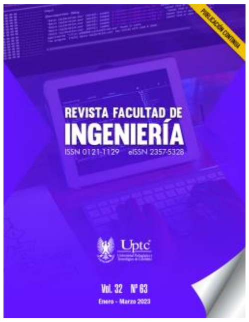
Revista Facultad de Ingeniería
Journal Homepage: https://revistas.uptc.edu.co/index.php/ingenieria



Automatic Extractive Single Document Summarization: A Systematic Mapping

Juan-David Yip-Herrera[1]
Martha-Eliana Mendoza-Becerra[2]
Francisco-Javier Rodríguez[3]

Received: November 13, 2022 **Accepted:** February 07, 2023 **Published:** February 28, 2023

Citation: J.-D. Yip-Herrera, M.-E. Mendoza-Becerra, F.-J. Rodríguez, "Automatic Extractive Single Document Summarization: A Systematic Mapping," *Revista Facultad de Ingeniería*, vol. 32 (63), e15232, 2023. https://doi.org/10.19053/01211129.v32.n63.2023.15232

Abstract

Automatic Extractive Single Document Summarization (AESDS) is a research area that aims to create a condensed version of a document with the most relevant information; it acquires more importance daily due to the need of users to obtain information on documents published on the Internet quickly. In automatic document summarization, each element must be evaluated and ranked to generate a summary. As such, there are three approaches considering the number of

[1] Universidad del Cauca (Popayán-Cauca, Colombia). juanyip@unicauca.edu.co. ORCID: 0000-0002-3206-6106
[2] Ph. D. Universidad del Cauca (Popayán-Cauca, Colombia). mmendoza@unicauca.edu.co. ORCID: 0000-0003-4033-2934
[3] Ph. D. Universidad de Granada (Granada, España). fjrodriguez@decsai.ugr.es. ORCID: 0000-0002-8965-9148

objectives they evaluate: single-objective, multi-objective, and many-objective. This systematic mapping aims to provide knowledge about the methods and techniques used in extractive techniques for AESDS, analyzing the number of objectives and characteristics evaluated, which can be helpful for future research. This mapping was carried out using a generic process for the realization of systematic reviews where a search string was built considering some research questions. A filter was then used with inclusion and exclusion criteria for selecting primary studies with which it will carry out the analysis. Additionally, these studies are sorted according to the relevance of their content. This process is summarized in three main steps: planning, execution, and result analysis. At the end of the mapping, the following observations were identified: (i) There is a preference for the use of machine learning methods and the use of clustering techniques, (ii) the importance of using both types of characteristics (statistics and semantics), and (iii) the need to explore the many-objective approach.

Keywords: automatic single document summarization; extractive; many-objective approach; systematic mapping.

Generación automática de resúmenes extractivos para un solo documento: un mapeo sistemático

Resumen

La Generación Automática de Resúmenes Extractivos para un Solo Documento (GAReUD) es un área de investigación que tiene como objetivo crear una versión corta de un documento con la información más relevante y adquiere mayor importancia a diario debido a la necesidad de los usuarios de obtener rápidamente información de documentos publicados en internet. En el área de generación automática de resúmenes cada elemento debe ser evaluado y luego rankeado para conformar un resumen, de acuerdo con esto, existen tres diferentes enfoques teniendo en cuenta la cantidad de objetivos que se evalúan, así: mono objetivo, multi objetivo y de muchos objetivos. El propósito de este mapeo sistemático es brindar conocimiento sobre los métodos y técnicas utilizadas en métodos extractivos de GAReUD, analizando la cantidad de objetivos y características evaluadas, que pueden ser útiles para futuras investigaciones. Este mapeo se realizó utilizando un proceso genérico para la realización de revisiones sistemáticas donde se construye una cadena de búsqueda considerando unas preguntas de investigación, luego se utiliza un filtro con unos criterios de inclusión y exclusión para la selección de los estudios primarios con los que se realizará el análisis, adicionalmente, estos estudios se ordenan de acuerdo con la relevancia de su contenido; este proceso se resume en tres pasos principales: Planificación, Ejecución y Análisis de resultados. Al final del mapeo se identificaron las siguientes observaciones: (i) existe una preferencia por la utilización de métodos basados en aprendizaje automático de máquina y también por el uso de técnicas de agrupamiento, (ii) la importancia de usar como objetivos ambos tipos de características (estadísticas y semánticas) y (iii) la necesidad de explorar el enfoque de muchos objetivos.

