 ISSN:2302-8149
e-ISSN:2540-9719

# A Recommendation System for University Discussion Committees

**[1]Zaid Mundher*, [2]Manar Talat Ahmad**
[1,2]Department of Computer Science, University of Mosul, Mosul, Iraq
*e-mail: zaidabdulelah@uomosul.edu.iq

(***received*:** 30 October 2024, ***revised*:** 31 October 2024, ***accepted*:** 16 November 2024)

## Abstract

Abstract - One of the topics that have emerged and gained popularity in recent years, due to the extensive availability of data, is recommendation systems. The concept of recommendation systems is based on saving users' time and effort while using the Internet for browsing, shopping, or other web activities. On the other hand, one of the routine tasks that is consistently performed in the academic community is the selection of committees' members for the defense of master's thesis or doctoral dissertations. These committees are responsible for evaluating the graduate students' work and assessment of the academic and research efforts. In general, naming discussion committees' members is one of the challenges that used to be solved manually. In this work, a recommendation system was built to propose a discussion committee's members at Computer Science department in University of Mosul based on a dataset that includes the committees that were previously named. Two methods were introduced, developed, and tested based on content-based recommendation system techniques and cluster-based recommendation system techniques.

**Keywords:** recommendation system, similarity measures, k-means clustering

## 1 Introduction

Due to the large data that is available today, it has become essential to develop mechanisms that save user's time and effort to make it easier for users to explore this vast amount of data efficiently and effectively. Thus, recommendation systems were found to provide different solutions for various needs. Recommendation systems, also called recommendation engines, are one of the most popular and widespread topics in the field of data science in recent times. Many of the websites we visit daily use recommendation systems in one way or another. Suggested videos on YouTube or recommended products on Amazon, are nothing but practical applications of recommendation systems. In practical terms, recommendation systems are an automatic way to introduce or delivering products and items to Internet users. In fact, it is a win-win solution. Companies that implement recommendation systems may improve their sales performance, while users can discover products they might be interested in with no effort.

In general, recommendation systems can be built in various ways based on different methods and algorithms. Content-based filtering, collaborative filtering, hybrid systems, and machine learning clustering are some of the most popular methods to build and develop recommendation systems. More details about each type and how they work are available in [1], [2], [3], [4], [5], [6].

On the other side, the process of manually selecting defense committees for graduate theses is a time-consuming task for academic institutions. Traditionally, after students submit their theses, university administrators are responsible for assigning committee members based on the subject area of the submitted work. The reliance on manual processes can lead to delays in defense committee formation, which negatively affects the thesis evaluation process. Furthermore, as with any manual task, this process carries a certain degree of potential errors. For instance, mismatching between committee members' expertise and the specific topics of the theses may occured, affecting the quality of evaluations. To address these issues, this work introduces a recommendation system designed to suggest appropriate defense committee members based on the specific research focus of each submitted thesis. The implementation was performed using two different methods. The first method based on finding the similarity between the new thesis title and the existing theses in the dataset using text similarity algorithms. The second method involves clustering the existing theses in the dataset based on their titles, then determining the cluster to which the new thesis belongs. More details are explained in the next

 ISSN:2302-8149
e-ISSN:2540-9719

sections. By automating the committee selection process, the proposed recommendation system aims to improve the efficiency and effectiveness of committee assignments. This method not only reduces the administrative burden but also improves the match between the committee's expertise and the research topics of students, thereby promoting a more favorable academic experience for graduate students.

## 2 Literature Review

In recent years, the concept of recommendation systems has gained widespread popularity and momentum. Many research papers have presented various examples of recommendation systems, ranging from traditional examples to more advanced ones. Papers, such as [7], [8], [9], [10], [11], [12] discussed a traditional recommendation system example which is movies recommender system. Different approaches were introduced in these papers such as K-Means clustering, K-Nearest neighbour, and decision trees. In contrast, in [13], a book recommender system was introduced using K-Means clustering method. However, in today's world, with the vast availability of data, recommendation systems are no longer limited to marketing or e-commerce; instead, they have been adopted in a wide range of fields. For instance, researchers in [14], [15] introduced recommendation systems in the food industry to suggest foods to users applying different algorithms and techniques. Additionally, researchers in [16], [17], [18], [19] implemented recommendation systems in the tourism field to suggest tourist destinations or specific travel routes. In the academic field, recommendation systems have also been developed to suggest scientific journals or publications, as demonstrated in [2], [18], [20], [21].

