ANALISIS DAN PERBANDINGAN KUALITAS PENGELOMPOKAN DOKUMEN (*DOCUMENT CLUSTERING*) DENGAN MENGGUNAKAN METODE K-MEANS DAN K-MEDIANS

BUSTAMI YUSUF
Fakultas Sains dan Teknologi, Universitas Islam Negeri Ar-Raniry
Banda Aceh, Indonesia
bustamiyusoef@ar-raniry.ac.id

Abstract: *Conducting data analysis on a large set of documents is not an easy task. The common stages are document filtering, document selection, and document clustering. Clustering is a technique used in data mining to find groups of data that do not have a natural grouping. There are many clustering algorithm have been introduced, and two of them are K-means and K-medians. Both methods classify documents based on the proximity of words weighting between documents. This study aims to compare the quality of the clusters produced by K-means and K-medians. The results show that K-medians obtain a better cluster quality when compared to K-means. However, it takes more time to cluster.*

Key Word: *Keywords: Data Mining, Clustering, K-means, and K-medians*

1. Pendahuluan

Proses analisa terhadap dokumen dalam jumlah besar bukanlah suatu pekerjaan yang mudah. Tahapan yang biasa dilakukan adalah dengan melakukan penyaringan data (*data filtering*), pemilihan data (*data selecting*), dan pengelompokan data (*data clustering*). Ketiga proses ini bertujuan untuk membuat data menjadi relevan dan lebih mudah untuk dianalisa. *Clustering* merupakan salah satu bentuk dari pengelompokan data.

Clustering adalah teknik dalam *data mining* yang digunakan untuk mencari kelompok dari data yang tidak memiliki kelompok secara alami. Algoritma *clustering* akan mengelompokkan data dalam beberapa *cluster* berdasarkan kemiripan antara satu data dengan data yang lainnya. Dalam hal ini, tidak ada patokan tertentu yang digunakan oleh pada algoritma *clustering* untuk mencari kelompok data. Data yang dikelompokkan dalam *cluster* yang sama diharapkan memiliki similaritas yang tinggi, sebaliknya, data dalam *cluster* yang berbeda diharapkan memiliki kesamaan yang rendah.

Banyak algoritma *clustering* yang sudah diperkenalkan, dua di antaranya adalah K-means dan K-medians. Kedua metode ini bisa mengatasi masalah pengelompokkan dokumen pada sejumlah dokumen teks berdasarkan kedekatan

bobot kata antar dokumen dan juga dapat mempermudah untuk melakukan analisis pada kumpulan dokumen yang berjumlah besar.

Dengan adanya tujuan yang sama, maka perbandingan terhadap kedua metode ini bisa dilakukan dalam mencari metode yang lebih sesuai untuk menangani masalah dalam menganalisa data yang berjumlah besar. Penilitian ini adalah lanjutan dari penilitian sebelumnya [3], dimana hasil yang perbandingkan masih meliputi kategori waktu, jumlah iterasi, dan bobot keakuratan dalam mengklasterisasi dokumen teks. Hasil perbandingan yang diperoleh diharapkan dapat membantu untuk menentukan metode yang tepat dan sesuai dengan kebutuhan untuk mengatasi masalah dalam mengelompokkan dokumen.

2. DASAR TEORI

a. *Latent Semantic Indexing*

Latent Semantic Indexing (LSI) adalah metode pengindeksian dan pencarian yang menggunakan teknik matematika yang disebut *Singular Value Decomposition (SVD)* untuk mengidentifikasi pola-pola hubungan antara istilah dan konsep-konsep yang terkandung dalam koleksi teks yang tidak terstruktur. Pada dasarnya, cara kerja *LSI* memegang prinsip bahwa kumpulan kata yang berada dalam satu konteks yang sama akan memiliki arti yang sama juga. *LSI* memiliki kemampuan untuk mengesktrak kerangka sebuah dokumen teks dengan menciptakan hubungan antar kelompok-kelompok istilah yang sama yang muncul di dalam dokumen[2].

