la Sección V se indican los conjuntos de datos usados, la configuración de los experimentos, las métricas de evaluación empleadas y los resultados obtenidos en los experimentos. Finalmente, la Sección VI recoge las conclusiones y el trabajo futuro.

## II. TRABAJO RELACIONADO

La revisión de los trabajos relacionados con el problema de resumen extractivo multidocumento se puede realizar desde distintos puntos de vista (p.e. trabajos basados en aprendizaje profundo [15], etc.), aquí nos centramos en los trabajos basados en optimización monoobjetivo y multiobjetivo.

Primero, los trabajos monoobjetivo se analizan uno a uno. En [16], el problema de resumen de texto multidocumento fue abordado mediante métodos de agrupación de oraciones. El algoritmo usado para la resolución fue una modificación del algoritmo de optimización de enjambre de partículas (PSO).

En [17] se propuso un modelo de resumen de texto fundamentado en la programación lineal entera, que fue resuelto mediante los algoritmos de ramificación y poda (B&B) y PSO.

En [5], el problema de resumen multidocumento fue planteado como un problema de optimización discreta, proponiendo un algoritmo de evolución diferencial (DE) adaptativo para la resolución.

En [18] y [19], el problema se estudió como un problema de la  $p$ -mediana modificada, desarrollando [18] un algoritmo DE autoadaptativo para resolverlo y [19] un algoritmo de evolución diferencial basado en mutación y cruce autoadaptativos (DESAMC).

Tanto en [20] como en [21] se consideró la programación no lineal para el problema de resumen de texto, definiendo la función objetivo como la media heroniana de los criterios. En ambos, el algoritmo propuesto para la resolución fue el algoritmo PSO.

En [22] y en [23] se abordó el problema de resumen multidocumento mediante el método de la programación booleana cuadrática, que fue resuelto mediante un algoritmo DE binario en ambos casos.

En [24] se planteó el problema de resumen de documentos con el método de la programación entera cuadrática, implementando un algoritmo PSO para su resolución.

En [25], la estrategia de optimización para el resumen de documentos se basó en un algoritmo de mejorado.

En [26], el problema se planteó como un problema de optimización lineal y no lineal, empleando para su resolución el algoritmo PSO.

En [27] se abordó el resumen multidocumento como un problema de optimización binaria, y fue resuelto con un algoritmo de selección elitista cruzada, recombinación heterogénea y mutación cataclísmica (CHC).

En [28], el problema de resumen de documentos se resolvió con el método de la programación booleana mediante el algoritmo DE.

En [29], el resumen multidocumento fue planteado mediante la proximidad tópica, implementando un algoritmo DE autoadaptativo para la resolución.

En [30] se presentó un modelo basado en patrones para el problema de resumen multidocumento.

En [31], el problema de resumen de texto fue abordado mediante un modelo de selección de oraciones de dos etapas basado en técnicas de agrupación y optimización, utilizando el algoritmo DE adaptativo con una nueva estrategia de mutación.

Finalmente, en [32] se propuso un método basado en la extracción de oraciones, implementando el algoritmo de optimización basado en los patrones de caza de los tiburones (SSO). La mayor debilidad de todos estos trabajos es que la única función objetivo incluye todos los criterios ponderados de forma subjetiva.

En segundo lugar, se revisan los trabajos multiobjetivo de la literatura. El problema de rendimiento de los trabajos monoobjetivo se resuelve en los multiobjetivo, ya que en estos se optimiza cada una de las funciones objetivo sin ponderaciones.

En [6] se propuso una estrategia basada en la optimización discreta para el problema de resumen extractivo, utilizando el algoritmo genético de ordenación no dominada II (NSGA-II) para su resolución. Y en [7] se presentó un modelo centrado en el algoritmo de colonia de abejas artificial multiobjetivo basado en dominancia. Ambos trabajos serán incluidos en la comparación de resultados, demostrando que nuestro enfoque basado en indicador obtiene mejores resultados.

Por otro lado, los diferentes objetivos utilizados en estos trabajos son: la cobertura del contenido, la reducción de la redundancia, la relevancia y la coherencia. Los dos primeros han estado presentes en todos los trabajos, mientras los otros dos se han utilizado muy puntualmente. Asimismo, en todos los trabajos se han utilizado los conjuntos de datos de DUC y las métricas de evaluación ROUGE para la experimentación.

## III. DEFINICIÓN DEL PROBLEMA

En esta sección se formula el problema de resumen de texto extractivo multidocumento desde el punto de vista de la optimización multiobjetivo.

Para el caso de resúmenes de texto, los métodos más utilizados se basan en vectores de términos, dado que cada oración se representa mediante un vector de términos. Además, la similitud entre las oraciones se compara por pares mediante algún criterio, siendo la similitud coseno el más empleado (como en [5], [6] y [7]).

### A. Similitud Coseno

En una colección de documentos  $D$ , el conjunto  $T = \{t_1, t_2, \dots, t_m\}$  es el conjunto que incluye los  $m$  términos distintos. Además,  $D$  contiene un total de  $n$  oraciones, siendo cada oración  $s_i$  representada como un vector donde cada elemento es el peso de su término correspondiente:  $s_i = (w_{i1}, w_{i2}, \dots, w_{im})$ ,  $i = 1, 2, \dots, n$ . Este peso  $w_{ik}$ , ligado al término  $t_k$  de la oración  $s_i$ , se calcula mediante el esquema *frecuencia del término - frecuencia inversa de la oración* ( $tf\_isf$ ) de la siguiente forma [33]:

$$w_{ik} = tf_{ik} \cdot \log(n/n_k). \quad (1)$$

El primer factor ( $tf$ ) cuenta las veces que aparece el término  $t_k$  en la oración  $s_i$ , representado como  $tf_{ik}$ . Y en el segundo factor ( $isf$ ),  $n_k$  es el número de oraciones que contienen el término  $t_k$ , y se representa como  $\log(n/n_k)$ .

