PERBANDINGAN METODE *CLUSTERING* MENGGUNAKAN METODE *SINGLE LINKAGE* DAN *K - MEANS* PADA PENGELOMPOKAN DOKUMEN

Rendy Handoyo[1], R. Rumani M[2], Surya Michrandi Nasution[3]
[1,2,3] Gedung N-203, Program Studi Sistem Komputer,
Fakultas Teknik Elektro - Universitas Telkom,
Jl. Telekomunikasi No. 1, Bandung 40257
[1]sianipar.rendy@yahoo.co.id, [2]rumani@telkomuniversity.ac.id, [3]smn@ittelkom.ac.id

Abstrak

Penyebaran berita saat ini semakin tersebar luas semenjak perkembangan dunia internet yang semakin pesat. Perkembangan dunia internet membuat berita yang tersebar semakin beragam dan berjumlah sangat besar. Pembaca berita akan kesulitan untuk memperoleh berita yang diinginkan jika berita tersebut tidak terkelompok dengan baik. Dan jika harus dikelompokkan secara manual membutuhkan waktu yang sangat lama. Oleh sebab itu, *Clustering* menjadi solusi untuk mengatasi masalah tersebut. *Clustering* akan mengelompokkan dokumen berita berdasarkan tingkat kemiripan dari dokumen tersebut.

Metode *Single Linkage* merupakan metode pengelompokan *hierarchical clustering*. Metode *Single Linkage* mengelompokkan dokumen didasarkan pada jarak terdekat antar dokumen. Komputasi *Single Linkage* merupakan komputasi yang mahal dan kompleks. Sedangkan metode *K-means* merupakan metode pengelompokan *partitioned clustering*. Metode *K-means* mengelompokkan dokumen didasarkan pada jarak terdekat dengan *centroid*-nya. *K-Means* merupakan metode pengelompokan yang sederhana dan dapat digunakan dengan mudah. Tetapi pada jenis data tertentu, *K-means* tidak dapat memberikan segementasi data dengan baik, sehingga kelompok yang terbentuk tidak murni data yang sama.

Metode pengujian yang digunakan untuk mengukur kualitas *cluster* adalah *Silhouette Coefficient* dan *Purity* . Berdasarkan hasil pengujian yang dilakukan, dapat disimpulkan, bahwa metode *Single Linkage* memiliki performansi yang lebih baik dibandingkan dengan metode *K-means*. Nilai *Silhouette Coefficient Single Linkage* selalu lebih unggul dibandingkan dengan *K-Means*. Pertambahan jumlah dokumen membuat nilai *Silhouette Coefficient single linkage* semakin kecil sedangkan *K-means* terkadang menghasilkan nilai yang negatif. Untuk nilai *Purity , Single Linkage* selalu bernilai 1 sedangkan *K-Means* tidak pernah bernilai 1. Hasil pertambahan jumlah *cluster* dan jumlah dokumen memberikan pengaruh terhadap nilai *Silhouette Coefficient* dan *Purity* . Hal ini berarti *single linkage* selalu menghasilkan dokumen yang sama, sedangkan *K-means* masih bercampur dengan dokumen yang lain.

Kata kunci : *Clustering*, *HAC, Partitioned, Single Linkage, K-Means, Silhouettte Coefficient, Purity*

1. Pendahuluan

Sistem pengelompokan dokumen berita merupakan sistem yang menggabungkan dokumen berita berdasarkan tingkat kemiripan dari dokumen berita tersebut. Sistem pengelompokan ini memberikan kemudahan dalam pencarian dokumen berita yang diinginkan. Pencarian dokumen berita yang diinginkan akan semakin cepat dilakukan.

