 ISBN: 978-602-60280-1-3

Sistem Rekomendasi Film menggunakan Bisecting K-Means dan Collaborative Filtering

Arwin Halim[1], Hernawati Gohzali[2], Dita Maria Panjaitan[3], Ilham Maulana[4]
[1,2,3,4] *Jurusan Teknik Informatika, STMIK Mikroskil, Medan*
Jl. Thamrin No. 112, 124, 140, Telp. (061) 4573767, Fax. (061) 4567789
[1]*arwin@mikroskil.ac.id,* [2]*hernawati.gohzali@mikroskil.ac.id,* [3]*ditamaria41@gmail.com,* [4]*Ilhamm49@gmail.com*

Abstrak

Penelitian ini bertujuan untuk mengembangkan sistem rekomendasi film menggunakan kombinasi dari bisecting K-Means dan Collaborative Filtering. Data film yang digunakan pada penelitian ini berasal dari MovieLens yang terdiri dari 100.000 rating dari 668 user untuk 10329 judul film dalam 18 genre film. Proses training terdiri dari proses kluster dengan algoritma bisecting K-Means dan perhitungan nilai similarity dengan collaborative filtering (item-based dan user-based). Proses testing dilakukan untuk menghitung nilai error sistem dengan menghitung nilai Mean Absolute Error (MAE). Hasil penelitian menunjukkan rekomendasi dengan bisecting K-Means dan user-based collaborative filtering mendapatkan nilai MAE yang lebih rendah dibandingkan dengan bisecting K-Means dan item-based collaborative filtering.

Kata kunci : rekomendasi film, bisecting k-means, user-based, item-based, collaborative filtering

Abstract

This study aims to develop a film recommendation system using a combination of bisecting K-Means and Collaborative Filtering. The data used in this research came from MovieLens that are 100,000 ratings from 668 users for 10329 movie titles in 18 film genres. The training process consists of clustering process with bisecting K-Means algorithm and calculate similarity value using collaborative filtering (item-based and user-based). The testing process is performed to get error value using Mean Absolute Error (MAE). The results showed that bisecting K-Means and user-based collaborative filtering recommendations received lower MAE values compared to bisecting K-Means and item-based collaborative filtering.

Keywords: film recommendation system, bisecting k-means, user-based, item-based, collaborative filtering

I. PENDAHULUAN

Film sudah menjadi salah satu media hiburan yang populer di kalangan masyarakat. Sejak tahun 1874 sampai 2015, sebanyak 3,361,741 judul film telah dikeluarkan oleh industri perfilman (http://imdb.com). Banyak-nya judul-judul film yang telah beredar membuat masyarakat sulit untuk menemukan film yang mereka inginkan. Data-data rating film yang terdapat dalam suatu website dapat diolah dan dimanfaatkan untuk merekomen-dasikan film kepada user lain. Pertimbangan-nya adalah menemukan film berdasarkan hubungan antara satu film dan film lainnya yang sudah diberi rating oleh user untuk dijadikan rekomendasi kepada user lain. Oleh karena itu, diperlukan suatu sistem yang dapat merekomendasikan film kepada user.

Sistem rekomendasi adalah suatu mekanisme yang dapat memberikan suatu informasi atau rekomendasi sesuai dengan kesukaan user berdasarkan informasi yang diperoleh dari user (Sarwar dkk, 2001). Oleh karena itu, diperlukan model rekomendasi yang tepat agar rekomendasi yang diberikan oleh sistem sesuai dengan kesukaan user, serta mempermudah user mengambil keputusan dalam menentukan item (film) yang akan dipilih (McGinty dan Smyth, 2006). Salah satu metode rekomendasi yang digunakan dalam sistem rekomendasi adalah Collaborative filltering. Collaborative filltering menghu-bungkan setiap user dengan kesukaan yang sama terhadap suatu item (film) berdasarkan rating yang diberikan user. Untuk meningkatkan keakurasian hubungan antara user dengan kesukaan yang sama terhadap suatu item (film) digunakan algoritma clustering (Gupta, 2009).

