Serendipity Identification Using Distance-Based Approach

Widhi Hartanto[1], Noor Akhmad Setiawan[2], Teguh Bharata Adji[3*]

***Abstract*—The recommendation system is a method for helping consumers to find products that fit their preferences. However, recommendations that are merely based on user preference are no longer satisfactory. Consumers expect recommendations that are novel, unexpected, and relevant. It requires the development of a serendipity recommendation system that matches the serendipity data character. However, there are still debates among researchers about the available common definition of serendipity. Therefore, our study proposes a work to identify serendipity data's character by directly using serendipity data ground truth from the famous Movielens dataset. The serendipity data identification is based on a distance-based approach using collaborative filtering and k-means clustering algorithms. Collaborative filtering is used to calculate the similarity value between data, while k-means is used to cluster the collaborative filtering data. The resulting clusters are used to determine the position of the serendipity cluster. The result of this study shows that the average distance between the recommended movie cluster and the serendipity movie cluster is 0.85 units, which is neither the closest cluster nor the farthest cluster from the recommended movie cluster.**

***Keyword*—Serendipity, Collaborative Filtering, K-Means.**

I. Introduction

The development of information technology increases the number of goods offered online. It is indicated by the increase in the number of sellers on the Tokopedia shopping site in 2018 that reached 4 million [1], while Bukalapak targeted to get a total number of sellers of 5 million by the end of 2018 [2]. The increase of the sellers is very beneficial for consumers because it will provide many product' choices. However, the more products offered by the marketplace, the more difficult for consumers to find products that suit their preferences. It has a negative impact on the online stores' performance since it will increase the time for consumers to select goods and reduce the online store's speed to acquire profit. This problem leads to the development of a method that can help consumers find products that suit their preferences. This method is known as a recommendation system which, with certain considerations, can provide advice to consumers about items that suit the consumer preference [3].

The recommendation system is generally considered to be successful if the recommended product gets a positive response from consumers. This positive response shows a positive comment on the recommended item or a high rating on the movie that has been seen.

Along with the increase in the number of products, the development of the recommendation system has shifted. The recommended item that is generated merely based on the consumer preference is no longer satisfactory. It is due to the recommendation system producing monotonous products which are similar to the products that have been purchased. This condition is known as the overspecialized problem. It has a negative outcome on consumer purchasing behavior. The consumer tends not to use the recommendation as purchasing reference.

The existence of overspecialized problem challenges the development of a recommendation system. Consumers expect alternative products that are different from the previous purchase character but still relevant to their preferences. The recommendation with this character is called serendipity recommendation. The serendipity recommendation will increase consumer loyalty to online stores as a recommendation provider [4]. Along with the increasing number of products, online consumers tend to choose products with serendipity criteria [5]. A recommendation system dealing with this serendipity criteria is known as a serendipity recommendation system.

Most serendipity recommendation system uses a popular basic algorithm for a recommendation system, i.e., collaborative filtering. This algorithm uses the principle of similarity between users in giving a rating to an item or the similarity in the number of ratings received by an item. Collaborative Filtering algorithms are often combined with machine learning algorithms to produce more accurate recommendations. One of the learning algorithms is the KNN algorithm. The KNN algorithm is an algorithm that uses the principle of the distance between data to determine recommendations. KNN algorithm produces recommendation items with high similarity values with items that match a certain user character based on the closest distance to the target item. The use of the KNN algorithm in determining recommendations can provide relatively fast and accurate recommendation results. However, the use of the KNN is not suitable for generating serendipity recommendations. Since the principle used in the KNN algorithm is intended to identify the data closest to the target data, this concept is not suitable with the novelty aspect in serendipity.

KFN algorithm is another machine learning algorithm that is used in recommendation system research. This algorithm determines the farthest data from the reference data. This algorithm produces recommendations that are very different from previous consumer preferences. This algorithm is intended to increase the diversity of the recommendations [6]. However, this principle is not appropriate when it is used in developing a serendipity recommendation system. It is because the serendipity criteria still consider the relevance aspect of the previous preference meaning that an object with a serendipity

[1,2,3] *Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada, Jln. Grafika No.2, Kampus UGM, Yogyakarta, 55281, INDONESIA (Tel. +62-274-552305, email: [1]widhihartanto@mail.ugm.ac.id, [2]noorwewe@ugm.ac.id, [3]adji@ugm.ac.id),*

[*] *corresponding author.*

character is not an object different from the previous preference.

