Perbandingan Algoritma Pengelompokan Non-Hierarki untuk *Dataset* Dokumen

Dyah Herawatie
Prodi Sistem Informasi
Fakultas Sains dan Teknologi Universitas Airlangga
Surabaya, Indonesia
dy4h_h3r4@yahoo.com
dyah.herawatie@fst.unair.ac.id

Eto Wuryanto, Purbandini
Prodi Sistem Informasi
Fakultas Sains dan Teknologi Universitas Airlangga
Surabaya, Indonesia
eto-w@fst.unair.ac.id, purbandini@fst.unair.ac.id
etowuryanto@gmail.com, purbandini03@gmail.com

***Abstract*—Tujuan penelitian ini adalah membandingkan beberapa algoritma pengelompokan non-hierarki : K-Means, Bisecting K-Means, K-Median dan K-Medoid untuk data dokumen. Perbandingan dilakukan dengan menggunakan F-*measure* dan akurasi sebagai ukuran kualitas ketepatan hasil pengelompokan. Dalam penelitian ini data yang dipakai untuk eksperimen adalah artikel media masa yang berbahasa Indonesia yang diambil dari website Kompas (www.kompas.com) dan Detik (www.detik.com).**

Tahapan yang digunakan dalam penelitian ini adalah sebagai berikut : Tahap pertama adalah pengambilan dan pemrosesan data. Sebelum dilakukan pengelompokan data, data dokumen terlebih dahulu harus melalui tahap pra proses, yaitu *Detagging*, *Stopword removal*, dan *Stemming*. Hal ini dimaksudkan agar kata-kata yang digunakan untuk membentuk *term-document matrix* hanyalah kata-kata khusus yang dapat merepresentasikan dokumen yang ada. Langkah selanjutnya adalah penentuan fitur dan penyusunan *term-document matrix*. Jumlah fitur yang digunakan untuk pengelompokan dinyatakan dengan persentase dari total fitur. Langkah terakhir : melakukan pengelompokan data set dokumen dengan menggunakan algoritma K-Mean, Bisecting K-Mean, K-Median, dan K-Medoid dengan menggunakan jarak Euclid dan cosinus.

Dari hasil eksperimen dapat diambil kesimpulan bahwa algoritma pengelompokan yang memberikan hasil yang terbaik adalah K-Mean. Disamping itu Bi-secting K-Mean juga menghasilkan pengelompokan yang memuaskan. Selain itu jarak cosinus juga memberikan hasil yang paling baik dibandingkan dengan jarak Euclid. Dari hasil ekperimen, hanya dengan menggunakan 10% sampai 30% fitur yang digunakan telah menghasilkan pengelompokan yang memuaskan.

***Keywords—Clustering* Dokumen, Algoritma K-Mean, Bisecting K-Mean, K-Median, K-Medoid, *F-Measure*, Akurasi.**

I. PENDAHULUAN

Perkembangan yang pesat dalam informasi digital telah menyebabkan semakin meningkat pula volume informasi yang berbentuk teks. Diantara berbagai bentuk informasi digital, diperkirakan 80% dokumen digital adalah dalam bentuk teks [1]. Menurut Bridge [2], hal yang lebih menyulitkan dalam analisis adalah sekitar 80% – 85% bentuk informasi tersebut dalam format tidak terstruktur. Melimpahnya informasi teks tidak terstruktur telah mendorong munculnya disiplin baru dalam analisis teks, yaitu *text mining* yang mencoba menemukan pola-pola informasi yang dapat digali dari suatu teks tidak terstruktur tersebut.

Salah satu masalah yang penting dan menarik untuk diteliti dalam *data mining* atau *knowledge discovery* adalah pengelompokan dokumen. Pengelompokan dokumen bertujuan membagi dokumen dalam beberapa kelompok (*cluster*) sedemikian hingga dokumen-dokumen dalam cluster yang sama (*intra-cluster*) memiliki kesamaan yang tinggi, sementara dokumen-dokumen dalam cluster yang berbeda (*inter-cluster*) memiliki kesamaan yang rendah.

Analisis kelompok (*Cluster Analysis*) merupakan salah satu metode yang ada dalam analisis multivariate. Tujuan dari analisis ini adalah mengelompokkan obyek-obyek pengamatan menjadi beberapa kelompok berdasarkan variabel-variabel yang diamati, sedemikian hingga obyek-obyek yang terdapat pada kelompok yang sama mempunyai karakteristik yang relatif homogen (mirip) dan obyek-obyek antar kelompok mempunyai karakteristik yang berbeda (tidak mirip) [3].

