

Pengelompokan Berita Indonesia Berdasarkan Histogram Kata Menggunakan *Self-Organizing Map*

Ambarwati*¹, Edi Winarko²

¹Jurusan Teknik Komputer, Instituto Profissionale de Canossa, Dili

²Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta
e-mail: *kristiyani.ambarwati@yahoo.com, ewinarko@gmail.com

Abstrak

Berita merupakan sumber informasi yang dinantikan oleh manusia setiap harinya. Manusia membaca berita dengan kategori yang diinginkan. Jika komputer mampu mengelompokkan berita secara otomatis maka tentunya manusia akan lebih mudah membaca berita sesuai dengan kategori yang diinginkan. Pengelompokan berita yang berupa artikel secara otomatis sangatlah menarik karena mengorganisir artikel berita secara manual membutuhkan waktu dan biaya yang tidak sedikit.

Tujuan penelitian ini adalah membuat sistem aplikasi untuk mengelompokkan artikel berita dengan menggunakan algoritma *Self Organizing Map*. Artikel berita digunakan sebagai input data. Kemudian sistem melakukan pemrosesan data untuk dikelompokkan. Proses yang dilakukan sistem meliputi *preprocessing*, *feature extraction*, *clustering* dan *visualize*. Sistem yang dikembangkan mampu menampilkan hasil *clustering* dengan algoritma *Self Organizing Map* dan memberikan visualisasi dengan *smoothed data histograms* berupa *island map* dari artikel berita. Selain itu sistem dapat menampilkan koleksi dokumen dari lima kategori berita yang ada pada tiap tahunnya dan banyaknya kata (*histogram kata*) yang sering muncul pada tiap artikel berita. Pengujian dari sistem ini dengan memasukan artikel berita, kemudian sistem memprosesnya dan mampu memberikan hasil *cluster* dari artikel berita yang dimasukan.

Kata kunci— Pengelompokan berita Indonesia, pengelompokan berdasar histogram kata, pengelompokan berita menggunakan SOM

Abstract

News is awaited information resources by humans every day. Human reading the news with the desired category. If the computer able to news clustering with automatically, humans of course will be easier to read the news according to the desired category. News clustering in the form of news articles with automatically very interesting because it organizes news articles manually takes time and costs not a little bit.

The purpose of this research is to create a system application for grouping news articles by using the *Self Organizing Map* algorithm. News article be used as input into the system. News articles used as input data. Then the system performs data processing until to be clustered. Processes performed by the system covers: *preprocessing*, *feature extraction*, *clustering* and *visualize*. The system developed is able to display the results clustering of the *Self Organizing Map* algorithm and gives visualization of the *Smoothed Data Histograms* in the form of *island map* from news articles. Additionally the system can display a word histogram and news articles from five categories news in each year. Testing of this system by entering the news articles, then the system performs data processing and gives results of a cluster from news articles that input.

Keywords— Indonesia news clustering, clustering based on words histograms, news clustering using SOM

1. PENDAHULUAN

Jumlah sumber berita berbahasa Indonesia yang tersedia di internet semakin besar [1] apalagi berita dalam jumlah besar setiap harinya. Pengelompokan berita yang berupa artikel secara otomatis sangatlah menarik karena mengorganisir artikel secara manual akan menjadi sangat mahal [2] atau tidak dapat dikerjakan dengan mudah karena keterbatasan waktu dari jumlah dokumen yang digunakan. Artikel dapat dikelompokkan dengan membagi artikel berita ke dalam kategori berita yang telah ditentukan sebelumnya. Setiap kategori berita mempunyai kemiripan isi berita yang yang tentunya bisa divisualisasikan dengan *cluster*.

Clustering dapat digunakan untuk membantu menganalisis berita dengan mengelompokkan secara otomatis berita yang memiliki kesamaan [3] isi. *Clustering* adalah aktifitas pengelompokan berita tanpa supervisi ke dalam *clusters*. Masalah mendasar dalam *clustering* artikel berita adalah mengelompokkan artikel berita yang mempunyai kemiripan isi diantara artikel berita satu dengan artikel berita yang lain [4]. Salah satu algoritma yang bisa digunakan untuk menyelesaikan *clustering* adalah algoritma SOM Kohonen. [5] mengemukakan bahwa koleksi dokumen yang besar dapat disusun sesuai dengan kesamaan tekstual dan statistik representasi dari isi tekstual digunakan sebagai fitur vektor dokumen dengan algoritma *Self-Organizing Map* (SOM).

