Paradigma, Vol. 27, No. 2, September 2025, P-ISSN 1410-5063, E-ISSN: 2579-3500, Page. 74-82
Published: LPPM Universitas Bina Sarana Informatika

PARADIGMA

website: http://jurnal.bsi.ac.id/index.php/paradigma/

K-Means++ and TF-IDF for Grouping Library Books by Topic

Jessicha Putrianingsih Pamput[1], Aindri Rizky Muthmainnah[2], Andi Akram Nur Risal[3], Dewi Fatmarani Surianto[4]

[1,2,3,4]Computer Engineering, Universitas Negeri Makassar, South Sulawesi, Indonesia

ARTICLE INFORMATION

Artikel History:
Received: February 6, 2025
Revised: March 18, 2025
Accepted: August 19, 2025
Available Online: Sept. 30, 2025

Keyword:

Cluster
K-Means++
Library
TF-IDF
Silhoutte Coefficient

ABSTRACT

The grouping of library materials in the Department of Informatics and Computer Engineering (JTIK) at Universitas Negeri Makassar (UNM) is still conducted using a conventional system that relies on predefined categories and librarian intuition. This approach often leads to inconsistencies in book categorization, making it difficult for users to find relevant references efficiently. To address this issue, this research applies the K-Means++ clustering method, which optimizes centroid initialization for more accurate cluster formation. Books are grouped based on the TF-IDF weighting matrix, resulting in six distinct clusters characterized by unique centroid values. Analysis of the top 10 words per cluster highlights dominant topics within each group. The clustering quality was evaluated using the Silhouette Coefficient, with the highest value of 0.04299, indicating a well-separated cluster structure. These findings demonstrate that K-Means++ effectively organizes books based on word similarity, enhancing library material management and improving information retrieval in the JTIK library.

Corresponding Author:

Dewi Fatmarani Surianto,
Computer Engineering, Faculty of Engineering,
Universitas Negeri Makassar,
Jl. Mallengkeri Raya, Makassar, Indonesia, 90224,
Email: dewifatmaranis@unm.ac.id

INTRODUCTION

Libraries act as centres of information services that are available to all, with the aim of facilitating access to the information needed. In addition, libraries support the development of knowledge and enrich the horizons of the communities they serve (Anggi Riyanto, Daryanto, and Abdurrahman 2022),(Anggraeni et al. 2021). In an academic context, libraries act as information centres that provide a variety of resources to support students' learning and research (Aditomo Mahardika Putra et al. 2023). A well-organised library not only offers a comprehensive collection of library materials, but also ensures that these materials are efficiently accessible and relevant according to user needs (Firmansyah, Poningsih, and Retno Andani 2022), (Anggraini, Sumantri, and Jamil. Khoirul 2024). However, library materials are sometimes still not well organised because the entire collection is still managed without a structured automated system.

Every progress of the library is inseparable from technological innovation and breakthroughs, such as the application of the Internet of Things, the application of big data technology, the application of cloud computing, the application of artificial intelligence, etc., to provide assistance for the library from passive service to active service (Xu and Shang 2024).

In the era of evolving technology, libraries face challenges in managing collections that continue to grow every time, thus requiring innovative solutions to improve the efficiency and effectiveness of resource management (Firmansyah et al. 2022),(Fransiska 2023). Grouping books by topic is one important aspect that affects the ease of searching for information (Anggi Riyanto et al. 2022),(Dea Mustika, Zakir, and Rizmi 2022). In this case study, the Department of Informatics and Computer Engineering (JTIK) at Makassar State University (UNM) also has challenges in terms of grouping its library materials. Based on observations and interviews, the JTIK library manager said that the process of grouping books in the JTIK library is still done without an automated system and tends to be based on intuition.

DOI: https://doi.org/10.31294/p.v27i2.8272

Based on the main problem faced, grouping books based on topics can help students find references that suit their needs. Some previous studies have shown that clustering books based on certain characteristics is more efficient in generating information, especially when dealing with large volumes of library data. For example, a study grouped books by title to help librarians manage book placement (Nur Afifah and Nurdiyanto 2023). Another study grouped books by title in several categories, with the aim of maximising the process of managing the book collection (Haryani, Nofriansyah, and Mariami 2021). Another study (Hasanah and Purnomo 2022) conducted book grouping based on the number of reading enthusiasts. In addition, there is an application of clustering methods on library materials based on the number of books that are often borrowed (Nasir 2021). In fact, there is research that performs clustering based on categories in 98 documents by giving Term Frequency-Inverse Document Frequency (TF-IDF) weighting to book abstracts (Widaningrum et al. 2022). Furthermore, research (Siburian et al. 2022) resulted in the clustering of library books based on borrowing data using the k-means method.

