JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

A User Recommendation Model for Answering Questions on Brainly Platform

Puji Winar Cahyo[1*], Kartikadyota Kusumaningtyas[2], Ulfi Saidata Aesyi[3]
[1,2] Informatics Department, FTTI UNJANI, Yogyakarta, Indonesia
[3] Information System Department, FTTI UNJANI, Yogyakarta, Indonesia
[1,2,3]Jl. Siliwangi, Ringroad Barat, Banyuraden, Gamping, Sleman, DIY, 55293
*Corresponding Email: *pwcahyo@gmail.com

Received 5 October 2020, Revised 14 January 2021, Accepted 1 February 2021

Abstract — Brainly is a Community Question Answer (CQA) application that allows students or parents to ask questions about their homework. The current mechanism is that users ask questions, then other users in the same subject interest can see and answer them. As a reward for answering questions, Brainly gives points. The number of points varies by question. The greater of total points users have, Brainly will automatically display them in the smartest user leaderboard on the site's front page. However, sometimes, some users do not have good activity in answering questions. Thus, it is possible to have an urgent question that no one has answered. This study implements the Fuzzy C-Means cluster method to improve Brainly's feature regarding the speed and accuracy of answers. The idea is to create student clusters by utilizing the most brilliant students' leaderboard, subject interest, and answering activities. The stages applied in this research started with Data Extraction, Preprocessing, Cluster Process, and User Recommender. The optimal number of clusters in the answerer recommendation in the Brainly platform is 2 clusters. The value of the fuzzy partition coefficient for two clusters reached 0.97 for Mathematics and 0.93 for Indonesian. Meanwhile, the results of the recommendations were influenced by answer ratings. Many answer numbers are not given ratings because the possibility of the answers is not appropriate or the user's insensitivity in giving ratings.

Keywords – clustering, fuzzy, recommender system, educational platform, Brainly

Copyright © 2021 JURNAL INFOTEL
All rights reserved.

I. INTRODUCTION

Brainly is a Community Question Answer (CQA) application that allows students or parents to ask questions about their homework. Currently, Brainly is spread across ten countries using different languages, including Indonesian. In the Brainly platform, anyone can ask or answer questions, whether students, material experts, or professional educators proficient in their fields. Brainly motivates users to be actively involved in the community by implementing gamification.

The current mechanism is that users ask questions, then other users in the same subject interest can see and answer them. As a reward for answering questions, Brainly gives points. The number of points varies by the level of suitability of the answer. The greater the total points users have, Brainly will automatically display them in the smartest user leaderboard on the site's front page. However, sometimes, some users do not have good activity in answering questions. Thus, it is possible to have an urgent question that no one has answered.

Determining user expectations in asking questions is influenced by quick responses, additional information, alternative information, and accurate information [1]. Thus, it is essential to continue to improve the relevance and satisfaction of answer information to every user. [2]. In other research, the increasing recommendation for answering a question is influenced by the correlation between answerer and asker through topic preference [3]. Grouping of answerer can be done by detecting some inherent similarities in features using the cluster method [4].

Some algorithms are usually used in the cluster method—K-Means clustering using for grouping data

Instagram users. The research can generate some related hashtags, using a suggestion hashtag in branding and marketing through Instagram Platform [5]. Fuzzy C-Means clustering can use for recommendation systems. In the movie recommendation system, fuzzy clustering can enhance the stability and robustness of the clustering process [6].

This study implements the Fuzzy C-Means cluster method to group active users in Brainly by utilizing the most brilliant students' leaderboard, subjects' interest, and answering activities. Before applying the system model, the model is usually tested with a fuzzy partition coefficient [7]. The result of testing is used for validating the number of clusters in Brainly user data. The recommendation system will use the student clusters' results to answer the questions [8]. Thus, the questions will be faster and more precise if they are recommended to the right student cluster.

II. RESEARCH METHODS

This research aims to cluster user data on the Brainly educational platform based on the point level and activities users on the answer question—the data collected from every user profile. The stages applied in this research started with Data Extraction, Preprocessing, Cluster Process, and User Recommender. The whole stage of this research can be seen in Fig. 1.

