

1
2
3
4
5
6
7
8
9
10

CONTENT-BASED FILTERING AND WEB SCRAPING IN WEBSITE FOR RECOMMENDED ANIME

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-scraping 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.

Result: 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 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%.

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

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

32 spread their work directly by selling and meeting people who have the same interests.
33 Comifuro is usually attended by fans of Japanese culture (anime, manga, or vtuber fans),
34 where there are cosplayers who play characters from certain anime or certain vtubers [6].
35

36 Anime is an English absorption word in Japanese from the word "Animation". Along with the
37 development of the times, anime has become a category for animated film series made in
38 Japan or having a visual style similar to animated films originating from Japan [6]. Along with
39 the development of the story model of anime, the categorization of anime genres is
40 increasing, there are more than 20 genres and subgenres [7]. MyAnimeList.net is a website
41 that allows users to follow and review anime and manga. This site has been operating since
42 2004 and currently has more than 5 million members. MyAnimeList includes a complete list
43 of anime and manga, new releases, articles, discussions, and much more. MyAnimeList also
44 offers a variety of other features, including wish lists, currently watching lists, text and video
45 articles, and more. In addition, the number of MyAnimeList users from Indonesia is ranked
46 2nd with a percentage of 7.2% of the total MyAnimeList users. Because of this, this website is
47 suitable as a source of data to be studied [5], [8].
48

49 This research was made using the content-based filtering method because there is not much
50 research on anime recommendation systems that only use the content-based filtering
51 method as the main method [9], by using this method as the main method the websites
52 created can be more focused on satisfying the choices or desires of the user [10], [11]. If
53 viewed from other anime recommendation system journals that use the filtering method in
54 their research [12], this research uses the collaborative-based filtering method as the main
55 method or as an additional method in the research conducted. In this case, this research will
56 implement the Content-Based Filtering method on an anime recommendation website and
57 find user satisfaction with the DeLone and McLean methods
58

59 **2. LITERATURE REVIEW**

60

61 **2.1 Content-Based Filtering**

62 Content-based filtering is a Machine Learning technique that uses attribute similarity to make
63 decisions. This technique is commonly used in building systems that provide a
64 recommendation, namely the design of an algorithm to advertise/recommend something to
65 users based on data collected about users [13]. This method generates a recommendation
66 by using the keywords and attributes assigned to objects in the database and matching them
67 with the user profile. User profiles are made based on data obtained from user activities,
68 such as ratings (likes and dislikes) or items searched for on websites [14]. This method is
69 used to recommend items to users based on their previous preferences and interests. It
70 uses item content (such as movie descriptions, book summaries, etc.) to recommend similar
71 items. This method is often used on online shopping sites and streaming services such as
72 Netflix and Amazon. A recommendation system that uses the content-based filtering method
73 will provide recommendations for items that have similarities to the items the user chooses
74 or likes [15]. The advantage of this model is that it does not require any data about other
75 users because the recommendations are specific to this user. This model can improve the
76 accuracy of the recommendation results, this model also has the ability to make more
77 specific recommendations, and also can make recommendations based on user preferences
78 [16], [17].
79

80 **2.2 Web Scraping**

81 Web Scraping is a technique of retrieving information or data from a website by utilizing the
82 HTML or XML structure of the website. The process of this technique is usually done using a
83 code that can retrieve data from the website automatically. Web Scraping is one kind of data
84 mining. Which is the step of Web Scraping is to get that information still unstructured data

85 from the website and turn it into a structure so that later it can be understood more easily
86 such as spreadsheets, databases, or comma-separated values (CSV) files [18]. Web
87 scraping is often used for collecting data needed for analysis, research, or other purposes
88 from one or several websites. Although web scraping can provide many benefits, especially
89 in collecting the necessary data, there are also some ethical considerations to consider.
90 Some websites may not allow web scraping, so be aware not to do web scraping illegally or
91 violate copyright from the website [19].