Dokumen berita yang telah dikelompokkan akan tersusun dengan terstrukrtur dan rapi sesuai dengan kemiripan dokumen tersebut. Dokumen yang akan dikelompokkan dalam sistem pengelompokan ini adalah berita berita *online* yang disimpan di dalam *notepad*. Berita merupakan informasi yang memiliki pembahasan yang bermacam macam. Karena itulah, dibutuhkan suatu sistem pengelompokan supaya berita berita tersebut tersusun berdasarkan topik - topik yang sama.

Sistem pengelompokan ini akan dikelompokkan menggunakan metode *single linkage clustering* dan *k-means clustering*. Metode *single linkage* merupakan teknik pengelompokan yang bekerja berdasarkan prinsip algoritma *Hierarchical Clustering*. Sedangkan *K-means* merupakan teknik pengelompokan yang bekerja berdasarkan *Partitioned Clustering*. Prinsip kerja dari pengelompokan *Hierarchical Clustering* dilakukan secara bertahap. Dan disetiap iterasi dari pengelompokan *Hierarchical Clustering* hanya ada satu pemilihan penggabungan suatu dokumen terhadap dokumen lainnya. Sedangkan prinsip kerja dari pengelompokan *Partitioned Clustering* adalah mengelompokkan dokumen secara acak karena dipengaruhi *centroid*. Dan disetiap iterasi dari pengelompokkan *Partitioned Clustering* memungkinkan untuk terjadinya lebih dari satu pemilihan dokumen yang akan digabungkan.

Sistem pengelompokan dokumen ini dilakukan dengan membandingkan performansi dari dua metode tersebut. Parameter yang digunakan untuk membandingkan performansi metode tersebut adalah *Silhouette Coefficient* dan *Purity* . *Silhouette Coefficient* diperoleh dengan menghitung jarak rata - rata antar dokumen dalam satu *cluster*. Setelah itu dilakukan penghitungan jarak antara suatu dokumen dengan dokumen lain yang berada dalam *cluster* lain, dan yang diambil adalah jarak yang terdekat. Sedangkan pengujian *Purity* , dilakukan secara *manual*. Pengujian *manual* yang dimaksud adalah, dengan memeriksa setiap dokumen dalam suatu *cluster* apakah merupakan dokumen yang mirip atau berada dalam satu kategori yang sama.

2. Kajian Pustaka

2.1 *Text Clustering*

Text Clustering adalah proses *unsupervised learning* (proses pembelajaran sendiri) yang mengelompokkan kumpulan dokumen berdasarkan hubungan kemiripannya dan memisahkannya ke dalam beberapa kelompok. [6]

2.2 *Preprocessing*

Preprocessing merupakan pemrosesan awal dokumen agar diperoleh suatu nilai yang dapat dipelajari oleh sistem *clustering*.[8]

2.2.1 Case Folding

Case folding merupakan suatu tahap yang mengubah huruf besar menjadi huruf kecil.[6]

2.2.2 Tokenization

Tokenization adalah proses pemotongan seluruh urutan karakter menjadi satu potongan kata.[6]

2.2.3 Stopword Removal

Stopword removal merupakan proses penghapusan semua kata yang tidak memiliki makna.[6]

2.2.4 Stemming

Stemming adalah proses membentuk suatu kata menjadi kata dasarnya. Algoritma *stemming* yang digunakan dalam sistem pengelompokkan ini adalah algoritma Nazief – Adriani [4]

2.2.5 *Term Weighting*

Term weighting merupakan proses pemberian bobot suatu *token* dalam suatu *term.*

2.2.5.1 Term Frequency

Term Frequency (TF) adalah pembobotan yang menghitung frekuensi kemunculan sebuah *token* pada suatu dokumen

$$TF\ (t_k, d_j) = f\ (t_k, d_j) \quad [6] \tag{1}$$

2.2.5.2 Document Frequency

Document Frequency (DF) adalah pembobotan yang menghitung frekuensi kemunculan sebuah *token* pada kumpulan dokumen

$$IDF\ (t_k) = log\ \frac{N}{df\ (t)} \quad [6] \tag{2}$$

2.2.5.3 Pembobotan TF.IDF

Pembobotan TF • IDF adalah suatu pengukuran statistik untuk mengukur seberapa penting sebuah *token* dalam kumpulan dokumen

$$TF \bullet IDF\ (t_k, d_j) = TF\ (t_k, d_j) \bullet IDF\ (t_k)\ [6] \tag{3}$$

2.2.5.4 Normalisasi

Normalisasi merupakan proses penyetaraan jumlah kata yang berbeda-beda pada setiap dokumen.

