Text documents clustering using data mining techniques

Ahmed Adeeb Jalal, Basheer Husham Ali
Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq

Article Info

Article history:

Received Apr 12, 2020
Revised Jul 10, 2020
Accepted Jul 25, 2020

Keywords:

Cosine similarity
Document clustering
TF-IDF
Topics
Web data mining

ABSTRACT

Increasing progress in numerous research fields and information technologies, led to an increase in the publication of research papers. Therefore, researchers take a lot of time to find interesting research papers that are close to their field of specialization. Consequently, in this paper we have proposed documents classification approach that can cluster the text documents of research papers into the meaningful categories in which contain a similar scientific field. Our presented approach based on essential focus and scopes of the target categories, where each of these categories includes many topics. Accordingly, we extract word tokens from these topics that relate to a specific category, separately. The frequency of word tokens in documents impacts on weight of document that calculated by using a numerical statistic of term frequency-inverse document frequency (TF-IDF). The proposed approach uses title, abstract, and keywords of the paper, in addition to the categories topics to perform the classification process. Subsequently, documents are classified and clustered into the primary categories based on the highest measure of cosine similarity between category weight and documents weights.

Corresponding Author:

Ahmed Adeeb Jalal,
Computer Engineering Department,
College of Engineering, Al-Iraqia University,
Baghdad, Iraq.
Email: ahmedadeeb@aliraqia.edu.iq

1. INTRODUCTION

Web document clustering is a suitable technique for collecting documents with similar content from a set of documents that spread on the web pages [1-3]. Document clustering provides one of useful and efficient techniques to find and understand the documents [4], where clustering can group the similar documents in one place. Accordingly, web documents can be classified according to a collection of topics for each category. These topics focus on word tokens that may appear during document analysis. The word tokens also refers to the repetition of terms in documents, where extracting terms from textual data helps in the classification of documents [5]. Consequently, the documents are classified by a cluster of terms into a set of categories, based on the number of occurrences with each word tokens for a specific topic in those documents [6].

The classification of documents is expedient for researchers who perform interdisciplinary research on various topics [7]. Ordinarily, document clustering is an important pillar in achieving this objective [8, 9]. Clustering will help the user to get all relevant documents in one category and the search can be limited to some important documents of his choice. Conversely, finding meaningful documents for researchers by normal search process, is a challenging and time-consuming problem especially in view of the steady increase in the number of documents. Moreover, diversity of the major sources of documents such as research papers, web pages, archives, technical reports, and digital repositories that available to the user over the internet.

Nowadays, a large number of people use the internet as their main source of information. Consequently, the users need to find their interesting requests easily and conveniently which represents the most relevant information that was queried [10-12]. However, the search engine retrieves more irrelevant pages based on a few keywords for a user’s query, resulting in long lists of URLs. Searching on the web pages to discover knowledge according to user query, is not an easy task to perform. Considering, the problem of information overload that facing internet data warehouses. Therefore, web data mining can be an easy and important technology for discovering and retrieving useful information and knowledge [13, 14]. Web data mining is a sub discipline of data mining applications to discover patterns that mainly deal with the internet. Web data mining can be categorized into three types: web structure mining, web content mining, and web usage mining [15]. All these types use a diversity of approaches, techniques, tools, and algorithms to discover the patterns of information [14]. Accordingly, improving search engine using data mining techniques aims to discover useful information from the large amount of data [16, 17].

Over the past decades, institutions, universities, and journals have published numerous research papers in various scientific fields. Ordinarily, research papers are not classified and clustering into categories. Consequently, there are many documents clustering approaches [8] and recommender systems [18] that proposed for classifying research papers based on the documents content characteristics or attributes. Each of these techniques differs in many parts, such as the types of attributes they used to characterize the documents, the similarity measure used, the representation of the clusters etc. The literature reviews of related works on research paper classification and its applications are as follows.