92

93 **2.3 Preprocessing**

94 Preprocessing is the process of preparing data for analysis by cleaning, changing, and
95 organizing it. This includes tasks such as removing outliers, normalizing data, and encoding
96 categorical variables. Preprocessing is an important step in a data science workflow
97 because it helps ensure that data is ready for analysis. Data that has gone through the
98 preprocessing stage will become more structured data [20], [21]. Preprocessing stages are
99 as follows:

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

114

115 **2.4 Cosine Similarity**

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

121

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

122

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

125

126 Information:

- 127 • $sim(A,B)$ = similarity value of item A and item B.
- 128 • $n(A)$ = the number of features in the content of item A.
- 129 • $n(B)$ = the number of features in the content of item B.
- 130 • $n(A,B)$ = the number of content features contained in item A and in item B.

131

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

135

136 **2.5 Top-N Recommendation**

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

142

143 **2.6 Confusion Matrix**

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

148

149 **2.7 Model Delone dan Mclean**

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

158

159 **3. METHODOLOGY**

160

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

163

164 **A. Data Collection**

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

- 166 1. Web scraping is used in this study to obtain anime data who want to research from
167 myanimelist.net. Data is retrieved by fetching data from the HTML file of the summer
168 2022 page myanimelist.net.
- 169 2. The questionnaire used in this study to determine which category will be added to the
170 application.

