Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2, No.2 : 2024
ISSN : **1906-9685**

TEXT DOCUMENTS CLUSTERING USING DATA MINING TECHNIQUES

Dr. M.MEENA KRITHIKA, ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, NGM COLLEGE, POLLACHI

Abstract:
Increasingprogressinnumerousresearchfieldsandinformationtechnologies,ledtoanincreaseinthepublica tionofresearchpapers.introduces a document classification approach, the key points of the proposed approach are **Problem Context** which aims to reduce the time researchers spend locating relevant papers. **Classification approach** clusters research papers into categories based on their scientific field. Each category includes various topics, and word tokens are extracted from these topics to represent each category.**TF-IDF**: The term frequency-inverse document frequency (TF-IDF) method is used to calculate the importance of words within a document relative to the entire dataset, thus determining the weight of the document.**Cosine Similarity**: The classification process relies on cosine similarity, which measures the closeness between the weight of a document (calculated using TF-IDF) and the category weight. Papers are classified into categories based on the highest similarity score.**Text Features Used**: The system uses key features from the paper, such as the title, abstract, keywords, and relevant category topics, to perform classification.

1. Introduction

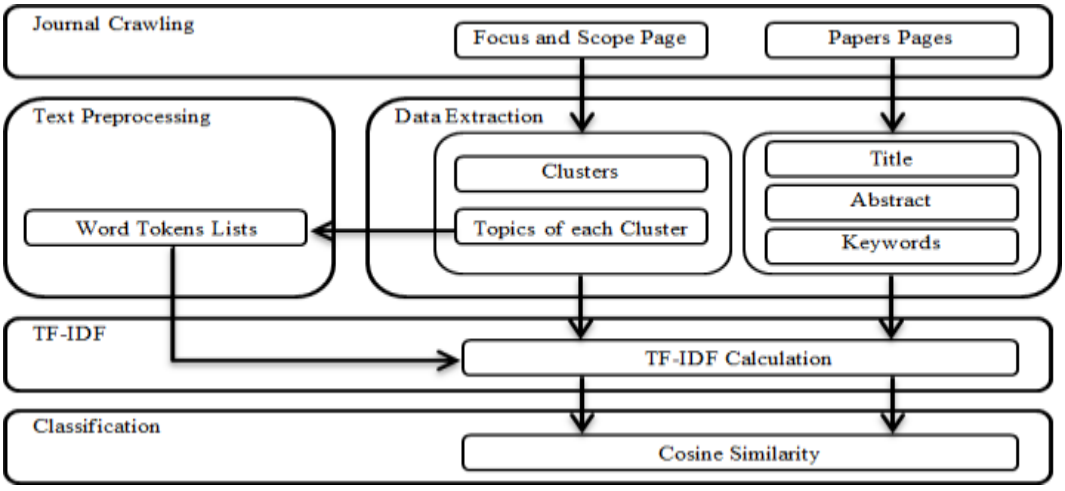
The challenges of finding relevant information on the internet and highlights web data mining as a solution. With the rapid increase in online information, users face information overload, making it difficult to efficiently retrieve the data they need. Search engines often return irrelevant results, which lengthens the search process. Web data mining, a subset of data mining, is seen as an effective approach for discovering useful patterns and knowledge from internet data. It includes three categories: web structure mining, web content mining, and web usage mining.
. It reviews various approaches and algorithms used in clustering, including K-means, TF-IDF, LDA (Latent Dirichlet Allocation), cosine similarity, and NLP-based techniques like RAKE (Rapid Automatic Keyword Extraction). These methods aim to organize research papers into meaningful categories, facilitating easier retrieval for researchers and improving search results.
Several research systems are discussed, such as those by Thushara et al., Kim and Gil, and Nahar et al., which use different techniques to extract key terms, cluster papers based on topics, and apply machine learning algorithms for classification. These approaches streamline the search and retrieval process, providing researchers with more relevant and accessible information.
This paper concludes by proposing a classification system based on TF, TF-IDF, and cosine similarity to cluster documents. This method aims to enhance user search experiences by categorizing research papers according to content similarity, addressing common challenges in retrieving relevant documents. The paper outlines its methodology, discusses the proposed classification system, and highlights the importance of clustering research papers to improve the search process.
This approach helps automate the classification of research papers, aiding researchers in quickly identifying relevant studies in their field. This passage introduces the concept of web document clustering as a method for grouping similar documents from a large set of web-based documents. Document clustering helps in understanding and locating documents based on shared content, making it particularly useful for researchers working on interdisciplinary topics. The clustering

