Enhanced Digital Library with Book Recommendations Based on Collaborative Filtering

Eki Nugraha[1,2*], Taufik Ardiansyah[2], Enjun Junaeti[2], Lala Septem Riza[2]**
[1]Program Studi Pendidikan Teknologi dan Kejuruan, Sekolah Pascasarjana, Universitas Pendidikan Indonesia.
[1,2]Departemen Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia,
[1]ekinugraha @upi.edu.
[2]taufardh@student.upi.edu; enjun@upi.edu; lala.s.riza@upi.edu.

Abstract: This research aims to develop a digital library system by implementing a recommendation system becomes a feature of the digital library system. The main problem with the digital library system is the large number of books that users will have difficulties finding interesting books. Therefore, Machine Learning is applied using the User-based Collaborative Filtering method to provide book recommendations for users. The book recommendation system being developed not only uses book value as input but also user behavior and book borrowing data. The resulting recommendations are divided into three parts according to the input which are book value, user behavior, and book borrowing data. The system developed is web-based using the PHP programming language, with modifications to the available open-source digital library system. Experiments are carried out by testing the system to university students and satisfaction questionnaire to the students. The total result of the average precision-recall calculation has an accuracy of 79% from 35 data users which includes an assessment of 1000 types of books with different categories. It means that books recommended by the system are relevant to users. The developed models and applications are potentially used as smart library applications.

Keywords: Collaborative Filtering, Digital Library, Pearson Correlation Coefficient, Recommendation System, User-based Collaborative Filtering

Corresponding Author
*Program Studi Pendidikan Teknologi dan Kejuruan, Universitas Pendidikan Indonesia,
*ekinugraha@upi.edu
**Departemen Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia
**lala.s.riza@upi.edu

1. Introduction

Technology is developing fast. People are generally interested in technology and times. Therefore, innovation is needed not only in entertainment or communication media but also in public facilities so that they are more attractive and can be used properly by the community. With adequate facilities, these public facilities will be more comfortable and easier for the public to access. One of the public facilities that we encounter a lot is a library and almost every educational institution or school has a library. Along with the development of information and communication technology, libraries have developed significantly and already have digital facilities called digital libraries or digital libraries (Firdaus, Wahyudin, and Nugroho, 2017; Widiaty, Riza, Abdullah, and Mubaroq, 2020). With the number of books circulating in the community and the development of science, the number of books is increasing and it becomes a challenge for the digital library to organize and process the book lending system.

Machine Learning plays a role in providing a system that is smart and able to provide better recommendations as data increases (Faruk and Cahyono, 2018). Learning considers and develops as tasks and process experiences are experienced. So, machine learning is a computer program that learns from experience by considering multiple tasks and results, as well as performance. If the performance on a task can improve along with the experience of the program or data being run (Mitchell, 1997). Machine Learning can be defined briefly, which is to enable a computer to successfully make predictions based on experience from previously processed data. The development of technology, data storage and the ability to process a computer causes machine learning to keep up with these developments to become increasingly sophisticated and smarter (Alpaydın, 2014). Machine Learning is basically a computer process to learn from existing data, without data, Machine Learning cannot learn. There are four types of Machine Learning, namely, supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Liu, 2018).

The recommendation system is mostly used by e-commerce websites, the system is used to attract visitors to buy goods offered by the sellers. Data to provide

recommendations is obtained from visitors' data and their interests through the behavior they perform when accessing the e-commerce website. With the development of a recommendation system to attract customers, it can also be implemented for other purposes including to attract public reading interest or library visitors with the recommendation system.

One of the e-commerce websites that has the advantage and success with its recommended features is amazon.com. These recommended features provide convenience to visitors because these features will give recommendations according to what the visitors are interested in, for example, amazon.com will provide recommendations for baby products for young mothers, or provide camera lens recommendations for photographers. All data is based on the input made by visitors, the more the visitors, the amazon.com recommendation system will be smarter (Linden, Smith, and York, 2003).