171

172 **B. System Design**

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

176

177 **C. System Building**

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

183

184 **D. System Testing**

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

187

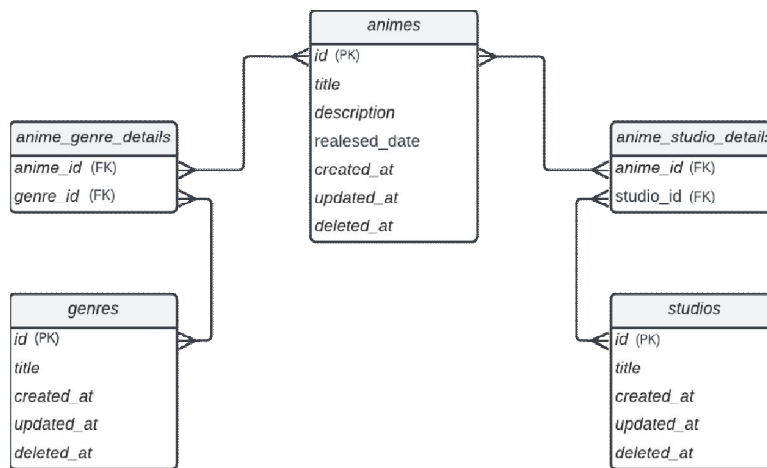
188 **E. User Satisfaction**

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

191

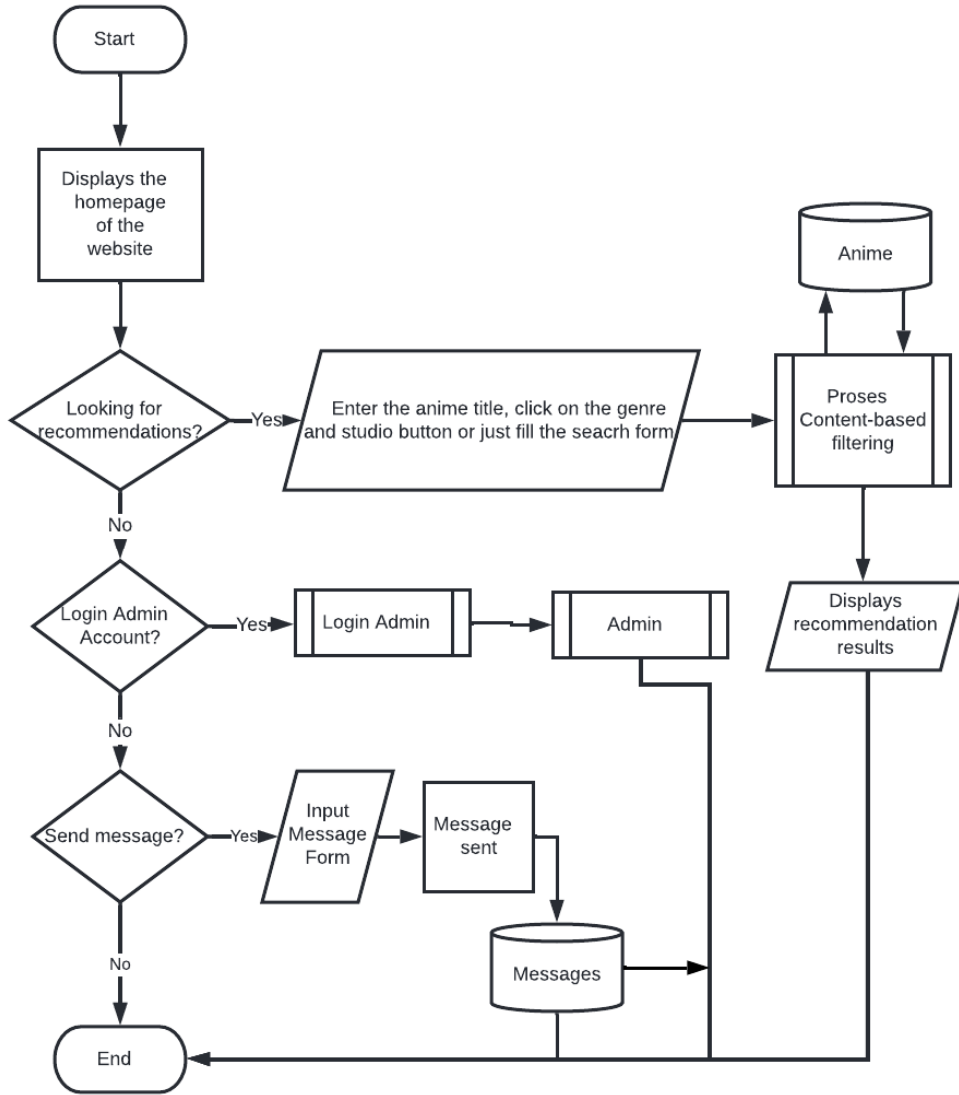
192 **3.1 Model Delone dan Mclean**

193 In designing an anime recommendation system using the content-based filtering method,
194 there are several main components including Entity relationship diagrams (ERD), flowcharts,
195 database system structures, and interface display designs. Fig 1 shows an ER diagram of
196 the database used in the development of the recommendation system, the diagram uses
197 Crow's Foot notation and will only display entity tables that have relationships with other
198 tables. [28].
199



200
201 **Fig. 1. Entity relationship diagram (ERD) recommendation system**
202

203 Fig 2 is a flowchart for the main page. On the main page, there is a feature to search for the
204 anime the user wants without the need to log in. users who input by searching or by clicking
205 the criteria button, the page will display anime recommendations according to the criteria
206 entered by the user.
207



208
 209
 210
 211
 212
 213
 214
 215

Fig 2. Flowchart home page

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.

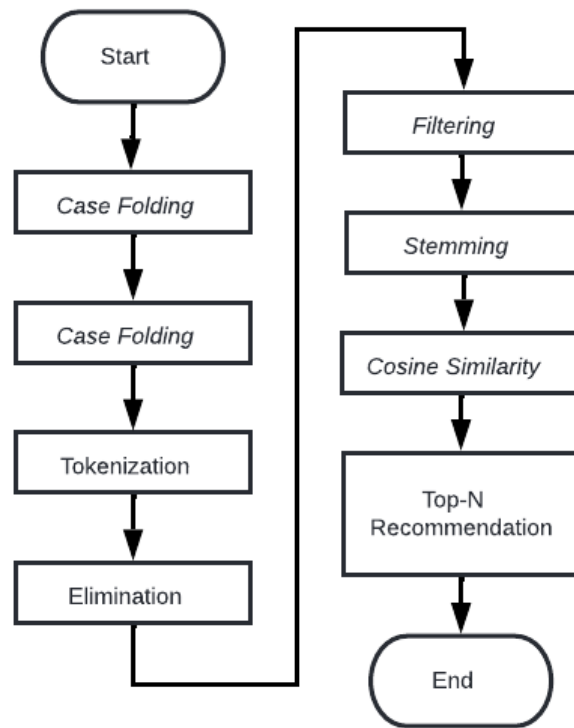


Fig 3. Flowchart content-based filtering

216
 217
 218
 219
 220
 221
 222
 223
 224

4. RESULT

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.