Clustering adalah mengelompokkan item data kedalam sejumlah kecil grup sedemikian sehingga masing-masing grup mempunyai sesuatu persamaan yang esensial (Garcia-Molina dkk, 2002). Pada penelitian Gupta (2009) telah mencoba menggabungkan Collaborative filltering dengan algoritma k-means yang menghasilkan sistem rekomendasi yang efisien untuk pemrosesan data dalam jumlah besar dan akurasi yang tinggi. Bisecting K-Means merupakan algoritma yang lebih baik dibandingkan algoritma K-Means karena memproduksi cluster yang seragam dan tidak memproduksi cluster kosong, tingkat keakurasian yang baik dan lebih efisien ketika jumlah cluster meningkat (Patil dkk, 2015). Penelitian ini menggabungkan Collaborative filtering dengan Bisecting K-Means untuk menghasilkan sistem rekomendasi yang baik.

II. METODE PENELITIAN

Penelitian ini dimulai dengan proses pengumpulan data, analisis proses, analisis kebutuhan sistem, perancangan sistem, implementasi dan pengujian sistem.

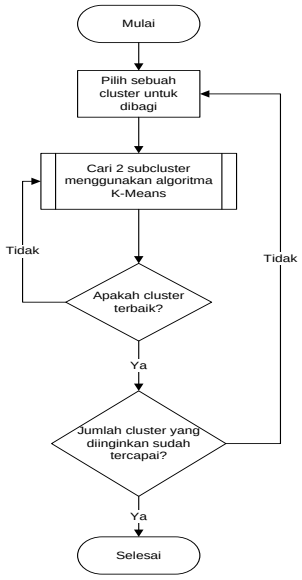
1. Pengumpulan Data

Data film diperoleh dari dataset MovieLens berisi 100.000 *rating*, 10.329 film dengan 668 *users.* Film dikelompok-kan dalam 18 *genre* yaitu *drama, comedy, short, dokumentary, talk-show, family, news, romance, animation, music, reality –Tv, crime, action, game-Show, adventure, thrailler, mystery, fantacy, sci-fi, adult, sport, horror, history, biography, western, war,* dan *film noir*. *User* dapat memberikan *rating* (skala 1-5) dengan rincian skala 1 adalah yang paling buruk dan skala 5 adalah yang paling bagus, terhadap film jika telah melakukan registrasi.

2. Analisis Proses

a. Algoritma Bisecting K-Means

Bisecting K-Means adalah variasi dari algoritma K-Means (Patil dkk, 2015). Kunci dari algoritma ini adalah satu cluster dibagi menjadi dua sub-cluster di setiap langkah. Algoritma bisecting K-Means dapat dijelaskan pada Gambar 1.



Gambar 1. Flowchart bisecting K-Means

b. Algoritma Collaborative Filtering (CF)

Ide utama dalam sistem rekomendasi collaborative filtering adalah untuk memanfaatkan opini *user* lain yang ada untuk memprediksi item yang mungkin akan disukai atau diminati oleh seorang *user* (Ricci dkk, 2011). Kualitas rekomendasi yang diberikan dengan menggunakan metode ini sangat bergantung dari opini *user* lain (*neighbor*) terhadap suatu item.

i. User-based Collaborative Filtering

User-Based Collaborative Filtering menemukan sekumpulan *user neighbour* memiliki sejarah kesukaan yang sama dengan *user* yang akan di jadikan sasaran rekomendasi. Setelah sekumpulan tetangga terbentuk, sistem menggabungkan kesukaan tetangga (*neighbour*) untuk menghasilkan rekomendasi kepada *user* yang sedang aktif. (Sarwar dkk, 2001)

a. Menghitung Similarity

Pearson Correlation digunakan untuk menghitung nilai kemiripan antara *user* dan *item*, seperti Persamaan 1.

$$S_{(i,j)} = \frac{\sum u \in U(R_{u,i}-\bar{R}_i)(R_{u,j}-\bar{R}_j)}{\sqrt{\sum u \in U(R_{u,i}-\bar{R}_i)^2}\sqrt{\sum u \in U(R_{u,j}-\bar{R}_j)^2}} \quad (1)$$

Keterangan:

$S_{(i,j)}$ = Nilai kemiripan antara *item* i dan *item* j

$u \in U$ = Himpunan *user* yang merating *item* i dan *item* j

$R_{u,i}$ = *Rating user* u pada *item* i

$R_{u,j}$ = *Rating user* u pada *item* j

R_i = *Rating* rata-rata *item* i

R_j = *Rating* rata-rata *item* j

b. Menghitung Nilai Prediksi

Langkah kedua yang dilakukan adalah menghitung prediksi *rating* dari item-item tersebut. Cara menghitung nilai prediksi untuk *user* atau item baru digunakan persamaan *Weighted Sum* (Sarwar dkk,2001) sesuai dengan Persamaan 2.

