Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 3, June 2023, pp. 467-475

DOI: https://doi.org/10.52436/1.jutif.2023.4.3.702
p-ISSN: 2723-3863
e-ISSN: 2723-3871

SCIENTIFIC ARTICLES RECOMMENDATION SYSTEM BASED ON USER'S RELATEDNESS USING ITEM-BASED COLLABORATIVE FILTERING METHOD

Ferzha Putra Utama[*1], Triska Mardiansyah[2], Ruvita Faurina[3], Arie Vatresia[4]

[1]Information Systems, Faculty of Engineering, Universitas Bengkulu, Indonesia
[2,3,4]Informatics, Faculty of Engineering, Universitas Bengkulu, Indonesia
Email: [1]fputama@unib.ac.id, [2]triskamardiansyah@gmail.com, [3]ruvita.faurina@unib.ac.id, [4]arie.vatresia@unib.ac.id

(Article received: November 25, 2022; Revision: December 30, 2022; published: June 26, 2023)

Abstract

Scientific article recommendation still remains one of the challenging issues in education, including learning process. Difficulties in finding related articles from research history and research interest have been experienced by students in collage affecting the duration of study and research time. This paper proposed a new solution by building a search engine to collect and to recommend articles related to student research topics. The system combined the web scraping method as an article data retrieval technique on google scholar and item-based collaborative filtering to recommend the article. Parameters result produced based on items of user's history, including item-searched, clicked, and downloaded. The system was built on a web-based scientific article recommendation system using python programming language. This system recommends articles based on the preferences of users and other users who are affiliated and who have an interest in the same item. This research showed that the validation result from the system obtained a recommendation accuracy value over 0.516801. The percentage of the RMSE error value of the recommendation system is 8.62%, or in other words that the accuracy of the recommendation system is 91.28%.

Keywords: *articles, item-based collaborative, recommendation system, scientific, web scraping.*

1. INTRODUCTION

Collage students face difficulties in finding references to support their research. Finding highly relevant articles or references requires experience and time to find relevant articles [1], [2]. Google Scholar is a useful and frequently used website to find references [3]. The major drawback of Google Scholar are shortcomings in the service, among which are articles that are newly added will never be recommended because only the articles that are referenced the most or have a high rating are displayed on the start page. One solution to this problem is using a recommendation system [4], [5].

Some previous research [6] utilizes the web scraping method to get data on the recapitulation of scientific articles on the Google Scholar site based on researcher's name at an institution. Suganeshwari [7] contains method analysis on a recommendation system that compares three methods, content-based, collaborative filtering and hybrid methods. This research proposed a combination of item-based collaborative filtering and user-based collaborative filtering methods. The results of combining the two methods have proven to generate high-quality recommendation.

In [8], the authors conclude that the item-based method can produce better recommendations than user-based because, in the prediction calculation from the user-based methods, many predictive values were found outside the range, besides that, the more items, the higher the MAE and NMAE values are produced and the recommendation process will take longer. Meanwhile, the item-based method does not find predictive values outside the range and has a better accuracy rate than the user-based method. Based on previous research in [7]–[9] proved that item-based collaborative filtering has a better performance than user-based collaborative filtering. In this research, we proposed item-based collaborative filtering and web scraping algorithms to search for scientific articles on Google Scholar. Determining spesific techniques will play a role in the effectiveness of a recommendation system in generating candidate items. The recommendation system analyzes item preferences and activities that occur in the system to predict items that will be recommended to users [10]–[12]. The main concept is to get recommendations based on the common interests of users. The proposed recommendation system analyzes user preferences based on their activity records and recommendations of other users with similar interests.

Web scraping (also called Screen Scraping, Web Data Extraction, or Web Harvesting) is a method used to extract some data from a website by extracting data and saving it to a database [13]. The web scraping process can be divided into four sequential steps, namely creating scraping templates, exploring web navigation, extracting information and storing the data obtained [14]. Web scraping does not provide a complete database, this technique obtains the requested information [15]. However, there are limits in the legality of web scraping that scrappers must comply with, namely 1) the data is not used for computer fraud and abuse, 2) the data taken is not copyrighted, 3) if the data is copyrighted, the content taken must comply with the standards by not modifying and including the copyright owner, 4) scraping measures do not burden and hinder the performance of the website services taken, 5) scraping must not violate the terms of use of the retrieved site and 6) scraping must not collect user information that is sensitive [13].

Collaborative filtering provides recommendations based on the collection of opinions and interests of several users through the rating given by the user to an item, then becomes a source of system knowledge to get the item of interest [16]. In addition to ratings, there are several sources of knowledge to obtain recommendations, namely implicit feedback, social tags, and context [4], [11], [17]. Collaborative filtering is basically divided into two methods: user-based collaborative filtering, also known as memory-based, and item-based collaborative filtering, also known as model-based [12]. In the user-based method, the system provides recommendations to users, items that are liked or rated by other users. For example, user *A* rates items 1, 2, and 3, then user *B* rates items 1, 2, and 4, then the system will recommend item 3 to user *B* and item 4 to user *A*. In comparison, the item-based method provides recommendations based on the similarity between items. The item-based method is a recommendation method with the similarity between rating an item and items that have been rated by users [18].