The main objective of this study is to address the challenge of naming appropriate defense committee members for graduate students. Although recommendation systems have been widely applied in various fields, their use in the academic context—specifically for selecting defense committees—has not been discussed. By developing a customized recommendation system, this study aims to simplify and improve the committee assignment process, ensuring that each student receives an expert committee relevant to their research area. To the best of our knowledge, no previous work has targeted this problem with a recommendation-based solution.

## 3 Research Method

The architecture of the proposed system can be explained in the steps mentioned below:

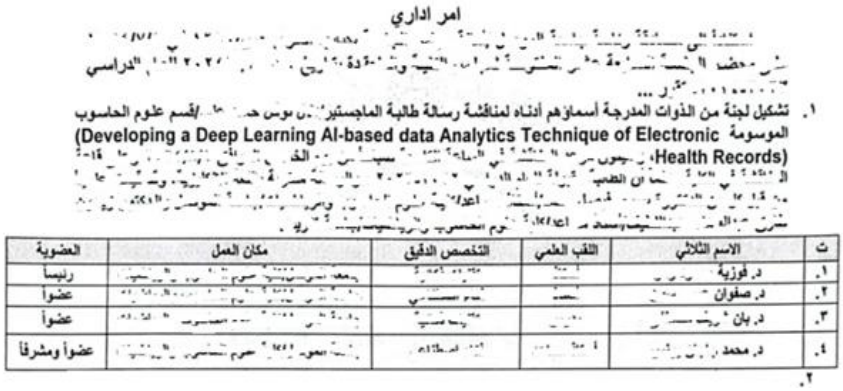
1. Data Collection
2. Data Pre-processing (Text Cleaning)
3. Feature Engineering
4. Model Building (Similarity Measures)
5. Recommend Members

Each phase involves a set of sub-steps that should be completed before progressing to the next phase.

### *A. Data Collection*

The main issue that is facing any data science project is data. In other words, data availability is the key to any data science related project. Therefore, the primary challenge of this work was to provide the required data so the next step can be implemented. As a case study, the previous defense committees from the Computer Science Department at the University of Mosul that formed over the last five years were used to build the necessary dataset. This dataset will serve as the starting point for this work. Figure 1 represents a sample of paper version of the official orders related to a defense committee for a master thesis. Based on our current knowledge, there is no dataset available that contain all the needed details, such as professor name, university affiliation, and specialization. Therefore, the initial step was to prepare and create the required dataset.

 ISSN:2302-8149
e-ISSN:2540-9719


امر اداري

دراسي

١. تشكيل لجنة من الذوات المدرجة أسماؤهم أدناه لمناقشة رسالة طالبة الماجستير
(Developing a Deep Learning AI-based data Analytics Technique of Electronic Health Records)

| ت | الاسم الثلاثي | اللقب العلمي | التخصص الدقيق | مكان العمل | العضوية |
|---|---|---|---|---|---|
| ١. | د. فوزية | | | | رئيساً |
| ٢. | د. صفوان | | | | عضواً |
| ٣. | د. بان | | | | عضواً |
| ٤. | د. محمد | | | | عضواً ومشرفاً |

٢.

**Figure 1. Paper version of the official orders of a defence committee [Some parts of the image have been intentionally obscured]**

*B. Data Pre-processing (Text Cleaning)*

The famous saying “Garbage in. Garbage out”, meaning that the output's quality depends on the quality of the input, is a fact in data science field. Thus, implementing pre-processing steps carefully and precisely is a very important issue which can have a substantial effect on the results. Therefore, before starting the actual steps of building the proposed recommendation system, pre-processing of the data was conducted. The "title" column, which contains the titles of the previous theses, is the main and most important column in this work. The following sections outline the specific text preprocessing steps applied to the "title" column, as well as their importance in the academic context and their influence on the accuracy of the recommendation system.