Una forma de resumir cuantitativamente el contenido principal de la colección de documentos  $D$  es mediante el vector medio  $O = (o_1, o_2, \dots, o_m)$ . Cada componente  $o_k$  de este vector es la media de los pesos del término  $t_k$  en el conjunto de oraciones, y se calcula de la manera siguiente:

$$o_k = \frac{1}{n} \sum_{i=1}^n w_{ik}, \quad k = 1, 2, \dots, m. \quad (2)$$

Por último, a partir de los pesos formulados anteriormente se puede calcular la similitud coseno. Esta similitud mide la semejanza entre un par de oraciones  $s_i = (w_{i1}, w_{i2}, \dots, w_{im})$  y  $s_j = (w_{j1}, w_{j2}, \dots, w_{jm})$  como sigue:

$$sim(s_i, s_j) = \frac{\sum_{k=1}^m w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^m w_{ik}^2 \cdot \sum_{k=1}^m w_{jk}^2}}, \quad i, j = 1, 2, \dots, n. \quad (3)$$

### B. Problema de Optimización

Tras establecer las bases matemáticas, se formula el problema de optimización. La colección de documentos  $D = \{d_1, d_2, d_3, \dots, d_N\}$ , que contiene  $N$  documentos, también se representa como  $D = \{s_1, s_2, \dots, s_n\}$ , que es un conjunto que comprende las  $n$  oraciones de la colección. El objetivo del problema es el de generar un resumen  $S \subset D$  teniendo en cuenta los siguientes tres aspectos:

- *Cobertura del contenido*: el resumen  $S$  debe cubrir el tema principal de la colección de documentos  $D$ .
- *Reducción de la redundancia*: el resumen  $S$  no debe ser redundante, esto es, las oraciones que lo forman no deben tener un contenido similar.
- *Longitud*: el resumen  $S$  debe cumplir unas restricciones de longitud  $L \pm \varepsilon$  (explicado a continuación).

La meta del problema de resumen de texto extractivo multidocumento supone la optimización de forma simultánea de la cobertura del contenido y de la reducción de la redundancia. No obstante, estos objetivos se contradicen entre sí, por lo que es necesario aplicar la técnica de optimización multiobjetivo.

Antes de formular los objetivos a optimizar, es preciso definir la representación del vector de decisión o solución  $X = (x_1, x_2, \dots, x_n)$ , siendo  $x_i \in \{0, 1\}$  una variable de

decisión binaria que considera la presencia o ausencia ( $x_i = 1$  o  $x_i = 0$ ) de la oración  $s_i$  en el resumen generado  $S$ .

El primer objetivo concierne al criterio de la cobertura del contenido,  $\Phi_{cob}(X)$ . La cobertura del contenido mide la similitud entre cada oración  $s_i \in S$  y el conjunto de oraciones de  $D$ , que se representa por el vector medio  $O$ . Por consiguiente, la siguiente función objetivo debe ser maximizada:

$$\Phi_{cob}(X) = \sum_{i=1}^n sim(s_i, O) \cdot x_i. \quad (4)$$

El segundo objetivo alude al criterio de la reducción de la redundancia,  $\Phi_{reR}(X)$ . Este objetivo necesita de otra variable de decisión binaria  $y_{ij}$  que tenga en cuenta la presencia o ausencia simultánea ( $y_{ij} = 1$  o  $y_{ij} = 0$ ) de las oraciones  $s_i$  y  $s_j$  en el resumen  $S$ . En este caso, la similitud coseno entre las dos oraciones debe ser minimizada, que es equivalente a decir que la reducción de la redundancia debe ser maximizada. Por lo tanto, la siguiente función objetivo debe ser maximizada:

$$\Phi_{reR}(X) = \frac{1}{\left( \sum_{i=1}^{n-1} \sum_{j=i+1}^n sim(s_i, s_j) \cdot y_{ij} \right) \cdot \sum_{i=1}^n x_i}. \quad (5)$$

Por último, el problema de resumen de texto extractivo multidocumento desde el punto de vista de la optimización multiobjetivo se formula de la siguiente manera:

$$\max \Phi(X) = \{\Phi_{cob}(X), \Phi_{reR}(X)\}, \quad (6)$$

$$\text{sujeto a } L - \varepsilon \leq \sum_{i=1}^n l_i \cdot x_i \leq L + \varepsilon. \quad (7)$$

$l_i$  es la longitud de la oración  $s_i$ , mientras que  $\varepsilon$  es la tolerancia para la restricción de longitud del resumen introducida anteriormente. Esta tolerancia se calcula de la siguiente forma:

$$\varepsilon = \max_{i=1,2,\dots,n} l_i - \min_{i=1,2,\dots,n} l_i. \quad (8)$$

## IV. ABC MULTI OBJETIVO BASADO EN INDICADOR

En esta sección se presenta el algoritmo ABC multiobjetivo basado en indicador (IMOABC). El algoritmo ABC estándar (*Artificial Bee Colony*) es un algoritmo que se basa en el comportamiento inteligente de una colonia de abejas. Los fundamentos de este algoritmo están detallados en [34]. El algoritmo ABC ha sido ampliamente utilizado con gran éxito para la resolución de problemas del mundo real en diferentes campos (ver p. ej. [35] y [36]), lo que ha motivado su selección como parte del estudio presentado en este trabajo.

### A. Optimización Multiobjetivo Basada en Indicador

Esta subsección define la técnica de optimización multiobjetivo empleada en este trabajo: el enfoque basado en indicador (obsérvese que en este campo se utilizan indistintamente los términos indicador y métrica).