Moreover, a problem arises because of no common agreement among researchers on the serendipity data character. It is due to the lack of available ground truth serendipity dataset only until 2018, when Movielens ground truth serendipity dataset became public. This dataset is the result of Movielens' research using real user feedback.

However, it is required to find the characteristic of serendipity points. Thus, in this paper, the researchers propose an approach for identifying the serendipity points using Movielens ground truth serendipity dataset and employ a distance-based approach. The distance-based approach will then find that the serendipity points are not very close to the reference point, nor is it the farthest distance from the reference point, which becomes the contribution of this work.

However, it is necessary to discover the characteristic of serendipity points. Thus, in this paper, the researchers propose an approach to identify the serendipity points using the Movielens ground truth serendipity dataset and employ a distance-based approach. The distance-based approach will then find that the serendipity points are neither too close to the reference point nor at the furthest distance from the reference point, which becomes the contribution of this work.

The algorithm that can be used to identify the relative position of data is the clustering algorithm. There are three subcategories of clustering algorithms:

1) Hierarchical Clustering (e.g., Divisive, Agglomerative): Hierarchical clustering is a clustering algorithm that produces clusters by grouping data based on a hierarchical relationship between existing data. The condition that must be met to use this algorithm is a hierarchical relationship between existing data.

2) Density-based clustering (e.g., DBSCAN): Density-based clustering is a clustering algorithm that produces clusters based on data density. This algorithm is very efficient to use for grouping data with an arbitration shape and finding outliers.

3) Partition-based clustering (e.g., k-median, k-means): Partition-based clustering is a clustering algorithm that produces clusters by dividing data based on the number of clusters that have been predetermined. This algorithm is very suitable for exploring data and then find new insights.

In this study, the most suitable clustering algorithm is partition-based clustering, i.e., k-means. The consideration of using k-means is because the data in the serendipity dataset has an equivalent level, and serendipity density is not required in finding the serendipity characteristic.

The k-means clustering algorithm will divide the data into groups based on the distance between the data and the centroids. This algorithm is suitable for identifying serendipity points since it will produce clusters containing data that have relatively high similarities with clear boundaries. The distance between centroids can be used as a consideration in determining the tendency of the serendipity points.

This paper consists of the following sections: Part II presents research related to developing a serendipity recommendation system. Part III describes the methodology in identifying serendipity. Part IV is the results and discussion. Part V is the conclusion.

II. Basic Theory

A. Collaborative Filtering in Recommender System

Collaborative Filtering algorithm uses ratings from consumers to provide the best-recommended item. For example, if two consumers have similar behavior in giving ratings on the same goods, the Collaborative Filtering algorithm will provide recommendations for the item highly rated by user 1 to user 2, and vice versa. Two algorithms are used in Collaborative Filtering, i.e., model-based collaborative filtering and memory-based collaborative filtering:

1) Model-Based Collaborative Filtering: In providing a recommendation, this algorithm uses a combination of data mining and a machine-learning algorithm to produce a predictive recommendation model. Some examples of model-based collaborative filtering algorithms are rule-based, Bayesian algorithm, latent factor, and decision tree models. A Latent factor model is one of the algorithms with a high level of coverage even for sparse rating matrices.

2) Memory-Based Collaborative Filtering: This filtering is also known as neighborhood-based collaborative filtering. This algorithm is the earliest algorithm to develop a recommendation system. It has two advantages, namely ease of implementation and ease of recommendation interpretation. This algorithm uses similarity values between products based on consumer rating records. This approach produces recommendations to target consumers based on the similarity in rating items or based on the similarity of objects that have been selected. The calculation of the similarity value between objects uses Pearson Correlation as in (1) [3]:

$$\mathrm{sim}(a,b) = \frac{\sum_{u\in U}\left(R(u,a)-\overline{R(a)}\right).\left(R(u,b)-\overline{R(b)}\right)}{\sqrt{\sum_{u\in U}\left(R(u,a)-\overline{R(a)}\right)^2}.\sqrt{\sum_{u\in U}\left(R(u,b)-\overline{R(b)}\right)^2}}. \qquad (1)$$

Referring to (1), a and b are the movies in which similarity levels, u is the user in U, and R is the rating value.

