Implementasi Content-based Filtering dan K-Medoids Clustering pada Aplikasi Layanan Penyedia Informasi Hobi

Felix[1], Syanti Irviantina[2], Firman[3], Aldian Syahri[4]
STMIK Mikroskil, Jl. Thamrin No. 112, 124, 140, Telp. (061) 4573767, Fax. (061) 4567789
Program Studi Teknik Informatika, STMIK Mikroskil, Medan
[1]felix.pandi@mikroskil.ac.id, [2]syanti@mikroskil.ac.id,
[3]151111402@students.mikroskil.ac.id, [4]151112874@students.mikroskil.ac.id

Abstrak

Hobi merupakan suatu hal yang merepresentasikan kesukaan individu berdasarkan aktivitas yang sering dilakukannya. Sangat minimnya informasi mengenai hobi seperti event, tempat, komunitas dan berita yang berkaitan dengan hobi pengguna menjadi kendala bagi setiap individu dalam mengembangkan hobi yang dimilikinya. Oleh sebab itu, dalam penelitian ini dibuat sebuah aplikasi yang dapat merekomendasikan informasi sesuai dengan hobi pengguna dengan menggunakan algoritma content-based filtering dan dikelompokkan berdasarkan urutan rating sebagai konten teratas yang direkomendasikan dengan menggunakan algoritma k-medoids clustering. Hasil pengujian sistem perangkat lunak yang sudah dilakukan untuk menguji fungsional sistem yang digambarkan dalam tabel skenario pengujian dengan menggunakan metode blackbox testing. Maka hasil kesimpulan yang bisa di dapat dari hasil pengujian bahwa fungsional sistem pada aplikasi berjalan sesuai dengan yang diharapkan. Pengujian hasil rekomendasi menunjukkan bahwa nilai error yang didapatkan melalui perhitungan keseluruhan rata-rata hasil akhir Mean Absolute Error (MAE) pada implementasi algoritma content-based filtering dan k-medoids clustering dalam sistem rekomendasi relatif rendah dengan hasil akhir yaitu 0.19233894129602 pada rentang 0-1 yang berarti hasil rekomendasi menunjukkan keakuratan yang baik.

Kata kunci— *Content-based filtering, k-medoids clustering, mean absolute error*

Abstract

Hobbies are things that represent individual preferences based on the activities he often does. Very minimal information about hobbies such as events, places, communities and news related to user hobbies is an obstacle for every individual in developing their hobbies. Therefore, in this study an application was made that could recommend information according to users' hobbies by using content-based filtering algorithms and grouped according to rating order as the top recommended content using the k-medoids clustering algorithm. The results of testing software systems that have been carried out to test the functional system described in the test scenario table using the blackbox testing method. Then the conclusion can be obtained from the test results that the functional system in the application runs as expected. Testing the recommended results show that the error value obtained through the overall calculation of the average final result Mean Absolute Error (MAE) in the implementation of the content-based filtering algorithm and k-medoids clustering in the recommendation system is relatively low with the final result of 0.19233894129602 in the range 0-1 which means the results of the recommendations show good accuracy.

Keywords— *Content-based filtering, k-medoids clustering, mean absolute error*

1. PENDAHULUAN

Setiap manusia pasti memiliki sebuah hobi, misalnya saja seperti bermain bola, bermain musik, membaca buku dan hobi lainnya. Berdasarkan hasil dari sebuah penelitian yang dikemukakan oleh Dimas Nurhariyadi bahwa, hobi merefleksikan kesukaan akan sesuatu maupun kegiatan pada setiap *individu*, apabila hobi tersebut ditekuni dan dikembangkan, akan menjadikan seseorang memiliki jiwa

profesional dalam menjalankan aktivitas suatu hobi [1]. Dalam menjalankan aktivitas hobi tersebut sangat dibutuhkan informasi mengenai hobi yang dimiliki baik halnya mengenai komunitas, maupun *event-event* yang diselenggarakan. Pihak promotor juga terkadang kurang efektif dalam menyampaikan informasi mengenai *event* yang akan diselenggarakan karena kurangnya media promosi *event* sesuai dengan minat dan hobi.