Dalam penelitian [6] yang dikembangkan ini menggunakan algoritma *Dynamic Time Warping* (DTW) untuk mencocokkan kurva linier dengan membandingkan gaya tulisan tangan dan SOM dipakai sebagai algoritma dalam pembentukan *cluster*. Algoritma SOM melakukan pemetaan nonlinear yang memelihara topologi lokal properti dari kumpulan *database* tulisan tangan tersebut. Hasilnya adalah menemukan bentuk gaya penulisan tangan dan dapat mengelompokkan tulisan tangan tersebut secara otomatis dengan sistem yang dibangun. Penelitian [7] mengenai pengintegrasian sistem Bveritas dan mengatur artikel berita pada situs-situs berita berbahasa Inggris. Sistem mengatur berita secara otomatis dari beberapa situs berita menggunakan menggunakan *Self-Organizing Map* (SOM). Pengelompokan dokumen didasarkan pada distribusi frekuensi dari kata-kata yang digunakan pada artikel berita tersebut. Sistem yang dibangun menyediakan topik dan kategorisasi regional dengan kombinasi *headlines* dan *breaking news*. Sistem juga menyediakan layanan yang bermanfaat seperti ringkasan berita dimana seorang pengguna dapat memilih tingkat ringkasan berita yang diinginkan pada suatu artikel berita. Penelitian [1] mengenai *clustering* untuk *corpus* berita berbahasa Indonesia. Penelitian ini menggunakan metode pembobotan log-tf (*logarithmic term frequency*) dan log-tf.idf (*logarithmic term frequency – inverse document frequency*). Algoritma yang digunakan dalam *clustering* yaitu algoritma K-Means. Sistem yang dibuat mampu menghasilkan *clustering* berita dapat digunakan untuk mencari berita yang terkait, mencari kategori baru dan memperlihatkan keterkaitan antara berita.

2. METODE PENELITIAN

SOM merupakan metode pengelompokan yang menyediakan penataan kelas-kelas berdasarkan topologinya. SOM dilatih secara iteratif melalui sejumlah *epoch*. Sebuah *epoch* didefinisikan sebagai proses dari semua pola input sehingga masing-masing pola input akan diproses sebanyak jumlah *epoch* [8]. SOM adalah jaringan saraf *single layer* dimana *neuron* ditetapkan sepanjang grid n-dimensi [8]. Jaringan SOM Kohonen terdiri dari dua lapisan (*layer*), yaitu lapisan *input* dan lapisan *output*. Setiap *output neuron* menerima *input* melalui bobot yang terhubung dengan masukan, sehingga vektor bobot memiliki dimensi yang sama dengan vektor *input*.

Awalnya algoritma SOM melakukan inisialisasi berupa bobot (w_{ij}) yang diperoleh secara acak untuk tiap node. Setelah bobot (w_{ij}) diberikan maka jaringan diberikan *input* (x_i). Setelah *input* diterima jaringan akan melakukan perhitungan jarak *vector* $D(j)$ yang didapat dengan menjumlahkan selisih antara *vector* bobot (w_{ij}) dengan *vector input* (x_i).

$$D(j) = \sum_i (w_{ij} - x_i)^2 \dots\dots\dots(1)$$

Setelah jarak antara node diketahui maka ditentukan nilai minimum dari perhitungan jarak *vector* $D(j)$, maka tahap selanjutnya melakukan perubahan bobot w_{ij} (*new*).

$$w_{ij}(\text{new}) = w_{ij}(\text{old}) + \alpha [x_i - w_{ij}(\text{old})] \dots\dots\dots(2)$$

Pada proses untuk mendapatkan bobot baru memerlukan nilai *learning rate* (α) yaitu $0 \leq \alpha \leq 1$. Nilai *learning rate* pada setiap *epoch* akan berkurang menjadi $\alpha_{(i+1)} = 0.5\alpha$ [8]. Kondisi penghentian pengujian dilakukan dengan menghitung selisih antara bobot w_{ij} (*new*) dengan w_{ij} (*old*), apabila nilai w_{ij} hanya berubah sedikit saja, berarti pengujian sudah mencapai konvergensi sehingga dapat dihentikan [9].

