

## Document classification using term frequency-inverse document frequency and K-means clustering

Wasseem N. Ibrahim Al-Obaydy<sup>1</sup>, Hala A. Hashim<sup>2</sup>, Yassen AbdulKhaleq Najm<sup>3</sup>, Ahmed Adeeb Jalal<sup>1</sup>

<sup>1</sup>Department of Computer Engineering, College of Engineering, Al-Iraqia University, Baghdad, Iraq

<sup>2</sup>Department of Dentistry, Dijlah University College, Baghdad, Iraq

<sup>3</sup>Department of English, College of Arts, Al-Iraqia University, Baghdad, Iraq

### Article Info

#### Article history:

Received Feb 23, 2022

Revised Jun 16, 2022

Accepted Jun 29, 2022

#### Keywords:

Data mining

Document classification

K-means clustering

TF-IDF

Topics

### ABSTRACT

Increased advancement in a variety of study subjects and information technologies, has increased the number of published research articles. However, researchers are facing difficulties and devote a significant time amount in locating scientific research publications relevant to their domain of expertise. In this article, an approach of document classification is presented to cluster the text documents of research articles into expressive groups that encompass a similar scientific field. The main focus and scopes of target groups were adopted in designing the proposed method, each group include several topics. The word tokens were separately extracted from topics related to a single group. The repeated appearance of word tokens in a document has an impact on the document's weight, which is computed using the term frequency-inverse document frequency (TF-IDF) numerical statistic. To perform the categorization process, the proposed approach employs the paper's title, abstract, and keywords, as well as the categories' topics. We exploited the K-means clustering algorithm for classifying and clustering the documents into primary categories. The K-means algorithm uses category weights to initialize the cluster centers (or centroids). Experimental results have shown that the suggested technique outperforms the k-nearest neighbors algorithm in terms of accuracy in retrieving information.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



### Corresponding Author:

Ahmed Adeeb Jalal

Department of Computer Engineering, College of Engineering, Al-Iraqia University

Baghdad, Iraq

Email: ahmedadeeb@aliraqia.edu.iq

## 1. INTRODUCTION

Web document clustering is a data mining technique that is used to gather a collection of documents having comparable content from a series of documents distributed over several websites [1]-[3]. Document clustering aims to locate and comprehend the documents [4] allowing comparable articles to be grouped in one place. Consequently, documents on the web can be categorized based on a set of subjects for each group. These subjects concentrate on special terms called word tokens that may be discovered through the analysis process of the document. The term "word tokens" denotes the frequency of specific idioms in papers, and the extraction of terms from textual data may assist in documents classification [5], [6]. As a result, the classification of documents into a set of categories can be achieved by a cluster of terms depending on how many times each word token appears for a specific subject in those papers [7].

Researchers who conduct multidisciplinary research on various topics can utilize documents classification. Usually, document and topic clustering may be exploited to achieve this objective [8]-[10].

All relevant documents can be clustered in a single category, and the search can be narrowed down to only the most significant documents selected by the user. Traditional search methods, on the other hand, make it difficult and time-consuming to find relevant documents for scholars particularly given the constant rise in the number of documents. Furthermore, there is a wide range of significant sources of documents accessible to users through the internet, including research papers, technical reports, webpages, digital repositories, and archives.

Nowadays, the Internet has become the primary source of information for a large number of individuals. As a result, the users should be able to quickly and readily find their relevant requests that reflect the queried information [11], [12]. However, depending on a limited number of terms in a user's query, the search engine returns more irrelevant sites, leading to extensive lists of URLs. The searching process through webpages for information that matches a user's query is not a trivial task taking into account the issue of data overload that encounters Internet data warehouses. To overcome this problem, web data mining may be adopted to develop techniques that discover and retrieve relevant information [13], [14]. The field of web data mining focuses on finding patterns on the Internet. There are three kinds of web data mining including web structure mining, web content mining, and web usage mining. Each type involves a variety of approaches to discover the information patterns [15]. Accordingly, data mining techniques can be exploited to enhance the ability of search engines to discover specific information in a huge amount of data [16], [17].

