Pemanfaatan Clustering dalam Pencarian Kemiripan Dokumen *Paper Conference*

Yan Puspitarani
Jurusan Teknik Informatika
Universitas Widyatama
Bandung, Indonesia
yan.puspitarani@widyatama.ac.id

Abstrak— Banyaknya penyimpanan informasi di Internet sangat membantu para penulis dalam menghasilkan karya tulis ilmiah. Penulisan karya ilmiah ini biasa dimanfaatkan kalangan akademik dalam kegiatan paper conference atau sebagai tugas kuliah bagi mahasiswa. Hal ini membuat pemeriksa kesulitan dalam memeriksa keunikan karya tulis yang dihasilkan. Pencarian kemiripan dokumen menjadi salah satu solusi yang dapat digunakan. Sehubungan dengan hal tersebut, proses clustering dalam text mining dapat dimanfaatkan untuk pencarian kemiripan dokumen agar lebih efektif.

Pada penelitian ini, dibuktikan dua buah hipotesis dalam pencarian kemiripan dokumen dan menghasilkan solusi pemanfaatan pencarian kemiripan dokumen-dokumen berbahasa Indonesia. Selain itu, akan dibuktikan pula hasil K-Means clustering dengan pemilihan feature terhadap isi dokumen berdasarkan judul, abstrak, pendahuluan, penutup, dan daftar pustaka, dapat lebih baik dibandingkan dengan hasil clustering biasa. Prototipe aplikasi pun dibangun untuk membuktikan hipotesis tersebut.

Hasil pengujian pada penelitian ini menunjukkan bahwa pemilihan feature untuk clustering menghasilkan akurasi yang paling tinggi, yaitu mencapai nilai 0.96. Selain itu, dibuktikan pula gap perhitungan waktu pencarian yang cukup besar antara pencarian terhadap dokumen ter-cluster dengan dokumen tanpa cluster.

Keywords—kemiripan dokumen; K-Means clustering; text mining

I. PENDAHULUAN

Information Retrieval sering dimanfaatkan untuk pencarian dokumen dengan tingkat kemiripan sangat tinggi. Ada beberapa metode yang dapat diterapkan dalam mencari kemiripan dokumen, seperti *Fuzzy C-Means*, *Vector Space Model* [2,7], *tree distance*[9], *manifold-ranking of blocks*[13], dan sebagainya. Diantara metode-metode tersebut, *Vector space model* merupakan metode pendekatan statistic yang paling banyak digunakan untuk mencari kemiripan dokumen[8].

Beberapa penelitian yang sudah dilakukan terkait hal ini adalah dengan menggabungkan teknik *clustering* yang ada pada text mining dalam proses *information retrieval*, seperti Scatter-Gather, Collection *clustering*, Language Modelling [11], dan Lingo[12]. Penelitian yang dilakukan Marti A. Hearst[6] dan Stanislaw Osi´ski[12] dengan Lingo-nya melakukan *clustering* terhadap hasil pencarian artikel, sedangkan Herny Februariyanti hanya menggunakan abstrak skripsi mahasiswa sebagai dataset dan tidak memanfaatkan *clustering* dalam prototipe yang dia buat[3]. Hasil penelitian-penelitian tersebut menunjukkan bahwa *cluster* pada *information retrieval* sangat berpengaruh terhadap performansi pencarian.

Besarnya dimensi yang dihasilkan dari dokumen, mempengaruhi performansi pencarian, sehingga pemilihan *feature* menggunakan berbagai teknik *feature selection* pun diperlukan. Kebanyakan teknik menggunakan pendekatan statistik yang masih memerlukan tambahan proses dalam alur proses pencarian, sehingga hal ini pun akan mempengaruhi performansi. Sementara itu, belum ada penelitian yang menggunakan pemilihan *feature* isi dokumen melalui intuisi manusia. Biasanya manusia memanfaatkan hal-hal yang singkat seperti judul, abstrak, pendahuluan, kesimpulan, dan daftar pustaka dalam mencari dokumen yang mereka inginkan, agar proses pengelompokkannya lebih cepat. Jika dikaitkan dengan *clustering* dan ukuran dimensi, penggunaan intuisi ini perlu diteliti karena sesuai dengan kebutuhan proses *clustering* dimana ukuran dimensi data menentukan akurasi hasil *clustering*.

