JURNAL TEKNOLOGI TECHNOSCIENTIA ISSN: 1979-8415
Vol. 1 No. 2 Februari 2009

EFEK PENAMBAHAN FRASA DALAM *FEATURE* KATA UNTUK *CLUSTERING* DOKUMEN TEKS

Amir Hamzah[1]

[1] Jurusan Teknik Informatika, IST AKPRIND, Jl. Kalisahak No. 28 Yogyakarta

Masuk: 15 Juni 2008 , revisi masuk: 17 Desember 2008, diterima: 20 Januari 2009

ABSTRACT

Text document clustering has been intensively studied because of its important role in text-mining and information retrieval. High dimensionality problem caused by high number of words is always happened in word-based clustering technique using vector space model. Although extracting words in the preprocessing phase is simple, the collection itself can not only be viewed as a set of words but also a set of partly more than one word phrase. Separating a phrase into its parts can eliminate the actual meaning of phrase. Therefore in order to maintain the context of words a phrase must be maintained as a phrase. It is assumed that by adding phrases to words as features in clustering will improve the performance. This paper will study the comparison of word-based and phrase-based clustering. Two clustering models were chosen i.e. hierarchical and partition. Four similarity techniques i.e.: Group Average, Complete Link, Single Link, and Cluster Center were tried for hierarchical, K-Means and Bisecting K-Mean and Buckshot for partition. A document collection from 200-800 news text that has been categorized ma-nually was used to test these algorithms by using F-measure as criteria of clustering performance. This value was derived from Recall and Precision and can be used to measure the performance of the algorithms to correctly classify the collections. Results show that by adding phrases or simply word pair, although it's still not statistically significant, it slightly improves the performance of clustering.

Keywords*: Word-base clustering, Phrase-based clustering, Clustering performance.*

INTISARI

Clustering dokumen merupakan teknik yang intensif dipelajari karena peran pentingnya dalam *text mining* dan sistem temu kembali informasi. Problem tingginya dimensi ruang vektor yang disebabkan banyaknya jumlah kata selalu terjadi dalam teknik *clustering* berbasis kata menggunakan model ruang vektor. Meskipun mengekstrak kata dalam tahap *preprocessing* cukup sederhana, tetapi koleksi dokumen itu sendiri tidak hanya dapat dilihat sebagai kumpulan kata tetapi juga sebagian adalah kumpulan frasa yang terdiri lebih dari satu kata. Pemisahan frasa menjadi kata terpisah dapat menghilangkan makna sebenarnya dari frasa tersebut. Dengan demikian untuk melindungi konteksnya sebuah frasa harus tetap ditangani sebagai frasa. Diasumsikan bahwa dengan menambah frasa ke dalam *feature* akan meningkatkan kinerja *clustering*. Tulisan ini akan membandingkan kinerja *clustering* berbasis kata dan frasa. Dua model *clustering* dipilih, yaitu model *hierarchi* dan partisi. Untuk model *hierarchi* dipilih empat teknik yaitu *Group Average*, *CompleteLink*, *SingleLink* dan *ClusterCenter, s*edangkan untuk model partisi diambil metode *K-Means* , *Bisecting K-Means* dan Buckshot. Koleksi dokumen teks berita bahasa Indonesia 200 sampai 800 dokumen yang telah dikluster manual digunakan sebagai uji coba. Parameter yang digunakan untuk membandingkan kinerja algoritma adalah F-*measure*, nilai yang diturunkan dari *recall* dan *precision.* Hasil penelitian menunjukkan bahwa penambahan frasa meningkatkan kinerja *clustering*, meskipun uji statistik belum menunjukkan signifikan.

Kata Kunci : *Clustering* Berbasis Kata, *Clustering* Berbasis Frasa, Kinerja *Clustering*

[1] Email: miramzah@yahoo.co.id,
Telp:(0274)-563029

PENDAHULUAN

Clustering dokumen teks menduduki posisi penting dalam *text data mining* dan *text information retrieval*. Hal ini karena dengan intensifnya teknologi digital dalam manajemen menyebabkan koleksi dokumen meningkat eksponensial. Saat ini dalam web lebih dari 36 Milyar dokumen teks dikoleksi google (www. google.com, 2008). Diperkirakan bahwa sebagian besar informasi (80% lebih) dalam suatu perusahaan adalah teks (Tan,1999). Hal ini mendorong kebutuhan riset untuk elaborasi koleksi teks (*text-mining*) dan riset untuk optimalisasi mesin pencari informasi (*IR-system*).