$$w\ (word_i) = \frac{w\ (wordi)}{\sqrt{w^2(wordi) + w^2(word2) + \ldots + w^2(wordn)}}\ [6] \tag{4}$$

2.3 *Vector Space Model* (VSM)

VSM merupakan metode yang merepresentasikan data atau *query* dalam bentuk vektor.

2.4 *Distance Space*

Distance Space adalah proses penghitungan jarak antara suatu dokumen dengan dokumen lainnya.

Distance space yang digunakan adalah *Euclidean distance,* dengan rumus sebagai berikut :

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{in} - x_{jn}|^2}\ [2] \tag{5}$$

$d(i,j)$ = jarak antara data ke i dan data ke j
x_{i1} = nilai atribut ke satu dari data ke i
x_{j1} = nilai atribut ke satu dari data ke j
n = jumlah atribut yang digunakan

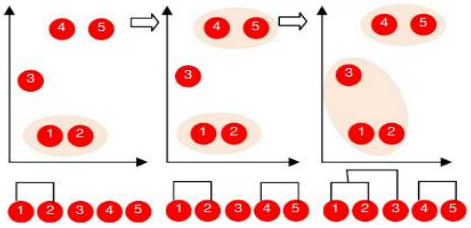
2.5 *Single Linkage*

Single Linkage Clustering merupakan contoh dari algoritma *Agglomerative Hierarchical Clustering.*

Langkah-langkah dari metode *Single Linkage Clustering* adalah sebagai berikut : [6]

1. Menentukan k sebagai jumlah *cluster* yang ingin dibentuk.

2. Setiap data dianggap sebagai *cluster*. Kalau n = jumlah data dan c = jumlah *cluster*, berarti ada c = n.
3. Menghitung jarak / *similarity* / *dissimilarity* antar *cluster*.
4. Cari dua *cluster* yang mempunyai jarak antar *cluster* yang minimal dan gabungkan (c = c-1). Setelah semua jarak diketahui, selanjutnya dikelompokkan dokumen-dokumen yang memiliki jarak terdekat.
5. Jika c > 3, kembali ke langkah 3



Gambar 1 Langkah-langkah metode *Single Linkage* [6]

2.6 *K-Means*

K-Means Clustering merupakan metode yang termasuk ke dalam golongan algoritma *Partitioning Clustering*.

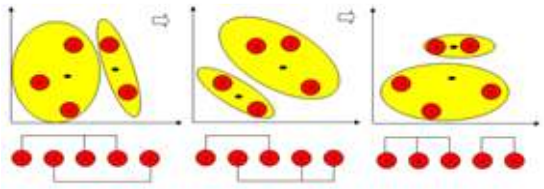
Langkah-langkah dari metode *K-Means* adalah sebagi berikut : [7]

1. Tentukan nilai k sebagai jumlah *cluster* yang ingin dibentuk.
2. Bangkitkan k *centroid* (titik pusat *cluster*) awal secara acak.
3. Hitung jarak setiap data ke masing-masing *centroid* menggunakan rumus korelasi antar dua objek (*Euclidean Distance*).
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroid*nya.
5. Tentukan posisi *centroid* baru (k C) dengan cara menghitung nilai rata-rata dari data yang ada pada *centroid* yang sama.

$$C_k = \left(\frac{1}{n_k}\right)\sum d_i$$

Dimana $\mathbf{n_k}$ adalah jumlah dokumen dalam *cluster* k dan $\mathbf{d_i}$ adalah dokumen dalam *cluster* k.

6. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama, tidak sama.



Gambar 2 Langkah-langkah metode K-Means [7]

2.7 Evaluasi *Cluster*

2.7.1 Silhouette Coefficient

Silhouette Coefficient digunakan untuk melihat kualitas dan kekuatan *cluster*, seberapa baik suatu objek ditempatkan dalam suatu *cluster*. Metode ini merupakan gabungan dari

metode *cohesion* dan *separation*. Tahapan perhitungan *Silhouette Coefficient* adalah sebagai berikut:

1. Hitung rata-rata jarak dari suatu dokumen misalkan i dengan semua dokumen lain yang berada dalam satu *cluster*

$$a(i) = \frac{1}{|A|-1} \sum_{j \in A,\ j \neq i} d(i,j) \quad [1] \tag{7}$$

dengan j adalah dokumen lain dalam satu *cluster* A dan d(i,j) adalah jarak antara dokumen i dengan j.

2. Hitung rata-rata jarak dari dokumen i tersebut dengan semua dokumen di *cluster* lain, dan diambil nilai terkecilnya.

$$d(i,C) = \frac{1}{|A|} \sum j \in C\ d(i,j) \quad [1] \tag{8}$$

dengan $d(i,C)$ adalah jarak rata-rata dokumen i dengan semua objek pada *cluster* lain C dimana A ≠ C.

$$b(i) = min\ C \neq A\ d(i,C)\ [3] \tag{9}$$

3. Nilai *Silhouette Coefficient* nya adalah :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i),b(i))} \quad [3] \tag{10}$$

2.7.2 *Purity* [5]

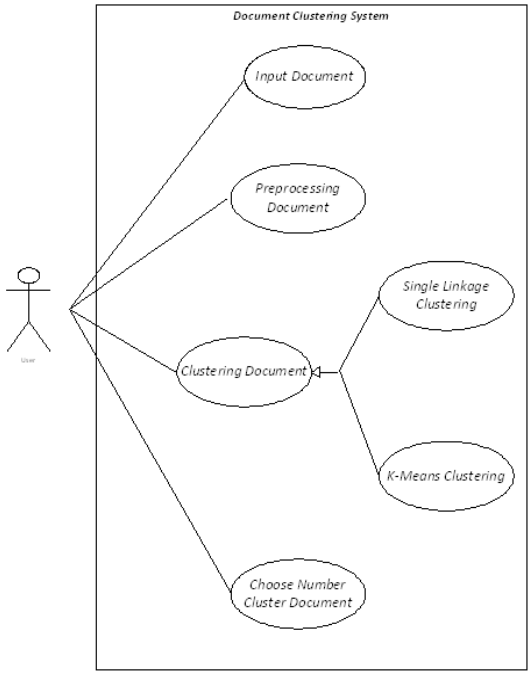
Purity (kemurnian) suatu *cluster* direpresentasikan sebagai anggota *cluster* yang paling banyak sesuai (cocok) di suatu kelas. *Purity* dapat dihitung dengan rumus berikut :

$$Purity\ (j) = \frac{1}{nj} max\ (n_{ij}) \tag{11}$$

Total nilai *Purity* dapat dihitung dengan rumusi berikut :