Oleh karena itu, penelitian ini dilakukan dengan tujuan untuk membuktikan dua buah hipotesis, yaitu:

1. Akurasi hasil *clustering* pada dokumen yang hanya berisi Judul, Abstrak, Pendahuluan, Kesimpulan dan Daftar Pustaka lebih baik dibandingkan dengan hasil *clustering* pada dokumen yang isinya lengkap, selanjutnya akan disebut sebagai hipotesis 1.
2. Pencarian dokumen terhadap dokumen ter-*cluster* lebih efektif dibandingkan dengan pencarian terhadap dokumen tanpa *cluster*, selanjutnya akan disebut sebagai hipotesis 2.

II. TEXT CLUSTERING DAN INFORMATION RETRIEVAL

Clustering adalah proses pengelompokan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan tujuan untuk mengelompokkannya ke dalam *cluster* yang sama jika objek tersebut memiliki kimiripan satu sama lain. Sedangkan objek yang berbeda akan dimasukkan ke dalam *cluster* yang berbeda pula. Objek yang dimaksud dalam hal ini adalah dokumen. Dengan kata lain, dokumen-dokumen dalam suatu *cluster* harus semirip mungkin dan dokumen-dokumen pada sebuah *cluster* harus tidak mirip sama sekali dengan dokumen-dokumen pada *cluster* yang lain [5].

Information retrieval adalah pencarian materi, yang biasanya berupa dokumen teks yang tidak terstruktur yang memenuhi kebutuhan informasi dari sekumpulan besar dokumen yang disimpan [11].

Berdasarkan hipotesis mengenai *clustering* yaitu, dokumen yang berada pada *cluster* yang sama akan berperilaku sama terhadap relevansinya dengan kebutuhan informasi. Maksud dari hipotesis tersebut jika dimanfaatkan untuk proses pencarian dalam *information retrieval* adalah jika ada sebuah dokumen dari sebuah *cluster* yang relevan dengan *search request*, maka ada kemungkinan dokumen-dokumen lain dalam *cluster* tersebut juga relevan [11]. Beberapa aplikasi yang didasari hipotesis tersebut adalah *Search result clustering*, Scatter-Gather, *Collection clustering*, Language modeling, dan Cluster-based retrieval [11]. Kelima aplikasi tersebut dibedakan berdasarkan data apa yang mereka *cluster* dan teknik seperti apa yang mereka gunakan. Keseluruhannya dilakukan untuk memperbaiki interaksi dengan user, perbaikan efektifitas, efisiensi, dan akurasi hasil pencarian.

A. Vector Space Model

Vector Space Model merupakan representasi sekumpulan dokumen sebagai vektor dimana $\vec{V}(d)$ sebagai notasi vektor dokumen *d*[11]. Hasil dari preproses dokumen yang menghasilkan *term* dan frekuensinya digunakan sebagai pemodelan vektor,

$$\vec{V}(d) = (x_1, x_2, x_3, \cdots, x_n) \quad (1)$$

dimana $x_1, x_2, x_3, \cdots, x_n$ merupakan frekuensi *term* terhadap dokumen *d* [11].

B. Euclidean Length

Jika $\vec{V}(d)$ merupakan vektor dokumen *d* dengan *n* komponen $x_1, \cdots, x_n$, maka Euclidean length *d* adalah[11]

$$\left|\vec{V}(d)\right| = \sqrt{\sum_{i=1}^{n} x_i^{\,2}} \quad (2)$$

Tujuan digunakannya eulidean length adalah untuk menormalisasi panjang setiap vektor agar seimbang [11]. Sedangkan perhitungan normalisasi vektor berdasarkan euclidean length dilakukan menggunakan persamaan berikut [11].

$$\vec{v}(d) = \frac{\vec{V}(d)}{\left|\vec{V}(d)\right|} \quad (3)$$

C. Cosine Similarity

Untuk mengukur kemiripan antara dua dokumen pada vector space adalah mengukur jarak vektor diantara kedua dokumen tersebut. Akan tetapi, perbedaan panjang vektor di setiap dokumen menjadi kendala. Oleh karena itu, cara standar untuk mengukur kemiripan antara *d*1 dan *d*2 adalah dengan menghitung *cosine similarity* antara $\vec{V}(d1)$ dan $\vec{V}(d2)$ [11].

$$sim(d_1, d_2) = \frac{\vec{V}(d_1) \bullet \vec{V}(d_2)}{\left|\vec{V}(d_1)\right|\left|\vec{V}(d_2)\right|}, \quad (4)$$

dimana pembilangnya berupa dot product dan penyebutnya berupa Euclidean lengths [11].