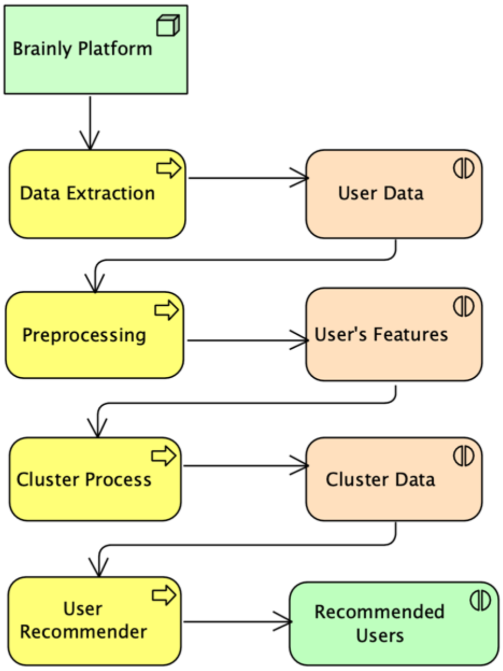
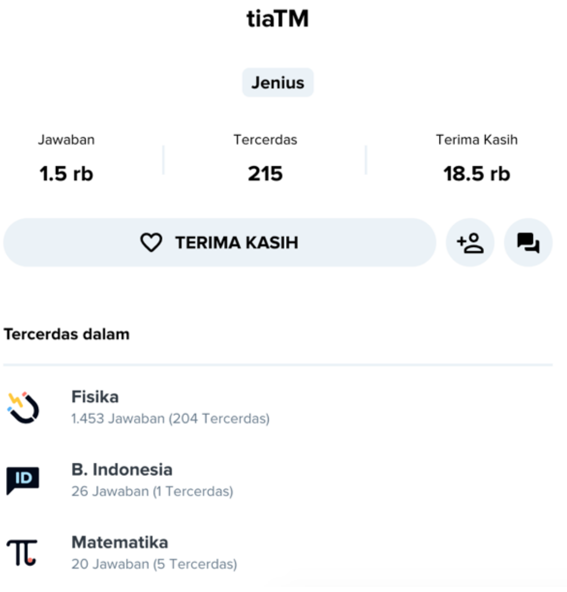


Fig. 1. Research Stages

A. Data Extraction

Data extraction is an essential part of getting valuable information [9]. The data extraction technique generally goes through unstructured, semi-structured, and structured data. A retrieve process is needed to obtain information from web pages [10]. For this reason, at this stage, we are trying to get user profile data for Brainly users taken from the Brainly.co.id web page, which can be seen in Fig. 2.



tiaTM

Jenius

Jawaban 1.5 rb | Tercerdas 215 | Terima Kasih 18.5 rb

TERIMA KASIH

Tercerdas dalam

Fisika
1.453 Jawaban (204 Tercerdas)

B. Indonesia
26 Jawaban (1 Tercerdas)

Matematika
20 Jawaban (5 Tercerdas)

