SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10047
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The User Personalization with KNN for Recommender System

Arie Satia Dharma
Faculty of Informatics and Electrical Engineering
Institut Teknologi Del
Toba Samosir 22381, Indonesia

***Abstract*—** Following the increase in of the information available on the Web, it is important to diversity of its users and the complexity of Web applications. One web application that has a diversity of users is a news website. Customizing a website with the characteristics of each user is called personalization. The purpose of this study is to study the methods used in giving news recommendations using user personalization. Collaborative filtering method (CF) is one method that groups users based on the nature of the user. This CF method can be applied using the k-nearest neighbor (KNN) algorithm. The proximity between users in this algorithm is sought using the Pearson correlation technique and cosine correlation. The best technique by considering the smallest value of prediction error evaluation will be applied to giving recommendations. Evaluation of these errors was tested by applying the formula Root Mean Square Error. The best evaluation results obtained in this study are the k-nearest neighbor algorithm with cosine correlation similiarity.

Keywords— personalization, website personalization, collaborative filtering, k-nearest neighbor, time spent, user interest.

A. INTRODUCTION

Following the increase in the information available on the web today has led to diversity of its users and the complexity of web applications. Therefore researchers started developing adaptive web systems that can adjust their appearance and behavior to each individual user or usergroup [1]. Personalization is website adjustments with each of website user. Personalization can be done by building readers' information by the website for a certain period of time to be used in learning behavior and recognizing readers while on the website [2]. The results of user personalization obtained can then be used to provide recommendations for readers of the website [3]. One of the applications on the web is a news website that has now become an alternative to news printed on paper [4]. Recommendations on news websites can be done in the form of news advice that might be desired so that readers no longer need to search for news and this will also save readers time [5]. In this paper, researcher will be implemented the personalization of users with collaborative filtering methods on news websites. News websites are chosen because their content is varied and can represent different reader behavior. Meanwhile collaborative filtering is the most widely used technique as a method of filtering data contained in websites based on user behavior [6]. User behavior that analyzed on news websites is category of the news, time duration of reading the news, and the similarity of the behavior of reading news between users.

B. LITERATURE

A. Website Personalization

Personalization provides recommendations in the form of web pages to users based on the user's search history that has previously been carried out [3]. For website administrators, personalization on the web can help to determine the user's choice based on the profile and behavior developed by the user [6]. And for website users, recommendations based on personalization are expected to make it easier for users to find information or content that suits the needs and desires of users [6]. In other words, personalization is done to make it easier for users to find the information needed from the web page.

This can happen because personalization will provide recommendations or information in accordance with the results of the website's ability to know how to manipulate user information [7].

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10047
e-ISSN : 2541-2019
p-ISSN : 2541-044X

B. Collaborative Filtering

Collaborative filtering is a method for recommendation systems where web pages are recommended to certain users if they have web similarities accessed by other users [8]. Collaborative Filtering is divided into 2 categories:

- Memory based, it is necessary to (rating) the user's ranking data from the user to the item to calculate the similarity between the user or item.
- In the Model Based generally the data used is not complete and a learning is needed in finding a model of available data which will be used to find similarities of users or items to be predicted.

In this paper, we use memory-based of collaborative filtering to get proximity between items and users. User or item proximity can be calculated using a neighborhood-based algorithm where the similarities between the two items or users will be resulted from the average weight of all ratings.

C. K-Nearest Neighbor(KNN)

K-nearest neighbor is an algorithm to classify objects based on learning data that is the closest distance or has similar characteristics from an object to another object [7]. The proximity of user characteristics can be found by using Pearson correlation, cosine similarity and adjusted cosine similarity.