Untuk itu diperlukan sebuah aplikasi yang memudahkan pengguna dalam menemukan informasi berita, acara dan komunitas sesuai dengan hobi yang di minati. Pengguna juga dapat melakukan proses pemesanan tiket acara maupun melakukan *booking* untuk fasilitas tempat yang tersedia secara langsung melalui aplikasi.

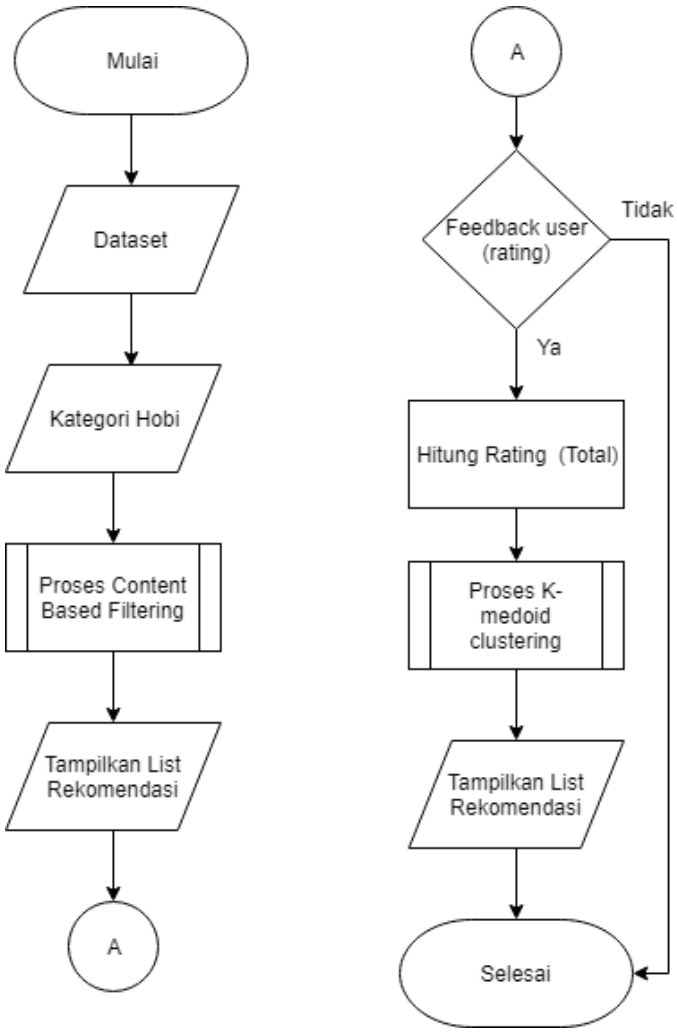
Pengguna yang telah memilih peminatan hobi dapat menemukan informasi mengenai komunitas maupun informasi event yang berkaitan dengan hobinya berdasarkan metode sistem rekomendasi yang banyak digunakan untuk menyarankan *item* kepada pengguna berdasarkan hobi yang telah dipilih. *Algoritma Content-Based Filtering (CBF)* merupakan salah satu tipe sistem rekomendasi tradisional. Akar dari penyaringan berbasis konten yang digunakan yaitu dalam pengambilan informasi dan penyaringan informasi. Sistem akan melakukan penilaian berdasarkan analisis kemiripan profil pengguna dengan vektor komponen pembentuk item. Jika item tersebut akan disukai oleh pengguna maka item tersebut akan direkomendasikan ke pengguna [2]. Pada informasi mengenai *event* dan tempat pertandingan seperti stadion, lapangan futsal maupun tempat-tempat lainnya yang akan menampilkan rekomendasi sesuai dari urutan teratas berdasarkan *rating* dan kategori hobi yang dipilih pengguna. Penerapan algoritma yang sesuai yaitu dengan menggunakan algoritma *K-Medoids Clustering* yang mengelompokkan kumpulan data dari *n* objek ke dalam *cluster.* Algoritma *K-Medoids* memiliki kelebihan untuk mengatasi kelemahan pada pada algoritma *K-Means* yang *sensitive* terhadap *noise* dan *outlier*, dimana objek dengan nilai yang besar yang memungkinkan menyimpang pada distribusi data. Kelebihan lainnya yaitu hasil proses *clustering* tidak bergantung pada urutan masuk *dataset* [3].