1. **Lowercasing**: Converting all characters to lowercase ensures uniformity and reduces redundancy caused by differences in letter case. This standardization step helps improve similarity among titles that are similar in content but may have case variations.
2. **Spelling Correction**: Typographical errors in titles were corrected to ensure linguistic accuracy, which is especially important in academic text that demands high precision.
3. **Punctuation Removal**: All punctuation was removed from titles, as it typically does not contribute to textual meaning in this context and can introduce unnecessary noise.
4. **Stop Words Removal**: Stop words (e.g., “a,” “an,” “and,” “are,” “as,” “at”) were removed as they do not add substantial semantic value and may introduce unwanted variance among texts. This step enhances model efficiency and accuracy by focusing on more meaningful words in the titles.
5. **Lemmatization**: Words were converted to their base forms through lemmatization, which reduces unnecessary variability between different forms of the same word.

By applying these text preprocessing steps, the similarity scores among academic titles are significantly improved, which in turn enhances the accuracy of the proposed recommendation system.

*C. Feature Engineering*

Generally, computers are effective when handling numerical data. Therefore, a primary and essential step to process natural language is transforming text into numerical data, referred to as features. Common algorithms for feature extraction in the domain of the Natural Language Processing (NLP) include TF-IDF, Word2Vec, FastText, and GloVe [22], [23].

In this work, TF-IDF (Term Frequency-Inverse Document Frequency), a statistical feature extraction technique, was selected to generate features and convert thesis titles into a vector-based format. TF-IDF was selected due to its effectiveness in clustering tasks, as it highlights the distinctiveness of words within documents compared to the entire dataset, thereby facilitating the differentiation of academic topics. In contrast to embedding methods such as Word2Vec or FastText, which focus on semantic similarities, TF-IDF offers a simple representation that improves the distinction of terms based on their frequency patterns. This feature is particularly beneficial for clustering, where unique term weights can more accurately reflect the differences between documents. Additional information about TF-IDF can be found in references [7], [24], [25].

 ISSN:2302-8149
e-ISSN:2540-9719

*D. Model Building (Similarity Measures)*

The primary step in this work involves comparing the title of a new thesis with previously stored theses in the dataset to identify the most similar one. Based on the closest match, a list of suggested committee members will be recommended. In this phase, two different approaches were applied, as described below

*1. Text similarity*

A fundamental method for implementing a content-based recommendation system is to apply a text similarity algorithm to calculate similarities among item descriptions. Various methods can measure similarity, including Euclidean distance, Pearson's correlation, and Cosine Similarity. In this study, Cosine Similarity was chosen due to its effectiveness in capturing the textual similarity between academic titles, which often rely on unique terms and phrases. Cosine Similarity calculates the angle between two vectors, representing the similarity without regard to the magnitude. Thus, a smaller angle between vectors signifies a higher similarity between thesis titles. The mathematical basis for Cosine Similarity is detailed in [5], [11], [25]. Figure (2) illustrates the primary behavior of this approach.