process involves grouping documents based on the occurrence of specific word tokens—repeated terms that help classify the documents into categories. This technique addresses the challenges researchers face when trying to find relevant documents using traditional search methods, which can be time-consuming given the massive and growing number of documents on the web.
The diversity of sources, such as research papers, web pages, archives, technical reports, and digital repositories, adds to the complexity, making clustering an important tool for streamlining document retrieval. This passage expands on the challenges of finding relevant information on the internet and highlights web data mining as a solution. With the rapid increase in online information, users face information overload, making it difficult to efficiently retrieve the data they need. Search engines often return irrelevant results, which lengthens the search process. Web data mining, a subset of data mining, is seen as an effective approach for discovering useful patterns and knowledge from internet data. It includes three categories: web structure mining, web content mining, and web usage mining.

2. Research Method

This section presents the research methodology for classifying research papers into relevant clusters, addressing the time-consuming task researchers face in identifying appropriate papers. The approach focuses on clustering papers based on three key components: title, abstract, and keywords. These elements are chosen because they provide a concise representation of the paper's content. The abstract, in particular, is highlighted as a crucial part, often read after the title, and contains essential terms that summarize the paper's direction and contents.
The dataset used for this research includes 518 papers published in the *Bulletin of Electrical Engineering and Informatics (BEEI)* journal. The papers span various topics, and the goal is to classify them into five clusters aligned with the journal's scope:
To achieve this, a basic crawler algorithm is applied to extract the content of these papers, particularly the title, abstract, and keywords. Classification is based on word tokens extracted from the papers within each of the five defined clusters. Techniques such as **Term Frequency-Inverse Document Frequency (TF-IDF)** and **cosine similarity** are used to facilitate the clustering process.
A flow diagram (Figure 1) illustrates the general steps of the proposed classification approach, emphasizing the extraction and processing techniques applied to the dataset to group papers into meaningful clusters. The proposed methodology aims to enhance the retrieval and classification of research papers, making it easier for users to find papers aligned with their interests.

Figure1.Classificationapproach flowdiagram



2.1. Text Preprocessing

Text preprocessing is a critical component in many text mining algorithms and involves several key tasks such as **tokenization**, **filtering**, **lemmatization**, and **stemming** . These steps prepare the text for further analysis, ensuring that clustering algorithms perform optimally by focusing on the most meaningful attributes (e.g., words, terms, or phrases) extracted from the documents.

As outlined in the text preprocessing stage of Figure 1, this step automatically generates lists of word tokens by executing the following tasks:

1. **Tokenization:** This involves breaking the character sequences in the crawled topics into smaller pieces known as tokens (words/terms). These tokens represent the basic units of the text to be processed.
2. **Filtering:** This step removes unnecessary words, such as stop words (e.g., "and", "the"), and similar redundant words. The aim is to reduce the size of the index and enhance the accuracy of the results by focusing only on relevant terms.
3. **Lemmatization:** This task groups different forms of related words together so they can be analyzed as a single item. It involves reducing words to their base or dictionary form (lemma), making it easier to analyze variations of the same word.
4. **Stemming:** Stemming reduces words to their root form, which is language-dependent. This process helps consolidate words with the same root, even if they appear in different forms.

After these steps, five separate lists of word tokens are generated, each corresponding to one of the predefined clusters. These token lists serve as the foundation for clustering the research papers based on their content.

2.2. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a widely used statistical method that assigns a weight to each word in a document, helping to identify the most important words based on their frequency of appearance. It serves as a key tool in information retrieval to calculate the importance of words, rank documents, and determine degrees of similarity among documents. In this approach, **TF-IDF** is used to extract word tokens from documents and calculate their significance within both clusters and individual documents.

1. Term Frequency (TF)

Term Frequency (TF) measures how often a particular word appears in a document. Words that occur frequently in a document are considered more important.

2. Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) measures how rare or common a word is across a set of documents. A word that appears in many documents has a low IDF value, while a rare word (present in only a few documents) has a high IDF.

3. TF-IDF Calculation

The final **TF-IDF** score combines both TF and IDF to provide a measure of how important a word is in a document, considering both its frequency in that document and its rarity across the entire document set.