$$P_{(a,j)} = \frac{\sum_{i \in I}(R_{a,i} * RS_{i,j})}{\sum_{i \in I}|S_{i,j}|} \quad (2)$$

Keterangan:

$P_{(a,j)}$ = Prediksi *rating* item j oleh *user* a

$i \in I$ = Himpunan item i yang mirip dengan item j

$R_{a,i}$ = *Rating user* a pada item i

$s_{i,j}$ = Nilai *similarity* antara item i dan j

ii. Item-based Collaborative Filtering

Item-Based Collaborative Filtering merupakan metode rekomendasi yang didasari atas adanya kesamaan antara pemberian *rating* terhadap suatu produk dengan produk yang dibeli. Dari tingkat kesamaan produk, kemudian dibagi dengan parameter kebutuhan pelanggan untuk memperoleh nilai kegunaan produk. Produk yang memiliki nilai kegunaan tertinggilah yang kemudian dijadikan rekomendasi (Sarwar dkk, 2001)

a. Menghitung Similarity

Untuk mengitung nilai *similarity* antar item dan *user* digunakan persamaan *adjusted cosine* sesuai Persamaan 3

$$s_{(i,j)} = \frac{\sum_{u \in U}(R_{u,i}-\bar{R}_u)(R_{u,j}-\bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{u \in U}(R_{u,j}-\bar{R}_u)^2}} \quad (3)$$

Keterangan :

$S_{(i,j)}$ = Nilai kemiripan antara item i dan item j

$u \in U$ = Himpunan *user* yang merating item i dan item j

$R_{u,i}$ = *Rating user* u pada item i

$R_{u,j}$ = *Rating user* u pada item j

$\bar{R}_u$ = Nilai *rating* rata-rata *user* u

b. Menghitung Nilai Prediksi

Langkah kedua yang dilakukan adalah menghitung prediksi *rating* dari item-item tersebut. Cara menghitung nilai prediksi untuk *user* atau item baru digunakan persamaan *Weighted Sum* (Sarwar dkk,2001) sesuai dengan Persamaan 4.

$$P_{(a,j)} = \frac{\sum_{i \in I}(R_{a,i} * RS_{i,j})}{\sum_{i \in I}|S_{i,j}|} \quad (4)$$

Keterangan :

$P_{(a,j)}$ = Prediksi *rating* item j oleh *user* a

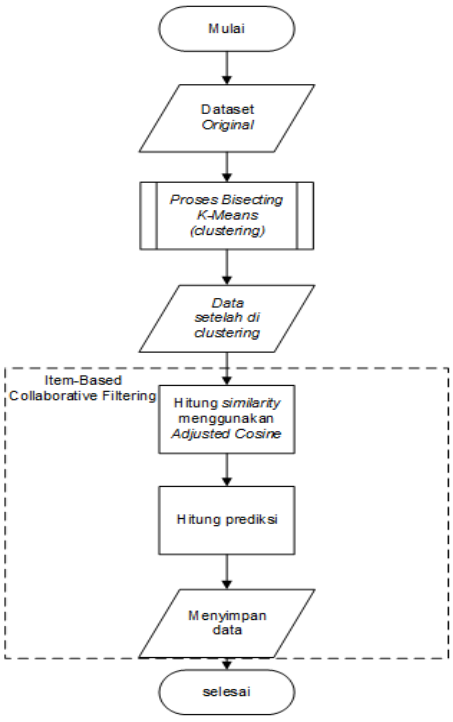
$i \in I$ = Himpunan item i yang mirip dengan item j

