PENGARUH USER PROFILING PADA REKOMENDASI SISTEM MENGGUNAKAN K MEANS DAN KNN

Hartatik [1] **, Rosyid** [2]

[1] *Manajemen Informatika Universitas Amikom Yogyakarta*
[2] *Teknik Informatika Universitas Amikom Yogyakarta*
email : hartatik@amikom.ac.id[1], *rosyid1267@gmail.com*[2]

Abstraksi

Sparsity adalah satu masalah yang sering terjadi pada teknik collaborative clustering dimana user sedikit sekali memiliki informasi (pada penelitian ini rating) yang menyebabkan sistem seringkali tidak akurat ketika memberikan rekomendasi. Banyak metode yang bisa digunakan untuk menyelesaikan masalah sparsity data, salah satunya adalah metode KNN. Namun metode KNN memiliki kelemahan yaitu scalability. Scalability terjadi ketika ketika data yang harus dicari kesamaannya semakin besar. Salah satu solusi yang mungkin diimplementasikan adalah dengan mencari profil dari user dan mengelompokkannya menjadi satu kelompok. Eksperimen yang dilakukan pada penelitian ini untuk mengatasi masalah sparsity dan scalability adalah dengan menggabungkan algoritma silhouette, k-means, K-Nearest Neighbour. Dataset yang dipakai di penelitian ini, berjumlah 700 rating yang di crawling melalui web traveloka. Data rating antara user dan item akan disimpan dalam database, untuk selanjutnya dirubah menjadi bentuk array user-item. Hasil pengujian dengan 5 data uji didapatkan nilai rata-rata RMSE 1,33% dengan rata-rata akurasi = 100% - 1,33% = 98,67 %.

***Kata Kunci* :**
Sparsity, Skalabilitas, Silhouette, k means, KNN

Abstract

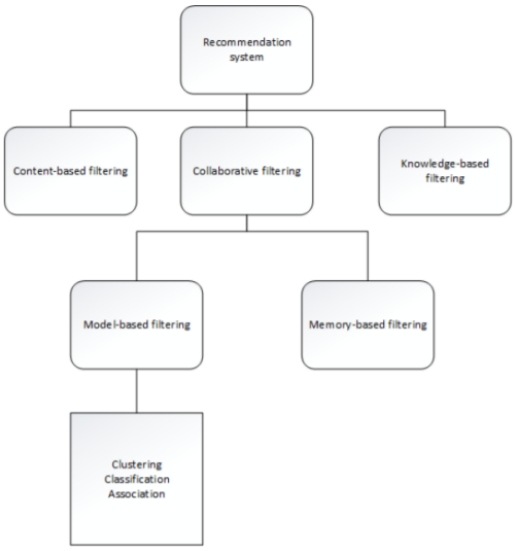
Sparsity is a problem that often occurs in collaborative clustering techniques where users have very little information (in this study rating) which causes the system is often inaccurate when making recommendations. Many methods can be used to solve data sparsity problems, one of which is the KNN method. But the KNN method has a weakness that is scalability. Scalability occurs when the data to be searched is large. One possible solution is to search for profiles of users and group them into one group. Experiments carried out in this study to overcome the problem of sparsity and scalability are by combining the silhouette method, k-means, and K-Nearest Neighbor algorithm. The dataset used in this study amounted to 700 ratings crawled through traveloka's web. Rating data between users and items will be stored in a database, then it will be converted into a user-item array. The test results with 5 test data obtained an average value of 1.33% RMSE with an average accuracy = 100% - 1.33% = 98.67%.

***Keywords* :**
Sparsity, Scalability, Silhouette, k means, KNN

Pendahuluan

Sistem rekomendasi merupakan salah satu riset yang cukup populer dalam dunia bisnis. Sistem rekomendasi terbangun atas beberapa algoritma atau metode yang mencoba untuk menemukan suatu pola dari data dan memberikan suatu rekomendasi dengan melakukan *filtering* berdasarkan keterhubungan pada minat atau kebutuhan [1]. Contoh paling mudah adalah pada toko online amazon yang menerapkan sistem rekomendasi untuk menawarkan buku-buku yang sejenis sesuai dengan histori pencarian *customer* pada saat mengunjungi toko online tersebut.

