Jurnal SimanteC
Vol. 6, No. 1 Juni 2017 ISSN 2088-2130;e-ISSN 2502-4884

CLUSTERING BERITA OLAHRAGA BERBAHASA INDONESIA MENGGUNAKAN METODE *K-MEDOID* BERSYARAT

Ach. Yasir Rofiqi
Teknik Informatika, Fakultas Teknik
Universitas Trunojoyo Madura
Jl. Raya Telang, Telang, Kamal, Madura, Jawa Timur 69162
E-mail: rofiqiachyasir@gmail.com

ABSTRAK

Berita merupakan fakta ataupun opini yang membuat banyak orang merasa tertarik untuk mengetahuinya. Berita olahraga merupakan salah satu berita yang memiliki *rating* pengaksesan cukup tinggi. Berita olahraga memiliki berbagai cabang seperti bola, raket, balap dan lain-lain. Oleh karena itu, berita olahraga memerlukan pengelompokkan berita. Pengelompokan dilakukan dengan tujuan agar berita tersebut terhimpun menjadi suatu kelompok sesuai dengan kesamaan berita. Pengelompokan dapat dilakukan dengan berbagai metode. Salah satu metode yang digunakan adalah *k-medoid*. *K-medoid* merupakan metode pengelompokkan yang bekerja secara optimal untuk mengelompokkan berita olahraga dalam jumlah yang sedikit. Setelah semua berita dikelompokkan sesuai dengan kemiripannya, hasil pengelompokkan tersebut perlu diukur nilai akurasinya. Pada saat uji coba dilakukan menghasilkan nilai akurasi sebesar 0,584 dengan jumlah cluster yang dinputkan sebanyak lima *cluster*. Perbaikan metode *k-medoid* dengan menambahkan metode *cosine similarity* mampu meningkatkan nilai akurasi yang semula rata-rata 0,44 bertambah optimal menjadi 0,55.

Kata kunci: Berita olahraga, *Cluster, K-medoid, medoid, cosine similarity*.

ABSTRACT

News is a fact or opinion that makes many people feel interested to know. Sports news is one of the news that has access rating is high enough. Sports news has various branches such as ball, racket, racing and others. Therefore, sports news needs the latest news. Grouping is done with the aim that the news is gathered into groups according to the similarity of news. Grouping can be done in various ways. One of the methods used is k-medoid. K-medoid is a grouping method that works optimally for grouping sports news in small quantities. After all the news is grouped according to the similarity, the result of grouping needs to be measured with accuracy. At the time of trial is done to produce an accuracy of 0.584 with the number of clusters in the input as many as five clusters. Improving k-medoid method by adding cosine similarity method can increase the original accuracy value of 0.44 to optimally increase to 0.55.

Keywords*: Sports news, cluster, k-medoids, medoid, cosine similarity*.

PENDAHULUAN

Dalam buku Sumadiria 2005 (hal 64-65) Assegaf mendefinisikan berita sebagai laporan mengenai fakta atau ide yang mampu menarik perhatian pembaca. Sedangkan menurut Mitchel V. Charnley dalam bukunya yang berjudul "Reporting" edisi ketiga (Holt-Reinhart & Winston, New York, 1975 halaman 44) menyatakan bahwa berita merupakan laporan mengenai fakta atau opini yang memiliki daya tarik atau hal penting bagi banyak orang. Kemudian A.M. Hoeta Soehoet dalam bukunya "Dasar-Dasar Jurnalistik" juga menuliskan bahwa berita merupakan keterangan mengenai peristiwa atau pendapat seseorang yang dianggap penting bagi pembaca untuk memenuhi kebutuhan hidupnya. Dari pengertian berita menurut tiga ahli di atas dapat disimpulkan bahwa berita merupakan fakta ataupun opini yang membuat banyak orang merasa tertarik untuk mengetahuinya. Berita dapat diperoleh dengan berbagai media seperti koran, surat kabar, televisi, internet dan lain-lain. Pada saat ini, media yang paling sering digunakan untuk memperoleh berita adalah internet.