A tremendous amount of research papers in several scientific fields have been published by institutions, universities, and journals during the past decades. However, these research articles are not clustered or classified into groups making their retrieval a difficult task. Many documents clustering approaches have been proposed to classify research articles depending on the features or attributes of documents content [18], [19]. The key differences among these techniques can be expressed in several aspects, for example, the types of features that are extracted to represent the papers, the measure of similarity, and the cluster representation. The following paragraphs present a literature review of the state-of-the-art approaches of research articles classification and their applications.

Buatoom *et al.* [20] proposed similarity-based constraints for the K-means clustering method based on the class information and collection distributions extracted from labeled data. The authors extracted three types of distribution statistics, namely inter-class term distribution, intra-class term distribution, and in-collection term distribution from labeled data to direct the clustering process towards the user preference. The proposed method captures behavioral patterns using statistics for clustering rather than relying on the prior knowledge of labeled data. The authors evaluated the efficacy of term weighting on clustering using five text datasets and three types of measurement, namely class-based, cluster-based, and similar-based measures.

Alsuhaim *et al.* [21] presented a clustering system that utilizes the enhanced K-means algorithm to cluster Arabic search results. In this approach, each cluster is labeled with the most recurrent word in the cluster. The proposed system is composed of seven stages: snippet extraction from the search engine, snippet text preprocessing, text features extraction, number of clusters estimation, applying the enhanced K-means clustering algorithm, clusters evaluation, and cluster's label creation. The system is intended to assist Arabic web users to identify each cluster's topic and access the required cluster directly.

Moreover, other mining techniques have been used to classify web documents. In [22], [23], the authors proposed text mining methods based on natural language processing. Other approaches [24], [25] were developed based on representing the articles semantically from their accompanying entities. The numerical statistic of term frequency-inverse document frequency was used in [26], [27] to determine the importance of a word to a document in a collection. These methods play the role of a weighting factor in the fields of text mining, information retrieval, and user modeling. Consequently, document clustering and classifying are considered essential to achieve user satisfaction and ease the documents retrieval process.

In this paper, we focus on classifying and clustering the research papers into groups to eliminate the search problems for the research community. According to the authors in [28], and [29], the main objective of clustering is to provide improved coverage and avoid complexity when used with research papers as well as other various domains. Thus, the main contribution of this work is to propose a research papers classification system based on term frequency-inverse document frequency (TF-IDF), and K-means clustering, to assist the researchers to find the relevant research papers in their field of expertise. The privileges of the proposed text documents clustering approach in this paper are outlined as follows. Firstly, it has a substantial influence to find useful information. Secondly, it addresses the shortage of comprehensibility of search engines. Finally, it improves search-ability for the researchers.

The remaining part of this article is structured as follows. Section 2 presents a detailed explanation of the methodologies used in the proposed text documents classification method, for example, web data mining, data extraction, TF-IDF, and K-means clustering. These approaches analyze scientific articles through data extraction to classify the articles based on the similarity score. Section 3 reports the

experimental results of the presented document classification approach and the techniques employed in its design. Lastly, the conclusion of the paper is outlined in section 4.

**2. PROPOSED METHOD**

In this section, a detailed description of the proposed document classification framework for clustering the research articles is presented. The presented method is intended to tackle the time-consuming search problem that encounters researchers when identifying the related cluster of the desired papers. Typically, existing methods follow the traditional approach of classifying the research articles into clusters according to concepts and contents. However, our method in this article employs three research components, namely title, abstract, and keywords to accomplish the clustering process. There are many reasons for adopting the abstract after the title in designing our classification approach. Firstly, the abstract represents the most significant portion that outlines the paper’s essence [30], [31]. Secondly, it is the next section that readers read frequently. Moreover, it involves a rich set of keywords and terms that describe the research direction of the paper. Finally, it summarizes the contents of the paper.

We collected a dataset of 518 papers published by the journal entitled Bulletin of Electrical Engineering and Informatics (BEEI) between 2012 and 2019. These articles are typed in English and cover various topic scopes. We aim to categorize the papers into five clusters depending on the topic scope of the journal, as demonstrated in Figure 1.