2.1. Analisa

Penelitian ini membahas masalah *clustering* pada artikel berita Indonesia yang diambil dari majalah Tempo. Diperlukan tahapan-tahapan yang harus dilakukan untuk mendapatkan *cluster* dari artikel berita tersebut. Tahapan-tahapan yang harus dilakukan adalah *preprocessing*, *feature extraction*, *clustering* dan *visualisasi*.

2.1.1. Preprocessing

Preprocessing merupakan tahap awal yang dilakukan untuk meningkatkan kualitas dari data. Sebelum proses *preprocessing* dilakukan, perlu disiapkan artikel berita yang akan diproses dalam bentuk *file* teks dan disimpan. Metodologi yang dilakukan dalam *preprocessing* yaitu *tokenization*, *stopword removing* dan *stemming*.

Proses yang terjadi dalam *tokenization* dalam penelitian ini sebagai berikut: artikel berita akan diproses dengan menjadikan semua huruf besar (*capital*) menjadi huruf kecil (*lower case*) proses ini dinamakan *case folding*. Kemudian sistem akan melakukan langkah untuk memisahkan deretan kata di dalam kalimat, paragraf atau halaman menjadi *token* atau potongan kata tunggal. Apabila terdapat deretan angka pada artikel sistem akan menghapusnya secara otomatis.

Pada *stopwords removing system* akan mencocokkan kata hasil dari proses *tokenization* dengan kata-kata yang terdapat dalam kamus *stopwords* yang telah dibuat, jika terdapat kata yang sama maka kata itu akan dihapus.

Stemming akan memproses kata tersebut satu persatu. Proses ini akan mengecek kata hasil dari *stopwords* dan membandingkannya dengan kamus kata dasar (*rootwords*) jika ditemukan sama maka kata tersebut akan disimpan. Apabila kata yang dicocokkan tidak diketemukan dalam kamus kata dasar maka akan sistem akan mengenali sebagai kata turunan. Kemudian sistem akan mengubah kata turunan menjadi kata dasar menggunakan algoritma Nazief & Adriani.

2.1.2. Feature Extraction

Feature extraction merupakan proses untuk menciptakan *feature vector* yang menggambarkan isi dari data. *Feature extraction* pada sistem ini memanfaatkan *package* SOMLib. Pada sistem yang dibangun proses yang dilakukan dalam *feature extraction* memerlukan 4 *class* yang ada dalam *package* SOMLib, yaitu: *wordsexc*, *templatevectorex*, *reducerexc* dan *extractorex*.

Pada *class wordsexc*, langkah-langkah yang dilakukan antara lain: menetapkan nilai minimum dari panjang kata dan menterjemahkan artikel berita (hasil dari proses *stemming*) ke dalam *file* histogram. Hasil dari proses dari *class wordsexc* adalah suatu *hashtable* berisi informasi yang mungkin mewakili kata-kata tunggal dan jumlah pemunculan kata. *Class templatevectorex* berfungsi untuk menterjemahkan *file* histogram kedalam *template vector* dengan menggabungkan *file-file* yang dibuat pada proses *wordhistogram* menjadi sebuah *file* tunggal dan membuat pengindeksan dari artikel berita secara lengkap. *Class reducerexc* berfungsi untuk menetapkan minimum dan maksimum pemunculan dari sebuah kata, dengan

nilai antara 0 dan 1 serta menghapus kata yang mempunyai pemunculan sedikit. *Class extractorex* menggunakan pembobotan *term frequency* (tf) yang dikalikan dengan *inverse document frequency* (idf). Artikel berita direpresentasikan dengan matriks *term-frequency* yang mempunyai nilai antara 1 dan 0. Nilai nol berarti bahwa *term* tersebut tidak hadir di dalam dokumen. Setiap sel dalam matriks bersesuaian dengan bobot yang diberikan dari suatu *term* dalam dokumen yang ditentukan

2.1.3. Clustering

Clustering pada penelitian ini menggunakan SOMToolbox untuk mendapatkan hasil *cluster*. Pembentukan *cluster* dimulai dengan memberikan *input file vector* dari *feature extraction* dan menkonversi format data tersebut dengan *method som_data_struct(Array)* untuk membuat struktur data SOM dengan format *file mat*. Untuk mendapatkan *file mat* yang akan digunakan untuk proses SOM *cluster*. Melakukan proses normalisasi, dimana normalisasi digunakan untuk menormalkan vektor dokumen. Kemudian sistem akan membuat *cluster* dengan metode *Self Organizing Map*.