Selain itu, *LSI* juga bisa digunakan untuk melakukan pengelompokan dokumen secara otomatis. Bahkan, dalam beberapa penelitian sebelumnya telah ditemukan kesamaan antara proses LSI dan proses pengelompokan dokumen teks[7]. Pengelompokan dokumen teks adalah cara untuk memisahkan dokumen kedalam beberapa kelompok yang didasari oleh kesamaan unsur yang dimikili oleh setiap dokumen[5]. Dalam proses pengelompokan, *LSI* membandingkan unsur yang dimilki oleh dokumen dengan unsur yang dimiliki oleh dokumen pembanding.

Teknik yang digunakan oleh *LSI* didalam mempelajari konsep hubungan antar dokumen teks tidak lain adalah teknik aljabar linier umum. Secara sederhana, prosesnya dimulai dengan membangun sebuah matrik bernilai dari dokumen teks yang diberikan dan menerapkan teknik *Singular Value Decomposition* (*SVD)* padanya, sehingga dengan menggunakan matriks tersebut bisa menyelidiki unsur-unsur yang dimiliki oleh dokumen teks.

b. *Singular Value Decomposition (SVD)*

Singular Value Decomposition (SVD) adalah salah satu teknik untuk mengolah matriks dari cabang ilmu aljabar linear yang diperkenalkan oleh Beltrami pada tahun 1873. *SVD* merupakan salah satu tahapan proses yang ada dalam metode *Latent Semantic Analysis (LSA*) yaitu teknik untuk mengolah

matriks dalam cabang ilmu aljabar linear sebagai salah satu alat matematis yang digunakan untuk merepresentasikan sebuah matriks dan mampu melakukan berbagai analisis dan komputasi matriks. *SVD* berguna untuk mendekomposisi suatu matriks menjadi 3 bagian matriks baru, yaitu matriks ortogonal *U*, matriks diagonal *S* dan *transpose* dari matriks ortogonal *D* atau dapat dirumuskan sebagai berikut:

$$\mathbf{A_{m\,x\,n} = U_{m\,x\,n}\, S_{n\,x\,n}\, V^{t}_{n\,x\,n}}$$

dimana,

U : matriks ortogonal berukuran m x n.

S : matriks diagonal berukuran n x n, dengan elemen matriks ≥ nol.

V^t: matriks ortogonal berukuran n x n [10]

c. ***Clustering***

Menurut Karel [6], analisis *cluster* adalah upaya menemukan sekelompok objek yang mewakili suatu karakter yang sama atau hampir sama (*similar*) antar satu objek dengan objek lainnya pada suatu kelompok dan memiliki perbedaan (*not similar*) dengan objek-objek pada kelompok lainnya. Tentunya persamaan dan perbedaan tersebut diperoleh berdasarkan informasi yang diberikan oleh objek-objek tersebut beserta hubungan (*relationship*) antar mereka. Dalam berbagai kesempatan, *clustering* juga sering disebut *Unsupervised Learning* yaitu, pengelompokan data yang tidak memiliki pengelompokan alami.

Dewasa ini sudah banyak metode-metode *clustering* dikembangkan dalam mengatasi permasalahan pengelompokan dokumen teks. Namun, dalam penelitian kami ini hanya fokus pada dua metode *clustering* saja, yaitu *K-Means* dan *K-Medians.*

1. K-Means

K-means adalah suatu metode analisa data atau metode *data mining* yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode K-means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu *cluster* dan memaksimalkan variasi dengan data yang ada di-*cluster* lainnya [1][4].

K-means merupakan metode klasterisasi yang paling terkenal dan banyak digunakan di berbagai bidang karena sederhana, mudah diimplementasikan, memiliki kemampuan untuk mengklaster data yang besar, dan kompleksitas waktunya *linear* O (nKT) dengan n adalah jumlah dokumen, K adalah jumlah *cluster*, dan T adalah jumlah iterasi. K-means merupakan metode pengklasteran