The principle of the Memory-Based Collaborative Filtering algorithm is suitable in identifying serendipity points because it can show the level of similarity between data based on predetermined features.

B. Serendipity Concept

Serendipity means surprisingly experiencing an unexpected event, and it turns out to be lucky. Researchers are then explaining different serendipity terminologies, which yet to be formalized.

A surprise is identified as a core component of serendipity. An item's quality is different from the user's profile; it shows the surprise quality level [7]. Another idea proposes three criteria in determining serendipity, namely novelty (novelty), unexpected (never thought of before), and relevance (having relevance) [3]. Following the shopping context, serendipity is defined as discovering unexpected goods and providing

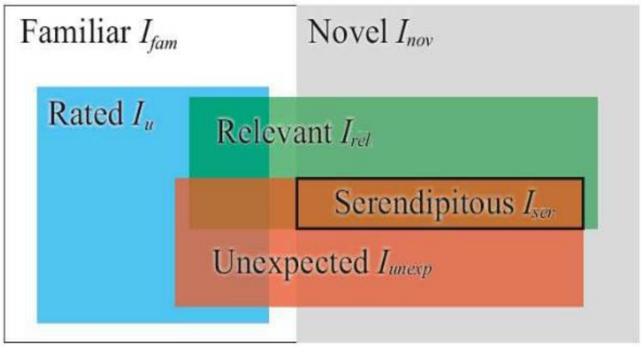

Fig. 1 Euler's diagram of an object from the point of consumer's view [8].

valuable information [5]. A concept of serendipity is also introduced as a concept consisting of three components, i.e., relevance, novelty, and unexpectedness [8], [9]. The three components are explained as follows.

1) Relevance: The level of suitability of an item to what is liked, consumed, and attractive to consumers.

2) Novelty: The novelty level of an object which has never been heard or seen by consumers.

3) Unexpectedness: The level of difference an object offered suits the consumer's profile; for example, a jazz song offering to classical music fans has a high unexpectedness rate compared to a classic song offering.

A study related to serendipity in movie selection found that all the variations in unexpectedness and novelty components expand consumer preferences in determining movie choices [10]. The visualization of the relationship between components in serendipity in the form of a Euler diagram is shown in Fig. 1.

Research in serendipity recommendation is performed online, offline, and the combination between online and offline methods. The online method is carried out by surveying consumers to obtain ground truth data. Meanwhile, the offline method is carried out using a mathematical approach to develop a serendipity model with certain assumptions to determine the character of serendipity.

Most serendipity recommendation research is conducted using the offline method. This is because the offline method requires fewer research resources. However, due to the assumption differences among researchers, the definitions of serendipity are still varied. Therefore, until now, there is still no common definition of serendipity.

An offline method, such as conducted in [11], produces a serendipity recommendation system algorithm called Serendipity Oriented Greedy (SOG) algorithm. This algorithm is based on the principle of topic diversification that aims to increase the resulting recommendation's diversity value. This research obtains a better result in finding serendipity items and obtaining a better value of recommendation diversity. Offline research was conducted by modifying the Collaborative Filtering algorithm using the Usage Context-Based Collaborative Filtering (UC-BCF) approach to increase customer satisfaction and provide positive surprises [12]. The experiment results show that this approach improves the accuracy of ranking predictions for niche objects that have a small amount of rating data and increase the diversity recommendations. In [13], an online experiment was conducted to get the serendipity values of tv programs. This study formulates a serendipity model using a hierarchical clustering algorithm to determine TV programs with potential serendipity values. The research used a survey of actual TV viewers. The experiment shows that the number of serendipity TV programs is 7-8% of the total existing TV programs. Online serendipity research in collaboration with the Movielens group is conducted using a survey of customers with certain criteria to answer several questions related to serendipity. The results of this study are a movie serendipity dataset. This dataset is provided to evaluate the serendipity algorithm produced from offline research.