Therefore, this research can focus on K-Means++ clustering by performing Term Frequency-Inverse Document Frequency (TF-IDF) weighting on book titles. Compared to other methods, such as standard K-Means which has less optimal initial centroid ownership and kierarchical clustering which tends to have high computational complexity, K-Means++ offers better stability and produces more accurate and efficient clustering (Du et al. 2021),(Zhao 2022). K-Means++ algorithm is a method that randomly selects the best initial centroid for K-Means clustering, then the next centroid is selected based on the farthest distance with a weighted probability approach, until the number of K centroids is reached for clustering (Aggarwal and Reddy 2018),(Daoudi et al. 2021), (Kenger et al. 2023). Meanwhile, the TF-IDF algorithm assesses the importance of terms in documents based on their frequency and rarity across documents (Manning, Raghavan, and Schütze 2008)(Lan 2022). Based on previous studies that generally cluster books based on borrowing data or simple categories, this study applies a combination of text weighting and clustering. Through this approach, clustering was performed on 701 book titles, resulting in a structured clustering system, improving efficiency in academic reference search, and reducing inconsistencies in the current book clustering process. From the application of these two algorithms, this research is expected to provide recommendations in optimising library management and improving information accessibility for students and other users, and can be a support in academic activities at JTIK UNM.

RESEARCH METHOD

The steps in this research were designed with the aim of achieving a comprehensive understanding and providing clear guidance in the implementation of the research. The sequence of steps is presented in Figure 1 below.

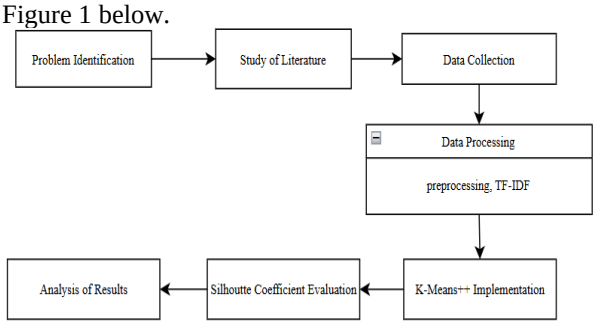

Figure 1. Research Stages

1. Problem Identification

At this stage, problem identification is carried out with the aim of gaining an understanding of book grouping in the library of the Department of Informatics and Computer Engineering (JTIK), Makassar State University (UNM) which is less efficient in organising book collections. This happens because the grouping method is still done without an automated system and based on intuition. In addition, the number of collections continues to increase every day, so it is necessary to implement a clustering method to optimise book grouping and improve library operational efficiency.

2. Literature Study

At this stage, a literature study is conducted by utilising sources from books, reviews of national and international journal articles relevant to the research topic. The purpose of this stage is to gain an in-depth understanding of the research topic under study. The articles analysed have an important role in understanding previous research, identifying its shortcomings, as well as presenting a strong knowledge base for continuing further research.

3. Data Collection

The data collection process in this research was conducted by observation and interviewing the library manager to understand the book management in JTIK library deeply. The main data used was obtained directly from the relevant library database, including information about 701 available books.

4. Data Processing

The data processing stage in this research was conducted in 2 stages, namely data preprocessing and TF-IDF weighting.

a. Data Preprocessing

Data preprocessing is an important first step in data processing to ensure the quality and consistency of data before further analysis, where this process implements lower case folding, stopword removal, regular expression (regex), and tokenisation. These steps aim to simplify the data so that it is easier to process.

b. TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is the second step after data preprocessing,

used to determine the importance of a word in a document relative to a set of documents (Apriliyani et al. 2024). It quantifies this importance by multiplying the term's frequency in the document by its inverse frequency across the entire collection (Bashir, Bichi, and Adamu 2024).

$$tf(t,d) = \frac{Number\ of\ occurrences\ of\ word\ t\ in\ doc}{Total\ number\ of\ words\ in\ documen} \quad (1)$$

$$idf(t) = log\frac{Total\ number\ of\ documents\ in}{Number\ of\ documents\ where\ word\ t} \quad (2)$$

$$tf - idf(t,d) = tf \times idf \quad (3)$$

5. K-Means++ Implementation

The K-Means++ method is an improvement of the k-means method by selecting initial cluster centroids based on distance probabilities, so that clustering results are more optimal and convergence is faster. After the initial centroid is selected, the standard k-means algorithm continues to cluster the data (Dea Mustika and Zakir 2022). The number of clusters was set at 6 to simplify the 12 categories of library books that have similar meanings, in order to form a more organised and efficient structure.