$$Purity\ = \sum_{i=0}^{j} \frac{nj}{n} Purity\ (j) \tag{12}$$

3 Diagram Use Case Sistem

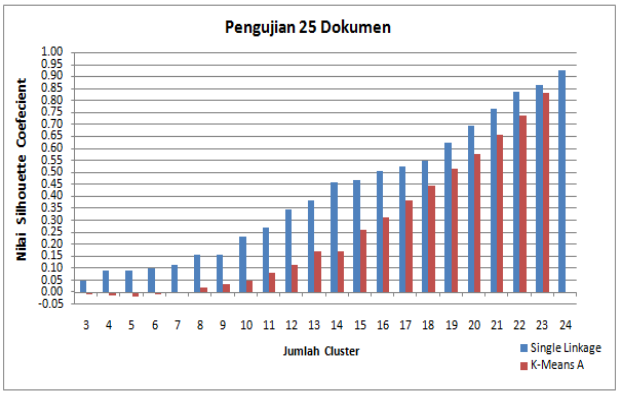


Gambar 3 Diagram *Use Case*

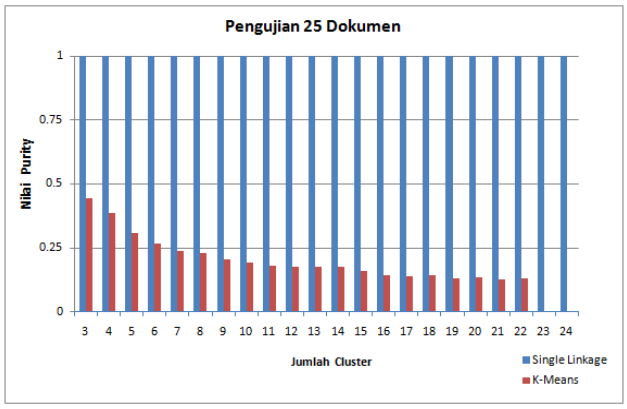
4 Pengujian Sistem

4.1 Hasil Pengujian

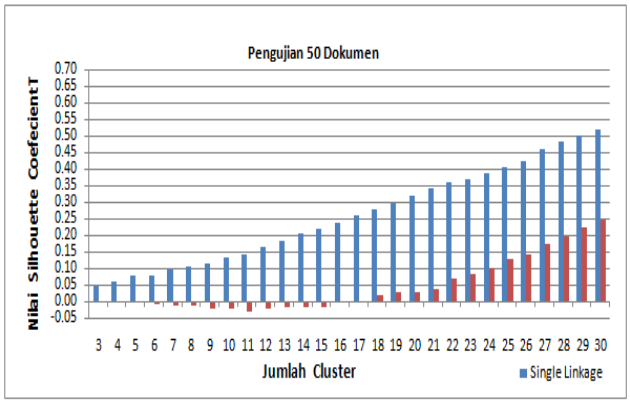
Pengujian ini bertujuan untuk melihat pengaruh jumlah dokumen, jumlah *cluster*, dan metode *clustering* dalam mengelompokkan dokumen. Skenario-skenario pengujiannya adalah menguji performansi *clustering* yang dibangun dengan *dataset* 4 kategori (25%, 50%, 75% dan 100% *dataset*) dengan jumlah *cluster* adalah 3, 4, 5, 6, 7, 8, 9, dan 10.



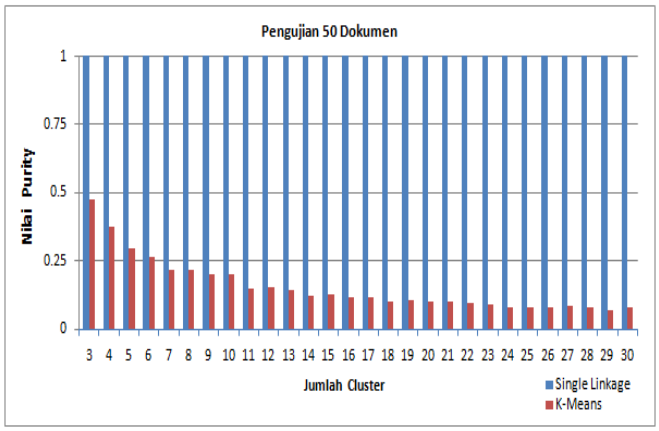
Gambar 4 Nilai *Silhouette Coefficient* untuk 25 Dokumen