For this reason, this study focuses on developing intelligent library applications by applying Machine Learning methods, namely collaborative filtering, which has the ability to provide book recommendations that may be of interest to users. These recommendations are given based on historical data/training data for previous users. With this recommendation, users will have an alternative or choice of books to borrow. Collaborative filtering, recommendations on e-commerce have also been used by several recommendation systems. The library book recommendation system with the collaborative filtering method is able to produce good recommendations with a rating system and uses the Adjusted Cossine Similarity and Weight Sum algorithm calculations so it can predict the rating of books that have never been rated before (Mathew, Kuriakose, and Hegde, 2016). Enhanced digital library with book recommendations based collaborative filtering to make recommendations to guide the visitors but the method of user-based collaborative filtering. Amazon.com uses this recommendation system to attract customer interest and customer convenience in finding and buying an item from amazon.com. By providing recommendations according to what they want. The same method is used to attract book readers in libraries and e-libraries so visitors or readers cannot only read the book they are looking for but also get book recommendations provided by the system based on the calculation results of the item-to-item collaborative filtering method (Tian, Zheng, Wang, Zhang, and Wu, 2019).

2. Method

The stages in this research include developing a collaborative filtering model to provide a rating and log-based recommendations as well as implementing a collaborative filtering recommendation system into an open-source digital library. The model built applies the User Base Collaborative Filtering method based on rating and log base. With the rating data on the book that has been obtained from the user and the log base of user behavior in the digital library system, it is entered, which then results in a recommendation for the user himself and is different from other users according to the data entered.

2.1 Recommended Model Development

The model development process includes making input for value data, log view data, and book lending data which will later be implemented into a digital library system based on users and rated books, user behavior as a log view, and book loan data. The recommendation system model as shown in Figure 1 consists of a dataset, collaborative filtering, Pearson correlation coefficient. In this study, input for the recommendation system was developed, not only from value data but also from log vie data, and book lending data.

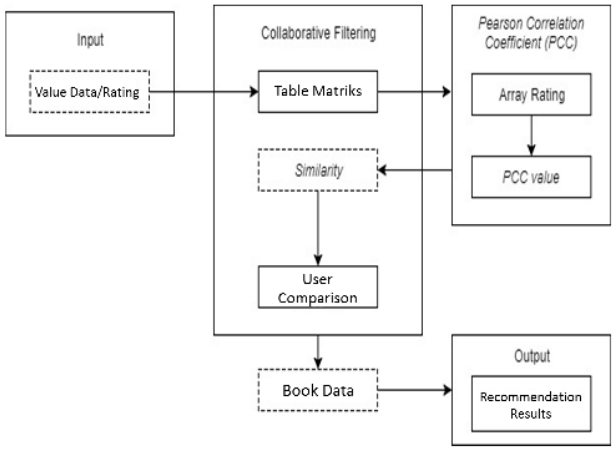

Fig.1 Model recommendation system

Figure 1 is digital library data, which is in the form of value data given by users to a book stored in a database in the form shown in Table 1. It can be seen that on each book represented by the column "Books Id", every user supplies a rating value illustrating how relevant the book is.

Table 1 Storage of value data

Users Id	Books Id	Rating/Value
1	2	4
2	2	4
2	4	9
….	...	...
n	n	n

Furthermore, the rating data is converted into a matrix table before calculating the similarity value. Example of data in the form of a matrix table shown in Table 2. So basically, we collect the values according to the User ID and Book Id.

Table 2 Value of matrix

User ID	Book Id				
	1	2	3	4	5
1	2	4	8	0	0
2	0	4	6	9	0
3	0	0	0	8	10
4	0	6	7	0	0

Table 2 is then compared, one user with another user and produces a similarity value. Similarity or the value of the relationship between two vectors where the similarity value has a range between -1 to 1, if it is zero then the two book value vectors have no relationship or equation at all, if it is negative then it has an opposite relationship, and if it is positive then two rank vector books have time, the more breakfast 1, the closer the time. In this study, calculating the

value similarity using one method, namely the Pearson Correlation Coefficient (PCC). One way to streamline the results of recommendations is to choose a method of calculating the value. Pearson Correlation Coefficient (PCC) is a popular method for calculating the similarity value for any collaborative filtering that is greater than two correlated users (Sheugh and Alizadeh, 2015).

$$pcc(u,u') = \frac{\sum_{i\in I}(r_{u,i} - \bar{r}_u)(r_{u',i} - \bar{r}_{u'})}{\sqrt{\sum_{i\in I}(r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_{i\in I}(r_{u',i} - \bar{r}_{u'})^2}} \quad (1)$$