2. METHODS

The workflow of implementing web scraping and item-based collaborative filtering on the article recommendation system is divided into three stages, as shown in Figure 1. In Figure 1.a, the user inputs the search keywords, then the article data mining is carried out on Google Scholar. The data cited in the article is then converted to a rating scale based on the normal curve. The data is then stored in the database. Figure 1.b shows the process of users rating on articles that are taken implicitly. This process counts user actions when clicking or downloading articles and then converts those actions as ratings stored in the database. Figure 1.c explains the process of determining recommendations based on rating item-based collaborative filtering. This stage begins by determining the item-neighborhood using a technique by looking for a group of users (neighbors) who have a history of choices related to the target user. The rating data collected is then formed a user-item rating to determine articles that have never been rated and then compared one by one with articles that have been rated to produce a similarity value. The value is then calculated and then sorted so that the rating prediction value is obtained as a recommendation.

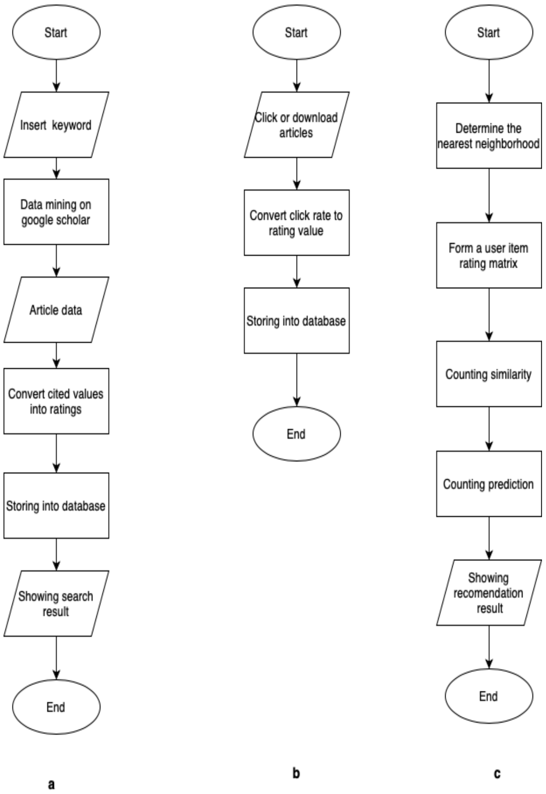

Figure 1. (a) Data mining and searching workflow (b) Rating workflow (c) Item-based collaborative filtering data workflow.

2.1. Data Collecting

Collecting data sets from Google Scholar uses three steps: mapping Google Scholar, extracting, and saving scraping information to the database [14]. These steps are shown in Figure 2.

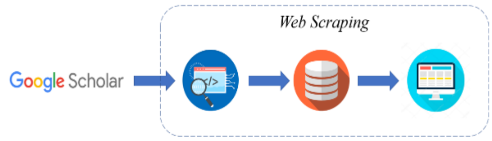

Figure 2. Web scraping steps on Google Scholar

The web scraping process starts with an HTTP request on the Google Scholar page. This request can be formatted into a URL containing a GET or HTTP request. After the request has been received and processed into an article search by Google Scholar, the Google Scholar search results will display the requested article data and then send it back to the system. Article data is obtained in HTML or JSON format. BeautifulSoup is a needed library to parse and extract information from HTML code. There are various data provided by Google Scholar, from the available data, attributes or data containing article information can be seen in Table 1.

Table 1. Google Scholar Data Object

No.	ID	Description
1	gs_r gs_or gs_scl	ID Artikel
2	gs_or_ggsm gs_press	Article download link
3	gs_rt	Title and source link
4	gs_a	Author, year, and publisher
5	gs_rs	Abstract
6	gs_or_cit	C*ited* and *Link*
7	gs_nph	Related article link

2.2. Recommendation Method

Item-based collaborative filtering is a recommendation method by calculating the similarity of items that have been rating with other items, then selected group of items that have similarity values with items that have been rated. The rated items will be used as a benchmark to find a number of other items that are correlated with items that were rated by other users [7], [19]. There are two steps that have to be taken to create a recommendation system using the item-based collaborative filtering method as follows [20]. This step is to calculate the similarity between one item with another item. The method that is often used to calculate the similarity of items is cosine similarity, but the weakness is that the difference in rating scales between various users will result in very different simarity [7], [19]. Adjusted cosine similarity overcomes the weakness of cosine similarity.

$$sim(i,j) = \frac{\Sigma u \epsilon U (R_{u,i} - \bar{R}u)(R_{u,j} - \bar{R}u)}{\sqrt{\Sigma u \epsilon U (R_{u,i} - \bar{R}u)^2}\sqrt{\Sigma u \epsilon U (R_{u,j} - \bar{R}u)^2}} \quad (1)$$

Descriptions:

sim(i,j) = The similarity value between article *i* and article j,
ΣuϵU = The set of *u* users who rate articles *i* and article *j*,
$R_{u,i}$ = *User rating u* on artikel *i*,
$R_{u,j}$ = *User rating u* on artikel *j*,
$\bar{R}_{u,i}$ = Average user rating for *u*

From the similarity calculation (1), the items will be sorted based on the similarity value. Items that have a high similarity will be at the top and vice versa. The result of the adjusted cosine similarity equation is in the range from -1 to 1. If the similarity value between the two items is close to +1, then the two items are considered to be more correlated. Conversely, if the similarity value is close to -1, then the two items will be increasingly uncorrelated.