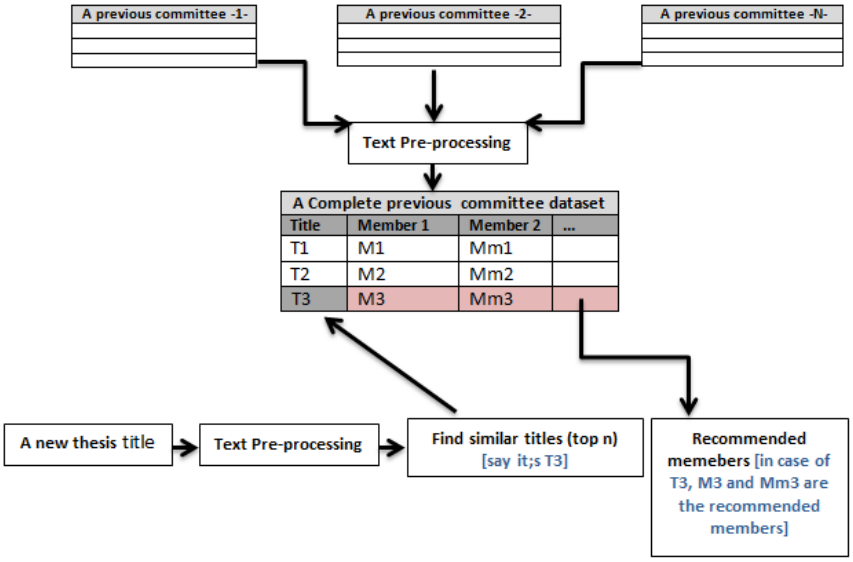



**Figure 2. Our content-based recommendation behaviour**

*2. K-means clustering*

In general, clustering is one of the most used methods to build recommendation systems. As a concept, clustering is an unsupervised machine learning technique that aims to divide data into different groups so that items in each group resemble each other more closely [13]. There are many clustered algorithms such as K-Means clustering, DBSCAN, and Mean Shift Clustering. K-Means can be considered as the most popular machine learning algorithms that groups items based on distance between points. The main concept of this approach is to cluster thesis title into different groups. In this work, K-Means clustering was applied to group thesis titles into cohesive clusters to support recommendation accuracy. After experimenting with various cluster sizes, five clusters were selected to provide a balance between interpretability and accuracy. This number captures the diversity of research topics without excessive fragmentation, allowing each cluster to represent a coherent theme. Determine the cluster that a new title belongs to, will be the job of the proposed recommendation engine.

*E. Recommend Members*

As a final step, a list of committee members is suggested. These names were recommended based on their prior involvement in discussing theses that are closely related to the new discussion topic for which the committee is being formed. In the case of using a text similarity algorithm, the names of the committee members for the most three similar theses are recommended based on the thesis title. In contrast, when using the K-means clustering algorithm, the names of the committee members from the cluster to which the new thesis belongs, are recommended.

 ISSN:2302-8149
e-ISSN:2540-9719

## 4 Implementation

To implement the proposed system, Python was employed along with several data analysis libraries such as Pandas and Matplotlib. The following steps outline the practical implementation of this work:

- *Required data to build the model was collected manually based on the hard copy (paper copy) of the past committees, and the data was stored in a CSV file.*
- *Vectors to represent text were created for every title using **tf-idf**.*
- *The pairwise cosine similarity value was calculated for every title*
- *The **K-Means** algorithm was implemented to cluster theses title to different five groups.*
- *A recommendation function was built that takes in a thesis title and recommend members of the most similar title.*

The CSV file that was created consists of eight features named as follows (Table 1):

**Table 1. Extracted features**

| Feature name | Description |
|---|---|
| **th_title** | the title of the thesis |
| **study** | Type of Study: PhD, Master's, Diploma |
| **date** | Date of Defence |
| **nam1** | Name of the first member |
| **nam2** | Name of the second member |
| **nam3** | Name of the third member |
| **nam4** | Name of the fourth member |
| **nam5** | Name of the fifth member |
| **sup** | Name of the supervisor |

## 5 Result and Discussion

The proposed recommendation system was fully coded, tested, and the results were analysed. Speaking of content-based recommendation approach that uses text similarity algorithm, Figure 3 shows a scenario when the thesis title "*Design a multi-level security model*" was given. . In response, the system identified the three most similar thesis titles to recommend suitable defense committee members based on topic relevance.