Menurut Halkidi dkk [4] algoritma pengelompokan dapat diklasifikasikan menjadi empat jenis, yaitu pengelompokan non-hierakri (partisi), pengelompokan hirarki, pengelompokan berdasarkan densitas, dan pengelompokan berdasarkan *grid*. Metode pengelompokan hirarki digunakan untuk mengelompokkan pengamatan secara terstruktur berdasarkan sifat kemiripannya, dan kelompok yang diinginkan belum diketahui banyaknya. Sedangkan metode pengelompokan nonhierarki, digunakan untuk mengelompokkan obyek-obyek pengamatan menjadi k kelompok.

Dalam pengelompokan dataset dokumen, telah banyak penelitian yang membahas tentang algoritma hierarki. Sedangkan untuk algoritma nonhierarki yang sering digunakan adalah algoritma K-Mean. Bahkan telah dikembangkan algoritma Bisecting K-Mean yang merupakan pengembangan dari K-Mean. Sedangkan algoritma

nonhierarki yang lain, belum banyak diulas. Algoritma yang lain di antaranya K-Mode, Partitioning Around Medoids (PAM), Clustering Large Applications (CLARA), Clustering LARge Aplications based on RANge Search (CLARANS), Fuzzy C-Means [4]. K-Median [5], dan Aproximated K-Median. Menurut Chu dkk [9], Partitioning Around Medoids (PAM) atau algoritma K-Medoid dikatakan lebih robust (tegar) terhadap outlier jika dibandingkan dengan algoritma yang lain.

Penelitian ini hanya membandingkan empat algoritma pengelompokan nonhierarki, yaitu K-Means, Bi-Secting K-Means, K-Median, K-Medoid. Hal ini disebabkan karena beberapa algoritma merupakan pengembangan dari algoritma yang dibahas. Seperti algoritma CLARA dan CLARANS merupakan pengembangan dari K-Medoid. Algoritma Approximated K-Median merupakan pengembangan dari K-Median. Algoritma-algoritma ini menggunakan prinsip sampling dalam proses pengelompokan datanya. Algoritma ini cocok digunakan untuk data dalam jumlah yang sangat besar. [6].

II. METODE PENELITIAN

Penelitian ini dilakukan dengan tahapan sebagai berikut:

A. Studi Literatur

Studi literatur dilakukan dengan melalui penelusuran, penelahaan literatur, yang membahas tentang analisis kelompok, data set dokumen dan bagaimana cara memprosesnya, serta algoritma teknik pengelompokan dokumen.

B. Pengambilan Data

Data yang digunakan untuk eksperimen adalah artikel media masa yang dalam Bahasa Indonesia yang diambil dari website Kompas (www.kompas.com) dan Detik (www.detik.com). Alasan dipilihnya artikel dari Kompas dan Antara adalah karena kemudahan akses dan pemakaian bahasa yang baku dalam setiap artikelnya. Dari website ini, artikel-artikel telah dikelompokkan berdasarkan kategorinya masing-masing. Artikel yang diambil dari Kompas terdiri dari enam kategori yaitu Ekonomi, *Female*, Kesehatan (*Health*), Properti, Sains, dan *Travel*.. Untuk artikel dari Detik, artikel yang digunakan hanyalah artikel dari kategori olahraga. Namun untuk keperluan eksperimen, artikel olahraga, dikelompokkan lagi ke dalam kategori yang lebih spesifik secara manual. Kategori tersebut antara lain : Balap, Basket, Sepakbola, dan Tenis. Untuk setiap kategori, masing-masing diambil sebanyak 30 artikel. Dalam penelitian, kategori ini nantinya diperlukan dalam penentuan kualitas suatu algoritma pengelompokan.

C. Pemrosesan Data

Sebelum dilakukan pengelompokan data, data dokumen terlebih dahulu harus melalui tahap pra proses, yaitu *Detagging*, *Stopword removal*, dan *Stemming*, Tahap *Stopword removal* diperlukan karena banyak kata-kata umum yang frekuensi kemunculannya tinggi tetapi muncul hampir dalam setiap kategori dokumen. Kata-kata inilah yang sering disebut dengan *stopword*. Dengan demikian, kata-kata yang digunakan untuk membentuk *term-document matrix* hanyalah kata-kata khusus yang dapat merepresentasikan dokumen yang ada.