Banyak teknik yang bisa digunakan untuk membuat sistem rekomendasi seperti *content based filtering*, *collaborative filtering* dan *knowledge based filtering* seperti yang digambarkan di gambar 1. Ada juga teknik lainnya yang menggabungkan beberapa teknik rekomendasi sehingga menghasilkan teknik yang *hybrid*. Salah satu komponen utama *content based filtering* adalah proses pemodelan pengguna, di mana minat pengguna disimpulkan dari item yang berinteraksi dengan pengguna. Item biasanya tekstual, misalnya, email atau halaman web.



Gambar 1. Teknik rekomendasi sistem [1]

Pada *content based filtering*, fitur yang paling deskriptif digunakan untuk memodelkan item dan pengguna. Fitur yang paling diskriminatif diidentifikasi, dan disimpan sebagai vektor yang berisi fitur dan bobotnya. Model pengguna biasanya terdiri dari fitur-fitur item pengguna. Untuk menghasilkan rekomendasi, model pengguna dan kandidat rekomendasi dibandingkan, misalnya menggunakan model ruang vektor dan koefisien kesamaan cosinus. *Over spesialisasi* adalah salah satu kekurangan dari *content based filtering*, dimana pengguna akan direkomendasikan fitur sejenis yang memiliki *similarity* tertinggi dengan profil dari penggunanya.

Tenik yang kedua adalah *Knowledge-based filtering*. Teknik ini merekomendasikan item kepada pengguna berdasarkan pengetahuan domain. Dengan kata lain, sistem memiliki beberapa pengetahuan tentang bagaimana item tertentu berkorelasi dengan pengguna tertentu. Teknik ini menggunakan penalaran berbasis kasus atau metode ontologis untuk menghasilkan rekomendasi. *Content based filtering* dan *knowledge-based filtering* tidak digunakan di penelitian ini karna data yang dipakai hanya nilai rating hotel di Yogyakarta yang didapat dengan melakukan *crawling* data dari situs traveloka.

Teknik yang ketiga adalah *collaborative-filtering*. Teknik ini akan merekomendasikan item populer kepada pengguna berdasarkan umpan balik dari pengguna lain yang memiliki atribut yang sama. Dua pendekatan yang paling umum untuk teknik penyaringan ini adalah *collaborative filtering* berbasis memori dan berbasis model. Pendekatan yang berbasis memori akan membandingkan catatan historis pengguna dengan catatan lain dalam database [2]. Pendekatan berbasis model menggunakan metode statistik atau pembelajaran, di mana model akan mengklasifikasikan histori pengguna dan membangun model yang dapat digunakan dalam proses rekomendasi. *Collaborative Filtering* berbasis memori memiliki dua kelemahan utama: masalah *sparsity* data dan *scalability*. *Sparsity* terjadi ketika rating yang diberikan oleh user sangat sedikit. Sedangkan *scalability* terjadi ketika data yang harus dicari *similarity* nya berjumlah besar.

Tinjauan Pustaka

Collaborative filtering bekerja dengan mengumpulkan dan menganalisis informasi-informasi dengan jumlah banyak yang mencerminkan perilaku pengguna, kegiatan dan rating pengguna pada suatu item serta memprediksi item berbeda yang berdasarkan kedekatan item yang dipilih oleh pengguna lain [3]. Kelemahan utama pada teknik ini yaitu *sparsity* dan *scalability* data.

Sparsity data disebabkan karena adanya suatu kondisi dimana item data belum banyak direferensikan sehingga item baru yang muncul sebagai rekomendasi bisa jadi kurang sesuai dengan keinginan dari pengguna [4].

Masalah sparsity data telah coba diselesaikan oleh beberapa peneliti melalui beberapa eksperimen [5,6,7,8]. Salah satunya dengan memanfaatkan algoritma KNN.

Diasumsikan ada seorang user pernah merating item i, maka item yang akan direkomendasikan ke user tersebut akan dihitung menggunakan rumus KNN. Algoritma KNN adalah salah satu algoritma dengan penggunaan rumus *similarity* yang cukup sederhana yaitu *euclidean distance* namun mampu menyelesaikan masalah klasifikasi seperti skalabilitas. Kelebihan lain dari metode KNN adalah mampu memberikan rekomendasi yang lebih cepat dan lebih akurat kepada user dengan kualitas yang baik [18]. KNN biasanya menggunakan 2 rumus dasar yang bisa dipilih untuk mencari kedekatan antara data latih dan data uji yaitu *euclidean distance* dan *cosine similarity* [18]. Pada penelitian yang dilakukan, *euclidean distance* digunakan untuk mencari nilai kedekatan ini. Tahapan yang dilakukan oleh algoritma KNN adalah :

1. Hitung jarak untuk setiap data di *cluster* yang sama menggunakan rumus

$$dist(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \text{.....(4)}$$

$dist(x_1, x_2)$ adalah jarak, x_{1i} adalah nilai data pertama dan x_{2i} adalah nilai data kedua.