Esta técnica se basa en una medida de rendimiento, denominada indicador  $I$ , que es usada para la selección de las soluciones del frente de Pareto [8]. A continuación se explica esta selección basada en indicador.

Primeramente, es preciso definir una serie de conceptos relacionados. Dado el conjunto  $P = \{X^1, X^2, \dots, X^{tam\_pob}\}$ , que es la población o colonia en un instante determinado, que contiene un total de  $tam\_pob$  vectores de decisión o soluciones.  $X^1$  y  $X^2$  son dos soluciones en un instante determinado de la población  $P$ , y a cada solución le corresponde un valor de aptitud  $F$ .

El indicador de selección empleado en este trabajo se basa en el concepto de hipervolumen,  $I_{HD}$  [8]:

$$I_{HD}(A^1, A^2) \begin{cases} I_H(A^2) - I_H(A^1) & \text{si } X^1 \succ X^2 \\ I_H(A^1 + A^2) - I_H(A^1) & \text{en caso contrario,} \end{cases} \quad (9)$$

donde  $A^1$  y  $A^2$  son la representación en el espacio objetivo  $Z$  de los vectores de decisión  $X^1$  y  $X^2$ , y  $X^1 \succ X^2$  significa que la solución  $X^1$  domina a la solución  $X^2$ . Además,  $I_H(A)$  es el hipervolumen del espacio objetivo dominado por  $A$ .

El operador de selección basado en indicador opera así:

1) Cálculo de los valores de aptitud. En primer lugar, es necesario calcular los valores de aptitud de cada solución siguiendo estos pasos:

- Los valores de los objetivos y del indicador se normalizan en el intervalo  $[0, 1]$ .
- Se calcula el valor del indicador  $I_{HD}(\{X^1\}, \{X^2\})$  y se obtiene el valor máximo del indicador  $c$  de la siguiente forma:  
 $c = \max_{X^1, X^2 \in P} |I_{HD}(\{X^1\}, \{X^2\})|$ .
- Para cada solución  $X^1 \in P$ , se calcula su valor de aptitud  $F(X^1)$ :

$$F(X^1) = \sum_{X^2 \in P \setminus \{X^1\}} -e^{-I_{HD}(\{X^2\}, \{X^1\})/(c \cdot \kappa)}, \quad (10)$$

donde  $\kappa$  es el factor de escala para el cálculo del valor de aptitud.

2) Aplicación de la selección basada en indicador. Los siguientes pasos se realizan hasta que el tamaño de la población  $P$  sea el tamaño inicial  $tam\_pob$ :

- Se selecciona la solución  $X^* \in P$  que tenga el valor de aptitud más pequeño.
- La solución  $X^*$  se elimina de la población  $P$ .
- Se actualizan los valores de aptitud de las soluciones restantes:

$$F(X) = F(X) + e^{-I_{HD}(\{X^*\}, \{X\})/(c \cdot \kappa)}, \quad \forall X \in P. \quad (11)$$

## B. Preprocesamiento

Previamente a la ejecución del algoritmo, es necesario realizar un preprocesamiento de la colección de documentos  $D$  que consta de las siguientes etapas:

- Segmentación de las oraciones. Cada una de las oraciones de la colección  $D$  se extrae individualmente, marcando su comienzo y su final.
- Tokenización de las palabras. Las palabras de las oraciones se separan mediante un token, como puede serlo el espacio en blanco. Además, en esta etapa se eliminan

los signos de puntuación, interrogación, exclamación, etc.

- Eliminación de las palabras vacías. Las palabras vacías, que son palabras sin significado como artículos, pronombres, preposiciones, conjunciones, etc., se eliminan de las oraciones. El paquete de las métricas de evaluación ROUGE [37] proporciona una lista de 598 palabras vacías, que será la utilizada.
- Stemming* de las palabras. A las palabras restantes se les aplica el algoritmo de Porter [38] para extraer las raíces de las mismas. Este algoritmo es el más empleado en el campo del análisis lexicográfico [39].

## C. Principales Pasos del Algoritmo

En esta subsección se detallan los pasos principales del algoritmo IMOABC. El Algoritmo 1 contiene el pseudocódigo del mismo, y seguidamente se explica su funcionamiento.

---

### Algoritmo 1 Pseudocódigo del algoritmo IMOABC.

---

```

1: Fichero_SND ← ∅
2: inicializar_colonia()
3: para ciclo = 1 de ciclos_max hacer
4:   enviar_abejas_obreras()
5:   calcular_valores_de_aptitud(tam_pob)
6:   calcular_probabilidades()
7:   enviar_abejas_observadoras()
8:   enviar_abejas_exploradoras()
9:   calcular_valores_de_aptitud(2 * tam_pob)
10:  aplicar_operador_de_seleccion(2 * tam_pob)
11:  exportar_colonia(Fichero_SND)
12: fin para

```

---

Antes de comenzar, es preciso inicializar el fichero de almacenamiento *Fichero\_SND*, que contiene las soluciones no dominadas (línea 1). Posteriormente, se inicializa la colonia de soluciones de forma aleatoria (línea 2). Esto es, las oraciones de cada solución se seleccionan aleatoriamente. Tras estas inicializaciones, los pasos de las líneas 3 a 12 se repiten durante un número máximo de ciclos *ciclos\_max* predeterminado.

En la línea 4 se envían las abejas obreras, que consiste en aplicar el operador de mutación (ver subsección siguiente) para tratar de mejorar la solución, la cual será seleccionada si mejora a la original. Tras esto se calculan los valores de aptitud de todas las soluciones de la población (línea 5).