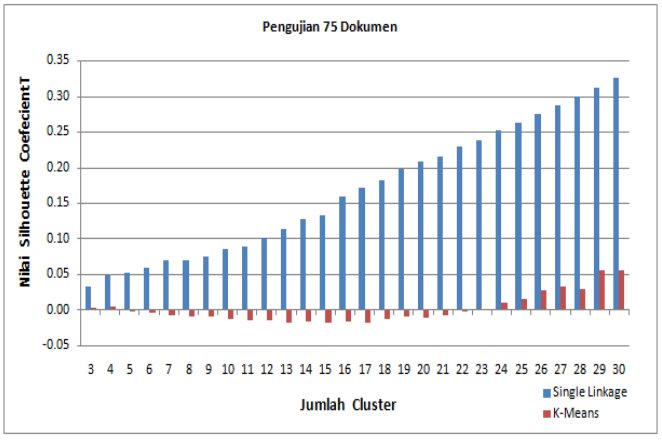
Gambar 5 Nilai *Purity* untuk 25 Dokumen



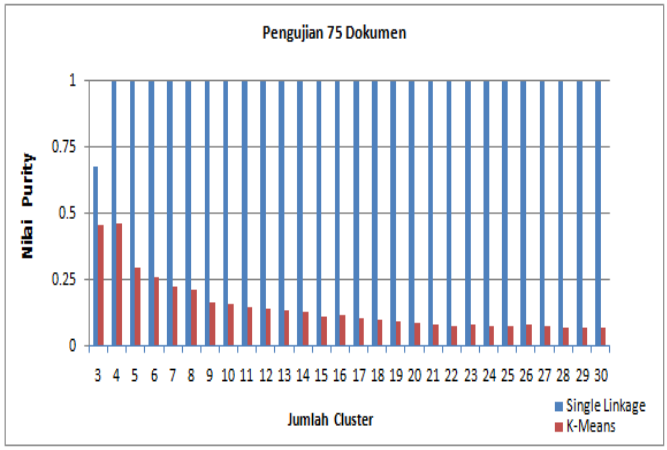
Gambar 6 Nilai *Silhouette Coefficient* untuk 50 Dokumen



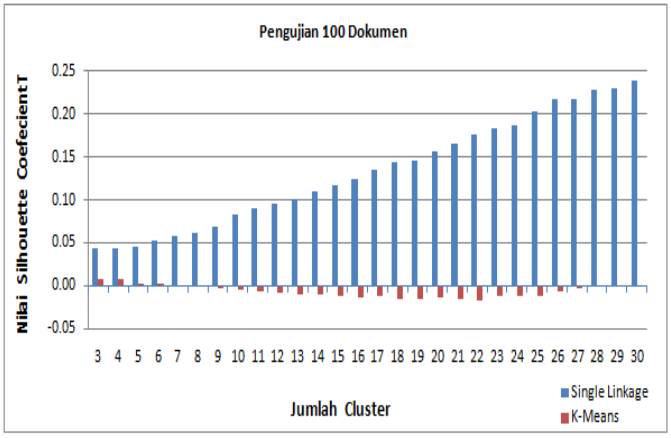
Gambar 7 Nilai *Purity* untuk 50 Dokumen



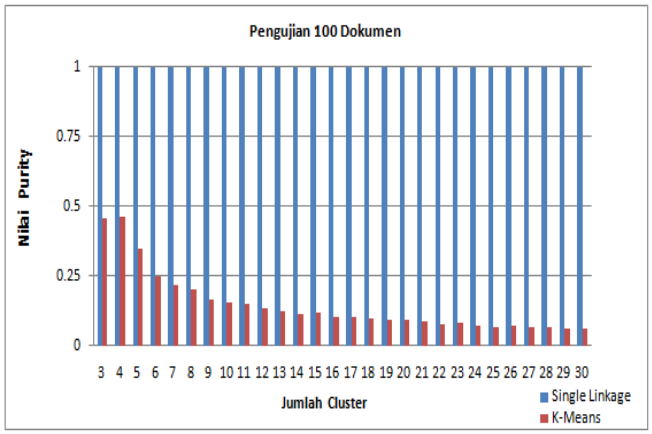
Gambar 8 Nilai *Silhouette Coefficient* untuk 75 Dokumen