Thushara *et al.*, [19] proposed a document-centered system for classifying research articles that published in the domains of computer science. It is based on automatic keywords extraction from research articles using rapid automatic keyword extraction (RAKE) algorithm to get best score-matrix of keywords. Moreover, the proposed system adopts a hybrid approach to the classification process by applying different methods at various phases of the system. This classification process relates to the semantic analysis by using score-matrix of keywords and cosine similarity for articles classification into relevant domain. Consequently, domain classification facilitates the identification and retrieval of important articles for researchers that are in line with their actual fields of interest.

Kim and Gil [20] proposed the paper classification system consists of four major processes: crawling, TF-IDF, topic modeling and data management, and classification. This proposed system aims to cluster the research papers into the meaningful categories in which contain similar topics. Accordingly, the proposed system creates a dictionary of keywords from the abstract and keywords data that crawled. These keywords consist of top-N of high frequency keywords among the entire keywords. Also, it extracts topics from the abstract data of each paper by latent dirichlet allocation (LDA) scheme. Finally, research papers are classified into similar subjects by using K-means clustering algorithm. The K-means clustering algorithm is based on the term frequency-inverse document frequency (TF-IDF) values of each paper.

Nahar *et al.*, [21] presented an approach for classifying and clustering the research’s papers into clusters based on concepts and contents. This clustering process uses title, keywords, and abstract of the paper for performing the classification process. The proposed approach is mainly depends on information retrieval (IR) as core process along with some natural language processing (NLP) techniques, latent dirichlet allocation (LDA), and latent semantic indexing (LSI). Moreover, it aims to improve the LDA model that is used for classification using the concept of topic modeling and the LSI model used for performing querying. Consequently, the presented approach provides an automatic, short time, and accurate solution for classifying research papers that published in the field of information technology.

Saad *et al.*, [22] presented emotions classification for Malay folklore from children short stories using four types of common emotions: happy, angry, fearful, and sad. This work based on term frequency-inverse document frequency (TF-IDF) that extracted from the text stories. Then, text stories will be classified by support vector machine (SVM) and decision tree (DT). This work aims to add emotions for a more natural storytelling.

In addition, there are also various other approaches for classifying the documents by applying different techniques such as using text mining based on the technology of natural language processing [23, 24], building a semantic representation of articles from their associated entities [25, 26], and using N-grams and efficient similarity measure that known as improved sqrt-cosine similarity measure [27]. As mentioned in the examples above, the importance of documents clustering and classifying is highlighted to satisfy users and facilitate the retrieval process of relevant documents.

This paper aims to classify and cluster the research papers into categories to overcome the respective difficulties for the search users. Moreover, clustering provides a better coverage while avoiding complexity, not only with research papers but with various domains as well [28-30]. Thus, this proposed approach of text documents clustering has a significant impact to find useful information, address the lack of understand-ability, and improve search-ability for users. Consequently, we proposed research papers classification system based on term frequency (TF), Term frequency-inverse document frequency (TF-IDF), and cosine similarity, to guide the users by their needs in the domain of research papers.

The second section explains the methodology and describes proposed methods for text documents clustering such as web mining, data extraction, TF-IDF, and cosine similarity. These techniques contribute to the analysis of scientific papers by extracting data from it, in order to classify the papers into groups organized according to similarity. The third section highlights on the results of the proposed classification approach and the algorithms that used to implement it. Finally, this research outlines the challenges of research papers classification and aims to provide a better clustering for the research papers.

2. RESEARCH METHOD

In this paper, a classification approach for clustering the research papers is presented, as researchers spend a lot of time to identifying the relevant cluster of the undertaken papers. Ordinarily, the papers are classified into clusters based on the concepts and the contents. Accordingly, our approach provides a clustering process depends on three major parts of the research papers: title, abstract, and keywords. The abstract was chosen as one of the important parts of the paper that describes its essence after the title [31, 32], and it is often the next part that users tend to read. Moreover, the abstract is enriched with interesting and fundamental words/terms that express the direction of the paper and a summary of all other contents of the paper.