Palabras clave: enfoque de muchos objetivos; extracción; generación automática de resúmenes para un documento; mapeo sistemático.

Geração automática de resumos extrativos para um único documento: Um Mapeamento Sistemático

Resumo

A geração automática de resumos extrativos para um único documento (GAReUD) é uma área de pesquisa que visa criar uma versão curta de um documento com as informações mais relevantes e ganha cada vez mais importância devido à necessidade de os usuários obterem informações rapidamente a partir de documentos publicados na internet. Na área da geração automática de resumos cada elemento deve ser avaliado e depois classificado para formar um resumo, de acordo com isso existem três abordagens diferentes tendo em conta o número de objetivos que se avaliam, sendo assim: objetivo único, objetivo múltiplo e objetivos múltiplos. A finalidade deste mapeamento sistemático é fornecer conhecimento sobre os métodos e técnicas utilizadas nos métodos extrativos GAReUD, analisando o número de objetivos e características avaliadas, o que pode ser útil para pesquisas futuras. Este mapeamento foi realizado através de um processo genérico para a realização de revisões sistemáticas onde uma cadeia de busca é construída considerando algumas questões de pesquisa, então um filtro com critérios de inclusão e exclusão é utilizado para a seleção dos estudos primários com os quais se baseia. fora da análise, adicionalmente, esses estudos são ordenados de acordo com a relevância de seu conteúdo; esse processo se resume em três etapas principais: Planejamento, Execução e Análise de resultados. Ao final do mapeamento, foram identificadas as seguintes observações: (i) há uma preferência pelo uso de métodos baseados em aprendizado de máquina e pelo uso de técnicas de clustering, (ii) a importância do uso de ambos os tipos de características (estatísticas e semântica) e (iii) a necessidade de explorar a abordagem de muitos objetivos.

Palavras-chave: abordagem de muitos alvos; extrativo; geração de resumo automático para um documento; mapeamento sistemático.

I. INTRODUCTION

The growth of public documents available on the Web has made it necessary to develop methods to generate summaries quickly and automatically with the most relevant document information. The aim of automatic single document summarization is to provide a short version of the document while preserving the main idea of its content, applying it to multiple areas of study in which documents of different types are considered, such as news, blogs, events, emails, movies, scientific documents and text comprehension, among others [1].

Automatic Document Summarization can be classified according to how the summary is generated: abstractive, extractive, or hybrid. Abstractive summaries are generated using natural language processing techniques that modify the sentences to make the summary more coherent. Extractive summaries are generated with the original sentences of the document by selecting those most relevant. These are faster and more straightforward than abstractive summaries [2]. Meanwhile, hybrid summaries are generated considering the most significant amount of original information, making minor modifications. For this mapping, automatic extractive single document summarization (AESDS) was considered since most studies in the state-of-the-art are of this type.

The majority of AESDS research is based on methods that use statistical features to evaluate the sentences of the document. Among these are position, frequency of terms, number of keywords, and length, among others. Semantic features have also been used to evaluate the meaning and sentiment of a sentence or the semantic distance between them. These include coverage, redundancy, relevance, and similarity with the title, among others. Accordingly, it is essential to analyze which characteristics are most used and what relationship exists in using them.

Additionally, automatic document summarization has been approached considering the number of objectives through three approaches: single-objective, multi-objective (2 or 3), or many-objective (4 or more). The single-objective approach assesses one or a combination of several characteristics. No studies related to the many-objective approach for AESDS were found in this systematic mapping.

This systematic mapping aims to provide knowledge about the methods and techniques used in AESDS by analyzing the different approaches and characteristics evaluated, with the purpose of providing information that can be used for future research.

