# Analisa Perbandingan Metode *Hierarchical Clustering*, K-means dan Gabungan Keduanya dalam *Cluster* Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)

Tahta Alfina, Budi Santosa, dan Ali Ridho Barakbah
Jurusan Teknik Industri, Fakultas Teknologi Industri, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111
*E-mail*: budi_s@ie.its.ac.id

***Abstrak*— Saat ini, konsep data mining semakin dikenal sebagai *tools* penting dalam manajemen informasi karena jumlah informasi yang semakin besar jumlahnya. Salah satu teknik yang dikenal dalam data mining adalah *clustering*, berupa proses pengelompokan sejumlah data atau objek ke dalam *cluster* (*group*) sehingga setiap dalam *cluster* tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam *cluster* yang lainnya. *Clustering* memiliki dua metode, yaitu partisi dan hierarki. Dua metode ini memiliki kelebihan dan kekurangan masing-masing, dan dengan menggabungkan keduanya dapat diperoleh hasil *cluster* yang lebih baik. Dari hasil *cluster* dengan menggunakan data problem Kerja Praktek Jurusan Teknik Industri ITS, maka diperoleh hasil bahwa gabungan metode *Single Linkage Clustering* dan K-means memberikan hasil *cluster* yang lebih baik dengan parameter uji *cluster variance* dan metode *silhouette* coefisien.**

*Kata Kunci*— Kerja Praktek, *Document Clustering*, *K-means, Hierarchical Clustering, Cluster Variance,* Metode *Silhouette Coeficient.*

## I. PENDAHULUAN

Saat ini, konsep data mining semakin dikenal sebagai *tools* penting dalam manajemen informasi karena jumlah informasi yang semakin besar jumlahnya. Data mining sendiri sering disebut sebagai *knowledge discovery in database (KDD)* adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola hubungan dalam set data berukuran besar [1]. *Output* dari data mining ini dapat digunakan untuk pengambilan keputusan di masa depan.

Salah satu teknik yang dikenal dalam data mining yaitu *clustering*. Pengertian *clustering* keilmuan dalam data mining adalah pengelompokan sejumlah data atau objek ke dalam *cluster* (*group*) sehingga setiap dalam *cluster* tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam *cluster* yang lainnya. Sampai saat ini, para ilmuwan masih terus melakukan berbagai usaha untuk melakukan perbaikan model *cluster* dan menghitung jumlah *cluster* yang optimal sehingga dapat dihasilkan *cluster* yang paling baik. Ada dua metode *clustering* yang kita kenal, yaitu *hierarchical clustering* dan *partitioning*. Metode *hierarchical clustering* sendiri terdiri dari *complete linkage clustering*, *single linkage clustering*, *average linkage clustering* dan *centroid linkage clustering*.Sedangkan metode *partitioning* sendiri terdiri dari k-means dan fuzzy k-means.

Metode *K-means* merupakan metode *clustering* yang paling sederhana dan umum [1]. Hal ini dikarenakan *K-means* mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien [2]. Namun, *K-means* mempunyai mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil *cluster* yang terbentuk dari metode *K-means* ini sangatlah tergantung pada inisiasi nilai pusat awal *cluster* yang diberikan [1]. Hal ini menyebabkan hasil *cluster*nya berupa solusi yang sifatnya *local optimal*. Untuk itu, maka *K-means* dikolaborasikan oleh metode hierarki untuk penentuan pusat awal *cluster*. Metode hierarki yang akan dicoba diterapkan dalam penelitian ini adalah kelima metode hirarchical *clustering* yang telah disebutkan sebelumnya. Kelima metode ini akan dibandingkan untuk melihat *cluster* mana yang memberikan hasil pengelompokan yang lebih baik. Dari proses pengelompokan ini nantinya diharapkan akan diketahui kemiripan atau kedekatan antar data sehingga dapat dikelompokkan ke dalam beberapa *cluster*, dimana antar anggota *cluster* memiliki tingkat kemiripan yang tinggi.

