

Asian Journal of Research in Computer Science

Volume 15, Issue 2, Page 32-42, 2023; Article no.AJRCOS.97066 ISSN: 2581-8260

Content-based Filtering and Web Scraping in Website for Recommended Anime

Reynaldi ^a and Wirawan Istiono ^{a*}

^a Universitas Multimedia Nusantara, Jl. Scientia Boulevard, Curug Sangereng, Tangerang, Banten-15810, Indonesia.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2023/v15i2318

Open Peer Review History:

This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here:

https://www.sdiarticle5.com/review-history/97066

Accepte

Received: 02/01/2023 Accepted: 04/03/2023 Published: 09/03/2023

Original Research Article

ABSTRACT

Aim: This research aim is to determine the level of user satisfaction using the Delone and Mclean models obtained from the implementation of the content-based filtering method in the anime recommendation system.

Study Design: This study was designed with Delone and Mclean and with a Content-based Filtering method and web-scrapping to build an anime recommendation system.

Place and Duration of Study: Department of Informatic Universitas Multimedia Nusantara, between July 2022 and December 2022.

Methodology: The initial step in this research was collecting data using web scrapping and questionnaires, then followed by a literature study, and after that continued with system design and application development. After the application is made the next step is to get the level of user satisfaction with Delone and Mclean, and the final step is writing a report from this research.

Results: The design and development of a system by implementing a content-based filtering method to the website-based have been successfully created, and the results of calculating the

*Corresponding author: E-mail: wirawan.istiono@umn.ac.id;

level of user satisfaction calculated from 43 respondents using the Delone and Mclean methods show, an anime recommendation system with content-based filtering methods has good result with a user satisfaction percentage of 74.23%.

Conclusion: The anime system recommendation application has been successfully made and the results of user satisfaction are 74.23%.

Keywords: Anime; content-based filtering; recommendation system; web scraping.

1. INTRODUCTION

In routine daily activities, there is time that is free to use and is outside daily activities which are called free time. A hobby can be referred to as one of the activities or activities that can be done to fill spare time. A hobby is an activity that is carried out as entertainment and to get pleasure in free time [1,2]. One of the entertainment that have now become a hobby that is in people loved today is watching anime. Interesting illustrations and lots of light but high-quality stories make anime easy for people to like [3]. Surabaya and Jakarta have been included in the list of 19 big cities that have the most anime fans in the world [4].

The growing level of popularity of Japanese animation in Indonesia is marked by the increasing number of people who want to learn Nihongo (Japanese language) or people who want to go to Japan because of watching anime and can also be seen in Muse Indonesia's YouTube channel, which currently has more than 7 million subscribers. Muse Indonesia is the Indonesian branch of the company that handles anime production and distribution [5]. Therefore Japanese culture which is currently popular makes big changes that occur to social values in society. Comic Frontier (Comifuro) is an exhibition event that is expected to be a place to channel the interests and talents of independent creators. Participants can spread their work directly by selling and meeting people who have the same interests. Comifuro is usually attended by fans of Japanese culture (anime, manga, or vtuber fans), where there are cosplayers who play characters from certain anime or certain vtubers [6].

Anime is an English absorption word in Japanese from the word "Animation". Along with the development of the times, anime has become a category for animated film series made in Japan or having a visual style similar to animated films originating from Japan [6]. Along with the development of the story model of anime, the categorization of anime genres is increasing,

there are more than 20 genres and subgenres [7]. MyAnimeList.net is a website that allows users to follow and review anime and manga. This site has been operating since 2004 and currently has more than 5 million members. MvAnimeList includes a complete list of anime and manga, new releases, articles, discussions, and much more. MyAnimeList also offers a variety of other features, including wish lists, currently watching lists, text and video articles, In addition, the number more. MyAnimeList users from Indonesia is ranked 2nd with a percentage of 7.2% of the total MyAnimeList users. Because of this, this website is suitable as a source of data to be studied [5,8].