According to the literature review in the previous section, the existing research works adopted the user’s query, semantic representation, or other techniques to categorize and retrieve articles. In our work, we extract the topical contents from all papers by applying the basic crawler algorithm separately to each cluster, in addition to the title, abstract, and keywords. We propose to extract a list of word tokens based on the topics of each cluster to classify the papers. Different statistical methods such TF-IDF, K-means algorithm, and K-nearest neighbors (KNN) algorithm were used in the classification step. The diagram in Figure 2 demonstrates the proposed document classification approach.

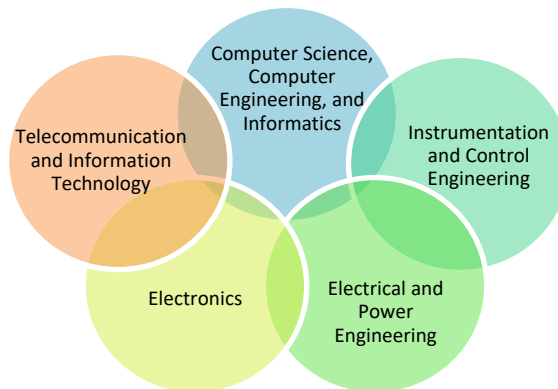


Figure 1. The five clusters of the topic scope

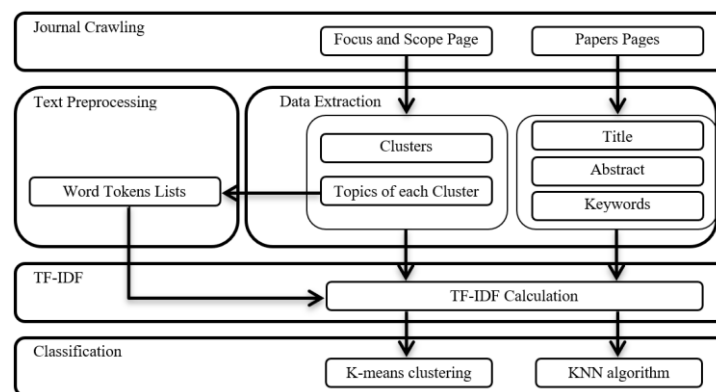


Figure 2. The proposed document classification approach

## 2.1. Text preprocessing

One of the key components of text mining algorithms is text preprocessing. The main tasks of the text preprocessing step include tokenization, filtering, lemmatization, and stemming [32]. Usually, the performance of the clustering algorithm can be reinforced based on the type of attributes, such as words, terms, and phrases that are extracted from the documents.

As can be seen in Figure 2, the role of the text preprocessing step is to extract word token lists using the aforementioned tasks. The purpose of the tokenization task is to break the series of characters in crawled topics into fragments called word tokens. The filtering step processes the lists of word tokens via removing comparable and stop words to decrease the amount of indexing in addition to improve the accuracy rates. The lemmatization task analyzes the words morphologically by grouping several related words so they can be analyzed as a single item. The goal of the stemming task is to generate the stem (root) of extracted words. As a result, we get five sets of word token lists based on the topics of the clusters, i.e., one list per cluster.

## 2.2. Term frequency-inverse document frequency (TF-IDF)

In the field of information retrieval, the most popular numerical and descriptive statistical technique is called TF-IDF that is widely used as a factor of weighting. The weighting mechanism of TF-IDF describes how important words are based on their presence in documents content. As a result, the TF-IDF can be employed for various tasks, such as the extraction of word tokens from articles, calculating the similarity degrees among documents, determining significant ranking, and others. In the proposed method, three statistical measures, namely TF, IDF, and TF-IDF were computed for each word token in the lists in both clusters and documents.

The term frequency (TF) is a metric that measures how frequently certain words appear in the content of a document. The TF may be defined as in (1). The high value of TF refers to a more important word in the document.

$$TF_{t,d} = \frac{f_{t,d}}{\max\{f_{t',d}: t' \in d\}} \quad (1)$$

Where,  $f_{t,d}$  denotes the recurring appearance of a word/term  $t$  in the document  $d$ .