Berita yang ada di internet memiliki berbagai macam topik. Salah satunya yaitu berita olahraga. Berita olahraga merupakan salah satu berita yang memiliki rating yang tinggi. Hal ini dibuktikan pada alexa.com di mana situs berita olahraga seperti sport.detik.com dan bola.net masuk ke dalam 25 *top site* di Indonesia. Sehingga berita olahraga termasuk berita favorit bagi warga Indonesia. Dalam penyajian berita, hampir semua situs memiliki cara yang sama. Berita dibagi menjadi beberapa kategori sesuai dengan cabang olahraga. Pada situs http://sport.detik.com, kategori yang digunakan sebanyak enam cabang olahraga yang meliputi: Basket, Raket, Formula 1, MotoGP, sepak bola dan sport lain. Pada situs http://sport.okezone.com, kategori yang digunakan sebanyak lima cabang olahraga yang meliputi F1, MotoGP, Netting, Basket dan Sport lain. Kemudian pada situs http://sports.sindonews.com, kategori yang digunakan Raket, MotoSpot, Tinju dan all sport. Dari ketiga situs di atas, dapat disimpulkan bahwa rata-rata setiap situs menyediakan kurang lebih lima cabang olahraga.

Namun, proses pengelompokkan masih dilakukan secara manual oleh manusia sesuai dengan kategori yang sudah ditentukan sebelumnya. Sehingga akan menjadi suatu permasalahan jika berita yang akan dikelompokkan berjumlah cukup banyak. Apabila pengelompokan dilakukan secara manual akan membutuhkan waktu yang lama. Oleh karena itu, pengelompokan dibutuhkan secara otomatis. Hal ini bisa dilakukan dengan menggunakan *text mining*.

Text mining merupakan sebuah proses *unsupervised learning* untuk mengelompokkan kemiripan suatu dokumen dengan dokumen yang lain sehingga dapat dipisahkan menjadi beberapa kelompok[1]. Salah satu contoh *text mining* adalah *clustering. Clustering* dapat diartikan sebagai salah satu teknik *text mining* yang digunakan untuk pengelompokan dokumen di mana dokumen dikelompokkan dengan konten yang memiliki kemiripan berita tanpa mendefinisikan kategori sebelumnya[2]. Misalnya koleksi dokumen yang berisi berita sepak bola dan bulu tangkis dapat dikelompokkan sedemikian rupa sehingga semua berita sepak bola dikelompokkan menjadi satu cluster sepak bola dan semua berita bulu tangkis dikelompokkan menjadi satu cluster bulu tangkis.

Untuk mengelompokkan berita dengan clustering dapat menggunakan beberapa metode. Salah satu metode yang sering digunakan adalah metode *k-means*. Metode *k-means* merupakan salah satu algoritma cluster yang sering digunakan karena algoritma ini cukup sederhana. Metode *k-means* mampu mengelompokkan data dengan jumlah yang cukup besar dan waktu komputasi

yang cukup singkat. Namun, *k-means* memiliki kelemahan pada penentuan pusat *cluster* awal. Karena *cluster-cluster* akhir yang dihasilkan sangat dipengaruhi oleh *cluster* awal yang telah didefinisikan[3].

K-medoid merupakan salah satu metode *clustering* yang memiliki efisiensi untuk menangani data set yang kecil. *K-medoid* memiliki kemiripan dengan *k-means*. Karena kedua metode ini tergolong pada metode *partitional clustering*. Namun, *K-meoid* lebih unggul dibadingkan dengan *k-means* karena *k-means* digunakan untuk *data set* dengan jumlah yang besar[4]. Ketika *data set* yang digunakan adalah data yang kecil, maka data yang dihasilkan kurang optimal. Sehingga penelitian ini mengunakan metode *K-meoid* untuk menangani masalah tersebut.