Dalam model ruang vektor dimana koleksi dokumen diwakili oleh matrik kata-dokumen dan sebuah dokumen diwakili oleh sebuah vektor dalam ruang dimensi t, dengan t jumlah kata dalam koleksi dokumen tersebut, umum dijumpai bahwa dimensi t sangat tinggi (Dhillon et al, 2001). Dalam dimensi tinggi jarak antar titik akan cenderung bernilai sama (Hinneburg and Keim, 1999). Hal ini berakibat algoritma *clustering* yang bertumpu pada fungsi jarak menghasilkan solusi yang bias. Reduksi dimensi ruang vektor dapat ditempuh pada tahap *clustering* atau tahap pre-*processing*. Pada tahap *clustering* reduksi ditempuh dengan pendekatan misalnya *projected clustering* (Aggarwal et al, 2000), analisis SVD atau PCA (Gao and Zhang, 2003). Reduksi tahap pre-*processing* ditempuh antara lain dengan seleksi kata (Dhillon, et al, 2002; Hamzah, dkk., 2006). Kata yang terlalu tinggi frekuensinya dibuang dengan cara *stop-word removal*, yaitu membuang kata seperti 'dan', 'ini','itu', 'dengan' dan lain-lain. Sedang kata frekuensi rendah dibuang dengan batas suatu *treshold* tertentu. Cara baku lain reduksi dimensi dalam tahap pre-*processing* adalah dengan *stemming* kata (Rijsbergen, 1979; Hamzah, 2006), yaitu mengembalikan kata ke dalam kata dasarnya.

Dalam model "bag of word" koleksi dokumen hanya diandaikan sebagai koleksi kata, padahal pada kenyataannya dalam dokumen sangat mungkin ada banyak frasa yang tersusun dari dua kata seperti "pasar modal", "kambing hitam", atau frasa tiga kata seperti "terapi tusuk jarum". Memisahkan semua frasa menjadi tinggal kata-kata penyusunnya bisa berakibat makna kata menyimpang jauh dari konteks sebenarnya. Oleh karena itu idealnya *feature* adalah kata dan frasa, seperti yang buktikan oleh Zhang et al (2004) bahwa *feature* frasa lebih baik dalam kinerja pembeda dokumen.

Tidak seperti ekstraksi kata dari dokumen yang secara teknis sangat mudah, ekstraksi frasa memerlukan algoritma yang tidak mudah. Dalam dokumen bahasa inggris riset dalam bidang ekstraksi frasa dari dokumen teks telah banyak dilakukan, antara lain oleh Maynard and Ananiadou (1999) dan Frantzi and Ananiadou (2003). Sayangnya dalam kata (teks) bahasa indonesia riset seperti ini belum banyak dilakukan karena riset bidang komputasi linguistik masih sangat minim (Nazief, 2000). Penelitian ini dimaksudkan sebagai penelitian awal untuk ekstraksi kata dari dokumen teks menggunakan teknik statistik pasangan kata. Selanjutnya pengaruh frasa yang diekstraksi dalam kinerja *clustering* dokumen teks berbahasa indonesia akan dilakukan.