Jika setiap vektor dihitung normalisasinya dengan

$$\vec{v}(d_1) = \frac{\vec{V}(d_1)}{\left|\vec{V}(d_1)\right|} \quad (5)$$

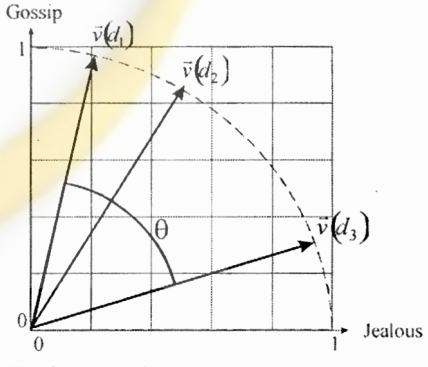
dan

$$\vec{v}(d_2) = \frac{\vec{V}(d_2)}{\left|\vec{V}(d_2)\right|}, \quad (6)$$

maka persamaan *cosine similarity* dapat disederhanakan menjadi [11]

$$sim(d_1, d_2) = \vec{v}(d_1) \bullet \vec{v}(d_2) \quad (7)$$

Berikut ini adalah gambaran *cosine similarity vector* untuk setiap dokumen dengan panjang yang telah dinormalisasi [11].



Gambar II-1 *Cosine similarity Vector*

(Digambar ulang dari [11])

Dalam proses pencarian dokumen menggunakan *cosine similarity*, maka setiap dokumen *d1...di* akan dicari kemiripannya dengan query dokumen *d* berdasarkan nilai *cosine similarity* terbesar [11].

Kumpulan *term* dan dokumen disimpan dalam bentuk matriks *M* x *N*, dimana *M* merupakan *term* dan *N* menunjukkan dokumen[11].

D. K-Means

K-Means merupakan salah satu algoritma *clustering* yang mudah untuk diimplementasikan, sederhana dan memiliki kompleksitas waktu yang linear. Pada algoritma ini, setiap *cluster* dihubungkan dengan *centroid* (center point) dan setiap point (dalam hal ini dokumen) dihubungkan dengan *centroid cluster* yang paling dekat. Agoritma *K-Means* dapat dijelaskan sebagai berikut [5]:

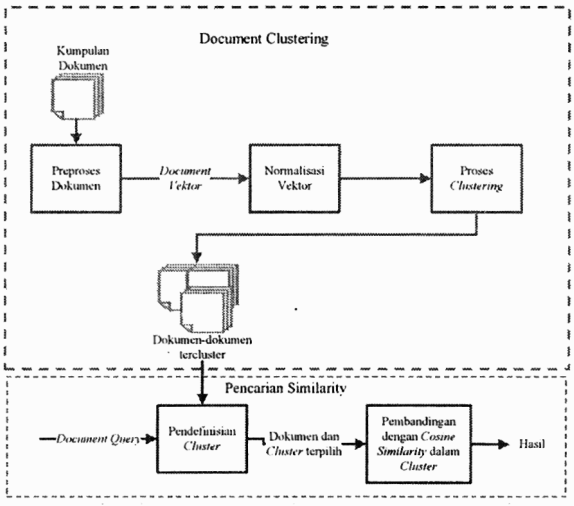
1. Tentukan nilai *k* sebagai jumlah *cluster* yang ingin dibentuk,
2. Bangkitkan *k centroid* (titik pusat *cluster*) awal secara random,
3. Hitung jarak setiap data ke masing-masing *centroid* berdasarkan ukuran kedekatan,
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroid*nya,
5. Tentukan posisi *centroid* baru dengan cara menghitung nilai rata-rata dari data-data yang ada pada *centroid* yang sama,
6. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.