Fig. 2. User profile

Table 1. Raw data

Data
https://Brainly.co.id/app/profile/3106136,MathTutor,"Brainly Bachelor,Jenius",pertanyaan 0,teman 61,jawaban 6.261,309,603.6 rb,"{'Matematika': '6.037 Jawaban (297 Tercerdas)', 'Fisika': '64 Jawaban (1 Tercerdas)', 'Ujian Nasional': '40 Jawaban (6 Tercerdas)', 'Biologi': '22 Jawaban (2 Tercerdas)', 'IPS': '18 Jawaban (1 Tercerdas)', 'B. Indonesia': '14 Jawaban', 'PPKn': '14 Jawaban', 'Sejarah': '13 Jawaban (1 Tercerdas)', 'SBMPTN': '12 Jawaban', 'Kimia': '7 Jawaban (1 Tercerdas)'}","{'Poin:': '97.576', 'Tingkat:': 'Sekolah Menengah Atas', 'Bergabung:': '25 Mei 2016'},{'terima_kasih': '658.522'}"
https://Brainly.co.id/app/profile/97314,jihannk99,Si Hebat,pertanyaan 288,teman 48,jawaban 870,108,8.7 rb,"{'B. Indonesia': '229 Jawaban (24 Tercerdas)', 'Matematika': '214 Jawaban (21 Tercerdas)', 'B. Daerah': '80 Jawaban (12 Tercerdas)', 'B. inggris': '74 Jawaban (5 Tercerdas)', 'Biologi': '43 Jawaban (8 Tercerdas)', 'IPS': '40 Jawaban (4 Tercerdas)', 'Fisika': '36 Jawaban (4 Tercerdas)', 'PPKn': '32 Jawaban (8 Tercerdas)', 'Seni': '26 Jawaban (6 Tercerdas)', 'Bahasa lain': '25 Jawaban (8 Tercerdas)'}","{'Poin:': '10.794', 'Tingkat:': 'Sekolah Menengah Atas', 'Bergabung:': '16 April 2014'},{'terima_kasih': '9.532'}"
https://Brainly.co.id/app/profile/60607,DiahYusi,Si Hebat,pertanyaan 55,teman 231,jawaban 1.030,113,14.6 rb,"{'Matematika': '865 Jawaban (83 Tercerdas)', 'Fisika': '79 Jawaban (15 Tercerdas)', 'TI': '17 Jawaban (1 Tercerdas)', 'B. inggris': '13 Jawaban (7 Tercerdas)', 'IPS': '11 Jawaban (1 Tercerdas)', 'B. Indonesia': '10 Jawaban (1 Tercerdas)', 'Seni': '7 Jawaban', 'B. jepang': '6 Jawaban (1 Tercerdas)', 'Sejarah': '5 Jawaban', 'Biologi': '5 Jawaban (3 Tercerdas)'}","{'Poin:': '164.503', 'Tingkat:': 'Sekolah Menengah Atas', 'Bergabung:': '9 Maret 2014'},{'terima_kasih': '14.743'}""

Table 2. Clean Data

User	Total Jawaban Matematika	Total jawaban Tercerdas Matematika	..
MathTutor	6037	297	..
jihannk99	214	21	..
DiahYusi	865	83	

B. *Preprocessing*

Preprocessing is when the data is prepared before being processed, the data must be clean and appropriate [11]. In this research, Data profile retrieved using data extraction from the Brainly website. The data profile has the text form with complete answers, total high rated answer, thanks note, etc. The features needed the total number of answers for each subject, the number of the brightest answers resulting from the asker's, appreciation, and users' thanks. Table 1 is an example of raw data taken from profile data on a Brainly account. Each account has attributes attached. From the raw data, it is processed according to research needs. The results of this processing produce clean data shown in Table 2.

C. *Cluster Process*

The clustering process is where clean data that has formed features are processed using the cluster method to produce a homogeneous data group for each cluster [12]. The clustering algorithm used in this study is Fuzzy C-Means, where the fundamental nature of this algorithm changes the discrete value from {0.1} to a constant value [0.1] [13]. The following is a step by the Fuzzy C-Means algorithm from [14] in [15].

1. Determine the number of clusters c $(2 \le c \le n)$ and the matrix value m' for the initial stage $U^{(0)}$, for each step labeled r where r = 0,1,2,..
2. Start calculating the cluster center $\{V_i{}^r\}$ for each step, whereby using (1)

$$v_{ij} = \frac{\sum_{k=1}^{n} \mu_{ik}^{m'} .\ X_{ki}}{\sum_{k=1}^{n} \mu_{ik}^{m'}} \qquad (1)$$

for *i* is the center of the cluster in the *i* feature, and *j* is the *j* feature