This research was made using the content-based filtering method because there is not much research on anime recommendation systems that only use the content-based filtering method as the main method [9], by using this method as the main method the websites created can be more focused on satisfying the choices or desires of the user [10,11]. If viewed from other anime recommendation system journals that use the filtering method in their research [12], this research uses the collaborative-based filtering method as the main method or as an additional method in the research conducted. In this case, this research will implement the Content-Based Filtering method on an anime recommendation website and find user satisfaction with the DeLone and McLean methods

2. LITERATURE REVIEW

2.1 Content-Based Filtering

Content-based filtering is a Machine Learning technique that uses attribute similarity to make decisions. This technique is commonly used in building systems that provide a recommendation, namely the design of an algorithm to advertise/recommend something to users based on data collected about users [13]. This method generates a recommendation by using the keywords and attributes assigned to objects in

the database and matching them with the user profile. User profiles are made based on data obtained from user activities, such as ratings (likes and dislikes) or items searched for on websites [14]. This method is used to recommend items to users based on their previous preferences and interests. It uses item content (such as movie descriptions, book summaries, etc.) to recommend similar items. This method is often used on online shopping sites and streaming services such as Netflix and Amazon. A recommendation system that uses the content-based filtering method will provide recommendations for items that have similarities to the items the user chooses or likes [15]. The advantage of this model is that it does not require any data about other users because the recommendations are specific to this user. This model can improve the accuracy of the recommendation results, this model also has the ability to make more specific recommendations. and also can make recommendations based on user preferences [16,17].

2.2 Web Scraping

Web Scraping is a technique of retrieving information or data from a website by utilizing the HTML or XML structure of the website. The process of this technique is usually done using a code that can retrieve data from the website automatically. Web Scraping is one kind of data mining. Which is the step of Web Scraping is to get that information still unstructured data from the website and turn it into a structure so that later it can be understood more easily such as spreadsheets, databases, or comma-separated values (CSV) files [18]. Web scraping is often used for collecting data needed for analysis, research, or other purposes from one or several websites. Although web scraping can provide many benefits, especially in collecting the necessary data, there are also some ethical considerations to consider. Some websites may not allow web scraping, so be aware not to do web scraping illegally or violate copyright from the website [19].

2.3 Preprocessing

Preprocessing is the process of preparing data for analysis by cleaning, changing, and organizing it. This includes tasks such as removing outliers, normalizing data, and encoding categorical variables. Preprocessing is an important step in a data science workflow because it helps ensure that data is ready for analysis. Data that has gone through the

preprocessing stage will become more structured data [20,21]. Preprocessing stages are as follows:

- Case Folding is the process of transforming all the letters and words in the anime data, be it the title, genre, or anime studio in the document into lowercase letters. This helps reduce vocabulary size and increases the accuracy of text classification algorithms.
- 2. Tokenization is the process of breaking text into words, phrases, symbols, or other elements called tokens.
- Elimination is a technique of preprocessing used to reduce the number of features in a dataset and remove duplicated words. Duplicated words are assumed to have the same features, only 1 word will be stored if there is the same word.
- Filtering is a preprocessing technique that involves a subset of data from the original dataset based on certain criteria. This filtering can be used to reduce dataset size, remove irrelevant data or focus on certain features.
- 5. Stemming is a preprocessing technique used to reduce the number of words in a document by removing prefixes & suffixes, in other words transforming a word that has a prefix or suffix to only basic words.

2.4 Cosine Similarity

In data mining, the similarity measure refers to the distance to the dimensions that represent the features of the data objects in the data set. If the distance is smaller then the level of similarity will be high, but if the distance is large then the level of similarity will be low. Cosine Similarity is the cosine of the angle between vectors. Vectors are usually non-zero and are in the product space in Cosine Similarity [22,23].

$$sim(A,B) = \frac{n(A \cap B)}{\sqrt{n(A)n(B)}}$$

The angle between two vectors is usually used to calculate the similarity between two objects. The cosine similarity function between item A and item B is shown as follows.

Information:

- sim(A,B) = similarity value of item A and item B.
- n(A) = the number of features in the content of item A.

- n(B) = the number of features in the content of item B.
- n(A,B) = the number of content features contained in item A and in item B.

The two objects that have a similarity value equal to 1 or the greater the value of the similarity function, the two objects being calculated are considered similar or identical and vice versa.

2.5 Top-N Recommendation

Top-N recommendation is a technique used in system recommendations to suggest the best number of items to the user. Values the results of cosine similarity calculations are used to provide rank recommendations to users. The value of the calculation results of cosine similarity with more similarity values predicted height will be the user's choice [24]. To determine the best items it will be suggested to users use the filtering method.