2. Penentuan titik rating tetangga terdekat dari user ditentukan dengan rumus :

$$C_i = \{x \in C_p; d(x, x_i) \le d(x, x_m), i \# m\} \text{....(5)}$$

dimana C_i adalah kelas prediksi.

Masalah yang kedua yaitu *scalability* bisa diatasi oleh metode *clustering* [5]. Algoritma *clustering* digunakan untuk mengelompokkan data k sebanyak k *cluster* dengan jumlah *cluster* k ≥2, dimana objek yang memiliki tingkat kesamaan yang tinggi akan

dikelompokkan dalam *cluster* yang sama sekaligus melebarkan jarak dengan objek yang ada di *cluster* lainnya [6]. Metode *cluster* biasanya digunakan di tahapan awal untuk mencari profil *cluster* pengguna yang memiliki nilai kesamaan yang tinggi. Profil cluster pengguna yang terbentuk, bisa dijadikan preferensi dan dapat mengurangi waktu komputasi dalam pembuatan rekomendasi [7]. Metode untuk pengelompokan data sendiri ada banyak macamnya seperti metode k means, DB Scan, Self Organizing Map, hierarchical clustering [6][7]. Diantara metode metode tersebut, yang paling sering digunakan adalah metode k means karena rumus yang digunakan sederhana dan memiliki iterasi yang relatif cepat [8].

K means akan mencoba melakukan grouping user profile yang ada di dataset berdasarkan atribut user, item hotel, dan nilai rating. Euclidean distance seperti di rumus 1, menjadi dasar k means dalam mencari jarak (d) sebuah data (X_n, dimana n=1,2,...,m) dengan centroidnya (Y_n, dimana y=1,2,....,n) [9].

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} .....(1)$$

Beberapa metode telah digunakan untuk menemukan jumlah cluster yang dianggap tepat pada algoritma k means, seperti silhouette method [10], Sum of Squared Errors (SSE) [11], Elbow Method [12]. Metode Silhouette method akan digunakan untuk penentuan jumlah cluster k means di penelitia ini.

Metode Silhouette digunakan untuk menilai validitas pengelompokan dengan memilih jumlah cluster optimal menggunakan data skala rasio [13]. Ketika diterapkan, algoritma silhouette akan mengukur jarak rata-rata suatu objek data i dengan semua objek data yang ada di cluster yang sama $a(i)$ dengan objek data di cluster lainnya $b(i)$ menggunakan rumus 2 [13]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} ....(2)$$

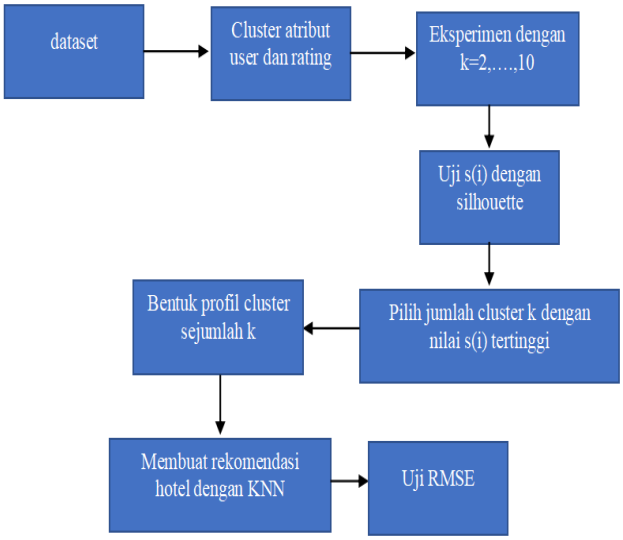
dimana $s(i)$ adalah nilai dissimilarities. Untuk menentukan dissimilarity suatu objek dengan objek lainnya digunakan rumus 3:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & if \quad a(i) < b(i), \\ 0 & if \quad a(i) = b(i), \\ b(i)/a(i) - 1 & if \quad a(i) > b(i), \end{cases} \quad (3)$$

Jumlah kluster $s(i)$ hasil rekomendasi algoritma silhouette akan diset di langkah awal algoritma k means. K means akan mengelompokkan data ke dalam sejumlah cluster $k(i)$ untuk membentuk cluster profil user. Nilai silhouette di penelitian dihitung untuk batasan k=1 sampai dengan k=10. Nilai Silhoutte yang paling mendekati angka 1 menunjukkan jumlah k yang paling optimal. Kemudian setelah dipilih jumlah k yang optimal, akan dipilih kluster yang didalamnya terdapat target (user) untuk dilakukan pencarian *similarity*.