1. *Pearson Correlation*

$$simil(u,v) = \frac{\sum_{c \in I_{u,v}}(r_{u,c} - \bar{r}_u)(r_{v,c} - \bar{r}_v)}{\sqrt{\sum_{c \in I_{u,v}}(r_{u,c} - \bar{r}_u)^2}\sqrt{\sum_{c \in I_{u,v}}(r_{v,c} - \bar{r}_v)^2}} \quad (1)$$

2. *Cosine similarity*

$$simil(x,y) = cos(\vec{u},\vec{v}) = \frac{\vec{u}.\vec{v}}{||\vec{u}||||\vec{v}||} = \frac{\sum_{c \in I_{u,v}} r_{u,c} r_{v,c}}{\sqrt{\sum_{c \in I_u} r_{u,c}^2}\sqrt{\sum_{c \in I_v} r_{v,c}^2}} \quad (2)$$

3. *Adjusted cosine similarity*

$$simil(u,v) = \frac{\sum_{c \in Iu,v}(r_{u,c} - \bar{r}_u)(r_{v,c} - \bar{r}_v)}{\sqrt{\sum_{c \in I_{u,v}}(r_{u,c} - \bar{r}_u)^2}\sqrt{\sum_{c \in I_{u,v}}(r_{v,c} - \bar{r}_v)^2}} \quad (3)$$

$\vec{u}, \vec{v}$ is a vector rating representation of the user *u* and its each neighbors *v*. $\vec{u}.\vec{v}$ is the inner product representation of the user *u* rating vector and v rating vector from its neighbors. $|\vec{u}|.|\vec{v}|$ represent the results of vector multiplication $\vec{u}$ dan $\vec{v}$ masing-masing. $I_{u,v} = I_u \cap I_v$ and for $r_{u,c}, r_{v,c}$ represent the rating of the item *c* by the user *u* and each the neighbor *v* while $\bar{r_u}$, $\bar{r_v}$ represent the average score of the user *u* and the average of each the neighbor's scores *v*. It is obtained $r_{u,i}$and $\bar{r}_u$ as shown below.

$$r_{u,i} = \frac{url_{i,time}}{url_{i,size}} \times N \quad (4)$$

$$\bar{r_u} = \frac{1}{|I_u|}\sum_{i \in I_u} r_{u,i} \quad (5)$$

$url_{i,time}$ represent the duration of user access u to url_i, $url_{i,size}$] represents the size of each page visited by the user u, N represent the number of visitors url_i by user *u* and $|I_u|$ represent user frequency *u* access the page.

The page size is affected by the amount of content that is contained in a news page. This is calculated because the size or amount of content in each news varies so that the user is interested to the news based on a more significant reading time.

C. METHODOLOGY

A. User Interest

The object study in this research is user interest. User interest is obtained from observing the behavior patterns of users when reading news on the website. The observed data came from 39 different users. These data are obtained when tracking from the web server log in the form of:

- IP address
- Access Time
- Status Code
- HTTP request method
- Page size/Number of byte

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10047
e-ISSN : 2541-2019
p-ISSN : 2541-044X

- Referer
- User Agent

The obtained data is then processed and analyzed to retrieve time spent data user when user accesses the news page on the website. With Eq. 4 and Eq.5, the time spent data is then processed to get user interest which is the main variable used in this research work.

B. User Similiarity

Collaborative filtering provides recommendations by looking at the similarity of user behavior in accessing news. Each user will be determined to read the behavior in order to produce recommendations based on personalization. User reading behavior is identified using parameters such as news categories that are accessed and the user interest.

At this stage an analysis is carried out for each user interest obtained from the user interest matrix for each content and content category that has value in the user interest to find the similarity of each user's interest.

User interest data is processed using the k-nearest neighbor algorithms. In the k-nearest neighbor algorithm, a pivot table will be formed containing user data, opened news data and user interest data on each news.

From the pivot table, the proximity between users will be calculated using Equation 1, Equation 2 and Equation 3. From the results of the calculation, 5 other users will be chosen with the greatest similarity value for each user. The greater the value of similiarity, the more interesting the two users / similar. Thus we can assume that the news accessed by user-1 is likely to be accessed also by user-2 or vice versa. The news that is accessed by 5 other users will be predicted by the user interest and then given as a recommendation to the user.