Berdasarkan uraian di atas, maka terciptalah sebuah inovasi untuk membuat aplikasi yang dapat memudahkan pengguna dalam menemukan informasi mengenai hobi maupun informasi *event*. Yang dalam penerapannya akan dirancang sebuah aplikasi *mobile* dan *web* disusun ke dalam bentuk sebuah penelitian dengan judul *"Aplikasi Layanan Penyedia Informasi Hobi dengan Algoritma Content-Based Filtering dan K-Medoids Clustering".*

2. METODE PENELITIAN

2.1 Analisis Proses

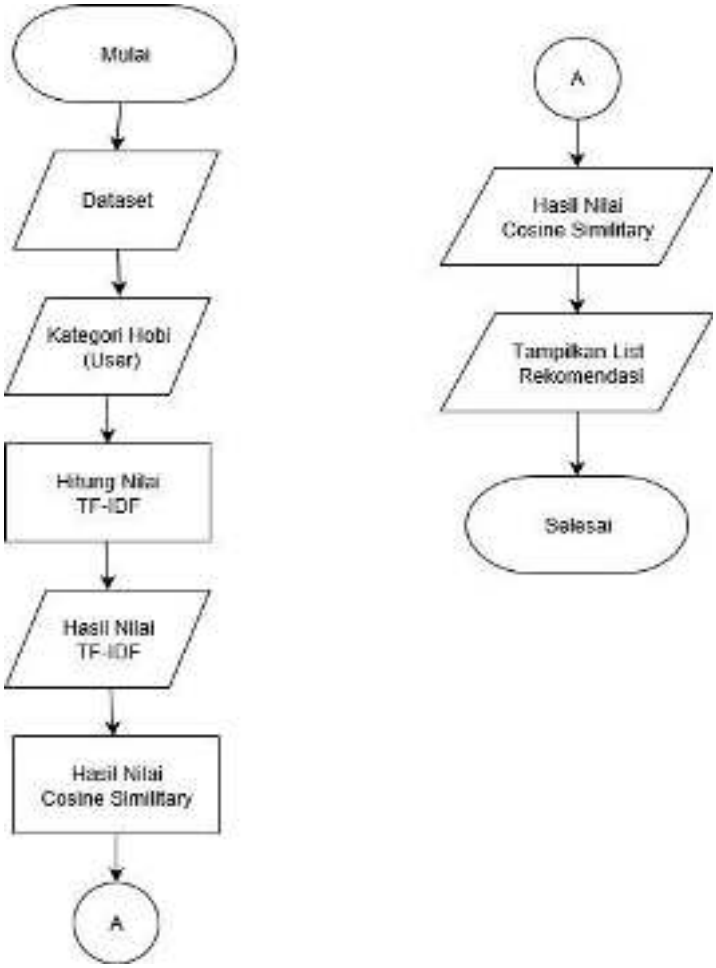
Analisis proses menjelaskan tentang alur kerja penerapan algoritma dalam proses pembuatan aplikasi untuk menyelesaikan permasalahan dalam merekomendasikan informasi yang sesuai dengan hobi pengguna dengan menggunakan metode *content-based filtering.* Selanjutnya, sistem akan merekomendasikan tempat dan acara berdasarkan *rating* dari pengguna dengan menggunakan metode *K-Medois Clustering.*



