TELKOMNIKA, Vol.17, No.3, June 2019, pp.1425~1431
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i3.12436

K-Nearest neighbor algorithm on implicit feedback to determine SOP

Muhammad Yusril Helmi Setyawan*, Rolly Maulana Awangga, Nadia Ayu Lestari
Applied Bachelor Program in Informatics Engineering, Politeknik Pos Indonesia
Sariasih St. No.54 Bandung Indonesia, telp/fax: 022-2009562/022-2009568
*Corresponding author, e-mail: yusrilhelmi@poltekpos.ac.id

Abstract

The availability of a lot of existing Standard Operating Procedures (SOP) document information, users often need time to find SOPs that fit their preference. Therefore, this requires a recommendation system based on user content consumption by personalized usage logs to support the establishment of SOP documents managed according to user preferences. The k-nearest neighbor (KNN) algorithm is used to identify the most relevant SOP document for the user by utilizing implicit feedback based on extraction data by monitoring the document search behavior. From the research results obtained 5 classifications as parameters, with a final value of 3:2 ratio that shows the best distance value with the majority of labels according to the concept of calculation KNN algorithm that sees from the nearest neighbor in the dataset. This shows the precision of applying the KNN algorithm in determining SOP documents according to user preferences based on implicit feedback resulting in 80% presentation for SOPs corresponding to profiles and 20% for SOPs that do not fit the user profile. To establish SOP documents to show more accurate results, it should be used in a broad SOP management system and utilize implicit feedback with parameters not only in search logs and more on performance evaluation evaluations.

Keywords: *implicit, K-nearest neighbors, preferences, recommendations, standard operating procedures*

1. Introduction

Standard Operating Procedures (SOP) is a documented process, which describes the step by step activities of the organization's need to run the business process [1]. SOP's operating procedures are vital for the company to maximize the effectiveness and efficiency of the company by its overall operational philosophy and policy [2] so that each implementer can easily understand and follow the company's ISO standards [3]. Moreover, in Regulation of the Minister of PER/21/M-PAN/11/2008 stated that SOP making must meet the principles between as is the ease and clarity, efficiency and effectiveness. If the organization of SOP does not match the preferences of each employee in each field in the library, then the SOP implementation is inefficient [4], because the speed of data processing is required to do the work [5]. Also, technology has become the company's primary requirement [6] in carrying out its work [7]. Involving the use of technology requires a transition system [8] example, a recommendation system that performs the process as needed. Moreover, it uses it to recommend objects according to preferences. Taking into account such scenarios designs an automated recommendation system [9], which will define SOP documents based on personalization of usage history [10]. The use of implicit feedback can be used as a landing to realize service improvement by applying information technology to monitor performance [11], especially for SOP implementers.

Business processes that previously used conventional methods began to shift to web-based or mobile devices [12]. So it is necessary to have a proposal to model a new system which concerns how to automatically push information according to user preferences. In search of SOP documents, whereas most existing recommendation systems use explicit feedback data [13], while in determining SOPs which corresponds to user preferences using explicit feedback, not in the appropriate context if an imbalance occurs, it can affect performance degradation. The framework for user SOP determination considers many attributes such as the interaction between users and content [14, 15] because the adding and modifying various and different relationship to make recommendations based on extracted data, by monitoring the accumulation of log data as implicit feedback such as user behavior [16]. Moreover, SOP

Received November 15, 2019; Revised January 30, 2019; Accepted February 28, 2019

classification to facilitate users in setting SOPs by their preferences, through data classification and into a list of suggestions or predictions in query completion on search [17]. So this research uses domain usage log data based on the SOP document search according to user preferences in the SOP system application.

Usage log data based on SOP document searches performed by employees has usually assumed in implicit feedback that the more employees look for the SOP document, closer to the employee preference. By applying the k-nearest neighbor algorithm to establish the operational standard of the procedure by worker activity with feedback can implicitly identify the most relevant SOPs for users. Bayesian algorithms based on reservation logs show significant results on the addition of context information with their implementation in restaurant ratings [18] and ranking e-commerce based on timestamp [19]. Additionally, monitoring user click behavior demonstrates the effectiveness of approaches that reflect user preferences which further indicate the potential for widespread application of applying to e-commerce purchase predictions [20]. Then by utilizing user interest for recommendations, such as applying a user location vector to its implementation to determine a tourist destination produces a 65-100% effectiveness of the prediction system as well as rank [21].