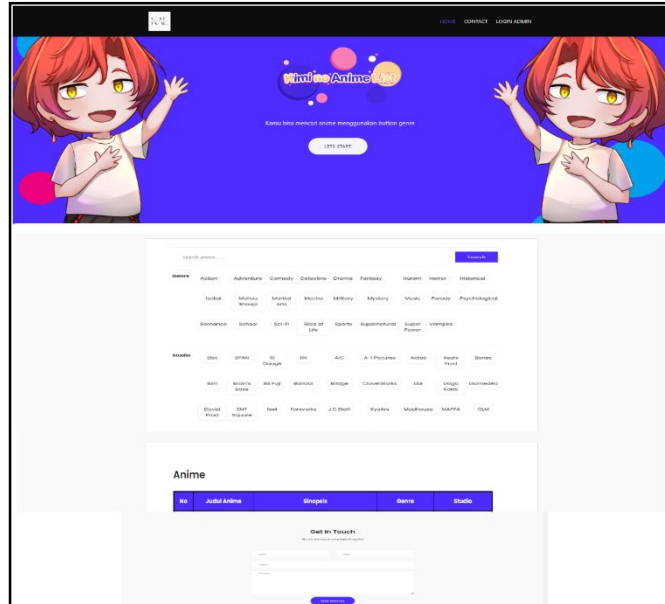


Fig 4. Home page

225
226
227
228
229
230

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.

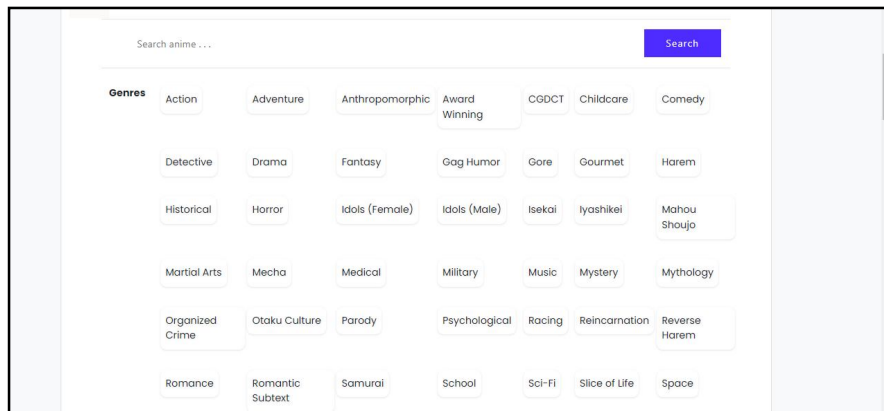


Fig 5. Search anime

231
232
233
234
235
236
237

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.

Anime

No	Judul Anime	Sinopsis	Genre	Studio	Score
1	Classroom of the Elite II	Life back on the cruise following the Island Special Examination is anything but smooth sailing. Almost immediately after their return, the first-year students of Tokyo Metropolitan Advanced Nurturing High School face yet another special exam, with both class and individual points on the line. In addition to the complicated ruleset, more issues arise in the form of Kakeru Ryuuken and Kei Karuzawa. Angered by the previous test's outcome, Ryuuken is dead set on outdoing every class in the new challenge using any means necessary. Meanwhile, Karuzawa, a crucial pillar of Class D, is close to crumbling under the pressure of her past. The stage is now set for Kiyotaka Ayanokouji to once again—using the full extent of his planning, foresight, and ruthless manipulation—steer Class D to victory as dangerously close enemy forces try to bring it down. [Written by MAL Rewrite]	Drama, Psychological, School, Suspense	Lerche	0.1072125348378