$$d_i = \max(j: 1 \rightarrow m) \,||\, xi - C_j\,||^{22} \quad (4)$$

6. Silhoutte Coefficient Evaluation

The evaluation stage of the optimal number of clusters is carried out by the silhouette coefficient method which will measure the quality of clustering by assessing how well the data points are grouped in clusters. The value will be calculated for each data based on the average distance into the same cluster and the average distance to the point in the nearest cluster (Pamput et al. 2024).

$$s = \frac{b-a}{\max(a,b)} \quad (5)$$

7. Analysis of Results

Result analysis is the process of evaluating the effectiveness of book clustering by the appropriateness of book clustering with relevant topics, using metrics such as silhouette coefficient to measure cluster quality and ensure books are effectively grouped based on topic similarity and provide insights for improved library collection management.

RESULTS AND DISCUSSION

At this stage, the discussion will focus on the data collection and processing process carried out to achieve the data grouping results. This process is very important to ensure that the data obtained has relevance and can be processed with the right method to produce groups that match the purpose of the analysis or research.

1. Data collection

Data collection in this research was conducted through observation and interviews with JTIK library administrators to obtain in-depth information related to the book management system and process. In addition, the main data used in this research was obtained directly from the book database in the library with a total of 701 books, thus ensuring that the data used is actual and relevant. The data collected is represented in Table 1 below.

Table 1. Data Collection Results

Code	Title	Author	Place and Year
TI-001	*Integrasi Teknologi Informasi Dengan Strategi*	*Dr. Ike Janita Dewi, MBA*	*Yogyakarta, 2005*
TI-002	*Perancangan Tata Kelola Teknologi Informasi*	*Bambang Gunawan*	*Yogyakarta, 2018*
TI-003	*Pengantar Teknologi Informasi Untuk Bisnis*	*M. Suyanto*	*Yogyakarta, 2005*
….	…	…	…
GM-006	*Gamification Membuat Belajar Seasyik Bermain Game*	*Noralia Purwa Yunita*	*Yogyakarta, 2022*
GM-007	*Aplikasi Android Game Pembelajaran Berbasis RPG Maker*	*Wahyu Hari Kristiyanto*	*Yogyakarta, 2020*
GM-008	*Aplikasi Android Game Pembelajaran Augmented Reality Berbasis Unity*	*Wahyu Hari Kristiyanto*	*Yogyakarta, 2020*

2. Data Processing

From the 701 books data collected, the data processing stage will be carried out by preprocessing the book title column, which includes lower case folding to convert all text into lowercase letters, removal of common words that are not meaningful using the stopword removal method, application of regular expression (regex) to clean data from irrelevant characters or symbols, and tokenisation to break the text into smaller word units. The following book title preprocessing results are presented in Table 2 below.

Table 2. Book Title Preprocessing Result

Title_Clean
integrasi teknologi informasi strategi
perancangan tata kelola teknologi informasi
pengantar teknologi informasi bisnis

Title_Clean
...
gamification belajar seasyik bermain game
aplikasi android game pembelajaran berbasis rpg maker
aplikasi android game pembelajaran augmented reality berbasis unity

After going through the preprocessing stage, the data is further processed using the TF-IDF (Term Frequency-Inverse Document Frequency) method to produce a weighting matrix. The matrix consists of 701 rows representing book titles and 368 columns representing unique words in the dataset. The weight of each word in the matrix indicates its level of relevance to a particular document, with higher weights indicating the word is more specific and significant in the context of that document.

3. K-Means++

After TF-IDF weighting, the K-Means++ method is used to cluster the data based on the similarity of the words in the TF-IDF matrix. The data was grouped into 6 clusters, which were determined based on the consideration that the number of book categories previously compiled by the library, namely 12 groups, was considered too large. Some of these categories had similar meanings, so they were simplified to 6 clusters to form a more organised structure, each reflecting a particular pattern in the book title documents. Each cluster has a centroid that represents the average characteristics of the words in the group. The result of clustering with the K-Means++ method is shown by the different patterns in each cluster, as reflected by its centroid value. This process can identify patterns in the data generated from the TF-IDF matrix, where each cluster represents the similarity of words in the analysed book titles.