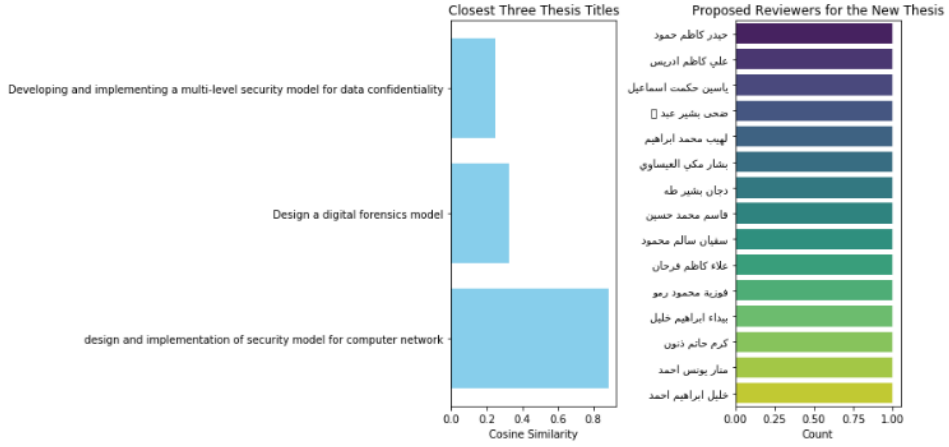



**Figure 3. A sample output of content-based recommendation system**

In contrast, our clustered-based recommendation system was also tested to evaluate the results. Figure 4 shows the K-Means clustering of theses title, while Figure 5 shows the distribution of theses across clusters.

 ISSN:2302-8149
e-ISSN:2540-9719

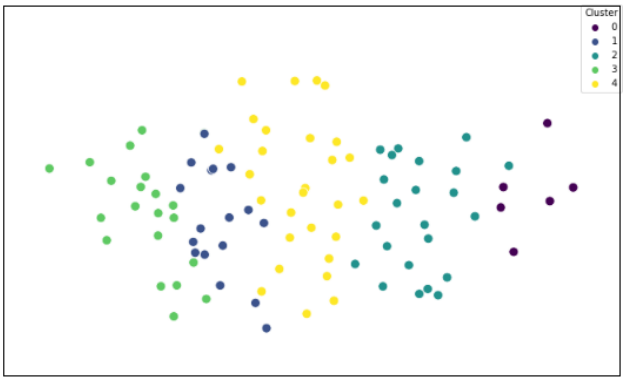



**Figure 4. The k-means clustering of theses title**

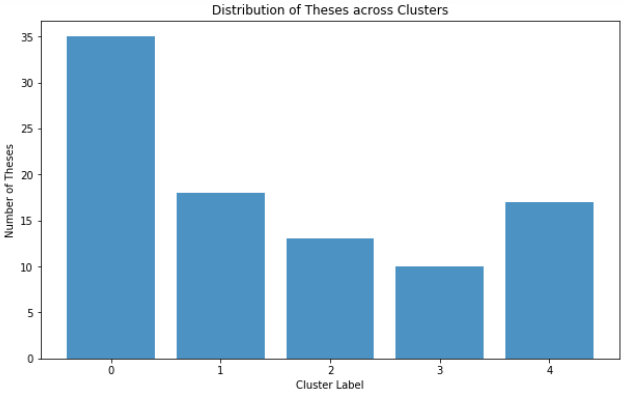



**Figure 5. The distribution of theses across clusters.**

In addition, Figures 6-7-8-9-10 illustrate top keywords in each cluster.