The data set contains 518 papers that published in Bulletin of Electrical Engineering and Informatics (BEEI) journal, since 2012 to 2019. These scientific papers include different topic scopes which are written in English. The BEEI journal is issued by the Institute of Advanced Engineering and Science (IAES) of Ahmad Dahlan University. Our goal is to classify these papers into five clusters according to the following scopes of the journal:

- Cluster 1: Computer Science, Computer Engineering, and Informatics.
- Cluster 2: Electronics.
- Cluster 3: Electrical and Power Engineering.
- Cluster 4: Telecommunication and Information Technology.
- Cluster 5: Instrumentation and Control Engineering.

Ordinarily, the research papers are often classified and retrieved according to the user’s query or by semantic representation and many other methods, as we mentioned in the literature reviews of related works in the first section. In our approach, we apply basic crawler algorithm [15] to extract the contents of the topics for each cluster separately, as well as the title, abstract, and keywords of all papers. Subsequently, we suggest classifying papers based on word tokens which extracted from the topics of the above five clusters that covered by the BEEI journal. Moreover, classification approach techniques include TF-IDF and cosine similarity. Figure 1 shows general steps of the flow diagram for techniques that used in the proposed classification approach.

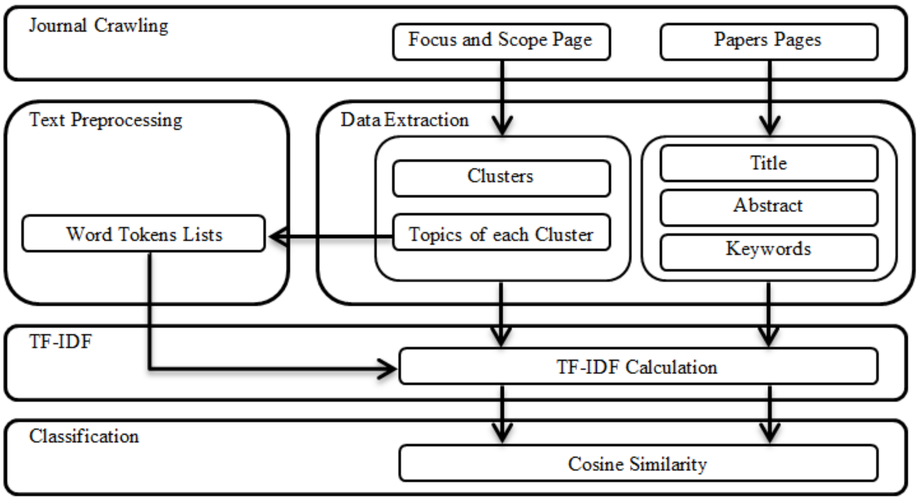


Figure 1. Classification approach flow diagram

2.1. Text preprocessing

Text preprocessing is a one of major component in many algorithms of text mining. It usually consists of the tasks such as tokenization, filtering, lemmatization, and stemming [33]. Ordinarily, clustering algorithms require to specifying the type of attributes (e.g. words, terms, or phrases) to extract from the documents that underpin the clustering algorithm performance.

As shown in the text preprocessing step of Figure 1, it automatically extracts word tokens lists using text preprocessing tasks. Tokenization is the task of breaking the character sequence in topics that are crawled into pieces (words/terms) called tokens. Filtering is a task intended to perform further processing on word tokens lists to remove stop and similar words to reduce the indexing size and increase the accuracy of results. Moreover, it necessary be taken into consideration the morphological analysis of words to group the various related words together to be analyzed as one item, lemmatization task is preferred in practice. Stemming task aims to get a stem (root) of derivative words that are actually language dependent. Consequently, we get five lists of word tokens from clusters topics, one for each cluster.

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

TF-IDF is a numerical and descriptive statistical mechanism that used as a weighting factor in the fields of information retrieval. The TF-IDF weighting provides a good insight of how important words are by the appearance of specific words in documents content. Consequently, the TF-IDF is used to extract word tokens from documents, calculate degrees of similarity among documents, determine important ranking, etc. In our approach, we calculate TF, IDF and TF-IDF for each word token in the lists on both clusters and documents.

The term frequency (TF) counts how often the specific words appear in document content, which can be calculated as in (1). The words with a high TF value are more importance in documents.