Parameter yang dibutuhkan dalam proses *clustering*: Lapisan *input* berupa matriks 2 dimensi berukuran XDIM x VEC_DIM, XDIM adalah jumlah data dan VEC_DIM adalah jumlah dimensi data sesuai dengan hasil *feature extraction*. Lapisan *output* berupa matriks 2 dimensi berukuran YDIM x VEC_DIM, YDIM adalah jumlah map unit dan VEC_DIM adalah jumlah dimensi data sesuai dengan hasil *feature extraction*. Menggunakan topologi hexagonal.

Pada proses SOM *cluster* ini sistem akan memanggil fungsi-fungsi yang terdapat dalam SOMToolbox yaitu fungsi *som_randinit*, *som_batchtrain* dan *som_autolabel*. Adapun langkah-langkah yang dilakukan antara lain: Inisialisasi awal dengan *som_randinit*. *Training* yang dilakukan menggunakan *som_batchtrain*. Proses pelabelan dilakukan dengan menggunakan *som_autolabel*.

2.1.4. Visualisasi

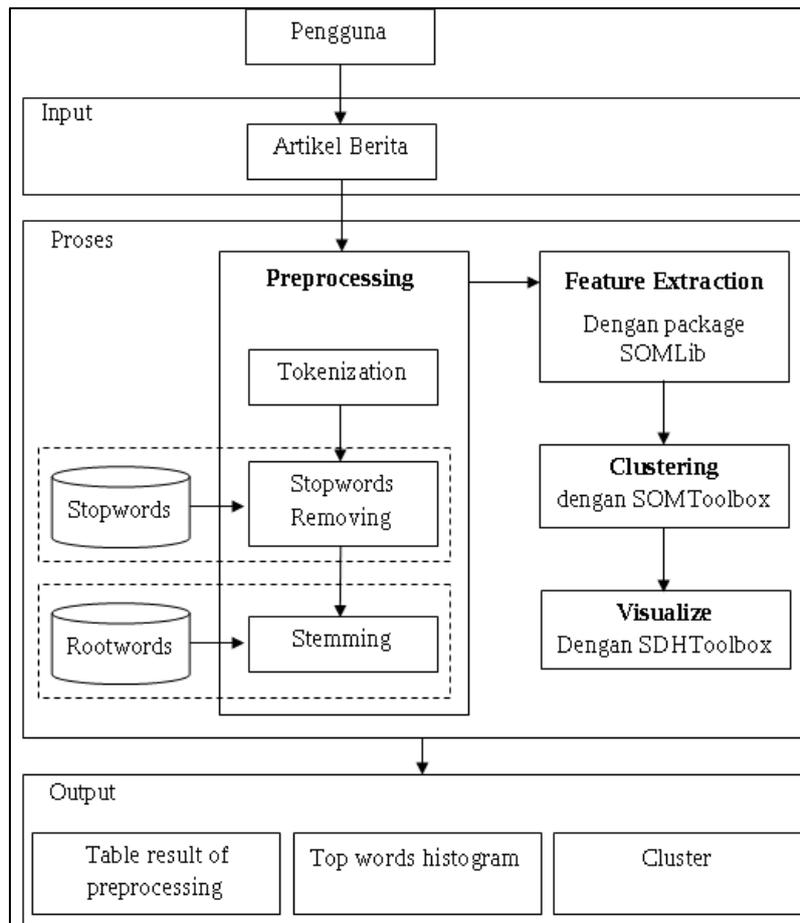
SOM *visualize* menggunakan SDHToolbox untuk memperoleh bentuk visualisasinya. SOM *visualize* merupakan suatu proses untuk menampilkan bentuk *cluster* seperti gugusan pulau-pulau (*islands*). Adapun langkah-langkah yang dilakukan antara lain:

- a. Melakukan perhitungan nilai penyebaran vektor dengan fungsi *sdh_calculate*.
- b. Menampilkan visualisasi hasil *clustering* berupa *island* dari *Smoothed Data Histogram* menggunakan fungsi *sdh_visualize*.