Estos valores de aptitud se utilizan para calcular las probabilidades de selección (línea 6), necesarias para el paso siguiente. De esta forma, las mejores soluciones tendrán más probabilidad de ser seleccionadas por las abejas observadoras.

En la línea 7 se envían las abejas observadoras, que se encargan de seleccionar su abeja obrera correspondiente (teniendo en cuenta las probabilidades de selección calculadas) para intentar mejorarla mediante la aplicación de nuevo del operador de mutación.

El tamaño de la colonia se duplica en este punto del algoritmo (2 \* *tam\_pob*). En la línea 8 se envían las abejas exploradoras, las cuales verifican las soluciones que están exhaustas, que son las que, tras un determinado número de

intentos, no han conseguido mejorar más. Las abejas agotadas se sustituyen por nuevas abejas exploradoras, que son soluciones generadas aleatoriamente.

A la terminación del ciclo, se aplican los operadores de selección basados en indicador (líneas 9 y 10), y el tamaño de la colonia vuelve a ser el tamaño original (*tam\_pob*).

Finalmente, en la línea 11 las soluciones finales se almacenan en el fichero *Fichero\_SND*, aplicándose el operador de reparación en aquellas soluciones que no cumplan la restricción de longitud (ver subsección siguiente).

#### D. Operadores Principales

El algoritmo IMOABC descrito anteriormente tiene dos operadores principales: el operador de mutación y el de reparación.

En primer lugar, el operador de mutación añade o elimina oraciones en el resumen. La probabilidad de mutación  $p_m \in (0, 1)$  gobierna este operador de la siguiente manera.

Se genera un número aleatorio  $r_i \sim U(0, 1)$  para cada una de las oraciones del resumen, y si  $r_i \leq p_m$  entonces la oración  $s_i$  es candidata para la mutación. Dada una oración  $s_i$  candidata a ser mutada, si se verifica la condición siguiente:

$$\text{sim}(s_i, O) \geq \frac{1}{n} \sum_{j=1}^n \text{sim}(s_j, O), \quad (12)$$

entonces  $s_i$  se añade en el resumen, eliminándose en el caso contrario. Obsérvese que esa condición busca mantener o incorporar al resumen solo las oraciones más similares, en término medio, al contenido principal de la colección de documentos  $D$  (es decir, a su vector medio  $O$ ).

En segundo lugar, el operador de reparación se encarga de chequear que el resumen generado no quebrante la restricción de longitud establecida en la Ecuación 7.

A la terminación de cada ciclo, un resumen generado puede infringir la restricción de la longitud, por lo que esta debe ser revisada.

Si la longitud del resumen generado es menor que  $L - \varepsilon$ , se descarta el resumen (pues este hecho ocurre muy raramente), y si la longitud es mayor que  $L + \varepsilon$ , entonces se realiza la operación de reparación, que se explica a continuación.

La operación de reparación elimina la o las oraciones más redundantes del resumen a reparar  $S$ , es decir, se descartan la o las oraciones con mayor grado de concordancia entre sí.

Este grado de concordancia se mide con el umbral de similitud  $\delta$ . Al igual que en [6], se ha utilizado como umbral un valor  $\delta = 0,9$ , esto es, un grado de concordancia del 90% entre las oraciones.

Por tanto, para cada par de oraciones  $s_i$  y  $s_j$  se verifican estas tres condiciones:

$$\{s_i, s_j \in S\} \wedge \{i \neq j\} \wedge \text{sim}(s_i, s_j) \geq \delta \quad i, j = 1, 2, \dots, n, \quad (13)$$

y si todas ellas se cumplen, entonces se realiza la reparación del resumen  $S$  mediante la exclusión de la oración  $s_i$  o  $s_j$ .

La forma de elegir la oración a excluir es mediante un valor de calidad *cal*, que se calcula de la siguiente manera:

$$\text{cal}_{s_i} = \text{sim}(s_i, O) + (\text{sim}(O^S, O) - \text{sim}(O^{S-s_i}, O)) \cdot 10, \quad (14)$$

donde  $\text{sim}(O^S, O)$  es la similitud entre el centro del resumen generado ( $O^S$ ) y el vector medio  $O$ , y  $\text{sim}(O^{S-s_i}, O)$  es la similitud entre el centro del resumen generado excluyendo la oración  $s_i$  ( $O^{S-s_i}$ ) y el vector medio  $O$ .

El segundo término de la ecuación mide la calidad del resumen cuando la oración  $s_i$  es excluida, por lo que se le asigna un orden de magnitud mayor multiplicándolo por 10.

Para terminar, la oración con el valor de calidad *cal* más bajo es descartada del resumen. Esta operación de reparación se aplica hasta que se cumpla la restricción de longitud exigida.

## V. RESULTADOS

Esta sección contiene los conjuntos de datos utilizados, las métricas de evaluación empleadas, la configuración de los experimentos y el análisis y comparación de los resultados.

### A. Conjuntos de Datos

Los conjuntos de datos que se han utilizado para los experimentos han sido proporcionados por DUC (*Document Understanding Conferences*), que es un banco de pruebas abierto de referencia para la evaluación de resúmenes automáticos [13].

Las comparaciones se han realizado con los trabajos [5], [6] y [7], donde se han utilizado los 10 temas del conjunto de datos DUC indicados en la Tabla I.

TABLA I  
ESTADÍSTICAS DE LOS 10 TEMAS UTILIZADOS DEL DUC2002

Tema	Nº de palabras	Nº de términos distintos	Nº de oraciones
d061j	3679	693	184
d062j	2691	630	118
d063j	4793	846	249
d064j	4080	924	183
d065j	5500	1080	284
d066j	3894	923	190
d067f	2805	644	122
d068f	2565	531	131
d069f	7767	1306	327
d070f	3116	620	148

### B. Métricas de Evaluación

Las métricas de evaluación empleadas en este trabajo han sido las métricas ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [14]. Estas métricas están consideradas oficiales por DUC para medir la calidad de resúmenes automáticos.

