STORAGE – Jurnal Ilmiah Teknik dan Ilmu Komputer, Vol. 5 No. 1, Februari 2026, 54 – 60
e-ISSN: 2828-5344 | DOI: 10.55123

PENINGKATAN SKALABILITAS SISTEM REKOMENDASI WEBSITE BERITA MENGGUNAKAN CONTENT-BASED FILTERING DAN K-MEANS

[1)]Muhamad Arldi Megantara, [2)]Ema Utami
[1,2)]Informatika, Pascasarjana, Universitas Amikom Yogyakarta
[1)]arldimegantara@students.amikom.ac.id

INFO ARTIKEL

Riwayat Artikel :
Diterima : 2 Februari 2026
Disetujui : 14 Februari 2026

Kata Kunci :
Content-Based Filtering, K-Means Clustering, Skalabilitas, *Mean Average Precision*, Sistem Rekomendasi Berita.

ABSTRAK

Perkembangan internet menghasilkan volume data yang besar termasuk website berita yang memunculkan tantangan pada model rekomendasi Content-based filtering (CBF) dengan kompleksitas komputasi linier O(N). Penelitian ini mengusulkan optimalisasi CBF dengan integrasi dengan algoritma K-Means yang akan menghasilkan partisi data dengan tujuan reduksi biaya komputasi. Digunakan 22.250 artikel berita dari antara news. Penentuan nilai cluster (K) optimal menggunakan elbow method dengan K terbaik K = 6 dan divalidasi dengan Silhouette Score dengan nilai 0,5201, yang mengindikasikan sebaran data baik. Hasil dari pengintegrasian CBF dan K-means menunjukan adanya efisiensi pada response time dan penggunaan memori hingga 6 kali lipat. Namun, ditemukan pula trade-off berupa penurunan Mean Average Precision (MAP) dari 0,89 (konvensional) menjadi 0,77 (K=6), yang masih baik pada kualitas rekomendasi. Selain itu, digunakan threshold dengan nilai 0,30 yang terbukti optimal dalam filter konten yang relevan. Kesimpulan yang dapat diambil dari penelitian adalah penggunaan partisi data mampu memberikan kualitas rekomendasi yang tetap andal namun pada beban komputasi yang ditekan.

ARTICLE INFO

Article History :
Received : Feb 2, 2026
Accepted : Feb 14, 2026

Keywords:
Content-Based Filtering, K-Means Clustering, Scalability, Mean Average Precision, News Recommendation System.

ABSTRACT

The development of the internet produces large volumes of data, including news websites, which pose challenges to the Content-based filtering (CBF) recommendation model with linear computational complexity of O(N). This study proposes CBF optimization by integrating with the K-Means algorithm that will produce data partitioning with the aim of reducing computational costs. 22,250 news articles from among the news were used. The optimal cluster value (K) was determined using the elbow method with the best K = 6 and validated with a Silhouette Score with a value of 0.5201, which indicates good data distribution. The results of the integration of CBF and K-means show efficiency in response time and memory usage up to 6 times. However, a trade-off was also found in the form of a decrease in Mean Average Precision (MAP) from 0.89 (conventional) to 0.77 (K = 6), which is still good for recommendation quality. In addition, a threshold value of 0.30 was used which was proven to be optimal in filtering relevant content. The conclusion that can be drawn from the research is that the use of data partitioning is able to provide reliable recommendation quality but with reduced computational load.

1. PENDAHULUAN

Internet telah mengalami berbagai perkembangan hingga hari ini. Seiring dengan perkembangan tersebut, data yang tercipta melalui internet mengalami peningkatan volume. Hal tersebut bukan hanya memberikan lebih banyak peluang yang dapat diambil, namun juga memunculkan tantangan terhadap pengguna seperti pemilihan konten yang paling sesuai. Oleh sebab itu muncul permasalahan mengenai pencarian serta rekomendasi berita yang relevan dengan pengguna yang makin sulit (Liu et al., 2021). Banyak perusahaan seperti Netflix, Amazon dan Youtube mengupayakan pengembangan metode rekomendasi yang dapat memberikan solusi kepada pengguna dalam pemilihan data dari dataset besar yang tersedia (Afoudi et al., 2021).