$R_{a,i}$ = *Rating user* a pada item i

$s_{i,j}$ = Nilai *similarity* antara item i dan j

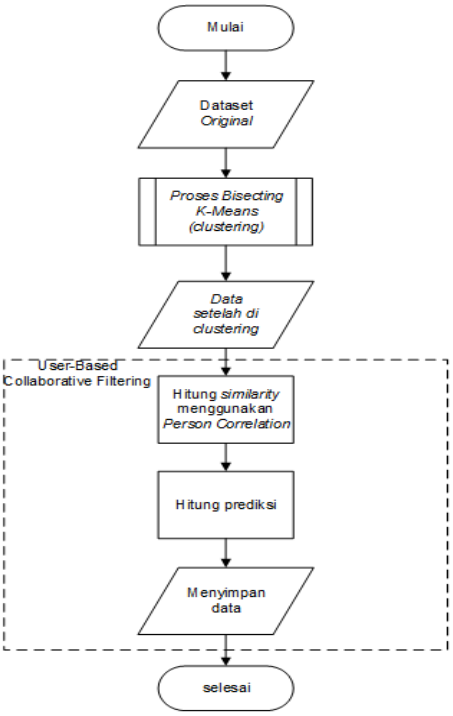
c. Kombinasi Bisecting K-Means dan Collaborative Filtering

Penelitian ini menggabungkan Bisecting K-Means dan Collaborative Filtering untuk mendapatkan hasil rekomendasi film. Gambar 2 menunjukkan gabungan bisecting K-Means dan item-based collaborative filtering.



Gambar 2. Flowchart bisecting K-Means dan item-based collaborative filtering

Gambar 3 menunjukkan gabungan bisecting K-Means dan user-based collabora-tive filtering.



Gambar 3. Flowchart bisecting K-Means dan user-based collaborative filtering

3. Perancangan Sistem

Perancangan yang dilakukan terdiri dari rancangan antar muka website sistem rekomendasi film dengan menggunakan balsamiq dan perancangan basis data dengan menggunakan Entity Relationship Diagram (ERD).

4. Implementasi

Sistem rekomendasi dibangun menggunakan bahasa pemrograman PHP, HTML dan CSS. Basis data berupa relational database yaitu MySQL.

5. Pengujian

Pengujian dilakukan dengan data set dari MovieLens. Untuk menguji sistem dilakukan dengan menguji hasil rekomendasi berdasarkan prediksi *rating*. Pengujian akan dihitung tingkat keakurasiannya menggunakan *Mean Absolute Error* (MAE) berdasarkan parameter *Neighbourhood Size* (NS), kemudian dilakukan analisis hasil. Nilai MAE dapat dihitung dengan menggunakan Persamaan 5.

$$MAE = \frac{\sum_{i=1}^{N} abs(p_i - r_i)}{N} \quad (5)$$

Keterangan:

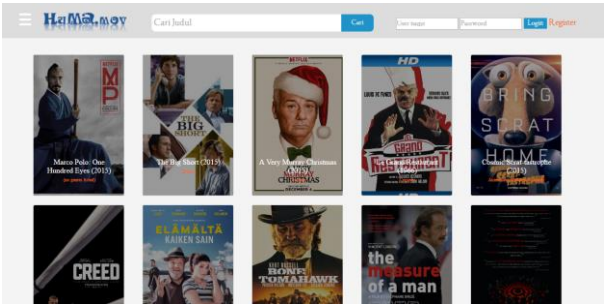
p_i = Nilai prediksi rating dari data ke-i

r_i = Nilai rating sebenarnya dari data ke-i

N = Jumlah data

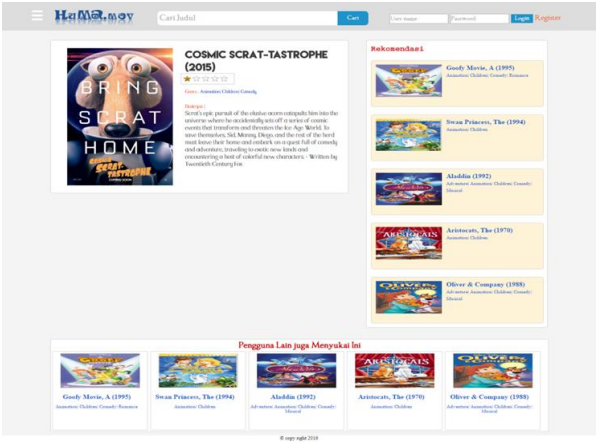
III. HASIL DAN PEMBAHASAN

Tampilan awal sistem rekomendasi film berbasis website yang akan dilihat oleh pengguna ditunjukkan pada Gambar 4.