Based on the Pearson correlation coefficient (PCC) formula as in (1), ru, i and ru ', i is the assessment score of two users and r̄u and r̄u' are the average ratings of two users, and I is a slice of the assessment of the two users, given by both u and u '. Then the two value vectors from each user are compared to produce similar value with each other user. Table 3 is the calculation result of the similarity value. Based on the calculation of the similarity, the relationship between active users, namely user 1 and user 2 has a positive value which means that it has similarities, as well as user 4 has a high enough similarity value, but in Table I it can be seen that user 4 does not have any other book besides the book that intersects with user 1.Therefore, the recommendation system will take another similarity value, namely from user 2 and provide a recommendation for book D because it does not coincide with the book rated by user 1. Whereas for user 3 it cannot provide recommendations because the similarity is negative.

Table 3 Result value of Pearson correlation

No	Active User	Comparison user	Similarity value
1	1	2	0.206935
2	1	3	-0.756024
3	1	4	0.910359

Apart from using traditional collaborative filtering, this research provides a little customization to the existing collaborative filtering methods, namely by log view or user behavior on a digital library system. By recording user behavior when operating the system, namely seeing the details of the book, the log view data is obtained in the form of the number of users displaying a book. The log view data becomes the rating data which is then calculated the similarity value.

The log-based data is unlimited, the log value will continue to increase every time the user displays the book details. In addition, log-based data is obtained automatically from user behavior, while ordinary rating data is obtained by means of users giving ratings to certain books, and not every user gives a rating to a book. An example of log view data can be seen in Table 4. It means that the user id 1 has read 8 times on the Book Id of 1.

Table 4 Data log view

Users Id	Books Id	Log
1	1	8
1	2	5
2	2	4
...	...	...
n	n	n

In addition to the additional log view data in this study, there is also a recommendation based on the data for borrowing books by users, which will also calculate the similarity value such as the value data which produces a similarity value. The book lending data is obtained from the simulation of the book borrowing process carried out by the user on the system. The following is an example of book data in Table 5.

Table 5 Book loan data

Users Id	Books Id	Loan Data
1	1	0
1	2	0
2	2	1
...	...	...
n	n	n

Book lending data consists of numbers 1 and 0, 1 for books that have been or are being borrowed, and 0 for books that have never been borrowed by a user. With the log view data and borrowed data, it is then converted as in the value data into a matrix table. Furthermore, the similarity value is calculated and provides book recommendations.

The process is the same as value data, but the data entered is in the form of a log view from the user, and data for borrowing books. The following is the pseudocode of the recommendation process:

Input:

- Data Rating
- Data Users
- Data Items

Output:

List of recommended items

Process:

Convert rating table to the matrix, user id - item id

If the id rating on the user is not equal to 0

a) Perform looping according to the number of user id without active user id
b) Calculate the similarity value of the active user id with another user id
c) Sort the similarity value based on the largest similarity value
d) Check the rating of each user id according to the similarity value
e) If there is a rating with an item id that is different from an active user id and is of good value, enter it into the recommended item id list
f) If there is no, it will continue to the user id with the next similarity value

Retrieves item data to be recommended based on the item id list displays a list of recommendations

Because the data input consists of three types of data, namely value data, log view data, and book borrowing data, each of which has a similarity value calculated. The result of the process is a list of book recommendations which is divided into three parts of recommendation features, namely:

- "*Recommended Books*", is a recommendation that is generated from the value data given by the user to a book.
- "*Most Viewed Books*", is a recommendation generated from the log view data that is obtained based on the behavior of displaying book details by the user on a book.
- "*Other People Also Borrow This Book*", while the last part is the recommendations generated from book lending data obtained from simulations of borrowing books by users in the digital library system.

Input is in the form of data consisting of value data, log view data, and book loan data. Then the similarity value is calculated and retrieves the book data based on the order of users with the greatest similarity and gives the results in the form of book recommendations.