Content-based filtering menjadi salah satu algoritma yang digunakan karena kemampuannya dalam mengatasi permasalahan cold-start pada item baru (Isinkaye et al., 2015). Algoritma ini menghasilkan rekomendasi yang baik namun mengalami kendala pada skalabilitas (Roy & Dutta, 2022). Cara kerja algoritma didasarkan pada perbandingan item terhadap kelompok item pada dataset. Secara teknis, proses ini melalui kompleksitas komputasi linear O(N), yang mana waktu pemrosesan akan meningkat seiring bertambahnya jumlah item (N). Jika dataset yang dikelola pada volume yang besar maka akan dibutuhkan sumber daya lebih dalam menjalankan metode tersebut. Oleh karena itu, dalam penelitian ini diusulkan penggunaan cluster untuk membagi data kedalam kelompok data yang lebih kecil (Yang et al., 2021).

Walaupun pada penelitian terdahulu clustering dengan menggunakan K-Means telah diusulkan (Jan et al., 2022), terdapat beberapa hal yang belum dieksplorasi lebih. Penelitian-penelitian terdahulu fokus pada peningkatan presisi model, namun kurang mengevaluasi trade-off antara perubahan waktu serta presisi model akibat pembatasan kelompok data dalam pencarian rekomendasi. Selain itu, pengelompokan data teks yang pada penelitian ini menggunakan tags berita berpotensi menghasilkan silhouette score yang sangat rendah karena masalah sparsity, dan mempengaruhi kualitas dari rekomendasi.

Penelitian memiliki tujuan untuk meningkatkan skalabilitas Content-Based Filtering dengan mengintegrasikan K-means sebagai pemfilter dataset. Fokus utama yang dibahas adalah menyajikan efisiensi penggunaan sumber daya baik berupa waktu maupun memori serta menentukan ambang batas kemiripan yang optimal untuk menjaga keandalan model. Penelitian ini akan menjawab 2 pertanyaan yaitu

1. Seberapa signifikan selisih waktu yang dicapai melalui partisi data dengan K-means dibandingkan dengan metode konvensional.
2. Bagaimana pengaruh penggunaan K-means terhadap score kemiripan yang dihasilkan

Dengan mengoptimalkan penggunaan K-means yang tidak terlalu kompleks dan mudah diimplementasikan (Ikotun et al., 2023), penelitian ini bertujuan untuk menghasilkan metode rekomendasi yang bukan hanya andal dengan score kemiripan maksimal namun juga efisien.

2. METODE

Penelitian ini termasuk ke dalam penelitian eksperimental dengan tujuan mengevaluasi efisiensi model rekomendasi dengan 2 skenario yang dijalankan yaitu : content based filtering konvensional tanpa partisi data dan content based filtering yang dioptimalisasi dengan K-means.

2.1 Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset publik. Data diambil dari website kaggle yang berisi kumpulan berita dari antara news dengan 22.250 data berita yang digunakan. Terdapat 4 fitur pada dataset yang dapat dimanfaatkan yaitu title, publish_date, article_text, dan tag.

Fokus utama fitur yang digunakan dalam penelitian ini adalah article_text yang akan digunakan untuk pembanding kemiripan antar data lalu tag yang digunakan untuk menentukan partisi dan irisan data antar data.

Karena dataset yang digunakan dalam penelitian ini berupa teks sehingga perlu beberapa langkah sebelum data dapat diolah dengan metode K-means maupun content-based filtering. Tahapan pra-pemrosesan tersebut antara lain

1. Tokenisasi & Pembersihan: Menghapus karakter yang tidak dibutuhkan lalu

mengubah kata menjadi huruf kecil (Imbar et al., 2014).
2. Representasi Fitur: Menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) untuk memberikan bobot pada article_text.
3. Reduksi Dimensi: Menggunakan Truncated Singular Value Decomposition (SVD) untuk mengatasi sparsity pada data biner tag, sehingga meningkatkan performa pengelompokan (Kumbhar et al., 2020).