4. Silhoutte Coefficient Evaluation

After calculating the centroid value for each cluster that has been determined, the next step is to evaluate the clusters formed using the silhouette coefficient method. This method is used to measure how good the data in a cluster is compared to other clusters. This evaluation can help in determining whether the number of clusters selected is optimal or needs to be adjusted. Table 3 below presents the calculation results of the average silhouette coefficient value obtained from evaluating the number of clusters tested.

Table 3. Average value of sillhoutte coefficient

Cluster	Average Value of Sillhoutte Coefficient
2	0.023009804634925433
3	0.027678041901485453
4	0.03441300017592933
5	0.04123615653859793
6	0.04298953240434177
7	0.04271530960938787
8	0.024558336200398042
9	0.02652943063091576
10	0.025137188167192973

The number of clusters was determined based on the highest Silhouette Coefficient value, which reached 0.04298953240434177 at six clusters. This indicates that six clusters provide the optimal balance between separation and cohesion. To clarify the visualisation of the average silhouette coefficient value, a graph illustrating the silhouette coefficient value for the number of clusters is presented in Figure 2 below.

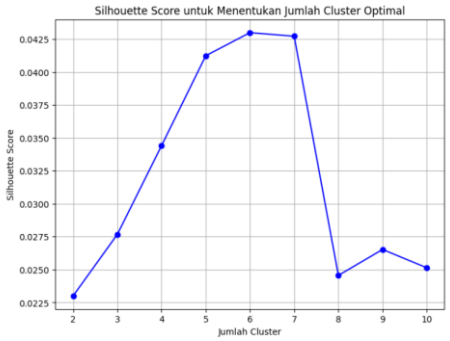

Figure 2. Graph of Average Value of Silhoutte Coefficient

After determining the optimal number of clusters, the number of clusters will be used as a reference in further analysis of the results of the implementation of the K-Means++ method. With the number of clusters that have been determined, in-depth analysis can be carried out to identify patterns contained in each cluster. Table 4 below presents the results of data grouping based on the clusters that have been determined.

Table 4. Cluster Calculation Result

Cluster	*Title_Clean*
Cluster 0	*integrasi teknologi informasi strategi*
	perancangan tata kelola teknologi informasi
	pengantar teknologi informasi bisnis
	...
	animasi game visual basic
	game dimensi game maker studio
	gamification belajar seasyik bermain game
Cluster 1	*panduan cepat belajar htmlphp mysql*
	perancangan sistem informasi aplikasinya
	sistem informasi kursus berbasis web
	...
	optimalisasi jaringan komputer kabel nirkabel
	konsep penerapan ip versi membangun jaringan komputer
	panduan praktis mini games android adobe animate cc
Cluster 2	*panduan implementasi model pembelajaran berbasis augmented virtual reality*
	asesmen pembelajaran
	penilaian pembelajaran bahasa sastra indonesia

Cluster	*Title_Clean*
	...
	modelmodel pembelajaran inovatif efektif
	aplikasi android game pembelajaran berbasis rpg maker
	aplikasi android game pembelajaran augmented reality berbasis unity
Cluster 3	*Penelitian Tindakan Kelas Edisi Revisi*
	Panduan Memahami Metodologi Pendidikan
	Metodologi Penitian
	...
	Metodologi Penelitian
	Kupas Tuntas Metode Penelitian
	Metodologi Penelitian
Cluster 4	*buku sakti pemrograman web*
	pemrograman dasar
	langka mudah pemrograman android app inventor ultimate
	...
	menguasai pemrograman arduino robotik
	dasardasar pemrograman net
	teknik penulisan tugas skripsi pemrograman
Cluster 5	*evaluasi pembelajaran*
	evaluasi kurikulum
	evaluasi pembejaran
	evaluasi hasil belajar
	dasardasar evaluasi pendidikan
	kecanduan internet panduan konseling petunjuk evaluasi penanngan

To determine the characteristics of each cluster, it is necessary to identify the top 10 words with the highest weight in each cluster. These words are taken based on the weight value at the centroid of each cluster generated by the K-Means++ method. These 10 words will reflect the themes or topics that dominate the documents in that cluster. Table 5 below presents the top 10 words for each cluster based on the highest weights generated by the K-Means++ method.