Gambar 1 *Flowchart* Analisis Proses Hasil Rekomendasi dengan *Content-based Filtering* dan *K-medoids clustering*

Proses Rekomendasi dengan Menggunakan Algoritma *Content-Based Filtering*

Pada tahapan ini dilakukan proses *filtering* untuk mengelompokkan *item* untuk data yang di uji berdasarkan nilai kecocokan antar *item,* menghitung nilai *TF-ID* dan menghitung nilai *similarity* antara dua *item a* dan *item b.* Pada proses ini dilakukan dengan mengambil 3 sampel data uji dari *dataset* untuk selanjutnya dilakukan proses *filtering*. Berikut langkah-langkah yang dilakukan dalam mencari hasil rekomendasi dengan metode *CBF.*



Gambar 2 *Flowchart* Pencarian Hasil Rekomendasi dengan Metode *CBF*

Content-Based Filtering

Content-based filtering memberikan rekomendasi dengan menganalisis deskripsi item-item itu telah diberi peringkat oleh pengguna dan deskripsi item yang akan direkomendasikan. Semakin banyak algoritma yang dimiliki telah diusulkan untuk menganalisis konten dokumen teks dan menemukan kesamaan dalam konten ini yang dapat ditayangkan sebagai dasar untuk membuat rekomendasi.

Dalam sistem pencarian informasi, langkah pertama adalah mengidentifikasi kata kunci untuk mewakili dokumen, Itumenghindari pengindeksan kata-kata yang tidak berguna, sistem pencarian teks sering mengaitkan daftar berhenti dengan seperangkat dokumen. Itu kata-kata yang tidak relevan disebut *stop list (the, of, for, with*, dll) [4].

Matriks frekuensi istilah (berat) TF (d.t) mengukur keterkaitan suatu istilah dengan sehubungan dengan dokumen yang diberikan d. Ini didefinisikan sebagai 0 jika dokumen tidak mengandung istilah dan bukan nol sebaliknya. Itu frekuensi istilah relatif diukur menggunakan frekuensi istilah versus jumlah total kemunculan semua ketentuan dalam dokumen. Frekuensi istilah dihitung.

$$TF(d.t) = \begin{cases} 0 & \text{if } freq(d,t) = 0 \\ 1+\log(1+\log(freq(d,t))) & \text{otherwise} \end{cases} \quad (1)$$

Keterangan :
TF :banyaknya kata yang dicari pada sebuah dokumen
d : dokumen ke-*d*

t : kata ke-*t* dari kata kunci

Ada ukuran penting lainnya, yang disebut frekuensi dokumen terbalik (IDF) dalam Persamaan.1 yang mewakili faktor penskalaan, atau pentingnya istilah t dan itu akan berkurang jika istilah t muncul di banyak dokumen. Untuk contoh istilah informasi mungkin kurang penting dalam banyak makalah penelitian. Formula untuk IDF (t) diberikan dalam Persamaan 2.

$$IDF(t) = \frac{\log 1 + |d|}{|dt|} \quad (2)$$

Keterangan :
IDF : Inversed Document Frequency
d : dokumen ke-*d*
t : kata ke-*t* dari kata kunci

Di mana, d adalah koleksi dokumen, dan dt adalah kumpulan dokumen yang berisi istilah t. Dalam ruang vektor lengkap
model, TF dan IDF digabungkan bersama, yang membentuk ukuran TF-IDF diberikan dalam Persamaan 3.

$$\text{TF-IDF}(d,t) = \text{TF}(d,t) * \text{IDF}(t) \quad (3)$$

Keterangan :
TF :banyaknya kata yang dicari pada sebuah dokumen
IDF : Inversed Document Frequency
d : dokumen ke-*d*
t : kata ke-*t* dari kata kunci