2.2 Pengelompokan Data (Clustering)

Fokus utama penelitian ini adalah meningkatkan skalabilitas dari content-based filtering dengan mempartisi dataset. Proses dalam pembuatan partisi dilakukan dengan memanfaatkan algoritma K-means. Penentuan jumlah cluster optimal menggunakan elbow method. Lalu untuk mengukur seberapa baik pembuatan cluster atau sebaran data dalam cluster digunakan Silhouette score (Naghizadeh & Metaxas, 2020). Fitur yang dipilih untuk pembuatan cluster adalah tag. Fitur tersebut berisi kata kunci yang menjadi topik dari berita. Karena fitur yang digunakan dalam clustering ini adalah tag maka dilakukan penyesuain ambang batas dari Silhouette score. Berdasarkan penelitian yang dilakukan oleh Kaufman dan Rousseeuw (Kaufman & Rousseeuw, 1990) ambang batas yang dapat digunakan sebagai acuan dengan fitur yang mirip dengan tag adalah di angka > 0,50. Angka tersebut mempertimbangkan separasi data yang terjadi pada tag.

2.3 Model Rekomendasi

Model rekomendasi yang diusulkan dalam penelitian ini mengintegrasikan hasil dari pengelompokan data ke algoritma content-based filtering. Hal tersebut yang membedakan metode konvensional dengan penelitian ini.

2.3.1. Metode pencarian

Secara teknis ada 4 tahapan yang dilalui dalam melakukan pencarian yaitu
- input data yang akan dicari kemiripannya
- identifikasi cluster dari data inputan
- mencari kemiripan dalam cluster terkait dengan perhitungan cosine similarity

2.3.2 Cosine similarity

Cosine similarity mengukur kemiripan antara 2 teks (Rinjeni et al., 2024). Dengan mengimplementasikan cosine similarity, teks diubah ke dalam bentuk vektor (Juni Permana & Agung Toto Wibowo, 2023). Proses mencari kemiripan digunakan term-frequency dari tiap kata pada teks. Hasil dari cosine similarity merupakan nilai dengan rentang antara 0-1. Jika mendekati 0 maka kedua teks semakin tidak mirip. Sebaliknya jika mendekati 1 maka teks semakin mirip (Sukestiyarno et al., 2023). Persamaan cosine tertampil sebagai berikut (Park et al., 2020) :

$$sim(A,B) = (A \cdot B) / (||A||\,||B||)$$

Keterangan :
- $sim(A,B)$: Nilai kemiripan antara dokumen kueri A dan dokumen target B
- $(A \cdot B)$: Hasil perkalian vektor A dengan B.
- $||A||$: panjang vektor A
- $||B||$: panjang vektor B

Atau jika lebih detail menjadi

$$sim(A,B) = \sum (Ai \times Bi) / (\sqrt{\sum Ai^2} \times \sqrt{\sum Bi^2})$$

Keterangan :
- i: Indeks untuk setiap fitur
- $\sum (Ai \times Bi)$: Hasil penjumlahan dari perkalian bobot fitur yang sama antara A dan B.
- $\sqrt{\sum Ai^2}$: Akar kuadrat dari jumlah kuadrat seluruh bobot fitur vektor A.
- $\sqrt{\sum Bi^2}$: Akar kuadrat dari jumlah kuadrat seluruh bobot fitur vektor B.

2.4 Evaluasi Model

Penelitian ini mengevaluasi performa model dengan mengacu pada 2 metrik utama berkaitan dengan tujuan penelitian ini dalam peningkatan skalabilitas. 2 matrik tersebut yaitu :
1. Efisiensi Komputasi dapat diukur dengan Response Time (Latency) dan penggunaan memori. Performa model diukur dengan menghitung waktu serta memori yang dibutuhkan model untuk memberikan

rekomendasi baik saat model menggunakan cluster atau tanpa cluster.