K-Nearest Neighbor (KNN) is an efficient algorithm for predicting implementations to identify bioluminescent proteins [22]. Then KNN's implementation shows that this method is used to identify relevant news content recommendations based on each user's implicit feedback. The use of the KNN algorithm to establish the freshness classification of fish based on the color of the fish to determine fish consumption is appropriate produces classification with 91.36% accuracy [23]. The application of KNN shows a high enough accuracy to predict the load the next day by merely utilizing the estimated temperature by monitoring the document search behavior. The use of KNN yields an accuracy of 94.95%, and the modified k-nearest neighbor produces 99.51% accuracy in classifying data [24] instruments [25]. Based on previous research, this research will determine the standard operating procedure (SOP) based on implicit feedback by utilizing user preferences so that it will determine the recommendation of an SOP according to consumption of SOP content on the system based on data extraction result and monitoring of user search behavior. In research by applying the KNN algorithm can identify the most relevant SOP for the user.

2. Methods

2.1. Implicit Feedback

Implicit Feedback is the feedback issued by the system automatically and by the wishes of the user. This feedback contains only positive values from one class. For example, like products that users, clicks, and bookmarks can be a single-class positive value. Feedback as well as automatic purchases and login access by the system, and those that are more easy to collect, and otherwise provided by the user to the item intentionally [26, 27]. Implicit feedback data is information that can use from users without user feedback, which can not only be used between users and content but also other available information [28]. The recommendation system with feedback requires settings to bring feedback to the user preference level.

2.2. K-Nearest Neighbor

The KNN algorithm is one of the methods to perform a classification analysis, but in the last few decades, the use of the KNN method is also for prediction [28]. KNN is a learning algorithm a typical KNN method classifies the sample request with the most voting strategy. For each query sample, KNN finds its nearest neighbor in the dataset and then assigns it to a class that is mostly owned by its neighbors. KNN is an algorithm that classifies objects based on similarities of data with other data [29]. The KNN is an approach to finding the case by calculating the proximity between the new case and the old case based on matching the weight of some existing features. Figure 1 is the steps for calculating the KNN algorithm. Here are the steps for calculating the KNN algorithm:

a. Weigthing with determining the characteristics in the dataset is the sum of the implicit feedback from the user personalization.
b. Calculate the number of implicit feedback within 5 weeks with the provisions if the appropriate profile then multiplied by the percentage of 80% if not the profile match multiplied by 20% percentage.

c. Calculate the KNN Algorithm with determining the parameter value k, by looking at the amount of data in the table, then taking the average to determine the value of k.
d. Calculate the distance of each training data to the data label, as described in (1):

$$dist(x_{1i,}\, x_{2i}) = \sqrt{\sum_{i=1}^{n}} (x_{1i} - x_{2i})^2 \quad (1)$$

Where, x_1 and x_2 show two samples, x_{1i} and x_{2i} are their variable values.
e. Determine the data label that has the minimum distance.
f. Line the training data into the data label in step 3.
g. Repeat steps 2 through 4 until the number of each class is k.