Fig 6. Search results display

238
239
240
241
242
243
244

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.

```
1  
2 public function search()  
3 {  
4     /*** Membuat genre dan studio menjadi string ***/  
5     $genre_string = implode(" ", $genre);  
6     $studio_string = implode(" ", $studio);  
7     /*** Proses case folding (merubah kata-kata menjadi lowercase)  
8     ***/  
9     $text_lowercase = strtolower($search);  
10    $genre_lowercase = strtolower($genre_string);  
11    $studio_lowercase = strtolower($studio_string);  
12    /*** filtering (stopword removal) ***/  
13    /* 1. mengambil text stopwords */  
14    $stopwords = array_column($this->mod->get_stopword(), 'word');  
15    $this->load->view('Home', $data);  
}
```

Fig 7. Code preprocessing snippets

245
246
247
248
249
250
251

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.

```

1
2 public function calculate_cosine($target, $data)
3 {
4     $count_target = count($target);
5     $count_data = count($data);
6
7     $count_same_array = count(array_intersect($target, $data));
8
9     return $count_same_array / (sqrt($count_target * $count_data));
10 }

```

Fig 8. Code cosine similarity snippet

252
253
254
255
256
257
258
259
260
261

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.

Table 1: Trial table

	n = 240	Actual Result	
		TRUE	FALSE
Prediction	TRUE	13	0
	FALSE	0	227

262
263
264
265
266
267
268
269
270
271
272

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.

273
274
275
276
277

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.

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

278
279

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

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

280

281 The trial of this system uses the DeLone and McLean method. The trial used a questionnaire
 282 containing seven questions and was distributed via Google Forms. From the questionnaire,
 283 the results of the questionnaire answers from 43 respondents can be seen in Table 2.
 284
 285

Table 2. List of the question for respondents

No	Question	Answer Choices				
		Very Bad	Bad	Neutral	Good	Excellent
	System Quality					
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?	1	3	13	16	10
	User Satisfaction					
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

286
 287 After all, the percentage score calculations have been carried out and the score percentages
 288 for each question variable have been obtained, the average value (mean) will then be sought
 289 to determine whether the "Anime Selection Recommendations" website can be considered
 290 successful or not. The calculation of the average (mean) is as follows.
 291

$$\text{Score Percentage} = \frac{(78.83 + 68.37 + 74.65 + 73.49 + 75.81)}{100} \times 100$$

$$= 74.23\%$$

292
 293
 294
 295 Based on the calculation result, it can be concluded that the average user satisfaction with
 296 the mean value is 74.23%. The results of the questionnaire data can be considered as a
 297 system that is accepted by users as a website that provides good anime recommendations.
 298
 299

300 **4. CONCLUSION**

301

302 The conclusions drawn based on the results of research conducted in building an anime
303 recommendation system using the Content-Based Filtering method, it can be concluded that
304 the design and development of the system by implementing a website-based content-based
305 filtering method for users has been successfully carried out and the results of calculating the
306 level of user satisfaction are calculated using the Delone and Mclean method, showing
307 anime recommendation websites with content-based filtering methods that are good with a
308 calculation result of 74.23%.