Model ruang vektor untuk koleksi dokumen mengandaikan dokumen sebagai sebuah vektor dalam ruang kata (*feature*). Klustering dokumen dipandang sebagai pengelompokan vektor berdasarkan suatu fungsi *similarity* antar dua vektor tersebut. Jika koleksi n buah dokumen dapat diindeks oleh t buah *term*/*feature* maka suatu dokumen dapat dipandang sebagai vektor berdimensi t dalam ruang term tersebut. Dengan demikian koleksi dokumen dapat dituliskan sebagai matrik kata-dokumen X, yang dapat ditulis :

$$X = \{x_{ij}\}\ i= 1,2,..t\ ;\ j =1,2,..\ n \qquad (1)$$

x_{ij} adalah bobot term i dalam dokumen ke j

Menurut Luhn (1958), kekuatan pembeda terkait dengan frekuensi term (*term-frequency*, *tf*). *Term* yang memiliki kekuatan diskriminasi adalah *term* dengan frekuensi sedang. Pemotongan term dengan frekuensi tinggi dilakukan dengan membuang *stop-word*, seperti 'ini','itu','yang', 'yaitu' dan lain-lain yang

dapat mengurangi frekuensi *feature* 30 sampai 40 persen (Rijsbergen, 1979).

Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan *term* dalam dokumen karena dipercaya bahwa frekuensi kemunculan *term* merupakan petunjuk sejauh mana *term* tersebut mewakili isi dokumen. Menurut Luhn (1958), kekuatan pembeda terkait dengan frekuensi term (*term-frequency*, *tf*), di mana *term* yang memiliki kekuatan diskriminasi adalah *term* dengan frekuensi sedang. Mempertimbangkan panjang dokumen dan kemunculan term dalam dokumen pembobotan baku yang digunakan adalah *term-frequency invers-document freqeuency* (TF-IDF) (Steinbach et al, 2000) sebagai berikut :

Kesamaan antara dokumen D_i dengan dokumen D_j umumnya diukur dengan fungsi similaritas tertentu. Menurut (Chisholm et al, 1999) untuk tujuan *clustering* dokumen fungsi yang baik adalah fungsi similaritas *Cosine*, berikut :

$$\textit{Cosine}\text{-sim}(D_i,D_j)= \frac{\sum_{i=1}^{t} D_i D_j}{\sqrt{\sum_{i=1}^{t}(D_i)^2 \sum_{i=1}^{t}(D_j)^2}} \qquad (2)$$

Jika vektor Di dan Dj masing-masing ternormalisasi sehingga masing-masing panjangnya satu, maka fungsi *cosine* menjadi :

$$\textit{Cosine}\text{-sim}(D_i,D_j) = \sum_{i=1}^{t} D_i D_j \qquad (3)$$

Secara umum *feature* yang digunakan untuk mewakili dokumen dalam model raung vektor adalah kata. Hal ini karena ekstraksi kata dari dokumen relatif mudah, yaitu hanya mendeteksi deretan karakter yang diakhiri dengan spasi. Jika dirancang bahwa angka tidak merupakan bagian dari kata maka dalam bahasa Indonesia karakter khusus yang mewakili kata hanya tanda hypen ("-"), yang menunjukkan kata ulang, selainnya adalah karakter abjad.Penelitian untuk teks bahasa inggris yang melibatkan frasa menunjukkan bahwa melibatkan frasa dalam *feature* dapat meningkatkna kinerja *clustering* (Zhang et al, 2004).

Penelitian tentang deteksi dan ekstraksi frasa dalam bahasa Inggris juga telah cukup banyak dilakukan (Frantzi and Ananiadou (2003). Metode seleksi beragam mulai dengan pendekatan statistik sampai pendekatan *natural language processing* (NLP). Untuk kasus bahasa Indonesia penelitian di bidang ini masih sangat minim (Nazief, 2000).

Dengan latar belakang itu dalam penelitian ini frasa didefinisikan sebagai dua kata yang saling berdekatan yang memiliki makna tertentu yang bisa berbeda dengan makna kata-kata tunggalnya, misalnya "kambing hitam". Teknik ekstraksi kata ditempuh dengan cara sederhana yaitu melakukan penghitungan frekuensi kemunculan dari pasangan dua kata. Selanjutnya seperti pada kata setelah dibatasi frekuensi minimal kemunculan, analisis variansi frekuensi dilakukan untuk melakukan seleksi. sebagai persamaan berikut (Dhillon et al, 2001; Dhillon et al, 2002) :

$$q_i(t)= \sum_{j=1}^{ni} f_j^{\,2} - \frac{1}{ni}\left[\sum_{j=1}^{ni} f_j\right]^2 \qquad (4)$$

dengan q_i adalah variansi jika frekuensi minimal kata/frasa muncul dalam analisis adalah i (i=0,1,2,...).