Gambar 9 Nilai *Purity* untuk 75 Dokumen



Gambar 10 Nilai *Silhouette Coefficient* untuk 100 Dokumen



Gambar 11 Nilai *Purity* untuk 100 Dokumen

4.2 Analisis Perbandingan Performansi Single Linkage Clustering dan K-Means

Setelah dilakukan pengujian secara keseluruhan, diperoleh hasil bahwa algoritma *single linkage* memiliki performansi yang lebih baik jika dibandingkan dengan algoritma *K-means*. Hal ini dapat dilihat dari perbandingan nilai *Silhouette Coefficient* dan *Purity* dari kedua metode tersebut. Nilai dari *Silhouette Coefficient* dan *Purity* metode *single linkage* selalu lebih tinggi jika dibandingkan dengan metode *K-means*. Kurang optimalnya metode *K-means* ini disebabkan karena proses inisialisasi *cluster* yang dilakukan secara acak pada saat pembangkitan *cluster* awal. Pembangkitan *cluster* awal tersebut ditujukan untuk memasukkan dokumen - dokumen ke setiap *cluster* sesuai dengan jumlah *cluster* yang telah ditentukan sebelumnya. Karenanya, sangat sulit untuk memperoleh hasil *cluster* awal yang unik. Walaupun untuk pengujian *K-means* dilakukan pengulangan sebanyak 30 kali, hasil pengujiaan-pengujian tersebut belum tentu merepresentasikan suatu *cluster* yang baik. Akibatnya *centroid* yang diperolehpun bukanlah *centroid* yang didominasi dengan dokumen-dokumen yang lebih banyak kemiripannya. Sehingga iterasi-iterasi setelah pembangkitan *cluster* awal tersebut akan menghasilkan *centroid* yang kurang tepat. Dengan demikian *cluster* yang akan terbentuk pada akhirnya bukanlah *cluster* yang memiliki nilai *Purity* yang mendekati 1.

Setelah melakukan variasi penambahan jumlah *cluster,* ternyata variasi tersebut memberikan pengaruh terhadap nilai *Silhouette Coefficient* dan *Purity* . Untuk nilai *Silhouette*

Coefficient, nilainya akan semakin besar jika jumlah *cluster* semakin bertambah. Dan sebaliknya jika jumlah *cluster* semakin kecil, maka nilai dari *Silhouette Coefficient* akan semakin kecil. Untuk nilai *Purity* , variasi penambahan jumlah *cluster* hanya berpengaruh terhadap metode *K-means*. Jika jumlah *cluster* semakin besar, maka nilai dari *Purity* mrtode *K-means* akan semakin kecil sedangkan sebaliknya, jika jumlah *cluster* semakin kecil, maka nilai dari *Purity* akan semakin besar. Nilai *Purity* untuk metode *single linkage* secara umum selalu bernilai 1. Artinya hampir disetiap *cluster* selalu dihasilkan anggota kelompok yang selalu mirip dengan anggota yang lainnya.

Sedangkan dengan melakukan variasi penambahan jumlah dokumen, perubahan yang tampak hanya terjadi untuk nilai *Silhouette Coefficient*. Penambahan jumlah dokumen tidak memberikan perubahan nilai yang sangat berbeda terhadap nilai *Purity* . Untuk nilai *Silhouette Coefficient* jika jumlah dokumen semakin besar, maka nilai dari *Silhouette Coefficient* semakin kecil, dan jika jumlah dokumennya semakin kecil maka nilai dari *Silhouette Coefficient*nya menjadi semakin besar.