K-medoid memiliki kekurangan yang sama dengan *k-means*. Di mana pada saat penginisialisasian *cluster* awal sangat mempengarui hasil *cluster*. Karena metode ini langsung mengambil secara acak sebanyak *cluster* yang diinginkan oleh pengguna. Untuk menangani hal itu, penelitian ini menambahkan proses pencegahan kesamaan atau tingkat kemiripan yang tinggi pada saat penginisialisasian *cluster* awal. Sehingga inisialisasi *cluster* awal memiliki tingkat kemiripan yang rendah.

Selain itu, jumlah *cluster* yang dimasukkan pengguna juga berpengaruh dalam pengujian akurasi hasil *cluster*. Sehingga penelitian ini diharapkan menghasilkan jumlah cluster yang optimal untuk mengelompokkan berita olahraga berbahasa Indonesia dengan akurasi yang tinggi. Selain itu, penelitian ini juga diharapkan pengelompokkan berita olahraga dengan metode *k-medoid* bersyarat mendapatkan hasil lebih baik dibandingkan dengan menggunakan metode *k-medoid* secara umum.

KAJIAN PUSTAKA

Text Mining

Text mining merupakan sebuah proses *unsupervised* learning untuk mengelompokkan kemiripan suatu dokumen dengan dokumen yang lain sehingga dapat dipisahkan menjadi beberapa kelompok[1]. *Text mining* dapat dilakukan dengan cara manual maupun dengan bantuan aplikasi. *Text mining* manual dapat dilakukan dengan cara mengunjung situs berita tertentu kemudian langsung mengambil isi berita tersebut secara manual. Sedangkan, *text mining* dengan bantuan aplikasi dapat dilakukan dengan bantuan aplikasi *mining* data seperti *crawler*.

Preprocessing

Preprocessing merupakaan tahap mengubah dokumen ke dalam format yang sesuai agar dapat diproses pada klasterisasi[1]. Tahap preprocessing dibagi menjadi beberapa bagian sebagai berikut:

1. *Case folding*
   Case folding merupakan tahap mengubah kapitalitas huruf yang ada pada berita ke dalam bentuk yang sama. Bentuk kapitalitas yang sering digunakan yaitu bentuk kapitalitas kecil. Perintah untuk mengubah huruf menjadi kecil menggunakan perintah *lowercase*.
2. Tokenisasi
   Tokenisasi merupakan tahap mengubah dokumen yang semula berbentuk teks menjadi kumpulan kata yang disebut dengan token. Pada tahap ini juga dilakukan penghapusan simbol atau tanda baca seperti titik(.), koma(,), titik dua(:), dan symbol lainnya. Sehingga berita yang semula berbentuk teks diubah menjadi sekumpulan kata dengan kata-kata yang dihasilkan tanpa mengandung simbol ataupun tanda baca.
3. Penghapusan *stopword*
   Penghapusan *stopword* merupakan sebuah proses untuk menyaring

token yang dihasilkan oleh proses tokenisasi agar tiding mengadung *stopword*. Hal ini dilakukan agar kata-kata yang akan dihitung kemiripannya merupakan kata-kata yang penting. *Stopword* sendiri dapat diartikan sebagai sekumpulan kata yang dianggap tidak penting untuk dihitung kemiripannya dengan kata yang lain.

4. *Stemming*
   Stemming merupakan proses mengubah token menjadi bentuk kata dasar yang disebut dengan *term*. Hal ini dilakukan agar token yang mengandung imbuhan dapat dikembalikan ke bentuk kata dasar sehingga pada saat proses perhitungan kemiripan dokumen, hasil yang didapatkan sangat optimal.
5. Pembobotan
   Pembobotan merupakan tahap menghitung frekuensi dan bobot dari setiap *term* yang dihasilkan oleh tahap *stemming*.