Clustering didefinisikan sebagai upaya pengelompokan data ke dalam kluster sehingga data-data didalam kluster yang sama memiliki lebih kesamaan dibandingkan dengan data-data pada kluster yang berbeda (Jain and Dubes, 1998). Dikenal dua pendekatan, yaitu *hierarchical* dan *partisional* dengan masing-masing memiliki banyak variasi.

Metode klustering secara *agglomerative* berawal dari n= cacah dokumen sebagai cluster. Dengan menggunakan fungsi similaritas antar kluster kemudian proses penggabungan kluster terdekat dilakukan. Ukuran similaritas antar kluster antara lain, misalnya: *UPGMA* CST dan IST *Single Link, Complete Link* (Jain and Dubes, 1998). Berikut ini ringkasan masing-masing teknik tersebut:

- *Unweighted Pair Group Method Average similarity* (UPGMA): Similaritas dua kluster diukur dengan rata-rata hi-

tung similaritas antar seluruh pasangan titik antara kedua kluster.

- *Single Link* (SL): jarak terbaik dua kluster diwakili oleh jarak terdekat (similaritas tertinggi) dari dua titik dari dua kluster.
- *Complete Link(CL)*: jarak terbaik dua kluster diwakili oleh jarak terjauh (similaritas terendah) dari dua titik dari dua kluster.
- *Centorid-Similarity Technique (CIST):* Jarak antar kluster ditentukan dengan jarak antar pusat kluster.

Secara teknis masukan bagi algoritma *hierarchical clustering* adalah matriks similaritas antar dokumen yang berukuran NxN. Iterasi yang setiap tahapnya melakukan penggabungan kluster dilakukan dengan melakukan *update* pada matrik similaritas. Hal inilah yang menyebabkan algoritma ini memiliki kompleksitas waktu dan ruang $O(N^2)$.

Algoritma K-*means clustering* merupakan algortima iteratif dengan meminimalkan jumlah kuadrat *error* antara vektor objek dengan pusat kluster terdekatnya (Jain and Dubes, 1998), yaitu :

$$\sum_{j=1}^{k}\sum_{x\in\pi_j}\left\|x-m_j\right\|^2 \qquad (5)$$

di mana $\mathbf{m}_j$ adalah pusat kluster (*mean vector*) dalam kluster ke j. Proses dimulai dengan mula-mula memilih secara random k buah dokumen sebagai pusat kluster awal.

Metode *Bisecting K-means* (Steinbach et al, 2000) mencoba menggabungkan pendekatan *partitional* dengan *divisive hierarchi*, yaitu mula-mula seluruh dokumen dibagi dua dengan cara *K-means* (*bisecting-step*). Selanjutnya cara itu dikenakan pada tiap-tiap kluster sampai diperoleh K buah kluster.

Algoritma *Buckshot* menggunakan pendekatan *hierarchie agglomerative* untuk mendapatkan k buah vektor sebagai pusat kluster awal. Langkah *Buckshot* mula-mula mengambil sampel acak sebesar √kn dokumen, dikluster dengan prosedur *hierarchie agglomerative* untuk mendapatkan k buah kluster. Selanjutnya dari partisi awal *Buckshot* proses *refinement* dilakukan sebagaimana dalam K-*means clustering* (Dhillon et al, 2001).

Validitas yang digunakan diturunkan dari *Confusion Matrix* yaitu matriks yang disusun berdasarkan berapa banyak objek yang diklasifikasikan dengan benar oleh proses *clustering*. Parameter kualitas *clustering* yang dapat diturunkan dari *confusion matrix* yang umum digunakan untuk document clustering adalah *F-measure* (persamaan (6)).

$$\text{F-}measure = \frac{2PR}{P+R} \qquad (6)$$

PEMBAHASAN

Koleksi dokumen yang digunakan untuk eksperimen adalah koleksi yang diambil dari koleksi Asian et al (2004). Koleksi tersebut dikemas menjadi 5 koleksi, yaitu 200, 300, 400, 500 dan 800 dokumen yang masing-masing telah dikluster secara manual. Adapun statistik koleksi tes tersaji dalam Tabel 1.