D. Penentuan Fitur

Pada *vector space model* untuk data teks, dokumen direpresentasikan sebagai vektor dalam dimensi m, dengan m adalah jumlah term yang digunakan. Setiap komponen dalam vektor merefleksikan tingkat di mana term yang bersangkutan memberikan arti (*semantic*) pada dokumen tersebut. Koleksi dokumen dapat dinyatakan dengan *term document matrix* yang merupakan kumpulan dari masing-masing vektor dokumen.

Pada penelitian ini digunakan fitur berupa term. Hal ini didasarkan pertimbangan bahwa term dapat merepresentasikan dokumen secara *semantic* sehingga membedakan antara dokumen yang satu dengan yang lain. Dokumen yang mempunyai kemiripan akan dikelompokkan ke dalam kelompok yang sama.

E. Penyusunan term-document matrix

Dalam teknik pengelompokan dokumen, input yang digunakan adalah *term- document matrix*. Matriks ini memberi informasi kemunculan fitur yang digunakan atau term dalam dokumen yang ada. Setiap baris ke-i merepresentasikan term yang ada dalam koleksi. Sedangkan setiap kolom ke-j merepresentasikan dokumen yang digunakan. Setiap sel kemudian ditentukan bobotnya, yaitu f*ij* yang menunjukkan bobot term ke-i pada dokumen j.

F. Pengelompokan Dataset Dokumen

Dari studi literatur, langkah selanjutnya adalah menyusun algoritma-algoritma pengelompokan nonhierarki untuk mengelompokkan dataset dokumen, antara lain K-Mean, Bisecting K-Mean, K-Median, dan K-Medoid. Dari algoritma yang telah disusun, akan diimplementasikan ke dalam program komputer JAVA. Setelah dilakukan uji coba terhadap program yang telah dibuat, selanjutnya program digunakan untuk mengelompokkan dataset dokumen. Untuk setiap algoritma akan digunakan dua ukuran ketakmiripan, yaitu Euclid dan Cosinus.

G. Perbandingan Hasil Pengelompokan

Hasil pengelompokkan dengan menggunakan berbagai algoritma pengelompokan nonhieraki dibandingkan, dan dievaluasi untuk mengetahui kualitas ketepatannya dalam mengelompokkan dokumen. Dalam penelitian ini digunakan *F-measures* dan Akurasi [10].

III. HASIL DAN PEMBAHASAN

A. Dataset

Untuk eksperimen, dalam penelitian ini akan digunakan tiga set data, antara lain :

1. Data Set Pertama : adalah data dengan hanya menggunakan 2 kategori. Data ini diambil dari artikel Kompas, dengan kategori Female dan Properti.
2. Data set Kedua : adalah data dengan menggunakan 6 kategori. Data ini diambil dari artikel Kompas, dengan mengambil keseluruhan kategori, yaitu Ekonomi, *Female*, Kesehatan (*Health*), Properti, Sains, dan *Travel*..
3. Data set Ketiga : adalah data dengan menggunakan 4 kategori. Data ini diambil dari artikel Detik, dengan kelas : Balap, Basket, Sepakbola, dan Tenis.

Pada data artikel Kompas, dilakukan pengelompokan sebanyak dua kali. Yang pertama hanya digunakan 2 kategori, sedangkan yang kedua sebanyak 6 kategori. Dengan bertambahnya kategori yang digunakan, diduga akan semakin menurunkan nilai F-measure dan akurasinya. Di samping itu dari data set pertama diharapkan dapat diperoleh hasil pengelompokan dengan nilai F-measure dan akurasi yang tinggi, karena kedua kategori (Female dan Properti) memang sangat berbeda karakteristiknya. Untuk Data set ketiga digunakan data dari satu kategori yang sama, yaitu Olahraga. Kemudian artikelnya dikelompokkan lagi ke dalam kategori yang lebih spesifik secara manual. Dari data ini ingin dilihat bagaimana kualitas hasil pengelompokkan jika data diperoleh dari kategori yang sama,

B. Implementasi Penentuan Fitur

Penentuan fitur dilakukan untuk semua dokumen dari semua kategori yang digunakan untuk setiap set data. Dalam pengelompokan dokumen, tidak semua fitur digunakan sebagai term. Penentuan fitur yang akan digunakan sebagai term dilakukan dengan mengidentifikasi setiap kata yang muncul dalam setiap dokumen. Setiap kata yang muncul, dihitung frekuensi kemunculannya pada setiap dokumen. Fitur ini selanjutnya diurutkan berdasarkan frekuensi kemunculannya. Fitur yang akan digunakan dalam pembuatan *term-document matrix* dalam eksperimen berupa persentase dari semua fitur yang teridentifikasi.

