PERANCANGAN DAN ANALISIS *CLUSTERING* DATA MENGGUNAKAN METODE *SINGLE LINKAGE* UNTUK BERITA BERBAHASA INGGRIS

DESIGN AND ANALYSIS OF DATA CLUSTERING USING SINGLE LINKAGE METHOD FOR ENGLISH NEWS

Fachri Nugraha Adhiatma[1], Surya Michrandi Nasution[2], Yudha Purwanto[3]

[1]Prodi S1 Sistem Komputer, Fakultas Teknik Elektro, Universitas Telkom
[2]Prodi S1 Sistem Komputer, Fakultas Teknik Elektro, Universitas Telkom
[3]Prodi S1 Sistem Komputer, Fakultas Teknik Elektro, Universitas Telkom
[1]fachrina15@gmail.com, [2]michrandi@telkomuniveristy.co.id, [3]omyudha@telkomuniversity.ac.id

Abstrak

Pada era sekarang dunia komputer dan informasi sudah berkembang sangat pesat. Terutama di bidang pengolahan data. Seiring dengan perkembangan zaman, data yang terbentuk di zaman sekarang sudah semakin banyak. Oleh karena itu dibutuhkan suatu solusi supaya data tersebut bisa diolah menjadi informasi yang berguna. Data mining merupakan suatu solusi untuk menangani banyaknya data. Data *mining* adalah sebuah alat yang memiliki peranan penting untuk mengatur dan mengolah banyaknya data.

Pada penelitian ini penulis membahas tentang *clustering* data untuk data teks, yaitu menggunakan data berita yang diambil dari situs *online* dan dipindah ke *notepad* dengan format .txt. Pada penelitian ini, penulis menggunakan metode *single lingkage clustering*. Metode *single linkage clustering* merupakan bagian dari *agglomerative hierarchical clustering.* Metode *single linkage* mengelompokkan data berdasarkan pada data atau objek terdekat. Metode yang digunakan untuk mengetahui kualitas *cluster* adalah dengan validitas berorientasi kemiripan.

Hasil yang sudah tercapai dari sistem *summarize* dan *single linkage clustering* cukup baik. *Summarize* dengan persentase 50% adalah yang terbaik, yaitu memiliki tingkat akurasi 74.10%. Sedangkan, *Single linkage clustering* dengan 12 *cluster* memiliki tingkat akurasi yang terbaik dibandingkan dengan *single linkage clustering* 6 *cluster* dan akurasi yang dihasilkan oleh jumlah *cluster* sebanyak 12 *cluster* adalah 79.8%.

Kata kunci : *Data mining, text mining, clustering, agglomerative hierarchical clustering, single linkage, summarize*

Abstract

In the current era, computing and information has been growing fast. Especially in the field of data processing . Along the time, the created data keeps increasing . Therefore we need a solution so that the data can be processed into useful information . Data mining is a solution to handle the amount of data. Data mining is a tool that has an important role to organize and process the amount of data.

In this research, the author discusses the clustering of data for text data, which uses data taken from the online site news and moved to a notepad with a .txt format. In this study, the authors use the Single Linkage clustering. Single linkage clustering method is part of agglomerative hierarchical clustering. Single linkage method of classifying the data based on the data or object nearby. The method used to determine the quality of the cluster is the validity of the similarity oriented.

The results from the system summarize and single linkage clustering is good enough. A percentage of 50% was the best in Summarize, which has a 74.10% accuracy rate. Meanwhile, Single linkage clustering with 12 clusters have better accuracy rate than single linkage clustering with 6 cluster which was 79.8%.

***Keywords** : Data mining, text mining, clustering, agglomerative hierarchical clustering, a single linkage, summarize*

1. Pendahuluan

Membaca koleksi dokumen yang berjumlah banyak tentu akan membutuhkan yang sangat lama. Begitu juga untuk melakukan pengelompokkan dokumen berdasarkan isi topik atau kesamaan topik dalam jumlah yang banyak. Oleh karena itu, penulis melakukan penelitian tentang peringkasan dokumen secara otomatis dan pengelompokkan dokumen dengan metode clustering.