309

310 **ACKNOWLEDGEMENTS**

311

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

315

316 **REFERENCES**

317

- 318 [1] H. K. Lee, "Participatory media fandom: A case study of anime fansubbing," *Media,*
319 *Culture and Society*, vol. 33, no. 8, pp. 1131–1147, 2011, doi:
320 10.1177/0163443711418271.doi:10.1177/0163443711418271
- 321 [2] M. Al Al Sada *et al.*, "From Anime to Reality: Embodying An Anime Character As A
322 Humanoid Robot," *Conference on Human Factors in Computing Systems -*
323 *Proceedings*, no. MI, 2021, doi:
324 10.1145/3411763.3451543.doi:10.1145/3411763.3451543
- 325 [3] E. Yamato, "Construction of discursive fandom and structural fandom through anime
326 comics and game fan conventions in Malaysia," *European Journal of Cultural Studies*,
327 vol. 21, no. 4, pp. 469–485, 2018, doi:
328 10.1177/1367549416682964.doi:10.1177/1367549416682964
- 329 [4] Y. Toi, "Kepopuleran dan Penerimaan Anime Jepang Di Indonesia," *Ayumi: Jurnal*
330 *Budaya, Bahasa dan Sastra*, vol. 7, no. 1, pp. 68–82, 2020, doi:
331 10.25139/ayumi.v7i1.2808.doi:10.25139/ayumi.v7i1.2808
- 332 [5] A. Wikayanto, "Analysis of the Potential Development of Technopark for Film , Game
333 and Animation Industry in Indonesia," *1st International Conference on Art, Craft,*
334 *Culture, and Design*, no. September 2017, 2017.
- 335 [6] R. Yuliani, R. M. Mulyadi, and M. Adji, "Japanese Soft Power in Indonesia on Anime
336 Entitled Ufo Baby: Study of Popular Culture," *Izumi*, vol. 10, no. 2, pp. 328–337, 2021,
337 doi: 10.14710/izumi.10.2.328-337.doi:10.14710/izumi.10.2.328-337
- 338 [7] P. Matthews and K. Glitre, "Genre analysis of movies using a topic model of plot
339 summaries," *Journal of the Association for Information Science and Technology*, vol.
340 72, no. 12, pp. 1511–1527, 2021, doi: 10.1002/asi.24525.doi:10.1002/asi.24525
- 341 [8] R. Mamat, R. A. Rashid, R. Pae, and N. Ahmad, "VTubers and anime culture: A case
342 study of Japanese learners in two public universities in Malaysia," *International journal*
343 *of health sciences*, no. August, pp. 11958–11974, 2022, doi:
344 10.53730/ijhs.v6ns2.8231.doi:10.53730/ijhs.v6ns2.8231
- 345 [9] W. I. Hansen Christian Zaputra, "Constructing a Culinary Tourism Recommendation
346 System Based in Palembang, Indonesian Using Weighted Product Method,"
347 *International Journal of Multidisciplinary Research and Analysis*, vol. 05, no. 08, pp.
348 1908–1917, 2022, doi: 10.47191/ijmra/v5-i8-01.doi:10.47191/ijmra/v5-i8-01
- 349 [10] S. Natarajan, S. Vairavasundaram, S. Natarajan, and A. H. Gandomi, "Resolving data
350 sparsity and cold start problem in collaborative filtering recommender system using
351 Linked Open Data," *Expert Systems with Applications*, vol. 149, 2020, doi:
352 10.1016/j.eswa.2020.113248.doi:10.1016/j.eswa.2020.113248