E. Evaluasi Cluster

Pengukuran kualitas hasil *cluster* memerlukan *human judgement* yang memiliki level subjektifitas yang tinggi [4]. Alat ukur yang paling umum digunakan dengan pendekatan ini adalah *purity*. Misalkan $\{L_1, L_2, \cdots, L_n\}$ merupakan *cluster-cluster* dokumen yang diberi label secara manual dan $\{C_1, C_2, \cdots, C_m\}$ merupakan *cluster-cluster* hasil proses *clustering*, maka pengukuran nilai *purity* menggunakan persamaan berikut ini [4].

$$Purity(C_i) = \frac{\max_j |L_j \cap C_i|}{|C_i|} \quad (8)$$

III. PERANCANGAN SISTEM



Gambar III-1 Sistem Pencarian Kemiripan Dokumen

Sistem pencarian dibuat menjadi dua bagian utama, yaitu:

A. Document Clustering

Bagian ini merupakan penerapan text mining menggunakan *clustering*. Pada bagian ini, dilakukan tahapan persiapan dokumen melalui preproses kemudian dilanjutkan dengan proses *clustering*. Bagian ini akan menghasilkan kumpulan dokumen-dokumen yang telah ter-*cluster*.

B. Pencarian Similarity

Bagian ini merupakan tahap pencarían dokumen berdasarkan *query* yang berupa dokumen pula. Pada tahap ini, diperlukan *centroid* dari dokumen-dokumen yang ter-*cluster* sebagai lokasi pencarían. Hasil dari tahap ini adalah dokumen-dokumen yang relevan dengan *query*, yaitu dokumen dalam satu *cluster* yang memiliki kemiripan dengan *query*.

IV. PENGUJIAN

A. Skenario Pengujian

Pengujian sistem dilakukan menggunakan prototipe aplikasi sederhana. Kriteria yang akan dianalisis dalam pengujian sistem, yaitu:

1. Pengukuran akurasi hasil pengelompokkan dokumen (*clustering*) sesuai kemiripan antara dokumen yang satu dengan yang lainnya, dan
2. Pengukuran waktu pencarian dan ketepatan hasil pencarian.

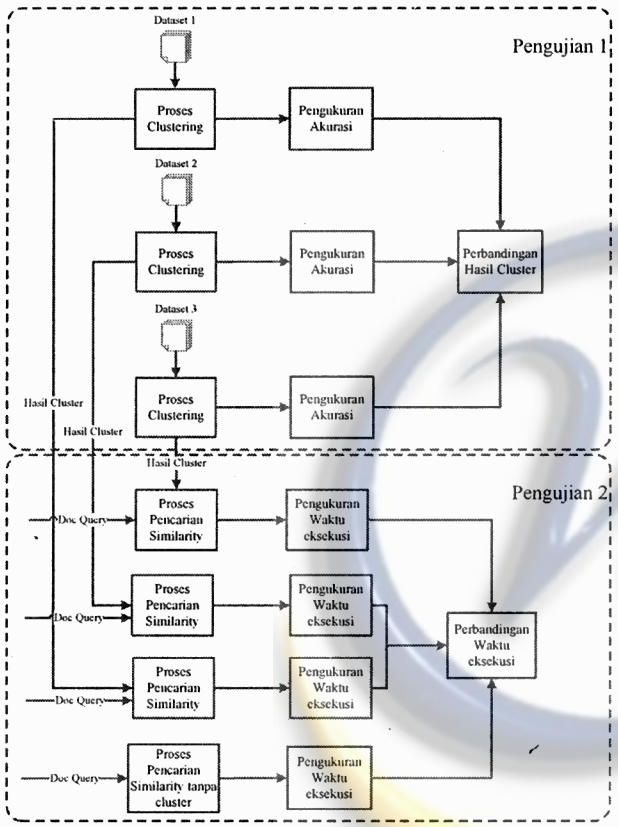
Untuk menguji kedua kriteria tersebut, dataset yang digunakan, yaitu:

1. Dataset 1, berisi dokumen paper utuh,
2. Dataset 2, berisi dokumen paper yang hanya terdiri dari judul, abstrak, pendahuluan, kesimpulan, dan daftar pustaka, dan
3. Dataset 3, berisi dokumen paper yang terdiri dari judul dan abstrak.

Ketiga jenis dataset tersebut berjumlah 50 dokumen paper dari berbagai sumber dengan 5 kategori. Komposisi jumlah dokumen untuk setiap kategori adalah sebagai berikut.

1. Kategori A berjumlah 6 dokumen,
2. Kategori B berjumlah 7 dokumen,
3. Kategori C berjumlah 12 dokumen,
4. Kategori D berjumlah 17 dokumen, dan
   Kategori E berjumlah 8 dokumen.