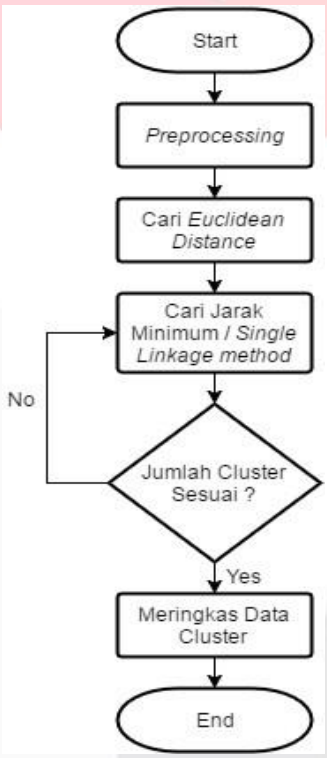
Pada penelitian ini penulis melakukan penelitian terhadap data berita yang diambil dari situs online, kemudian disalin ke notepad dengan format .txt. Data berita tersebut berupa dokumen teks yang memiliki isi

tulisan atau kumpulan kata. Metode yang digunakan untuk meringkas dokumen teks adalah metode TF-IDF. Ringkasan tersebut akan diambil berdasarkan scoring TF-IDF tertinggi. Peringkasan terbagi menjadi dua metode yaitu abstraktif dan ekstraktif. Abstraktif adalah meringkas dokumen teks menggunakan teknologi natural language generation yang hasil ringkasannya akan tercipta secara alami, yaitu seperti meringkas secara manual. Metode ekstraktif adalah metode peringkasan teks dengan menampilkan kembali kalimat yang dibaca sistem paling penting pada dokumen teks yang diringkas.

Clustering adalah metode untuk menganalisa data. Clustering berguna untuk mengelompokan objek atau data berdasarkan tingkat kemiripan ke dalam suatu cluster sehingga cluster atau grup tersebut berisi data yang mirip dan berbeda dengan data pada cluster lainnya. Pengelompokkan dokumen teks dilakukan menggunakan metode single linkage clustering. Metode tersebut termasuk dalam hierarchical clustering. Pada kasus ini clustering digunakan untuk mengelompokkan dokumen teks berdasarkan isi atau kesamaan topik antar dokumen. Single linkage clustering adalah metode clustering yang mengelompokkan data atau objek berdasarkan jarak terdekat antar data atau objek.

2. Perancangan Sistem

2.1 Deskripsi Umum Sistem



Gambar 2.1 *Flowchart* Umum Sistem

Tahapan sistem yang dibangun terdiri atas :

1. Sistem menirima inputan dataset berupa file .txt yang berisi berita berbahasa inggris. Kemudian dilakukan proses *preprocessing* yang terdiri atas *tokenization*, *stopwords removal*, *case folding*, *stemming*, *TF-IDF*. Proses *TF-IDF* dilakukan untuk mendapatkan nilai dan mengetahui seberapa besar bobot suatu kata. Hasil *summarize* / ringkasan didapatkan dengan cara melakukan *sorting* terhadap *TF-IDF*, dan hasil *sorting* diurutkan yang terbesar untuk mendapatkan hasil ringkasan sesuai dengan presentase yang diinginkan.

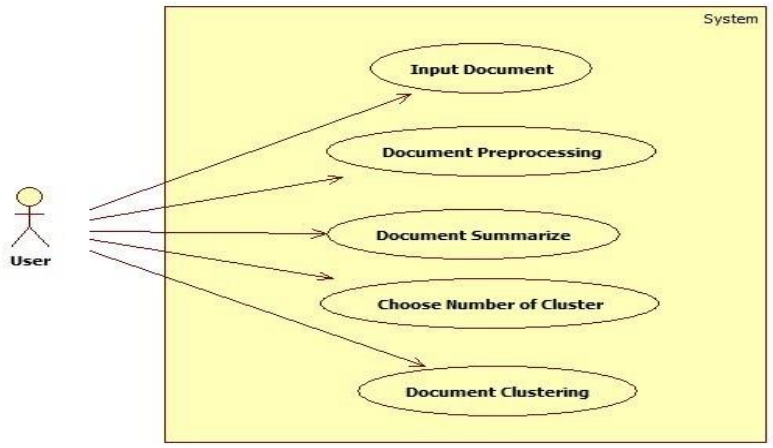
2. Dilakukan perhitungan jarak antar dokumen menggunakan metode *euclidean distance*.