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

where, $f_{t,d}$ denotes to the frequency of word/term t that occurs in document d.

On the other hand, the inverse document frequency (IDF) measures the rarity and importance of a word/term across all documents, which can be calculated as in (2). The words with a high IDF value are considered rare in all documents.

$$IDF_{t,D} = log \frac{D}{\{d \in D : t \in d\}} \tag{2}$$

where, $IDF_{t,D}$ is a logarithmic scale for dividing the total number of documents D by the number of documents in which the word/term t appears.

Consequently, the Term Frequency–Inverse Document Frequency (TF-IDF) weighting is calculated as in (3). The TF-IDF weighting value increases when a specific word/term has high frequency in a document and the number of documents in which the word/term appears is low.

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

2.3. Cosine similarity

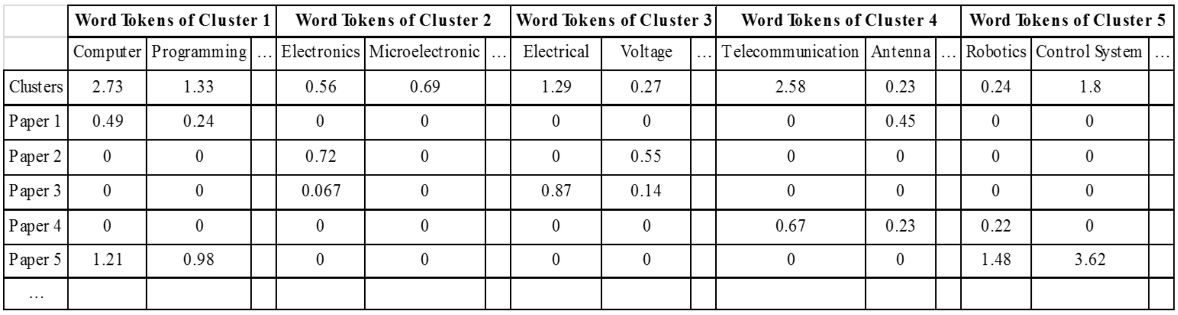
Cosine similarity is a one of the powerful similarity measures compared to all other techniques, that used to measure similarities between two vectors based on the cosine of the angle as in (4). Moreover, the cosine similarity is widely used in document clustering in the field of data mining. Ordinarily, the cosine similarity method measures the similarity between a user query and retrieved documents based on the terms that extracted from the user query. Nevertheless, in our approach we suggest measuring the similarity between the content of clusters and documents based on the word tokens lists, as shown in the classification step of Figure 1.

$$Cosine_{C,D} = \frac{\sum_{i=1}^{n} C_i \times D_i}{\sqrt{\sum_{i=1}^{n} C_i^2} \times \sqrt{\sum_{i=1}^{n} D_i^2}} \tag{4}$$

where, C and D are denote to the cluster and document vectors, respectively. The higher-ranking documents are more relevant to the cluster.