2.6 Confusion Matrix

A confusion Matrix is a popular method used when solving classification problems and can be used to determine the performance of a system by comparing the classification results of the system with the actual classification [25]. This method can be applied to binary classification as well as to multiclass classification problems.

2.7 Model Delone Dan Mclean

The Delone and Mclean model is a model for determining the success of information systems developed by DeLone and McLean in 1992. It is based on the premise that information system success is a function of five main dimensions: System Quality, Information Quality, Service Quality, User Satisfaction, and Net Benefits. The model explains that system quality will affect system use and user satisfaction. Information quality will also affect the use and user satisfaction. User usage and satisfaction will ultimately affect the Individual Impact, and the aggregates of the Individual Impact will ultimately affect the Organizational Impact [26].

3. METHODOLOGY

In the research process "Implementation of the Content-Based Filtering Method in the Anime Recommendation System" was carried out in the following stages.

3.1 Data Collection

The data collection method used is Web Scraping and a Questionnaire.

- Web scraping is used in this study to obtain anime data who want to research from myanimelist.net. Data is retrieved by fetching data from the HTML file of the summer 2022 page myanimelist.net.
- The questionnaire used in this study to determine which category will be added to the application.

3.2 System Design

Application design starts from designing the user interface design, designing the flow of content-based filtering that is used to calculate values to produce a ranking in a recommendation system with a flowchart and database structure.

3.3 System Building

At this stage, an anime recommendation system will be built and the data used by the system will be taken from answers to questionnaires that have been distributed using the Google form. This system will be made into a website. At this stage, the development of the user interface is carried out using the bootstrap framework, writing code using the PHP language for HTML, and implementing content-based filtering.

3.4 System Testing

The system testing process is carried out to test the successful implementation of the anime recommendation system using the confusion matrix method.

3.5 User Satisfaction

The process of finding the level of user satisfaction with the Delone and Mclean models uses a questionnaire as a method of collecting data on user satisfaction.

3.6 Model Delone dan Mclean

In designing an anime recommendation system using the content-based filtering method, there are several main components including Entity relationship diagrams (ERD), flowcharts, database system structures, and interface display designs. Fig. 1 shows an ER diagram of the database used in the development of the

recommendation system, the diagram uses Crow's Foot notation and will only display entity tables that have relationships with other tables [28].

Fig. 2 is a flowchart for the main page. On the main page, there is a feature to search for the anime the user wants without the need to log in. users who input by searching or by clicking the criteria button, the page will display anime

recommendations according to the criteria entered by the user.

Fig. 3 is the process of calculating the content-based filtering method, this process occurs after inputting criteria that begins with the preprocessing step (case folding, tokenization, elimination, filtering, and stemming), cosine similarity, and ranking the cosine similarity score using the Top method -N Recommendations.

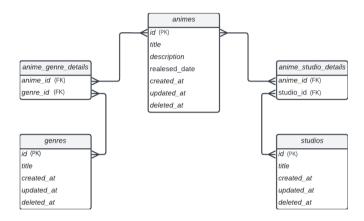


Fig. 1. Entity relationship diagram (ERD) recommendation system

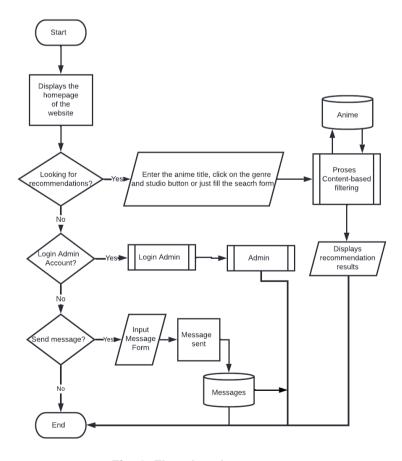


Fig. 2. Flowchart home page

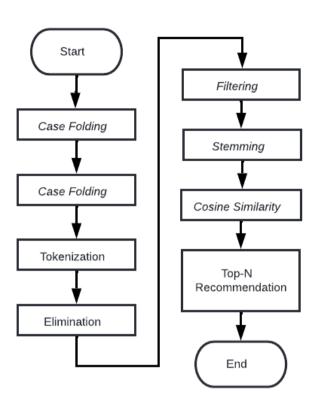


Fig. 3. Flowchart content-based filtering

4. RESULTS

Fig. 4 is a display of the website's main page. The main page has a logo display, a menu in the header, and a feature to search for anime.