Untuk menentukan kesamaan antara dua dokumen, cosine similarity digunakan. Salah satu yang paling jelas. Kelebihan dari algoritma *content-based filtering* adalah algoritma ini tidak perlu domain pengetahuan. Inicukup untuk mengumpulkan umpan balik dari pelanggan tentang prioritas mereka. Keuntungan selanjutnya dari *content-based filtering* yang dapat kita pertimbangkan adalah, algoritma ini lebih baik daripada *Collaborative Filtering (CF)*.

a. *Vector Cosine-Based Similarity*

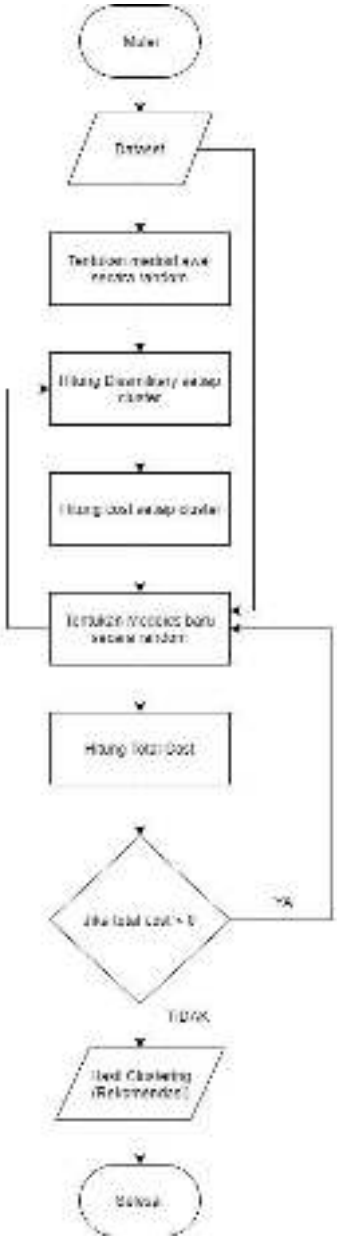
Vector cosine similarity antara *item i* dan *j* diberikan dimana "•"menunjukkan titik-produk dari dua vektor. Sebuah matriks kesamaan $n \times n$ dihitung untuk mendapatkan perhitungan kesamaan yang diinginkan, untuk *n item*. Jika :
Persamaan vector kosinus antara A dan B yang diberikan dalam persamaan 4.

$$W_{A,B} = \cos(\vec{A} \cdot \vec{B}) = \frac{x1x2 + y1y2}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \quad (4)$$

Keterangan :
Vektor A=*{x1,y1},*vektor B=*{x2,y2}*.

Proses Pembentukan *Clustering* dengan Menggunakan Algoritma *K-Medoids Clustering*

Dalam tahapan ini dilakukan proses *clustering* untuk mengelompokkan *item* kedalam *cluster* berdasarkan nilai *rating* yang dilakukan oleh *user*. Berikut langkah-langkah dalam proses rekomendasi dengan *k-medoids.*