Las dos variantes que se han utilizado son el ROUGE-2 y el ROUGE-L. Además de la media, también se han considerado otras dos estadísticas para la evaluación de la dispersión, como son el rango y el coeficiente de variación estadístico.

El motivo de esta elección es que algunos de los trabajos con los que se compara posteriormente solo informan sobre la media, el mejor y el peor valor ROUGE obtenido en cada tema, y mediante estas dos estadísticas es posible realizar comparaciones justas. Por un lado, el rango se calcula como la diferencia entre el mejor y el peor valor ROUGE obtenido.

TABLA II  
 MEDIAS, RANGOS Y COEFICIENTES DE VARIACIÓN (CV) DE LOS VALORES ROUGE-2. LOS MEJORES VALORES SE MUESTRAN SOMBREADOS

Tema	DE adaptativo [5]			NSGA-II [6]			MOABC [7]			Enfoque propuesto		
	Media	Rango	CV (%)	Media	Rango	CV (%)	Media	Rango	CV (%)	Media	Rango	CV (%)
d061j	0,266	0,290	109,02	0,306	0,263	85,95	0,365	0,093	25,43	0,522	0,026	4,91
d062j	0,188	0,275	146,28	0,200	0,422	211,00	0,342	0,023	6,60	0,424	0,026	6,07
d063j	0,245	0,208	84,90	0,275	0,279	101,45	0,272	0,005	1,84	0,314	0,022	6,91
d064j	0,194	0,280	144,33	0,233	0,356	152,79	0,308	0,009	2,83	0,339	0,013	3,79
d065j	0,144	0,209	145,14	0,182	0,208	114,29	0,198	0,026	13,32	0,222	0,025	11,09
d066j	0,201	0,257	127,86	0,181	0,245	135,36	0,290	0,019	6,54	0,320	0,013	3,99
d067f	0,239	0,235	98,33	0,260	0,298	114,62	0,356	0,005	1,39	0,341	0,024	7,06
d068f	0,491	0,384	78,21	0,496	0,281	56,65	0,444	0,084	18,83	0,356	0,048	13,46
d069f	0,184	0,166	90,22	0,232	0,239	103,02	0,240	0,008	3,15	0,245	0,030	12,36
d070f	0,224	0,260	116,07	0,262	0,215	82,03	0,305	0,002	0,76	0,265	0,016	6,00
<b>Media</b>	<b>0,238</b>	<b>0,256</b>	<b>114,03</b>	<b>0,263</b>	<b>0,281</b>	<b>115,72</b>	<b>0,312</b>	<b>0,027</b>	<b>8,07</b>	<b>0,335</b>	<b>0,024</b>	<b>7,56</b>

Sin embargo, el rango no es adimensional, por lo que se necesita de una adaptación del coeficiente de variación de Pearson (CV), que mide la relación entre el rango y la media en términos porcentuales de la siguiente forma:

$$CV = \frac{Rango}{Media} \cdot 100. \quad (15)$$

### C. Configuración de los Experimentos

La configuración de los parámetros del algoritmo IMOABC es la siguiente: tamaño de la población o colonia,  $tam\_pob = 64$ ; número de ciclos,  $ciclos_{max} = 1000$ ; probabilidad de mutación,  $p_m = 0,1$ ; y número de repeticiones,  $reps_{max} = 20$ .

Estos valores han sido elegidos para realizar una comparación justa con los resultados obtenidos en los trabajos de la literatura ([5], [6] y [7]).

Además, el enfoque basado en indicador necesita de otros dos parámetros: el factor de escala para el cálculo del valor de aptitud,  $\kappa = 0,05$ ; y el punto de referencia para el cálculo del indicador de hipervolumen,  $\rho = 3$ .

### D. Comparación de Resultados

Los resultados obtenidos por el algoritmo IMOABC se han comparado con los obtenidos en los siguientes trabajos: con [5], que es el mejor trabajo monoobjetivo; con [6], el que era el mejor trabajo multiobjetivo; y con [7], que es actualmente el mejor trabajo multiobjetivo.

Las Tablas II y III muestran, para el ROUGE-2 y el ROUGE-L respectivamente, la media, el rango y el CV de los resultados obtenidos en cada tema y los valores promedio para los cuatro trabajos comparados.

En primer lugar, los resultados de la Tabla II revelan que para el ROUGE-2 el enfoque propuesto obtiene mejores resultados que los otros tres trabajos: supera en 9 de 10 temas a [5] y a [6], y en 7 de 10 temas a [7], obteniendo además el mejor valor promedio.

En cuanto a los rangos y los CV, los valores mostrados indican que el enfoque propuesto es el más robusto.

Teniendo en cuenta los valores promedio, mientras que en [5] y en [6] el CV está en 114,03% y en 115,72%

respectivamente, el valor en el enfoque propuesto es de 7,56%, mejorando incluso a [7], que obtiene un 8,07%.

En los resultados de la Tabla III se puede apreciar que, de igual forma, el enfoque propuesto consigue los mejores resultados para el ROUGE-L. Gana en 9 de los 10 temas con respecto a [5], en 8 de los 10 temas con respecto a [6], y en 7 de 10 temas con respecto a [7], consiguiendo el mejor valor promedio. Considerando el rango y el CV, los valores obtenidos llevan a la conclusión de que el enfoque propuesto es el más robusto.