2.3. Calculating Rating Prediction

This step is done to calculate the predicted rating of these items by comparing the rating that the user has given to an item with the similarity between the item and other items. The method used is the weighted-sum method [7], [19].

$$P(u,j) = \frac{\Sigma i \epsilon I (R_{u,i} * S_{i,j})}{\Sigma i \epsilon I |S_{i,j}|} \quad (2)$$

Descriptions:

P(u,j) = Rating prediction on article *j* by user *u*,
iϵI = Collection of articles similar to *j* article,
$R_{u,i}$ = *Rating user u* for *i* article,
$S_{i,j}$ = The similarity value between article *i* and article *j*

2.4. Evaluation

System testing is done through an evaluation matrix and error calculation function test. The purpose of the evaluation matrix is to measure the quality of proximity to the truth or the actual value achieved by a system. Predicting the rating that the user will give to an item is the main optimization carried out in item-based collaborative filtering. User rating predictions are calculated to find errors or deviations between the predicted rating and the actual rating. There are three matrices in this evaluation phase: Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE), the equation is as follows:

$$MAE = \frac{1}{|Q|}\sum_{(u,i)\in Q}|r_{ui} - \hat{r}_{ui}| \quad (3)$$

Mean Square Error:

$$MSE = \frac{1}{|Q|}\sum_{(u,i)\in Q}(r_{ui} - \hat{r}_{ui})^2 \quad (4)$$

Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{|Q|}\sum_{(u,i)\in Q}(r_{ui} - \hat{r}_{ui})^2} \quad (5)$$

where Q is the test data, r_{ui} as the actual user rating, $\hat{r}_{ui}$ represents the prediction rating of the recommendation system. The MAE value represents the average absolute error between the predicted value and the actual value [21]. MAE is the simplest, but does not take into account the direction of the error (error positive or error negative). MSE has a greater penalty on large errors and squared errors have no intuitive meaning. Therefore, RMSE is more widely used in calculating the prediction accuracy of the recommendation system [22].

The testing process on the recommendation system uses two approaches, namely MAP@K (Mean Average Precision at top K) for evaluation of the recommendation results:

$$MAP = \sqrt{\frac{\sum_{q=1}^{Q} Avep(q)}{Q}} \quad (6)$$

$$Avep(q) = \frac{\sum_{k=1}^{n} p(k) * rel(k)}{\#relevant\ item} \quad (7)$$

where Q is the number of recommendations, k is the power, $rel(k)$ represents the relativity function assigned a rating of k, $p(k)$ represents the precision assigned a rating of k. MAR@K (Mean Average Recall at top K) for evaluating prediction results:

$$MAP = \sqrt{\frac{\sum_{q=1}^{Q} Aver(q)}{Q}} \tag{8}$$

$$Aver(q) = \frac{\sum_{k=1}^{n} rel(r)}{\#relevant\ item} \tag{9}$$

where Q represents the number of user test cases against the item, r denotes the rating assigned, and $rel(r)$ is a binary function on the relevance of the given rating.

Bottegal and Pillonetto [23] consider the items stored in the test data set from the cross-validation evaluation as the relevant item set. They state that these matrices may be useful, but should be used with caution. Based on this matrix the items in the test data set are only a sample of items that can be considered relevant. In addition, these recall estimates should be used comparatively on the same data set and not as an absolute measure [17].

In the recommendation system, the coverage-item refers to the proportion of recommended items to the total items. Coverage-item is defined as follows:

$$coverage = \frac{U_{u \in U} I(u)}{I} \tag{10}$$

where $I(u)$ is the number of items recommended for the user, I is the number of items. Coverage is an important evaluation metric of a recommendation system because it can describe the ability to explore various recommendation items.

3. HASIL DAN PEMBAHASAN

3.1. Calculating of Item-based Collaborative Filtering

Table 2 shows an example of a case in the calculation process in providing recommendations using the item-based collaborative filtering method consisting of 5 affiliated users, namely $U_i = \{U_1, U_2, U_3, U_4, U_5\}$ and items consisting of 5 articles $A_j = \{A_1, A_2, A_3, A_4, A_5\}$ form a user-item ratings matrix to collect the rating values assigned by the user to each article.

Table 2. User-item rating matrix

		Item					
		A_1	A_2	A_3	A_4	A_5	$\bar{R}_u$
User	U_1	4	2		4	3	3,25
	U_2	2	3	2		4	2,75
	U_3				3	3	3
	U_4	2			1	2	1,66667
	U_5	1	2	5	2	4	2,8

User U_6 accessed the system for the first time and has not done ratings and searches for scientific articles. The system will display recommendations based on available articles or have been rated by users who are affiliated in the field of science with U_6 user, it is assumed that U_6 are interested in A_2 article from user activity by searching, clicking and downloading, getting three ratings on the article. This schematic is shown in Table 3.