This article is developed as follows: Section 2 explains the research methodology for the systematic mapping; Section 3 shows the results obtained from the mapping; Section 4 presents a discussion based on the results obtained; and Section 5 presents the conclusions.

II. METHODOLOGY

Systematic mapping is a method that allows the identification and classification of studies containing knowledge related to a specific area, such as AESDS. This mapping was carried out based on the methodological guide proposed in [3]–[6] and in its three central stages Planning, Execution, and Result Analysis (Fig. 1).

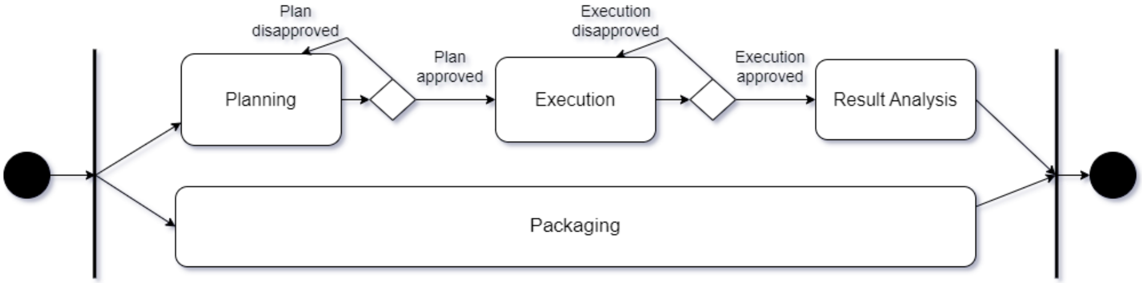


Fig. 1. Stages for systematic mapping [3].

A. Planning

This stage includes the activities described below:

1) Objectives and Research Questions. The set of research questions is designed to align with the objectives posed for this mapping, as presented in Table 1.

Table 1. Research questions.

Questions	Motivation
Q1. What types of methods have been used for AESDS recently?	Identify techniques and approaches to solve the AESDS problem.

DOI: https://doi.org/10.19053/01211129.v32.n63.2023.15232

Questions	Motivation
Q2. Which methods in AESDS have a many-objective approach?	Determine if there are methods with a many-objective approach.
Q3. What characteristics are considered for AESDS?	Identify the characteristics most used to evaluate sentences.
Q4. What types of clustering have been applied in AESDS?	Identify the clustering techniques used in AESDS.

These questions and motivations are used to define the search string used in the Science Direct, Springer Link, Scopus, and IEEE search engines.

2) Research Strategy. A search string was designed to find the primary studies (Table 2). This table omits studies whose titles contained the terms multi-document or abstractive.

Table 2. Search string.

Title, Abstract or Keywords	Title
Single AND (text OR document) AND summarization AND extractive	NOT (abstractive OR multi-document)

Then, to select the primary studies, a selection process was applied where the study title, the abstract, and the complete document were analyzed (Fig. 2).

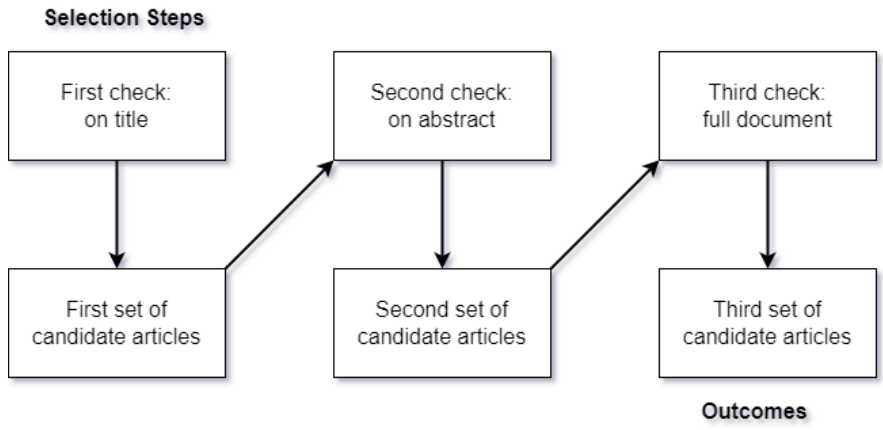

Selection Steps
First check: on title
Second check: on abstract
Third check: full document
First set of candidate articles
Second set of candidate articles
Third set of candidate articles
Outcomes

Fig. 2. Classification of files. Adapted from [4].