2.2 Implementation of Recommendation Systems into the Digital Library System

At this stage, the recommendation system model is implemented into the digital library. In this research, the digital library system is taken from phpgurukul.com for free and is open source. The following is the process of implementing a recommendation system into a digital library.
(i) *Identification*. At this stage, identification of the digital library system is carried out in relation to the architecture, work methods, business flow, and the database used. The thing that must be considered is the table structure and database in the digital library because this digital library system does not yet have a recommendation feature, additional tables are needed to store input values and logs from users. Adding recommendation features, after identifying the digital library structure, the next step is to add a recommendation system on the user display page so that users can see recommended books. In addition, a feature is also added to provide value to each existing book and also a log recording system for each activity that displays book details.
(ii) *Testing*. After the recommendation system is implemented, the recommendation system will be tested with input in the form of assessment input on the book and user behavior on the system whether it can operate properly and in accordance with the pseudocode and calculations on the recommendation system model.

3. Results and Discussions

Collaborative filtering is a method of determining or providing recommendations to users. The most important part of collaborative filtering is a rating or assessment of an item made by the user in the system. Implementations carried out in digital libraries produce recommendations based on collaborative filtering for books. The rating and log data are divided into two, namely training data and test data. Training data is rating data and logs that already exist on the system and are obtained from users who have provided input or logs. While the test data is in the form of users who are currently active, both new users who logged in for the first time and have not provided input data, either ratings or logs or old users who access the library system again. Data collection techniques for training and testing data can also be seen in references (Riza, Handian, Megasari, Abdullah, Nandiyanto, and Nazir, 2018; Riza, Pradini, and Rahman, 2017; Riza, Zainafif, and Rasim, 2018; Riza, Utama, Putra, Simatupang, and Nugroho, 2018; Riza, Asyari, Prabawa, Kusnendar, and Rahman, 2018).

The research was conducted by testing the program towards several students of the Department of Computer Science, Universitas Pendidikan Indonesia. Students are asked to register for the digital library application, then operate the application including doing activities to display book details, assessing books, and simulating books. So the data is obtained in the form of user data during user registration and assessment data consisting of value data, log view data, and book lending data. Every user who signs in to the system becomes data-testing and data-training, namely other users who have provided data into the application system.

The book data used for the research are obtained manually from the gramedia.com site in the form of book titles, book covers, book authors, book publishers, year of publication, and brief book details. The book data consists of 1000 books consisting of several categories, namely, 98 agricultural books, 22 art and design books, 93 comic books, 295 computer and technology books, 30 engineering books, 20 financial books, 46 history books, 41 humanitarian books, 13 lifestyle books, 20 literacy books, 59 health books, 158 novels, 20 special skills books, 44 religious' books, 11 spiritual books, and 30 travel books. Every behavior performed by the user is recorded by the system so that it can provide recommendations according to that behavior. In addition, logs of behavior in library applications are also stored along with the results of similarity calculations for all users in the digital library application system. Similarity calculation by measuring the match distance between two data can be studied in literature (Riza, Awaludin, Sutarno, Munir and Wibawa, 2017; Riza, Pertiwi, Rahman, Munir, and Abdullah, 2019).

Experiments conducted with students of the Department of Computer Science Education. Table 6 is the result of one of the experiments conducted on several users.

Table 6 Experiment result

Recommendation	Recommended amount	The amount that the user is interested in	Number of books according to user category	Incorrect book
"Recommended Books"	10	8	9	2
"Most Viewed Books"	10	9	9	1
"Other People Also Borrow This Books"	10	5	9	5

Testing is done by calculating precision recall (Riza, Zainafif, and Rasim, 2018), in which it is shown in the experimental results in Table 6. So, as to produce a precision recall diagram as shown in Figure 2.

Based on Figure 2, in the experiment the log view recommendation section has a high value of 1.00 in the recall

section and 0.90 in the precision section. While the accuracy obtained is 73%. The results of all research experiments can be seen in Table 7, which consists of 8 experiments based on books that are of interest to users and based on categories that are of interest to users.

The system created can provide recommendations according to users because the system gets data to be recommended to these users based on the behavior of the users themselves (Farasyi, Setiawan, Fahsi, and Riza, 2018). The more data used, the better for the system in providing recommendations because the data obtained varies greatly depending on the user's behavior and preferences for the book.