Distance Space

Distance space atau yang dikenal dengan pengukuran jarak merupakan tahap untuk menghitung jarak setiap dokumen dengan dokumen lainnya. Dokumen yang memiliki kemiripan dengan dokumen yang lain dapat diketahui dengan menggunakan *distance space*. Misalkan ada lima dokumen, yaitu dokumen A, B, C, D, dan E. Jika *distance space* antara dokumen A terhadap dokumen B lebih kecil dibandingkan dengan *distance space* dokumen A terhadap dokumen C, maka dapat dikatakan bahwa jarak antara dokumen A terhadap dokumen B lebih dekat dibandingkan dengan jarak antara dokumen A terhadap dokumen C. *Distance space* memiliki beberapa metode antara lain *Euclidean Distance*, *Manhattan Distance*, *Canberra Distance*, dan lain-lain. Pada penelitian ini, perhitungan jarak yang digunakan adalah *Euclidean Distance Space*. Rumus dari *Euclidean Distance* Space dapat dilihat pada Persamaan 1.

$$d(o,m) = \sum_{i=1}^{n} |o_i - m_i|^2 \quad (1)$$

Dimana d merupakan jarak antar dokumen, o adalah data dari dokumen pertama, m adalah data dari dokumen kedua, i adalah inisialisasi index, n adalah index terakhir data dari dokumen.

K-medoid

Metode *k-medoid* dikembangkan oleh Leonard Kaufman dan Peter J. Rousseeuw pada tahun 1987. Algoritma *k-medoid* sering disebut juga algoritma *Partitoning Around Medoid* (PAM). Metode *k-medoid* memiliki kesamaan dengan metode *k-means* yaitu sama-sama termasuk metode *partitioning*. Metode *partitioning* merupakan metode pengelompokan data ke dalam sejumlah *cluster* tanpa adanya struktur hirarki antara satu dengan yang lainnya. Metode *k-medoid* memiliki keunggulan dibandingkan dengan metode *k-means*. *K-medoid* memiliki kinerja yang lebih optimal jika jumlah data yang digunakan berjumlah sedikit. Algoritma ini menggunakan objek pada kumpulan objek untuk mewakili sebuah *cluster*. Objek yang terpilih untuk mewakili sebuah *cluster* disebut *medoid*.

Namun, metode *k-medoid* memiliki kekurangan pada saat inisialisasi *medoid* awal. *Medoid* diambil secara acak tanpa melihat kemiripan antar *medoid*. Hal ini akan berpengaruh pada hasil *cluster*. Karena ada kemungkinan dari *medoid* yang diambil secara acak tersebut memiliki kemiripan yang sangat tinggi. Sehingga hasil *cluster* bisa tidak optimal. Oleh karena itu, penilitian ini sedikit memodifikasi metode *k-medoid.* Sehingga penelitian ini bisa menghasilkan *cluster* yang optimal.

Cosine similarity

Cosine similarity merupakan metode pengukuran jarak atau kemiripan antar objek A dengan objek B[5]. Dengan menambahkan metode *cosine similarity*

pada saat pengambilan *medoid* awal akan membuat sistem lebih akurat dalam mencegah kemiripan antar *medoid*. Sehingga, inisialisasi *medoid* lebih akurat digunakan dalam pengelompokan berita olahraga. Persamaan 2 merupakan metode *cosine similarity*.

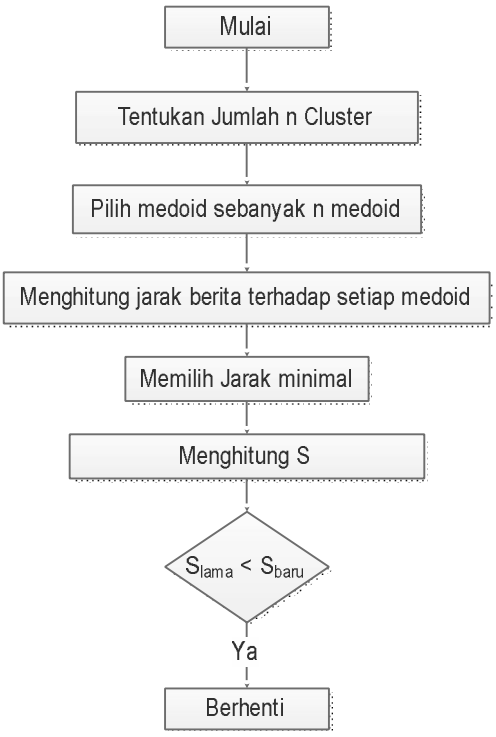
$$similarity(A_j, B) = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iB})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{t} w_{iB}^2}} \quad (2)$$

