KLASTERING BERITA ONLINE TENTANG BENCANA DENGAN ALGORITMA SINGLE PASS CLUSTERING

Herny Februariyanti, Eri Zuliarso, Mardi Siswo Utomo

Abstract

Too many type of natural disaster that came and went over to Indonesia. The volume of electronic news on natural disasters in Indonesian language is the greater source valuable information. Clustering text documents is one of the operations on text mining to classify documents that have similar content. Grouping news documents needed to facilitate the search for information about a particular disaster.

Our research focus on a strict on-line setting, in that the system must indicate whether the current document contains or does not contain discussion of a new event before looking at the next document. Our approach to the problem uses a single pass clustering algorithm and linkage between the news is measured by the similarity between documents (similarity).

The algorithm was tested using a sample of news media online. The testing results show that the algorithm can be applied to grouping news in Indonesian language.

Keyword : Information Retrieval, Stemming, Single Pass Clustering, Cosine Similarity.

Abstrak

Terlalu banyak untuk menyebut jenis bencana alam yang datang silih berganti menghampiri Indonesia. Volume berita elektronik berbahasa Indonesia tentang bencana alam yang semakin besar merupakan sumber informasi yang berharga. *Clustering* dokumen teks adalah salah satu operasi pada text mining untuk mengelompokkan dokumen yang memiliki kesamaan isi. Pengelompokan dokumen berita dibutuhkan untuk mempermudah pencarian informasi mengenai suatu *bencana* tertentu.

Penelitian ini menitik beratkan pada berita on-line, sehingga system harus dapat mengindikasikan apakah dokumen saat ini memuat atau tidak memuat topic bencana sebelum melihat pada dokumen berikutnya. Pendekatan yang dilakukan untuk menyelesaikan masalah menggunakan keterkaitan antar berita ini diukur berdasarkan kemiripan antar dokumen (*similarity*).

Algoritma ini diuji coba dengan menggunakan sampel berita dari media online. Hasil uji coba menunjukkan bahwa algoritma ini dapat diaplikasikan untuk pengelompokan berita-berita berbahasa Indonesia.

Kata Kunci : Information Retrieval, Stemming, Single Pass Clustering, Cosine Similarity.

I. PENDAHULUAN

Terlalu banyak untuk menyebut jenis bencana alam yang datang silih berganti menghampiri Indonesia. Kondisi ini membuat pemerintah harus berpikir ekstra keras untuk mengatasi dampak buruk bencana alam tersebut.

Dalam konteks ini, teknologi informasi (TI) hadir memainkan peran yang cukup penting. Sebagaimana manusia, eksistensi TI tidak untuk menghalau suatu bencana alam yang datang secara tiba-tiba melainkan untuk menyampaikan informasi sebelum dan sesudah bencana alam itu terjadi. Mengantisipasi bencana dalam waktu singkat dapat dilakukan dengan menerapkan sistem peringatan dini. Sistem itu bekerja dengan memanfaatkan basis data dari berbagai situs di Internet. Yang diperlukan dalam suatu Sistem yang cerdas (*Intelligent System*) adalah secara otomatis sistem ini dapat :

1. Mendeteksi kejadian penting dari sekian banyak berita-berita baru
2. Memasukkan berita-berita kejadian baru pada kelompok berita yang sudah ada.
3. Membuat kelompok berita baru apabila suatu berita tidak memenuhi kelompok berita yang sudah ada
4. Menampilkan kejadian yang diperlukan oleh user dalam bentuk berita-berita terkait yang ada hubungannya dengan keinginan user

Inilah yang menjadi tujuan utama dari penelitian yang *Topic Detection and Tracking (TDT).* Topik (*Topic*) disini adalah perubahan kejadian secara dinamis. Dalam metode ini kita menggunakan beberapa cara untuk pencarian informasi (*Information Retrieval*) dan teknik *machine learning* untuk keefektifan pendeteksian (detection) dan Klastering (Allan et al ,1998). Pendeteksian berita dalam *Topic Detection and Tracking* (TDT) merupakan suatu cara untuk mendapatkan suatu solusi dalam pencarian informasi (*Information Retrieval*). Dari permasalahan yang mungkin muncul inilah maka dikembangkan suatu metode untuk menangani masalah-masalah di atas, dalam hal ini contoh kasus yang diangkat adalah pendeteksian dan pengelompokan suatu kejadian yang merupakan

bagian dari *Event Detecting* dalam TDT. Pada dasarnya TDT dapat dibagi menjadi 2 tahap (Arifin, AZ., Setiono, AN,2002) :

1. *Penyiapan Dokumen*

   Pada tahap ini akan dilakukan manipulasi teks pada dokumen yang selanjutnya tiap dokumen akan direpresentasikan dalam bobot tertentu

2. *Klastering dokumen*

   Untuk klastering dokumen diperlukan nilai batas (*Treshold value*).Untuk mendapatkan nilai batas (*Treshold value)* diperlukan suatu data training (*restrospective document*).