The inverse document frequency (IDF), on the other hand, calculates the infrequency and significance of a word/term across all documents. The IDF is defined as (2). The high value of IDF refers to an infrequent word in all documents.

$$IDF_{t,D} = \log \frac{D}{\{d \in D: t \in d\}} \quad (2)$$

Where,  $IDF_{t,D}$  is a logarithmic scale that divides the total number of documents  $D$  by the number of documents that include the word/term  $t$ .

Finally, the (3) defines the TF-IDF weighting. When the frequency of a certain word/term in a document is high and the number of documents that include the word/term is low, the value of TF-IDF weighting rises.

$$TF - IDF = TF_{t,d} \times IDF_{t,D} \quad (3)$$

## 2.3. K-means clustering

One of the most effective approaches to unsupervised learning is the K-means algorithm. This technique uses iterative clustering for grouping a set of objects in K clusters based on their attributes. In our approach, the documents are clustered and then the centroid of each cluster is specified. The average mean of the objects in the cluster is used to define the centroid. The major step in the algorithm is the assignment of centroid to each cluster. The centroid's assignment to each cluster is accomplished by applying Euclidean distance-based similarity function to all objects in that cluster. The K-means clustering strategy is described by the following steps:

- The cluster centers (or centroids) are initialized using the K-means algorithm by category weights. In our paper, we have five clusters  $C$ . Utilizing Euclidean distance, each of the training documents  $d$  is assigned to their nearest centroids as (4).

$$W(C) = \sum_{i=1}^n \sum_{d \in C_i} \|d - m_i\|^2 \quad (4)$$

- The centroid of each of the  $k$  clusters is estimated as the average of the training documents  $d$  allocated to that cluster, as (5).

$$m_i^{(i+1)} = \frac{1}{|c_i|} \sum_{d \in c_i} d \tag{5}$$

- If the difference between the new centroids and the previous centroids is remarkable, return to step 2 and repeat until convergence.
- It's worth noting that K-means is a repetitive process for recalculating centroids and reassigning training instances to the clusters represented by these centroids until the centroids converge.

### 3. RESULTS AND DISCUSSIONS

In this section, we present a brief explanation of the collected research papers dataset and evaluate the performance of the proposed document classification approach in comparison with the existing recent approaches. A set of 518 research papers published by the BEEI Journal in several subjects and scopes was collected. The papers were selected from different scientific disciplines, for example, computer science, informatics, information technology, electronics, electrical, power engineering, computer engineering, control engineering, telecommunication, and instrumentation. Many topics, for example, programming, computer security, computer architecture, electrical engineering materials, electronic materials, microelectronic systems, distributed platforms, wave propagation and antenna, and robotics are included in each of these scopes. We aim to classify the extracted papers into five categories according to these domains. As mentioned previously in section 2, the result of the crawling process applied on the title, keywords, and abstract of each article was used as essential data for classification. In the meantime, five lists of word tokens were extracted from the topics of scopes. The whole corpus, then, was passed as input to the TF-IDF calculation module to compute the weight of each word token for both clusters and papers. Some examples are illustrated in Table 1. The last step in the proposed approach consists of applying the K-means clustering and KNN algorithms on the TF-IDF weights. KNN is a supervised classification technique that generates new data points based on the  $k$  number of the nearest data points, whereas K-means clustering represents an unsupervised clustering approach that collects and assembles data into a  $k$  number of centroid-based clusters.

Table 1. TF-IDF weights

	Word tokens of cluster 1		Word tokens of cluster 2		Word tokens of cluster 3		Word tokens of cluster 4		Word tokens of cluster 5	
	Computer science, computer engineering, and informatics		Electronics		Electrical and power engineering		Telecommunication and information technology		Instrumentation and control engineering	
	Computer	Programming	Electronics	Microelectronic	Electrical	Voltage	Telecomm unication	Antenna	Robotics	Control system
Clusters	2.3	1.4	0.52	0.61	1.39	0.29	2.54	0.33	0.34	1.5
Paper 1	0.41	0.44	0	0	0	0	0	0.55	0	0
Paper 2	0	0	0.32	0	0	0.65	0	0	0	0
Paper 3	0	0	0.167	0	0.77	0.24	0	0	0	0
Paper 4	0	0	0	0	0	0	0.77	0.53	0.22	0
Paper 5	1.27	0.93	0	0	0	0	0	0	1.58	3.32
...										