3. Dilakukan proses *clustering* menggunakan metode *single linkage* yaitu mencari jarak minimum dari dokumen satu ke dokumen lainnya atau antar dokumen, apabila memiliki jarak terdekat, maka akan digabung menjadi satu *cluster*.

4. Jika jumlah cluster yang kita inginkan belum tercapai, maka akan mengulang proses *clustering* sampai cluster yang kita minta bisa tercapai, dan jika jumlah cluster yang kita inginkan sudah sesuai, maka akan terbentuk *cluster*.

5. Setelah terbentuk cluster, masing-masing dokumen pada setiap cluster akan digabungkan, dan kemudian diambil ringkasannya dari masing-masing cluster dengan presentase ringkasan 50%. Jadi, setiap cluster memiliki sebuah ringkasan.

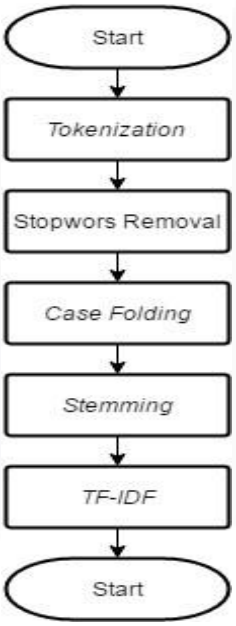
6. **2.2 *Use Case Diagram* Sistem**



Gambar 2.2 *Use Case Diagram* sistem

2.3 *Preprocessing* Sistem

Preprocessing dilakukan untuk mengubah data mentah untuk dapat diolah atau dieksekusi pada proses berikutnya. Proses *Preprocessing* sendiri terdiri atas *tokenization*, *stopwords removal*, *case folding*, *stemming*, *TF-IDF*.



Gambar 2.3 *Flowchart Preprocessing*

Preprocessing merupakan pemrosesan awal dokumen agar diperoleh suatu nilai yang dapat dipelajari oleh sistem *clustering* [1]. Data *preprocessing* merupakan langkah untuk mengubah data mentah menjadi data atau format yang sesuai untuk tahap analisis berikutnya. Transformasi data, seperti normalisasi, yaitu sebagai proses penyetaraan jumlah kata yang berbeda-beda pada setiap dokumen dapat diterapkan. Sebagai contoh, normalisasi dapat meningkatkan akurasi dan efisiensi algoritma *mining* yang melibatkan pengukuran jarak.

2.3.1 *Case Folding*

Case folding merupakan suatu tahap yang mengubah huruf besar menjadi huruf kecil [1].

2.3.2 *Tokenization*

Tokenization adalah proses pemotongan seluruh urutan karakter menjadi satu potongan kata [1].

2.3.3 *Stopword Removal*

Stopword removal merupakan proses penghapusan semua kata yang tidak memiliki makna [1].

2.3.4 *Stemming*

Stemming adalah proses membentuk suatu kata menjadi kata dasarnya [1]. Proses penghilangan semua imbuhan (*affix*) yang terdiri dari awalan (*prefix*), sisipan (*infix*), akhiran (*suffix*) dan duplikasi [2]. Penelitian ini menggunakan algoritma *porter stemmer*. Algoritma *Porter Stemmer* adalah proses penghilangan akhiran *morphological* dan *inflexional* yang umumnya terdapat dalam bahasa Inggris (Porter,1980). Proses *stemming* dilakukan untuk mendapatkan kata dasar dari kata berimbuhan.

2.3.5 *Term Weighting*

Term weighting merupakan proses pemberian bobot suatu *token* dalam suatu *term* [1].