   Permasalahan yang timbul dari klastering dokumen berita adalah :

   a. Perlunya algoritma klastering dokumen.
   b. Penentuan tingkat kemiripan (*similarity)* antar dokumen berdasarkan komposisi term.
   c. Pilihan *term* dari kata-kata dalam Bahasa Indonesia yang relevan sebagai pembeda.

II. METODOLOGI

1. Pengumpulan data sampel

Dokumen yang digunakan sebagai sampel adalah berita-berita yang diambil dari Surat Kabar *Online* Kompas. Jumlah sampel adalah 400 dokumen berita yang diklasteringkan secara manual menjadi 10 *event*. Klastering manual ini nantinya digunakan untuk mengevaluasi tingkat keberhasilan Klastering.

2. *Ekstraksi dokumen*

   Proses ekstraksi ini bertujuan untuk menghasilkan *term-term* yang akan digunakan sebagai *prototype* bagi setiap dokumen.

3. *Penghitungan Bobot TF-IDF*

   Pada tahap ini, tiap dokumen diwujudkan sebagai sebuah *vector* dengan elemen sebanyak *term* yang berhasil dikenali dari tahap ekstrasi dokumen di atas. Vektor tersebut beranggotakan bobot dari tiap term yang dihitung berdasarkan metode TF-IDF. Metode TF-IDF ini merupakan metode pembobotan dalam bentuk sebuah metode yang merupakan integrasi antar

term frequency (tf), dan *inverse document frequency (idf)* (Allan, et all, 1998)(Utami E., dkk., 2008), adapun rumusnya adalah :

w(t,d) = tf(t,d)*log2(N/nt)

Simbol w(t,d) adalah bobot dari *term* t dalam dokumen d sedangkan tf(t,d) adalah frekuensi *term* dalam dokumen(tf) dimana N merupakan ukuran data training yang digunakan untuk penghitungan IDF. Adapun *nt* adalah jumlah dari dokumen yang detraining yang mengandung nilai t. Fungsi metode ini adalah untuk mencari representasi nilai dari tiap-tiap dokumen dari suatu kumpulan data training (*training set*). Dari sini akan dibentuk suatu vektor antara dokumen dengan kata (*documents with terms*) yang kemudian untuk kesamaan antar dokumen dengan cluster akan ditentukan oleh sebuah *prototype* vektor yang disebut juga dengan *cluster centroid*.

4. *Penghitungan tingkat kemiripan*

   Perbandingan kemiripan (*similarity*) yang digunakan disini adalah *standard cosine similarity* dengan rumus :

$$S_{D_i D_j} = \frac{\sum_{k-1}^{L} (weight_{ik} weight_{jk})}{\sqrt{\sum_{k-1}^{L} weight_{ik}^2} \sum_{k-1}^{L} weight_{jk}^2}$$

   S_{DiDj} = Similarity Dokumen ke I dan Ke J

5. *Klastering*

   Similarity yang telah dihasilkan selanjutnya dievaluasi untuk menentukan pasangan-pasangan dokumen yang dinyatakan mirip berdasarkan nilai *treshold* tertentu. Pengklasteringan dokumen berita dengan menggunakan Algoritma *Single Pass Clustering*. Tujuan klastering dokumen adalah untuk memisahkan dokumen yang relevan dari dokumen yang tidak relevan (Zhang J., et all, 2001). Atau dengan kata lain, dokumen-dokumen yang relevan

dengan suatu *query* cenderung memiliki kemiripan satu sama lain dari pada dokumen yang tidak relevan, sehingga dapat dikelompokkan ke dalam suatu klaster.

6. *Evaluasi Klastering*

   Evaluasi ini dilakukan untuk mengetahui kinerja algoritma Klastering pada tahap uji coba. Pengukuran ini didasarkan pada dua parameter, yakni *recall* dan *precision.*

III. CRAWLER

Web Crawler, juga sering dikenal sebagai Web Spider atau Web Robot adalah salah satu komponen penting dalam sebuah mesin pencari modern. Fungsi utama Web Crawler adalah untuk melakukan penjelajahan dan pengambilan halaman-halaman Web yang ada di Internet. Hasil pengumpulan situs Web selanjutnya akan diindeks oleh mesin pencari sehingga mempermudah pencarian informasi di Internet.