Data yang digunakan dalam penelitian ini adalah data teks. Dimana data ini merupakan data problem kerja praktek Jurusan Teknik Industri ITS yang disampaikan oleh mahasiswanya melalui forum diskusi jejaring sosial *facebook*. Sehingga dalam penelitian ini nantinya akan dijelaskan bagaimana cara mengelompokkan problem kerja praktek berdasarkan *posting* problem yang ada pada forum diskusi *online* SI-KP yang ada di jejaring sosial *facebook.* Metode yang digunakan adalah metode *document clustering* dengan *K-means* dan *hierarchical clustering*.

Sebelumnya, penelitian mengenai problem pengelompokan dokumen telah banyak dilakukan melalui berbagai metode.

Misalnya penggunaan metode *K-Nearest Neigbour* (KNN) untuk kategorisasi teks [3], klasifikasi dokumen berbahasa indonesia dengan algoritma *single pass clustering* [4], *clustering based on frequent word and sequence* dan *K-means* [5], pengelompokan data teks dengan *fuzzy c-means* [6] dan beberapa penelitian lain dengan metode yang hampir serupa.

## II. METODOLOGI PENELITIAN

Pada bagian metodologi penelitian ini akan diuraikan langkah-langkah sistematis dan terarah yang akan dijadikan acuan sebagai kerangka penelitian penentuan kemiripan problem kerja praktek di Jurusan Teknik Industri dengan menggunakan kombinasi semua metode *Hierarchical clustering* dan *K-means* sehingga dapat diketahui metode manakah yang menghasilkan hasil *cluster* yang terbaik.

### *A. Tahap Pengolahan Data Teks ke dalam Metadata*

Dari semua data yang diperoleh, dipilih *keyword-keyword* yang dapat mewakili problem Kerja Praktek Jurusan Teknik Industri ITS pada tahun 2011. *Keyword* yang terpilih nantinya akan digunakan digunakan untuk membentuk matriks metadata yang menunjukkan frekuensi dari setiap *keywords* dalam setiap problem yang disampaikan oleh mahasiswa. Pemilihan *keywords* dapat dilakukan dengan berbagai metode. Diantaranya dengan menggunakan metode *document clustering* atau menggunakan algoritma *text* mining. Dalam penelitian ini digunakan algoritma *document clustering* sederhana karena domain teks yang akan dibawa kedalam suatu *cluster* bersifat spesifik, yaitu problem kerja praktek Jurusan Teknik Industri ITS. Sehingga *keywords* yang akan digunakan dalam metadata dapat ditentukan secara manual oleh peneliti. Berbeda halnya dengan algoritma *text* mining. *Text* mining digunakan untuk mengelompokkan data dimana domainnya bersifat bebas. Sehingga harus melewati proses-proses dalam *text* mining seperti *tokenizing* adalah proses penghilangan tanda baca pada kalimat yang ada dalam dokumen sehingga menghasilkan kata-kata yang berdiri sendiri-sendiri dan tahap *filtering* adalah tahap pengambilan kata-kata yang penting dari hasil tokenizing dengan menggunakan algoritma *stoplist* atau *wordlist*.

Pada umumnya, penetuan *keywords* ditentukan oleh subyektifitas peneliti karena peneliti yang mengerti dan tahu tujuan yang ingin dicapai dalam penelitiannya. Namun, problem kerja praktek ini melibatkan banyak stakeholder yang ikut berperan disana. Oleh karena itulah, *keywords* yang dipilih pada permasalahan ini ditentukan dengan melakukan interview yang dibantu dengan sebuah kuisioner (terlampir). Adapun pihak yang diinterview adalah admin Kerja Praktek tahun 2011, Koordinator Kerja Praktek Jurusan Teknik Industri dan beberapa mahasiswa yang mengambil Kerja Praktek pada tahun 2011. Berikut ini adalah *list keywords* yang digunakan untuk membentuk matriks metadata :