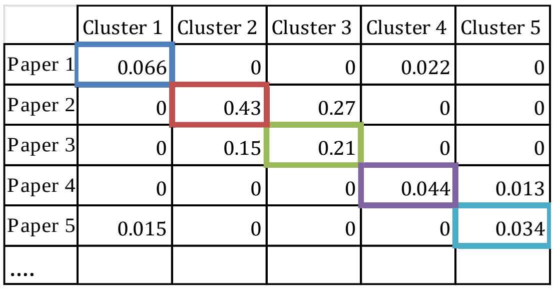
3. RESULTS AND DISCUSSIONS

The proposed research papers classification system is based on web data mining techniques to manage and process research papers data. In this section, we will describe the data sets collected and the steps taken while running the experiments along with discussing the results down to the evaluation. We collected 518 research papers for use in experiments, that are published in BEEI journal in various subject and scopes. The papers are related to the field of computer science, computer engineering, informatics, electronics, electrical, power engineering, telecommunication, information technology, instrumentation, control engineering. Each of these scope contains several topics such as computer architecture, programming, computer security, electronic materials, microelectronic system, electrical engineering materials, antenna and wave propagation, distributed platform, and robotics. Our goal is to classify these papers into five clusters according to those scopes. Consequently, as we explained early in research method section, we crawled the title, keywords and abstract for each paper to use as core data for classification. Meanwhile, we extract five lists of word tokens from the topics of scopes. Once these steps are completed, the corpus became ready to be used as input for TF-IDF calculation module to calculate the weight for each word token for both clusters and papers, as shown in Figure 2. Subsequently, the cosine similarity algorithm is implemented based on TF-IDF weights, as shown in Figure 3. Typically, the cosine similarity value ranges from 0 to 1, where a high value indicates that data are well-matched to their own cluster and poorly matched to neighboring clusters.



	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	Programming	…	Electronics	Microelectronic	…	Electrical	Voltage	…	Telecommunication	Antenna	…	Robotics	Control System	…
Clusters	2.73	1.33		0.56	0.69		1.29	0.27		2.58	0.23		0.24	1.8	
Paper 1	0.49	0.24		0	0		0	0		0	0.45		0	0	
Paper 2	0	0		0.72	0		0	0.55		0	0		0	0	
Paper 3	0	0		0.067	0		0.87	0.14		0	0		0	0	
Paper 4	0	0		0	0		0	0		0.67	0.23		0.22	0	
Paper 5	1.21	0.98		0	0		0	0		0	0		1.48	3.62	
…															

Figure 2. TF-IDF weights



	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Paper 1	0.066	0	0	0.022	0
Paper 2	0	0.43	0.27	0	0
Paper 3	0	0.15	0.21	0	0
Paper 4	0	0	0	0.044	0.013
Paper 5	0.015	0	0	0	0.034
….					

Figure 3. Cosine similarity results

As we see in Figure 2, there are five different clusters. For instance, the first cluster revolves about computer science, computer engineering, and informatics. The first cluster consists of many word tokens such as computer, programming, computing, and security, to mention a few. Similarly, we can examine the rest of clusters by analyzing the set of extracted topics. The results showed that most of the papers have been linked to the right cluster, depending on the results of cosine similarity algorithm. Figure 4 shows the classification, number, and distribution of over 96% of papers, since 2012 to 2019. These results constitute the efficiency of the proposed approach.

The validation factor allows evaluating the classification of papers according to the selected algorithms. We evaluate the proposed system using precision and recall metrics which are one of the most common validation metrics that based on separation between relevant and irrelevant items. As shown in Figure 5, the validation gives more accurate labeling for the papers classification. We found that some papers contain mixed subjects, which means that many different module, contribution, and tools have been employed in the paper.