2.3.5.1 *Term Frequency*

Term Frequency (TF) adalah pembobotan yang menghitung frekuensi kemunculan sebuah *token* pada suatu dokumen [1].

2.3.5.2 *Document Frequency*

Document Frequency (DF) adalah pembobotan yang menghitung frekuensi kemunculan sebuah *token* pada kumpulan dokumen [1].

$$IDF\ (word) = log \frac{total\ document}{document\ frequency} \quad (2.1)$$

2.3.5.3 Pembobotan *TF-IDF*

Pembobotan TF-IDF adalah suatu pengukuran statistik untuk mengukur seberapa penting sebuah token dalam kumpulan dokumen [1]. Pada tugas akhir ini, metode yang digunakan untuk peringkasan teks otomatis adalah menggunakan metode *TF-IDF*, metode *TF-IDF* digunakan untuk meringkas dokumen sebelum dokumen tersebut di *cluster*.

$$w(word\ i) = TF(word\ i) * IDF\ (word\ i) \quad (2.2)$$

2.3.5.4 Normalisasi

Normalisasi merupakan proses penyetaraan jumlah kata yang berbeda-beda pada setiap dokumen [1].

$$w(word_i) = \frac{w(wordi)}{\sqrt{w^2(wordi)+w^2(word2)+\ldots+w^2(wordn)}} \quad (2.3)$$

2.3.6 *Vector Space Model* (VSM)

Vector Space Model (VSM) adalah metode yang digunakan untuk merepresentasikan data atau query dalam bentuk vecktor [1].

2.3.7 *Distance Space*

Distance Space adalah proses perhitungan jarak antara suatu dokumen dengan dokumen lainnya [1]. Pada tugas akhir ini, *Distance Space* yang digunakan adalah *Euclidean Distance*, dengan rumus sebagai berikut :

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2} + \sqrt{|x_{i2} - x_{j2}|^2} + \cdots + \sqrt{|x_{in} - x_{jn}|^2} \quad (2.4)$$

2.4 *Summarize*

Ringkasan adalah suatu teks yang dihasilkan dari satu atau lebih teks yang berisi bagian informasi yang signifikan dalam teks asal, dan yang tidak lebih dari setengah teks aslinya (Hovy, Mitkov, 2005). Ringkasan teks (*text summarization*) adalah suatu proses penyulingan sebagian besar infromasi penting dari sumber (beberapa sumber) untuk menghasilkan suatu ringkasan bagi pengguna (Mani, House, Klein, 1999) [3].

Peringkasan multi-dokumen adalah suatu prosedur untuk menghasilkan informasi yang dianggap penting dari sejumlah dokumen dari *cluster* yang sama, dan disajikan dalam bentuk dokumen yang lebih ringkas dengan tidak mengurangi atau menghilangkan informasi utama pada dokumen tersebut. Dokumen yang ringkas dapat diukur dengan melihat ukurannya yang lebih pendek dari dokumen sumber dan padat akan informasi [4].

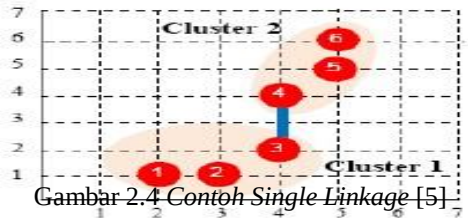
2.5 *Single Linkage Clustering*

Single linkage merupakan bagian dari *agglomerative hierarchical clustering.* Implementasi *single linkage* dilakukan dengan cara menghitung tingkat kemiripan antar cluster untuk jarak terdekat pada setiap cluster. Dan akan didapatkan hasil tingkat kemiripan antar cluster hingga terbentuk sebuah cluster tunggal. Hasil suatu clusternya dapat digambarkan secara grafik yang ditampilkan dalam bentuk diagram pohon atau dendogram [3].