Fig. 5 is a display of the main page search section of the website. This section allows you to search for anime by typing a title and selecting a genre or studio.

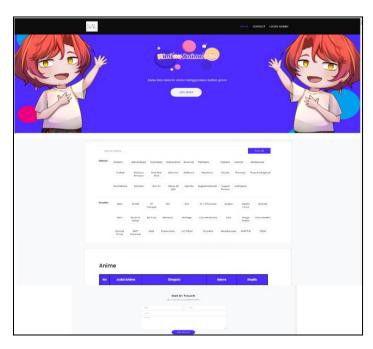


Fig. 4. Home page



Fig. 5. Search anime

Fig. 6 is part of the website's main page to display data on anime recommendations that have been searched. The main page has a logo display, a menu in the header, and a feature to search for anime.

The implementation of the Content-Based Filtering method used can be seen in Fig. 7, the code explained carrying out the preprocessing process starting from the case folding stage to the steaming stage.

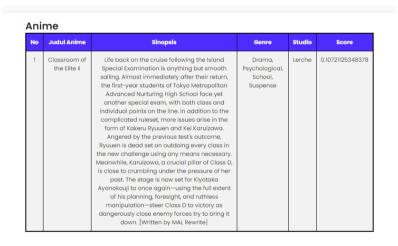


Fig. 6. Search results display

```
public function search()

{
    /*** Membuat genre dan studio menjadi string ***/
    $genre_string = implode(" ", $genre);
    $studio_string = implode(" ", $studio);
    /*** Proses case folding (merubah kata-kata menjadi lowercase)
    ***/

$ text_tolowercase = strtolower($search);
    $genre_tolowercase = strtolower($genre_string);
    $studio_tolowercase = strtolower($studio_string);
    /*** filtering (stopword removal) ***/
    /* 1. mengambil text stopword */
    $stopwords = array_column($this->mod->get_stopword(), 'word');
    $this->load->view('Home', $data);
}
```

Fig. 7. Code preprocessing snippets

```
public function calculate_cosine($target, $data)

{
    $count_target = count($target);
    $count_data = count($data);

$count_same_array = count(array_intersect($target, $data));

return $count_same_array / (sqrt($count_target * $count_data));

;
}
```

Fig. 8. Code cosine similarity snippet

The code snippet shown in Fig. 8 is the code for carrying out the cosine similarity process to get a similarity score which will later be used as a reference for sorting the recommended anime displays.

To determine the precision and accuracy of the algorithm used. Precision and accuracy are tested by predicting the results of recommendations that have the word 'night' in the title or anime synopsis. Testing is done by searching for anime through the website created, then comparing the recommendation results with the actual values in the database. Furthermore, the data that has been obtained is inputted into Table 1.

In this case:

- True Positive (TP): The prediction results show 13 anime that have the word "Night" in the title or synopsis and the result is correct.
- True Negative (TN): The results of the prediction are 227 anime that don't have the words "Night" in the title or synopsis and the result is correct.
- False Positive (FP): The prediction result shows 0 anime shown had the word "Night" and did not have the word.
- False Negative (FN): The prediction results show 0 anime that are not shown as having the word "Night" and apparently did.

After the prediction results that are shown in Table 1, the next step can be to calculate Accuracy and Precision values. The following is a step of the calculation, where Precision and accuracy for the first test each have a value of 1.0 and 1.0.

1. Precision describes the accuracy between the data sought and the predictions/ recommendations of the results provided by the model.

Precision =
$$\frac{13}{13+0}$$
 x100 = 100%

2. Accuracy describes how accurate the model is in classifying appropriately

Accuracy =
$$\frac{13+230}{13+230+0+0}$$
 $x100 = 100\%$

The trial of this system uses the DeLone and McLean method. The trial used a questionnaire containing seven questions and was distributed via Google Forms. From the questionnaire, the results of the questionnaire answers from 43 respondents can be seen in Table 2.