En los trabajos [5] y [6], el valor promedio del CV es del 39,13% y del 41,68% respectivamente, mientras que para el enfoque propuesto es del 4,02%, mejorando también el 5,43% obtenido por [7].

Finalmente, la Tabla IV muestra la mejora conseguida en términos porcentuales por el enfoque propuesto.

De los resultados de la Tabla IV se obtienen las siguientes conclusiones sobre el enfoque propuesto. Primero, que mejora a la aproximación monoobjetivo con el algoritmo DE adaptativo de [5].

El ROUGE-2 y ROUGE-L promedios se han mejorado un 40,76% y un 11,24% respectivamente. Segundo, que también mejora a la estrategia multiobjetivo con el algoritmo NSGA-II de [6].

Los valores ROUGE-2 y ROUGE-L promedios son mejorados un 27,38% y un 8,84% respectivamente.

Y tercero, que ha mejorado al enfoque presentado en [7], en el que el algoritmo ABC multiobjetivo está basado en dominancia y no en indicadores. En este caso, el ROUGE-2 promedio ha sido mejorado en un 7,37%, y el ROUGE-L promedio en un 2,59%.

Además, los rangos y los CV obtenidos demuestran que es el enfoque más robusto de todos, esto es, que es el más estable y el que menos dispersión produce. Para el ROUGE-2, la mejora en promedio de los CV es de más de un 1400% (14 veces más robusto) que [5] y [6], y en un 6,75% a [7].

Y para el ROUGE-L, la mejora del CV promedio es de en torno el 900% (9 veces más robusto) con respecto a [5] y [6], y del 35,07% con respecto a [7].

La Figura 1 muestra de forma gráfica la mejora obtenida en los valores ROUGE-2 y ROUGE-L.

TABLA III  
 MEDIAS, RANGOS Y COEFICIENTES DE VARIACIÓN (CV) DE LOS VALORES ROUGE-L. LOS MEJORES VALORES SE MUESTRAN SOMBREADOS

Tema	DE adaptativo [5]			NSGA-II [6]			MOABC [7]			Enfoque propuesto		
	Media	Rango	CV (%)	Media	Rango	CV (%)	Media	Rango	CV (%)	Media	Rango	CV (%)
d061j	0,542	0,208	38,38	0,554	0,205	37,00	0,590	0,056	9,46	0,654	0,022	3,43
d062j	0,473	0,239	50,53	0,481	0,306	63,62	0,536	0,017	3,26	0,583	0,030	5,14
d063j	0,493	0,156	31,64	0,528	0,171	32,39	0,509	0,050	9,74	0,532	0,006	1,12
d064j	0,462	0,235	50,87	0,488	0,287	58,81	0,495	0,011	2,19	0,541	0,030	5,60
d065j	0,431	0,141	32,71	0,457	0,174	38,07	0,464	0,057	12,29	0,486	0,031	6,30
d066j	0,455	0,196	43,08	0,441	0,149	33,79	0,519	0,007	1,36	0,538	0,029	5,46
d067f	0,509	0,232	45,58	0,529	0,244	46,12	0,580	0,012	2,15	0,537	0,013	2,51
d068f	0,666	0,226	33,93	0,626	0,226	36,10	0,639	0,071	11,09	0,604	0,023	3,85
d069f	0,454	0,135	29,74	0,476	0,191	40,13	0,554	0,010	1,81	0,555	0,022	3,89
d070f	0,496	0,173	34,88	0,513	0,158	30,80	0,515	0,005	0,90	0,512	0,015	2,94
<b>Media</b>	<b>0,498</b>	<b>0,194</b>	<b>39,13</b>	<b>0,509</b>	<b>0,211</b>	<b>41,68</b>	<b>0,540</b>	<b>0,030</b>	<b>5,43</b>	<b>0,554</b>	<b>0,022</b>	<b>4,02</b>

TABLA IV  
 COMPARACIÓN DE LA MEDIA, DEL RANGO Y DEL COEFICIENTE DE VARIACIÓN (CV) ENTRE EL ENFOQUE PROPUESTO Y LOS OTROS

Trabajo	Mejora del enfoque propuesto (%)					
	ROUGE-2			ROUGE-L		
	Media	Rango	CV	Media	Rango	CV
DE adaptativo [5]	40,76	966,67	1408,33	11,24	781,82	873,38
NSGA-II [6]	27,38	1070,83	1430,69	8,84	859,09	936,82
MOABC [7]	7,37	12,50	6,75	2,59	36,36	35,07

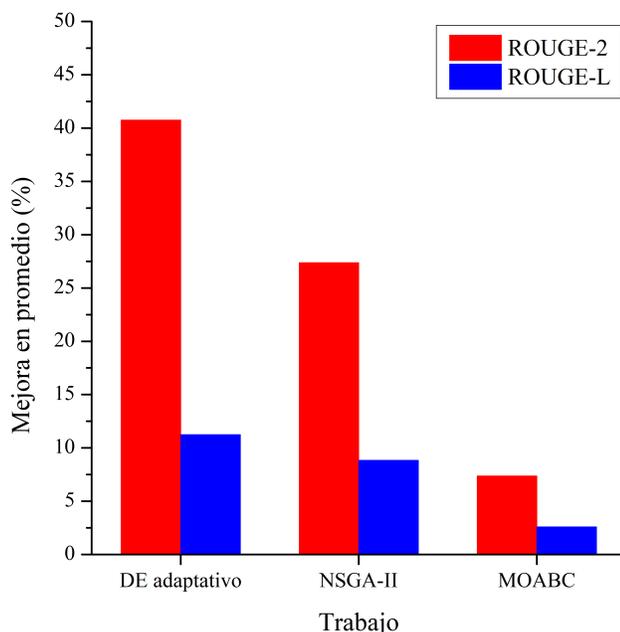


Fig. 1. Mejora en promedio de los valores ROUGE del algoritmo IMOABC con respecto a los otros trabajos.