As can be seen in Table 1, five different clusters were constructed. The first cluster, for example, is composed of three scopes, namely computer science, computer engineering, and informatics. Several word tokens were extracted from this cluster such as computer, programming, and security. Similarly, the remaining four clusters can be scrutinized by studying the extracted topics. According to the experimental results of K-means clustering, the majority of the papers were assigned to the correct cluster. The classification, number, and distribution of most papers between 2012 and 2019 are demonstrated in Figure 3. In K-means, following the symmetry concept, both seeded and unseeded centroid initializations are investigated and compared to the centroid-based classification. These outcomes show the promising performance of the proposed approach.

The performance of the proposed approach was evaluated using two common validation metrics, namely precision, and recall as illustrated in Figures 4 and 5, respectively. These metrics rely on the separation between related and unrelated items. The validation results in both figures show the accurate labeling of the papers' classification. The performance of the proposed system using K-means clustering and

KNN algorithms was compared to that of cosine similarity [33] based on the five clusters. As it is evident in the two figures, K-means clustering achieved outstanding performance in the resultant clusters, compared to the KNN algorithm, and cosine similarity.

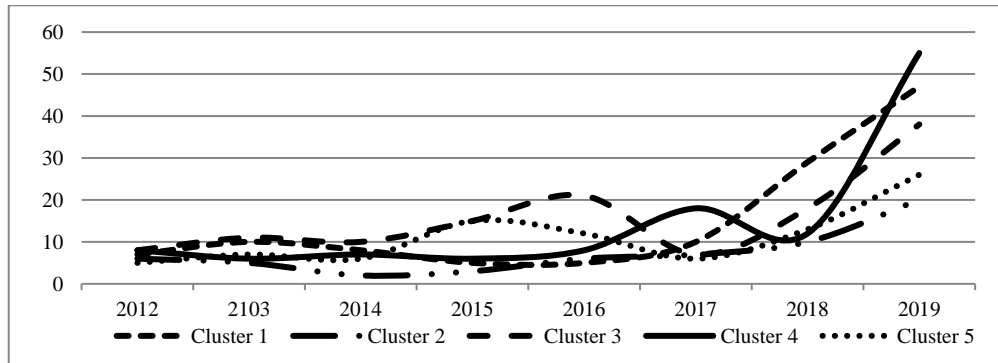


Figure 3. Papers classification and distribution

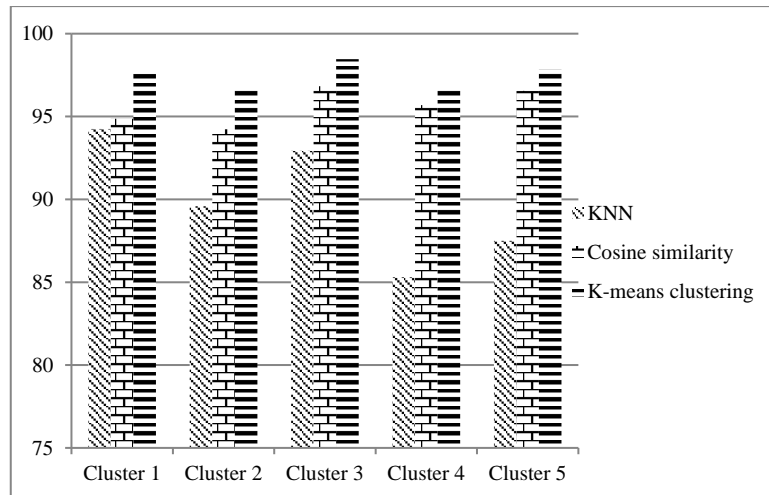


Figure 4. Precision results

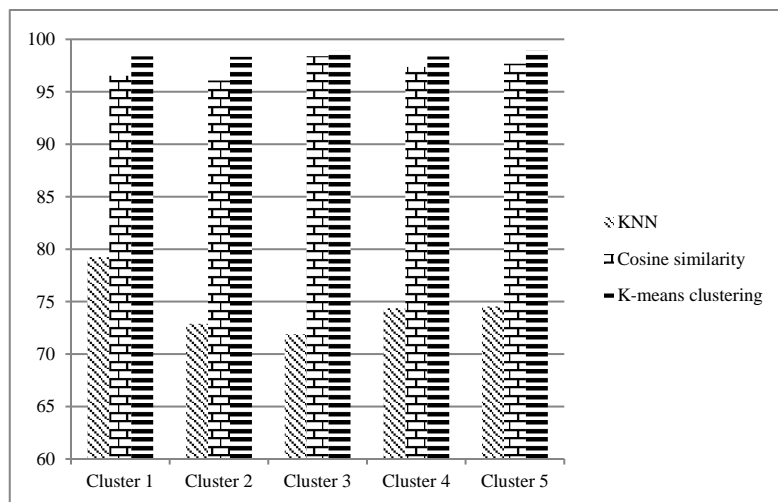


Figure 5. Recall results

#### 4. CONCLUSION

Classifying the research papers is crucial to ease the search for scientific research and meet the needs of the research community. A document classification approach for clustering the research articles has been proposed in this paper. The objective of this system is to make the organization and classification process of scientific papers fully automatic and more enhanced. Various web data mining algorithms were employed in designing the proposed approach to categorize the research articles based on the scope topics. Such techniques include TF-IDF, K-means algorithm, and KNN algorithm. These techniques achieved accurate and reliable classification performance based on several predefined clusters. The experimental results measured by precision and recall metrics showed the outstanding performance of the proposed system in classifying 98% of the articles in similar scopes using the K-means clustering.