Method yang digunakan untuk menghitung *smoothened data histogram* yaitu *sdh_calculate(sData:SOM, sMap:SOM)*. Setelah selesai dihitung sistem akan menampilkan kluster-kluster ke dalam model *islands* dengan *method sdh_visualize(islands:SDH)*.

2.2. Arsitektur Sistem

Alur proses yang dilakukan oleh seorang pengguna digambarkan dalam bentuk diagram sistem seperti terlihat pada Gambar 1. Sistem yang dibuat dalam penelitian ini terdiri dari *input*, proses dan *output*. *Input* yang dipakai dalam penelitian ini berupa *file* teks yang akan diinputkan oleh *user* pada waktu akan memulai proses. Proses yang terjadi dalam sistem ini meliputi *preprocessing*, *feature extraction*, *clustering* dan *visualisasi*. Hasil yang didapat berupa tabel yang berisi pengurangan kata, kata dasar dari hasil *preprocessing*, *cluster* dari artikel berita dan informasi mengenai jumlah daftar kata terbanyak yang ditampilkan dengan 10 urutan kata. Pada arsitektur sistem ini terdapat 2 kamus data yaitu kamus data yang berisi kata-kata (*stopwords*) yang akan dibuang dan kamus kata dasar (*rootwords*) yang digunakan untuk mencocokkan kata.



Gambar 1 Arsitektur sistem

2.3. Implementasi

Implementasi sistem ini merupakan hasil akhir dari rancangan antarmuka utama untuk pengguna. Antarmuka utama untuk pengguna ini berisi seluruh kegiatan yang dapat dilakukan pengguna untuk membentuk *cluster*. Antarmuka yang telah dibuat ini mempunyai empat buah menu yaitu menu *pages*, menu *table*, menu *data statistics* dan menu *about* seperti terlihat pada implementasi tampilan *process* pada Gambar 3. Antarmuka utama untuk pengguna mempunyai 2 form isian yang pertama untuk *preprocessing* dan *feature extraction*. Form isian kedua untuk *clustering* dan *visualize*.

Preprocessing diimplementasikan dengan membuat kode sendiri dari *tokenization*, *stopwords removing* dan *stemming*. Implementasi *tokenization* intinya dibuat untuk memisahkan deretan kata di dalam kalimat, paragraf atau halaman menjadi *token*. Implementasi *stopwords removing* pada intinya dibuat untuk mencocokkan kata hasil dari proses *tokenization* dengan kata-kata yang terdapat dalam kamus *stopwords* yang telah dibuat, jika terdapat kata yang sama maka kata itu akan dihapus. Implementasi *stemming* dibuat untuk membentuk kata dasar dengan algoritma Nazief dan Adriani.