Table 3. User-item rating matrix after adding U_6

		Item					
		A_1	A_2	A_3	A_4	A_5	$\bar{R}_u$
User	U_1	4	2		4	3	3,25
	U_2	2	3	2		4	2,75
	U_3				3	3	3
	U_4	2			1	2	1,66667
	U_5	1	2	5	2	4	2,8
New user	U_6		3				3

To determine the recommended articles on U_6, the following steps are:

1) Finding Adjusted Cosine Similarity

The step in the adjusted cosine similarity algorithm is to find the similarity value between the articles being compared. Based on Table 2, it can be explained as follows:

a. Checking between the rating values owned by articles A_1 and A_2
b. When U_1 rates articles A_1 and A_2, it finds values of 4 and 2. If one of the articles does not have a rating value, the similarity will not be calculated. Look at the second column and row of articles A_4 and A_5 and get an empty rating value and 4.
c. When U_4 the article rates A_1 and A_2, it gets a rating value of 2 and an empty value. This is because one of the articles does not have a rating value, so the similarities cannot be calculated.
d. Column and row checking will continue until the last row (U_6 row).
e. After getting the rating value between articles, the next step is to calculate the similarity value from the rating value that has been obtained by using the equation (1).

2) Adjusted Cosine Similarity Calculation Results

Table 4 shows the article rating for each user who rates the two articles, before calculating the similarity with equation (1).

Table 4. Representation of adjusted cosine similarity user rating in article A_1 and A_2

User	$R_{u,A1}$	$R_{u,A2}$	$\bar{R}_u$
U_1	4	2	3,25
U_2	2	3	2,75
U_5	1	2	2,8

$$Sim(A_1, A_2) = \frac{(4-3{,}25)(2-3{,}25)+(2-2{,}75)(3-2{,}75)\ldots}{\sqrt{(4-3{,}25)^2+(2-2{,}75)^2+(1-2{,}8)^2\ldots}}$$

$$\frac{\ldots + (1-2{,}8)(2-2{,}8)}{\ldots\sqrt{(2-3{,}25)^2+(3-2{,}75)^2+(2-2{,}8)^2}}$$

$$Sim(A_1, A_2) = \frac{-0{,}9375 + -0{,}1875 + 1{,}44}{\sqrt{4{,}365}\sqrt{2{,}265}}$$

$$Sim(A_1, A_2) = \frac{0{,}315}{\sqrt{4{,}365}\sqrt{2{,}265}}$$

$$= 0{,}10019$$

After calculating the adjusted cosine similarity equation to find the similarity value between articles A_1 and A_2, the similarity results obtained with a value of 0,10019. The results of the similarity between articles are in Table 5.

The next stage is to sort the similarity values in descending order from large to small data. This ranking is to take as many as *n* data with the highest similarity value. After that, calculate the predictions for each article that has never been rated by users and will be recommended to users using the *weight sum* algorithm.

3) Weighted Sum

Weighted sum will be used to find the predictive value of articles that will be recommended to users. First it will find the value of the new user U_6. The calculation starts from the user column that has never been rated. There are 4 article columns that have not been rated by the user, namely the article columns A_1, A_3, A_4 and A_5. The calculation stages are:

a. To find the predictions for A_1, A_3, A_4 and A_5 ratings with a non-empty rating value (U_6), the non-empty rating value in the A_2 article is 3.

b. After that, the article rating value is calculated using the formula (Aj * similarity (Ai, Aj)/similarity (Ai, Aj)) or

$$P(u,j) = \frac{\sum i\epsilon I(R_{u,i} * S_{i,j})}{\sum i\epsilon I\left|S_{i,j}\right|}$$

c. After getting the user rating value on the article, $R_{u,i}$ and the similarity value between articles $S_{i,j}$, then calculated using equation (2).

$$P(u,j) \quad = \frac{\Sigma i\epsilon I(Ru,i \times Si,j)}{\Sigma i\epsilon I|Si,j|}$$

$$P(U_6, A_1) = \frac{(3 * 0{,}1002) + (0 * -0{,}7496) + (0 * 0{,}7012) \ldots}{|0{,}1002 - 0{,}7496 + -0{,}7496 - 0{,}8418|}$$

$$\frac{\ldots + (0 * -0{,}8418)}{}$$

$$= \frac{0{,}3006}{|-0{,}79|}$$

$$= 0{,}1256$$

The result of the prediction calculation for article A_1 is 0,1256. After all the rating prediction data is collected, then they are sorted based on the predicted value from large to small data. Prediction results can be seen in Table 6. There are two problems that can affect the accuracy of rating predictions in the Item-based collaborative filtering method, namely the cold start problem and data sparsity [11], [12].

Table 6. Result prediction

User	Article	Prediction Result (*u,j*)	Round
U_1	A_3	3	3
U_2	A_4	2	2
	A_1	1,7185	2
U_3	A_2	0,2456	0
	A_3	1,036	1
U_4	A_2	0,6833	1
	A_3	1,6168	2
	A_1	0,1256	0
U_6	A_5	-0,1708	0
	A_4	-0,2013	0
	A_3	-0,9455	-1

The sparsity problem is a scourge in the recommendation system, a problem that occurs when predicting ratings for pairs of items that are not rated by users. Evaluating the accuracy of predictions, the system generally treats a subset of unknown ratings for those with known ratings, calculating predictions and comparing them with the actual rating using the Root Mean Squared Error (RMSE). RMSE is a very helpful matrix in measuring inaccuracies in all ratings [24]. Calculation of the comparison of the predicted rating value with the actual RMSE rating with equation (5). Table 7 shows the quality of the closeness of the rating prediction to the actual rating value.