Table 5. Top 10 Words of Each Cluster

Cluster 0	*belajar*	0.0346
	aplikasi	0.0259
	pemula	0.0229
	photoshop	0.0229
	mudah	0.0223
	microsoft	0.0222
	komputer	0.0189
	pendidikan	0.0189
	data	0.0186
	learning	0.0186
Cluster 1	*sistem*	0.2242
	panduan	0.1598
	jaringan	0.1533
	komputer	0.1005
	membangun	0.0812
	praktis	0.0550
	informasi	0.0544
	android	0.0458
	lengkap	0.0448
	berbasis	0.0395
Cluster 2	*pembelajaran*	0.4349
	berbasis	0.0871
	media	0.0814
	multimedia	0.0582
	modelmodel	0.0579
	inovatif	0.0575
	strategi	0.0572
	teknologi	0.0561
	model	0.0533
	aplikasi	0.0512
Cluster 3	*penelitian*	0.4278
	metodologi	0.2504
	metode	0.1674
	kualitatif	0.1233
	kuantitatif	0.0919
	pendidikan	0.0850
	praktis	0.0385
	pendekatan	0.0367
	dasardasar	0.0364
	tuntas	0.0350
Cluster 4	*pemrograman*	0.4034
	web	0.1108
	algoritma	0.1017
	php	0.0705
	belajar	0.0697
	java	0.0536
	pemula	0.0353
	buku	0.0340
	arduino	0.0333
	mudah	0.0328
Cluster 5	*evaluasi*	0.7513
	kurikulum	0.1258
	petunjuk	0.0996
	dasardasar	0.0990
	pembelajaran	0.0930
	belajar	0.0860
	internet	0.0810
	pendidikan	0.0798
	panduan	0.0617
	bahasa	0.0000

Based on the 10 words in each cluster helps identify dominant themes or topics, providing a clearer picture of the patterns and relationships in the data. These key words also support the interpretation of the clustering results and uncover potential applications of the data. In addition, visualisation with word clouds is used to represent words based on their frequency or weight, where words with higher weights are displayed in larger sizes.

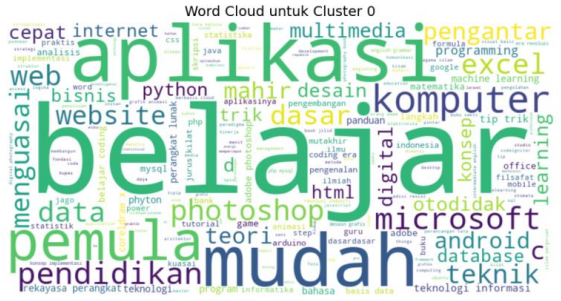

Figure 3. Word Cloud Cluster 0

The results of identifying cluster 0 based on the top 10 words and word cloud show the dominance of words such as learning, application, beginner, photoshop, easy, microsoft, computer, education, data, and learning. This pattern illustrates that cluster 0 includes 478 books that focus on introducing and learning basic computer technology for beginners. The word cloud visualisation for this cluster is presented in Figure 3.

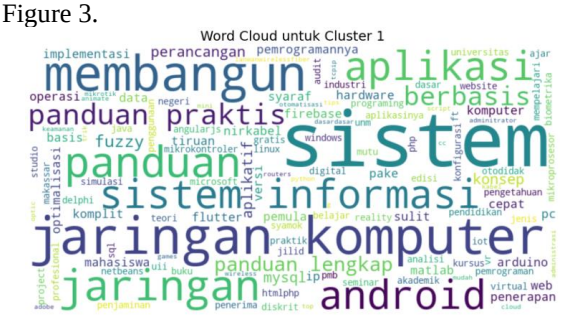

Figure 4. Word Cloud Cluster 1

Cluster 1, with 61 books, covers the topic of building systems based on information technology and computer networks. Based on the top 10 words and the word cloud in Figure 4, words such as system, guide, network, computer, build, practical, information, android, complete, and based indicate a focus on practical guidance as well as implementation of network-based systems, android applications, and management of information technology. This cluster is relevant for learning and professional applications in the IT field.

Figure 5. Word Cloud Cluster 2

Cluster 2, which consists of 42 books, focuses on the development of innovative multimedia and technology-based learning. Based on the analysis of the top 10 words and the word cloud in Figure 5, words such as learning, based, media, multimedia, model, models, innovative, strategy, technology, and application show the main focus on the integration of modern technology in the learning process. Topics in this cluster include the use of interactive media, the development of technology-based learning models, innovative strategies in teaching, and multimedia applications to improve the effectiveness and quality of education.