Gambar 4. Tampilan Awal Website

Pada halaman ini, ditampilkan logo aplikasi, form login, link registrasi, kolom *search* yang dapat menampilkan beberapa judul film sesuai kata kunci yang dimasukkan pengguna dan menu daftar genre. Selain itu, terdapat beberapa film yang disertai dengan gambar, judul dan deskripsi singkat dari film tersebut dimana ketika dipilih salah satu akan menuju ke rincian film, seperti pada Gambar 5.



Gambar 5. Tampilan Web Rincian Film

Tampilan rincian film terdiri dari gambar film yang dipilih beserta judul, genre, deksripsi dan rating yang dapat diisi oleh member. Rating tidak akan bisa diisi oleh non-member. Sidebar kanan menampilkan 5 (lima) rekomendasi film berdasarkan genre dimana rekomendasi ini ditampilkan berdasarkan metode Item-based collaborative filtering dan pada sidebar bawah ditampilkan 5 (lima) rekomendasi film berdasarkan pengguna lain menyukai dimana rekomendasi ini ditampilkan berdasarkan metode User based collaborative filtering.

Gambar 6 menunjukkan hasil dari proses training dengan gabungan bisecting K-Means dan collaborative filtering untuk setiap film pada dataset.

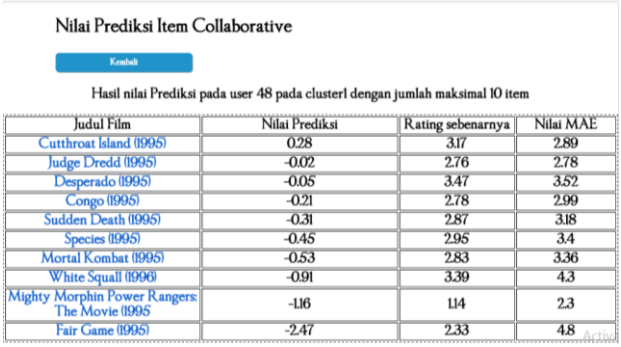
Menu Training Sistem

Kembali

iduser	Cluster	item collaborative	user collaborative
48	cluster1	82319, 15560, 3943, 5898, 26269, 62803, 132046, 2307, 371, 2698	2808, 2528, 741, 3375, 1373, 610, 3698, 3444, 2642, 2275
48	cluster2	32551, 7048, 5720, 6233, 47117, 3836, 5247, 4121, 8037, 7394	27193, 2471, 5247, 97836, 4121, 370, 367, 8531, 104241, 886
48	cluster5	109850, 4965, 5915, 27497, 62801, 107159, 72554, 4438, 74510, 26681	72554, 2410, 1047, 996, 390, 79224, 3275, 493, 315, 2411
48	cluster6	6923, 6924, 6925, 8773, 890, 73211, 200, 57520, 2856, 27784	60832, 50736, 1656, 1971, 457, 50, 43936, 5266, 47, 1061
48	cluster7	716, 68522, 98121, 128360, 6598, 5003, 6938, 60137, 13124, 6433	1771, 6667, 3142, 46347, 46850, 363, 27878, 1987, 1501, 3017
48	cluster8	95311, 58972, 97757, 27368, 4016, 60816, 7304, 8580, 65577, 72356	488, 87529, 3740, 258, 26776, 314, 4016, 2005, 4899, 197
48	cluster9	62970, 58364, 78967, 118696, 5797, 7389, 25793, 31950, 47493, 74089	70305, 2140, 2017, 2399, 5952, 2093, 4993, 2, 5816, 118696
48	cluster10	27800, 101962, 117444, 26403, 7099, 5146, 6093, 8015, 74553, 95167	2099, 2089, 7099, 551, 2142, 5146, 5882, 1881, 1030, 2119
48	cluster11	5, 3315, 2046, 6239, 2056, 61698, 362, 1017, 7743, 62376	362, 155, 1582, 2046, 986, 711, 828, 8, 1017, 62376
48	cluster12	136598, 7770, 3612, 2483, 3492, 49422, 2037, 2794, 5399, 95543	212, 3070, 2133, 231, 2470, 4247, 5596, 4734, 96281, 2794
48	cluster13	5092, 85780, 4777, 5978, 33090, 103655, 58347, 3397, 53993, 7106	7106, 74698, 1646, 3087, 1007, 2038, 1241, 2014, 1126, 2798

Gambar 6. Tampilan Website Hasil Training

Gambar 7 menunjukkan contoh hasil nilai prediksi dari user 48 pada cluster1 untuk 10 film pertama dengan gabungan bisecting k-means dan item-based collaborative filtering.