Ada dua skenario pengujian. Pengujian tersebut digambarkan melalui diagram blok pada Gambar IV-1.



Gambar IV-1 Skenario pengujian Sistem

Preproses terhadap semua dataset dilakukan sebelum pengujian. Preproses ini pun menghasilkan ukuran dimensi setiap dataset sebagai berikut.

1. Dataset 1 menghasilkan 9643 *term*,
2. Dataset 2 menghasilkan 5526 *term*, dan
3. Dataset 3 menghasilkan 1613 *term*.

B. Hasil Pengujian

Pada pengujian 1, dilakukan perhitungan akurasi terhadap hasil *clustering* dengan enam percobaan karena inisialisasi *centroid* dilakukan secara acak setiap eksekusi. Berikut ini adalah hasil pengujian 1:

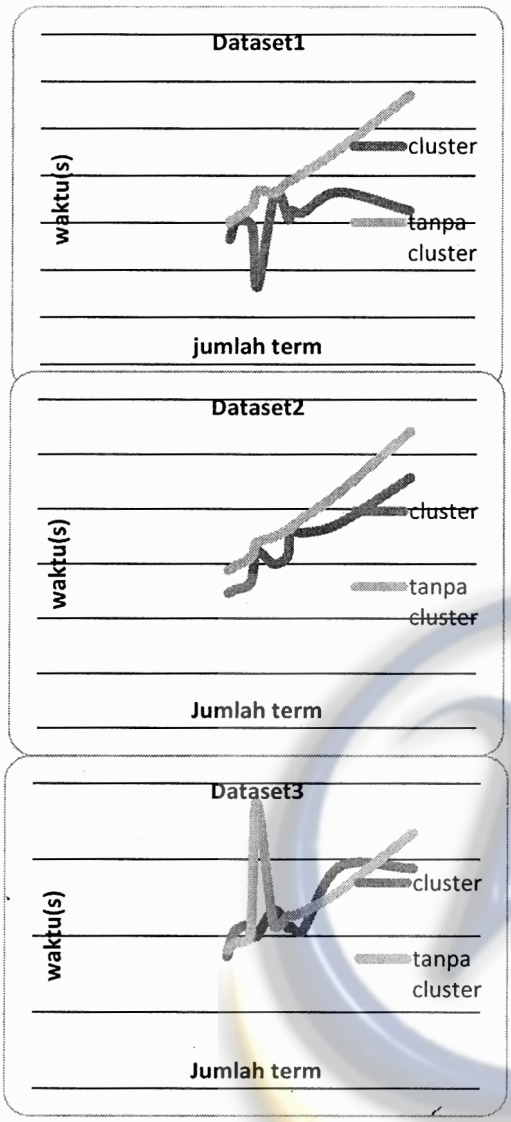
TABEL IV-1 AKURASI HASIL *CLUSTERING*

Percobaan	Akurasi		
	Dataset 1	Dataset 2	Dataset 3
1	0.82	0.86	0.86
2	0.86	0.76	0.68
3	0.7	0.78	0.74
4	0.8	0.96	0.62
5	0.7	0.82	0.52
6	0.78	0.86	0.72
rata-rata	**0.78**	**0.84**	**0.70**

Algoritma *clustering* dengan *K-Means* hanya memperhatikan nilai *term* yang berkaitan dengan dokumen. Nilai *term* tersebut menentukan seberapa penting informasi yang dimiliki suatu dokumen. Oleh karena itu, akurasi tidak akan mencapai nilai terbaik jika dimensi datanya sangat besar atau sangat kecil. Berdasarkan Tabel IV-1, terlihat bahwa akurasi hasil *clustering* pada setiap percobaan memiliki gap yang cukup besar. Hal ini dapat diakibatkan oleh inisialisasi *centroid* yang dilakukan secara random menghasilkan *cluster* yang kurang tepat. Akan tetapi, rata-rata akurasi hasil *clustering* pada dataset 2 paling tinggi. Bahkan, pada percobaan ke-4, nilai akurasi tertinggi dicapai oleh hasil *clustering* pada dataset 2 dengan nilai 0.96.