C. K-Means Clustering Algorithm

Clustering is an unsupervised learning algorithm that groups data based on their distance. In the development of recommendation systems, clustering algorithms generally aim to solve the scalability problem. Clustering can be used to reduce the size of the data that must be processed [14]. By reducing the size of the data, the computational speed will be improved. In [15], the clustering algorithm was combined with a Collaborative Filtering algorithm to improve the recommendation's quality.

Apart from addressing data size problems, clustering algorithms can be used to identify the user character in order to deal with the dynamic changes of consumer preferences [16].

In our study, the clustering algorithm will be used to identify the ground truth serendipity data's character based on the cluster distances. K-means clustering algorithm, the most widely used clustering algorithm, is employed because k-means divides the data optimally into several, which is done by calculating the average distance between the data and the centroid points.

To identify serendipity, k-means is used to cluster similarity values generated from the collaborative filtering calculation. K-means is calculated using (2). J is the objective function, k is the number of clusters, n is the number of objects, x_i is the i^{th}, and c_j is the j^{th} centroid.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left| x_i^{(j)} - c_j \right|^2 \qquad (2)$$

III. Method

D. Proposed Method (Serendipity Identification Method)

In this study, the identification of serendipity was carried out by analyzing the serendipity character on the ground truth serendipity data provided by Movielens. The algorithm for identifying serendipity character was the Collaborative Filtering algorithm and k-means Clustering algorithm, as shown in Collaborative Filtering block and k-means block in Fig. 2, respectively. A collaborative filtering algorithm was used to calculate the similarity values among movies. While the similarity values among movies were calculated based on ratings and tag relevance values, the result of this similarity calculation was processed using the k-means clustering algorithm.

The consideration in determining the value of *k* was using the silhouette method. This method aims to determine the level of consistency of membership values in each cluster based on variations in the value of *k*. The higher the silhouette value indicates that the cluster members have consistent values or have a high degree of similarity between data.

The result of the clustering process was clusters that contain movies with high similarity values. In addition, the result of this clustering showed the clusters containing the recommended movies and the cluster containing the ground truth serendipity movies. The serendipity character is shown by the distance between the recommended movie cluster centroids and the serendipity cluster centroids. The farther the distance between the recommended movie cluster centroids and the serendipity cluster centroids, the lower the similarity and the more serendipity movie is different from the previous user preference.

This study aims to explore the serendipity character using the ground truth dataset. The study evaluation was done by calculating the distance between the serendipity movie cluster and the recommended movie cluster (see the bottom block of Fig. 2). The flowchart diagram to find the serendipity character is shown in Fig. 2.

E. Experimental Setup

This study used the serendipity 2018 dataset obtained from Movielens. This dataset was the result of serendipity research using real user feedback [10]. The dataset was ground truth serendipity data used as an evaluation material for research in serendipity recommendation systems. The 182 MB dataset contained user answers to the serendipity level, movie titles, and recommended movies by Movielens to a certain user, tag-relevance data, movie label data, and movie rating data. Overall, there were ten million rating data, consisting of 49,151 movie title data and 104,661 user data.

The dataset preparation was intended to obtain the serendipity data test. The serendipity data test consisted of serendipity user data, serendipity movie data, and movies recommended by Movielens to the serendipity users. The serendipity users were users with serendipity answers. The movies recommended by Movielens to the serendipity users were assumed to be the most suitable movies according to the previous user preferences.

This dataset preparation produced 1,000 data test records. From this number of data, 200 sample data were taken randomly using a systematic random sampling method. The method was applied using interval sampling of 100 and a sample size of 20. The sample size was 1/5 of the original data and was therefore sufficient to represent the entire data. Another consideration of using this sample was the computation time, which was approximately 10 minutes for processing 1 data test. These data tests were used as references in calculating the similarity values with the other movies using (1).