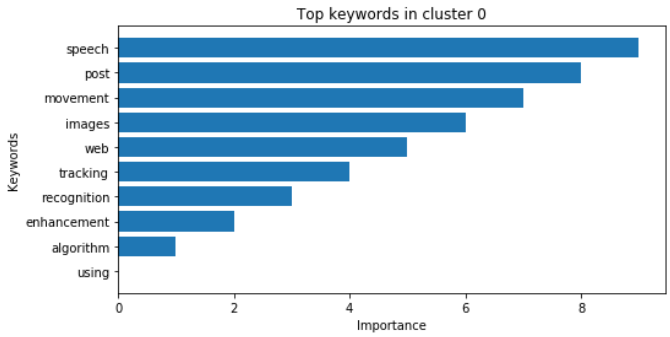



**Figure 6. Top keywords in cluster 0**

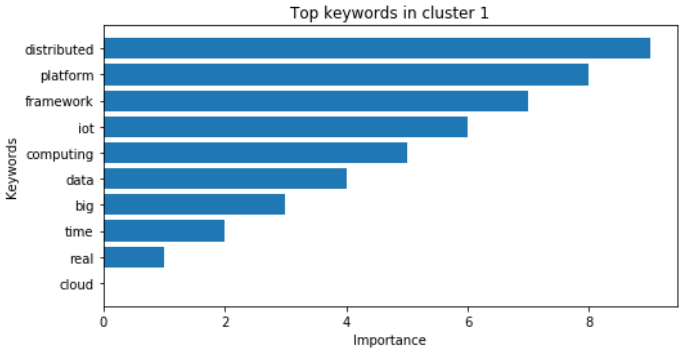



**Figure 7. Top keywords in cluster 1**

 ISSN:2302-8149
e-ISSN:2540-9719

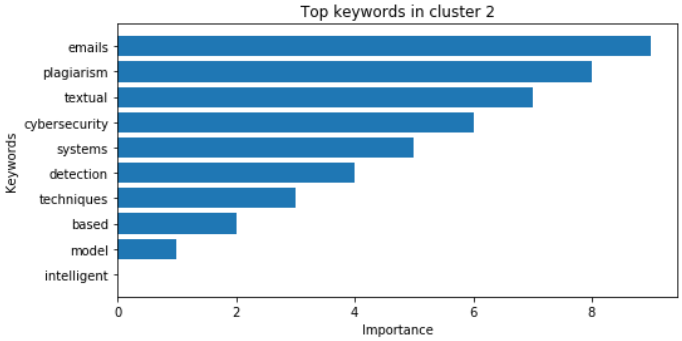



**Figure 8. Top keywords in cluster 2**

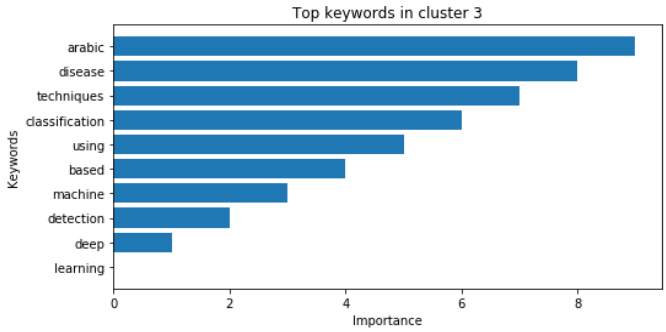



**Figure 9 Top keywords in cluster 3**

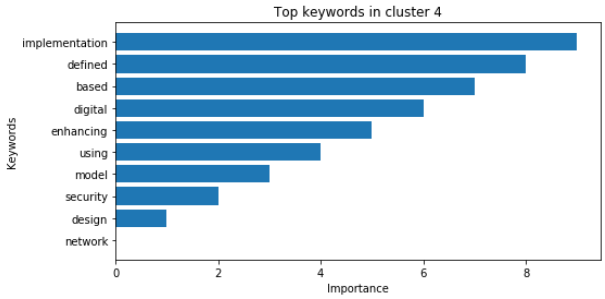



**Figure 10. Top keywords in cluster 4**

To evaluate the quality of the clustering algorithm, Silhouette Score was calculated. Silhouette Score is a value from -1 to 1 where a closer value to 1 indicates that data points that belong to same cluster are very similar. For this work, a value of 0.014 was obtained, which is relatively low. As mentioned earlier, a value close to one indicates well-defined clusters without overlap, while the value we obtained indicates considerable overlap between clusters.

*A. Evaluating the Low Silhouette Score*

The low Silhouette Score suggests that clustering based on thesis titles alone may lack sufficient distinction, leading to clusters that overlap considerably. Thesis titles are typically brief and may not capture enough unique information to create well-separated clusters, especially in academic domains where topics often merge boundaries across research areas. While the TF-IDF technique was first used to vectorize thesis titles, further experiments were conducted with Word2Vec, a technique that encodes semantic relationships between words. The Word2Vec model achieved a Silhouette Score of 0.049, showing only a slight improvement. This minor increase reflects the challenge of generating well-defined clusters based solely on title text, as Word2Vec also struggled to capture sufficient context within these short phrases. To address these limitations and achieve more accurate clustering, incorporating additional features beyond the thesis title is advisable. Features such as sub-discipline, research abstract, and even keywords or methodology descriptions could provide a more comprehensive context for clustering.

 ISSN:2302-8149
e-ISSN:2540-9719

*B. Limitations of the Study*

While this research presents a foundational approach to a content-based recommendation system for thesis committees, several limitations must be acknowledged that may affect the outcomes and generalizability of the findings.