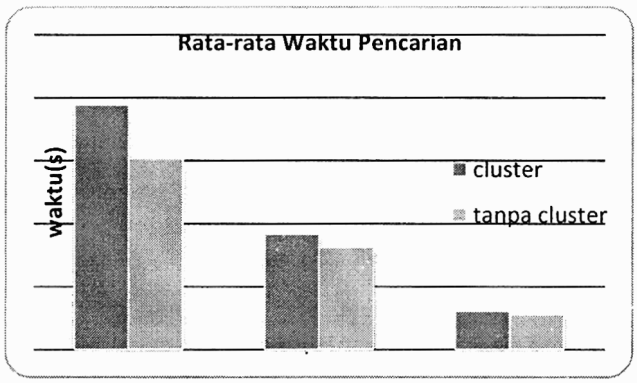
Hal ini menunjukkan bahwa hasil *clustering* dengan pemilihan *feature* terhadap dokumen berupa judul, abstrak, pendahuluan, kesimpulan dan daftar pustaka lebih baik dibandingkan proses *clustering* terhadap dokumen utuh dan abstrak dokumen.

Pada pengujian 2, dilakukan perhitungan waktu eksekusi terhadap pencarian dokumen yang memiliki nilai similaritas tinggi dengan *query* dokumen. *Cluster* yang digunakan untuk setiap dataset dipilih berdasarkan hasil *cluster* yang menghasilkan akurasi paling tinggi. Ada 10 *query* dokumen yang digunakan sebagai test set. Berikut ini hasil perbandingan waktu eksekusi untuk masing-masing berdasarkan jumlah *term* test set antara dataset ter-*cluster* dan tanpa *cluster*.



Gambar IV-2 Grafik Perbandingan Waktu Eksekusi setiap Dataset

Berdasarkan Gambar IV-2, hampir semua pencarian dokumen berdasarkan test set terhadap dataset ter-*cluster* menghasilkan performa yang lebih baik. Dari grafik pun terlihat bahwa gap waktu pencarian pada dataset 1 jauh lebih besar. Hal ini dapat diakibatkan oleh ukuran dimensi dari dataset 1 yang besar. Gap yang semakin besar pun diperlihatkan oleh dataset 2 dan dataset 3 seiring ukuran dimensi test set yang semakin besar. Berdasarkan hasil tersebut, dapat dikatakan bahwa waktu eksekusi pencarian menggunakan *cluster* akan lebih baik pada dataset dan test set berdimensi besar.



Gambar IV-3 Grafik Rata-rata Waktu Pencarian setiap Dataset

Berdasarkan Gambar IV-3, dapat diketahui bahwa rata-rata waktu eksekusi untuk semua test set pada semua dataset ter-*cluster* lebih baik daripada waktu eksekusi terhadap dataset tanpa *cluster*. Jika dilihat dari rata-rata waktu eksekusi, hasil pencarian pada dataset 3 akan lebih cepat karena jumlah *term* pada dataset tersebut paling sedikit di antara dataset lainnya. Hal ini sesuai dengan penjelasan pada Bab II, *cosine similarity* memperhitungkan kemiripan berdasarkan nilai dari kumpulan *term* yang sama di antara dua dokumen.

TABEL IV-2 KETEPATAN HASIL PENCARIAN DOKUMEN TER-*CLUSTER* DENGAN TANPA CLUSTER

Test set	dataset1	dataset2	dataset3
test1.doc	tepat	tepat	tepat
test2.doc	tepat	tidak tepat	tepat
test3.doc	tidak tepat	tepat	tidak tepat
test4.doc	tepat	tepat	tepat
test5.doc	tepat	tepat	tepat
test6.doc	tidak tepat	tidak tepat	tidak tepat
test7.doc	tepat	tepat	tepat
test8.doc	tepat	tepat	tepat
test9.doc	tepat	tepat	tidak tepat
test10.doc	tepat	tidak tepat	tepat
Jumlah test set tepat	**8**	**7**	**7**

Pengukuran ketepatan hasil pencarian antara dataset ter-*cluster* dengan tanpa *cluster* juga diperhitungkan. Tabel IV-2 menunjukkan ketepatan hasil pencarian antara dokumen ter-*cluster* dengan tanpa *cluster*. Hasilnya menunjukkan bahwa pada dataset 1, ada 8 dari 10 dokumen test yang menunjukkan dokumen yang sama (tepat). Sedangkan pada dataset 2 dan dataset 3, ada 7 dari 10 dokumen test yang tepat.