Jumlah fitur yang digunakan untuk pengelompokan dinyatakan dengan persentase dari total fitur. Untuk keperluan pengelompokan di antara fitur yang ada pada keseluruhan dokumen, banyaknya fitur yang digunakan untuk pengelompokan dokumen dalam eksperimen digunakan sebanyak 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% dan 100%. Semakin besar persentase fitur berarti semakin banyak informasi yang digunakan untuk mengelompokkan dokumen. Dengan semakin banyaknya pengetahuan untuk membedakan satu dokumen dengan dokumen yang lain, sehingga pengelompokan dokumen diharapkan dapat dilakukan dengan lebih akurat.

C. Implementasi Pembuatan Term-Document Matrix

Setelah melakukan penentuan fitur, langkah selanjutnya adalam membuat *term-document matrix*. Matriks ini merupakan input yang akan digunakan dalam pengelompokan dokumen. *Term-document matrix* merupakan representasi dari setiap dokumen yang digunakan dalam pengelompokan. Untuk setiap variasi eksperimen yang dilakukan dibentuk satu *term-document matrix*. Matriks ini merupakan matriks dua dimensi, yang terdiri dari dimensi term dan dimensi dokumen. Sedangkan data yang diisikan dalam matriks itu adalah frekuensi atau kemunculan sebuah term dalam suatu dokumen. Untuk keperluan evaluasi, label kategori dari tiap dokumen. Disimpan. Informasi ini nantinya akan digunakan untuk pencocokan hasil pengelompokan.

D. Implementasi Pengelompokan dokumen.

Setiap set data dokumen, akan dilakukan pengelompokan dengan menggunakan algoritma K-Mean, Bisecting K-Mean, K-Median, dan K-Medoid. Untuk setiap algoritma digunakan dua jarak yaitu Euclid dan Cosinus. Setiap algoritma pengelompokan diulang sebanyak 10 kali.

Hasil pengelompokan dengan menggunakan data set yang pertama, rata-rata nilai F-measure dan akurasinya bisa dilihat pada tabel I dan II. Sedangkan secara grafik, ditunjukkan pada gambar I dan II. Dari hasil ini terlihat bahwa algoritma yang menghasilkan pengelompokan terbaik adalah algoritma Bisecting K-Mean dan K-Mean, dengan menggunakan jarak cosinus. Hal ini ditunjukkan dengan tingginya baik nilai F-Measure maupun akurasi. Algoritma yang terbaik selanjutnya ditunjukkan oleh Bisecting K-Mean dengan jarak Euclid. Sedangkan hasil pengelompokan dengan nilai F-Measure dan akurasi terendah dihasilkan oleh algoritma pengelompokan K-Medoid, baik dengan menggunakan jarak euclid maupun cosinus. Untuk hasil pengelompokan dengan jarak euclid, baik K-Mean maupun K-Median juga menunjukkan hasil yang kurang memuaskan. Jika dilihat dari persentase fitur yang akan digunakan dalam pengelompokkan, dengan hanya 10% – 20% dari keselurahan fitur, sudah bisa mengelompokkan data dokumen dengan baik.

Untuk data set yang kedua, statistika deskriptif tentang nilai F-Measure dan akurasi hasil pengelompokan bisa dilihat secara lengkap pada tabel III dan IV. Sedangkan grafik yang terkait rata-rata nilai F-Measure dan Akurasinya bisa dilihat pada gambar III dan IV.

Jika dilihat dari rata-rata nilai F-Measure, algoritma K-Mean dengan menggunakan jarak cosinus menunjukkan hasil yang paling memuaskan. dibandingkan dengan algoritma yang lain. Terbaik berikutnya adalah algoritma K-Mean dengan jarak euclid, dan kemudian diikuti oleh algoritma Bisecting K-Mean dengan jarak cosinus. Jika dilihat dari rata-rata Akurasinya, juga menunjukkan hasil yang serupa. Pebedaan dari nilai keduanya adalah jika pada rata-rata nilai F-Measure semua metode menunjukkan nilai yang hampir stabil, sedangkan nilai akurasi menunjukkan nilai yang berfluktuasi khususnya untuk algoritma K-Medain dan K-Medoid. Untuk

algoritma K-Medoid dengan jarak Euclid, kurang bisa memberikan hasil pengelompokan. Dari *running* program diperoleh hasil yang tidak konvergen. Jika diteliti lebih lanjut, diperoleh hasil pengelompokan dengan kelompok yang tidak ada anggotanya.