## VI. CONCLUSIONES

El resumen de texto extractivo multidocumento es un problema que necesita de la optimización de más de una función objetivo, por lo que requiere la aplicación de técnicas

de optimización multiobjetivo. Existen múltiples formas de abordar este tipo de problemas en el campo de la optimización multiobjetivo, y una de ellas es el enfoque basado en indicador, el cual se ha popularizado en los últimos años. Por este motivo, en este trabajo se ha propuesto un algoritmo de colonia de abejas artificial multiobjetivo basado en indicador (IMOABC).

Para los conjuntos de datos utilizados, los resultados obtenidos por el algoritmo IMOABC han mejorado a los resultados de otros trabajos de la literatura científica. Se ha comparado con el mejor trabajo monoobjetivo, [5] (basado en un algoritmo adaptativo DE), con el que fue el mejor trabajo multiobjetivo, [6] (basado en el algoritmo NSGA-II), y con el mejor trabajo multiobjetivo hasta la fecha, [7] (basado en un ABC multiobjetivo basado en dominancia). La mejora conseguida en término promedio ha oscilado entre el 7,37% y el 40,76% para el ROUGE-2 y entre el 2,59% y el 11,24% para el ROUGE-L. Además, los resultados obtenidos muestran más robustez que los de los otros trabajos, demostrando que es estadísticamente más estable que ellos.

Como línea de trabajo futuro, se realizará un estudio comparativo de distintos tipos de similitudes (Jaro-Winkler, Levenshtein, Jaccard, Sorensen-Dice, etc.) además de la similitud coseno, analizando cómo cambian los resultados según la similitud usada. Además, este modelo será implementado en la plataforma de *e-learning* NeuroK<sup>1</sup> [40].

El modelo se encargará de generar resúmenes automáticos de los contenidos escritos por los estudiantes en las unidades de aprendizaje. Esto será de utilidad para los profesores con varios propósitos, como la evaluación automática de los alumnos.

## AGRADECIMIENTOS

Esta investigación ha sido apoyada por la Agencia Estatal de Investigación - España (proyectos TIN2016-76259-P y MTM2017-86875-C3-2-R), la Junta de Extremadura - España (proyectos GR18090 y GR18108) y la Unión Europea (Fondo Europeo de Desarrollo Regional). Jesus M. Sanchez-Gomez

<sup>1</sup><https://neurok.es/>

está apoyado por un Contrato Predoctoral financiado por la Junta de Extremadura - España (contrato PD18057) y la Unión Europea (Fondo Social Europeo).