Gambar 3 *Flowchart* Hasil *Clustering* Metode *K-Medoids Clustering*

K-medoids Clustering

Algoritma *k-Means* sensitif terhadap *outlier* karena objek dengan nilai yang sangat besar mungkin secara substansial mendistorsi distribusi data. Alih-alih mengambil nilai rata-rata objek dalam *a* klaster sebagai titik referensi, *medoid* dapat digunakan, yang merupakan objek paling terpusat terletak di sebuah *cluster*. Dengan demikian, metode partisi masih bisa dilakukan berdasarkan prinsip meminimalkan jumlah perbedaan antara setiap objek dan objeknya titik referensi yang sesuai [5]. Ini membentuk dasar dari metode *k-medoids*. Strategi dasar algoritma pengelompokan *k-medoids* adalah menemukan *k cluster* di *n* keberatan dengan terlebih dahulu menemukan perwakilan secara sewenang-wenang objek *(medoid)* untuk setiap *cluster*. Masing-masing tersisa objek dikelompokkan dengan *medoid* yang itu paling mirip. Metode *k-medoid* menggunakan objek representatif sebagai titik referensi alih-alih mengambil nilai rata-rata objek di setiap cluster adalah titik kunci dari metode ini. Algoritma mengambil input parameter *k*, jumlah *cluster* menjadi dipartisi di antara seperangkat objek *n*. Algoritma *k-medoids* untuk partisi berdasarkan objek *medoid* atau pusat adalah sebagai berikut:

```
Input: k: Jumlah cluster
D: Kumpulan data yang berisi n objek
Output: Satu set k cluster yang meminimalkan jumlah ketidaksamaan semua objek dengan medoid terdekat.
Method: Pilih objek k dalam D sebagai objek representatif awal;
Ulangi Penetapan setiap objek yang tersisa ke cluster dengan medoid terdekat;
pilih secara acak objek non medoid Orandom;
Hitung total poin S dari objek swap Oj dengan Orandom;
jika S < 0 maka tukar Oj dengan Orandom untuk membentuk set baru k medoid;
Sampai tidak ada perubahan;
```

Tentukan partisi *k* untuk *n* objek. Setelah pemilihan acak awal *k medoid*, the algoritma berulang kali mencoba membuat pilihan yang lebih baik *medoid*. Karena itu, algoritmanya sering disebut sebagai algoritma berbasis objek representatif.

3. HASIL DAN PEMBAHASAN

3.1 Pengujian

Untuk melihat hasil aplikasi yang telah dibangun, maka dilakukan 2 (dua) pengujian, yaitu pengujian terhadap sistem perangkat lunak dengan menggunakan metode *Blackbox Testing* dan pengujian keakuratan hasil rekomendasi dari sistem dengan menggunakan *Mean Absolute Error* (MAE).

3.1.1. Pengujian Sistem dengan Blackbox Testing

Berdasarkan hasil pengujian sistem perangkat lunak yang sudah dilakukan untuk menguji fungsional sistem yang digambarkan dalam tabel skenario pengujian dengan menggunakan metode *blackbox testing*. Maka hasil kesimpulan yang bisa di dapat dari hasil pengujian bahwa fungsional sistem pada aplikasi berjalan sesuai dengan yang diharapkan.

3.1.2. Pengujian Keakuratan Hasil Rekomendasi dengan MAE

Rekomendasi pada hasil peminatan hobi yang telah dipilih oleh pengguna pada saat menggunakan aplikasi agar merekomendasikan konten yang ditampilkan sesuai dengan hobi yang telah dipilih pengguna berdasarkan konten yang telah dikelompokkan sesuai dengan urutan rekomendasi teratas. Maka perlu dilakukan pengujian keakuratan terhadap nilai prediksi yang dihasilkan untuk dilihat sejauh mana keakuratan hasil rekomendasi yang diberikan oleh sistem. Pengujian menggunakan *Mean Absolute Error* (MAE) untuk menghitung rata-rata *error* dari nilai prediksi yang dihasilkan. Pengujian dilakukan terhadap 6 *user* berdasarkan hobi yang berbeda-beda dan untuk hobi yang diuji ada 3 kategori yaitu musik, *e-sport* dan olahraga. Untuk data awalnya yaitu sebagai berikut :

Tabel 1 Sampel Data Hobi dari 6 *User*

id_user	Username	Hobi yang dipilih
1	Mantri99	Musik, *esport*, olahraga
2	Sadan33	*E-sport*, futsal, olahraga
3	Budi88	Musik, olahraga
4	yogi11	Musik, *e-sport*
5	Ewin11	Badminton, musik
6	Anastasya55	Badminton, olahraga, swimming