3. Update the matrix partition for step *r*, $U^{(r)}$ following (2)

$$\mu_{ik}^{(r+1)} = \left[\sum_{j=1}^{c}\left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}}\right)^{2/(m'-1)}\right]^{-1} ;\ I_k = \emptyset \qquad (2.a)$$

or $$\mu_{ik}^{(r+1)} = 0 \text{ for } i \in I_k \qquad (2.b)$$

where $$I_k = \{i|2 \le c \le n; d_{ik}^{(r)} = 0\} \qquad (3)$$

and $$\underset{\sim k}{I} = \{1,2,\dots,c\} - I_k \qquad (4)$$

and $$\sum_{i\in I_k} \mu_{ik}^{(r+1)} = 1 \qquad (5)$$

4. If $\left\|\underset{\sim}{U}^{(r+1)} - \underset{\sim}{U}^{(r)}\right\| \le \varepsilon_L$ then stop, if not, then *r = r + 1* and return to the calculation stage 2.

Step 4, comparing two fuzzy partition matrices consecutively to reach a level of accuracy ε_L (good error rate). In step 3, several equations are shown, including (2) to (5). For (2.a) run if the matrix partition set is empty or still in the phase r = 0, unless the result d_{jk} is 0 then (2.b) will run in anticipation of (3) and (4) by setting the partition membership value to 0 for each class. Whereas (5) to ensure that the sum of all columns on the fuzzy partition is not 0, symbol $\underset{\sim}{U}$ is the total number of partitions. Equation (6) below is used to find the valued_{ik}.

$$d_{ik} = d(x_k - v_i)\ ;$$

$$d(x_k - v_i) = \left[\sum_{j=1}^{m}(x_{kj} - v_{ij})^2\right]^{1/2} \qquad (6)$$

Determining the optimal number of clusters in this study accommodates the Fuzzy Partition Coefficient (FPC). To calculate FPC can use (7).

$$FPC = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{nc} u_{ij}^2 \qquad (7)$$

PC values are based on the range [1/nc, 1] where nc is the number of clusters. u_{ij} is a fuzzy membership partition on the cluster. At the same time, *N* is the number of clustered data sets.

D. *User Recommender*

Recommender Systems is a process where data and information will be reviewed based on the appropriate order for each type of recommendation so that the results can be used for decision making [16]. User Recommender is part of Recommender Systems, which is used to determine the appropriate user groups. The user group is generated from the cluster stage using Fuzzy C-Means. The appropriate user selected to help answer the question (according to the level and type of question to be answered).

III. RESULTS

From the stages and methods that have been determined, the process's application at each stage. The web data extraction process resulted in 1225 total users of the Brainly education platform in Indonesia. The user data is complete with profiles and the most brilliant total answers in answering questions.

At the clustering process, we use three features, including Feature 1, the number of answers to a subject (this feature is used to determine how often users answer questions on that subject). Feature 2 is the brightest number of answers to a lesson (the brightest number feature is the number of answers with the highest rating). Feature 3 is the number of thanks given by other users. All features can be seen in Table 3.

Table 3. Data Features

User	Feature 1	Feature 2	Feature 3
MathTutor	6037	297	658522
jihannk99	214	21	9532
DiahYusi	865	83	14743

After obtaining three main features, the process towards the clustering stage can be carried out using the Fuzzy C-Means method. Before completing the clustering stage, finding the optimal number of clusters suitable for 1225 data with the three features mentioned earlier. The search for the optimal number of clusters uses the Fuzzy Partition Coefficient calculation by conducting 9 test trials for Math and Indonesian Subjects, starting from 2 to 10 clusters. From the test results, graphs and test data are obtained as in Fig. 3.

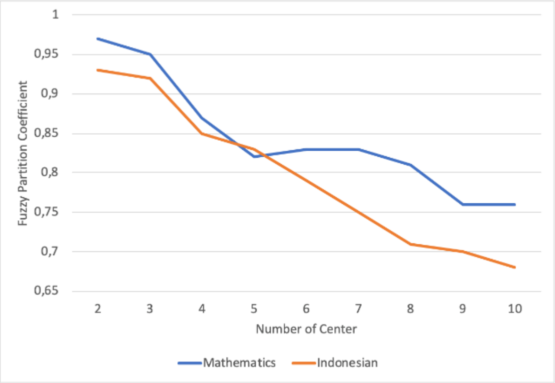


Fig. 3. Fuzzy Partition Coefficient

As seen in Fig. 3, testing the optimal number of clusters is tested nine times with cluster variations ranging from 2, 3, 4, 5, 6, 7, 8, 9, 10. From nine tests, the optimal number of clusters is 2 clusters for Math and Indonesian.