Pada metode *single linkage* (MIN), kedekatan di antara dua *cluster* ditentukan dari jarak terdekat di antara pasangan di antara dua data *cluster* berbeda (satu dari *cluster* pertama satu dari *cluster* yang lain) atau disebut juga nilai kemiripan yang maksimal. Maka dengan cara ini kita memulainya dari masing-masing data sebagai *cluster*, kemudian mencari tetangga terdekat dan menggunakan *single linkage* untuk menggabungkan dua *cluster* berikutnya sampai semuanya bergabung menjadi satu *cluster*. Metode ini bagus untuk menangani set data yang bentuk distribusi datanya non-elips (*non-eliptical shapes*), tapi sangat sensitif terhadap *noise* dan *outlier*. Pengukuran jarak dua cluster dalam *single linkage* menggunakan formula jarak minimal (*minimal proximity*) [3].

Cara dari metode *single linkage clustering* adalah : [3]

a. Menentukan k sebagai jumlah *cluster* yang ingin dibentuk.
b. Setiap data dianggap sebagai *cluster*, jika n = jumlah data dan c = jumlah *cluster*, maka c = n.
c. Menghitung jarak / *similarity / dissimilarity* antar *cluster*.
d. Cari dua *cluster* yang mempunyai jarak antar *cluster* yang minimal dan gabungkan (c = c-1). Setelah semua jarak diketahui, selanjutnya adalah dikelompokkan dokumen-dokumen yang memiliki jarak terdekat.
e. Jika c > 3, kembali ke langkah 3.



Gambar 2.4 *Contoh Single Linkage* [5]

3. Pengujian dan Analisis

3.1 Pengjujian *Summarize*

Skenario yang dilakukan untuk melakukan pengujian *summarize* adalah dengan cara memberikan sepuluh dokumen teks yang diambil dari situs online *www.goal.com* dan *www.cnn.com* kepada seorang responden. Kemudian, sepuluh dokumen tersebut akan diringkas secara manual oleh responden secara manual dengan persentase ringkasan atau rangkuman sebesar 30%, 40%, dan 50% dari semua jumlah kalimat dalam satu dokumen teks. Setelah itu hasil ringkasan dari responden akan dibandingkan akurasinya dengan ringkasan yang tercipta oleh sistem.

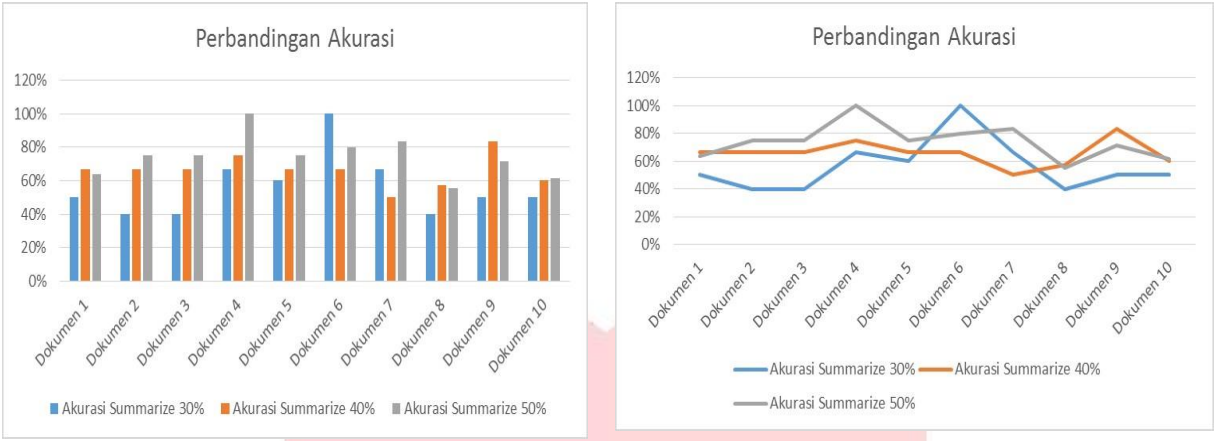
$$\text{Akurasi Ringkasan} = \frac{Jumlah\ Kalimat\ Sistem}{Jumlah\ Kalimat\ Manual} * 100\% \quad (3.1)$$

Dimana,

a. Jumlah kalimat sistem = Jumlah kalimat ringkasan yang tercipta oleh sistem dan relevan terhadap hasil ringkasan manual.
b. Jumlah kalimat manual = Jumlah kalimat ringkasan yang tercipta secara manual yaitu oleh responden, dan hasil ringkasan manual dianggap yang paling benar.