Hasil tersebut menunjukkan bahwa penggunaan *cluster* pun dapat menghasilkan dokumen hasil pencarian yang tepat sama dengan pencarian tanpa *cluster*. Hasil pengujian 2 ini pun berhasil menunjukkan hipotesis 2 bahwa, sebuah dokumen yang relevan dengan dokumen lain dalam satu *cluster* akan relevan

juga dengan dokumen-dokumen dalam *cluster* tersebut sehingga proses pencarian dokumen yang mirip akan lebih efektif jika hanya mencari dalam *cluster*-nya saja.

V. KESIMPULAN

Hipotesis 1 dapat dibuktikan berdasarkan hasil pengujian 1. Hal ini dapat terlihat dari Tabel IV-1, bahwa rata-rata akurasi hasil clustering terhadap dataset 2 mencapai nilai tertinggi, bahkan pada percobaan ke-4 berhasil mencapai akurasi hingga 0.96.

Hipotesis 2 dapat dibuktikan berdasarkan hasil pengujian 2. Hasil pengujian 2 memperlihatkan rata-rata waktu pencarían pada semua dataset ter-*cluster* lebih baik daripada waktu pencarían terhadap dataset tanpa *cluster*. Hal ini terlihat dari Gambar IV-3 yang memperlihatkan gap waktu pencarían yang cukup besar antara dataset ter-*cluster* dengan tanpa *cluster*. Selain itu, berdasarkan Tabel IV-2, penggunaan *cluster* dalam pencarían, sebagian besar menghasilkan dokumen yang tepat sama dengan pencarían tanpa cluster.

REFERENCES

[1] Asian, Jelita, E., Hugh and Tahaghoghi, S.M.M., "Stemming Indonesian." s.l.: ACM, 2007. ACM Transactions on Asian Language Information Processing (TALIP).

[2] Bhuyan, Chandrani Ray Chowdhury Prachet., "Information retrieval using fuzzy c-means clustering and modified vector space model." 2010. ICCSIT 3rd IEEE International Conference.

[3] Februariyanti, Herny, Zuliarso, Eri and Utomo, Mardi Siswo., "Prototipe Mesin Pencari Dokumen Teks." Jurnal Teknologi Informasi DINAMIK, 2010, Issue 2, Vol. XV, pp. 115-120. ISSN: 0854-9524.

[4] Feldman, Ronen and Sanger, James., *The Text Mining Handbook.* New York: Cambridge University Press, 2007.

[5] Han, J. and Kamber, M., *Data Mining: Concept and Technique.* s.l.: Morgan Kaufman, 2001.

[6] Hearst, Marti A. and Pedersen, Jan O., "Reexamining the cluster hypothesis: scatter/gather on retrieval results." New York: ACM, 1996. SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval . ISBN:0-89791-792-8.

[7] Kashefi, Omid, Mohseni, Nina and Minaei, Behrouz., "Optimizing Document Similarity Detection in Persian Information Retrieval." Journal of Convergence Information Technology, 2010, Vol. 5. 2.

[8] Kumar, Atul and Sanyal, Sudip., "Effect of Pronoun Resolution on Document Similarity." International Journal of Computer Applications, s.l.: Foundation of Computer Science, 2010, Issue 16, Vol. 1, pp. 60-64.

[9] Lakkaraju, Praveen, Gauch, Susan and Speretta, Micro., "Document similarity based on concept tree distance." s.l.: ACM, 2008. Proceedings of the 33rd international conference on Very large data bases. pp. 127-132.

[10] Leuski, Anton., "Evaluating document clustering for interactive information retrieval." New York: ACM, 2001. CIKM '01 Proceedings of the tenth international conference on Information and knowledge management. ISBN: 1-58113-436-3.

[11] Manning, Christopher D., Raghavan, Prabhakar and Shutze, Hinrich., *Introduction to Information Retrieval.* New York: Cambridge University Press, 2008.

[12] Osinski, Stanislaw, Stefanowski, Jerzy and Weiss, Dawid., "Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition." s.l.: Springer, 2004.

[13] Wan, Xiaojun, Yang, Jianwu and Xiao, Jianguo., "Towards a unified approach to document similarity search using manifold-ranking of blocks." Information Processing & Management, s.l.: Elsevier Ltd, 2007, Issue 3, Vol. 44, pp. 1032-1048.