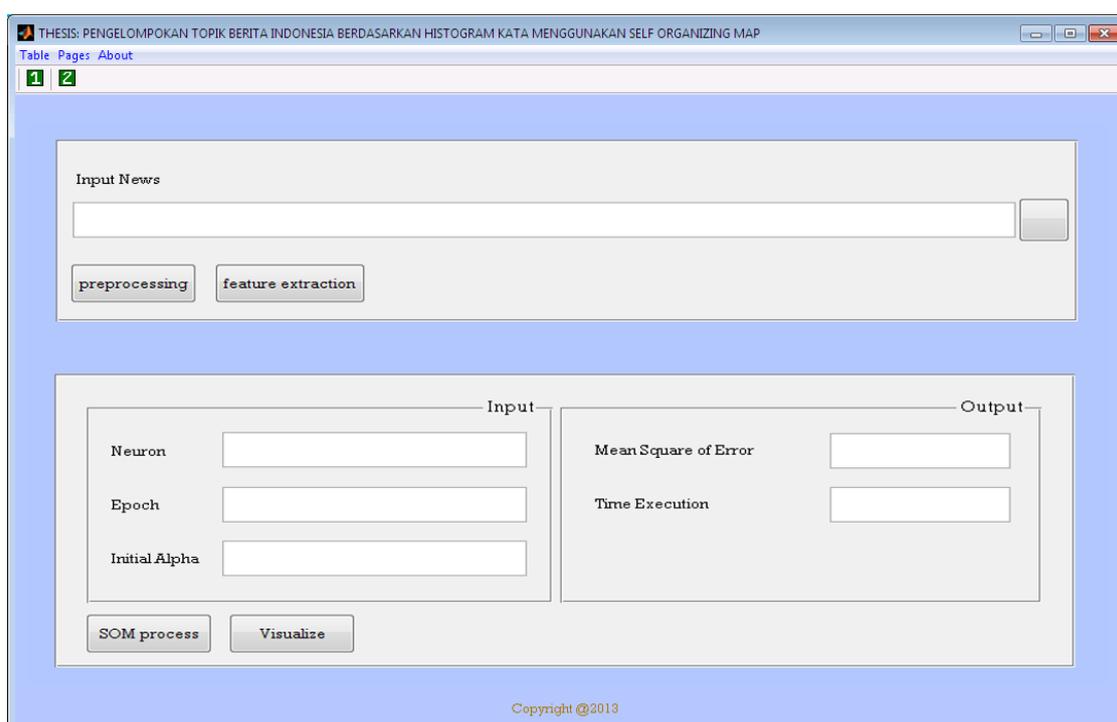
Feature extraction diimplementasikan dengan menggunakan *package* SOMLib. *Class* yang digunakan pada *package* SOMLib dapat melakukan proses ekstraksi fitur pada artikel berita yang tersimpan dalam satu direktori. *Class* yang dipakai pada *package* SOMLib adalah *class wordsexc*, *templatevectorex*, *reducerexc* dan *extractorex*. *Class* tersebut akan dipanggil dengan menggunakan perintah *cmd* pada *matlab*.

Proses *clustering* diimplementasikan dengan memanfaatkan SOMToolbox dengan metode *Self Organizing Map*. Gambar 2 adalah struktur dari data yang dipakai sebagai masukan oleh SOM. Oleh sebab itu data yang akan dikelompokkan harus diubah ke format matrik dengan

fungsi *InputSOM* bertujuan untuk membaca bobot hasil proses *extractorexc*. Pada Gambar 3 menampilkan antarmuka utama untuk pengguna.

ab	type	'som_data'
	data	<570x2357 double>
	labels	<570x1 cell>
ab	name	'News Structure'
	comp_names	<2357x1 cell>
	comp_norm	<2357x1 cell>
	label_names	[]

Gambar 2 Struktur data SOMToolbox



Gambar 3. Antarmuka utama untuk pengguna

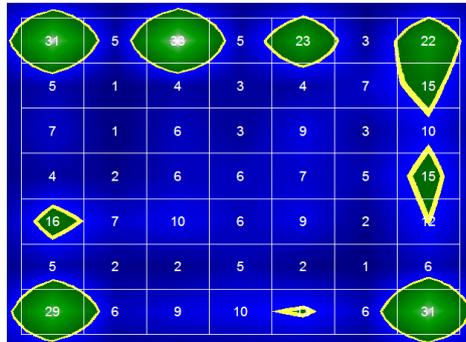
Proses *visualize* diimplementasikan dengan memanfaatkan SDHToolbox untuk mendapatkan model *island* dari *cluster* yang terbentuk. Implementasi *visualize* berfungsi untuk menghasilkan visualisasi hasil klusterisasi dalam bentuk *island map*. Ketika *user* menekan tombol SOM Visualize maka sistem akan menjalankan fungsi *Smooth Data Histogram* (SDH), berikut adalah fungsi dari SDH:

```
sdh_visualize(S, 'labels', sMap.labels, 'sofn', 0, 'type', 'continuous', 'grid', 'on')
```

3. HASIL DAN PEMBAHASAN

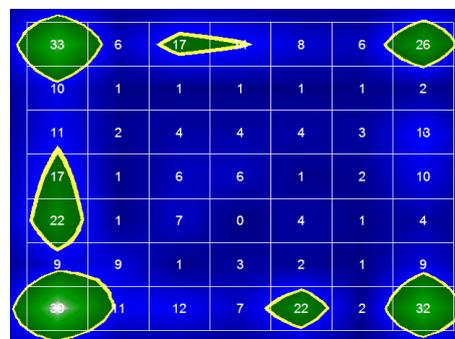
Hasil yang diperoleh dari penelitian ini terbagi atas tahap-tahap proses yang dilakukan sistem. Pembahasan pertama dari hasil yang diperoleh pada *preprocessing* kata. Dari artikel dokumen yang diinputkan sebanyak 570 *file* dimana *file* eko010102.txt mempunyai jumlah kata sebanyak 1267 pada tahap *preprocessing* sistem menghasilkan 1111 kata pada proses *tokenization*, 677 kata pada proses *stopwords removing* dan 677 kata dasar pada proses *stemming*. Hasil perbandingan banyaknya kata bisa dilihat pada perbandingan antar tabel.