This weighting increases when a word is frequent in a particular document but rare across other documents, making it a valuable indicator for clustering and classification.

2.3. Cosine Similarity

Cosine similarity is a powerful technique commonly used to measure the similarity between two vectors by calculating the cosine of the angle between them. It is especially useful in document clustering and information retrieval, where it helps in determining how similar two documents are based on their content. In this approach, cosine similarity is applied to measure the similarity between clusters and individual documents, using word token lists as the basis for comparison.

Ordinarily, cosine similarity is used to compare a user query with retrieved documents. However, in this paper's method, cosine similarity is used to compare the content of each cluster with the documents to identify the most relevant ones for each cluster.

The resulting cosine similarity score ranges from 0 to 1:

- A score of **1** means the vectors (i.e., the cluster and document) are identical in terms of content.
- A score of **0** indicates no similarity.

Higher cosine similarity scores imply that the document is more relevant to the cluster, and thus, a better match for classification.

3. Results and Discussions

The proposed research paper classification system leverages web data mining techniques to process and organize research papers. In this section, we detail the dataset used, experimental procedures, and discuss the results of the classification process.

3.1. Dataset Overview

A dataset of **518 research papers** published in the *Bulletin of Electrical Engineering and Informatics (BEEI)* journal from 2012 to 2019. These papers span a variety of fields, including:Each field contains several topics such as computer architecture, programming, computer security, microelectronic systems, antenna propagation, robotics, etc. The goal of this experiment is to classify these papers into **five clusters** corresponding to these fields.

3.2. Data Preparation

To begin the classification process, we **crawled the titles, keywords, and abstracts** of all the papers. These components were essential for extracting word tokens related to the topics within each cluster. After this step, the corpus was ready for the **TF-IDF calculation module**, which assigned weights to each word token based on its significance to both clusters and papers. The results of the TF-IDF calculations are shown in **Figure 2**.

	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	
...															

Figure2.TF-IDFweights

3.3. Cosine Similarity Results

Following the TF-IDF calculation, I implemented the **cosine similarity algorithm** to measure the relevance of each paper to its respective cluster, as displayed in **Figure 3**. Typically, cosine similarity values range from **0 to 1**, where a higher value indicates a stronger match to the cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Paper1	0.066	0	0	0.022	0
Paper2	0	0.43	0.27	0	0
Paper3	0	0.15	0.21	0	0
Paper4	0	0	0	0.044	0.013
Paper5	0.015	0	0	0	0.034
….					

Figure3.Cosinesimilarityresults

For example, in **Figure 2**:

- **Paper 1** has a TF-IDF score of **0.066** for Cluster 1, making it most relevant to that cluster.
- **Paper 2** shows higher relevance to Cluster 2, with a TF-IDF score of **0.43**.

These results confirm that most papers were appropriately classified into the correct cluster based on their content, as indicated by the high cosine similarity values.

3.4. Classification and Distribution

The system successfully classified over **96% of papers** into the correct clusters, demonstrating the efficiency of the proposed approach. **Figure 4** shows the distribution of papers across the five clusters from 2012 to 2019. The clusters were well-defined, and the classification accurately reflected the underlying topics of the research papers.

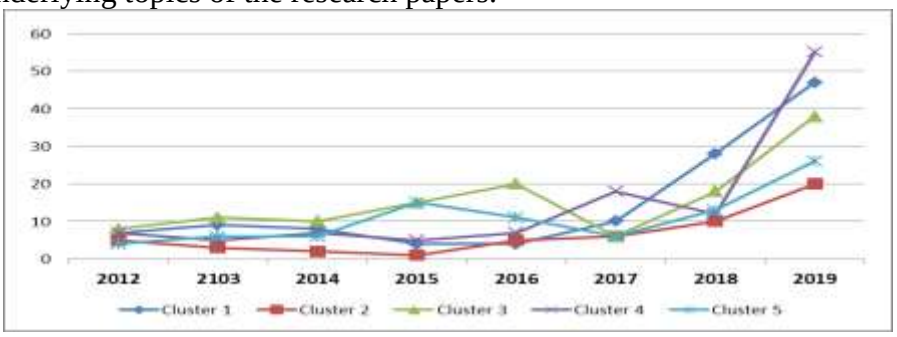


Figure4.Papersclassificationanddistribution

3.5. Validation and Evaluation

To evaluate the performance of the classification system, we used **precision** and **recall** metrics, which are common validation measures for assessing the accuracy of classifications:

- **Precision** measures the proportion of correctly classified papers relative to the total number of papers classified.