#### REFERENCIAS

- [1] W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1–5, 2013.
- [2] H. Hashimi, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," *Computers in Human Behavior*, vol. 51, pp. 729–733, 2015.
- [3] X. Wan, "An exploration of document impact on graph-based multi-document summarization," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 755–762.
- [4] D. M. Zajic, B. J. Dorr, and J. Lin, "Single-document and multi-document summarization techniques for email threads using sentence compression," *Information Processing & Management*, vol. 44, no. 4, pp. 1600–1610, 2008.
- [5] R. M. Alguliev, R. M. Aliguliyev, and C. A. Mehdiyev, "Sentence selection for generic document summarization using an adaptive differential evolution algorithm," *Swarm and Evolutionary Computation*, vol. 1, no. 4, pp. 213–222, 2011.
- [6] H. H. Saleh, N. J. Kadhim, and B. A. Attea, "A Genetic Based Optimization Model for Extractive Multi-Document Text Summarization," *Iraqi Journal of Science*, vol. 56, no. 2, pp. 1489–1498, 2015.
- [7] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach," *Knowledge-Based Systems*, vol. 159, pp. 1–8, 2018.
- [8] E. Zitzler and S. Künzli, "Indicator-based selection in multiobjective search," in *International Conference on Parallel Problem Solving from Nature*. Springer, 2004, pp. 832–842.
- [9] T. Wagner, N. Beume, and B. Naujoks, "Pareto-, Aggregation-, and Indicator-based Methods in Many-Objective Optimization," in *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2007, pp. 742–756.
- [10] M. Basseur and E. K. Burke, "Indicator-based multi-objective local search," in *2007 IEEE Congress on Evolutionary Computation*. IEEE, 2007, pp. 3100–3107.
- [11] M. Basseur, A. Liefoghe, K. Le, and E. K. Burke, "The efficiency of indicator-based local search for multi-objective combinatorial optimisation problems," *Journal of Heuristics*, vol. 18, pp. 263–296, 2012.
- [12] K. Van Moffaert, M. M. Drugan, and A. Nowé, "Hypervolume-based multi-objective reinforcement learning," in *International Conference on Evolutionary Multi-Criterion Optimization*, 2013, pp. 352–366.
- [13] "Document Understanding Conferences," <http://duc.nist.gov>, 2014, último acceso: 26-Agosto-2019.
- [14] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text summarization branches out: Proceedings of the ACL-04 workshop*, vol. 8. Barcelona, Spain, 2004, pp. 74–81.
- [15] Y. Zhang, M. J. Er, R. Zhao, and M. Pratama, "Multiview convolutional neural networks for multidocument extractive summarization," *IEEE Transactions on Cybernetics*, vol. 47, no. 10, pp. 3230–3242, 2017.
- [16] R. M. Aliguliyev, "Clustering Techniques and Discrete Particle Swarm Optimization Algorithm for Multi-Document Summarization," *Computational Intelligence*, vol. 26, no. 4, pp. 420–448, 2010.
- [17] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. A. Mehdiyev, "MCMR: Maximum coverage and minimum redundant text summarization model," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14 514–14 522, 2011.
- [18] R. M. Alguliev, R. M. Aliguliyev, and C. A. Mehdiyev, "pSum-Sade: a modified p-median problem and self-adaptive differential evolution algorithm for text summarization," *Applied Computational Intelligence and Soft Computing*, vol. 2011, no. 351498, pp. 1–13, 2011.
- [19] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "DE-SAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization," *Knowledge-Based Systems*, vol. 36, pp. 21–38, 2012.
- [20] R. M. Alguliev, R. M. Aliguliyev, and C. A. Mehdiyev, "An optimization model and DPSO-EDA for document summarization," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 3, no. 5, pp. 59–68, 2011.
- [21] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "Formulation of document summarization as a 0–1 nonlinear programming problem," *Computers & Industrial Engineering*, vol. 64, no. 1, pp. 94–102, 2013.
- [22] R. M. Alguliev, R. M. Aliguliyev, and M. S. Hajirahimova, "Quadratic boolean programming model and binary differential evolution algorithm for text summarization," *Problems of Information Technology*, vol. 3, no. 2, pp. 20–29, 2012.
- [23] —, "GenDocSum+MCLR: Generic document summarization based on maximum coverage and less redundancy," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 460–12 473, 2012.
- [24] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "CDDS: Constraint-driven document summarization models," *Expert Systems with Applications*, vol. 40, no. 2, pp. 458–465, 2013.
- [25] —, "Multiple documents summarization based on evolutionary optimization algorithm," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1675–1689, 2013.
- [26] R. M. Alguliev, R. M. Aliguliyev, and C. A. Mehdiyev, "An optimization approach to automatic generic document summarization," *Computational Intelligence*, vol. 29, no. 1, pp. 129–155, 2013.
- [27] M. Mendoza, C. Cobos, E. Leon, M. Lozano, F. Rodríguez, and E. Herrera-Viedma, "A new memetic algorithm for multi-document summarization based on CHC algorithm and greedy search," in *Mexican International Conference on Artificial Intelligence*, 2014, pp. 125–138.
- [28] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "An unsupervised approach to generating generic summaries of documents," *Applied Soft Computing*, vol. 34, pp. 236–250, 2015.
- [29] K. Umam, F. W. Putro, G. Q. O. Pratamasunu, A. Z. Arifin, and D. Purwitasari, "Coverage, Diversity, and Coherence Optimization for Multi-Document Summarization," *Jurnal Ilmu Komputer dan Informasi*, vol. 8, no. 1, pp. 1–10, 2015.
- [30] J.-P. Qiang, P. Chen, W. Ding, F. Xie, and X. Wu, "Multi-document summarization using closed patterns," *Knowledge-Based Systems*, vol. 99, pp. 28–38, 2016.
- [31] R. M. Alguliev, R. M. Aliguliyev, N. R. Isazade, A. Abdi, and N. Idris, "COSUM: Text summarization based on clustering and optimization," *Expert Systems*, vol. 36, no. 1, p. e12340, 2019.
- [32] P. Verma and H. Om, "MCRM: Maximum coverage and relevancy with minimal redundancy based multi-document summarization," *Expert Systems with Applications*, vol. 120, pp. 43–56, 2019.
- [33] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [34] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Technical Report-TR06. Erciyes University, Engineering Faculty, Computer Engineering Department, Tech. Rep., 2005.
- [35] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of Global Optimization*, vol. 39, no. 3, pp. 459–471, 2007.
- [36] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: artificial bee colony (ABC) algorithm and applications," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 21–57, 2014.
- [37] ROUGE, "Summary Evaluation Package," <http://www.berouge.com/>, 2017, último acceso: 14-Julio-2017.
- [38] M. Porter, "The Porter Stemming Algorithm," <http://www.tartarus.org/martin/PorterStemmer/>, 2016, último acceso: 26-Agosto-2019.
- [39] P. Willett, "The Porter stemming algorithm: then and now," *Program*, vol. 40, no. 3, pp. 219–223, 2006.
- [40] F. Calle-Alonso, A. Cuenca-Guevara, D. de la Mata Lara, J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Perez Sanchez, "NeuroK: A Collaborative e-Learning Platform based on Pedagogical Principles from Neuroscience," in *Proceedings of the 9th International Conference on Computer Supported Education (CSEDE)*, vol. 1, 2017, pp. 550–555.



**Jesus M. Sanchez-Gomez** received the Master's degree in Informatics Engineering from the University of Extremadura, Spain, in 2018. He is conducting research in the Department of Computer and Communications Technologies, University of Extremadura, exploring the multi-document text summarization field from different approaches. His main research interests include parallel and distributed computing, evolutionary computation, and multi-objective optimization.



**Miguel A. Vega-Rodríguez** received the Ph.D. degree in Computer Engineering from the University of Extremadura, Spain, in 2003. He is currently an Associate Professor (accredited as Full Professor) of computer architecture in the Department of Computer and Communications Technologies, University of Extremadura. He has authored or co-authored more than 680 publications including journal papers (more than 140 JCR-indexed journal papers), book chapters, and peer-reviewed conference proceedings. His main research interests include parallel and

distributed computing, evolutionary computation, and bioinformatics.



**Carlos J. Pérez** is an Associate Professor (accredited as Full Professor) of Statistics in the Department of Mathematics at the University of Extremadura, Spain. He received his Master's degree in Mathematical Science in 1996 and his Ph.D. degree in Mathematics in 2003. His main research interests are in the area of Bayesian statistical inference and classification. He has authored or co-authored more than 50 JCR-indexed journal papers about statistical methodology and applications in diverse knowledge fields. He has participated in many research projects

from competitive calls and contracts.