- 353 [11] R. Logesh, V. Subramaniaswamy, D. Malathi, N. Sivaramakrishnan, and V.
354 Vijayakumar, "Enhancing recommendation stability of collaborative filtering
355 recommender system through bio-inspired clustering ensemble method," *Neural*
356 *Computing and Applications*, vol. 32, no. 7, pp. 2141–2164, 2020, doi: 10.1007/s00521-
357 018-3891-5.doi:10.1007/s00521-018-3891-5
- 358 [12] C. Prasetyo, W. Istiono, U. M. Nusantara, and U. M. Nusantara, "Fitness Exercise
359 Recommendation System Using Weighted Products," *International Journal of Emerging*
360 *Trends in Engineering Research*, vol. 9, no. 9, pp. 1234–1238, 2021, doi:
361 10.30534/ijeter/2021/05992021.doi:10.30534/ijeter/2021/05992021
- 362 [13] N. Nassar, A. Jafar, and Y. Rahhal, "A novel deep multi-criteria collaborative filtering
363 model for recommendation system," *Knowledge-Based Systems*, vol. 187, p. 104811,
364 2020, doi: 10.1016/j.knosys.2019.06.019.doi:10.1016/j.knosys.2019.06.019
- 365 [14] A. S. Girsang, B. Al Faruq, H. R. Herlianto, and S. Simbolon, "Collaborative
366 Recommendation System in Users of Anime Films," *Journal of Physics: Conference*
367 *Series*, vol. 1566, no. 1, 2020, doi: 10.1088/1742-
368 6596/1566/1/012057.doi:10.1088/1742-6596/1566/1/012057
- 369 [15] N. Tadi Bani and S. Fekri-Ershad, "Content-based image retrieval based on
370 combination of texture and colour information extracted in spatial and frequency
371 domains," *Electronic Library*, vol. 37, no. 4, pp. 650–666, 2019, doi: 10.1108/EL-03-
372 2019-0067.doi:10.1108/EL-03-2019-0067
- 373 [16] A. Latif *et al.*, "Content-based image retrieval and feature extraction: A comprehensive
374 review," *Mathematical Problems in Engineering*, vol. 2019, 2019, doi:
375 10.1155/2019/9658350.doi:10.1155/2019/9658350
- 376 [17] C. Jia *et al.*, "Content-Aware Convolutional Neural Network for In-Loop Filtering in High
377 Efficiency Video Coding," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp.
378 3343–3356, 2019, doi: 10.1109/TIP.2019.2896489.doi:10.1109/TIP.2019.2896489
- 379 [18] V. Singrodia, A. Mitra, and S. Paul, "A Review on Web Scrapping and its Applications,"
380 *2019 International Conference on Computer Communication and Informatics, ICCCI*
381 *2019*, pp. 1–6, 2019, doi:
382 10.1109/ICCCI.2019.8821809.doi:10.1109/ICCCI.2019.8821809
- 383 [19] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and
384 application," *International Journal of Advances in Soft Computing and its Applications*,
385 vol. 13, no. 3, pp. 144–168, 2021, doi:
386 10.15849/ijasca.211128.11.doi:10.15849/ijasca.211128.11
- 387 [20] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, *Data preprocessing in*
388 *predictive data mining*, vol. 34, 2019. doi:
389 10.1017/S026988891800036X.doi:10.1017/S026988891800036X
- 390 [21] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on
391 machine learning," *Proceedings - International Conference on Data Engineering*, vol.
392 2019-April, pp. 2086–2090, 2019, doi:
393 10.1109/ICDE.2019.00245.doi:10.1109/ICDE.2019.00245
- 394 [22] M. Abdel-Basset, M. Mohamed, M. Elhoseny, L. H. Son, F. Chiclana, and A. E. N. H.
395 Zaied, "Cosine similarity measures of bipolar neutrosophic set for diagnosis of bipolar
396 disorder diseases," *Artificial Intelligence in Medicine*, vol. 101, p. 101735, 2019, doi:
397 10.1016/j.artmed.2019.101735.doi:10.1016/j.artmed.2019.101735
- 398 [23] T. Thongtan and T. Phientrakul, "Sentiment classification using document embeddings
399 trained with cosine similarity," *ACL 2019 - 57th Annual Meeting of the Association for*
400 *Computational Linguistics, Proceedings of the Student Research Workshop*, pp. 407–
401 414, 2019, doi: 10.18653/v1/p19-2057.doi:10.18653/v1/p19-2057
- 402 [24] V. W. Anelli, A. Bellogin, T. Di Noia, D. Jannach, and C. Pomo, "Top-N
403 Recommendation Algorithms: A Quest for the State-of-the-Art," *UMAP2022 -*
404 *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and*
405 *Personalization*, pp. 121–131, 2022, doi:

- 406 10.1145/3503252.3531292.doi:10.1145/3503252.3531292
407 [25] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto,
408 "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services
409 Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi:
410 10.1109/ACCESS.2020.2994222.doi:10.1109/ACCESS.2020.2994222
411 [26] Karnita, A. Kurniawan, and A. Suangga, "Analysis of Online BPHTB Application
412 Success System Using Information System Success Models DeLone and McLean
413 (Case Study of the Revenue Service, Financial Management, and Regional Assets of
414 Subang Regency)," *JPSAM (Journal of Public Sector Accounting and Management)*,
415 vol. 1, no. 1, pp. 55–69, 2019.
416
417
418