Hasil *cluster* terbesar pada Gambar 6 mempunyai 37 *file* yang terletak dalam 2 kotak. Pada kotak pertama didominasi berita yang berisikan kriminal dan pada kotak kedua didominasi berita yang berisikan teknologi, karena kedua kotak ini merupakan satu *cluster* maka kecenderungan berita didalamnya mempunyai banyak kemiripan kata. Cluster terbesar kedua mempunyai 33 *file* dimana isi berita didalamnya adalah berita kesehatan. Ada dua cluster terbesar ketiga dengan jumlah *file* 31 *file* berisi berita kesehatan dan 31 *file* berisi berita ekonomi.



Gambar 6 Hasil *cluster* pada majalah Tempo tahun 2008

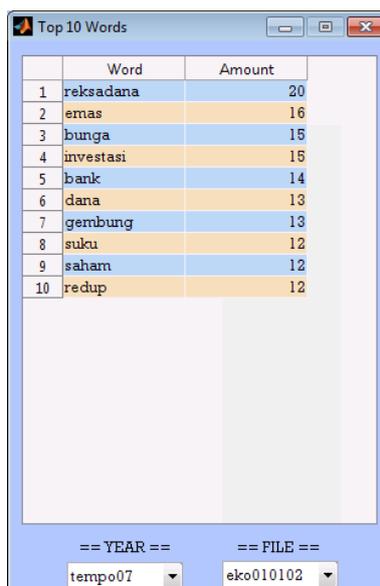
Hasil *cluster* terbesar pada Gambar 7 *file* Tempo tahun 2009 ada 7 *island* yang terbentuk pada proses *clustering* ini, dimana *island* terbesar mempunyai 39 *file* dan *island* terkecil mempunyai 17 *file*. Pada Gambar 7 terlihat ada dua buah *cluster* yang mempunyai jumlah berita yang sama banyak yaitu 39 *file*, dimana kedua *cluster* ini mempunyai kesamaan isi berita yaitu berita mengenai ekonomi. Perbedaan kedua *cluster* ini terletak pada *mountain* yang satu *cluster* untuk mempunyai *mountain* sedangkan *cluster* dari dua kotak tidak terdapat *mountain*. Cluster terbesar kedua berisi berita ekonomi dengan jumlah *file* sebanyak 33 *file*. Sedangkan cluster terbesar ke tiga berjumlah 32 *file* dengan isi berita yang mengulas mengenai kesehatan.



Gambar 7 Hasil *cluster* pada majalah Tempo tahun 2009

Dari tahun 2007 sampai dengan tahun 2009 hasil proses *clustering* memiliki kecenderungan yang berbeda, pada tahun 2007 mempunyai kecenderungan berita ekonomi, sedangkan pada tahun 2008 kecenderungan berita kriminal dan teknologi, yang terakhir uji coba data teks pada tahun 2009 kembali lagi mempunyai kecenderungan berita ekonomi.

Pembahasan keempat diperoleh dari hasil dari proses *wordsex* dapat digunakan sebagai analisa data statistik untuk kemunculan kata terbanyak tiap jenis artikel berita (top 10 words). Gambar 8 adalah salah satu gambar yang memberikan keterangan data statistik dari 10 kata yang muncul dengan nama artikel eko10102 tahun 2007. Dari isi artikel eko10102 kata-kata yang banyak muncul adalah reksadana dengan frekuensi kemunculan sebanyak 20 kali. Tentunya data statistik 10 kata yang sering muncul akan berbeda dalam tiap artikelnya karena kemungkinan isi dari artikel beritanya berbeda. Dari kata-kata yang sering muncul ini dapat dibandingkan untuk mengetahui kemiripan isi dari tiap jenis artikel berita.