Untuk data set ketiga, statistik deskriptif hasil pengelompokan dengan menggunakan semua metode dapat dilihat pada tabel V dan VI. Sedangkan secara grafik tentang rata-rata nilai F-Measure dan Akurasinya ditunjukkan pada gambar V dan VI. Dari rata-rata nilai F-Measure dan Akurasi, algoritma yang memberikan hasil yang paling memuaskan berturut-turut K-Mean dengan jarak cosinus, Biscecting K-Mean dengan jarak cosinus dan K-Median dengan jarak cosinus. Sedangkan metode K-Medoid dengan jarak Euclid memberikan hasil yang paling tidak memuaskan.

Berdasarkan hasil eksperimen, secara umum memberikan hasil pengelompokan yang paling memuaskan adalah K-Mean. Hal ini dievaluasi berdasarkan nilai *F-measures* dan Akurasi. Selain itu jarak cosinus juga memberikan hasil yang paling baik. Hal ini disebabkan karena data yang digunakan sebagai dasar pengelompokan adalah data *counting* atau data hasil penjumlahan atau data frekuensi.

Dari hasil ekperimen, persentase fitur yang semakin besar secara umum tidak mengakibatkan peningkatan kualitas hasil pengelompokan. Berdasarkan hasil eksperimen, dengan menggunakan 10% sampai 30% fitur yang digunakan, menghasilkan pengelompokan yang memuaskan. Hal ini disebabkan karena fitur yang hanya sedikit muncul dalam dokumen kurang memiliki arti yang penting dalam pengelompokan data.

KESIMPULAN

Dari hasil perbandingan algoritma pengelompokan K-Mean, Bisecting K-Mean, K-Median, dan K-Medoid untuk data set dokumen, bisa diambil kesimpulan:

1. Algoritma pengelompokan yang memberikan hasil pengelompokan yang terbaik adalah K-Mean. Disamping itu Bi-secting K-Mean juga menghasilkan pengelompokan yang memuaskan. Sedangkan hasil pengelompokan K-Medoid memberikan hasil yang tidak memuaskan.
2. Jarak cosinus memberikan hasil yang lebih baik dibandingkan dengan jarak Euclid. Hal ini disebabkan karena data yang digunakan sebagai dasar pengelompokan adalah data *counting* atau data hasil penjumlahan atau data frekuensi.
3. Persentase fitur yang semakin besar secara umum tidak mengakibatkan peningkatan kualitas hasil pengelompokan. Berdasarkan hasil eksperimen, dengan menggunakan 10% sampai 30% fitur yang digunakan, menghasilkan pengelompokan yang memuaskan. Hal ini disebabkan karena fitur yang hanya sedikit muncul dalam dokumen kurang memiliki arti yang penting dalam pengelompokan data.

REFERENCES

[1] Hamzah, A. (2012). Klasifikasi Teks dengan Naive Bayes Clasifier (NBC) untuk Pengelompokan teks Berita dan Abstrak Akademik, Prosiding Seminar Nasional Aplikasi Sains & Teknologi III, Yogyakarta.

[2] Bridge, C., 2011, *Unstructured Data and the 80 Percent Rule*. http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551 ;

[3] Sharma, S., 1996. *Applied Multivariate Techniques*. John Wiley & Sons Inc, New York.

[4] Halkidi, M., Y. Batistakis, M. Vazirgiannis, 2001. *On Clustering Validation Techniques. Journal of Intelligent Information Systems.* http://citeseer.nj.nec.com/513619.html. Tanggal akses 31/3/03.

[5] Gomez-Ballester, E., Luisa Mico dan Jose Oncina (2002). A Fast Approximated k-Median Algorithm. http://citeseer.nj.nec.com/547596.html.

[6] *Perbandingan Algoritma Pengelompokan Non-Hierarki dengan Indeks Validitas Kelompok*, MATEMATIKA, Jurnal Matematika atau Pembelajarannya, Tahun XI, Nomor 1, April 2005.

[7] Luo, C., Yanjun Lie, Soon M. Chung (2009), Text Document Clustering Based on Neighbors, Journal Elsivier : Data & Knowledge Enginering.