3) Inclusion/Exclusion Criteria. A filter was applied to the studies found considering as relevant those that met the following inclusion criteria: (i) published in English, (ii) contains the keywords defined in the search string, and (iii) falls within the date range between 2018 and 2022. Similarly, those that met any exclusion criteria were discarded: (i) not relevant to the AESDS problem, (ii) published in unrecognized journals, books, congresses, or conferences, (iii) without detailed information or access, and (iv) repeated studies.

4) Quality Assessment Criteria. In addition, to measure the quality of the primary studies, a questionnaire was defined using a 3-value scoring system (-1, 0, +1) according to their content. The first four questions are associated with the research questions, and the remaining questions analyze the importance of the articles (see Table 3). The total score for each item corresponds to the sum of values for each question, obtaining values between -6 and +6. The scores obtained indicate the studies that could be relevant in the future, and the results can be found at the following link.

Table 3. Evaluation criteria questionnaire.

Id	Question	Assigned score		
		-1	**0**	**+1**
Q1	Is the proposed method ranked in the state-of-the-art in AESDS?	Top 7+	Top 4 - 6	Top 3
Q2	How many objectives are evaluated in the proposed method?	1	2 or 3	4+
Q3	How many characteristics are considered to evaluate the sentences?	1 to 3	4 to 6	7+
Q4	In the paper, was at least one clustering technique used?	No	Partial	Yes
Q5	The impact level of the journal, book, conference, or congress	Q4	Q2 or Q3	Q1
Q6	Number of references in the study	0	1 to 10	10+

5) Execution Stage. The selection of studies was carried out in four iterations considering the query databases as shown in this link, as follows: (i) matching the search string, (ii) considering the inclusion and exclusion criteria, (iii) answering the research questions, and (iv) eliminating repeated studies. At the end of the iterations, out of the 571 studies found, 24 primary studies were obtained (Table 4).

Table 4. Selection of studies.

Database	**Studies**				
	Found	**Relevant**	**Repeated**	**Criteria**	**Primary**
Scopus	175	124	24	19	13
Science Direct	45	18	16	7	7
Springer Link	264	22	5	4	4
IEEE	87	12	3	0	0
Total	**571**	**176**	**48**	**30**	**24**

III. Results

The results obtained for each mapping research question related to the quality criteria are presented below:

Q1. What types of methods have been used for AESDS recently? A significant group of 13 studies (54%) based their proposals on machine learning (ML) [7]–[13] or hybridizing neural networks with techniques such as graphs [14], metaheuristics [15], differential evolution [16], NLP[17], [18], or genetic algorithms [19], and most of the methods present in the ranking of state of the art are among these proposals.

Q2. Which methods in AESDS have a many-objective approach? From the set of 24 studies obtained in the execution stage and from the results obtained from the evaluation questionnaire, it is evident that there are no studies where a many-objective approach is considered. Thirteen studies (54%) are multi-objective, while the remaining 11 (46%) are single-objective.

Q3. What characteristics are considered for AESDS? In the studies reviewed, one feature or the combination of several features (naive multi-objective approach) can be evaluated in a single objective. From these, the most commonly used features were: sentence position 54%, sentence length 46%, title similarity 29%, proper nouns 29%, TF-IDF measure 29%, numerical data 21%, keywords 17%, coherence 17%, antiredundancy 17%, sentiment (positive or negative) 17%, aggregate similarity 13%, diversity 8%, readability 8%, coverage 8%, date-type data 4%, centrality 4%, among others such as number of verbs, importance, number of entities, relevance, number of verbs, bigrams, and trigrams.