After the prediction is obtained then an evaluation of these results is done by using an RMSE formula (*root mean square error*). This RMSE applies a function to find out the error rate of each processing algorithm. The smaller the RMSE value obtained, the better the model will be.

The RMSE formula is:

$$RMSE = \sqrt{\frac{\sum(x_t - f_t)^2}{n}} \quad (6)$$

n = jumlah data
X_t = Data aktual

$\hat{f}_t = Data\ prediksi$

This evaluation is used to calculate the error value by comparing predictive user interest values and actual user interest values from user access data

D. RESULT

Based on the implementation that has been done, the following results are obtained from tracking the web server logs that are carried out by the user:

TABLE 1. DATA FROM WEB LOG SERVER

	idSite	idVisit	visitIp	visitorId	actionDetails/0/type	actionDetails/0/url	actionDetails/0/pageTitle	acti
0	1	172	114.125.0.0	8235ce7c4996e0b4	action	https://detik.xyz/wordpress/	detik news – Sumber Berita Terpercaya Anda	2.0
1	1	171	43.243.0.0	b09474e50ba644ee	action	https://detik.xyz/wordpress/	detik news – Sumber Berita Terpercaya Anda	2.0
2	1	170	114.125.0.0	9930e377afd3cbff	action	https://detik.xyz/wordpress/	detik news – Sumber Berita Terpercaya Anda	2.0
3	1	169	114.125.0.0	e41c8e9aa805a524	action	https://detik.xyz/wordpress/kriminal/pukuli-p...	Pukuli Pria dan Curi Motor Driver Ojek Online,...	182.
4	1	168	36.78.0.0	12e03f58031198cc	action	https://detik.xyz/wordpress/	detik news – Sumber Berita Terpercaya Anda	2.0
5	1	167	182.1.0.0	1b41457ec183706f	search	NaN	NaN	NaN

As we state previously that ini our research user interest to certain pages or news are determined based on the page the user opens and the access period that the user spends while reading the news. So we get user interest as following:

TABLE II. USER INTEREST

	id_visitor	id_page_action	time_spent	item_id	user_id	Page_size(KB)	n_access	rating(interest)
0	1b41457ec183706f	92	7	0	2	39.02	23	0.179395
1	1b41457ec183706f	94	13	1	2	39.31	27	0.330705
2	1b41457ec183706f	101	31	3	2	40.45	21	0.766378
3	1b41457ec183706f	107	23	7	2	39.49	14	0.582426
4	1b41457ec183706f	157	5	28	2	41.78	3	0.119674
5	1b41457ec183706f	183	8	35	2	39.13	1	0.204447

After obtaining data on the user's interest above, then the data is processed into a table pivot where item_id becomes the x axis, user_id becomes the y axis. Then the data in the cell is filled with the value of each user's interest in the item / news. So that the picture of the pivot table obtained is as follows:

TABLE III. PIVOT TABLE OF USER INTEREST

item_id	0	1	2	3	4	5
user_id						
0	0.000000	0.483338	0.000000	0.000000	0.655407	0.000000
1	0.000000	0.457899	0.000000	0.123609	0.000000	0.248077
2	0.179395	0.330705	0.000000	0.766378	0.000000	0.000000
3	0.179395	0.152633	0.000000	0.000000	0.126040	0.000000
4	0.512558	0.152633	0.126422	0.123609	0.201664	0.148846
5	0.102512	0.203511	0.101138	0.000000	0.176456	0.148846

The similarity of patterns in reading news from each user can be resulted by looking for the proximity between users using the KNN algorithm.