[8] Kusuma, JH., Kiki Maulana, Warih Maharani (2011). Analisis Active Fuzzy Constrained Clustering dengan menggunakan Vektor Model untuk Pengelompokan Dokumen, Prosiding Konferensi Nasional Sistem & Informatika, Bali.

[9] Chu, S., John F. Roddick, J.S. Pan (2002). *Efficient K-Medoids Algorthms Using Multi-Centroids With Multi-Runs Sampling Scheme*. http://citeseer.nj.nec.com/537999..html.

[10] Steinbach,M, Geroge Karypis, Vipin Kumar (2000), *A Comparison of Document Clustering Techniques*, technical report Departemen of Computer Science $ Engineering University of Minesota.

TABEL I. RATA-RATA NILAI F-MEASURE UNTUK DATASET PERTAMA

	Metode	Jarak	Persentase Fitur									
			10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
F-Measure	K-Mean	Cosinus	0,947	0,971	0,983	0,954	0,955	0,973	0,947	0,964	0,977	0,961
		Euclid	0,634	0,634	0,644	0,635	0,663	0,642	0,638	0,638	0,646	0,656
	Bisecting K-Mean	Cosinus	0,984	0,983	0,960	0,974	0,983	0,983	0,962	0,986	0,983	0,978
		Euclid	0,792	0,819	0,813	0,828	0,859	0,867	0,862	0,859	0,847	0,892
	K-Median	Cosinus	0,689	0,710	0,682	0,701	0,708	0,697	0,695	0,702	0,701	0,704
		Euclid	0,665	0,636	0,684	0,692	0,643	0,646	0,662	0,668	0,642	0,660
	K-Medoid	Cosinus	0,622	0,655	0,671	0,669	0,608	0,623	0,713	0,683	0,646	0,632
		Euclid	0,635	0,646	0,639	0,640	0,638	0,636	0,635	0,634	0,640	0,641

TABEL II. RATA-RATA AKURASI UNTUK DATASET PERTAMA

	Metode	Jarak	Persentase Fitur									
			10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Akurasi	K-Mean	Cosinus	0,984	0,971	0,984	0,955	0,955	0,974	0,947	0,965	0,977	0,962
		Euclid	0,634	0,501	0,521	0,508	0,570	0,522	0,518	0,499	0,540	0,563
	Bisecting K-Mean	Cosinus	0,984	0,984	0,961	0,974	0,984	0,984	0,962	0,987	0,984	0,978
		Euclid	0,789	0,818	0,811	0,827	0,859	0,868	0,863	0,860	0,846	0,893
	K-Median	Cosinus	0,636	0,672	0,628	0,632	0,670	0,627	0,625	0,665	0,645	0,650
		Euclid	0,568	0,523	0,572	0,594	0,503	0,507	0,543	0,568	0,506	0,537
	K-Medoid	Cosinus	0,607	0,642	0,648	0,640	0,600	0,612	0,700	0,612	0,559	0,619
		Euclid	0,504	0,498	0,497	0,525	0,499	0,501	0,502	0,512	0,498	0,496

TABEL III. RATA-RATA NILAI F-MEASURE UNTUK DATASET KEDUA

	Metode	Jarak	Persentase Fitur									
			10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
F-Measure	K-Mean	Cosinus	0,679	0,651	0,697	0,74	0,718	0,732	0,703	0,779	0,688	0,728
		Euclid	0,405	0,457	0,446	0,444	0,448	0,446	0,472	0,445	0,424	0,468
	Bisecting K-Mean	Cosinus	0,539	0,358	0,337	0,356	0,34	0,354	0,354	0,365	0,363	0,352
		Euclid	0,317	0,316	0,328	0,32	0,328	0,321	0,321	0,303	0,327	0,333
	K-Median	Cosinus	0,285	0,305	0,277	0,283	0,276	0,276	0,281	0,291	0,294	0,287
		Euclid	0,275	0,278	0,282	0,267	0,274	0,274	0,286	0,269	0,274	0,276
	K-Medoid	Cosinus	0,282	0,29	0,279	0,29	0,279	0,283	0,279	0,285	0,283	0,288
		Euclid	0,273	0,277	0,277	0,277	0,277	0,277	0,277	0,276	0,276	0,276