3.1.1 Perbandingan Akurasi *Summarize*

Perbandingan akurasi *summarize* dilakukan untuk mengetahui sejauh mana kemampuan sistem dalam mengolah data. Kemudian persentase mana yang terbaik dalam mengolah data menjadi ringkasan tanpa menghilangkan intisari yang terdapat dalam suatu dokumen teks. Perbandingan akurasi *summarize* dilakukan dengan membandingkan persentase 30%, 40%, dan 50%. Masing-masing persentase dilakukan terhadap sepuluh dokumen teks. Grafik batang dan grafik garis antara ketiga persentase tersebut adalah sebagai berikut :



Gambar 3.1 Grafik Perbandingan Akurasi *Summarize*

Dari grafik diatas dapat dilihat bahwa tingkat akurasi terkecil dari ketiga persentase, paling banyak terdapat pada *summarize* dengan persentase 30%, kemudian diikuti oleh *summarize* dengan persentase 40%, dan tingkat akurasi terbesar adalah *summarize* dengan persentase 50%.

a. *Summarize* dengan persentase 30% memiliki nilai rata-rata sebesar 56.40% untuk sepuluh dokumen teks.

b. *Summarize* dengan persentase 40% memiliki nilai rata-rata sebesar 66% untuk sepuluh dokumen teks.

c. *Summarize* dengan persentase 50% memiliki nilai rata-rata sebesar 74.10% untuk sepuluh dokumen teks.

Dapat disimpulkan bahwa *summarize* dengan persentase 50% untuk sepuluh dokumen teks dengan total kalimat rata-rata 16.2 kalimat adalah yang terbaik dari ketiganya karena memiliki tingkat akurasi rata-rata paling tinggi, yaitu sebesar 74.10%.

3.2 Pengujian *Clustering*

Skenario yang akan diujikan pada pengujian ini adalah dengan melakukan *clustering* sebanyak 30 dokumen teks. Masing-masing dokumen teks memiliki kelas berbeda, yaitu lima berita olah raga, lima berita teknologi, lima berita *fashion*, lima berita ekonomi, lima berita *entertainment*, dan lima berita *traveling*. Pada pengujian ini akan dilakukan sebanyak dua kali percobaan dan dengan dua skenario yang berbeda. Skenario yang dilakukan adalah memasukkan jumlah *cluster* sebanyak 6 *cluster* dan 12 *cluster* yang nantinya akan dibandingkan tingkat akurasinya dengan metode matriks similaritas.

Tabel 3.1 Matriks Kontingen Menentukan Kelas Sama dan Berbeda [6]

fij	*Cluster* Sama	*Cluster* Beda
Kelas Sama	*f11*	*f10*
Kelas Beda	*f01*	*f00*

Matriks yang dihitung berdasarkan matriks kontigen tersebut adalah nilai statistik rand dengan rumus sebagai berikut : [6]

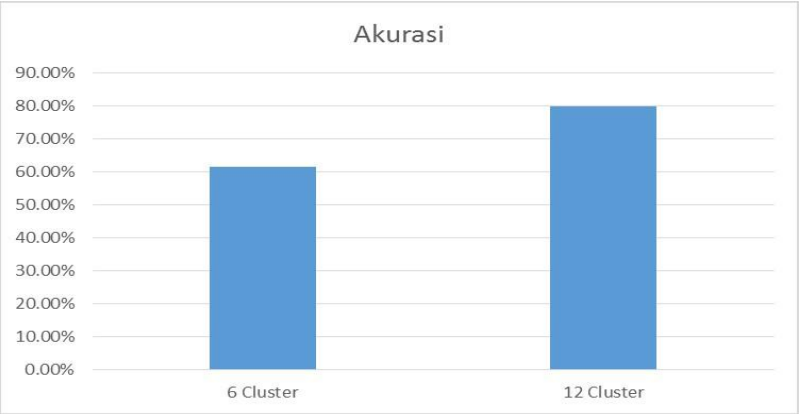
$$\text{Statistik Rand} = \frac{f00 + f11}{f00 + f01 + f10 + f11}$$