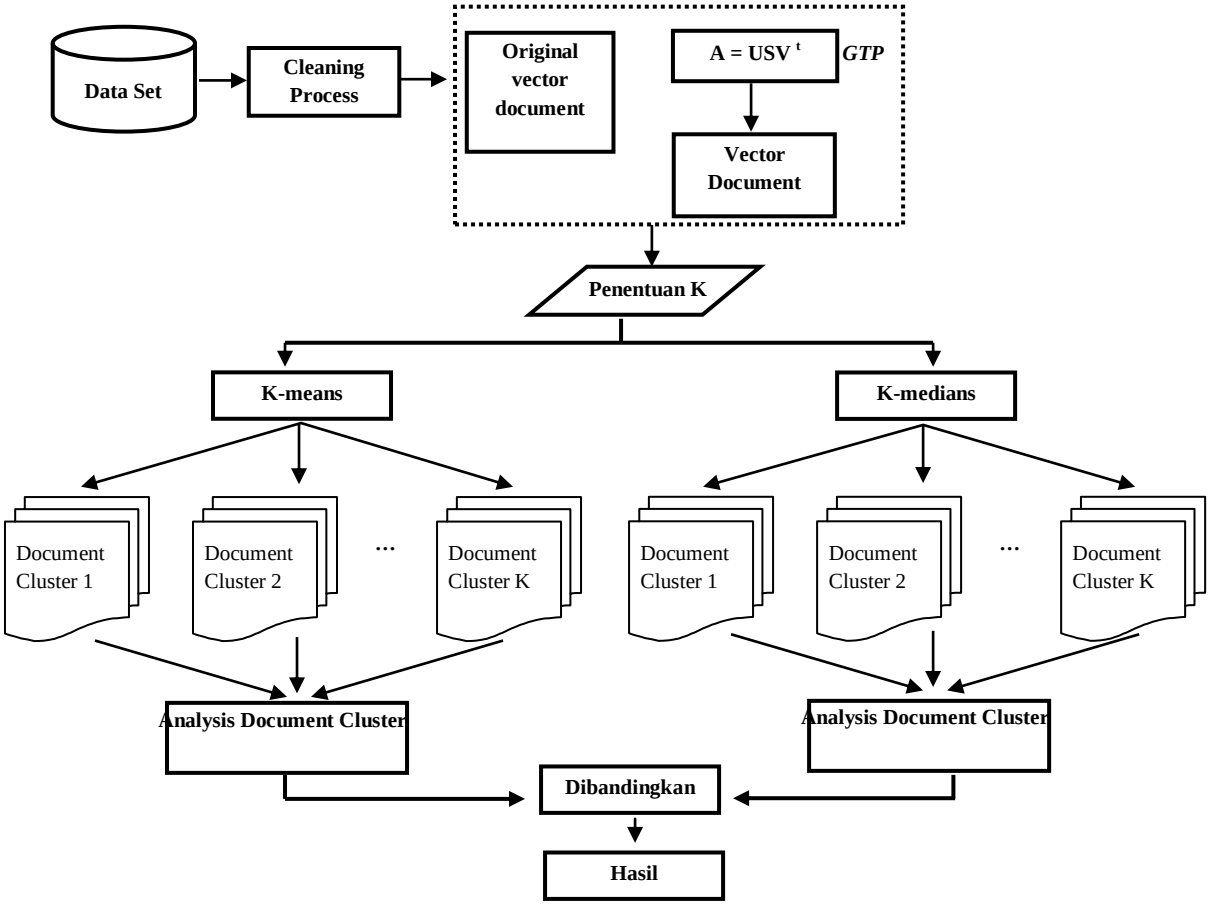
secara *partitioning* yang memisahkan data ke dalam kelompok yang berbeda. Dengan *partitioning* secara iteratif, K-means mampu meminimalkan rata-rata jarak setiap data ke *cluster*-nya. Metode ini dikembangkan oleh MacQueen pada tahun 1967 [9].

2. K-Medians

Algoritma K-medians adalah algoritma *clustering* yang terkait dengan algoritma K-means dan termasuk partisi *clustering* yang berusaha untuk mencari jarak antara titik berlabel yang berada dalam *cluster* dan menetapkannya sebagai sebuah titik pusat *cluster*[5].

3. METODOLOGI PENELITIAN

Prosedur penelitian yang dilakukan adalah seperti yang terlihat pada skema penelitian dalam gambar 1.