Jika dilihat dari nilai *Silhouette Coefficient*nya, metode *K-means* membutuhkan jumlah *cluster* yang lebih banyak untuk menyamai nilai *Silhouette Coefficient* metode *single linkage*. Sebagai contoh, pada pengujian 25 dokumen dengan 3 *cluster*, metode *single linkage* memiliki nilai *Silhouette Coefficient* sebesar 0.047325055. Untuk mencapai nilai *Silhouette Coefficient* yang hampir sama dengan nilai *Silhouette Coefficient single linkage* tersebut diatas, diperlukan jumlah *cluster* metode *K-means* sebanyak 10 *cluster*. Dengan kondisi ini (10 *cluster*), hasil pengujian nilai *Silhouette Coefficient* metode *K-means* adalah 0.0456749737981.

5 Kesimpulan

5.1 Kesimpulan

1. Metode *Single Linkage* memilliki performansi yang lebih baik dibandingkan dengan metode *K-means*; berdasarkan data yang ada, hubungan antara nilai performansi dengan jumlah dokumen dapat dinyatakan dengan persamaan berikut : **y = ((1200 – 17x) / 25) – 6,** dimana y menyatakan nilai performansi dalam %, dan x menyatakan jumlah dokumen.
2. Jumlah *cluster* memberikan pengaruh terhadap nilai *silhouette* dan *Purity* . Jika jumlah *cluster* bertambah, maka nilai *sillhouette coefficient* akan bertambah, sedangkan nilai *Purity* akan mengecil. Jika jumlah *cluster* berkurang, maka nilai *Silhouette Coefficient* akan semakin mengecil, sedangkan *Purity* semakin membesar.
3. Jumlah dokumen memberikan pengaruh terhadap nilai *Silhouette Coefficient*. Jika jumlah dokumen berkurang, maka nilai *Silhouette Coefficient* menjadi semakin membesar. Sedangkan untuk *Purity ,* jumlah dokumen memberikan pengaruh yang acak terhadap *Silhouette Coefficient*, bervariasi antara 0.21 - 0.24.

5.2 Saran

1. Untuk metode *K-means* dapat dikembangkan lebih lanjut suatu metode untuk inisialisasi *cluster* awal.
2. Sistem pengelompokan ini dapat dikembangkan menjadi sistem pemeriksaan kemungkinan adanya plagiarisme di dalam penulisan karya tulis ilmiah, seperti jurnal ilmiah, tesis, desertasi, dan sebagainya.
3. Penelitian selanjutnya dapat dikembangkan untuk membandingkan metode-metode *hierarchical* dan *partitioned* yang lainnya seperti *average linkage, complete linkage*, *k-medoids,* dan lain-lain.

Referensi

[1] Al-Zoubi, Moh'd Belal, Mohammad al Rawi, 2008, *An Efficient Approach for Computing Silhouette Coefficients*, Journal of Computer Science.
[2] Feldman, Ronen., James Sanger, 2006. *The Text Mining Handbook*, New York: Cambridge University Press.
[3] Han, Jiawei., Micheline Kamber, Jian Pei, 2012, *Data Mining Concepts and Techniques,* USA : Morgan Kaufmann.
[4] Keke, Dyan., Rian Chikita, Agus Dwi Kuncoro, 2012, *Algoritma Nazief dan Adriani*, Universitas Gajah Mada.
[5] Li, Yanjun., Congnan Luo, Soon M. Chung, May 2008, *Text Clustering with Feature Selection by Using Statistical Data,* IEEE. Volume : 20, No.5
[6] Manning, Christopher D., Prabhakar Raghavan, Hinrich Schutze, 2009, *An Introduction to Information Retrieval*, Cambridge: Cambridge University Presss.
[7] Srivastava, Ashok., Mehran Sahami, 2009, *Text Mining Classification, Clustering, and Applications,* USA : Taylor and Francis Group, LLC.
[8] Wu, Junjie, 2012, *Advanced in K-means Clustering,* London : Springer