Tabel 1. Koleksi Dokumen Untuk Penguian algoritma *clustering*

Coll Name	∑ doc	∑ clus	Clust Size	∑uniq Word	avg word /doc
T200	200	10	Sama	6.652	372
T300	300	10	Beda	8.472	373
T400	400	11	Beda	10.153	388
T500	500	13	Beda	11.637	385
T800	800	14	Beda	15.752	410

Setiap koleksi terdiri dari sejumlah dokumen dengan format setiap dokumen seperti gambar 1.

```
<DOC>
<DOCNO>news035-html</DOCNO>
   banyaknya calhaj kalsel
  bukan indikator membaiknya
   perekonomian ..........
</DOC>
```

Gambar 1. Format koleksi dokumen untuk Tes

Proses *pre-processing* berupa ekstrak kata, frasa, komputasi statistik frekuensi sampai dengan penyusunan matrik dilakukan dengan kode program JAVA (jdk1.4.2). Frase dalam eksperimen ini adalah dua buah kata yang muncul berdampingan dengan frekuensi tertentu. Selanjutnya diujikan metode-metode *clustering* yaitu: metode *hierarchi ag-*

lomerative (strategi similaritas: *Single Link, Complete Link, Group Average, centroid similarity, intra cluster similarity*), metode *partitional* (K-means, *bisecting k-means, Buckshot*). Program dirancang dengan script MATLAB.

Hasil pengujian kinerja *feature* kata dan frasa diukur melalui nilai F-*measure* yang membandingkan *feature* kata saja, frasa saja dan *feature* campuran. Uji statistik hasil dengan uji t untuk pengamatan berpasangan. Pada semua koleksi dilakukan pre-*processing* dengan batas minimal nilai tertentu. Ditentukan 3 macam penggunaan *feature*, yaitu campuran, kata saja dan frasa saja. Selanjutnya clustering dilakukan dengan 100% *feature* yang didapat dengan pembatasan minimal tersebut.

Tabel 2. Statistik Kata+Frasa, Kata dan Frasa

Kol	Min f	Σ Kt+Fr	Σ Kata	Σ Frasa
T200	3	3037	1852	1069
T300	4	3306	2067	1142
T400	5	3588	2247	1242
T500	6	3748	2377	1237
T800	10	3680	2488	1108

Seleksi *feature* dengan prosentase 20%,15%,10% dan 5% dengan analisis varian frekuensi kemunculan juga dilakukan. Statistik *feature* berdasarkan *threshold* minimal pada koleksi tersaji dalam Tabel 2.

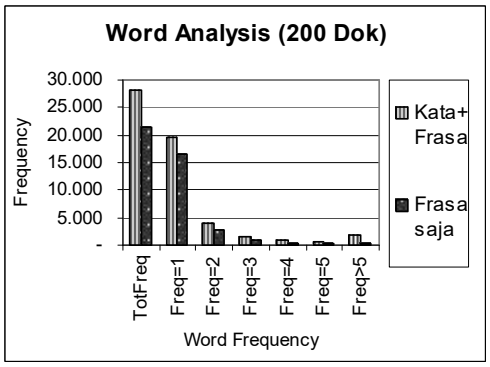
Frasa didefinisikan sebagai pasangan dua kata yang berturutan dalam teks yang sering muncul. Asumsinya jika dua kata tersebut merupakan frasa maka kemungkinan besar frekeunsi kemunculannya cukup tinggi karena jika pasangan kata tersebut acak maka kemunculannya akan rendah sehingga ia akan dominant pada frekuensi rendah. Hal ini dapat dijelaskan pada pola kemunculan kata dan pasangan kata yang sama pada seluruh koleksi, salah satunya untuk koleksi 200 kata seperti dalam Tabel 3 Gambar 2.