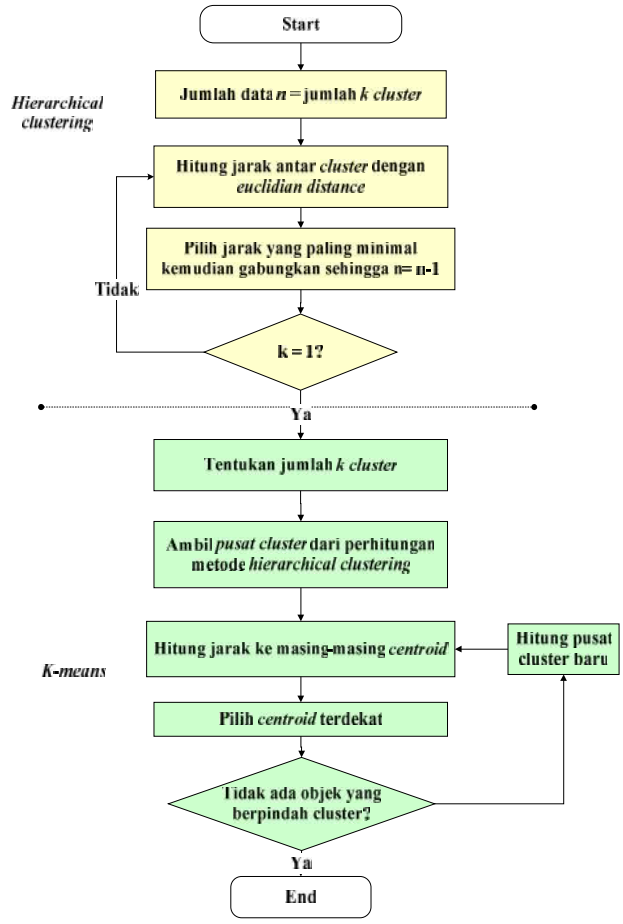
Tabel 1 List *Keywords* untuk Metadata

| No. | Keywords | No. | Keywords | No. | Keywords | No. | Keywords |
|---|---|---|---|---|---|---|---|
| **1** | SI-KP | **11** | Jurusan | **21** | eksternal | **31** | Username |
| **2** | KP | **12** | email | **22** | logbook | **32** | Approve |
| **3** | registrasi | **13** | Prosedur | **23** | PDF | **33** | Delete |
| **4** | Mahasiswa | **14** | aktivasi | **24** | Manual | **34** | Hapus |
| **5** | Tanggal | **15** | Arahan | **25** | Dashboard | **35** | Edit |
| **6** | User | **16** | Notifikasi | **26** | Login | **36** | Input |
| **7** | Akun | **17** | Dosen | **27** | online | **37** | Laporan |
| **8** | Admin | **18** | Pembimbing | **28** | password | **38** | Periode |
| **9** | Koordinator | **19** | nilai | **29** | group | **39** | Sosialisasi |
| **10** | perusahaan | **20** | Internal | **30** | Kelompok | **40** | Cetak |

Dari hasil interview dan survey yang dilakukan, terdapat 40 *keywords* yang akan digunakan untuk membentuk kolom matriks metadata. Sedangkan banyaknya data yang digunakan sebagai input sebanyak 327 data yang membentuk baris matriks metadata.

### *B. Tahap Clustering*

Pada tahap ini dilakukan pengelompokan data menggunakan kombinasi dua algoritma *clustering*, yaitu *hierarchical clustering* dan metode *K-means*. Dari algoritma *hierarchical clustering* ini digunakan untuk menentukan pusat *cluster*. Selanjutnya, pusat *cluster* yang diperoleh *hierarchical clustering* tersebut digunakan untuk proses pengelompokan data dengan menggunakan metode *K-means*. Gambar 1 adalah flowchart yang menjelaskan urutan pengerjaan penelitian dengan menggunakan metode *hierarchical clustering* dan *K-means*.