Table 7. The comparison of rating predicted value with actual rating

User	*Item*	r_{ui}	$\hat{r}_{ui}$	$r_{ui} - \hat{r}_{ui}$	$(r_{ui} - \hat{r}_{ui})^2$
	A_1	4	2,25	1,75	3,0625
	A_2	2	2,5	-0,5	0,25
U_1	A_3	3	2,33	0,67	0,4489
	A_4	4	2,5	1,5	2,25
	A_5	3	3,2	-0,2	0,04
Total		16	12,78	3,22	6,0514

$$RMSE = \sqrt{\frac{1}{|Q|}\sum_{(u,i)\in Q}(r_{ui} - \hat{r}_{ui})^2}$$

$$RMSE = \sqrt{\frac{6{,}0514}{6}} = \sqrt{1{,}0086} = 1{,}0041$$

According to the RMSE equation, the accuracy value of the U_1 recommendation is 1.0041, in percentage the RMSE error value of the recommendation system is 16.8%, in other words the accuracy of the recommendation system for U_1 users is 83.2%.

3.2. Shilling Attack on Recommendation

In a shilling attack, users randomly rate the recommendation system so that the profile contains a number of possible ratings related to other active users, can affect the results of recommendations from other users. There are several classifications of rating models as attacks in the recommended articles described in [24].

3.3. Evaluation Result of Articles Recommendation

Evaluation of the recommendation system is done per user, the aim is to see the accuracy of the recommendations for each user who accesses the system. The data evaluated for recommendations, namely user preference data on items, scrapped articles, the total number of users, and affiliated users are users who have similarities in the field of science and users who have similarities in rating articles. It can be seen in Figure 3 as an example of evaluating the results of recommendations to users with ID G1B017001.

Evaluasi Rekomendasi Artikel MAP@K DAN MAR@K

Data

Nama	Mohammad aziz	
NPM	G1B017001	
Jumlah Artikel	390	DETAIL
Jumlah Pengguna	367	DETAIL
Pengguna Terafiliasi (Aktif)	4	G1B017002, G1B017001, G1B017003, G1B017004,

Figure 3. User preference data

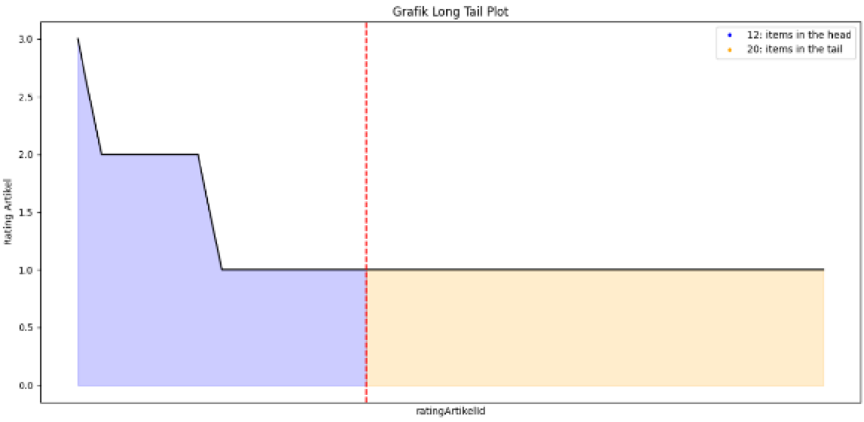

Figure 4. Plot long tail graph

The long tail plot in Figure 4 is used to explore the pattern of popularity in the interaction or rating of articles based on the items searched for, clicked on and downloaded by the user. Generally, only a small number of items have high interaction, and this is referred to as "head". Most of the items are in the "tail", which has little interaction with the user [25].

No	ID Artikel	PrediksiRating	Rating Rata-Rata	$r_{ui} - \hat{r}_{ui}$	$(r_{ui} - \hat{r}_{ui})^2$
1	XizCQ-hYadkJ	4.81	1.0	3.81	14.51
2	mSAsNrAiI58J	4.69	5.0	-0.31	0.1
3	jODCYwE68sgJ	3.93	4.0	-0.07	0.01
4	mu_gIK9sB4MJ	3.15	1.0	2.15	4.63
5	4xNeTz2JByUJ	2.56	2.0	0.56	0.31
6	EWTaUqHkLgEJ	2.44	3.0	-0.56	0.32
7	xoZmAGwOOdoJ	2.44	1.0	1.44	2.06
8	J8PpZap41WUJ	2.3	2.0	0.3	0.09
9	XezUUc-eluEJ	2.26	3.0	-0.74	0.55
10	J8hLARJZIVoJ	2.16	1.0	1.16	1.33
	Total	162.9238760248915	139.0	23.92	49.78243615205548

Figure 5. Rating prediction calculation results using item-based collaborative filtering