2. Kualitas Rekomendasi (Relevansi) Dataset yang digunakan pada penelitian ini tidak memiliki ground truth. Oleh sebab itu, pengukuran kualitas model yang dievaluasi dapat menggunakan similarity score dengan nilai tertinggi lalu dikomparasikan dengan threshold. Fungsi dari threshold adalah untuk membedakan antara rekomendasi sesuai atau tidak sesuai. Dalam penentuan nilai threshold yang diambil, beberapa penelitian digunakan sebagai rujukan.

3. HASIL DAN PEMBAHASAN

3.1 analisis skalabilitas dan biaya komputasi

Masalah yang dihadapi dalam penerapan content-based filtering adalah saat pertumbuhan data menyebabkan beban biaya komputasi yang meningkat. Hal tersebut terjadi karena model bersifat linier terhadap jumlah item (N) dengan membandingkan suatu item terhadap seluruh item pada dataset. Namun penggunaan K-means memberikan cara baru rekomendasi dari yang melakukan pencarian konvensional ke seluruh data pada dataset menjadi pencarian terfragmentasi. Dengan membagi N item ke sejumlah cluster maka akan terjadi reduksi beban kerja. Secara matematis dengan jumlah dataset yang digunakan yaitu 22.500 data maka dengan operasi konvensional model akan melakukan 22.500 kali operasi perbandingan. Sedangkan jika menggunakan cluster dengan jumlah k adalah 6 maka akan didapat rata-rata 3.750 item setiap cluster.

3.2 Eksperimen

Percobaan dilakukan dengan 4 skala dataset yaitu 10.000, 15.000, 20.000 serta 22.500 data. Terdapat 2 indikator yang menjadi acuan yaitu response time dan penggunaan memori. Semakin sedikit response time dan memori maka metode bekerja lebih efisien.

Karena dalam penelitian diusulkan K-means sebagai metode untuk pembuatan partisi data, maka perlu dilakukan pengujian terhadap hasil clustering yang dilakukan. Pada tahap awal, penentuan jumlah cluster (K) optimal yang menggunakan elbow method menjadi kunci untuk mendapatkan hasil clustering terbaik. Didapatkan nilai silhouette score 0,5201 pada K = 6 sebagai K optimal yang menurut teori Kaufman & Rousseeuw masuk dalam kategori reasonable structure (sebaran data yang baik).

Tabel 3.1 Perbandingan Performa Model

Jumlah Data	MetrikCBF	Konvensional	CBF + (K6)	CBF + (K3)
10.000	Response Time (s)	0.067	0.009	0.020
	Memory (KB)	1255.889	155.545	418.631
15.000	Response Time (s)	0.102	0.016	0.031
	Memory (KB)	1935.612	268.283	685.624
20.000	Response Time (s)	0.100	0.022	0.046
	Memory (KB)	2630.542	391.902	950.616
22.250	Response Time (s)	0.147	0.025	0.050
	Memory (KB)	2973.890	455.591	1083.701

3.3. Penanganan Cold-Start

Penanganan cold-start sangat umum ditemui ketika berhadapan dengan pengguna baru. Hal tersebut dikarenakan user baru belum memiliki data historis mengenai minat atau profil pengguna. Metode yang diusulkan pada penelitian ini tidak membutuhkan data dari pengguna sebelumnya. Metode bekerja dengan memberikan rekomendasi terhadap data yang saat itu dipilih yang selanjutnya akan dicari kemiripannya terhadap keseluruhan atau sebagian dataset. Oleh sebab itu metode yang diusulkan bisa diterapkan dalam memberikan rekomendasi untuk pengguna baru.

3.4. Evaluasi Kualitas Rekomendasi (MAP)

Menambahkan partisi data dalam merekomendasi akan menimbulkan probabilitas penurunan performa karena item yang seharusnya direkomendasikan tidak terdapat pada cluster yang sama dengan data uji.

Memperhatikan hal tersebut maka kualitas clustering sangat mempengaruhi performa model agar tetap sama atau mendekati model secara konvensional.