Gambar 1. Prosedur penelitian

a. *Dataset*

Ada 2 (dua) jenis *dataset* yang digunakan dalam penelitian ini yaitu IRIS dan Reuters-21578 *dataset*.

1. IRIS *dataset*

IRIS *dataset* atau data IRIS merupakan data multivariat yang diperkenalkan oleh Sir Ronald Aylmer Fisher (1936) sebagai contoh analisis diskriminan. Data ini juga disebut Anderson IRIS *dataset* karena Edgar Anderson mengumpulkan data ini untuk mengukur variasi geografis bunga IRIS di Semenanjung Gaspé.

Data IRIS berasal dari 3 (tiga) jenis bunga IRIS yaitu *setosa*, *versicolor*, dan *virginica*. Masing-masing jenis bunga terdiri dari 50 sampel. Ada empat fitur yang diukur dari masing-masing sampel tersebut yaitu, *sepal length*, *sepal width*, *petal length*, dan *petal width*.

2. Reuters-21578 *dataset*

Reuters-21578 (V 1.3) *dataset* adalah data dokumen teks yang sudah dikategorikan dan dipublikasikan oleh David D. Lewis pada tanggal 14 Mei 2004 (*http://www.research.att.com/~lewis*). Reuters-21578 *dataset* memiliki enam kategori asli yaitu *exchanges*, *orgs*, *people*, *places*, *topics*, dan *companies*, tetapi karena jumlah artikel untuk kategori *companies* hanya berjumlah 6 dan tidak memiliki isi maka kategori ini tidak disertakan dalam penelitian ini. Dokumen-dokumen dalam data Reuters disusun dalam format seperti yang terlihat pada Gambar. 2.

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5544" NEWID="1">
<DATE>26-FEB-1987 15:01:01.79</DATE>
<TOPICS><D>cocoa</D></TOPICS>
<PLACES><D>el-salvador</D><D>usa</D><D>uruguay</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;C T
&#22;&#22;&#1;f0704&#31;reute
u f BC-BAHIA-COCOA-REVIEW   02-26 0105</UNKNOWN>
<TEXT>&#2;
<TITLE>BAHIA COCOA REVIEW</TITLE>
<DATELINE>    SALVADOR, Feb 26 - </DATELINE><BODY>Showers continued
throughout the week in
the Bahia cocoa zone, alleviating the drought since early
January and improving prospects for the coming temporao,
although normal humidity levels have not been restored,
Comissaria Smith said in its weekly review…
    Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Gambar 2. Contoh format dokumen data Reuters

b. *Cleaning*

Sebelum proses pembersihan dilakukan, semua data Reuters dalam format *SGML* yang berjumlah 22 *file* digabungkan terlebih dahulu menjadi sebuah *file* besar dalam format txt. Dalam proses *cleaning*, *file* besar tersebut dipecahkan

menjadi beberapa *file* kecil sejumlah dokumen yang terdapat dalam data Reuters. Bersamaan dengan proses itu, tag-tag *HTML* seperti <title>, </title>, <date>, </date>, <dateline>, </dateline>, <unknown>, </unknow>, <body>, dan </body> dibuang. Selain itu, karakter dan simbol-simbol yang tidak bermakna, seperti &#5;, &#5;C, T&#22;, &#1;, f0704&#31;reute, dan &#2 juga dihapus.

Proses penghapusan tag-tag *HTML* dan simbol yang tidak bermakna juga disertakan dengan proses penghapusan dokumen dan penentuan kategori dokumen. Dokumen yang dihapus adalah dokumen-dokumen yang memiliki kata kurang dari 100, sedangkan penentuan kategori dokumen dilakukan dengan cara menghitung jumlah sub kategori terbanyak yang terdapat pada setiap kategori, misalnya pada gambar 2, kategori *places* memiliki lebih banyak sub kategori bila dibandingkan dengan kategori yang lainya, sehingga dokumen tersebut dianggap berkategori *places*. Gambar 3 menunjukkan dokumen data router yang sudah dibersihkan dan dikategorikan.

```
<PLACES>
Showers continued throughout the week in the Bahia cocoa zone,
alleviating the drought since early January and improving
prospects for the coming temporao, although normal humidity
levels have not been restored, Comissaria Smith said in its
weekly review. The dry period means the temporao will be late
this year.
...
```