#### REFERENCES

- [1] M. Gopianand and P. Jaganathan, "An effective quality analysis of XML web data using hybrid clustering and classification approach," *Soft Computing*, vol. 24, pp. 2139–2150, 2020, doi: 10.1007/s00500-019-04045-9.
- [2] A. A. Amer and H. I. Abdalla, "A set theory based similarity measure for text clustering and classification," *Journal of Big Data*, vol. 7, no. 74, pp. 1-43, 2020, doi: 10.1186/s40537-020-00344-3.
- [3] N. Kumar, S. K. Yadav, and D. S. Yadav, "Similarity measure approaches applied in text document clustering for information retrieval," in *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 2020, pp. 88-92, doi: 10.1109/PDGC50313.2020.9315851.
- [4] R. K. Roul and J. K. Sahoo, "A novel approach for ranking web documents based on query-optimized personalized pagerank," *International Journal of Data Science and Analytics*, vol. 7, pp. 37-55, 2021, doi: 10.1007/s41060-020-00232-2.
- [5] I. G. Gossen, T. Risse, and E. Demidova, "Towards extracting event-centric collections from web archives," *International Journal on Digital Libraries*, vol. 21, pp. 3–45, 2020, doi: 10.1007/s00799-018-0258-6.
- [6] S. Sharma, V. Gupta, and M. Juneja, "Diverse feature set based keyphrase extraction and indexing techniques," *Multimedia Tools and Applications*, vol. 80, pp. 4111–4142, 2021, doi: 10.1007/s11042-020-09423-2.
- [7] S. B. Marzijarani and H. Sajedi, "Opinion mining with reviews summarization based on clustering," *International Journal of Information Technology*, vol. 12, pp. 1299–1310, 2020, doi: 10.1007/s41870-020-00511-y.
- [8] A. Dhar, H. Mukherjee, N. S. Dash, and Kaushik Roy, "Text categorization: past and present," *Artificial Intelligence Review*, vol. 54, pp. 3007–3054, 2021, doi: 10.1007/s10462-020-09919-1.
- [9] N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, "Improving classification accuracy using clustering technique," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 7, no. 3, pp. 465-470, 2018, doi: 10.11591/eei.v7i3.1272.
- [10] J. Singh and A. K. Singh, "NSLPCD: topic based tweets clustering using node significance based label propagation community detection algorithm," *Annals of Mathematics and Artificial Intelligence*, vol. 89, pp. 371–407, 2021, doi: 10.1007/s10472-020-09709-z.
- [11] M. Gayatri, P. Satheesh, and R. R. Rao, "Towards an efficient framework for web user behavioural pattern mining," *International Journal of System Assurance Engineering and Management*, 2021, doi: 10.1007/s13198-021-01212-w.
- [12] Y. S. Su and S. Y. Wu, "Applying data mining techniques to explore user behaviors and watching video patterns in converged IT environments," *Journal of Ambient Intelligence and Humanized Computing*, 2021, doi: 10.1007/s12652-020-02712-6.
- [13] M. C. d. Menezes, et al., "Web data mining: validity of data from google earth for food retail evaluation," *Journal of Urban Health*, vol. 98, pp. 285–295, 2021, doi: 10.1007/s11524-020-00495-x.