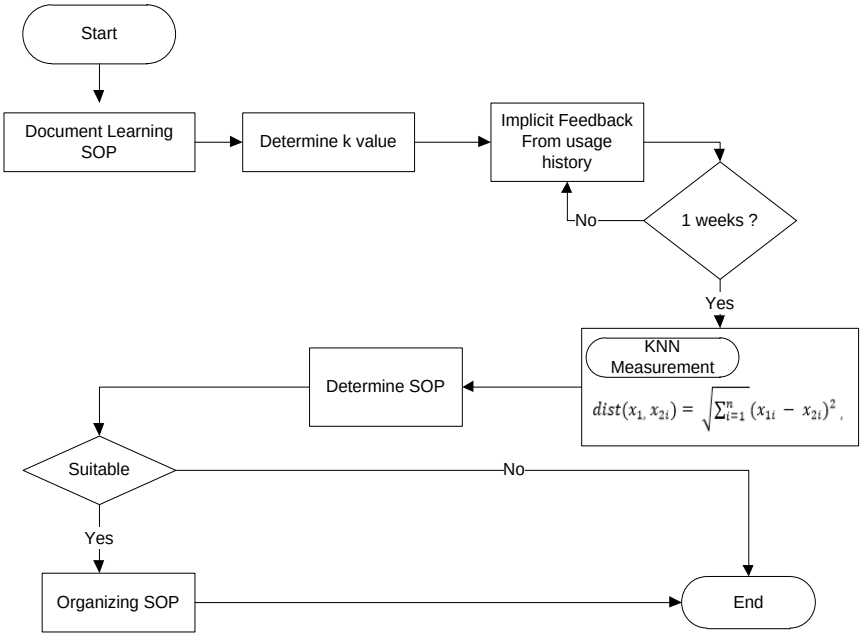

Figure 1. K-Nearest neighbor algorithm

With the test document, the system finds the k-nearest neighbor in the classified training location and obtains the category of test document according to the class distribution of these neighbors. Which can be used to measure similarities between neighbors and the test documents for weighting in getting a better classification effect [30]. In this research based on implicit feedback data from usage log that is search SOP document. First calculate the amount of implicit feedback each user has on the SOP, then measure the distance between the user and the SOP document to determine the relevant SOP. In general, the process flow of the application of the KNN method to determine the relevant SOP plan in Figure 2.

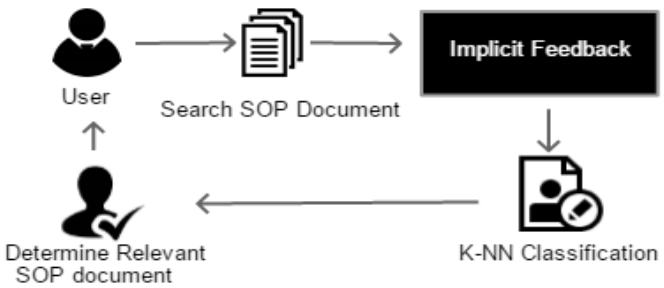

Figure 2. K-Nearest neighbor process

3. Result and Analysis

The calculation of K-NN done by searching the k group of objects in the closest learning data (similar) to the object on the new data or test data, this process is to calculate the similarity between documents calculated based on implicit feedback data in the form of log usage is the standard document search procedure. This study uses SOP document from the Library of Politeknik Pos Indonesia which amounts to 11 SOP documents; SOP is a document containing guidelines or set of rules that are made to facilitate the task of employees to work by their primary tasks and functions in their respective institutions. This study aims to establish the SOP document that best suits the user preferences through user usage history.

3.1. Input Identification

The data used in this study amounted to 11 SOP documents. For input identification process, 10 existing SOP documents will go through the preprocessing stage to calculate the rating or value based on the implicit feedback. Provided if accessed by the profile then multiplied by the percentage of 80% and if not by the profile of the meal will be multiplied by a percentage of 20% this is done for minimizing the possibility of user access that does not match the preferences view from the profile. However, it also not a rule out the possibility of the SOP being recommended to be incompatible with its profile because it relies on the user's preference from its implicit feedback. According to the source of the number of SOP access within 5 weeks at least 5 times access, then to determine the dataset seen from the value of the paid difference of 5 times access that is 3. Next to determine the appropriate label and not fit on the view dataset of the total amount of implicit feedback if the number more or equal to 2.4 then the label given is appropriate but if below 2.4 then the label given is not appropriate, the value 2.4 obtained from the multiplication of 3 and 80%, so 2.4 used as a benchmark to determine the label on the dataset for SOP in accordance with the profile users. The existing SOP documents represent as documents with SOP38, SOP39, ... SOPn and document queries will also be considerate as comparable documents with SOPq. Table 1 is the preprocessing phase by determining the value of each SOP to 1 user by observing the history of using SOP for 5 weeks.

Table 1. The Dataset Base on the Rating of the Implicit Feedback

No	SOP ID	1 Week	2 Week	3 Week	4 Week	5 Week	Explanation
1	*SOP*48	0.8	0.8	0	0	0	Unsuitable
2	*SOP*39	0	0.2	0.2	0	0.2	Suitable
3	*SOP*40	0.4	0.2	0	0.4	0.2	Suitable
4	*SOP*41	0.2	0.2	0	0.2	0	Suitable
5	*SOP*38	0.2	0.2	0	0	0	UnSuitable
6	*SOP*43	0.2	0	0	0	0	Unsuitable
7	*SOP*44	0.8	0	0	0	0	Unsuitable
8	*SOP*45	0	1.6	0.8	0.8	0.8	Suitable
9	*SOP*46	0.8	0	0.8	1.6	0	Suitable
10	*SOP*47	0	0.8	0.8	0.8	1.6	Suitable
11	*SOP*42	0.2	0.4	0	0.8	0.2	?