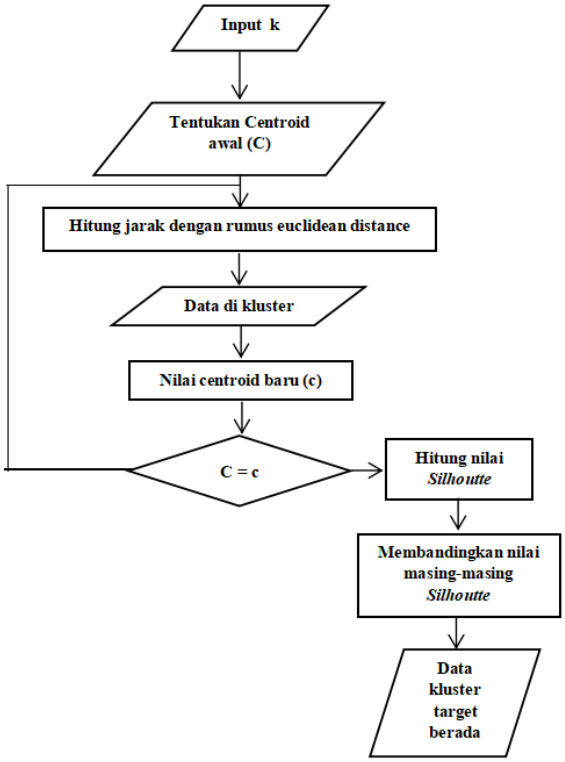
Metode Penelitian

Penelitian ini dilakukan dengan langkah sepeti tergambar di gambar 2 dan gambar 3.



Gambar 2. Tahapan Penelitian

Dataset yang dipakai di penelitian ini, berjumlah 700 rating yang di crawling melalui web traveloka. Data rating antara user dan item akan disimpan dalam database, untuk selanjutnya dirubah menjadi bentuk array user-item. Data ini lalu di *cluster* oleh algoritma k menas untuk mendapatkan profil dari user. Pembentukan profil dari user dilakukan melalui beberapa tahapan seperti tergambar di gambar 3.



Gambar 3. Skema pembentukan profil user oleh algoritma k means dan silhouett

K means akan mencoba melakukan grouping user profile yang ada di dataset berdasarkan atribut user, item hotel, dan nilai rating.

Nilai centroid menjadi acuan algoritma k means dalam menentukan cluster dari suatu data. Jumlah cluster (k), centroid ditentukan di awal. Jumlah cluster yang terlalu sedikit ataupun terlalu banyak akan mempengaruhi *similarity* dan *dissimilarity* objek datanya. Pada paper ini, jumlah cluster ditentukan oleh metode silhouette.

Prediksi rating antar user, dihitung menggunakan algoritma KNN. Tahapan KNN dalam mencari simililarity dapat dilihat digambar 4.

Gambar 4. Skema prediksi rating dengan algoritma KNN

Uji tingkat akurasi dari item-item yang akan menjadi rekomendasi menggunakan teknik RMSE. Proses ini dilakukan untuk mengetahui berapa besaran presentase error dari penerapan algoritma dalam membuat rekomendasi.

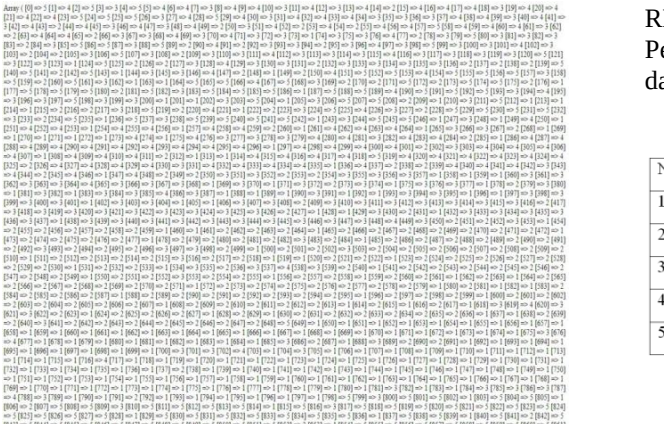
Hasil dan Pembahasan

Dataset diambil langsung melalui proses crawling web traveloka. Data yang dikumpulkan berjumlah 700 data seperti pada tabel 1.