- [14] L. D. C. S. Subhashini, Y. Li, J. Zhang, A. S. Atukorale, and Y. Wu, "Mining and classifying customer reviews: a survey," *Artificial Intelligence Review*, 2021, doi: 10.1007/s10462-021-09955-5.
- [15] V.-T. Luu, G. Forestier, J. Weber, P. Bourgeois, F. Djelil and P. -A. Muller, "A review of alignment based similarity measures for web usage mining," *Artificial Intelligence Review*, vol. 53, pp. 1529–1551, 2020, doi: 10.1007/s10462-019-09712-9.
- [16] H.-P. Chan, L. Xu, H.-H. Liu, R.-T. Zhang and A. K. Sangaiah, "System design of cloud search engine based on rich text content," *Mobile Networks and Applications*, vol. 26, pp. 459–472, 2021, doi: 10.1007/s11036-020-01676-3.
- [17] C. D. Schultz, "Informational, transactional, and navigational need of information: relevance of search intention in search engine advertising," *Information Retrieval Journal*, vol. 23, pp. 117–135, 2020, doi: 10.1007/s10791-019-09368-7.
- [18] K. Haruna, et al., "Research paper recommender system based on public contextual metadata," *Scientometrics*, vol. 125, pp. 101–114, 2020, doi: 10.1007/s11192-020-03642-y.
- [19] M. B. Magara, S. O. Ojo, and T. Zuva, "Towards a serendipitous research paper recommender system using bisociative information networks (BisoNets)," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, 2018, pp. 1-6, doi: 10.1109/ICABCD.2018.8465475.
- [20] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, "Document clustering using K-means with term weighting as similarity-based constraints," *Symmetry*, vol. 12, no. 6, pp. 1-25, 2020, doi: 10.3390/sym12060967.
- [21] A. F. Alsuhaime, A. M. Azmi, and M. Hussain, "Improving the retrieval of arabic web search results using enhanced K-means clustering algorithm," *Entropy*, vol. 23, no. 4, pp. 1-19, 2021, doi: 10.3390/e23040449.
- [22] K. Prudhvi, A. B. Chowdary, P. S. R. Reddy and P. L. Prasanna, "Text summarization using natural language processing," *Intelligent System Design, Advances in Intelligent Systems and Computing*, vol. 1171, pp. 535-547, 2021, doi: 10.1007/978-981-15-5400-1\_54.
- [23] U. Mahender, M. K. Swamy, H. Shaik and S. Kumar, "Improved approach to extract knowledge from unstructured data using applied natural language processing techniques," *Humanin Proceedings of the Third International Conference on Computational Intelligence and Informatics, Advances in Intelligent Systems and Computing*, 2020, vol. 1090, pp. 145-152, doi: 10.1007/978-981-15-1480-7\_12.
- [24] D. Kozłowski, J. Dusdal, J. Pang, and A. Zilian, "Semantic and relational spaces in science of science: deep learning models for article vectorisation," *Journal Scientometrics*, vol. 126, pp. 5881–5910, 2021, doi: 10.1007/s11192-021-03984-1.
- [25] P. Fafalios, H. Holzmann, V. Kasturia, and W. Nejdl, "Building and querying semantic layers for web archives (extended version)," *International Journal on Digital Libraries*, vol. 21, pp. 149–167, 2020, doi: 10.1007/s00799-018-0251-0.