Dimana :

A = medoid sebelumnya
B = medoid sementara
t = term
w_{ij} = TF-IDF kata ke i dari A ke j
w_{iq} = TF-IDF kata ke i dari B

METODE

Pada penelitian ini menerapkan metode *k-medoid* untuk mengelompokkan berita olahraga sehingga berita dapat dikelompokkan sesuai dengan kemiripan antar berita. Rancangan sistem pada aplikasi penelitian ini sebagaimana diagram Gambar 1.



Gambar 1. Diagram Alir Metode *K-medoid*

Langkah-langkah untuk mengelompokkan objek dengan menggunakan algoritma *K-medoid*[6].

1. Inisialisasi: secara acak pilih sebanyak k objek dari n objek sebagai *medoid*.
2. Hitung jarak dokumen ke *medoid* menggunakan rumus *Euclidean Distance*.
3. Ganti *medoid* dengan data *non-medoid*
4. Hitung jarak total.
5. Ulangi langkah 2 hingga 4 sampai tidak ada perubahan jarak total tersebut.

Pada langkah pertama, penentuan *medoid* awal dilakukan secara acak. Sehingga dapat menimbulkan kemungkinan *medoid* yang terilah merupakan berita dengan tingkat kemiripan yang tinggi. oleh karena itu, peneliti menambahkan suatu metode untuk pengecekan kemiripan antar *medoid* saat dilakukan pemilihan secara acak.

Metode yang digunakan untuk mengecek kemiripan antar medoid ada metode *cosine similarity*. Dengan menambahkan metode *cosine similarity*, *medoid* yang dihasilkan merupakan *medoid-medoid* yang memiliki tingkat kemiripan yang sangat rendah. Sehingga hasil pengelompokkan berita akan lebih optimal jika dibandingkan tanpa menambahkan metode *cosine similarity*.

HASIL DAN PEMBAHASAN

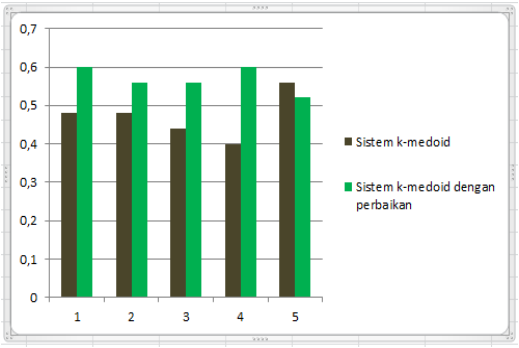
Pada hasil dan pembahasan, penelitian ini akan membandingkan analisa nilai akurasi sistem pengelompokan berita olahraga menggunakan metode *k-medoid* dengan sistem pengelompokan berita olahraga menggunakan metode *k-medoid* yang ditambahkan metode *cosine similarity* pada saat penentuan *medoid*. Uji coba dilakukan dengan melibatkan pengguna sebanyak lima orang. Uji coba dilakukan untuk membandingkan nilai akurasi yang dihasilkan oleh kedua sistem dengan

mengggunakan empat jenis *cluster* masukan.

Hasil dan analisa untuk uji coba pada kedua sistem

Hasil uji coba

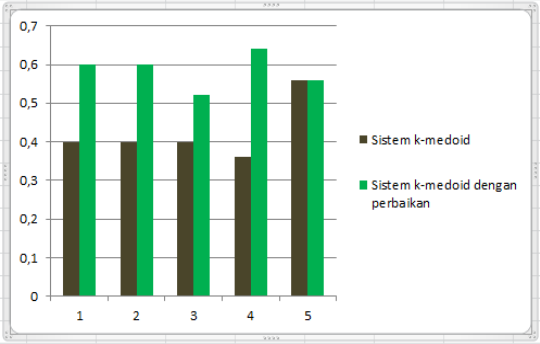
Uji coba dilakukan untuk membandingkan nilai akurasi yang dihasilkan pada pengelompokan berita olahraga dengan menggunakan sistem *clustering* metode *k-medoid* dan sistem *clustering* metode *k-medoid* bersyarat di mana kedua sistem tersebut menggunakan empat jenis *cluster* masukkan. Uji coba pertama dilakukan dengan menggunakan masukkan sebesar 3 *cluster* dengan hasil uji coba pada Gambar 2.