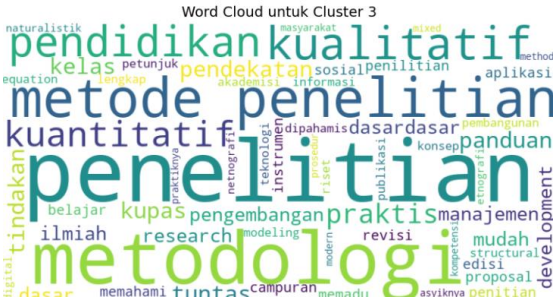

Figure 6. Word Cloud Cluster 3

Cluster 3, consisting of 34 books, focuses on research methodology in education, with an emphasis on qualitative and quantitative methods. Based on the analysis of the top 10 words and the word cloud in Figure 6, words such as research, methodology, method, qualitative, quantitative, education, practical, approach, fundamentals, and thorough reflect the depth of coverage. Topics in this cluster cover basic theories, practical approaches, as well as the systematic application of research methodologies to produce thorough, targeted, and relevant studies in the context of education.

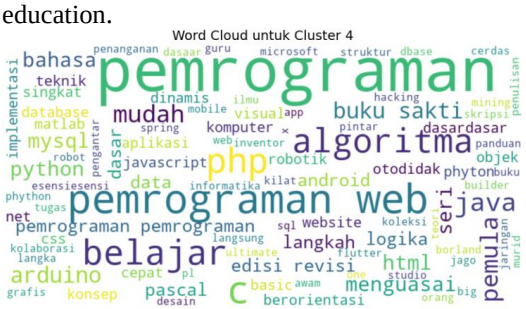

Figure 7. Word Cloud Cluster 4

Cluster 4, consisting of 80 books, focuses on computer programming and web development, covering various programming languages such as PHP and Java, as well as algorithm concepts. Based on the analysis of the top 10 words and the word cloud in Figure 7, words such as programming, web, algorithm, PHP, learn, Java, beginner, book, Arduino, and easy show a wide range of coverage, from programming basics for beginners to advanced levels. Topics in this cluster include practical guides to learning programming, project development such as Arduino, as well as easy-to-understand approaches to learning algorithms and building web applications.

Figure 8. Word Cloud Cluster 5

Cluster 5, which consists of 6 books, focuses on developing teachers' skills in technology-based learning and modern education. Based on the top 10 words and the word cloud in Figure 8, words such as evaluation, curriculum, instruction, basics, learning, learning, internet, education, guide, and language indicate discussions on technology-based curriculum development, learning evaluation, as well as the use of the internet as a medium to support teaching. The books also provide practical guidance for teachers to improve their competence in utilising technology and developing more effective teaching strategies. This topic is very relevant in supporting the transformation of education towards a more interactive and adaptive to the digital era.

Figure 9 below presents a scatter visualisation of the clustering results, which shows the distribution of data in 6 clusters based on TF-IDF matrix analysis. Cluster 0 shows a predominance of high data density in the centre, while cluster 1 and cluster 2 have a more dispersed distribution. Cluster 3 covers a more isolated area, showing unique characteristics, while cluster 4 consists of a small amount of data concentrated in a specific area. Cluster 5 has a very limited distribution of data. This visualisation reflects the pattern of data grouping based on different characteristics between clusters.

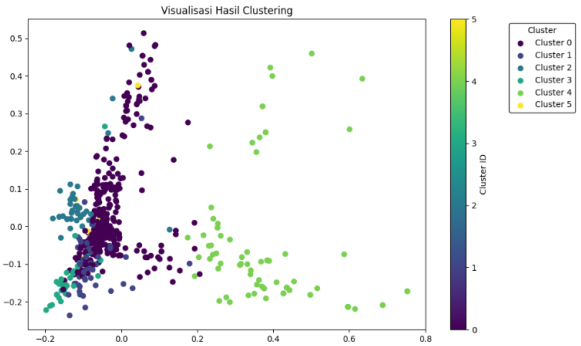

Figure 9. Scatter Visualisation of Clustering Results

The results of the clustering analysis on the 6 groups of books show that each cluster covers topics related to technology and education. Cluster 0 covers a basic introduction to computer applications, while cluster 1 focuses on the development of information technology-based systems. Cluster 2 examines the application of technology in learning and innovative curriculum development, and cluster 3 contains books that discuss research methodology in education. Cluster 4 covers computer programming and web application development, while cluster 5 provides guidance for teachers to integrate technology in the learning process. The scatter clustering visualisation shows the distribution of data that illustrates the relationship between clusters, helping to understand the characteristics and patterns that exist in each group of books.