1. **Incompleteness of Data**: One of the significant challenges encountered was the difficulty in acquiring comprehensive data on each committee member. The lack of detailed information regarding their research interests and past committee roles may undermine the accuracy of the recommendations, as the system relies on potentially incomplete data.
2. **Lack of Detailed Features**: The study primarily utilized thesis titles for clustering, which may not provide sufficient context to accurately reflect the complexity of the research topics. Incorporating additional features such as abstracts or specific keywords could enhance the model's capability to make more accurate recommendations.
3. **Dynamic Nature of Research Topics**: The static nature of the dataset may not capture the rapid developments in academic research fields. As new areas emerge and existing ones evolve, the system must be regularly updated to ensure it remains relevant and effective in its recommendations.

By clearly outlining these limitations, this study highlights areas that warrant further research and improvement. Addressing these challenges in future work will enhance the reliability and applicability of the recommendation system, providing more accurate and context-aware suggestions for thesis committee compositions.

## 6 Conclusion

In recent years, the topic of recommendation systems has gained significant importance, becoming integral to many websites. This work aims to develop a recommendation system that suggests names of defense committee members, effectively replacing what was previously a manual process. The system is designed to utilize data from committees formed in recent years, employing two widely-used methodologies: a content-based approach using cosine similarity and a clustering method utilizing the K-Means algorithm. This work can be improved by adding other factors, such as specifying the research's sub-discipline (e.g., AI, networks, etc.) and then relying on researcher data that includes more detailed academic information about the researchers, such as their specific specialties and research interests. Moreover, the results can also be improved by adding the 'Abstract' attribute, which includes a summary of a thesis, to the dataset. This attribute could be used to cluster the theses, leading to more distinct and differentiated clusters compared to relying solely on the title. By implementing these improvements, the recommendation system could become a more powerful tool for supporting graduate students in selecting appropriate defense committee members, ultimately contributing to more informed and effective academic evaluations.

## Acknowledgement

The researcher would like to express his thanks and gratitude to the College of Computer Science and Mathematics at the University of Mosul for its support and facilitation of the requirements for completing the work, especially with regard to obtaining paper version of the official orders related to a defense committee that were previously named.

## Reference

[1] B. Alhijawi and Y. Kilani, "The Recommender System: A Survey," *International Journal of Advanced Intelligence Paradigms*, vol. 10, no. 1, p. 1, 2018, doi: 10.1504/ijaip.2018.10010164.

[2] A. Al-Badarenah and J. Alsakran, "An Automated Recommender System for Course Selection," 2016. [Online]. Available: www.ijacsa.thesai.org

[3] P. M. Chawan, "Recommendation System using Machine Learning Techniques," 2022. [Online]. Available: www.irjet.net

 ISSN:2302-8149
e-ISSN:2540-9719

[4] C. Mediani, S. Harous, and M. Djoudi, *Content-Based Recommender System using Word Embeddings for Pedagogical Resources*. 2023. doi: 10.1109/PAIS60821.2023.10321989.

[5] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web, WWW 2001*, Association for Computing Machinery, Inc, Apr. 2001, pp. 285–295. doi: 10.1145/371920.372071.

[6] M. Sharma, D. Saxena, and A. K. Singh, "A Survey and Classification on Recommendation Systems." [Online]. Available: https://www.researchgate.net/publication/357562406

[7] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[8] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using k-means clustering and k-nearest neighbor," in *Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019*, Institute of Electrical and Electronics Engineers Inc., Jan. 2019, pp. 263–268. doi: 10.1109/CONFLUENCE.2019.8776969.

[9] M. B. R. Azaki and Z. K. A. Baizal, "Movie Recommender System Using Decision Tree Method," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, pp. 729–735, Aug. 2023, doi: 10.29100/jipi.v8i3.3867.

[10] M. K. Delimayanti *et al.*, "Web-Based Movie Recommendation System using Content-Based Filtering and KNN Algorithm," in *Proceedings - 2022 9th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 314–318. doi: 10.1109/ICITACEE55701.2022.9923974.

[11] Omkar Kunde, Omkar Gaikwad, Prathamesh Kelgandre, Rohan Damodhar, and Prof. Mrs. M. M. Swami, "The Movie Recommendation System using Content Based Filtering with TF-IDF¬¬-Vectorization and Levenshtein Distance," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 257–263, May 2022, doi: 10.48175/ijarsct-3648.