Tabel 3. Statistik frekuensi term 200 dokumen

	Kata dan Pas Kata	Hanya Kata	% Pas Kata
TotFreq	28,106	21,454	76.33%
Freq=1	19,694	16,609	84.34%
Freq=2	3,981	2,853	71.67%
Freq=3	1,394	811	58.18%
Freq=4	767	415	54.11%
Freq=5	467	208	44.54%
Freq>5	1,803	225	12.48%

Tabel 4. Perbandingan kinerja clustering dengan *feature* Kata dan Kata+Frasa diukur dari F-*Measure* untuk koleksi dokumen 200 dokumen

	100% term		20% term		15 % term		10% term		5% term	
Metode	Kt+Frs	Kata	Kt+Frs	Kata	Kt+Frs	Kata	Kt+Frs	Kata	Kt+Frs	Kata
hcaUPGMA	0.93	0.92	0.86	0.86	0.86	0.85	0.86	0.84	0.90	0.96
hcaCST	0.60	0.50	0.80	0.63	0.69	0.71	0.80	0.78	0.83	0.85
hcaIST	0.74	0.72	0.95	0.89	0.81	0.84	0.95	0.88	0.91	0.76
hcaSL	0.41	0.28	0.41	0.28	0.28	0.28	0.41	0.28	0.36	0.50
hcaCL	0.98	0.95	0.88	0.86	0.98	0.96	0.88	0.90	0.80	0.82
spherekm	0.75	0.73	0.63	0.71	0.66	0.65	0.61	0.61	0.70	0.72
bisectkm	0.99	0.90	0.93	0.97	0.90	0.98	0.98	0.93	0.98	0.99
buckshot	0.64	0.74	0.77	0.84	0.79	0.86	0.67	0.86	0.77	0.71
Rata-rata	0.73	0.67	0.78	0.54	0.72	0.71	0.78	0.73	0.77	0.76



Gambar 2. Frekuensi Kata+Frasa dengan Frasa

Terlihat dari Tabel 3 dan Gambar 2 bahwa diatas frekuensi 5 pasangan kata hanya 15% dari campuran kata dan pasangan kata.

Analisis kinerja *clustering* berdasar nilai F-*measure* menggunakan seluruh metode dilakukan pada tiap koleksi. Perbandingan dilakukan antara *feature* campuran (Kata+Frasa) dan *feature* hanya kata. Pada koleksi T200 (200 dokumen) hasil perbandingan tersaji pada Tabel 4. Karena sempitnya ruang perbandingan untuk koleksi T300,T400, T500 dan T800 tidak ditampilkan dan hanya akan ditampilkan hasil uji statistik perbandingan kinerja tersebut.

Terlihat dari Tabel 4 bahwa kinerja *clustering* pada berbagai metode terkadang unggul untuk *feature* kata+frasa dan kadang unggul untuk kata saja. Pola seperti ini terjadi tidak hanya pada koleksi T200 tetapi pada semua koleksi yang diujikan. Secara rata-rata *feature* campuran bernilai lebih tinggi dari *feature* kata saja, tetapi dari uji statistika *rank wilcoxon* untuk sampel berpasangan menghasilkan uji beda tidak signifikan pada seluruh koleksi yang diuji (Tabel 5). Semua menghasilkan uji *non-sig*, yang berarti belum dapat dikatakan bahwa ada perbedaan kinerja *clustering* karena penambahan pasangan kata pada *feature* kata.

Kenyataan ini dapat diduga disebabkan secara statistik kemunculan frasa (pasangan kata) dengan analisis frekuensi yang sama dengan kata paling tinggi adalah 38% dari seluruh *feature* (kata+frasa). Jika dilakukan seleksi maka frasa akan semakin mengecil pada seleksi *feature* sampai 5%, maka hanya terdapat paling tinggi 9% *feature* adalah pasangan kata. Secara rinci prosentasi frasa (pasangan kata) akan berubah jika seleksi *feature* dilakukan seperti pada Tabel 6.