- [26] S. Choudhary, H. Guttikonda, D. R. Chowdhury, and G. P. Learmonth, "Document retrieval using deep learning," *2020 Systems and Information Engineering Design Symposium (SIEDS)*, 2020, pp. 1-6, doi: 10.1109/SIEDS49339.2020.9106632.
- [27] L. Zhang, "Research on case reasoning method based on TF-IDF," *International Journal of System Assurance Engineering and Management*, vol. 12, pp. 608–615, 2021, doi: 10.1007/s13198-021-01135-6.
- [28] K. G. Dhal, A. Das, S. Ray, K. Sarkar and J. Gálvez, "An analytical review on rough set based image clustering," *Archives of Computational Methods in Engineering*, 2021, doi: 10.1007/s11831-021-09629-z.
- [29] G. Yasmin, S. Chowdhury, J. Nayak, P. Das and A. K. Das, "Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework," *Neural Computing and Applications*, 2021, doi: 10.1007/s00521-021-06132-1.
- [30] V. K. Kotte, S. Rajavelu, and E. B. Rajsingh, "A similarity function for feature pattern clustering and high dimensional text document classification," *Foundations of Science*, vol. 25, pp. 1077–1094, 2020, doi: 10.1007/s10699-019-09592-w.
- [31] C. Reyes-Peña, M. T. Vidal, and J. d. J. L. Martínez, "Document clustering by relevant terms: an approach," in *Proceedings of the Future Technologies Conference (FTC) 2019, Advances in Intelligent Systems and Computing*, 2020, vol. 1069, pp. 610-617, doi: 10.1007/978-3-030-32520-6\_44.
- [32] C. Zong, R. Xia, and J. Zhang, *Text data mining*, Singapore: Springer, 2021.
- [33] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, No. 1, pp. 664-670, 2021, doi: 10.11591/ijece.v11i1.pp664-670.

## BIOGRAPHIES OF AUTHORS






**Wasseem N. Ibrahim Al-Obaydy**    received the B.Sc. degree in computer and software engineering from University of Technology, Iraq in 2003. He received the Master's degree in computing from University of Buckingham, United Kingdom in 2011. Currently, he is working as a lecturer at the computer engineering department, college of engineering, Al-Iraqia University, Iraq. His research interests include face recognition, feature extraction, image processing, computer vision and pattern recognition. He has served as a reviewer for IEEE Access journal. He can be contacted at email: wasseem.nahi@aliraqia.edu.iq.






**Hala A. Hashim**    received a B.Sc. degree in mathematics and computer applications from Al-Nahrain University, Iraq in 2001. She received her Master's degree in Applied Mathematics from Al-Nahrain University, Iraq in 2005. Currently, she is working as a lecturer at the Dentistry Department at Dijlah University College, Iraq. Her research interests include but are not limited to numerical methods, integral equations, differential equations, smart algorithms, applied mathematics, distance education, and pattern recognition. She can be contacted at email: hala.adnan@duc.edu.iq.



**Yassen AbdulKhaleq Najm**    received the BSc degree in statistics and computer science from Al-Rafidain University College, Iraq in 1994. He received the Master degree in computer science from Universiti Teknikal Malaysia Melaka, Malaysia in 2020. Currently, he is a lecturer of Department of English, College of Arts, Al-Iraqia University, Baghdad, Iraq. His research interests include network technology, cloud computing security, and database applications. He can be contacted at email: yasin.abdulkhaliq@aliraqia.edu.iq.



**Ahmed Adeeb Jalal**    received the Engineer degree in Software Engineering from Al-Rafidain University College, Iraq in 2002. He received the Master degree in Computer Engineering from Yildiz Technical University, Turkey in 2016. Currently, he is a lecturer of Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq. His research interests include data mining, hybrid recommendation systems design, and web applications. He can be contacted at email: ahmedadeeb@aliraqia.edu.iq.