Q4. What types of clustering have been applied in AESDS? It was found that 16 studies (64%) included the use of some grouping or clustering technique such as K-means [13], [20], [21]; K-medoids [15], [22], [23]; neural network-based clustering, e.g., self-organizing maps (SOM) [7]–[9], [15], [16]; topic detection such as LDA [14]; clustering by keyword identification [11], [18], [24] or using non-negative factorization matrices [12], [25].

IV. Discussion

This AESDS systematic mapping identified a preference for machine learning methods or hybridizing with machine learning. These methods report excellent results and are well-placed within the state-of-the-art ranking. For example, in [15], [16], self-organizing maps (SOM neural networks) combined with metaheuristics are used, and in [14], neural networks are used together with graphs, among others.

Although semantic features evaluate the meaning and sentiment of a sentence or the semantic distance between sentences, a high percentage of research found uses statistical features (sentence position, sentence length, proper names, numerical data), showing that these are still useful for the generation of extractive summaries. In addition, the best state-of-the-art methods [16], [21], [26] show the importance of considering both types of features (statistical and semantic) independently or combined.

Concerning the many-objective approach, no previous study considered four or more objectives, providing a space for future research using this approach.

Similarly, there is growing attention to the inclusion of clustering techniques in AESDS methods, also achieving an improvement in the results compared to other similar or equal methods in the state of the art that do not include these techniques.

V. Conclusions

This AESDS mapping identified specific methods that tend to obtain the best results from the state-of-the-art. These are machine learning-based or hybrids that include some clustering techniques.

It was also found that statistical features remain relevant for assessing the quality of summaries in AESDS methods, as most of the methods found in this systematic mapping use this type of feature.

In future work, we propose to address the AESDS problem with a many-objective approach because, in this research area, there are combinations of features that can be evaluated independently. In addition, we intend to continue mapping this research area because it is very active, and new articles are constantly being published. We further hope to consider mapping the automatic generation of summaries for multiple documents and abstractive methods for summarization.

ACKNOWLEDGMENTS

The Universidad del Cauca (Colombia) partially supported this research.

AUTHORS' CONTRIBUTIONS

Juan-David Yip-Herrera: Conceptualization, Methodology, Resources, Writing - original draft.

Martha-Eliana Mendoza-Becerra: Supervision, Conceptualization, Methodology, Resources, Writing - original draft, Project administration.

Francisco-Javier Rodríguez: Supervision, Writing - editing and review.

REFERENCES

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, e113679, 2021. https://doi.org/10.1016/j.eswa.2020.113679

[2] A. Nenkova, K. McKeown, "A Survey of Text Summarization Techniques," in *Mining Text Data*, Boston, MA: Springer US, 2012, pp. 43–76.

[3] P. Mian, T. Conte, A. Natali, J. Biolchini, G. Travassos, "A systematic review process to software engineering," *ESELAW*, vol. 32, 2005.

[4] T. Marew, J. Kim, D. H. Bae, "Systematic Mapping Studies in Software," *International Journal of Software Engineering and Knowledge Engineering*, vol. 17, no. 1, pp. 33–55, 2007.

[5] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, "Systematic literature reviews in software engineering - A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, 2009. https://doi.org/10.1016/j.infsof.2008.09.009

[6] O. Kaiwartya *et al.*, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *IEEE Access*, vol. 4, pp. 5356–5373, 2016. https://doi.org/10.1109/ACCESS.2016.2603219

[7] M. Gambhir, V. Gupta, "Deep learning-based extractive text summarization with word-level attention mechanism," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 20829–20852, 2022. https://doi.org/10.1007/s11042-022-12729-y

[8] X. Han, Q. Wang, Z. Chen, L. Hu, P. Hu, "OnSum: Extractive Single Document Summarization Using

Ordered Neuron LSTM," *Lecture Notes in Computer Science*, vol. 12837, pp. 605–615, 2021. https://doi.org/10.1007/978-3-030-84529-2_51

[9] M. Rahul Raj, R. P. Haroon, N. V Sobhana, "A novel extractive text summarization system with self-organizing map clustering and entity recognition," *Sadhana.*, vol. 45, no. 1, e32, 2020. https://doi.org/10.1007/s12046-019-1248-0