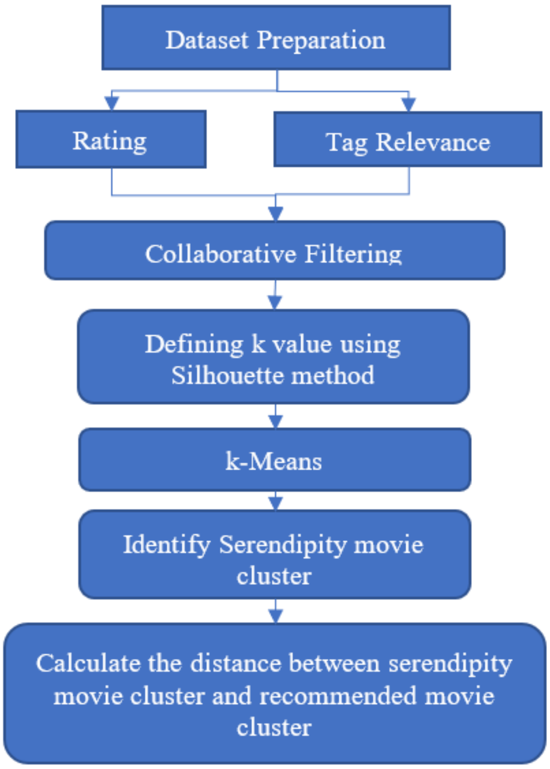
Dataset Preparation
Rating
Tag Relevance
Collaborative Filtering
Defining k value using Silhouette method
k-Means
Identify Serendipity movie cluster
Calculate the distance between serendipity movie cluster and recommended movie cluster

Fig. 2 Serendipity identification method flowchart.

The result of collaborative filtering was processed with k-means clustering. The determination of the value of *k* was done by using the silhouette method with the variation of the *k* value is 2 - 15.

The recommended movie cluster and the serendipity movie cluster were then identified uses the clustering result. The distance between the recommended movie cluster and the serendipity movie cluster was calculated using the Pythagorean theorem.

The tools used in this study were a PC with a 7th generation Intel Core i5 processor, DDR4 memory of 24 GB, 1 terabyte hard disk, and Python 3.7 software.

IV. Result and Discussion

The similarity value calculation results using the Collaborative Filtering method obtained two similarity values, i.e., similarity based on rating and similarity based on the relevance tag. These results were then processed using the silhouette method to obtain a recommendation for the value of *k*. The average of silhouette value from $k = 2$ to $k = 15$ was 0.36, with a variation of 0.03. This silhouette value began to stabilize at the value of $k = 10$. This silhouette value shows that the resulting variation in silhouette values is very thin, which means that the variation in the value of k does not have much effect on changes in silhouette values. Therefore, in this study, the value of $k = 10$ was chosen, which was the lower limit of the silhouette value.

The example of the clustering result from the collaborative filtering data using $k = 10$ is shown in Fig. 3. The data for this example was the user with id 106850, to which a movie with id