Tabel 5. Statistik Uji t *rank-wilcoxon* beda sample berpasangan pada alpa 5%

Kol	Rata-rata beda	T-value	T-tabel	Hasil Uji beda
T200	0.007	0.56	1.96	Non sig
T300	0.015	0.879	1.96	Non sig
T400	0.009	0.770	1.96	Non sig
T500	0.018	0.885	1.96	Non sig
T800	0.012	0.812	1.96	Non sig

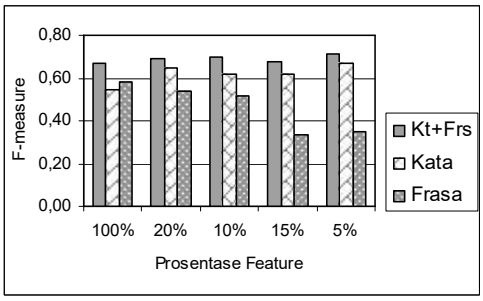
Tabel 6. Penurunan prosentasi frasa (pasangan kata) dalam *feature* campuran oleh seleksi *feature*

Kol	100% featr	20% featr	15% featr	10% featr	10% featr
T200	38%	17%	15%	13%	9%
T300	37%	18%	13%	10%	7%
T400	38%	16%	13%	9%	8%
T500	36%	16%	11%	9%	9%
T800	33%	15%	11%	9%	9%

Penggunaan *feature* sepenuhnya frasa menunjukkan hasil yang relatif lebih rendah, baik pada 100% *feature*, maupun 20%,15%,10% atau 5%. Hasil lebih rendah ini konsisten pada semua koleksi yang diujikan. Sebagai contoh untuk koleksi T500 rata-rata F-*measure* untuk seluruh metode tersaji pada Gambar 3. Kinerja hanya frasa semakin menurun jika digunakan seleksi *feature* yang rendah. Pada prosentasi 10% dan 5% perbedaan antara frasa dan kata serta antara frasa dan campuran terlihat signifikan. Dari gambar tersebut juga terlihat jelas bahwa kinerja yang terbaik adalah manakala digunakan *feature* campuran. Ini terlihat pada semua prosentasi *feature* yang diambil.

Salah satu penyebab menurunnya nilai F-*measure* yang berarti menurunnya kinerja *clustering* karena pada kenyataannya *feature* kata dan frasa akan cenderung lebih banyak kata. Pemilihan frasa dengan hanya mengambil dua kata yang berdekatan dan dengan frekuensi tertentu, tanpa melihat makna be-

lum menjamin bahwa pasangan kata tersebut benar-benar sebuah frasa. Tabel 7 menyajikan contoh 20 pasangan kata terbaik yang diekstraksi dari koleksi T300. Terlihat beberapa pasangan kata seperti "per gram", "per dolar", "juara piala" adalah bukan frasa yang benar. Pasangan kata yang "abdullah syaf**ei"** dengan **"**abdullah syaf**iie"** adalah contoh pasangan yang sebenarnya sama tetapi diidentifikasi berbeda karena ketidak konsistenan wartawan dalam menulis berita. Hal ini juga terjadi pada pasangan kata "j**a**maah haji" yang kadang ditulis sebagai "j**e**maah haji". Tentu saja ini merupakan noise yang menurunkan kinerja *clustering*.



Gambar 3. Rata-rata F-measure pada koleksi T500

Tabel 7. Contoh pasangan kata terseks-trak dari koleksi

per dolar	**manchester united**
jemaah haji	**arab saudi**
abdullah syafei	**menko polkam**
piala dunia	**pasar uang**
terhadap dolar	**tenaga kerja**
mata uang	**kota ambon**
banda aceh	**pasukan tni**
liga utama	juara piala
abdullah syafiie	**jamaah haji**
juara liga	per gram

KESIMPULAN

Penambahan *feature* frasa yang diambil dari pasangan kata dengan frekuensi tertentu meningkatkan hasil kinerja *clustering*, meskipun pengujian secara statistik peningkatan belum signifikan.

Jika digunakan seleksi *term* atas *feature* campuran dengan hanya mengambil beberapa persen dari total *feature* campuran, maka jumlah frasa yang terlibat akan menurun sampai dibawah 10%.