A. *Clustering for Mathematics subjects*

A clustering process is carried out with two the number of clusters for Math. The first feature is the number of answers in mathematics. The second feature is the total of the smartest answers in mathematics. The third feature is the number of thanks by other users. The resulting graph is shown in Fig. 4.

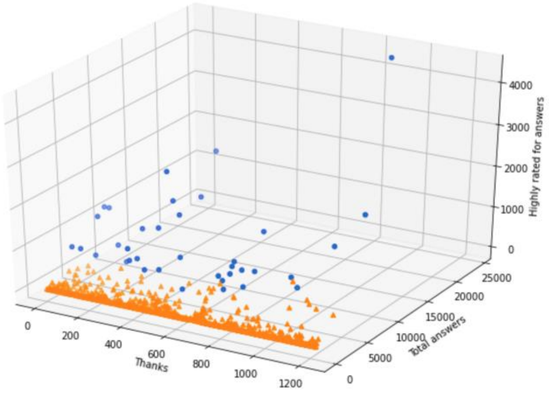


Fig. 4. The result of math cluster

From Fig. 4, it can be seen that the orange color is the first cluster. It is not a user recommendation with the best answer with 1188 data. Meanwhile, the blue color is the second cluster. It is a user's recommendation with the best answer with 37 user data.

In Fig. 5, the total data for mathematics answers are above 10000 answers, the number of highly-rated math answers is reached 4320, and thanks reaches 1664. Different from cluster 2, The data of cluster 2 is not an answerer recommendation for mathematics, as shown in Fig. 6.

In Fig. 6, it can be seen that the total mathematics answers reach 3399, the number of highly-rated math answers does not reach 600, while the thanks are not up to 1000. The two clusters' results show that the number of answer features in subjects has a sufficient value to affect cluster data grouping.

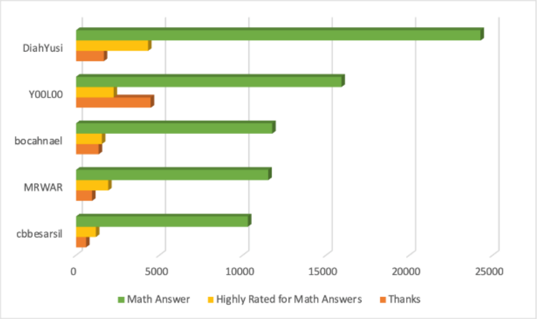


Fig. 5. Answerer recommendations for Math

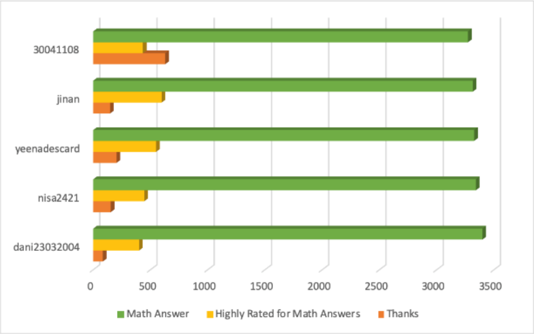


Fig. 6. Not answerer recommendations for Math

B. Clustering for Indonesian subjects

In Indonesian, 2 clusters are divided, with the first feature the number of Indonesian answers. The second feature is the number of answers with a high rating, and the third feature is the number of thanks by other users. The number of clusters and features that have been determined produces a cluster graphic, according to Fig. 7.