TABEL IV. RATA-RATA AKURASI UNTUK DATASET KEDUA

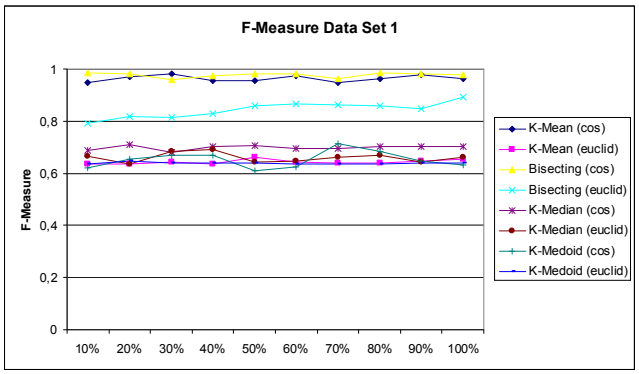
	Metode	Jarak	Persentase Fitur									
			10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Akurasi	K-Mean	Cosinus	0,893	0,884	0,9	0,915	0,907	0,912	0,902	0,928	0,896	0,911
		Euclid	0,718	0,754	0,747	0,746	0,758	0,765	0,77	0,747	0,726	0,772
	Bisecting K-Mean	Cosinus	0,822	0,712	0,702	0,731	0,698	0,74	0,729	0,725	0,739	0,711
		Euclid	0,570	0,532	0,605	0,545	0,563	0,545	0,544	0,519	0,575	0,551
	K-Median	Cosinus	0,748	0,755	0,615	0,691	0,509	0,613	0,447	0,501	0,457	0,450
		Euclid	0,583	0,446	0,488	0,402	0,503	0,395	0,454	0,445	0,457	0,396
	K-Medoid	Cosinus	0,222	0,28	0,162	0,393	0,162	0,281	0,162	0,334	0,218	0,278
		Euclid	0,289	0,199	0,202	0,199	0,191	0,193	0,194	0,222	0,216	0,224

TABEL V. RATA-RATA NILAI F-MEASURE UNTUK DATASET KETIGA

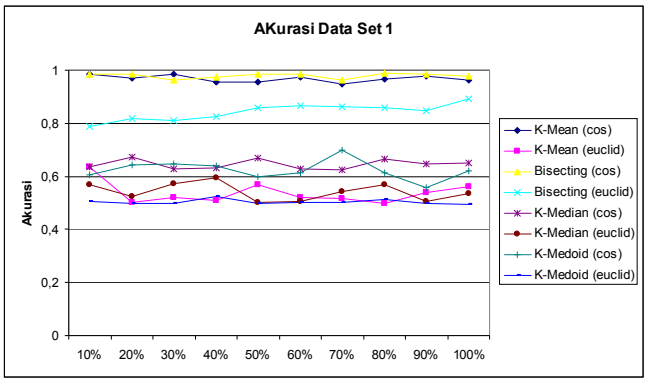
	Metode	Jarak	Persentase Fitur									
			10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
F-Measure	K-Mean	Cosinus	0,943	0,729	0,655	0,638	0,636	0,581	0,660	0,605	0,611	0,632
		Euclid	0,848	0,472	0,467	0,457	0,475	0,433	0,439	0,476	0,440	0,440
	Bisecting K-Mean	Cosinus	0,876	0,589	0,612	0,585	0,655	0,615	0,632	0,591	0,572	0,574
		Euclid	0,816	0,543	0,527	0,555	0,552	0,525	0,539	0,545	0,518	0,536
	K-Median	Cosinus	0,780	0,560	0,591	0,576	0,564	0,575	0,511	0,594	0,542	0,609
		Euclid	0,699	0,450	0,416	0,427	0,427	0,446	0,445	0,429	0,430	0,428
	K-Medoid	Cosinus	0,438	0,531	0,474	0,525	0,459	0,486	0,458	0,476	0,530	0,531
		Euclid	0,401	0,409	0,409	0,411	0,398	0,402	0,405	0,404	0,404	0,409