After all, the percentage score calculations have been carried out and the score percentages for each question variable have been obtained, the average value (mean) will then be sought to determine whether the "Anime Selection Recommendations" website can be considered successful or not. The calculation of the average (mean) is as follows.

Score Percentage

$$= (78.83 + 68.37 + 74.65 + 73.49 + 75.81) \times 100$$

= 74.23%

Table 1. Trial table

n = 240		Actual Result			
Prediction		True	False		
	True	13	0		
	False	0	227		

Table 2. List of the question for respondents

No.	Question	Answer Choices					
	System Quality	Very Bad	Bad	Neutral	Good	Excellent	
1	Is the "Anime Recommendation" website easy to use?	0	2	7	21	13	
2	Does the "Anime Recommendation" website already have a good/attractive appearance?	1	2	12	16	12	
	Information Quality						
3	Does the "Recommendation Anime" website have pretty much-recommended anime?	3	8	12	8	12	
	Service Quality						
4	Can this "Anime Recommendation" website be run quickly and easily?	0	7	9	15	12	
5	Does the "Anime" Recommendation" website give you the right recommendations? User Satisfaction	1	3	13	16	10	
6	Can this "Anime Recommendation" website help you choose anime?	3	3	13	10	14	
	Net Benefit						
7	Is this "Anime Recommendation" website useful for you?	1	2	15	12	13	

Based on the calculation result, it can be concluded that the average user satisfaction with the mean value is 74.23%. The results of the questionnaire data can be considered as a system that is accepted by users as a website that provides good anime recommendations.

5. CONCLUSION

The conclusions drawn based on the results of research conducted in building an anime recommendation system using the Content-Based Filtering method, it can be concluded that the design and development of the system by implementing a website-based content-based filtering method for users has been successfully carried out and the results of calculating the level of user satisfaction are calculated using the Delone and Mclean method, showing anime recommendation websites with content-based

filtering methods that are good with a calculation result of 74.23%.

ACKNOWLEDGEMENTS

Thank you to the Universitas Multimedia Nusantara, Indonesia which has become a place for researchers to develop this journal research. Hopefully, this research can make a major contribution to the advancement of technology in Indonesia.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Lee HK. Participatory media fandom: A case study of anime fansubbing, Media,

- Culture and Society. 2011;33(8):1131–1147.
- DOI: 10.1177/0163443711418271
- Al Al Sada M et al. From Anime to Reality: Embodying An Anime Character As A Humanoid Robot, Conference on Human Factors in Computing Systems -Proceedings, no. Ml; 2021. DOI: 10.1145/3411763.3451543
- Yamato E. Construction of discursive fandom and structural fandom through anime comics and game fan conventions in Malaysia, European Journal of Cultural Studies. 2018;21(4):469–485.
 DOI: 10.1177/1367549416682964
- 4. Toi Y. Kepopuleran dan Penerimaan Anime Jepang Di Indonesia, Ayumi: Jurnal Budaya, Bahasa dan Sastra. 2020;7(1): 68–82.
 - DOI: 10.25139/ayumi.v7i1.2808
- Wikayanto A. Analysis of the Potential Development of Technopark for Film, Game and Animation Industry in Indonesia, 1st International Conference on Art, Craft, Culture, and Design; 2017.
- 6. Yuliani R, Mulyadi RM, Adji M. Japanese Soft Power in Indonesia on Anime Entitled Ufo Baby: Study of Popular Culture, Izumi. 2021;10(2):328–337.
 - DOI: 10.14710/izumi.10.2.328-337
- Matthews P, Glitre K. Genre analysis of movies using a topic model of plot summaries," Journal of the Association for Information Science and Technology. 2021;72(12):1511–1527.
 - DOI: 10.1002/asi.24525
- 8. Mamat R, Rashid RA, Paee R, Ahmad N. VTubers and anime culture: A case study of Japanese learners in two public universities in Malaysia, International Journal of Health Sciences. 2022;11958–11974.
 - DOI: 10.53730/ijhs.v6ns2.8231
- Hansen Christian Zaputra WI. Constructing a Culinary Tourism Recommendation System Based in Palembang, Indonesian Using Weighted Product Method, International Journal of Multidisciplinary Research and Analysis. 2022;05(08):1908–1917.
 - DOI: 10.47191/ijmra/v5-i8-01
- Natarajan S, Vairavasundaram S, Natarajan S, Gandomi AH. Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data, Expert Systems with Applications. 2020;149.