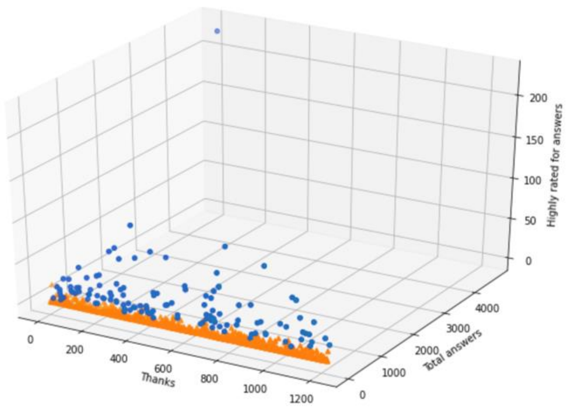


Fig. 7. The result of Indonesian cluster

Fig. 7 shows the first cluster in orange, which is not recommended for answer assistance with total 1100 users. The second cluster in blue is the user recommendation cluster for answer assistance with 125 users. As for the top users in the cluster, the recommendation for answer assistance can be seen in Fig. 8.

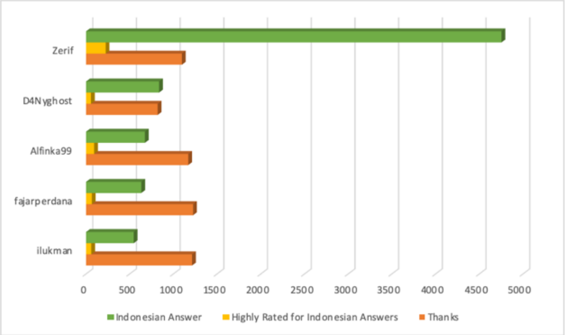


Fig. 8. Answerer recommendations for Indonesian

In Fig. 8, users with the name Zerif have the top answerer in Indonesian subjects with a total answer exceeding 4500 and the total thank you exceeding 1000. Unlike other users, the total answers are not more than 1000. The answers with the highest rating in the first cluster reach 224, and the lowest reaches 56. At the same time, users who are not recommended for answer assistance can be seen in Fig. 9.

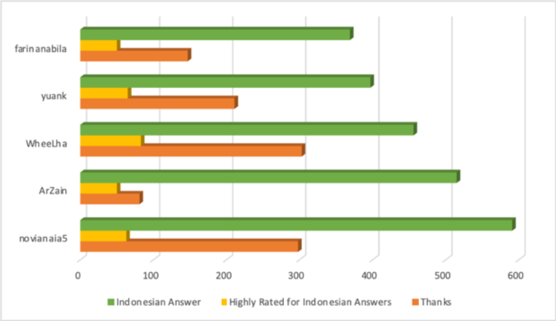


Fig. 9. Not answerer recommendations for Indonesian

Figure 9 shows that users who are not recommended answerers have total answering do not reach 600 in Indonesian subjects. Meanwhile, thanks only reached 300, and answers with high ratings did not reach 100 answers.

IV. DISCUSSION

User data clustering on the Brainly education platform has been carried out by showing that the optimal number of clusters in Math and Indonesian is 2 clusters. These 2 clusters were obtained from testing 9 cluster variations, starting from 2, 3, 4, 5, 6, 7, 8, 9, and 10. From nine tests of 9 cluster variations, two optimal clusters were obtained. More increases in the number of clusters will be less optimal.

The answering cluster in Math recommends 37 answering users, with a profile of the highest rating answers in Math is reaches 4320 from exceeding 20000 answers. Meanwhile, the non-answerer cluster is reached 1188 answerers, with the profile is given the highest rating answers is not more than 600 the number of answers from 3399 total answers.

Clustering of the answerer in the Indonesian subject provided 125 answerers from 1225 users of the Brainly education platform. The recommended answerer cluster has the highest total number of answers reaching more than 4500 answers, and the lowest is still above 500. For the highest rating, user answers are reached 224 out of 4500 answers. In contrast, not a recommended answering cluster had the highest rating of less than 100 from the total Indonesian answers.