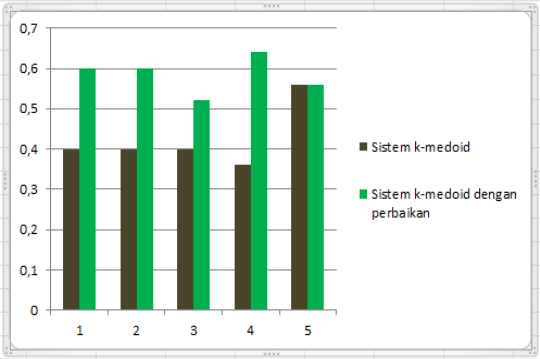
Gambar 2. Perbandingan nilai akurasi dengan memasukkan 3 cluster

Hasil uji coba dengan menginputkan 3 cluster pada kedua aplikasi menghasilkan perbandingan bahwa sistem clustering berita olahraga dengan menggunakan metode k-medoid yang dilakukan perbaikan dengan menambahkan metode cosine similarity memiliki nilai akurasi yang lebih besar. Pada sistem pertama menghasilkan nilai akurasi sebesar 0,472. Sedangkan pada sistem kedua menghasilkan nilai akurasi sebesar 0,568.

Pada Gambar 3, rata-rata nilai akurasi tertinggi dihasilkan oleh sistem menggunakan metode *k-medoid* yang dilakukan perbaikan. Pada sistem pertama menghasilkan nilai akurasi sebesar 0,392. Sedangkan pada sistem kedua menghasilkan nilai akurasi sebesar 0,512.

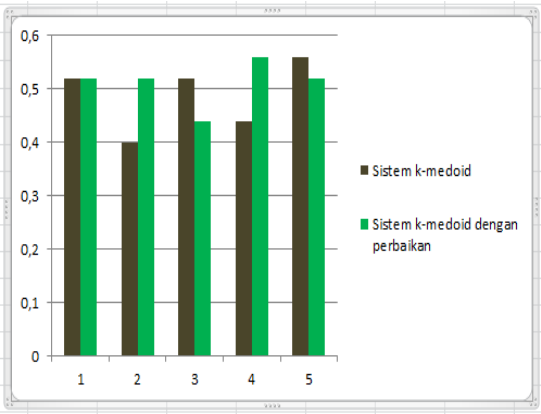


Gambar 3. Perbandingan nilai akurasi dengan memasukkan 4 cluster



Gambar 4. Perbandingan nilai akurasi dengan memasukkan 5 cluster

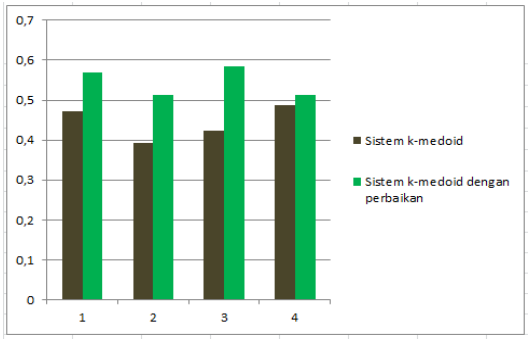
Pada Gambar 4, rata-rata nilai akurasi tertinggi dihasilkan oleh sistem menggunakan metode k-medoid yang dilakukan perbaikan juga. Pada sistem pertama menghasilkan nilai akurasi sebesar 0,424. Sedangkan, pada sistem kedua menghasilkan nilai akurasi sebesar 0,584.



