K-Means and J48 Algorithms to Categorize Student Research Abstracts

Lee Kyung Choi[1], Kim Beom Rii[2], Han Woo Park[3]
Faculty of Information and Communication Engineering, Sungkyunkwan University[1,2,3]
25-2 Sungkyunkwan-ro, Jongno-gu
South Korea
e-mail: leekyungchoi@yahoo.com[1], krii742@yahoo.com[2], wooparkh@gmail.com[3]

Author Notification
06 March 2023
Final Revised
15 March 2023
Published
29 March 2023

To cite this document:
Choi, L. K. ., Rii, K. B. ., & Park, H. W. . (2023). K-Means and J48 Algorithms to Categorize Student Research Abstracts. International Journal of Cyber and IT Service Management, 3(1), 61–64.
Retrieved from https://iiast.iaic-publisher.org/ijcitsm/index.php/IJCITSM/article/view/125

DOI:
https://doi.org/10.34306/ijcitsm.v3i1.125

Abstract

Text mining is a rapidly growing field in computer science that is used to extract meaningful information from text data. This information can be used for various applications, such as categorizing research abstracts based on their content. This study focuses on the use of text mining techniques. The goal was to determine which algorithm was more accurate in categorizing the research abstracts. The results of the study indicated that the J48 algorithm outperformed the K-Means algorithm in terms of accuracy. This suggests that the J48 algorithm is a more effective method for categorizing research abstracts based on their content. Additionally, the findings provide insight into the use of text mining techniques for categorizing research abstracts in specific fields, such as computer science. Overall, the study demonstrates the potential of text mining techniques for analyzing and categorizing large volumes of text data. As the field of text mining continues to grow, it is likely that more applications will emerge, making it easier to extract valuable information from unstructured text data. The findings of this study can be used to improve the efficiency and accuracy of text mining techniques, particularly for categorizing research abstracts in specific fields.

***Keywords**: Text Mining, K-Means Algorithm, J48 Algorithm, Classification Method, Research Abstracts*

1. Introduction

Text mining is a new field that seeks to gather meaningful information from natural language text [1]. Text mining is different from accessing information. The main purpose of accessing information is to find text that suits your needs [2]. While the purpose of text mining is to find or obtain hidden information or new patterns from unstructured data [3]. Text mining has several algorithms, including K-Means and J48. The K-Means algorithm is an unsupervised classification method that divides data items into one or more clusters [4]. Unlike the J48 algorithm, this algorithm is a supervised algorithm that modifies the C4.5 algorithm which is one of the models of the decision tree algorithm. The K-Means algorithm is considered quite efficient [5] and the J48 algorithm has a fairly high accuracy [6]. The K-Means algorithm and the J48 algorithm in this study were used to classify student research abstracts. In this study, the classification of student abstract text is limited to the Faculty of Science and Technology, University of Raharja. This is done to provide information about the current state of research. This is because the abstract contains a brief statement about the method, results, and prospects of the research conducted [7]. The classification of abstract text is based on the object of research, namely Software Engineering, Business Intelligence, System Architecture, and Information System.

2. Research Method

The data mining procedure is divided into six stages. The six stages are document collection, text pre-processing, text transformation, attribute selection, data mining, and interpretation. Figure 1 shows these stages.

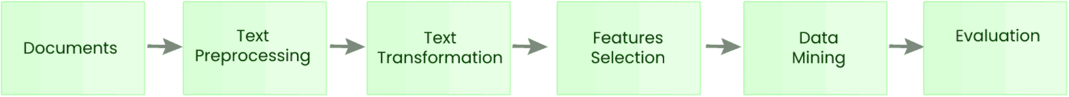

Figure 1. Text Mining Stages

The flow of this research includes a review of the literature, data collection, data analysis, and conclusions. For data analysis in this study, Weka software is used. The analysis begins with preparing research abstracts as research data, saving data in csv format, converting data into arff format, using data in arff format to be analyzed by WEKA, using modules with K-Means and J48 algorithms, and comparing the analysis results. The abstracts used in this study were written by students at the University of Raharja's Faculty of Science and Technology.

2.1 Algorithms

The K-Means algorithm has the following steps:

1. Determine the number of clusters.
2. Allocate objects into clusters randomly.
3. Calculating the centroid/average of the data in each cluster.
4. Allocate each data to the nearest centroid.
5. Return to step 3, if there is data that moves clusters or changes the centroid value, some are above the threshold value or changes in the value of the objective function used above the threshold value [4].

The J48 algorithm is a variant of the C4.5 algorithm, which is based on the ID3 algorithm. The ID3 algorithm works by constructing a tree with initial branching based on attributes that best partition objects into appropriate classes. The abstract class of IT is divided into three branches, each of which only includes abstract Business Intelligence, abstract Software Engineering, and abstract Information System.

3. Findings

This study relied on 42 abstracts from the Faculty of Science and Technology's 2018 and 2019 classes. The data is then converted to csv format using Microsoft Excel. WEKA recognizes the csv format data as primary data. Following that, the data is converted to arff format. After the data is converted to arff format, it is prepared using WEKA for text pre-processing, text transformation, and attribute selection. The data was then processed using WEKA modules to generate Tables 1 and 2.