The clustering results not only identify patterns of word similarity in book titles, but also have practical implications for library management and ease of access for users. For librarians, the information from the clustering can be used to structure the collection layout in a more structured manner and serve as a basis for decision-making regarding book procurement. In addition, the clustering results provide a structured organization of books, allowing librarians to better manage book placement and support user searches. The practical implication includes improving user experience and supporting the development of a recommendation system. Meanwhile, for library users, this system can improve efficiency in reference search and support the development of a more targeted book recommendation system.

CONCLUSION

The results showed that the K-Means++ method successfully grouped the book data into 7 clusters with unique characteristics, reflected by their centroid values and thematic patterns. Analysis of the top 10 words in each cluster revealed the dominant topic, while word cloud visualisation helped to clarify the central theme of each group. This method has the potential to be applied in digital library systems to improve accuracy in collection organisation and optimise automatic book searches, thus supporting the efficiency of large-scale library management. In addition, the clustering results can be used as a basis in the development of a topic-based book recommendation system, contributing to an improved user experience in accessing more relevant literature.

Future research can integrate additional metadata, such as abstracts and keywords, and apply the Latent Dirichlet Allocation (LDA) method to improve clustering accuracy. In addition, exploration of more complex machine learning models and implementation in real-time digital library systems can be done to optimise collection management and book search.

REFERENCES

Aditomo Mahardika Putra, Rio, Dian Pratiwi, Galuh Pramita, and Fajar Dewantoro. 2023. "Implementasi Perpustakaan Digital Di SMK Negeri 1 Trimurjo, Kabupaten Lampung Tengah." *JEIT-CS* 1(3):180–86. doi: 10.33365/jeit-cs.v1i3.230.

Aggarwal, Charu C., and Chandan K. Reddy. 2018. *Data Clustering Data Clustering Algorithms and Applications Chapman & Hall/CRC Data Mining and Knowledge Discovery Series Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*. Taylor & Francis Group.

Anggi Riyanto, Alfathan, Daryanto Daryanto, and Ginanjar Abdurrahman. 2022. "Text Mining Untuk Clustering Buku Di Perpustakaan Menggunakan Metode K-Means." *National Multidisciplinary Sciences* 1(6):835–45. doi: 10.32528/nms.v1i6.239.

Anggraeni, Diah Bekti, Widyastuti Widyastuti, Fitri Puji Rahmawati, and Madya Giri Aditama. 2021.

"Pengembangan Sistem Klasifikasi Kepustakaan Dengan Dewey Decimal Classification (DDC)." *Buletin KKN Pendidikan* 3(2):152–60. doi: 10.23917/bkkndik.v3i2.15734.

Anggraini, Tripani, Solihah Titin Sumantri, and Jamil. Khoirul. 2024. "Implementasi Pengklasifikasian Dan Penataan Bahan Pustaka Di Perpustakaan Sekolah Menengah Pertama It Al-Hijrah 2 Kecamatan Percut Sei Tuan Kabupaten Deli." *AL-IMAN: Jurnal Keislaman Dan Kemasyarakatan* 8(2):491–516.

Apriliyani, Meli, Mirza Izzal Musyaffaq, Siti Nur'Aini, Maya Rini Handayani, and Khotibul Umam. 2024. "Implementasi Analisis Sentimen Pada Ulasan Aplikasi Duolingo Di Google Playstore Menggunakan Algoritma Naïve Bayes." *AITI* 21(2):298–311. doi: 10.24246/aiti.v21i2.298-311.

Bashir, Abubakar Salisu, Abdulkadir Abubakar Bichi, and Alhassan Adamu. 2024. "Automatic Construction of Generic Hausa Language Stop Words List Using Term Frequency-Inverse Document Frequency." *Journal of Electrical Systems and Information Technology* 11(1):58. doi: 10.1186/s43067-024-00187-5.

Daoudi, Sara, Chakib Mustapha Anouar Zouaoui, Miloud Chikr El-Mezouar, and Nasreddine Taleb. 2021. "Parallelization of the K-Means++ Clustering Algorithm." *Ingenierie Des Systemes d'Information* 26(1):59–66. doi: 10.18280/isi.260106.

Dea Mustika, Rizky, and Ahmad Zakir. 2022. *Jurnal Media Informatika [JUMIN] Implementasi Algoritma K-Means Untuk Clustering Judul Skripsi Universitas Harapan Medan*.