3.2. Process Identification

After preprocessing done, next is to determine the classification parameter, by looking at the number of datasets that exist then calculate the amount of difference, then from the above dataset obtained k=5. After determining the parameter value, then the Euclidean distance calculation as described from (2) to (6) of each test document against document Q.

$$dist(x_{1,}\, x_{2i}) = \sqrt{\sum_{i=1}^{n}}(x_{1i} - x_{2i})^2 \quad (2)$$

$$dist = \sqrt{(0.2 - 0.8)^2 + (0.4 - 0.8)^2 + (0 - 0)^2 + (0 - 0)^2 + (0.2 - 0)^2} \quad (3)$$

$$dist = \sqrt{(0.36)^2 + (0.16)^2 + (0)^2 + (0)^2 + (0.04)^2} \quad (4)$$

$$dist = \sqrt{0.56} \quad (5)$$

$$dist = 0.75 \quad (6)$$

Furthermore, after calculating the distance between each Euclidean object and document Q, then the ranking for each value on the document, so that obtained Euclidean distance and the overall ranking of documents in Table 2.

3.3. Identification of Output

In the identification output is to classify the document based on the ranking with the provisions of parameter 5, this provision is taken based because of datasets used is 10 so the parameter value is taken from half the number of datasets as the difference in comparison. The majority category results where the suitable values are shown by documents with ID D2, D4, and D6, while for un-suitable values are shown in documents with ID D5 and D3, obtaining a ratio of 2:3, for more suitable values. Using the majority category, then based on the calculation of KNN obtained the classification results of the 5 parameters of the data. That document Q is a document corresponding to the user in the field of service by looking at the value of comparison between the 5 parameters with the best distance value with the majority indicating the appropriate label.

Figure 3 shows the process to be built on the system to determine the standard document application procedure. Administrator will input the SOP document data; then the SOP document will be accessed by the user by monitoring the user's search behavior as implicit feedback stored in the search log. Next, calculate by using the KNN algorithm to determine the suitability of SOP documents with user preferences. The next user will get recommendations for their preferences.

Table 2. Euclidean Distance and Ranking

No	Name	Distance	Ranking
1	*D*11, *D*1	0.75	4
2	*D*11, *D*2	0.35	8
3	*D*11, *D*3	0.49	6
4	*D*11, *D*4	0.35	8
5	*D*11, *D*5	0.28	10
6	*D*11, *D*6	0.45	7
7	*D*11, *D*7	0.75	4
8	*D*11, *D*8	1.77	3
9	*D*11, *D*9	1.94	1
10	*D*11, *D*10	1.85	2

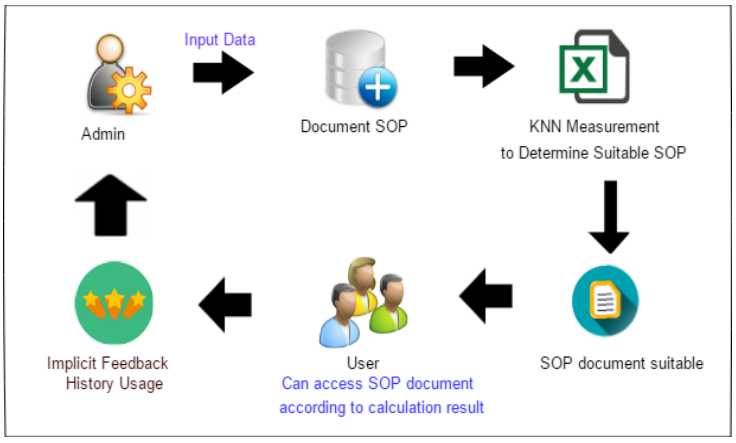

Figure 3. System illustration

3.4. Result

Experiments using the KNN method with k=5 count similarities between the test data and the center of each label data. This test determines the prosecution of standard operational documents by using feedback from user log-out with greater provision if the selected document is a document that corresponds to the field located on the user profile. Based on the test results, obtained 5 SOP documents with 2:3 presentation, where the ratings of 1.4, and 5 indicate the label according to the rank 2 and 3 label is not appropriate.