	Word	Amount
1	reksadana	20
2	emas	16
3	bunga	15
4	investasi	15
5	bank	14
6	dana	13
7	gembung	13
8	suku	12
9	saham	12
10	redup	12

== YEAR == tempo07 == FILE == eko010102

Gambar 8 Informasi top 10 words yang muncul dari artikel eko10102 tahun 2007

4. KESIMPULAN

1. Dari tahun 2007 sampai dengan tahun 2009 hasil proses *clustering* memiliki kecenderungan yang berbeda, pada tahun 2007 mempunyai kecenderungan berita ekonomi, sedangkan pada tahun 2008 kecenderungan berita kriminal dan teknologi, yang terakhir uji coba data teks pada tahun 2009 kembali lagi mempunyai kecenderungan berita ekonomi.
2. Hasil pulau dari data dengan *Smoothed Data Histogram* (SDH) sangat bergantung pada nilai atribut '*spread*' pada SDH.
3. Sistem dapat menampilkan hasil *clustering* dengan algoritma *Self Organizing Map* dan menampilkan visualisasi dengan *moothed data histograms* berupa *island map* dari artikel berita pada majalah Tempo yang telah diproses.
4. Sistem dapat menampilkan koleksi dokumen dari lima kategori berita yang ada pada tiap tahunnya, selain itu sistem ini mampu menampilkan banyaknya kata yang sering muncul pada tiap artikel berita (*top 10 words*).

5. SARAN

1. Perlu menambahkan idiom atau frase karena dapat memperbaiki performansi *cluster*, karena *tokenisasi* pada sistem ini tidak memperhatikan bentuk idiom atau frasa.
2. *Feature extraction* yang digunakan dalam sistem ini didasarkan pada histogram kata, yang mempunyai banyak kelemahan karena seluruh kategori berita didasarkan pada banyaknya kata yang sama. Jadi bisa dicoba untuk *feature extraction* yang lain.
3. Artikel yang bisa diproses dalam sistem ini hanya berupa *file* tipe *.txt*, untuk itu perlu dikembangkan penelitian yang bisa menghendel lebih banyak tipe *file* seperti, *html*, *doc*, *pdf* dan lain sebagainya.
4. Perlu dibuat *database* untuk pengembangan sistem yang lebih lanjut agar data dapat terdistribusi dengan baik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Instituto Proffisional de Canossa yang telah memberi dukungan finansial terhadap penelitian ini.

DAFTAR PUSTAKA

- [1]. Wibisono Y., dan Khodra, M. L., 2005, Clustering Berita Berbahasa Indonesia, *Jurnal FPMIPA UPI dan KK Informatika ITB*, Bandung.
- [2]. Ramadan, R., 2007, Penerapan Pohon Untuk Klasifikasi Dokumen Teks Berbahasa Inggris, *Seminar Nasional*, Bandung.
- [3]. Langgeni, D. P., Baizal A., dan Firdaus, Y., 2010, Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection, *Seminar Nasional Informatika*, Yogyakarta.
- [4]. Anggraeni, D., 2008, Klasifikasi Topik Menggunakan Metode Naive Bayes dan Maximum Entropi pada Artikel Media Masa dan Abstrak Tulisan, *Skripsi*, Program Ilmu Komputer, Universitas Indonesia, Depok.
- [5]. Kohonen, T., 2000, Self-Organizing Maps of Massive Document Collections, *IEEE-INNS-ENNS International Joint Conference on Neural Networks*, Finland.
- [6]. Vuori, V., 2002 Clustering Writing Styles with a Self-Organizing Map, *IEEE Eighth International Workshop on Frontiers in Handwriting Recognition*, Finland.
- [7]. Perelomov, I., Azcarraga, A.P., Tan, J., dan Chua, T.S., 2002, *Using Structured Self-Organizing Maps in News Integration Websites*, School of Computing, National University of Singapore, Singapore.
- [8]. Bação, F., dan Lobo, V., 2010, *Introduction to Kohonen's Self-Organizing Maps*, Instituto Superior de Estatística E Gestao de Informacao, Universidade Nova de Lisboa, Portugal.
- [9]. Siang, J. J., 2009, *Jaringan Saraf Tiruan dan Pemrogramannya Menggunakan Matlab*, Andi Offset, Yogyakarta.