Gambar 1 Algoritma *Hierarchical clustering* dan *K-means*

Dari gambar flowchart tersebut dapat diketahui tentang urutan metode *clustering* dengan *hierarchical clustering* yang ditandai dengan warna kuning dan metode *K-means* ditandai dengan warna hijau. Pada metode *hierarchical clustering*, sebelum dilakukan pengelompokan, setiap data yang ada diasumsikan sebagai *cluster*. Hal ini jika terdapat jumlah data sebanyak *n,* dan *k* dianggap sebagai jumlah *cluster*, maka besarnya *n* = *k.* Kemudian, dihitung jarak antar *cluster*nya dengan menggunakan *Euclidian distance* berdasarkan jarak rata-rata antar objek. Selanjutnya, dari hasil perhitungan tadi dipilih jarak yang paling minimal dan digabungkan sehingga besarnya *n* = *n* -1. Hal ini akan terus dilakukan dan akan berhenti jika memenuhi kondisi jumlah *k* = 1. Pada akhir tahap *hierarchical clustering* ini akan diperoleh sebuah gambar dendrogram yang menunjukkan urutan pengelompokan masing-masing anggota dalam *cluster*.

Setelah sampai pada kondisi *k* = 1, maka dilanjutkan dengan metode *K-means*. Pada metode ini, seharusnya diawali oleh penentuan jumlah *k cluster* yang akan dibentuk, kemudian dilanjutkan dengan penentuan pusat awal *cluster* secara random. Namun, karena metode ini merupakan kombinasi antara *hierarchical clustering* dan *K-means*, maka penentuan pusat *cluster* untuk metode K-means ditentukan dengan mencari rata-rata dari data yang berada pada sebuah *cluster* hasil dari *hierarchical k-means*. Sehingga pada tahapan ini, pusat *cluster* metode *K-means* langsung dapat ditentukan. Selanjutnya, dihitung jarak anggota *cluster* ke setiap *centroid*nya. Setelah didapatkan hasilnya, anggota *cluster* dimasukkan kedalam *cluster* yang memiliki jarak yang paing dekat dengan *centroid*nya. Iterasi pada K-Means akan berhenti ketika semua data yang berada pada sebuah *cluster* tertentu tidak berpindah ke *cluster* yang lainnya.

### *C. Pengujian Performansi Algoritma*

Pada Subbab ini akan dilakukan pengujian tehadap hasil algoritma metode *clustering*. Pengujian ini dilakukan untuk melihat apakah kombinasi algoritma *hierarchical clustering* dengan K-means menghasilkan pengelompokan data yang lebih baik jika dibandingkan dengan metode hierarki itu sendiri maupun K-means. Adapun pengujian yang dilakukan adalah sebagai berikut :

1. Cluster Variance

Analisa ini digunakan untuk nilai penyebaran dari data-data hasil *clustering* dengan metode K-means. *Cluster variance* ini hanya digunakan untuk data yang bersifat *unsupervised*. Sedangkan pada data *supervised* digunakan *error ratio analysis [7].* Besarnya nilai varian sebuah *cluster* dapat dihitung dengan rumus berikut :

$$v_c^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} (d_i - \bar{d}_i)^2 \qquad (1)$$

Dimana :

$v_c^2$ = *variance* pada *cluster* ke *c*

$c$ = *1...k,*dimana *k* =jumlah *cluster*

$n_c$ = jumlah data pada *cluster c*

$d_i$ = data ke-*i* pada suatu *cluster*

$\bar{d}_i$ = rata-rata dari data pada suatu *cluster*

Ada dua macam *cluster variance*, yaitu *varian within cluster* ($V_w$) *dan varian between cluster* ($V_b$). $V_w$ digunakan untuk melihat hasil variansi penyebaran data yang ada pada sebuah *cluster* (*internal homogenity*). Semakin kecil nilai $V_w$, maka semakin baik *cluster*nya. Besarnya nilai $V_w$ dapat dihitung dengan rumus :

$$v_w = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1).v_i^2 \ldots \qquad (2)$$

Dimana :