Tabel 1. Dataset

No	User	Spesifikasi Hotel
1	U1	Hotel Tentrem Yogyakarta, "Jetis, Yogyakarta", 1512500, "HAS_24_HOUR_FRONT_DESK, RESTAURANT, WIFI_PUBLIC_AREA", 5
2	U2	Crystal Lotus Hotel, "Mlati, Yogyakarta", 850000, "HAS_24_HOUR_FRONT_DESK, RESTAURANT, WIFI_PUBLIC_AREA", 4
3	U1	Royal Ambarrukmo Yogyakarta, "Depok, Yogyakarta", 1206551, "HAS_24_HOUR_FRONT_DESK, RESTAURANT, WIFI_PUBLIC_AREA", 5
….		….

Data ini kemudian dilakukan preprocessing dengan menghilangkan semua bagian kecuali user yang diidentifikasi melalui email, nama hotel dan rating.
Data ini kemudian di rubah kedalam bentuk array matriks user dan item seperti yang terlihat di gambar 5.

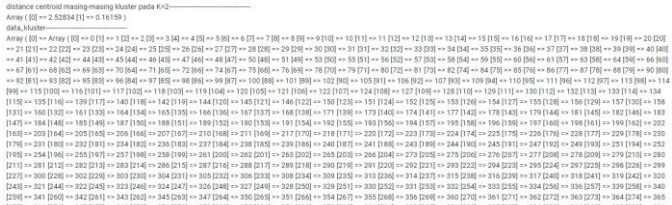


Gambar 5. Array dataset

Data ini kemudian dikelompokkan oleh algoritma k means untuk mendaptkan profil dari usernya. Jumlah kelompok (k) ditentukan menggunakan metode silhouette.

Nilai silhoutte pada k = 2 adalah 0,71 sedangkan nilai silhoutte pada k = 3 adalah 0,65. Nilai k dikatakan baik jika menghasilkan angka yang mendekati 1, sedangkan semakin menjauhi 1 dan mendekati -1 maka nilai silhoette buruk. Dari hasil perhitungan awal, dapat disimpulkan k optimal yang dapat ditemukan adalah k = 2.

Perhitungan nilai silhouette pada aplikasi dapat dilihat pada gambar 6.



Gambar 6. Hasil perhitungan nilai silhouette

Sistem yang berjalan kemudian mencari kedekatan rating antar neighbour menggunakan KNN. Tidak semua data diprediksi ratingnya. Hanya user yang ada dalam satu kelompok user profile saja yang dihitung. *Screenshoot* perhitungan prediksi rating oleh algoritma KNN dapat dilihat pada gambar 7.

Gambar 7. Pencarian nilai rating oleh KNN

Hasil perhitungan diatas merupakan hasil dengan data input rating sebesar 3. Kemudian memberikan data rekomendasi berupa indeks data dari item yang direkomendasikan. Data item yang direkomendasikan dihitung nilai RMSE nya terhadap data input rating sebenarnya ke 2. Didapat nilai RMSE sebesar 1,06.

Pengujian dilakukan pada 5 data inputan user dengan data sebagai berikut:

Tabel 1. Data *Testing*

No	Id_*user*	*user*name	1	Input ke 2
1	20	rosyid	3	4
2	21	ros	4	4
3	22	day	1	2
4	23	mad	2	1
5	24	frendy	5	3

Berdasarkan data di tabel 1, maka di dapat nilai RMSE sebagai berikut :

Tabel 2. Nilai *RMSE*

Id_*user*	Input ke 1	Input ke 2	Nilai *RMSE*
20	3	4	1
21	4	4	1,06
22	1	2	1,22
23	2	1	2,39
24	5	3	0,96

Berdasarkan data nilai RMSE maka didapat nilai akurasi sebagai berikut:

RMSE = 1 = 1%
Akurasi = 100% - 1% = 99%

RMSE = 1,06 = 1,06%
Akurasi = 100% - 1,06% = 98,94%

RMSE = 1,22 = 1,22%
Akurasi = 100% - 1,22% = 98,78%

RMSE = 2,39 = 2,39%
Akurasi = 100% - 2,39% = 97,61%

RMSE = 0,96 = 0,96%
Akurasi = 100% - 0,96% = 99, 04%

Maka didapat rata-rata sebagai berikut:

Rata-rata RMSE = (1 + 1,06 + 1,22 + 2,39 + 0,96) / 5 = 1,326 = 1,33 = 1,33%

Rata-rata akurasi = 100% - 1,33% = 98,67 %.