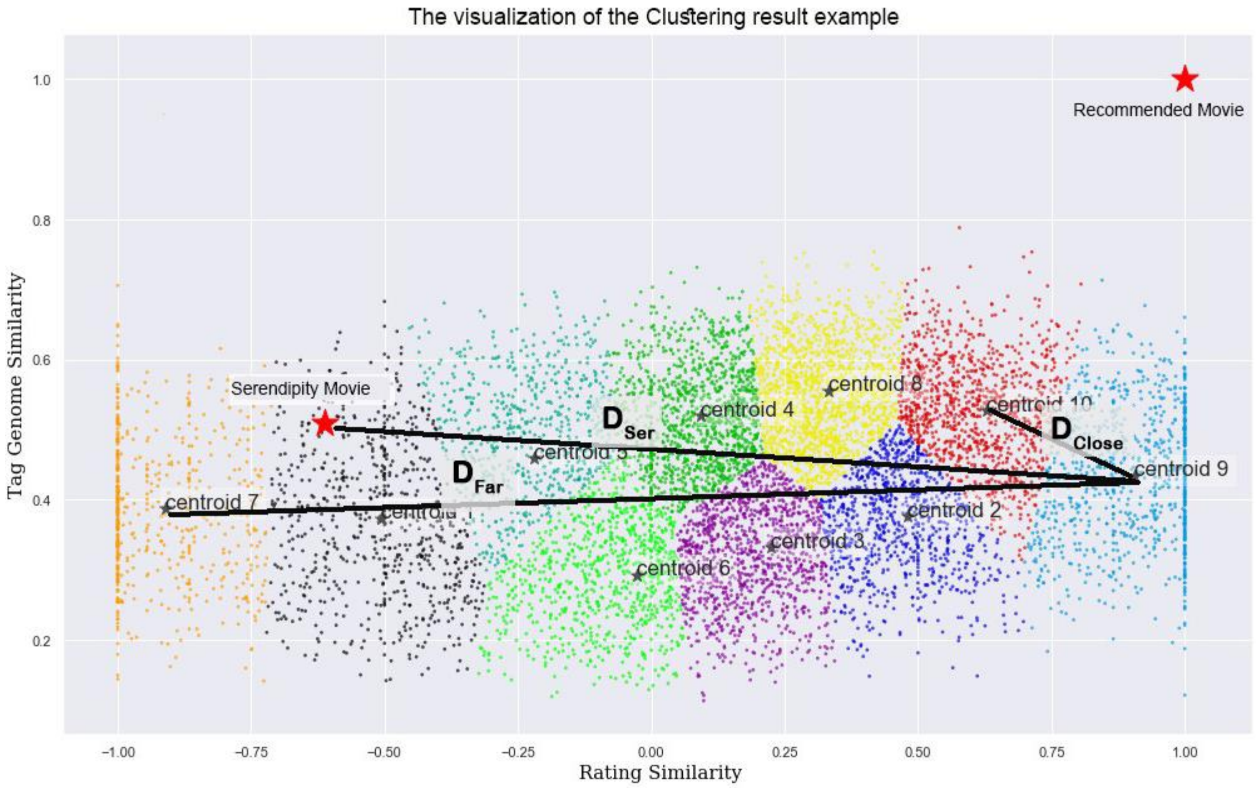

Fig. 3 The visualization of the clustering result example.

111235 was recommended and for which another movie with id 1897 was the serendipity item. It can be seen that the nearest cluster and the farthest cluster to the recommended movie are cluster 10 (with a distance D_{Close} of 0.37 units) and cluster 7 (with a distance D_{Far} of 1.9 units), respectively. Meanwhile, the ground truth serendipity movie is in cluster 1, and the recommended movie is in cluster 9, both of which have a distance (D_{Ser}) of 1.5 units. Hence, the cluster containing the serendipity movie is quite far from the cluster containing the recommended movie. However, the serendipity movie is not in the farthest cluster. This course of steps is repeated for all 200-sample data.

Meanwhile, the average distance between the nearest cluster ($\overline{D_{Close}}$) and the farthest cluster ($\overline{D_{Far}}$) to the recommended movie cluster is 0.27 units and 1.87 units, respectively. Meanwhile, the average distance between the centroid of the cluster containing ground truth serendipity movie and the centroid of the cluster containing the recommended movie ($\overline{D_{Ser}}$) is 0.85 units.

The result of this study showed that ($\overline{D_{Far}}$) > ($\overline{D_{Ser}}$) > ($\overline{D_{Close}}$), which concludes into two consequences. The first consequence shows that the serendipity movie is not much different from the previous user's preferences according to the relevance serendipity's relevance character and consequence suggests that the serendipity movie is also not very similar to the previous user's preference, corresponding to the novelty character of the serendipity.

V. Conclusion

This study explores the character of serendipity in the Movielens ground truth serendipity dataset. This exploration was carried out using a Collaborative Filtering algorithm to calculate the levels of similarity between items. The clustering process was then performed to obtain the recommendation cluster. The result is the closest cluster to the recommendation cluster, the farthest cluster to the recommendation cluster, and the serendipity cluster. The distance between the clusters shows the characteristics of serendipity.

This study indicates that the serendipity movie cluster is neither located in the closest cluster nor located in the farthest cluster from the recommendation movie cluster. This result shows that the character of Movielens ground truth data set matches with two serendipity main characters, i.e., relevance and novelty.