N : jumlah semua data

k : jumlah *cluster*

$n_i$ : jumlah anggota dalam *cluster* ke-*i*

Sedangkan nilai ($V_b$) merupakan nilai yang digunakan untuk melihat hasil variansi penyebaran data antar *cluster* (*external homogenity*). Semakin besar nilai ($V_b$), maka semakin baik hasil *cluster*nya.besarnya nilai ($V_b$) dapat dihitung dengan rumus :

$$V_b = \frac{1}{k-1} \sum_{i=1}^{k} ni(\bar{i} - \bar{d})^2 \ldots\ldots \qquad (3)$$

Dimana

k : jumlah *cluster*

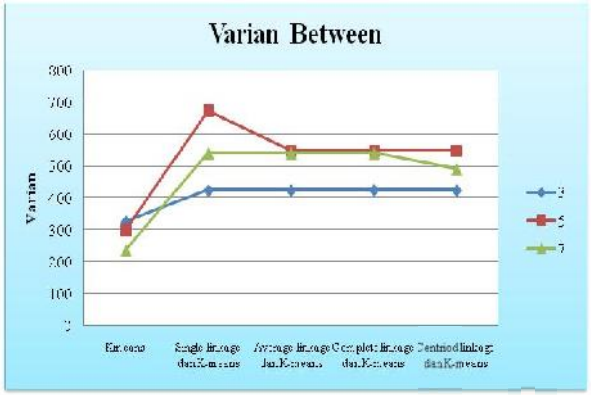
$\bar{d}$ : rata-rata dari $\bar{d}i$

Sedangkan untuk melihat varian dari semua *cluster* maka diukur dengan membandingkan nilai ($V_w$) dan ($V_b$) yaitu

$$V = \frac{V_w}{V_b} \qquad (4)$$

Nilai $V_w$ akan menunjukkan hasil yang semakin baik ketika nilainya semakin kecil. Sedangkan nilai $V_b$ akan menunjukkan hasil yang baik ketika nilainya semakin besar. Maka dari sini, nilai *V* dari semua *cluster* akan semakin baik jika nilainya semakin kecil.

Dengan menggunakan rumus-rumus diatas, maka diperoleh hasil cluster variance untuk 3,5 dan 7 cluster sebagai berikut :




Gambar 2 Grafik perbandingan nilai $V_b$

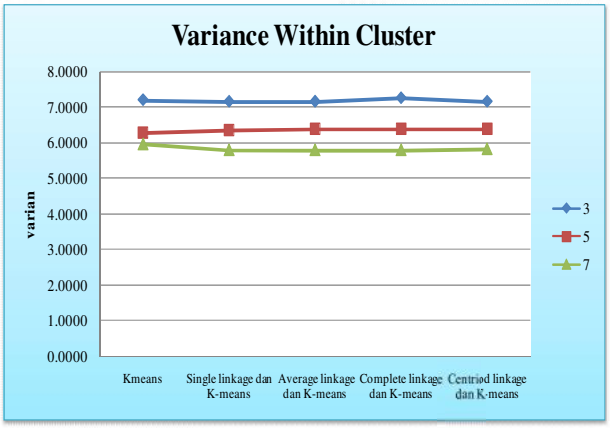
Gambar 2 menunjukkan perbandingan nilai ($V_b$) dari masing-masing metode dengan 3 skenario *cluster* yang telah dibuat. Jika ditinjau dari segi jumlah *cluster*nya, proses *clustering* yang menghasilkan ($V_b$) terbaik adalah dihasilkan oleh data yang di*cluster*kan kedalam 5 *cluster*. Hal ini menunjukkan bahwa jumlah *cluster* sebanyak 5 ini memberikan hasil penyebaran data yang baik dibandingkan dengan 3 *cluster* atau 7 *cluster*. Sedangkan jika ditinjau dari segi metode, nilai $V_b$terbesar dihasilkan oleh metode *single linkage clustering*. Pada umumnya, metode average *linkage* clustering memberikan hasil yang lebih baik jika dibandingka dengan metode yang lainnya. Namun hal tersebut berlaku jika

data yang digunakan dalam menguji *cluster* berupa data set, misalnya data *iris,* data *russpini* dan lain sebagainya. Sedangkan data yang digunakan dalam penelitian ini adalah data yang bersifat non-globular atau menyebar.