Kesimpulan dan Saran

Berdasarkan hasil perhitungan diatas maka dapat disimpulkan :

1. Algoritma K-Means memiliki performa baik untuk membantu proses klustering data sehingga waktu proses rekomendasi tidak memakan waktu dan memori besar.
2. Average Silhoutte Method mampu mengoptimalkan fungsi K-Means dalam menentukan jumlah kluster.

Algoritma KNN memiliki performa cukup baik dengan rata-rata error 1,33% dan rata-rata akurasi 98,67% pada pengujian data sebanyak 5.

Daftar Pustaka

[1] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, Recommendation systems: Principles, methods and evaluation, *Egyptian Informatics Journal* Volume 16, Issue 3, November 2015, Pages 261-273.

[2] S. Schiaffino, A. Amandi, Intelligent User Profiling, M. Bramer (Ed.): Artificial Intelligence, *LNAI* 5640, pp. 193 – 216, 2009.

[3] J.S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, *UAI'98: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, July 1998, Pages 43–52.

[4] A.F Jain, S.K Vishwakarma, P. Jain, An Efficient Collaborative Recommender System for Removing Sparsity Problem, *ICT Analysis and Applications* pp 131-141

[5] J. Cheng, L. Zhang, Jaccard Coefficient-Based Bi-clustering and Fusion Recommender System for Solving Data Sparsity, *Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD 2019: Advances in Knowledge Discovery and Data Mining* pp 369-380

[6] J.P Ortega, N.N.A Ortega, D. Romero, Balancing effort and benefit of K-means clustering algorithms in Big Data realms, *PLoS ONE* 13(9): 2018

[7] U. Kużelewska, Clustering Algorithms in Hybrid Recommender System on MovieLens Data, *STUDIES IN LOGIC, GRAMMAR AND RHETORIC* 37 (50) 2014

[8] Theodoridis, S., Pikrakis, A., Koutroumbas, K., Cavouras, D. (2010). *An Introduction to Pattern Reccognition : A MATLAB Approach*. Academic Press, USA

[9] S. Awawdeh , A. Edinat, A. Sleit, An Enhanced K-means Clustering Algorithm for Multi-attributes Data, *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 17, No. 2, February 2019

[10] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987, Vol.20, pp.53-65.

[11] Kwedlo, W. A clustering method combining differential evolution with the k-means algorithm. *Pattern Recognition Letters*. 2011, Vol.32, pp.1613–1621.

[12] M A Syakur, B K Khotimah, E M S Rochman, B D Satoto, Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster, *IOP Conf. Series: Materials Science and Engineering* 336 (2018) 012017 doi:10.1088/1757-899X/336/1/012017

[13] T. Thinsungnoena, N. Kaoungkub, P. Durongdumronchaib, K. Kerdprasopb, N. Kerdprasopb, The Clustering Validity with Silhouette and Sum of Squared Errors, *Proceedings of the 3rd International Conference on Industrial Application Engineering* 2015

[14] Zhang, D., Hsu, C.H., Chen, M., Chen, Q., Xiong, N., Lloret, J.: Cold-start recommendation using bi-clustering and fusion for large-scale social recommender systems. *IEEE Trans. Emerg. Top. Comput*. 2(2), 239–250 (2014)

[15] Xie, F., Xu, M., Chen, Z.: RBRA: a simple and efficient rating-based recommender algorithm to cope with sparsity in recommender systems. *In: International Conference on Advanced Information NETWORKING and Applications Workshops*, pp. 306–311 (2012)

[16] Ricci, F., Rokach, L., Shapira, B., Kantor, *P.B.: Recommender Systems Handbook*. Springer, US (2011). https://doi.org/10.1007/978-0-387-85820-3

[17] A. Gong, Y. Gao, Z. Gao, W. Gong, H. Li, H. Gao, A Slope One and Clustering based Collaborative Filtering Algorithm, *International Journal of Hybrid Information Technology* Vol.9, No.4 (2016), pp. 437-446

[18] D.A. Adeniyi, Z. Wei, Y. Yongquan, Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, *Applied Computing and Informatics* Volume 12, Issue 1, January 2016, Pages 90-108.