Penggunaan *feature* hanya frasa memiliki kinerja yang rendah dibandingkan dengan *feature* campuran (kata dan frasa) ataupun *feature* kata saja. Kinerja ini semakin jika digunakan seleksi *feature* frasa pada prosentase 10% atau 5%. Hal ini dapat dipahami karena pada kenyataannya suatu dokumen teks bukanlah kumpulan frasa tetapi kata dan frasa dengan frekuensi kata yang jauh lebih besar dari pada frasa.

Diperlukan penelitian lebih jauh untuk melakukan ekstraksi frasa dengan teknik yang lebih baik dari sekedar melakukan statistik pada kemunculan pasangan kata sebagai *feature*.

DAFTAR PUSTAKA

Aggarwal, C.C. and Yu, P.S., 2000, Finding Generalized Projected Cluster in High Dimensional Spaces, *Proc.ACM SIGMOD Conf*.

Asian, J., Williams, H.E., and Tahaghoghi, S.M.M., 2004, Tesbed for Indonesian Text Retrieval, *9th Australian Document Computing Symposiom*, Melbourne December, 13, 2004.

Chisholm, E. and Kolda, T.G. , 1999, *New Term Weighting Formula for the Vector Space Method in Information Retrieval*, Research Report, Computer Science and Mathematics Division, Oak Ridge National Library, Oak Ridge, TN 3781-6367, March 1999.

Dhillon, S. I., J. Fan, and Guan, Y., 2001, *Efficient Clustering of Very Large Document Collection*, www. citeseer.ist.psu.edu/dhillon01.html.

Dhillon, I., Kogan, J. and Nicholas, C., 2002, *Feature Selection and Document Clustering*, www.csee. umbc.edu/cadip/2002Symposim/ koghan.pdf.

Jain, A.K. and Dubes, R. C. , 1998, *Algorithms for Clustering Data*, Prentice-Hall.

Frantzi K.T. and Annaniadou, S., 2003, *Automatic Term Recognition Using Contextual Cues*, DELOS'03,www.ercim.org DELOS03/ frantzi.pdf.

Gao, J. and Zhang, J., 2003, *Clustered SVD Strategies in Latent Semantic Indexing*, Technical Report No. 382–03, Department of Computer Science, University of Kentucky, Lexington, KY.

Hamzah, A, 2006, Pengaruh Stemming Kata Dalam Peningkatan Unjuk Kerja Document Clustering Untuk Dokumen Berbahasa Indonesia , *Prosiding Seminar Nasional Riset Teknologi Informasi*, AKAKOM, Juli, 2006.

Hamzah, A., Soesianto, F., Susanto, A., Istiyanto, J.E., 2006, Seleksi Feature Kata Berdasarkan Variansi Kemunculan Kata Dalam Peningkatan Unjuk Kerja Document Clustering Untuk Dokumen Berbahasa Indonesia, *Pakar, Jurnal Teknologi Informasi dan Bisnis* , Vol.7, No.3. , pp. 181-190.

Hinneburg, A. and Keim, D.K., 1999, Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering, *Proceeding of 25th VLDB Conference*, Edinburg, Scotland.

Luhn, H.P.,1958, The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development,* 2:159-165.

Maynard, D. and Ananiadou, S., 1999, *Incorporating Linguistic Information for Multi-Word Term Extraction*, Dept.of Computing & Mathematics, Manchester, MI 5GD, UK.

Nazief, B., 2000, *Development of Computational Linguistic Research: a Challenge for Indonesia*, Computer Science Center, University of Indonesia.

Rijsbergen, C. J., 1979, *Information Retrieval*, Information Retrieval Group, University of Glasgow , UK.

Steinbach, M., Karypis, G., Kumar, V., 2000, *A Comparison of Document Clustering Techniques,* University of Minnesota, Technical Report #00-034, at http://www. cs.umn. edu/tech_reports.

Tan, AH, 1999, *Text Mining: The state of the art and the challenges*, *Kent Ridge Digital Labs* 21 Heng Mui Keng Terrace Singapore 119613.

www.google.com

Zhang, Y., E. Milios and Heywood, N. Z., 2004, *A Comparison of Key-word and Keyterm-based Methods for Automatic Web Site Summarization*, Tecnical Report, Faculty of Computer Science, University Ave. Halivax, Nova Scotia, 2004.