Gambar 3. Contoh format dokumen data Reuters yang sudah bersih

c. *General Text Parser* (*GTP*)

Program *General Text Parser* (*GTP*) adalah software text mining yang digunakan sebagai tool untuk membangun matriks *term dokuments* yang menggunakan *SVD*. *GTP* adalah suatu paket program dalam bahasa C yang dibuat oleh Michael W. Berry dari *University of Tennessee, USA*, yang mengimplementasikan *SVD* di dalamnya. *GTP* menghasilkan data vektor yang dapat digunakan untuk menyelesaikan masalah pencarian informasi (*information retrival*).

GTP mengizinkan pengguna untuk menentukan ukuran dimensi dari matriks singular dalam penerapan *SVD*. Ada 3 dimensi yang dipilih dalam penelitian ini yaitu, 8738 x 200 (D200), 8738 x 400 (D400), dan 8738 x 600 (D600).

b. Algoritma K-Means dan K-Medians

Metode *k*-means *clustering* mengelompokkan data ke dalam *k* buah *cluster*, dimana tiap-tiap *cluster* memiliki sebuah titik *centroid* yang dipilih secara random. Dalam setiap *cluster* akan dicari sebuah *centroid* baru dengan mencari nilai rata-rata dari semua anggota masing-masing *cluster*. Proses penentuan *centroid* dan pengelompokan data dalam *cluster* diulangi sampai semua titik *centroid* tidak lagi

berubah. Penelitian ini menggunakan *prototype k-mean* penelitian sebelumnya[11]. Gambar 4 menunjukkan cara kerja algoritma cara kerja K-means.

```
Input: vektor dokumen D, k
Output: k cluster dokumen
1. Pilih secara acak k vektor sebagai centroid
2.    tempatkan data (vektor) dalam cluster atau centroid
      terdekat
3.    hitung centroid baru dari cluster yang terbentuk
4. sampai centroid tidak berubah lagi
5. catat nilai jarak antara centroid baru dan centroid lama (λ).
6. ulangi dari langkah 3 apabila perulangan < 100 kali
7. akhiri apabila nilai jarak baru (λB) >  nilai jarak lama (λL).
```

Gambar 4. Algoritma K-means.

Algoritma K-medians juga mengelompokkan data ke dalam *k cluster.* Data yang ditempatkan ke dalam suatu *cluster* memiliki jarak yang yang paling dekat ke *cluster* tersebut dari pada ke *cluster* yang lain. Perbedaan K-means dengan K-medians terletak pada penentuan titik *centroid* baru, dimana *centroid* baru dalam K-medians ditentukan dengan mencari titik tengah dari data yang sudah diurutkan dari setiap *cluster*. Proses penentuan titik *centroid* dan penempatan data ke dalam *cluster* terdekat akan diakhiri apabila semua data dari setiap *cluster* tidak lagi berpindah. Berikut ini adalah cara kerja algoritma K-medians.

```
Input: vektor dokumen D, k
Output: k cluster dokumen
1. Pilih secara acak k vektor sebagai centroid
2.    tempatkan data (vektor) dalam cluster atau terdekat
3.    hitung centroid baru dari cluster yang terbentuk
4. akhiri apabila setiap centroid baru = setiap centroid lama
5. catat nilai jarak antara centroid baru dan centroid lama (λ).
6. ulangi dari langkah 3 apabila perulangan < 100 kali
7. akhiri apabila nilai jarak baru (λB) >  nilai jarak lama
```

Gambar 5. Algoritma K-medians.

e. Evaluasi Hasil Pengelompokkan

Evaluasi Hasil Pengelompokkan sangat berguna untuk mengukur kualitas dari *cluster* yang dihasilkan oleh sebuah algoritma *clustering.* Beberapa metode yang biasa digunakan dalam mengukur kualitas *clustering*, di antaranya metode *entropy*, *purity*, *recall* dan *F-score*.