(3.2) Nilai elemen matriks kontingen ada 4 macam, dengan rincian sebagai berikut

: [6]

a. *f00* = Jumlah pasangan objek yang mempunjai kelas berbeda dan *cluster* berbeda.
b. *f01* = Jumlah pasangan objek yang mempunyai kelas berbeda dan *cluster* sama.
c. *f10* = Jumlah pasangan objek yang mempunyai kelas sama dan *cluster* berbeda.
d. *f11* = Jumlah pasangan objek yang memiliki kelas sama dan *cluster* sama.

3.2.1 Perbandingan Akurasi *Clustering*



Gambar 3.2 Perbandingan Akurasi *Cluster*

Berdasarkan grafik diatas menunjukkan bahwa akurasi yang dihasilkan pada percobaan 12 *cluster* lebih tinggi daripada percobaan 6 *cluster*. Pada percobaan 12 *cluster* menghasilkan nilai akurasi sebesar 79.8%, sedangkan pada percobaan 6 *cluster* menghasilkan nilai akurasi sebesar 61.5%. Artinya percobaan 12 *cluster* dengan total 30 dokumen teks adalah yang terbaik, karena tingkat akurasinya lebih tinggi. Hal tersebut membuktikan bahwa metode *single linkage clustering* cukup baik dalam mengelompokkan dokumen teks.

4. Kesimpulan

a Dapat disimpulkan bahwa summarize dengan persentase 50% untuk sepuluh dokumen teks dengan total kalimat rata-rata 16.2 kalimat adalah yang terbaik dari ketiganya karena memiliki tingkat akurasi rata-rata paling tinggi, yaitu sebesar 74.10%.

b Menurut penulis summarize dengan metode TF-IDF cukup baik karena memliki tingkat akurasi sebesar 74.10%

c Dari hasil pengujian clustering dengan dataset sebanyak 30 dokumen teks, tingkat akurasi single linkage clustering dengan jumlah 12 cluster mendapatkan hasil yang lebih baik dibandingkan dengan single linkage clustering dengan jumlah 6 cluster. Tingkat akurasi yang dihasilkan adalah sebesar 79.8%. Hal tersebut terjadi karena jumlah cluster yang diinputkan lebih banyak yaitu sebesar 12 cluster dari total 30 dokumen teks. Artinya cluster yang diinputkan hampir mendekati jumlah total dokumen.

d Menurut penulis metode single linkage cukup baik untuk digunakan karena memiliki tingkat akurasi 79.8%.

Daftar Pustaka

[1] Handoyo, Rendy,. R. Rumani, dan M, Surya Michrandi Nasution. 2014. ***Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K - Means Pada Pengelompokan Dokumen.*** Bandung : Universitas Telkom.

[2] Widiantoro, Agustinus. 2014. ***Peringkasan Teks Otomatis Pada Dokumen Berbahasa Jawa Menggunakan Metode Tf-Idf.*** Yogyakarta : Universitas Sanata Dharma.

[3] Noor, M. Helmi, dan Moch. Hariadi. 2009. ***Image Cluster Berdasarkan Warna Untuk Identifikasi Kematangan Buah Tomat Dengan Metode Valley Tracing.*** Surabaya : ITS.

[4] Akbar, Angga Aulia. 2015. ***Peringakasan Multi-dokumen Berita Berbahasa Indonesia menggunakan Conditional Random Fields (CRF).*** Bandung : Universitas Telkom.

[5] Berry, M.W. & Kogan, J. 2010. ***Text Mining Aplication and Theory.*** WILEY : United Kingdom.

[6] Prasetyo, Eko. 2014. ***Data Mining : Mengolah Data Menjadi Informasi Menggunakan Matlab.*** Yogyakarta : Penerbit Andi.