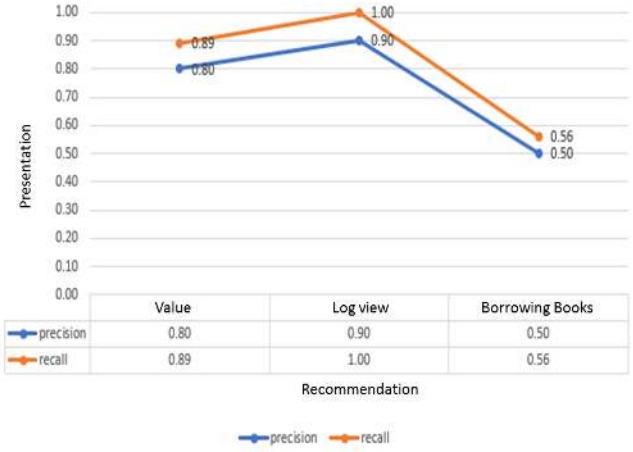

Fig. 2 Diagram of the Recall Precision Results

Table 7 Results all experiment

Experiments	Based on books that are of interest to users	Based on the categories that the user is interested in	Average
1	0.63	0.83	0.73
2	0.67	0.80	0.74
3	0.73	0.90	0.82
4	0.70	0.80	0.75
5	0.67	0.77	0.72
6	0.73	0.87	0.80
7	0.97	1.00	0.98
8	0.77	0.87	0.82
Average			0.79

Based on the experimental results in Table 7, it can be concluded that the accuracy of the recommendation results with the user's interests based on precision recall is 79%. This result is quite good because of the limited data book, user data, and assessment data in the system. In addition, limited experimentation can affect these results. Therefore, by this research, we have improved the existing digital library by adding the recommendation feature so that users can easily find other relevant books automatically.

As the next research, several other methods can also be used, such as fuzzy rule-based system (Riza, Bergmeir, Herrera Triguero, and Benítez Sánchez, 2015), rough set (Riza, Janusz, Bergmeir, Cornelis, Herrera, Ślezak, and Benítez, 2014), and gradient descent (Riza, Nasrulloh, Junaeti, Zain, and Nandiyanto, 2016). In addition, an approach using parallel computing can also be applied to speed up the computation process (Riza, Utama, Putra, Simatupang, and Nugroho, 2018; Riza, Asyari, Prabawa, Kusnendar, and Rahman, 2018; Riza, Rachmat, Munir, and Nazir, 2019; Riza, Anwar, Rahman, Abdullah, and Nazir, 2020; Pratama and Atmi, 2020).

4. Conclusion

From this research it can be seen that the digital library system can implement features in the form of book recommendations that make it easier for users to find books they like, besides that the recommendation features in digital library applications can increase the curiosity of users because displaying recommendations, users will be curious to explore the books in libraries and digital libraries. This study obtained an average precision-recall result based on the user's interest category of the recommended book and the user interest category, namely 0.79 or an accuracy value of 79%. These results are obtained from 35 data users that provide an assessment of 1000 types of books with different categories. Most of the recommendations given by the system are in demand by users.

Acknowledgements

We would like to thank Program Studi Pendidikan Teknologi dan Kejuruan Sekolah Pasca Sarjana Universitas Pendidikan Indonesia. We also thanked to Lecturer of Departemen Pendidikan Ilmu Komputer for assisting this experiment.

References

Alpaydın, E. (2014). Introduction to Machine Learning. Methods in Molecular Biology, 1107, 105–128

Farasyi, G., Setiawan, W., Fahsi, M., and Riza, L. S. (2018). The Association Rule Method for Mapping and Recommendation System on Students' Difficulties. Transylvanian Review, 1(1).

Faruk, A., and Cahyono, E. S. (2018). Prediction and classification of low birth weight data using machine learning techniques. Indonesian Journal of Science and Technology, 3(1), 18-28.

Firdaus, C., Wahyudin, W., and Nugroho, E. P. (2017). Monitoring system with two central facilities protocol. Indonesian Journal of Science and Technology, 2(1), 8-25.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computer, 7(1), 76–80.

Liu, Y. (2018). Data mining of university library management based on improved collaborative filtering association rules algorithm. Wireless Personal Communications, 102(4), 3781-3790.

Li, L., Zhou, Y., Xiong, H., Hu, C., and Wei, X. (2017, March). Collaborative filtering based on user attributes and user ratings for restaurant recommendation. In 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (pp. 2592-2597). IEEE.