The use of k = 10 on k-means has shown the Movielens ground truth dataset's serendipity character. The use of other variations in the value of k will be carried out in future studies to see the serendipity character's consistency. Hence, the other future work is to find an accurate method to produce serendipity recommendations.

Acknowledgment

This research was supported by The Ministry of Communication and Information Technology master's degree scholarship program.

REFERENCES

[1] (2018) "Setiap Bulan Tokopedia Catat 300 Juta Kunjungan pada Situsnya" [Online], https://industri.kontan.co.id/news/setiap-bulan-tokopedia-catat-300-juta-kunjungan-pada-situsnya, access date: 7-Oct-2020.

[2] (2018) "Bukalapak Targetkan 5 Juta Pelapak Hingga Akhir 2018." [Online], https://ekonomi.kompas.com/read/2018/07/27/083000526/bukalapak- targetkan-5-juta-pelapak-hingga-akhir-2018, access date: 23-Aug-2020.

[3] C.C. Aggarwal, *Recommender System*, Cham, Switzerland: Springer International Publishing, 2016.

[4] D. Kotkov, "Serendipity in Recommender Systems," Dissertation, University of Jyväskylä, Jyvaskyla, Finland, 2018.

[5] C. Grange, I. Benbasat, and A. Burton-Jones, "With a Little Help from My Friends: Cultivating Serendipity in Online Shopping Environments," *Inf. Manag.*, Vol. 56, No. 2, pp. 225-235, 2018.

[6] A. Said and B. Fields, "User-Centric Evaluation of a K-Furthest Neighbor Collaborative Filtering Recommender Algorithm," *Proc. of the 2013 Conference on Computer Supported Cooperative Work*, 2013, pp.1399-1408.

[7] M. Kaminskas, D. Bridge, and I. Centre, "Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems," *Trans. Manag. Inf. Syst.*, Vol. 7, No. 1, pp. 1–42, 2016.

[8] D. Kotkov, S. Wang, and J. Veijalainen, "A Survey of Serendipity in Recommender Systems," *Knowledge-Based Syst.*, Vol. 111, pp. 180–192, 2016.

[9] N. Izyan, Y. Saat, S. Azman, M. Noah, and M. Mohd, "Towards Serendipity for Content-Based Recommender Systems," *Int. J. on Adv. Sci., Eng. and Inf. Technol.*, Vol. 8, No. 4, pp. 1762–1769, 2018.

[10] D. Kotkov, J.A. Konstan, Q. Zhao, and J. Veijalainen, "Investigating Serendipity in Recommender Systems Based on Real User Feedback," *Proc. 33rd Annual ACM Symposium Applied Computing - SAC '18*, 2018, pp. 1341–1350.

[11] D. Kotkov, J. Veijalainen, and S. Wang, "A Serendipity-Oriented Greedy Algorithm for Recommendations," *Proc. 13th International Conference on Web Information Systems and Technologies (WEBIST 2017)*, 2017, pp. 32–40.

[12] K. Niemann and M. Wolpers, "A New Collaborative Filtering Approach for Increasing the Aggregate Diversity of Recommender Systems," *Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*, 2013, pp. 955-963.

[13] T. Akiyama, K. Obara, and M. Tanizaki, "Proposal and Evaluation of Serendipitous Recommendation Method Using General Unexpectedness," *Proc. of Workshop on the Practical Use of Recommender Systems, Algorithms and Technologies (PRSAT 2010)*, 2010, pp. 3–10.

[14] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering," *Proc. Fifth International Conference on Computer and Information Technology*, 2002, pp. 1-6.

[15] U. Kuzelewska, "Collaborative Filtering Recommender Systems Based on K-Means Multi-clustering," in *Contemporary Complex Systems and Their Dependability. DepCoS-RELCOMEX 2018. Advances in Intelligent Systems and Computing*, Vol. 761, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, J. Kacprzyk, Eds., Cham, Switzerland: Springer, 2019, pp. 316-325.

[16] Y.S. Sneha and G. Mahadevan, "A Study on Clustering Techniques in Recommender Systems," *International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2011)*, 2011, pp. 97-100.