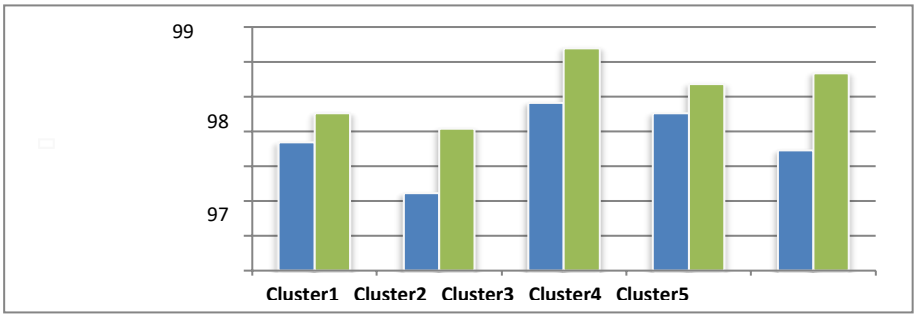


Figure5.Validationresults

- **Recall** measures the proportion of relevant papers successfully retrieved by the system.

As shown in **Figure 5**, the system performed well in labeling papers accurately. However, some papers presented challenges due to **mixed subjects**, where multiple topics or contributions were involved. These mixed-subject papers required additional consideration, as they may belong to multiple clusters.

Conclusion

The proposed classification system provides a robust method for clustering research papers using TF-IDF and cosine similarity. The results demonstrate high accuracy in matching papers to the correct clusters, offering an efficient solution for managing and retrieving research papers across various domains.

REFERENCES

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

[2] A. P. Singh, et al., “Phrase based Web Document Clustering: an Indexing Approach,” *Computer Communication,NetworkingandInternet 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 *IJCAProceedingsonInternationalConferenceonAdvancesinComputerEngineeringandApplications(ICACEA)*,vol.2, pp. 34-39, 2014.

[4] N. K. Nagwani, “Summarizing Large Text Collection Using Topic Modeling and Clustering based on MapreduceFramework,”*Journal ofBigData*,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 -ComputerandInformationSciences,*vol.27,no. 1, pp. 46-57, 2015.

[6] P.Gurung and R.Wagh,"A Study onTopic IdentificationUsing K MeansClustering Algorithm: Big vs.SmallDocuments,"*AdvancesinComputationalSciencesandTechnology,* vol.10,no. 2,pp.221-233,2017.

[7] P.B.Bafna,etal.,"DocumentClustering:TF-IDFApproach,"in*IEEE2016InternationalConferenceonElectrical,Electronics,andOptimizationTechniques(ICEEOT),*pp. 61-66,2016.

[8] N.OikonomakouandM.Vazirgiannis,"AReviewofWebDocumentClusteringApproaches,"*DataMiningandKnowledgeDiscoveryHandbook,*pp. 931-948, 2010.

[9] N.M.N.Mathivanan,etal.,"ImprovingClassificationAccuracyUsingClusteringTechnique,"*BulletinofElectricalEngineeringandInformatics,*vol.7, no.3, pp. 465-470,2018.

[10] A.S.Al-Hegamiand H.H.Al-Omaisi,"DataMining Techniques forMining QueryLogsin Web SearchEngines," *InternationalJournalofComputerScienceand Network,*vol.6,no.2,pp.2277-5420,2017.

[11] S.Girish,etal.,"MiningtheWebDataforClassifyingandPredictingUsers'Requests,"*InternationalJournalofElectricalandComputerEngineering,*vol.8, no.4, pp. 2088-8708,2018.

[12] S.Khan,etal.,"WebMining inSearchEnginesforImproving PageRank,"*InternationalJournalofSoftComputingandEngineering,*vol.5, no. 4,pp.2231-2307, 2015.

[13] M.J.H.Mughal,"DataMining:WebDataMiningTechniques,ToolsandAlgorithms:AnOverview," *InternationalJournalofAdvancedComputerScienceand 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 IntelligentEngineeringSystems,* vol.22,no. 3,pp. 177-193,2018.

[15] B.Liu,"WebDataMining:ExploringHyperlinks,Contents,andUsageData,"Springer,2011