V. CONCLUSION

The optimal number of clusters in the answerer recommendation in the Brainly platform is 2 clusters. The value of the fuzzy partition coefficient for two clusters reached 0.97 for Mathematics and 0.93 for

Indonesian. These results are obtained from optimal testing using the Fuzzy Partition Coefficient on two clusters. Meanwhile, the results of the recommendations were influenced by answers rating. Possibility because the answers are not appropriate or the user's insensitivity in giving ratings.

ACKNOWLEDGMENT

The research funding comes from PKRA Faculty of Engineering and Information Technology Universitas Jenderal Achmad Yani, Yogyakarta. The data comes from research collaboration of Pusat Studi dan Layanan Analitik Data (PusLAD) Universitas Jenderal Achmad Yani, Yogyakarta with Brainly.

REFERENCES

[1] C. Xu, W. Xin, and Y. Guo, “Collaborative Expert Recommendation for Community-Based Question Answering,” in *Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 378–393.

[2] J.-J. Huang, “Simple Additive Weighting Method,” *Mult. Attrib. Decis. Mak.*, pp. 55–67, 2011.

[3] C. Fu, “User correlation model for question recommendation in community question answering,” *Appl. Intell.*, vol. 50, no. 2, pp. 634–645, 2020.

[4] N. John, Blooma Mohan; Chua, Alton Y.K.; Goh, Dion Hoe Lian; and Wickramasinghe, “Graph-based Cluster Analysis to Identify Similar Questions: A Design Science Approach,” *J. Assoc. Inf. Syst.*, vol. 17, no. 9, pp. 590–613, 2016.

[5] M. Habibi and P. W. Cahyo, “Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 4, pp. 399–408, 2019.

[6] S. V Vimala and K. Vivekanandan, “A Kullback–Leibler divergence-based fuzzy C-means clustering for enhancing the potential of an movie recommendation system,” *SN Appl. Sci.*, vol. 1, no. 7, p. 698, 2019.

[7] K. V. Rajkumar, A. Yesubabu, and K. Subrahmanyam, “Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset,” *Int. J. Adv. Technol. Eng. Explor.*, vol. 9, no. 4, pp. 2760–2770, 2019.

[8] F. Eustáquio and T. Nogueira, “On Monotonic Tendency of Some Fuzzy Cluster Validity Indices for High-Dimensional Data,” in *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, 2018, pp. 558–563.

[9] W. Nadee and K. Prutsachainimmit, “Towards data extraction of dynamic content from JavaScript Web applications,” in *2018 International Conference on Information Networking (ICOIN)*, 2018, pp. 750–754.

[10] M. A. Bin Mohd Azir and K. B. Ahmad, “Wrapper approaches for web data extraction : A review,” in *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, 2017, pp. 1–6.

[11] P. W. Cahyo and E. Winarko, “Model Monitoring Sebaran Penyakit Demam Berdarah di Indonesia Berdasarkan Analisis Pesan Twitter,” Universitas Gadjah Mada Yogyakarta, 2017.

[12] P. W. Cahyo, “Klasterisasi Tipe Pembelajar Sebagai Parameter Evaluasi Kualitas Pendidikan Di Perguruan Tinggi,” *Teknomatika*, vol. 11, no. 1, pp. 49–55, 2018.

[13] R. Winkler, F. Klawonn, and R. Kruse, “Fuzzy C-Means in High Dimensional Spaces Fuzzy c-means in high dimensional spaces,” *Int. J. Fuzzy Syst. Appl.*, vol. 11, no. 1, 2010.

[14] J. C. Bezdek, “FCM : THE FUZZY c-MEANS CLUSTERING ALGORITHM,” vol. 10, no. 2, pp. 191–203, 1984.

[15] P. W. Cahyo and M. Habibi, “Clustering followers of influencers accounts based on likes and comments on Instagram Platform,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 14, no. 2, pp. 199–208, 2020.

[16] M. Quadrana, P. Cremonesi, and D. Jannach, “Sequence-Aware Recommender Systems,” *ACM Comput. Surv.*, vol. 51, no. 4, Jul. 2018.