Pada tabel di atas terdapat data awal digunakan dari 6 *user* yang memiliki hobi berbeda. Perhitungan pengujian *Mean Absolute Error* dapat dilakukan dengan cara menjumlahkan nilai aktual

dan nilai hasil prediksi setiap konten yang direkomendasikan kemudian dibagi dengan jumlah konten yang diuji yang penjabarannya dirumuskan sebagai berikut :

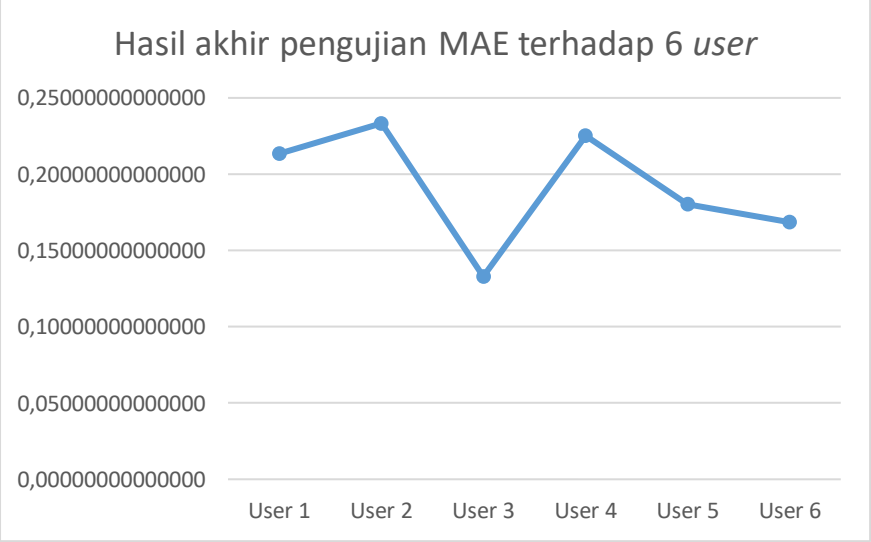
$$\text{MAE} = \frac{\sum_t^n |Y_t - \hat{Y}_t|}{n} \quad (5)$$

Dimana,
Y_t = Nilai Aktual
$\hat{Y}_t$ = Nilai Prediksi
n = Jumlah Konten

Tabel 2 Hasil Akhir Nilai Pengujian *MAE* dari 6 *User*

Id_User	Hasil rata-rata
User 1	0.21362136214026
User 2	0.23333333333333
User 3	0.13292893218813
User 4	0.22526906895103
User 5	0.18031091537756
User 6	0.16857003578577
Hasil akhir prediksi MAE=	**0.19233894129602**

Pada tabel di atas merupakan hasil akhir prediksi MAE dari 6 *user* yang diuji. Hasil tersebut merupakan nilai rata-rata pengujian yang di ambil pada setiap *user* dan selanjutnya dilakukan perhitungan untuk mencari nilai rata-rata keseluruhannya. Dari hasil pengujian menunjukkan bahwa nilai *error* yang didapatkan melalui perhitungan keseluruhan rata-rata hasil akhir MAE pada implementasi algoritma *content-based filtering* dan *k-medoids clustering* dalam sistem rekomendasi relatif rendah dengan hasil akhir yaitu 0.19233894129602 pada rentang 0-1, yang dimana jika hasil akhir yang didapat mendekati angka 1 maka semakin besar nilai *error* yang didapat dan juga nilai akurasi yang di dapat akan semakin rendah. Sehingga dapat disimpulkan bahwa penggunaan algoritma *content-based filtering* dan algoritma *k-medoids clustering* memiliki keakuratan yang baik.



Gambar 4 Grafik Hasil Pengujian *MAE* Terhadap 6 *User*

3.2. Pembahasan

Berdasarkan hasil pengujian sistem perangkat lunak yang sudah dilakukan untuk menguji fungsional sistem yang digambarkan dalam tabel skenario pengujian dengan menggunakan metode *blackbox testing*. Maka hasil kesimpulan yang bisa di dapat dari hasil pengujian bahwa fungsional sistem pada aplikasi berjalan sesuai dengan yang diharapkan.