Consider the fact that document Q is a document that matches the user by looking at a comparison between the parameters with the relevant value with the KNN method to retrieve from the nearest neighbor that must be obtained, so the following is graph 5 which shows the results. Based on Figure 4, the presentation of the ranking of the SOP document to be known or the test document by looking at the classification parameters of the 5. Documents obtained a presentation of each of the rank 1 until 5 as in the graph shows the majority of labels appear for test data that is appropriate for the service.

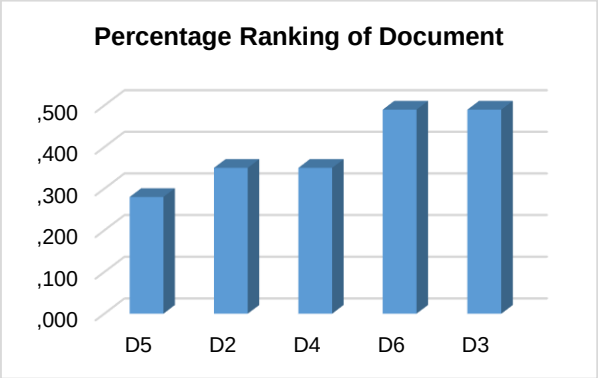

Figure 4. Percentage ranking of document

4. Conclusion

The present study presents an approach to determine the SOP documentation using the KNN algorithm for document classification by utilizing implicit feedback from the search log data with the percentage of documents corresponding to the 80% and unsuitable profiles of 20%. KNN is one of the most popular classifier, easy to use and efficient enough. By finding the value of K representing the number of neighbors as the class for the classification. Previous research using the KNN algorithm to perform web text analysis can automatically improve accuracy automatically while in this study there is not many classifications of documents but also predictions or user rates.

5. Discussion

From the research, however, in the establishment of SOP documents are still in the case study of SOP libraries by SOP determination within the scope of the division, and only utilize implicit feedback in the form of search logs. To establish SOP documents to show more accurate results it should be used in a broad SOP management system and utilize implicit feedback with parameters not only in the search log but maybe user behavior, usage logs and more and discuss performance evaluation. So it is expected that the determination in SOP document retrieval can show more accurate results.

References

[1] Kourdali HK, Sherry L. A comparison of two takeoffs and climb out flap retraction standard operating procedures. *Integrated Communications Navigation and Surveillance (ICNS).* 2016: 4A2–1.
[2] *Simulation of time-on-procedure (ToP) for evaluating airline procedures*. 2017 Integrated Communications, Navigation and Surveillance Conference (ICNS). 2017. [Online]. Available: https://doi.org/10.1109/icnsurv.2017.8011892
[3] N Wangskarn, J Siritantitam, N Meesri, P Chiravirakul. *Flowty-flow: A web application for preparation and distribution of standard operating procedures*. Student Project Conference (ICT-ISPC), 2016 Fifth ICT International. 2016: 21–24.
[4] Z Jing, L Boang, Y Hao. Fault diagnosis strategy for startup process based on standard operating procedures. *2013 25th Chinese Control and Decision Conference (CCDC)*, Guiyang. 2013: 4221-4226.
[5] Simulation of time-on-procedure (ToP) for evaluating airline procedures. in *2017 Integrated Communications, Navigation and Surveillance Conference (ICNS)*. IEEE. 2017. [Online]. Available: https://doi.org/10.1109/icnsurv.2017.8011892.
[6] S Hasegawa. *How ICT changes quality assuarance in graduate education and research? From knowledge transfe]ll improvement 2015.* International Conference on Information & Communication Technology and Systems (ICTS), Surabaya. 2015; 1-2.
[7] The Republic of Indonesia Government, Minister of Administrative and Bureaucratic Reform, Guidelines for the preparation of operational standards for government administrative procedures. 2012: 63.
[8] The Republic of Indonesia Government. Regulation of Minister of Administrative and Bureaucratic Reform Number 15 concerning Guidelines for Service Standards. 2014: 14.

[9] MY Helmi Setyawan, RM Awangga, SR Efendi. *Comparison of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot*. 2018 International Conference on Applied Engineering (ICAE). Batam, 2018: 1-5.

[10] NH Harani, AA Arman, RM Awangga. Improving togaf adm 9.1 migration planning phase by ITIL v3 service transition. *Journal of Physics: Conference Series*. 2018; 1007(1): 012036.