Penelitian ini akan digunakan 2 (dua) metode dalam menguji kualitas *clustering*, yaitu metode *entropy* dan *purity*. Untuk setiap *cluster*, *entropy* dapat diukur sebagai berikut .

$$Entropy\ (D_i) = -\sum_{j=1}^{k} \Pr_i(c_j) \log_2 \Pr_i(c_j),$$

Dimana, $\Pr_i(c_j)$ adalah rasio data dengan kategori c_j dalam cluster D_i. Total *entropy* dari seluruh cluster dihitung dengan :

$$Entropy_{total}(D_i) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \text{ x } entropy(D_i),$$

dengan D_i adalah jumlah data *cluster i* dan *D* adalah total jumlah data.

Nilai *purity* dan total *purity* setiap *cluster* dapat diukur sebagai berikut.

$$Purity\ (D_i) = \max(\Pr_i(c_j)) \text{ dan}$$

$$Purity_{total}(D_i) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \text{ x } purity(D_i),$$

dimana, $max(Pr_i(c_j))$ adalah rasio kategori tertinggi dalam *cluster* D_i [8].

4. Hasil Penelitian

a. Pengujian dengan Data IRIS

Data yang pertama diuji dalam penelitian ini adalah data IRIS. Pengujian tehadap data ini dilakukan sebanyak 6 (enam) kali percobaan. Seperti yang telah dijelaskan pada bab sebelumnya, ada 3 (tiga) jenis data IRIS yang digunakan yaitu data IRIS All (data IRIS asli) dan 2 jenis data IRIS yang diperkecil dimensinya dengan metode *SVD* (IRIS 3D dan IRIS 2D).

Tabel 1 menunjukkan perbandingan kualitas *cluster* yang didapatkan dengan data IRIS untuk kedua metode *clustering* yang digunakan. Pada baris pertama memperlihatkan kualitas *cluster* untuk data IRIS All, sedangkan baris kedua dan ketiga adalah kualitas *cluster* untuk data IRIS 3D dan 2D

Tabel 1 Hasil rata-rata perbandingan kualitas *cluster* dengan data IRIS untuk K=3

No	DATA	Metode K-means				Metode K-medians			
		Entropy Total	**Purity Total**	**Jumlah Iterasi**	**Waktu (dtk)**	**Entropy Total**	**Purity Total**	**Jumlah Iterasi**	**Waktu (dtk)**
1	IRIS All (Rank=4)	0.453	0.858	6.143	0.054	0.429	0.866	8.714	0.079
2	IRIS 3D (Rank=3)	0.457	0.857	6.857	0.048	0.424	0.861	6.429	0.046
3	IRIS 2D (Rank=2)	0.464	0.855	6.571	0.032	0.436	0.855	5.714	0.030
Rata-rata		0.458	0.857	6.524	0.045	0.414	0.872	6.952	0.052

b. Pengujian dengan Data Reuters-21578

Pengujian yang kedua dilakukan dengan data Reuters-21578. Sama seperti data IRIS, pengujian dengan data ini juga dilakukan sebanyak 6 kali percobaan untuk tiap-tiap jenis data, kecuali pada data Reuters All, hanya dilakukan sebannyak 2 kali percobaan. Tabel 2 menunjukkan kualitas *cluster* yang diperoleh dari setiap metode *clustering* yang digunakan. Baris pertama memperlihatkan kualitas *cluster* ketika data Reuters All (32675 kata dari 8738 dokumen), sedangkan baris kedua, ketiga, dan keempat menunjukkan hasil ketika data diperkecil dimensinya (D600, D400, dan D200).

Kualitas *cluster* yang diperoleh dari pengujian dengan data Reuters-21578 untuk kedua jenis metode *clustering* tidak begitu bagus. Hal ini ditunjukkan

dengan tingginya nilai *entropy* dan rendahnya nilai *purity* yang diperoleh dari kedua jenis metode *clustering* untuk tiap-tiap data Reuters yang digunakan. Salah satu penyebab tidak bagusnya nilai *entropy* dan *purity* yang diperoleh dari pengujian dengan data Reuters diduga karena banyaknya dokumen yang terdapat dalam data tersebut memiliki lebih dari satu jenis kategori.