Figure 5 is the result of recommendations using the Item-Based Collaborative Filtering method. The information contained in the table are:

1) The article ID that represents the entire article information obtained from Google Scholar, which has been described in Table 1.
2) The average rating is the result of the average overall rating on the article.
3) The predictive value of item-based collaborative filtering with weighted sum calculations is described in Table 5.
4) $r_{ui} - \hat{r}_{ui}$ and $(r_{ui} - \hat{r}_{ui})^2$ are used to calculate the difference between the predicted value and the actual value

Recommendations to the user (G1B017001) obtained MAE of 0,08, MSE of 1,08 and RMSE of 1,04. The MAE, MSE, and RMSE error values in percent are 0.06%, 0.81%, and 0.78%, respectively. In other words, the accuracy of the recommendation system is 99.94%, 99.19% and 99.22%. In order to evaluate the total results of article recommendations, compared the results of recommendations using the item-based collaborative filtering method, popular articles. Tests with MAP@K to evaluate the results of the recommendations can be seen in Figure 6. Tests with MAR@K to evaluate the results of predictions can be seen in Figure 7.

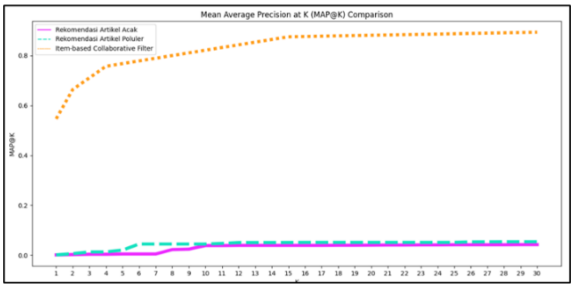

Figure 6. MAP@K performance of each recommended approach

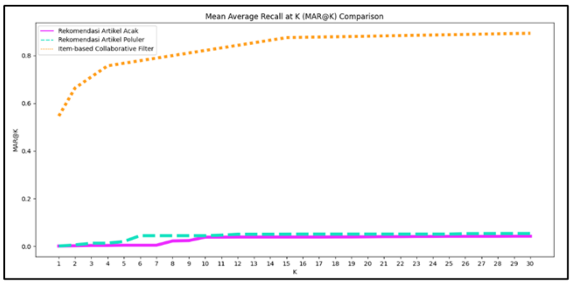

Figure 7. MAR@K performance of each recommended approach

Evaluation Comparing the predicted results of MAP@K and MAR@K, this study compared the thirty articles that had the highest predictive value. It is known from the results that the item-based collaborative filtering method consistently outperforms other baselines. The test results show that the increase in MAP@K and MAR@K item-based collaborative filtering is more statistically significant compared to other baselines. This is shown in Figure 8.

Item-Based Collaborative Filtering	Artikel	PT9m8niWbqMAxNeTz2JByUJ v41Z3oAze2EJ Ts16g7GETBEJLgUFND990akJGuLFVL3TrV4J ZwZGhbZsgqUJcetiKewYxBwJ dusm6OSuIgoJ e3S8XDEB_1IJ ef9tIeFm_fUJ ettftxPC8mIJ -0prdZjGyj8J fCNQcxShJPsJ cZy66KN1NsYJgO22LvmyqYcJhhtRPfWao9AJ hmBejdvrqJ4J i-ljhuH3u5aJ ft80ZjkNVIcJ bvyv_IfVY9cJ cLlv9dpRsHkJ jODCYwE68sgJbh3Ox3yxK88J bL9CAsXp9RgJbFB-hoNw7kAJanhecSriaQYJ aVk0cbn_QdgJ aPX7L8bc_DwJ_1hNyRgdZZoJ
	MAR@K	1) 0.344581 2) 0.489163 3) 0.533744 4) 0.578325 5) 0.622906 6) 0.667488 7) 0.712069 8) 0.728079 9) 0.744089 10) 0.760099 11) 0.776108 12) 0.792118 13) 0.808128 14) 0.824138 15) 0.825862 16) 0.827586 17) 0.82931 18) 0.831034 19) 0.832759 20) 0.834483 21) 0.836207 22) 0.837931 23) 0.839655 24) 0.841379 25) 0.843103 26) 0.844828 27) 0.846552 28) 0.848276 29) 0.85 30) 0.851724
Artikel Populer	Artikel	mSAsNrAiI58J jODCYwE68sgJXezUUc-eluEJ vfy2Szz4RKcJ XizCQ-hYadkJ 0yeUeoWOpL0J XaBn0ZMpYxw9WlZxcYK4vcJCzd76oKIDf4J e3aOW2j8f5cJ dusm6OSuIgoJ _1hNyRgdZZoJ blaa6JmSVA0J q-JNB0BXlAsJ zGOaysPdNB8Jf6e7WqvLXdcJ ft80ZjkNVIcJ BSwPFKPokDQJ LeCaD5WbctoJ d9g1iIrT08wJ kI3TFwiVjx0J PFRFqPKVPK0JTZFa7RnynDwJuA-m8OhI2G8J x272sA1OVmwJgO22LvmyqYcJUEhmB_XIA2sJMEZku-7EaoQJQNzfr0LoT4kJ F66f_nLDgQcJ
	MAR@K	1) 0.014286 2) 0.029433 3) 0.029433 4) 0.030296 5) 0.03133 6) 0.065813 7) 0.065813 8) 0.065813 9) 0.065813 10) 0.065813 11) 0.066597 12) 0.070168 13) 0.070168 14) 0.070168 15) 0.070858 16) 0.070858 17) 0.071568 18) 0.072334 19) 0.072334 20) 0.072334 21) 0.072334 22) 0.072334 23) 0.073008 24) 0.073008 25) 0.073008 26) 0.073672 27) 0.073672 28) 0.073672 29) 0.073672 30) 0.073672
Artikel Acak	Artikel	8XUsOsePZR0JovbP8QRIIPAJ AeWnB6m-1YIJ svSxLfUwTu0J 7mtCbI-k6IQJ 9q2yxH4JdokJ xsUSHhwkVXIJQTC1uYkQtYcJB3rIZNrZ_kwJ QolErl6BlQIJ IcgFjhAnghIJ v71Oax0Nq74J W0NkZATZFsUJQQv9k6C3AgJ vc4rC0Zb6U8J ettftxPC8mIJ ub2aYQga5JEJ vfy2Szz4RKcJ tD0CAeo2hoYJ TZFa7RnynDwJmBuKdirUoHQJ AEy58aqQkSUJUno-6lbqWxEJ 0rw6T-jSXcwJ S5RSLkNeU34JQolErl6BlQIJ a-m3IzPL0FIJ PUL2tQU1-qsJ wyc6nV9My8MJsxeTP-H6WoJ
	MAR@K	1) 0.0 2) 0.0 3) 0.0 4) 0.0 5) 0.0 6) 0.0 7) 0.000246 8) 0.000246 9) 0.000629 10) 0.020629 11) 0.020629 12) 0.020629 13) 0.020629 14) 0.020629 15) 0.020629 16) 0.020629 17) 0.020629 18) 0.020629 19) 0.021654 20) 0.021654 21) 0.021982 22) 0.021982 23) 0.021982 24) 0.021982 25) 0.02347 26) 0.023868 27) 0.024315 28) 0.024315 29) 0.02479 30) 0.02479