Sedangkan nilai $V_w$ akan menunjukkan hasil yang semakin baik ketika nilainya semakin kecil. Untuk lebih mudah dalam melakukan analisa, nilai $V_w$ untuk semua metode dan semua *cluster* yang telah diperoleh dari perhitungan pada bab 4 diplotkan ke dalam sebuah grafik seperti yang ditampilkan pada gambar 3 dibawah.

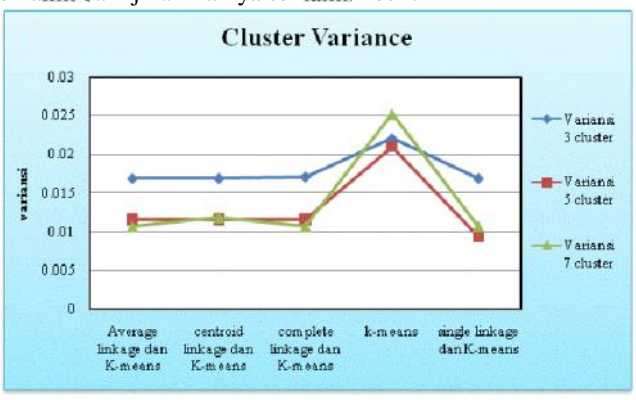
Jika dilihat dari segi metode yang digunakan, besarnya nilai $V_w$ yang dihasilkan tidak menunjukkan perbedaan yang terlalu signifikan. Sedangkan jika ditinjau dari jumlah *cluster*, nilai $V_w$ terkecil dihasilkan oleh data yang terbentuk kedalam 7 *cluster*. Hal ini dikarenakan semakin banyak *cluster* yang dibentuk, semakin banyak data yang dapat masuk ke dalam data *cluster* yang berbeda-beda sehingga menghasilkan variansi cluster yang semakin kecil.

Selanjutnya, nilai kedua varian tersebut dapat dibandingkan sehingga nantinya dapat digunakan untuk melihat variansi dari semua *cluster* maka diukur dengan membandingkan nilai $(V_w)$ dan $(V_b)$. Maka dari sini, nilai $V$ dari semua *cluster* akan semakin baik jika nilainya semakin kecil.




Gambar 3 Grafik perbandingan nilai $V_w$

Selanjutnya, nilai kedua varian tersebut dapat dibandingkan sehingga nantinya dapat digunakan untuk melihat variansi dari semua *cluster* maka diukur dengan membandingkan nilai $(V_w)$ dan $(V_b)$. Maka dari sini, nilai $V$ dari semua *cluster* akan semakin baik jika nilainya semakin kecil.