- DOI: 10.1016/j.eswa.2020.113248
- Logesh R, Subramaniyaswamy V, Malathi D, Sivaramakrishnan N, Vijayakumar V. Enhancing recommendation stability of collaborative filtering recommender system through bio-inspired clustering ensemble method, Neural Computing and Applications. 2020;32(7):2141–2164. DOI: 10.1007/s00521-018-3891-5
- Prasetyo C, Istiono W, Nusantara UM, Nusantara UM. Fitness Exercise Recommendation System Using Weighted Products, International Journal of Emerging Trends in Engineering Research. 2021;9(9):1234–1238. DOI: 10.30534/ijeter/2021/05992021
- Nassar N, Jafar A, Rahhal Y. A novel deep multi-criteria collaborative filtering model for recommendation system, Knowledge-Based Systems. 2020;187:104811. DOI: 10.1016/j.knosys.2019.06.019
- 14. Girsang AS, Al Faruq B, Herlianto HR, Simbolon S. Collaborative Recommendation System in Users of Anime Films, Journal of Physics: Conference Series. 2020;1566(1). DOI: 10.1088/1742-6596/1566/1/012057
- Tadi Bani N, Fekri-Ershad S. Content-based image retrieval based on combination of texture and colour information extracted in spatial and frequency domains, Electronic Library. 2019;37(4):650–666
 DOI: 10.1108/EL-03-2019-0067
- Latif A et al. Content-based image retrieval and feature extraction: A comprehensive review, Mathematical Problems in Engineering; 2019. DOI: 10.1155/2019/9658350
- Jia C et al. Content-Aware Convolutional Neural Network for In-Loop Filtering in High Efficiency Video Coding, IEEE Transactions on Image Processing. 2019;28(7):3343–3356.
 DOI: 10.1109/TIP.2019.2896489
- Singrodia V, Mitra A, Paul S. A Review on Web Scrapping and its Applications. International Conference on Computer Communication and Informatics, ICCCI. 2019;1–6.
 - DOI: 10.1109/ICCCI.2019.8821809
- Khder MA. Web scraping or web crawling: State of art, techniques, approaches and application, International Journal of Advances in Soft Computing and its Applications. 2021;13(3):144–168.
 DOI: 10.15849/ijasca.211128.11

- Alexandropoulos SAN, Kotsiantis SB, Vrahatis MN. Data preprocessing in predictive data Mining. 2019;34.
 DOI: 10.1017/S026988891800036X
- 21. Gonzalez Zelaya CV. Towards explaining the effects of data preprocessing on machine learning, Proceedings International Conference on Data Engineering. 2019:2086–2090. DOI:10.1109/ICDE.2019.00245
- Abdel-Basset M, Mohamed M, Elhoseny M, Son LH, Chiclana F, Zaied AENH. Cosine similarity measures of bipolar neutrosophic set for diagnosis of bipolar disorder diseases, Artificial Intelligence in Medicine. 2019;101:101735.
 DOI: 10.1016/j.artmed.2019.101735
- 23. Thongtan T, Phienthrakul T. Sentiment classification using document embeddings trained with cosine similarity, ACL 2019 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop. 2019; 407–414.

- DOI: 10.18653/v1/p19-2057
- 24. Anelli VW, Bellogín A, Di Noia T, Jannach D, Pomo C. Top-N Recommendation Algorithms: A Quest for the State-of-the-Art, UMAP2022 Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. 2022; 121–131.
 - DOI: 10.1145/3503252.3531292
- Hasnain M, Pasha MF, Ghani I, Imran M, Alzahrani MY, Budiarto R. Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking, IEEE Access. 2020;8:90847–90861.
 - DOI: 10.1109/ACCESS.2020.2994222
- 26. Karnita A, Kurniawan, Suangga A. Analysis of Online BPHTB Application Success System Using Information System Success Models DeLone and McLean (Case Study of the Revenue Service, Financial Management, and Regional Assets of Subang Regency), JPSAM (Journal of Public Sector Accounting and Management). 2019;1(1):55–69.

© 2023 Reynaldi and Istiono; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/97066