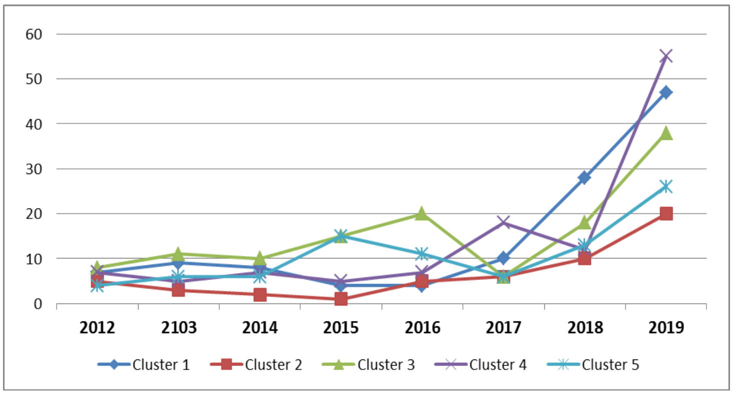


Figure 4. Papers classification and distribution

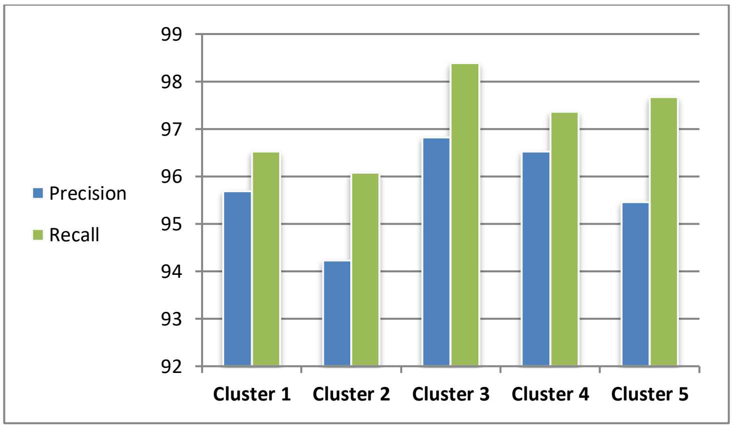


Figure 5. Validation results

4. CONCLUSION

In this paper, we proposed a classification approach for clustering the research papers to improve and automate the process of organizing and classifying scientific papers. The classification approach that introduced in this paper uses web data mining techniques to classify research papers depending on the focus and scope topics. The selected algorithms have shown accurate and reliable results in the classification according to predefined clusters. Ordinarily, classification of papers is essential to facilitate the finding of scientific research and increase the effectiveness of identifying the needs of researchers. The experimental results showed that it is possible to classify more than 96% of the papers in similar scopes using the cosine similarity algorithm, as these results were verified by precision and recall metrics. This paper mainly focuses on developing and analyzing the classification of research papers based on clusters topics. Future work should be extended to include various topics extracted from the papers to classify the whole papers accurately and efficiently.

REFERENCES

[1] J. Avanija, et al., “Semantic Similarity based Web Document Clustering Using Hybrid Swarm Intelligence and Fuzzy C-Means,” *HELIX -The Scientific Explorer*, vol. 7, no. 5, pp. 2007-2012, 2017.

[2] A. P. Singh, et al., “Phrase based Web Document Clustering: an Indexing Approach,” *Computer Communication, Networking and Internet Security*, vol. 5, pp. 481-492, 2017.

[3] R. K. Roul, et al., “Web Document Clustering and Ranking Using TF-IDF based Apriori Approach,” in *IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications (ICACEA)*, vol. 2, pp. 34-39, 2014.

[4] N. K. Nagwani, “Summarizing Large Text Collection Using Topic Modeling and Clustering based on Mapreduce Framework,” *Journal of Big Data*, vol. 2, no. 1, pp. 1-18, 2015.

[5] I. Alsmadi and I. Alhami, “Clustering and Classification of Email Contents,” *Journal of King Saud University - Computer and Information Sciences*, vol. 27, no. 1, pp. 46-57, 2015.

[6] P. Gurung and R. Wagh, “A Study on Topic Identification Using K Means Clustering Algorithm: Big vs. Small Documents,” *Advances in Computational Sciences and Technology*, vol. 10, no. 2, pp. 221-233, 2017.

[7] P. B. Bafna, et al., “Document Clustering: TF-IDF Approach,” in *IEEE 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),* pp. 61-66, 2016.

[8] N. Oikonomakou and M. Vazirgiannis, “A Review of Web Document Clustering Approaches,” *Data Mining and Knowledge Discovery Handbook,* pp. 931-948, 2010.

[9] N. M. N. Mathivanan, et al., “Improving Classification Accuracy Using Clustering Technique,” *Bulletin of Electrical Engineering and Informatics,* vol. 7, no. 3, pp. 465-470, 2018.

[10] A. S. Al-Hegami and H. H. Al-Omaisi, “Data Mining Techniques for Mining Query Logs in Web Search Engines,” *International Journal of Computer Science and Network,* vol. 6, no. 2, pp. 2277-5420, 2017.

[11] S. Girish, et al., “Mining the Web Data for Classifying and Predicting Users’ Requests,” *International Journal of Electrical and Computer Engineering,* vol. 8, no. 4, pp. 2088-8708, 2018.

[12] S. Khan, et al., “Web Mining in Search Engines for Improving Page Rank,” *International Journal of Soft Computing and Engineering,* vol. 5, no. 4, pp. 2231-2307, 2015.