Nilai Prediksi Item Collaborative

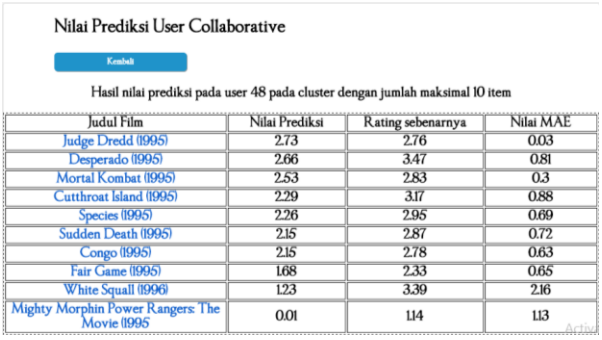
Kembali

Hasil nilai Prediksi pada user 48 pada cluster1 dengan jumlah maksimal 10 item

Judul Film	Nilai Prediksi	Rating sebenarnya	Nilai MAE
Cutthroat Island (1995)	0.28	3.17	2.89
Judge Dredd (1995)	-0.02	2.76	2.78
Desperado (1995)	-0.05	3.47	3.52
Congo (1995)	-0.21	2.78	2.99
Sudden Death (1995)	-0.31	2.87	3.18
Species (1995)	-0.45	2.95	3.4
Mortal Kombat (1995)	-0.53	2.83	3.36
White Squall (1996)	-0.91	3.39	4.3
Mighty Morphin Power Rangers: The Movie (1995	-1.16	1.14	2.3
Fair Game (1995)	-2.47	2.33	4.8

Gambar 7. Contoh Hasil Prediksi Bisecting K-Means dan Item-based CF

Gambar 8 menunjukkan contoh hasil nilai prediksi dari user 48 pada cluster1 untuk 10 film pertama dengan gabungan bisecting k-means dan user-based collaborative filtering.



Nilai Prediksi User Collaborative

Kembali

Hasil nilai prediksi pada user 48 pada cluster dengan jumlah maksimal 10 item

Judul Film	Nilai Prediksi	Rating sebenarnya	Nilai MAE
Judge Dredd (1995)	2.73	2.76	0.03
Desperado (1995)	2.66	3.47	0.81
Mortal Kombat (1995)	2.53	2.83	0.3
Cutthroat Island (1995)	2.29	3.17	0.88
Species (1995)	2.26	2.95	0.69
Sudden Death (1995)	2.15	2.87	0.72
Congo (1995)	2.15	2.78	0.63
Fair Game (1995)	1.68	2.33	0.65
White Squall (1996)	1.23	3.39	2.16
Mighty Morphin Power Rangers: The Movie (1995	0.01	1.14	1.13

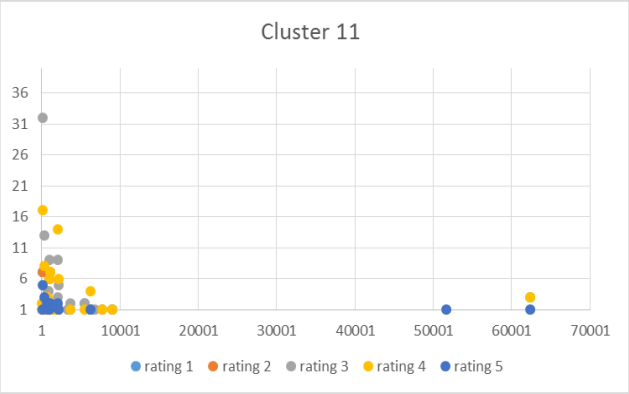
Gambar 8. Contoh hasil prediksi Bisecting K-Means dan user-based CF

Tabel 1 menunjukkan hasil perhitungan nilai MAE setelah proses testing untuk semua film pada setiap cluster terhadap 5 user dengan spesifikasi 3 user pernah melakukan rating dan 2 user yang belum pernah melakukan rating sama sekali. Jumlah cluster pada pengujian ini adalah 18 (sesuai dengan jumlah genre).