Table 1. Results using J48 Algorithm

No.	Component	Result
1.	Incorrectly Clustered	9.313%
2.	Correctly Classified	90.687%
3.	Number leaves	8
4.	Size of tree	20
5.	Time build model	1.51 seconds

Table 2. Results using K-Means Algorithm

No.	Component	Result
1.	Incorrectly Classified	70.333%
2.	Iterations	2
3.	Time building model	0.31 seconds

Based on Tables 1 and 2, it can be concluded that the J48 algorithm is more accurate than the K-Means algorithm for classifying abstracts. However, the K-Means algorithm has a faster model building time of 0.31 seconds than J48 which is 1.51 seconds.

4. Conclusion

In conclusion, the study examined the effectiveness of two classification algorithms, J48 and K-Means The research abstracts were categorized based on their research objects, which include Software Engineering, Business Intelligence, System Architecture, and Information System. The findings reveal that the J48 algorithm outperformed the K-Means algorithm in terms of accuracy, achieving a percentage of 90.687%.

The study demonstrates the potential of text mining techniques and classification algorithms in analyzing large volumes of text data, particularly for categorizing research abstracts in specific fields. The use of WEKA software provides an effective approach to text mining and classification, and the findings can be applied to other research areas beyond the field of computer science.

Overall, the study highlights the importance of accurate categorization of research abstracts and the potential of text mining techniques to improve the efficiency of the categorization process. Future research in this area could explore the use of other classification algorithms and text mining techniques, as well as expanding the scope to include research abstracts from other faculties or universities.

References

[1] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics, 9*(8), 1295.

[2] Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access: Practical Innovations, Open Solutions, 8*, 80716–80727.

[3] Tian, K., Li, J., Zeng, J., Evans, A., & Zhang, L. (2019). Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm. *Computers and Electronics in Agriculture, 165*(104962), 104962.

[4] Zhu, E., Zhang, Y., Wen, P., & Liu, F. (2019). Fast and stable clustering analysis based on Grid-mapping K-means algorithm and new clustering validity index. *Neurocomputing, 363*, 149–170.

[5] Song, K., Yao, X., Nie, F., Li, X., & Xu, M. (2021). Weighted bilateral K-means algorithm for fast co-clustering and fast spectral clustering. *Pattern Recognition, 109*(107560), 107560.

[6] Rahardja, U., Harahap, E. P., & Dewi, S. R. (2019). The strategy of enhancing article citation and H-index on SINTA to improve tertiary reputation. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 17*(2), 683.

[7] Zarlis, M., Harahap, E. P., & Husna, L. N. (2019). Test appraisal system application based on YII Framework as media input student value final project and thesis session at higher education. *Aptisi Transactions On Technopreneurship (ATT), 1*(1), 73–81.

[8] Kartini, Santoso, S., Harahap, E. P., Khoirunisa, A., & Zelina, K. (2021). A systematic review through intellectual based blockchain-intermediary. *2021 9th International Conference on Cyber and IT Service Management (CITSM)*, 1–7.

[9] Krishnakumar, S., & Manivannan, K. (2021). RETRACTED ARTICLE: Effective segmentation and classification of brain tumor using rough K means algorithm and multi kernel SVM in MR images. *Journal of Ambient Intelligence and Humanized Computing, 12*(6), 6751–6760.

[10] Bienvenido-Huertas, D., Nieto-Julián, J. E., Moyano, J. J., Macías-Bernal, J. M., & Castro, J. (2020). Implementing artificial intelligence in H-BIM using the J48 algorithm to manage historic buildings. *International Journal of Architectural Heritage: Conservation, Analysis, and Restoration, 14*(8), 1148–1160.

[11] Adnan, M., Sarno, R., & Sungkono, K. R. (2019). Sentiment analysis of restaurant review with classification approach in the decision tree-J48 algorithm. *2019 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 121–126.

[12] Azizah, N. N., & Mariyanti, T. (2022). Education and technology management policies and practices in madarasah. *International Transactions on Education Technology, 1*(1), 29–34.

[13] Hermawan, D. R., Fahrio Ghanial Fatihah, M., Kurniawati, L., & Helen, A. (2021). Comparative study of J48 decision tree classification algorithm, random tree, and random forest on in-vehicle CouponRecommendation data. 2021 International Conference on Artificial Intelligence and Big Data Analytics, 1–6.

[14] Moodi, F., & Saadatfar, H. (2022). An improved K-means algorithm for big data. *IET Software, 16*(1), 48–59.

[15] Nandapala, E. Y. L., & Jayasena, K. P. N. (2020). The practical approach in Customers segmentation by using the K-Means Algorithm. *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 344–349.

[16] Razdan, S., Gupta, H., & Seth, A. (2021). Performance analysis of network intrusion detection systems using J48 and naive Bayes algorithms. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–7.