Tabel 2 Hasil rata-rata perbandingan kualitas *cluster* dengan data Reuters-21578 untuk K=5

No	Data	Metode K-means				Metode K-medians			
		Entropy Total	Purity Total	Jumlah Iterasi	Waktu (dtk)	Entropy Total	Purity Total	Jumlah Iterasi	Waktu (dtk)
1	Data Reuters All (Rank=32675)	1.078	0.527	28.00	2100	1.073	0.584	62.50	6144.7
2	Data Reuters D600 (Rank=600)	1.107	0.511	54.14	70	1.108	0.512	101.00	197.0
3	Data Reuters D400 (Rank=400)	1.107	0.513	56.28	48	1.108	0.512	101.00	108
4	Data Reuters D200 (Rank=200)	1.107	0.511	49.71	21	1.105	0.511	101.00	61

5. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan beserta pembahasan pada bab sebelumnya maka dapat disimpulkan bahwa:

1. Pengujian dengan data IRIS jelas memperlihatkan bahwa kualitas *cluster* yang diperoleh dari metode K-medians lebih baik dibandingkan dengan metode K-means.
2. Berdasarkan pengujian pada data Reuters All, waktu yang dibutuhkan metode K-means jauh labih cepat dibandingkan dengan metode K-medians. K-means hanya membutuhkan waktu sekitar 35 jam sedangkan metode K-medians membutuhkan waktu sebanyak 102,411 jam (4 hari 6 jam).
3. Penggunaan metode *SVD* dalam kasus pengelompokan dokumen adalah solusi yang baik untuk memperkecil waktu dalam melakukan *clustering*.
4. Data Reuters-21578 V.1 kurang cocok digunakan untuk kasus *clustering* dokumen karena terdapat dokumen-dokumen yang tidak memiliki kategori tunggal.
5. Kualitas *cluster* yang diperoleh dari metode *clustering* juga dipengaruhi oleh kualitas dokumen.

Daftar Pustaka

[1.] Agus, E.A., 2008, *Subspace Clustering Pada Data Multidimensi Menggunakan Algoritma Mafia Subspace Clustering On Multidimensional Data Using Mafia Algorithm*, *Skripsi*, IT TELKOM, Jakarta.

[2.] Agusta, Y., 2007, *K-Means – Penerapan, Permasalahan dan Metode Terkait*, Jurnal Sistem dan Informatika Vol. 3 (Pebruari 2007), 47-60.

[3.] Deerwester, S., et al, 1988, *Improving Information Retrieval with Latent Semantic Indexing*, Proceedings 51 American Society for Information Science 25, USA, hlm. 36-40.
[4.] Fuadi Abidin, T. et al., 2010, *Singular Value Decomposition for Dimensionality Reduction in Unsupervised Text Learning Problems,* proceeding of the International Conference on Education Technology and Computer, China.
[5.] Garcia, E., 2006, *Singular Value Decomposition (SVD) A Fast Track Tutorial*, (http://www.miislita.com., diakses 2 Juni 2010).
[6.] Han, J., Micheline, K., 2006. *Data Mining: Concepts and Technique,* Morgan Kaufmann Publishers, San Francisco.
[7.] Karel, R.H. , 2005, *Pembuatan Aplikasi Data Mining untuk Clustering Item dengan Menggunakan Metode Clarans pada Perusahaan X*, *Skripsi*, Universitas Kristen Petra, Surabaya.
[8.] Landauer, T., et al.,1998, *Learning like Human Knowledge with Singular Value Decomposition*, Advances in Neural Information Processing Systems 10, Cambridge: MIT Press, hal. 45-51.
[9.] Liu, B., 2007, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer Berlin Heidelberg, New York.
[10.] MacQueen, J. B., 1967, *Some Methods for Classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1: 281-297.
[11.] Subekti, B., 2000, *Perbandingan Metode-metode Penyelesaian dari Sistem Persamaan Linier yang Singular*, Jurnal Surveying dan Geodesi, Vol.X, No.3.
[12.] Umran, M, et al., 2009, *Pengelompokan Dokumen Menggunakan K-Means dan Singular Value Decomposition: Studi Kasus Menggunakan Data Blog*. Prosiding Seminar Sistem Informasi Indonesia 2009 (Sesindo 2009), Institut Teknologi Surabaya (ITS), Indonesia.