Dea Mustika, Rizky, Ahmad Zakir, and Alkhowa Rizmi. 2022. "Implementasi Algoritma K-Means Untuk Clustering Judul Skripsi Universitas Harapan Medan." *Jurnal Media Informatika* 4(1):40–47. doi: 10.55338/jumin.v4i1.405.

Du, Guoyu, Xuehua Li, Lanjie Zhang, Libo Liu, and Chaohua Zhao. 2021. "Novel Automated K-Means++ Algorithm for Financial Data Sets." *Mathematical Problems in Engineering* 2021:1–12. doi: 10.1155/2021/5521119.

Firmansyah, Taufik, Poningsih, and Sundari Retno Andani. 2022. "Analisis Clustering Algoritma K-Means Sebagai Rekomendasi Penambahan Koleksi Buku Di Perpustakaan Madrasah Tsanawiyah Negeri 2 Simalungun." *ZAHRA: Bulletin Big Data, Data Science, and Artificial Intelligence* 1(1).

Fransiska, Andien. 2023. "Penataan Koleksi Bahan Pustaka Di Perpustakaan Politeknik Sriwijaya Sebagai Upaya Mempermudah Menemukan Buku Yang Diperlukan Oleh Pemustaka." *Jurnal Multidisipliner Bharasumba* 2(3).

Haryani, Dicku Nofriansyah, and Ita Mariami. 2021. "Implementasi Data Mining Untuk Pengelempokan Buku Di Perpustakaan Yayasan Nurul Islam Indonesia Baru Dengan Metode K-Means Clustering." *Jurnal CyberTech* 1(1):1–12.

Hasanah, Nisriina Nuur, and Agus Sidiq Purnomo. 2022. "Implementasi Data Mining Untuk Pengelompokan Buku Menggunakan Algoritma K-Means Clustering (Studi Kasus : Perpustakaan Politeknik LPP Yogyakarta)." *Jurnal Teknologi Dan Sistem Informasi Bisnis* 4(2):300–311. doi: 10.47233/jteksis.v4i2.499.

Kenger, Omer N., Zulal Diri Kenger, Eren Ozceylan, and Beata Mrugalska. 2023. "Clustering of Cities Based on Their Smart Performances: A Comparative Approach of Fuzzy C-Means, K-Means, and K-Medoids." *IEEE Access* 11:134446–59. doi: 10.1109/ACCESS.2023.3333753.

Lan, Fei. 2022. "Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method." *Advances in Multimedia* 2022:1–11. doi: 10.1155/2022/7923262.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Nasir, Januardi. 2021. "Penerapan Data Mining Clustering Dalam Mengelompokan Buku Dengan Metode K-Means." *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer* 11(2):690–703. doi: 10.24176/simet.v11i2.5482.

Nur Afifah, Inas Ajeng, and Heri Nurdiyanto. 2023. "Data Mining Clustering Dalam Pengelompokan Buku Perpustakaan Mengunakan Algoritma K-Means." *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)* 8(3):802–14. doi: 10.29100/jipi.v8i3.3891.

Pamput, Jessicha Putrianingsih, Salsa Dillah, Aindri Rizky Muthmainnah, and Dewi Fatmarani Surianto. 2024. "Analysis of Fuzzy C-Means In Personality Clustering Based On The Ocean Model." *JIKO (Jurnal Informatika Dan Komputer)* 7(3):158–64. doi: 10.33387/jiko.v7i3.8369.

Siburian, Daud, Sundari Retno Andani, Ika Purnama Sari, and Genesis Artikel. 2022. "Implementasi Algoritma K-Means Untuk Pengelompokkan Peminjaman Buku Pada Perpustakaan Sekolah Implementation of K-Means Algorithm for Clustering Books Borrowing in School Libraries." *JOMLAI: Journal of Machine Learning and Artificial Intelligence* 1(2):2828–9099. doi: 10.55123/jomlai.v1i2.725.

Widaningrum, Ida, Dyah Mustikasari, Rizal Arifin, Siti Lathifah Tsaqila, and Dwiyunia Fatmawati. 2022. "Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) Dan K-Means Clustering Untuk Menentukan Kategori Dokumen." *SISFOTEK: Sistem Informasi Dan Teknologi*.

Xu, Xiaobo, and Jin Shang. 2024. "Research on the Construction Scheme of Smart Library Based on Blockchain Technology." *Measurement: Sensors* 31. doi: 10.1016/j.measen.2023.100943.

Zhao, Huiling. 2022. "Design and Implementation of an Improved K-Means Clustering Algorithm." *Mobile Information Systems* 2022:1–10. doi: 10.1155/2022/6041484.