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 3, Number 2, April 2019
DOI : https://doi.org/10.33395/sinkron.v3i1.10047
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The identification of the user's proximity is applied by using the formula of user similiarity as explained in Eq.1 and Eq.2 obtained as follows:

```
5 user yang paling besar kedekatannya untuk User 1:

1: User 35, dengan kedekatannya yaitu 0.7508238599484947
2: User 37, dengan kedekatannya yaitu 0.6560079283313361
3: User 26, dengan kedekatannya yaitu 0.6085265420656001
4: User 4, dengan kedekatannya yaitu 0.5690778189183309
5: User 28, dengan kedekatannya yaitu 0.4891315946044108
```

Fig 1. Similiar Results with The Pearson Correlation

```
5 user yang paling besar kedekatannya untuk User 1:

1: User 35, dengan kedekatannya yaitu 0.7790642602179731
2: User 37, dengan kedekatannya yaitu 0.7033260952225182
3: User 26, dengan kedekatannya yaitu 0.6688735234423067
4: User 4, dengan kedekatannya yaitu 0.6194880024773226
5: User 6, dengan kedekatannya yaitu 0.5569767952122435
```

Fig 2. Similiar Results with The Cosine Correlation

The Evaluation for KNN algorithm that has been done with RMSE formula, the predicted value is obtained with the following error:

1. *RMSE KNN cosine similiarity = 1.237*
2. *RMSE KNN pearson correlation = 1.346*

The smallest RMSE value obtained from the results of the implementation and processing of data of a certain size in this research is found in the RMSE KNN value at cosine similiarity with a value of 1,237. So, news of recommendations to users will be given using predictions from the KNN cosine similiarity algorithm.

E. CONCLUSIONS

1. The smallest RMSE value obtained from the results of the implementation and processing of data with a certain size in this research is found in the RMSE KNN value on cosine similiarity with a value of 1,237.
2. According to the original purpose, the recommendation to the user can be done using the best algorithm in accordance with the results of the experiment using the KNN cosine similiarity algorithm.

REFERENCES

[1] Gorakala, K. S, Building Recommendation Engines, Birmingham: Packt Publishing, 2016.

[2] Saranya.K.G, & G.Sudha Sadhasivam, P, "A personalized online news recommendation system," International Journal of Computer Applications, pp. 6-13, 2012.

[3] F. Isinkaye, Y. Folajimi, and B. Ojokoh, "Recommendation systems: principles, methods and evaluation," Egyptian Informatics Journal, pp. 261-273, 2015.

[4] H. H. Sarirah, "News recommendation based on web usage and web content mining." IEEE Transl. International Conference on Data Engineering Workshops (ICDEW), pp. 326-329, 2013.

[5] Chaturvedi, A. K, "Recommender system for news articles using supervised learning," Department of Information and Communications Technologies - Universitat Pompeu Fabra, 2017.

[6] S.Kaur, and E. M. Rashid, "Web news mining using Back Propagation Neural Network and clustering using K-Means algorithm in big data," Indian Journal of Science and Technology, pp. 1-8, 2016.

[7] H. Hasija and D. Chaurasia. "Recommender system with web usage mining based on Fuzzy C Means and Neural Networks," IEEE transl. International Conference on Next Generation Computing Technologies (NGCT), pp. 768-772, 2015.

[8] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," Advances in Artificial Intelligence, pp. 2-2, January 2009.

[9] R. Suguna, and D. Sharmila, "An efficient web recommendation system using collaborative filtering and pattern discovery algorithms," International Journal of Computer Applications, pp. 37-42, 2013.

[10] Sridhar, B., & Khan, M. Z, "RMSE comparison of Path Loss Models for UHF/VHF bands in India," *IEEE*, pp. 330-335, 2014.

[11] S. Yang , M. Korayem , K. AlJadda , T. Grainger , and S. Natarajan, "Combining content-based and collaborative filtering for job recommendation system," Knowledge-Based Systems, pp. 37-45, November 2017.

[12] Chintan R. Varnagar, N. N, "Web usage mining: A review on process methods and techniques," IEEE Transl. Proceedings of the International Conference on Information Communication and Embedded Systems (ICICES), pp. 40-46, Feb 2013.

[13] Pandya, R, "Web usage mining with personalization on social web," International Journal of Engineering Trends and Technology, pp. 325-328, 2015.