Tabel 3.2 Perbandingan Nilai Mean Average Precision (MAP)

Skenario Model	Jumlah Cluster (K)	Mean Average Precision (MAP)	Selisih terhadap Konvensional
CBF Konvensional	-	0.89	-
CBF + K-Means	K = 3	0.84	0.05
CBF + K-Means	K = 6	0.77	0.12

Pada k=6 model mampu untuk memberikan rekomendasi dengan MAP 0,77 atau turun 0,12 dari hasil yang didapatkan ketika menggunakan metode konvensional. Penurunan tersebut adalah trade-off yang wajar dan menguntungkan untuk penurunan waktu serta memori yang cukup signifikan hingga 6 kali lipat atau sesuai dengan jumlah cluster yang digunakan.

Namun perlu menjadi catatan ketika penggunaan MAP dalam penentuan ground truth merupakan tindakan komparatif dan bukan absolut. Penggunaan MAP didorong karena ketiadaan ground truth pada dataset. Selain itu, harus diakui bahwa penentuan ground truth pada penelitian ini merepresentasikan kemiripan item, bukan kepuasan dari pengguna secara langsung.

3.5. Analisis Threshold Optimal

Dalam menjaga kualitas dari model, dilakukan pengujian dalam pencarian nilai ambang batas (threshold) dengan nilai kemiripan pada nilai antara 0,05 hingga 1,00. Threshold pada penelitian ini digunakan sebagai alat untuk menyaring apakah rekomendasi yang dihasilkan model layak atau justru tidak relevan.

Tabel 3.3 Perbandingan Nilai Threshold

Threshold	Mean Precision	Status
0,05 - 0,20	Rendah	Noise
0,30	Tinggi	Optimal
0,40 - 0,60	Sangat Tinggi	Strict
0,70 - 1,00	Maksimal	Identical

Penggunaan threshold di angka 0,30 mampu memberikan hasil yang baik walaupun terjadi penurunan MAP akibat clustering. Oleh sebab itu angka tersebut dipilih untuk menghasilkan model yang akurat dengan ruang pencarian yang diperkecil.

4. PENUTUP

4.1. Kesimpulan

Penggunaan K-means terbukti secara efektif dan efisien memberikan hasil peningkatan skalabilitas dari model content-based filtering. Implementasi clustering dengan titik optimal cluster K = 6 mampu mengurangi beban komputasi baik secara response time maupun penggunaan memori hingga 6 kali lipat dari metode konvensional. Hal tersebut terbukti dari waktu respons tereduksi dari 0,1472 detik menjadi 0,0248 detik, serta penggunaan memori berkurang dari 2973,89 KB menjadi 455,59 KB. Hubungan antara penggunaan clustering yang berimplikasi pada partisi dataset mengakibatkan potensi penurunan kualitas dari rekomendasi, dan dalam penelitian ini ditemukan trade-off antara kecepatan komputasi dan kualitas rekomendasi. Penggunaan K = 6 mengakibatkan penurunan MAP dari metode konvensional di angka 0,12. Dengan menggunakan metode konvensional didapatkan MAP dengan nilai 0,89 sedangkan metode clustering dengan K = 6 mendapatkan MAP 0,77 yang mana masih tergolong pada nilai yang ideal. Namun, penurunan yang terjadi terbayarkan dengan reduksi response time serta penggunaan memori. Penentuan similarity threshold pada angka 0,30 merupakan titik optimal dalam menjaga relevansi rekomendasi.

Baik dalam penentuan jumlah cluster (K) ataupun ambang batas (threshold) terdapat trade-off yang harus diterima. Jumlah cluster yang besar akan menguntungkan dalam segi

pemangkasan beban komputasi namun akan menurunkan presisi dari model. Sedangkan nilai ambang batas yang terlalu tinggi akan menyebabkan model tidak menghasilkan rekomendasi atau justru jika terlalu rendah berpotensi menghasilkan rekomendasi yang tidak relevan. Hasil yang didapat pada penelitian ini tidak bersifat absolut. Penurunan beban komputasi ataupun nilai presisi bergantung pada toleransi dalam kehilangan salah satu sisi, biaya komputasi atau relevansi hasil.