Figure 8. MAR@K calculation results.

4. DISCUSSION

Previous research has recommended articles and journals. Research by [26] recommend the most suitable journals based on references. While the research [27] recommend the article as a legal reference at trial. We observed the results of searching articles using different tools with the exact keywords showing that it is possible to get different search results. This result is due to Google's Search Engine Optimization (SEO), which displays an index of articles according to user activity patterns [28]. The keywords used as input are essential in determining search results. Ideally, keywords consist of two or more keywords [29]. The use of a single keyword will result in a decrease in the index ranking of the article. However, future research can develop a more independent recommendation system by scraping article information from scientific article provider sites.

5. CONCLUSION

Based on the results of research and discussion on the application of web scraping and item-based collaborative filtering on a web-based scientific article recommendation system based on the relevance of thesis topics, this research has succeeded in implementing web scraping and item-based collaborative filtering methods and obtaining test results using RMSE with an accuracy value. The recommendation result is 0.516801. The percentage of the RMSE error value of the recommendation system is 8.62% or in other words that the accuracy of the recommendation system is 91.28%.

ACKNOWLEDGEMENT

Special thanks to the Bengkulu University Research and Community Service, which has funded this development research scheme with contract number 850/UN30.15/PG/2021.

REFERENCES

[1] Ö. F. SÖNMEZ, "Bibliometric analysis of educational research articles published in the field of social study education based on Web of Science Database," *Particip. Educ. Res.*, vol. 7, no. 2, pp. 216–229, 2020, doi: 10.17275/per.20.30.7.2.

[2] V. Chaitanya and P. K. Singh, "Research articles suggestion using topic modelling," in *2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 2017, pp. 178–182.

[3] S. Malik, A. Rana, and M. Bansal, "A survey of recommendation systems," *Inf. Resour. Manag. J.*, vol. 33, no. 4, pp. 53–73, 2020, doi: 10.4018/IRMJ.2020100104.

[4] N. Nassar, A. Jafar, and Y. Rahhal, "A novel deep multi-criteria collaborative filtering model for recommendation system," *Knowledge-Based Syst.*, vol. 187, p. 104811, 2020, doi: 10.1016/j.knosys.2019.06.019.

[5] S. Gupta and V. Kant, "An aggregation approach to multi-criteria recommender system using genetic programming," *Evol. Syst.*, vol. 11, no. 1, pp. 29–44, 2020, doi: 10.1007/s12530-019-09296-3.

[6] A. Rahmatulloh and R. Gunawan, "Web Scraping with HTML DOM Method for Data

Collection of Scientific Articles from Google Scholar," *Indones. J. Inf. Syst.*, vol. 2, no. 2, p. 16, 2020, doi: 10.24002/ijis.v2i2.3029.

[7] G. S. and S. P. S. Ibrahim, "A Survey on Collaborative Filtering Based Recommendation System," *Smart Innov. Syst. Technol.*, vol. 49, pp. v–vii, 2016, doi: 10.1007/978-3-319-30348-2.