Lalu dari hasil pengujian *MAE* menunjukkan bahwa nilai *error* yang didapatkan melalui perhitungan keseluruhan rata-rata hasil akhir *MAE* pada implementasi algoritma *content-based filtering* dan *k-medoids clustering* dalam sistem rekomendasi relatif rendah dengan hasil akhir yaitu 0.19233894129602 pada rentang 0-1, yang dimana jika hasil akhir yang didapat mendekati angka 1 maka semakin besar nilai *error* yang didapat dan juga nilai akurasi yang di dapat akan semakin rendah. Sehingga dapat disimpulkan bahwa penggunaan algoritma *content-based filtering* dan algoritma *k-medoids clustering* memiliki keakuratan yang baik.

4. KESIMPULAN

Berdasarkan pengujian yang dilakukan terhadap implementasi algoritma *Content-based filtering* dan algoritma *K-medoids clustering*. Beberapa kesimpulan yang diperoleh, yaitu :

1. Pengujian sistem perangkat lunak yang sudah dilakukan untuk menguji fungsional sistem yang digambarkan dalam tabel skenario pengujian dengan menggunakan metode *blackbox testing*. Maka hasil kesimpulan yang bisa di dapat dari hasil pengujian bahwa fungsional sistem pada aplikasi berjalan sesuai dengan yang diharapkan.
2. Dari hasil pengujian menunjukkan bahwa nilai *error* yang didapatkan melalui perhitungan keseluruhan rata-rata hasil akhir MAE pada implementasi algoritma *content-based filtering* dan *k-medoids clustering* dalam sistem rekomendasi relatif rendah dengan hasil akhir yaitu 0.19233894129602 pada rentang 0-1, yang dimana jika hasil akhir yang didapat mendekati angka 1 maka semakin besar nilai *error* yang didapat dan juga nilai akurasi yang di dapat akan semakin rendah. Sehingga dapat disimpulkan bahwa penggunaan algoritma *content-based filtering* dan algoritma *k-medoids clustering* memiliki keakuratan yang baik.

5. SARAN

Di dalam penelitian ini masih ada yang dapat diteliti dan dikembangkan, diantaranya :

1. Menambahkan proses yang mampu untuk mempertemukan antar pengguna berdasarkan ketertarikan hobi yang sama dengan membagikan informasi lokasi dari pengguna berdasarkan jarak terdekat dengan algoritma *neighborhood-based clustering.*
2. Menambahkan fitur *matching team* yang bertujuan untuk mencari lawan bertanding. Misalnya pada kategori komunitas sepakbola, fitur *matching team* akan merekomendasikan lawan bertanding sesuai dengan lokasi disekitar pengguna berdasarkan hobi yang sama antar pengguna.

DAFTAR PUSTAKA

[1] Nurhariyadi, Dimas. 2016. "Preferensi Ruang Hobi." *Prosiding Temu Ilmiah IPLBI 2016* (2016): hal 135-137.

[2] Moattari, Mahta. 2013. *Implementing a Content-based Recommender System for News Readers*. Diss. University of New Brunswick.

[3] Patel, Abhishek, dan P. Singh. 2013. "New Approach for K-mean and K-medoids algorithm." *International Journal of Computer Applications Technology and Research* 2.1 (2013): hal 1-5.

[4] Manjula, R., and A. Chilambuchelvan. 2016. "*Content Based Filtering Techniques in Recommendation System using user preferences*." http://ijiet.com/wp-content/uploads/2016/12/20.pdf

[5] Velmurugan, T. 2012. "Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points." *Int. J. Computer Technology & Applications* 3.5 (2012): hal 1758-1764. https://pdfs.semanticscholar.org/5d87/ef6efa31e9aa4f0b63aa1dabe31f9693a93b.pdf