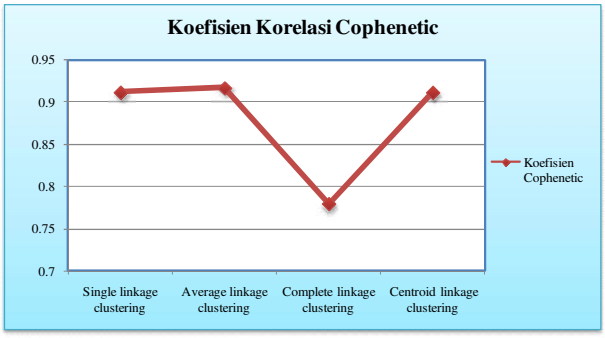



Gambar 4 Grafik perbandingan nilai V

Nilai $V$ yang dihasilkan disini menunjukkan variansi total dari setiap *cluster* yang dihasilkan. Dari Gambar 19, terlihat bahwa hasil *cluster* terbaik dihasilkan oleh data yang dibagi ke dalam 5 *cluster* untuk semua metode karena mempunyai nilai $V$ yang paling kecil jika dibandingkan dengan jumlah *cluster* yang lain. Sedangkan jika ditinjau dari segi metode, algoritma K-Means menghasilkan varian *cluster* yang terbesar jika dibandingkan metode K-means yang pusat *cluster*nya diinisiasi dari algoritma *hierarchical clustering*. Ini berarti metode k-means tidak dapat menghasilkan *cluster* yang lebih baik jika dibandingkan dengan kombinasi metode *hierarchical clustering* dan K-means. Sedangkan untuk metode kombinasi *hierarchical clustering* dan K-means, pada kasus ini hasil terbaiknya dihasilkan oleh K-means yang pusat awalnya diinisiasi oleh *single linkage clustering*.

2. Koefisien Korelasi Cophenetic

Selain menggunakan analisa *cluster variance*, analisa *cluster* juga dapat dilakukan dengan menggunakan koefisien korelasi *cophenetic*. Namun analisa ini hanya terbatas pada *cluster* yang dibentuk dengan menggunakan algoritma *hierarchical clustering*. Nilai koefisien korelasi cophenentic terbesar adalah *Average Linkage Clustering* yaitu sebesar 0.9171, sedangkan nilai koefisien korelasi *cophenetic* yang paling kecil adalah pada *Complete Linkage Clustering* yaitu sebesar 0.7798. Sedangkan nilai dua metode *linkage* yang lain yaitu *Single Linkage Clustering* dan *Centroid Linkage Clustering* masing-masing sebesar 0.9118 dan 0.9117. untuk lebih mudahnya, perbandingan nilai koefisien korelasi *Cophenetic* ini dapat dilihat pada gambar Grafik berikut ini :

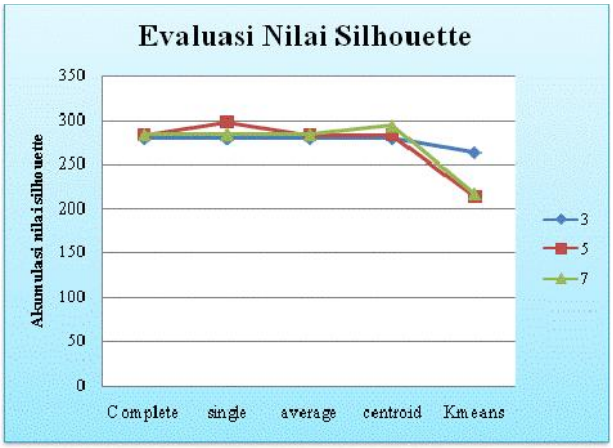



Gambar 5 Grafik Koefisien Korelasi *Cophenetic*

Besarnya nilai ini harus sangat dekat dengan 1 untuk solusi yang lebih baik. Ukuran ini dapat digunakan untuk membandingkan solusi *cluster* alternatif diperoleh dengan menggunakan algoritma yang berbeda. Sehingga, dari penjelasan tersebut dapat disimpulkan semakin besar (semakin mendekati 1) nilai koefisien korelasi *cophnenetic*nya, maka semakin baik pula hasil *cluster*nya. Dari sini, dapat kita simpulkan bahwa metode hierarki yang paling baik dalam membentuk suatu *cluster* adalah *average linkage clustering*. Hal ini dikarenakan metode ini merupakan satu-satunya metode *clustering* yang memperhitungan setiap jarak antar titiknya dalam menentukan urutan membentuk *cluster*.