[12] S. Rakesh, "Movie Recommendation System Using Content Based Filtering," *Al-Bahir Journal for Engineering and Pure Sciences*, vol. 4, no. 1, Dec. 2023, doi: 10.55810/2313-0083.1043.

[13] R. Rani and R. Sahu, "Ijesrt International Journal Of Engineering Sciences & Research Technology Book Recommendation Using K-Mean Clustering And Collaborative Filtering," *Int J Eng Sci Res Technol*, doi: 10.5281/zenodo.1042100.

[14] M. B. S. Siddik and A. T. Wibowo, "Collaborative Filtering Based Food Recommendation System Using Matrix Factorization," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, p. 1041, Jul. 2023, doi: 10.30865/mib.v7i3.6049.

[15] R. Singh and P. Dwivedi, "Food Recommendation Systems Based On Content-based and Collaborative Filtering Techniques," in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCNT56998.2023.10307080.

[16] D. F. Gimnastian, R. Yasirandi, and D. Oktaria, "Analysis and Design of a Route Recommendation System and Bicycle Rental Fees at Tourist Destinations with Genetic

 ISSN:2302-8149
e-ISSN:2540-9719

Algorithms," *Jurnal Media Informatika Budidarma*, vol. 6, no. 2, p. 837, Apr. 2022, doi: 10.30865/mib.v6i2.3749.

[17] Y. Jiang, Y. Zhang, Z. Li, W. Yu, H. Wei, and L. Yuan, "Tourist Attraction Recommendation System Based on Django and Collaborative Filtering," 2024, pp. 226–235. doi: 10.1007/978-981-97-0827-7_20.

[18] A. A. Khan and M. Chowdhury, "TourMate: A Personalized Multi-factor Based Tourist Place Recommendation System Using Machine Learning," *International Journal of Intelligent Systems and Applications*, vol. 16, no. 4, pp. 55–71, Aug. 2024, doi: 10.5815/ijisa.2024.04.04.

[19] G. Ratnakanth and S. Poonkuzhali, "Indian Tourist Recommendation System Using Collaborative Filtering and Deep Autoencoder," 2022, pp. 341–356. doi: 10.1007/978-981-19-0098-3_34.

[20] T. Ajagbe, B. S. Aribisala, O. Olabanjo, and B. Aribisala, "A Recommender System For Academic Publications Using Content-Based Filtering Techniques Full Paper A Recommender System For Academic Publications Using Content-Based Filtering Techniques." [Online]. Available: https://www.researchgate.net/publication/357240718

[21] D. Wang, Y. Liang, D. Xu, X. Feng, and R. Guan, "A content-based recommender system for computer science publications," *Knowl Based Syst*, vol. 157, pp. 1–9, Oct. 2018, doi: 10.1016/j.knosys.2018.05.001.

[22] A. Amalia, O. Salim Sitompul, E. Budhiarti Nababan, and T. Mantoro, "A Comparison Study Of Document Clustering Using Doc2vec Versus Tfidf Combined With Lsa For Small Corpora," *J Theor Appl Inf Technol*, vol. 15, p. 17, 2020, [Online]. Available: www.jatit.org

[23] A. D. Susanto, S. Andrian Pradita, C. Stryadhi, K. E. Setiawan, and M. Fikri Hasani, "Text Vectorization Techniques for Trending Topic Clustering on Twitter: A Comparative Evaluation of TF-IDF, Doc2Vec, and Sentence-BERT," in *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICORIS60118.2023.10352228.

[24] P. Desai, N. Telis, B. Lehmann, K. Bettinger, J. K. Pritchard, and S. Datta, "SciReader: A Cloud-based Recommender System for Biomedical Literature," May 30, 2018. doi: 10.1101/333922.

[25] Z. Mundher, W. Khater, and L. Ganeem, "Adopting Text Similarity Methods and Cloud Computing to Build a College Chatbot Model," *Journal Of Education And Science*, vol. 30, no. 1, pp. 117–125, Mar. 2021, doi: 10.33899/edusj.2020.127244.1079.