Riza, L. S., Handian, D., Megasari, R., Abdullah, A. G., Nandiyanto, A. B. D., and Nazir, S. (2018). Development of R package and experimental analysis on prediction of the CO2 compressibility factor using gradient descent. Journal of Engineering Science and Technology, 13(8), 2342-2351.

Riza, L. S., Pradini, M., and Rahman, E.F. (2017). An expert system for diagnosis of sleep disorder using fuzzy rule-based classification systems. IOP Materials Science and Engineering Conference Series, 185(1), 012011.

Riza, L. S., Firmansyah, M. I., Siregar, H., Budiana, D., and Rosales-Pérez, A. (2018). Determining Strategies on Playing Badminton using the Knuth-Morris-Pratt Algorithm. Telkomnika, 16(6), 2763-2770.

Mathew, P., Kuriakose, B., and Hegde, V. (2016). Book Recommendation System through content based and collaborative filtering method. In 2016 International conference on data mining and advanced computing (SAPIENCE). 47-52.

Pratama, E. E., and Atmi, R. L. A. (2020). Text mining implementation based on twitter data to analyse information regarding corona virus in Indonesia. Journal of Computers for Society, 1(1), 91-100.

Riza, L. S., Awaludin, R., Sutarno, H., Munir, and Wibawa, A. P. (2017). A model for auto generating sets of examination items in educational assessment by using fuzzy c-means. World Transactions on Engineering and Technology Education, 15, 114-119.

Riza, L. S., Pertiwi, A. D., Rahman, E. F., Munir, M., and Abdullah, C. U. (2019). Question Generator System of Sentence Completion in TOEFL Using NLP and K-Nearest Neighbor. Indonesian Journal of Science and Technology, 4(2), 294-311.

Riza, L. S., Zainafif, A., and Rasim, S. N. (2018). Fuzzy rule-based classification systems for the gender prediction from handwriting. Telkomnika, 16(6), 2725-2732.

Riza, L. S., Bergmeir, C. N., Herrera Triguero, F., and Benítez Sánchez, J. M. (2015). FRBS: Fuzzy rule-based systems for classification and regression in R. American Statistical Association, 65(6), 1-30.

Riza, L. S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślezak, D., and Benítez, J. M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets". Information Sciences, 287, 68-89.

Riza, L. S., Nasrulloh, I. F., Junaeti, E., Zain, R., and Nandiyanto, A. B. D. (2016). gradDescentR: An R package implementing gradient descent and its variants for regression tasks. In 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). 125-129.

Riza, L. S., Utama, J. A., Putra, S. M., Simatupang, F. M., and Nugroho, E. P. (2018). Parallel Exponential Smoothing Using the Bootstrap Method in R for Forecasting Asteroid's Orbital Elements. Pertanika Journal of Science and Technology, 26(1), 441-462.

Riza, L. S., Asyari, A. H., Prabawa, H. W., Kusnendar, J., and Rahman, E. F. (2018). Parallel particle swarm optimization for determining pressure on water distribution systems in R. Advanced Science Letters, 24(10), 7501-7506.

Riza, L. S., Rachmat, A. B., Munir, T. H., and Nazir, S. (2019). Genomic Repeat Detection Using the Knuth-Morris-Pratt Algorithm on R High-Performance-Computing Package. International Journal of Advance Soft Computer. Application, 11(1), 94-111.

Riza, L. S., Anwar, F. S., Rahman, E. F., Abdullah, C. U., and Nazir, S. (2020). Natural Language Processing and Levenshtein Distance for Generating Error Identification Typed Questions on TOEFL. Journal of Computers for Society, 1(1), 1-23.

Sheugh, L., and Alizadeh, S. H. (2015). A note on pearson correlation coefficient as a metric of similarity in recommender system. In 2015 AI & Robotics (IRANOPEN). 1-6.

T. M. Mitchell. (1997). Machine Learning. McGraw-Hill Science/Engineering/Math.

Tian, Y., Zheng, B., Wang, Y., Zhang, Y., and Wu, Q. (2019). College library personalized recommendation system based on hybrid recommendation algorithm. Procedia CIRP, 83, 490-494.

Widiaty, I., Riza, L. S., Abdullah, A. G., & Mubaroq, S. R. (2020). Multiplatform application technology–based heutagogy on learning batik: A curriculum development framework. Indonesian Journal of Science and Technology, 5(1), 45-61.