[8] Y. E. Christanti, "Perbandingan metode user-item based dan item-based collaborative filtering pada studi kasus sistem rekomendasi tempat wisata untuk wilayah Solo Dan Yogyakarta," UNS, 2013.

[9] A. N. Nikolakopoulos and G. Karypis, "Boosting Item-based collaborative filtering via nearly uncoupled random walks," *ACM Trans. Knowl. Discov. from Data*, vol. 14, no. 6, pp. 1–26, 2020.

[10] M. Fadelillah, "Sistem Rekomendasi Pencarian Artikel Jurnal Indonesia Menggunakan Metode Jaccard's Coefficient," Unissula, 2017.

[11] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, "Scientific paper recommendation: A survey," *IEEE Access*, vol. 7, pp. 9324–9339, 2019, doi: 10.1109/ACCESS.2018.2890388.

[12] I. S. Wahyudi, "Mesin Rekomendasi Film Menggunakan Metode Kemiripan Genre Berbasis Collaborative Filtering," Institut Teknologi Sepuluh Nopember Surabaya, 2017.

[13] V. A. Flores, P. A. Permatasari, and L. Jasa, "Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword," *Maj. Ilm. Teknol. Elektro*, vol. 19, no. 2, p. 157, 2020, doi: 10.24843/mite.2020.v19i02.p06.

[14] F. Maria Rosario B, Yovi Pratama, "Penerapan Web Scraping Pada Website Company Profile," *Kntia*, vol. 4, no. 4, pp. 37–43, 2017.

[15] C. Adamiak, "Current state and development of Airbnb accommodation offer in 167 countries," *Curr. Issues Tour.*, vol. 25, no. 19, pp. 3131–3149, 2022, doi: 10.1080/13683500.2019.1696758.

[16] F. A. Prayoga, F. I. Komputer, and U. A. Yogyakarta, "Smartphone Recommendation System Using Model-Based Sistem Rekomendasi Smartphone Menggunakan Metode Model-Based Collaborative Filltering," *J. Tek. Inform.*, vol. 3, no. 6, pp. 1613–1622, 2022, doi: 10.20884/1.jutif.2022.3.6.413.

[17] D. Parra and S. Sahebi, "Recommender systems: Sources of knowledge and evaluation metrics," *Stud. Comput. Intell.*, vol. 452, pp. 149–175, 2013, doi: 10.1007/978-3-642-33326-2_7.

[18] A. E. Wijaya and D. Alfian, "Sistem Rekomendasi Laptop Menggunakan Collaborative Filtering Dan Content-Based Filtering," *J. Comput. Bisnis*, vol. 12, no. 1, pp. 11–27, 2018.

[19] Y. V. L. Jaja, B. Susanto, and L. R. Sasongko, "Penerapan Meode Item-Based Collaborative Filtering Umtuk Sistem Rekomendasi Data MovieLens," *J. Mat. dan Apl.*, 2020.

[20] Y. Setiawan, A. Nurwanto, and A. Erlansari, "Implementasi Item Based Collaborative Filtering Dalam Pemberian Rekomendasi Agenda Wisata Berbasis Android," *Pseudocode*, vol. 6, no. 1, pp. 13–20, 2019, doi: 10.33369/pseudocode.6.1.13-20.

[21] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," in *IOP conference series: materials science and engineering*, 2018, vol. 324, no. 1, p. 12049.

[22] M. Chen and P. Liu, "Performance evaluation of recommender systems," *Int. J. Performability Eng.*, vol. 13, no. 8, pp. 1246–1256, 2017, doi: 10.23940/ijpe.17.08.p7.12461256.

[23] G. Bottegal and G. Pillonetto, "The generalized cross validation filter," *Automatica*, vol. 90, pp. 130–137, 2018, doi: 10.1016/j.automatica.2017.12.054.

[24] C. Panagiotakis, H. Papadakis, and P. Fragopoulou, "Unsupervised and supervised methods for the detection of hurriedly created profiles in recommender systems," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 9, pp. 2165–2179, 2020, doi: 10.1007/s13042-020-01108-4.

[25] X. Hu, C. Zhang, M. Wu, and Y. Zeng, "Research on Long Tail Recommendation Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 261, no. 1, 2017, doi: 10.1088/1757-899X/261/1/012019.

[26] A. O. Ogunde *et al.*, "The design of a hybrid model-based journal recommendation system," *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 6, pp. 1153–1162, 2020, doi: 10.25046/aj0506139.

[27] M. Zheng, B. Liu, and L. Sun, "LawRec: Automatic Recommendation of Legal Provisions Based on Legal Text Analysis," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/6313161.

[28] M. Nagpal and J. A. Petersen, "Keyword selection strategies in search engine optimization: how relevant is relevance?," *J.*

Retail., 2020, doi: 10.1016/j.jretai.2020.12.002.

[29] A. Setiawan, Z. Harahap, D. Syamsuar, and Y. N. Kunang, "The optimization of website visibility and traffic by implementing search engine optimization (SEO) in Palembang Polytechnic of tourism," *CommIT (Communication Inf. Technol. J.*, vol. 14, no. 1, pp. 31–44, 2020, doi: 10.21512/commit.v14i1.5953..