3. Metode Silhouette Coefficient

Analisa metode *silhouette* ini dilakukan dengan melihat besar nilai *s* dari hasil perhitungan dengan menggunakan bantuan *software* MatLab. Hasil perhitungan nilai *silhouette* coeffisien dapat bervariasi antara -1 hingga 1. Jika $s_i = 1$ berarti objek *i* sudah berada dalam *cluster* yang tepat. Jika nilai $s_i = 0$ maka objek *i* berada di antara dua *cluster* sehingga objek tersebut tidak jelas harus dimasukan ke dalam *cluster* A atau *cluster* B. Akan tetapi, jika $s_i = -1$ artinya struktur *cluster* yang dihasilkan *overlapping*, sehingga objek *i* lebih tepat dimasukan ke dalam *cluster* yang lain. Untuk mempermudah dalam melakukan analisa, nilai *s* dikonversikan ke dalam dua nilai, yaitu 1 jika nilai *silhouette*nya lebih besar dari 0 dan bernilai 0 jika nilai *silhouette*nya lebih kecil dari nol. Sehingga ketika hasil dari penjumlahan nilai *s* yang dikonversikan tadi jumlahnya paling besar diantara *cluster* yang lainnya, maka artinya hasil *cluster* yang dihasilkan merupakan *cluster* yang terbaik karena semakin sedikit nilai *s* yang nilainya dibawah 0. Berikut ini adalah gambar berikut ini :




Gambar 6 Evaluasi Nilai *Silhouette* dengan Jumlah *Cluster* sebanyak 3

Untuk skenario jumlah cluster sebanyak 3, semua metode hierarki yang digabungkan dengan K-means memberikan hasil *cluster* yang sama dan lebih baik jika dibandingkan dengan metode K-means itu sendiri. Untuk skenario jumlah cluster yang digunakan sebanyak 5, dapat dilihat bahwa penjumlahan nilai *s* terbesar diperoleh ketika peng*cluster*an dilakukan dengan menggunakan metode *single* lingkage *clustering* yang dikombinasikan dengan K-means, diikuti oleh 3 metode Hierarchical *Clustering* yang lainnya yang digabungkan dengan K-means dan penjumlahan nilai *s* yang paling kecil dihasilkan oleh metode K-means.

Berbeda dengan evaluasi nilai *s* yang dihasilkan oleh 5 *cluster*, pada evaluasi nilai *s* dengan 7 *cluster*, metode yang memberikan hasil yang terbaik adalah metode gabungan *Centroid Linkage Clustering* dan K-means sebesar 294 data di*cluster*kan pada *cluster* yang tepat. Kemudian baru diikuti oleh 3 metode Hierarchical *Clustering* yang lainnya yang digabungkan dengan K-means dan penjumlahan nilai *s* yang paling kecil dihasilkan oleh metode K-means.

## III. KESIMPULAN

Adapun kesimpulan dari penelitian ini adalah sebagai berikut :

1. Kombinasi algoritma *hierarchical clustering* dan k-means menghasilkan pengelompokan data yang lebih baik jika dibandingkan dengan k-means dalam semua pengujian.
2. Dengan evalusi koefisien *cophenetic*, metode *clustering* terbaik dihasilkan oleh *average linkage clustering*
3. Dalam studi kasus Problem Kerja Praktek Jurusan Teknik Industri ITS, dari kombinasi *hierarchical clustering* dan K-means yang ada, kombinasi *single linkage clustering* dan K-means menghasilkan pengelompokan data yang terbaik dibandingkan dengan metode hierarki yang lainnya.

## DAFTAR PUSTAKA

[1] B. Santosa, *Data Mining. Teknik Pemanfaatan Data untuk Keperluan Bisnis*, First Edition ed. Yogyakarta: Graha Ilmu, (2007).

[2] K. Arai and A. R. Barakbah, "Hierarchical K-means: an algorithm for centroids initialization for K-means," (2007).

[3] S. Jiang, *et al.*, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications,* vol. 39, pp. 1503-1509, (2011).

[4] A. Z. Arifin and A. N. Setiono, "*Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*," (2002).

[5] N. R. Widyawati, *et al.*, *Perbandingan Clustering Based on Frequent Word Sequnce dan K-Means untuk Pengelompokan Dokumen Berbahasa Indonesia*, ( 2011).

[6] C.-x. Li and N. Lin, "A Novel Text Clustering Algorithm," *Energy Procedia,* vol. 13, pp. 3583-3588, (2011).

[7] A. R. Barakbah and Y. Kiyoki, "A pillar algorithm for k-means optimization by distance maximization for initial centroid designation," (2009), pp. 61-68.