Gambar 5. Perbandingan nilai akurasi dengan memasukkan 5 cluster

Pada Gambar 5, rata-rata nilai akurasi tertinggi dihasilkan oleh sistem menggunakan metode k-medoid yang dilakukan perbaikan. Pada sistem pertama menghasilkan nilai akurasi sebesar 0,488. Sedangkan, pada sistem kedua menghasilkan nilai akurasi sebesar 0,512.

Analisa uji coba



Gambar 6. Perbandingan nilai akurasi dengan memasukkan 5 cluster

Dari Gambar 6 dapat diketahui bahwa metode *k-medoid* dapat digunakan untuk mengelompokkan berita olahraga dengan tingkat akurasi rata-rata 0,44. Jumlah *cluster* yang dapat digunakan untuk mengelompokkan berita olahraga dengan hasil yang optimal adalah percobaan yang dilakukan dengan menggunakan enam *cluster* dengan nilai akurasi sebesar 0,488. Tetapi metode ini belum menghasilkan nilai akurasi yang tinggi. Sehingga diperlukan perbaikan atau penyempurnaan untuk mendapatkan hasil yang optimal. Oleh karena itu, peneliti menambahkan metode *cosine similarity* pada saat pendeklarasian *medoid*. Hal ini dilakukan untuk mencegah adanya inisialisasi *medoid* yang memiliki tingkat kemiripan yang tinggi.

Pada hasil uji coba membuktikan bahwa penambahan metode *cosim* sangat membantu untuk mengoptimalkan hasil *cluster*. Di mana nilai akurasi metode *k-medoid* bersyarat lebih tinggi dibandingkan dengan metode *k-medoid*. Nilai rata-rata akurasi metode *k-medoid* bersyarat memperoleh nilai sebesar 0,54 dengan jumlah *cluster* yang menghasilkan akurasi yang paling optimal adalah percobaan yang dilakukan dengan menggunakan lima *cluster* dengan akurasi sebesar 0,584.

SIMPULAN

Dari pengujian 2 sistem tersebut, didapatkan perbedaan rata-rata nilai akurasi sebesar 0,44 untuk sistem dengan menggunakan metode *k-medoid* dan 0,54 untuk sistem dengan menggunakan metode *k-medoid* yang ditambahkan metode *cosine similarity* pada saat pendeklarasian *medoid*. Sehingga dapat disimpulkan bahwa perbaikan metode *k-medoid* dengan menambahkan metode *cosine similarity* pada saat pendeklarasian *medoid* sangat membantu untuk meningkatkan nilai akurasi.

DAFTAR PUSTAKA

[1] R. Handoyo, R. Mangkudjaja, S.N. Michrandi. "Perbandingan Metode Clustering Menggunakan Metode Single Linkage dan K-Means Pada Pengelompokan Dokumen", *Jurnal Sifo Mikroskil*, Vol. 15, No. 2, pp. 73-82, 2014.

[2] S. Raharjo, E. Winarko, "Klasterisasi, Klasifikasi Dan Peringkasan Teks Berbahasa Indonesia", *Prosiding seminar ilmiah nasional komputer dan sistem intelijen*, Vol.8, pp. 391-401, 2014.

[3] T. Alfina, B. Santosa, A.B. Ridho, "Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)", *Jurnal Teknik ITS*, Vol. 1, No. 1, pp. 521-525, 2012.

[4] W. T. Agus, "Algoritma K-Medoids Untuk Penentuan Strategi Pemasaran Produk", *Jurnal Simetris*, Vol. 6, No. 1, pp. 183-188, 2015.

[5] S. D. Rahardjo, F. Solihin, I. Santoso, *Perancangan dan Pembuatan Mesin Pencari Lowongan Pekerjaan menggunakan Metode Cosine Similarity*, 2014

[6] Y. Wibisono, "Perbandingan Partition Around Medoids (PAM) dan K-means Clustering untuk Tweet", *Prosiding Konferensi Nasional Sistem Informasi*, pp. 25-26, 2011.