TABEL VI. RATA-RATA AKURASI UNTUK DATASET KETIGA

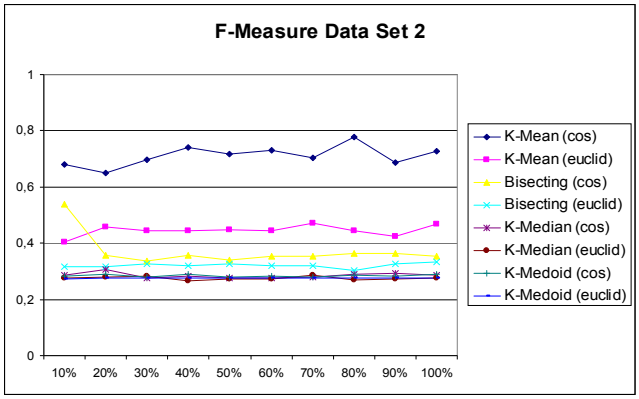
	Metode	Jarak	Persentase Fitur									
			10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Akurasi	K-Mean	Cosinus	0,885	0,855	0,812	0,812	0,810	0,780	0,824	0,786	0,795	0,806
		Euclid	0,723	0,555	0,562	0,545	0,558	0,501	0,490	0,575	0,502	0,494
	Bisecting K-Mean	Cosinus	0,754	0,777	0,784	0,764	0,814	0,785	0,796	0,775	0,756	0,776
		Euclid	0,677	0,696	0,648	0,693	0,700	0,694	0,681	0,669	0,649	0,676
	K-Median	Cosinus	0,571	0,776	0,797	0,744	0,779	0,782	0,657	0,751	0,758	0,807
		Euclid	0,488	0,623	0,542	0,507	0,622	0,556	0,580	0,553	0,546	0,499
	K-Medoid	Cosinus	0,437	0,663	0,631	0,671	0,491	0,551	0,531	0,624	0,669	0,701
		Euclid	0,390	0,454	0,452	0,465	0,419	0,439	0,443	0,447	0,455	0,443



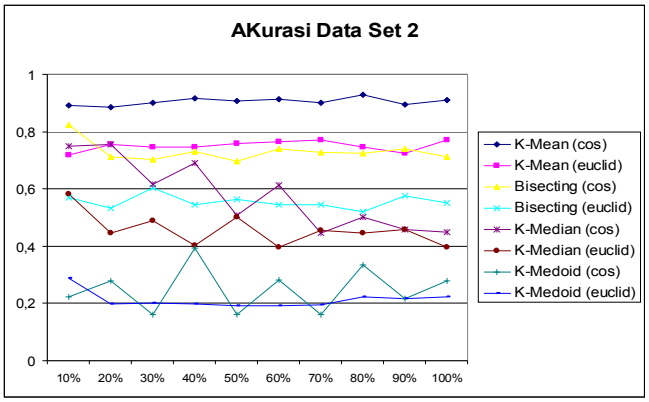
Gambar 1. Rata-rata F-Measure untuk Data Set Pertama



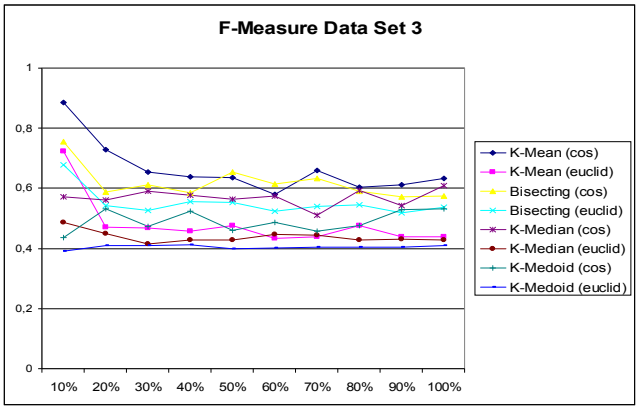
Gambar 2. Rata-rata Akurasi untuk Data Set Pertama



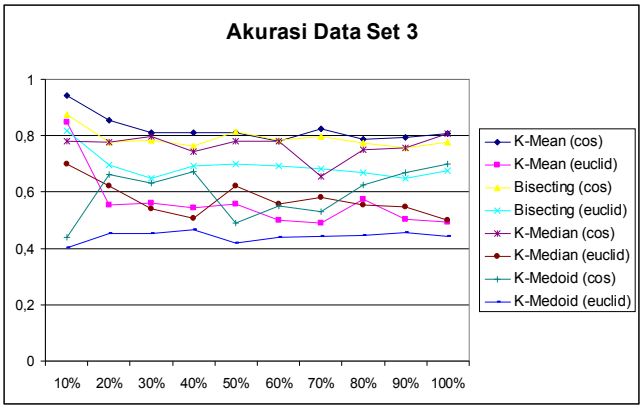
Gambar 3. Rata-rata F-Measure untuk Data Set Kedua



Gambar 4. Rata-rata Akurasi untuk Data Set Kedua



Gambar 5. Rata-rata F-Measure untuk Data Set Ketiga



Gambar 6. Rata-rata Akurasi untuk Data Set Ketiga