Mendesain sebuah crawler yang baik saat ini menemui banyak tantangan. Secara eksternal, crawler harus mengatasi besarnya situs Web dan link jaringan. Secara internal , crawler harus mengatasi besarnya volume data. Sehubungan dengan terbatasnya sumber daya komputasi dan keterbatasan waktu, maka harus hati-hati memutuskan URL apa yang harus di scan dan bagaimana urutannya. Crawler tidak dapat mengunduh semua halaman web. Penting bagi crawler untuk memilih halaman dan mengunjungi halaman yang penting dulu dengan memprioritaskan URL yang penting tersebut dalam antrian. Crawler juga harus memutuskan berapa frekuensi untuk merevisi halaman yang pernah dilihat, untuk memberikan informasi ke client perubahan yang terjadi di Web. Zuliarso E dan Mustofa, K., 2009a, telah menguji algoritma kunjungan crawler berdasarkan isi halaman web. Dalam Zuliarso E., Mustofa, K., 2009b telah menguji algoritma penelusuran berdasarkan breadth first search, banyaknya backlink, dan ontologi.

IV. PENGINDEKSAN

Inverted file atau index inverted adalah mekanisme untuk pengindeksan kata dari koleksi teks yang digunakan untuk mempercepat proses pencarian. Struktur inverted file terdiri dari dua elemen, yaitu: kata (*vocabulary*) dan kemunculan (*occurences*). Kata-kata tersebut adalah himpunan dari kata-kata yang ada pada teks, atau merupakan ekstraksi dari kumpulan teks yang ada. Februariyanti H., 2010 melakukan penelitian menggunakan algoritma indeks inverted untuk proses indeks kata (*term*), cosine similaritas untuk menghitung kesamaan kata dalam dokumen.

V. STEMMING BAHASA INDONESIA

Proses *stemming* adalah proses pembentukan kata dasar. Term yang diperoleh dari tahap pembuangan stop word akan dilakukan proses stemming. Algoritma stemming yang digunakan adalah modifikasi Porter stemmer dari (Tala FZ., 2003). Stemming digunakan untuk mereduksi bentuk term untuk menghindari ketidak cocokan yang dapat mengurangi recall, di mana term-term yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk.

VI. KLASTERING DOKUMEN

Klastering biasa digunakan pada banyak bidang, seperti : data mining, pattern recognition (pengenalan pola), image classification (pengklasifikasian gambar), ilmu biologi, pemasaran, perencanaan kota, pencarian dokumen, dan lain sebagainya.

Tujuan dari klastering adalah untuk menentukan pengelompokan dari suatu set data. Akan tetapi tidak ada ”ukuran terbaik” untuk pengelompokan data. Untuk pengelompokkan data tergantung tujuan akhir dari klastering, maka diperlukan suatu kriteria sehingga hasil klastering seperti yang diinginkan.

Penelitian tentang clustering document (klastering dokumen) telah banyak dilakukan. Secara umum klastering dokumen adalah proses mengelompokkan

dokumen berdasarkan kemiripan antara satu dengan yang lain dalam satu klaster (Februariyanti, H., Winarko, E., 2010)

Tujuan klastering dokumen adalah untuk memisahkan dokumen yang relevan dari dokumen yang tidak relevan (Zhang J., et all., 2001). Atau dengan kata lain, dokumen-dokumen yang relevan dengan suatu query cenderung memiliki kemiripan satu sama lain dari pada dokumen yang tidak relevan, sehingga dapat dikelompokkan ke dalam suatu klaster.

Klastering dokumen dapat dilakukan sebelum atau sesudah proses temu kembali (Zhang J., et all., 2001). Pada klastering dokumen yang dilakukan sebelum proses temu kembali informasi, koleksi dokumen dikelompokkan ke dalam klaster berdasarkan kemiripan (similarity) antar dokumen. Selanjutnya dalam proses temu kembali informasi, apabila suatu dokumen ditemukan maka seluruh dokumen yang berada dalam klaster yang sama dengan dokumen tersebut juga dapat ditemukan.

Pada algoritma klastering, dokumen akan dikelompokkan menjadi *klaster-klaster* berdasarkan kemiripan satu data dengan yang lain. Prinsip dari *klastering* adalah memaksimalkan kesamaan antar anggota satu klaster dan meminimumkan kesamaan antar anggota *klaster* yang berbeda.

VII. ALGORITMA SINGLE PASS CLUSTERING

Single Pass Clustering merupakan suatu tipe clustering yang berusaha melakukan pengelompokan data satu demi satu dan pembentukan kelompok dilakukan seiring dengan pengevaluasian setiap data yang dimasukkan ke dalam proses cluster. Pengevaluasian tingkat kesamaan antar data dan cluster dilakukan dengan berbagai macam cara termasuk menggunakan fungsi jarak, vectors similarity, dan lain-lain.