Tabel 1. Nilai MAE untuk setiap kluster

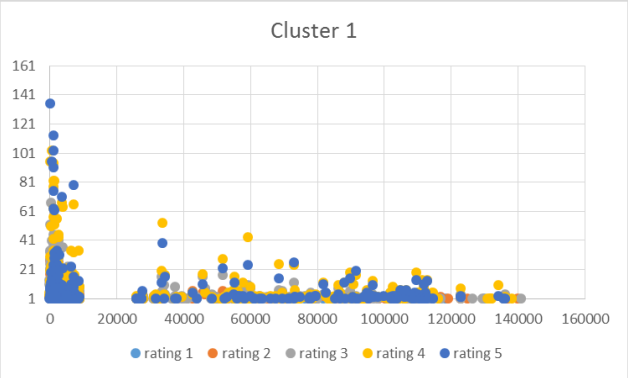
Cluster	*MAE Item-based*	*MAE User-based*
1	1.18	1.33
2	1.60	1.35
3	1.46	1.48
4	1.58	1.78
5	1.72	1.64
6	1.73	1.83
7	1.26	1.54
8	1.90	1.49
9	1.84	1.82
10	2.07	1.67
11	2.56	2.00
12	1.71	1.40
13	1.47	1.29
14	1.87	1.90
15	1.76	1.54
16	1.60	1.55
17	2.31	2.43
18	1.40	1.38
Rata-rata	1.72	1.63

Nilai rata-rata MAE untuk setiap cluster pada item-based lebih tinggi daripada user-based, berarti kombinasi bisecting K-Means dan user-based menghasilkan nilai rekomendasi yang lebih baik. Nilai rekomendasi terburuk terjadi pada cluster 11 dan 17 dibandingkan dengan cluster lainnya. Hal ini disebabkan oleh distribusi nilai rating user yang tidak baik, sehingga menghasilkan nilai prediksi yang tidak baik. Gambar 9 menunjukkan distribusi nilai rating pada dataset untuk cluster 11 yang tidak merata, sehingga nilai error pada sistem rekomendasi relatif lebih tinggi.



Gambar 9. Distribusi nilai rating yang tidak merata pada cluster 11

Gambar 10 menunjukkan distribusi nilai rating pada dataset untuk cluster 1 yang merata sehingga nilai error pada sistem rekomendasi relatif lebih rendah.



Gambar 10. Distribusi nilai rating yang merata pada cluster 1

IV KESIMPULAN

Penelitian ini telah menghasilkan sebuah sistem rekomendasi film berbasis website dengan menggunakan kombinasi algoritma bisecting K-Means dan Collaborative Filtering. Sistem rekomendasi berbasis website yang telah dikembangkan menggunakan informasi dari dataset MovieLens. Tingkat error pada sistem rekomendasi telah dihitung dengan menggunakan nilai MAE. Nilai rata-rata MAE kombinasi dari bisecting K-Means dan user-based CF adalah 1.63, lebih rendah dibandingkan dengan nilai rata-rata MAE kombinasi dari bisecting K-Means dan item-based. Selain metode rekomendasi, distribusi nilai rating pada dataset juga sangat mempengaruhi nilai MAE. Hal ini juga ditunjukkan pada cluster 11 dan 17 dengan distribusi nilai rating yang tidak merata, akan menghasilkan nilai error yang lebih tinggi pada sistem rekomendasi.

DAFTAR PUSTAKA

Garcia- Molina, Hector; Ullman, JD., & Widom, Jennifer. 2002. Database systems the complete book, International edition. New Jersey, Prentice Hall

Gupta, S. 2009, Collaboration Filtering using K-mean Algorithm, Vision Tech, In 6th National Level Technical Symposium, University of Rajiv Gandhi Proudyogiki Shwavidyalaya, Bhopal.

Mcginty, L., & Smyth, B., 2006, Adaptive selection : An analysis of critiquing and preference-based feedback in conversational recomender systems. International Journal of Electronic Commerce, volume 11, nomor 2, halaman 35-57

Patil, R., Ruchika, Khaan, A., et al, 2015, Bisecting K-Means For Clustering Web Log Data, Department of Computer Technology, Nagpur, India, International Journal of Computer Applications (0975 – 8887) Vol 116, 19 April 2015.

Ricci, F., Rokach, L., Shapira, B., Kantor, P.B., 2011, Recommender Systems Handbook.

Sarwar, B., Karypis, G., Konstan, J., & Rield, J., 2001, Item-based collaborative filtering recommendation algorithm. Proceedings of the 10th international conference on World Wide Web, (pp. 285-295)