4.2 Saran

Meskipun penelitian ini menemukan tujuan awalnya, ada beberapa hal yang bisa menjadi peluang dalam pengembangan di masa depan antara lain

Penggunaan Algoritma Clustering Lain: Terdapat banyak algoritma clustering lain yang dapat dieksplorasi seperti Bisecting K-Means atau DBSCAN yang berpotensi dapat meningkatkan presisi dari model yang diusulkan.

Pendekatan Hybrid: Melakukan peningkatan model dengan melakukan penggabungan dengan algoritma Collaborative Filtering (pendekatan hybrid) untuk memperkaya relevansi rekomendasi berdasarkan perilaku pengguna. Dengan penggabungan tersebut, terdapat potensi model akan bisa menangani rekomendasi berdasarkan data historis dari pengguna sekaligus menangani cold-start pada pengguna baru.

5. DAFTAR PUSTAKA

Afoudi, Y., Lazaar, M., & Al Achhab, M. (2021). Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. Simulation Modelling Practice and Theory, 113, 102375. https://doi.org/10.1016/j.simpat.2021.102375

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Imbar, R. V., Adelia, Ayub, M., & Rehatta, A. (2014). Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16(3), 261–273. https://doi.org/10.1016/j.eij.2015.06.005

Jan, M., Khusro, S., Alam, I., Khan, I., & Niazi, B. (2022). Interest-Based Content Clustering for Enhancing Searching and Recommendations on Smart TV. Wireless Communications and Mobile Computing, 2022, 1–14. https://doi.org/10.1155/2022/3896840

Juni Permana, A. H. J. P. & Agung Toto Wibowo. (2023). Movie Recommendation System Based on Synopsis Using Content-Based Filtering with TF-IDF and Cosine Similarity. International Journal on Information and Communication Technology (IJoICT), 9(2), 1–14. https://doi.org/10.21108/ijoict.v9i2.747

Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis (1st ed.). Wiley. https://doi.org/10.1002/9780470316801

Kumbhar, R., Mhamane, S., Patil, H., Patil, S., & Kale, S. (2020). Text Document Clustering Using K-means Algorithm with Dimension Reduction Techniques. 2020 5th International Conference on Communication and Electronics Systems (ICCES), 1222–1228. https://doi.org/10.1109/ICCES48766.2020.9137928

Liu, J., Song, J., Li, C., Zhu, X., & Deng, R. (2021). A Hybrid News Recommendation Algorithm Based On K-means Clustering and Collaborative Filtering. Journal of Physics: Conference Series, 1881(3), Article 3. https://doi.org/10.1088/1742-6596/1881/3/032050

Naghizadeh, A., & Metaxas, D. N. (2020). Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means. Procedia Computer Science, 176, 205–214. https://doi.org/10.1016/j.procs.2020.08.022

Park, K., Hong, J. S., & Kim, W. (2020). A Methodology Combining Cosine Similarity with Classifier for Text Classification. Applied Artificial Intelligence, 34(5), 396–411. https://doi.org/10.1080/08839514.2020.1723868

Rinjeni, T. P., Indriawan, A., & Rakhmawati, N. A. (2024). Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm. Procedia Computer Science, 234, 553–560. https://doi.org/10.1016/j.procs.2024.03.039

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. Journal of Big Data, 9(1), 59. https://doi.org/10.1186/s40537-022-00592-5

Sukestiyarno, Y. L., Sapolo, H. A., & Sofyan, H. (2023). Application of Recommendation System on E-Learning Platform Using Content-Based Filtering with Jaccard Similarity and Cosine Similarity Algorithms. Computer Science and Mathematics. https://doi.org/10.20944/preprints202306.1672.v1

Yang, Y., Yao, H., Li, R., & Wang, S. (2021). A collaborative filtering recommendation algorithm based on user clustering with preference types. Journal of Physics: Conference Series, 1848(1), Article 1. https://doi.org/10.1088/1742-6596/1848/1/012043