[13] M. J. H. Mughal, “Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview,” *International Journal of Advanced Computer Science and Applications,* vol. 9, no. 6, pp. 208-215, 2018.

[14] A. A. Jalal, “Big Data and Intelligent Software Systems,” *International Journal of Knowledge-based and Intelligent Engineering Systems,* vol. 22, no. 3, pp. 177-193, 2018.

[15] B. Liu, “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2011.

[16] K. O. Khorsheed, “Search Engine Optimization Using Data Mining Approach,” *International Journal of Computer Engineering and Applications,* vol. 9, no. 6, pp. 184-200, 2015.

[17] N. Duklan, et al., “Classification of Search Engine Optimization Techniques: A Data Mining Approach,” in *2nd International Conference on System Modeling & Advancement in Research Trends (SMART),* 2015.

[18] B. M. Maake, et al., “Information Processing in Research Paper Recommender System Classes,” *Research Data Access and Management in Modern Libraries,* pp. 90-118, 2019.

[19] M. G. Thushara, et al., “Domain Classification of Research Papers Using Hybrid Keyphrase Extraction Method,” *Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing,* vol. 708, pp. 387-398, 2018.

[20] S. Kim and J. Gil, “Research Paper Classification Systems based on TF-IDF and LDA Schemes,” *Human-centric Computing and Information Sciences,* vol. 9, no. 30, pp. 1-21, 2019.

[21] K. M. O. Nahar, et al., “NLP and IR Based Solution for Confirming Classification,” *Journal of Theoretical and Applied Information Technology,* vol. 96, no. 16, pp. 5269-5279, 2018.

[22] M. M. Saad, et al., “Evaluation of Support Vector Machine and Decision Tree for Emotion Recognition of Malay Folklores,” *Bulletin of Electrical Engineering and Informatics,* vol. 7, no. 3, pp. 479-486, 2018.

[23] S. Sulova, et al., “Using Text Mining to Classify Research Papers,” in *17th International Multidisciplinary Scientific GeoConference SGEM 2017,* vol. 17, no. 21, pp. 647-654, 2017.

[24] S. A. Salloum, et al., “Using Text Mining Techniques for Extracting Information from Research Articles,” *Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence,* vol. 740, pp. 373-397, 2018.

[25] S. Wang and R. Koopman, “Clustering Articles based on Semantic Similarity,” *Scientometrics,* vol. 111, pp. 1017-1031, 2017.

[26] R. K. Ibrahim, et al., “Survey on Semantic Similarity Based on Document Clustering,” *Advances in Science, Technology and Engineering Systems Journal,* vol. 4, no. 5, pp. 115-122, 2019.

[27] D. B. Bisandu, et al., “Clustering News Articles Using Efficient Similarity Measure and N-grams,” *International Journal of Knowledge Engineering and Data Mining,* vol. 5, no. 4, pp. 333-348, 2018.

[28] A. H. Nasution, et al., “Generating Similarity Cluster of Indonesian Languages with Semi-supervised Clustering,” *International Journal of Electrical and Computer Engineering,* vol. 9, no. 1, pp. 531-538, 2019.

[29] H. M. Alghamdi and A. Selamat, “Arabic Web Page Clustering: A Review,” *Journal of King Saud University - Computer and Information Sciences,* vol. 31, no. 1, pp. 1-14, 2019.

[30] S. Singh and P. Singh, “Speaker Specific Feature Based Clustering and Its Applications in Language Independent Forensic Speaker Recognition,” *International Journal of Electrical and Computer Engineering,* vol. 10, no. 4, pp. 3508-3518, 2020.

[31] J. Petrus, et al., “Soft and Hard Clustering for Abstract Scientific Paper in Indonesian,” in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS),* pp. 131-136, 2019.

[32] B. Nie and S. Sun, “Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research,” *Applied Sciences,* vol. 7, no. 4, pp. 401:1-21, 2017.

[33] M. Allahyari, et al., “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” in *Proceedings of KDD Bigdas,* 2017.