[11] SW Byun, SM Lee, SP Lee, KYKim, C Kee-Seong. *A recommendation system based on the object of the interest.* Advanced Communication Technology (ICACT), 2016 18th International Conference. 2016: 689–691.

[12] P Patil, S Gore. Study of recommendation system for yoga and raga for personalized health based on the constitution (Prakriti). *International Journal of Computer Applications*; 2016; 136(4): 25–27.

[13] Y Zhang, C Yang, Z Niu. *A research of job recommendation system based on collaborative filtering.* Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium. 2014: 533–538.

[14] K Lin, Y Chen, X Li, Q Wu, Z Xu. *Friend recommendation algorithm based on location-based social networks*. Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference. 2016: 233–236.

[15] J Wang, L Yu, H Zhang. *Brand recommendation leveraging heterogeneous implicit feedbacks*. Computer Science and Engineering (APWC on CSE), 2015 2nd Asia-Pacific World Congress. 2015: 1–6.

[16] IN Yulita, S Purwani, R Rosadi, RM Awangga. A quantization of deep belief networks for long short-term memory in sleep stage detection. in *Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017 International Conference on* IEEE, 2017: 1–5.

[17] D Ajantha, J Vijay, R Sridhar. *A user-location vector based approach for personalized tourism and travel recommendation*. Big Data Analytics and Computational Intelligence (ICBDAC), 2017 International Conference. 2017: 440–446.

[18] RM Awangga, M Yusril, H Setyawan. Ontology design of influential people identification using centrality. *Journal of Physics: Conference Series*. 2018; 1007(1): 012012.

[19] S Gupta, U Ojha, VS Dixit. *Personalized web recommendations analyzing sequential behavior using implicit data streams: A survey*. 2017 8th International Conference on Computing, Communication, and Networking Technologies (ICCCNT). 2017: 1–7.

[20] P Mathew, B Kuriakose, V Hegde. *Book recommendation system through content-based and collaborative filtering method*. Data Mining and Advanced Computing (SAPI- ENCE), International Conference. 2016: 47–52.

[21] G Wu, V Swaminathan, S Mitra, R Kumar. *Digital content recommendation system using implicit feedback data*. 2017 IEEE International Conference on Big Data (Big Data). 2017: 2766–2771.

[22] R Jia, R Li, M Yu, S Wang. *E-commerce purchase prediction approach by user behavior data.* Computer, Information and Telecommunication Systems (CITS), 2017 International Conference. 2017: 1–5.

[23] W-T Kuo, Y-C. Wang, RT-H. Tsai, JY-j. Hsu. *Contextual restaurant recommendation utilizing implicit feedback*. Wireless and Optical Communication Conference (WOCC), 2015 24th. 2015: 170–174.

[24] C Chen, D Wang, Y Ding. *User actions and timestamp based personalized recommendation for e-commerce system.* Computer and Information Technology (CIT), 2016 IEEE International Conference. 2016: 183–189.

[25] J Hu. Blknn: A *k-nearest neighbors method for predicting bioluminescent proteins*. Computational Intelligence in Bioinformatics and Computational Biology 2014 IEEE conference. 2014: 1–6.

[26] H Al-Shehri, A Al-Qarni, L Al-Saati, A Batoaq, H Badukhen, S Alrashed, J Alhiyafi, SO Olatunji. *Student performance prediction using support vector machine and k-nearest neighbor*. Electrical and Computer Engineering (CCECE), 2017 IEEE 30th Canadian Conference. 2017: 1–4.

[27] NMS Iswari, Wella, Ranny. *Fish freshness classification method based on fish image using k-nearest neighbor*. 2017 4th International Conference on New Media Studies (CONMEDIA). 2017: 87–91.

[28] R Zhang, Y Xu, Z Y Dong, W Kong, KP Wong. *A composite k-nearest neighbor model for day-ahead load forecasting with limited temperature forecasts*. *Power and Energy Society General Meeting (PESGM).* 2016: 1–5.

[29] I Gazalba, Mustakim, NGI Reza. *Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification.* 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). 2017: 294–298.

[30] C-Y Lin, L-C Wang, K.-H. Tsai. Hybrid real-time matrix factorization for implicit feedback recommendation systems. *IEEE Access*; 6: 21 369–21 380, 2018.