[10] A. Joshi, E. Fidalgo, E. Alegre, L. Fernández-Robles, "SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders," *Expert Systems with Applications*, vol. 129, pp. 200–215, 2019. https://doi.org/10.1016/j.eswa.2019.03.045

[11] A. Qaroush, I. Abu Farha, W. Ghanem, M. Washaha, E. Maali, "An efficient single document Arabic text summarization using a combination of statistical and semantic features," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 677–692, 2021. https://doi.org/10.1016/j.jksuci.2019.03.010

[12] A. Khurana, V. Bhatnagar, "Investigating Entropy for Extractive Document Summarization," *Expert Systems with Applications*, vol. 187, e115820, 2022. https://doi.org/10.1016/j.eswa.2021.115820

[13] S. Agarwal, N. K. Singh, P. Meel, "Single-Document Summarization Using Sentence Embeddings and K-Means Clustering," in *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking,* 2018, pp. 162–165. https://doi.org/10.1109/ICACCCN.2018.8748762

[14] A. Joshi, E. Fidalgo, E. Alegre, R. Alaiz-Rodriguez, "RankSum—An unsupervised extractive text summarization based on rank fusion," *Expert Systems with Applications*, vol. 200, e116846, 2022. https://doi.org/10.1016/j.eswa.2022.116846

[15] N. Saini, S. Saha, A. Jangra, P. Bhattacharyya, "Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm," *Knowledge-Based Systems*, vol. 164, pp. 45–67, 2019. https://doi.org/10.1016/j.knosys.2018.10.021

[16] N. Saini, S. Saha, D. Chakraborty, P. Bhattacharyya, "Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures," *PLoS One*, vol. 14, no. 11, e0223477, 2019. https://doi.org/10.1371/journal.pone.0223477

[17] F. S. Tabak, V. Evrim, "Event-based summarization of news articles," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 2, pp. 850–864, 2020. https://doi.org/10.3906/elk-1904-98

[18] N. Kindo, G. Bhuyan, R. Padhy, *A New Technique for Extrinsic Text Summarization*, Springer, 2019.

[19] K. Arai, S. Kapoor, R. Bhatia, *Single Document Extractive Text Summarization Using Neural Networks and Genetic Algorithm*, Cham: Springer International Publishing, 2019.

[20] A. Sharaff, M. Jain, G. Modugula, "Feature based cluster ranking approach for single document summarization," *International Journal of Information Technology*, vol. 14, no. 4, pp. 2057–2065, 2022. https://doi.org/10.1007/s41870-021-00853-1

[21] W. S. El-Kassas, C. R. Salama, A. A. Rafea, H. K. Mohamed, "EdgeSumm: Graph-based framework for automatic text summarization," *Information Processing & Management*, vol. 57, no. 6, e102264, 2020. https://doi.org/10.1016/j.ipm.2020.102264

[22] R. Srivastava, P. Singh, K. P. S. Rana, V. Kumar, "A topic modeled unsupervised approach to single document extractive text summarization," *Knowledge-Based Systems*, vol. 246, e108636, 2022.

https://doi.org/10.1016/j.knosys.2022.108636

[23] S. Kumar, M. Naveen, S. Sriparna, S. Pushpak, *Scientific document summarization in multi-objective clustering framework*," 2021.

[24] X. Mao, H. Yang, S. Huang, Y. Liu, R. Li, "Extractive summarization using supervised and unsupervised learning," *Expert Systems with Applications*, vol. 133, pp. 173–181, 2019. https://doi.org/10.1016/j.eswa.2019.05.011

[25] A. Khurana, V. Bhatnagar, "Extractive Document Summarization using Non-negative Matrix Factorization," in *Lecture Notes in Computer Science*, vol. 11707, pp. 76–90, 2019.

[26] D. Debnath, R. Das, P. Pakray, "Extractive single document summarization using multi-objective modified cat swarm optimization approach: ESDS-MCSO," *Neural Computing and Applications*, vol. 4, e06337, 2021. https://doi.org/10.1007/s00521-021-06337-4