Algoritma Single Pass Clustering dapat dilakukan dengan langkah-langkah sebagai berikut (Frakes WB and Baeza Yates, R., 1992):

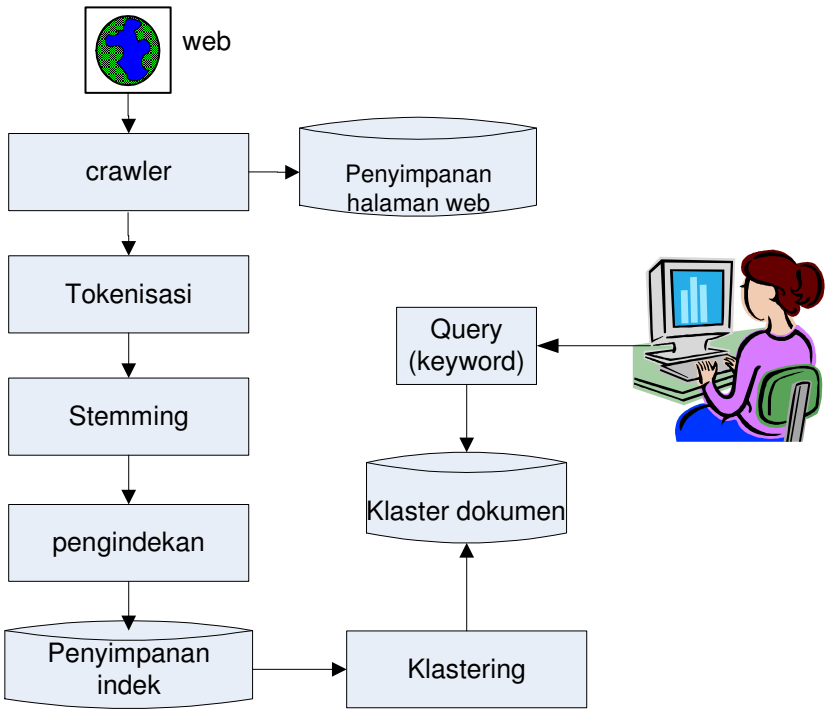
1. Masukkan (dokumen pertama) D1 representasi (*Cluster* pertama) C1
2. Untuk (dokumen ke-i) Di hitung kesamaan (*similarity*) dengan setiap wakil dari masing-masing *cluster*.

3. Jika (*Maximum Similarity*) Smax lebih besar dari batas nilai (*threshold value*) ST, tambahkan tambahkan item kepada cluster yang bersesuaian dan hitung kembali representasi *cluster*, sebaliknya gunakan Di untuk inisialisasi *cluster* baru.
4. Jika masih ada sebuah item Di yang belum dikelompokkan, kembali ke langkah ke-2 Algoritma ini digunakan untuk restrospective data detection maupun on-line detection.

VIII. HASIL DAN PEMBAHASAN

Arsitektur Sistem

Pada Gambar 1 diperlihatkan arsitektur sistem klastering dokumen berita secara online. Aplikasi dimulai dari mendapatkan dokumen berita bencana secara otomatis dengan menggunakan modul crawler. Dilanjutkan dengan proses preprosesing yaitu proses pembersihan dokumen yaitu dengan proses tokenisasi dan stemming dokumen. Setelah dokumen dilakukan proses preprosesing dilanjutkan dengan proses hitung similaritas dan klastering dokumen .



Gambar 1 Arsitektus Sistem Klastering Dokumen Berita

Modul Crawler

Proses Crawler adalah untuk melakukan penjelajahan dan pengambilan halaman-halaman Web yang ada di Internet. Hasil pengumpulan situs Web selanjutnya akan diindeks oleh mesin pencari sehingga mempermudah pencarian informasi di Internet. Crawler digunakan untuk mengambil konten pada situs target secara berkala. Proses dilakukan terus-menerus dan dilakukan secara otomatis. Proses indeks akan disimpan dalam table seperti terlihat pada Gambar 2

alamat	html	isi	id
http://worldcup.kompas.com/welcome/chgcomment/3081...	1.htm		195
http://www.kompas.com/read/xml/2008/05/24/06434832...	gempabumiguncangsumlakimaluku.htm	JAKARTA, SABTU - Badan Meteorologi dan Geofisika m...	196
http://kelaskaryawansabtuminggu.com/search/%2Btopi...	%2Btopik%2Btim%2Bf1%2Bhormati%2Bkorban%2Bgempa%2Bb...		197
http://internasional.kompas.com/read/2012/01/24/07...	GempaBesarAkanMelandaTokyo.htm	TOKYO, KOMPAS.com — Para peneliti Jepang mempering...	198
http://sains.kompas.com/read/2009/09/04/08214439/M...	MencariCaraMemprediksiGempaBumi.htm	Yuni IkawatiKOMPAS.com — Apakah gempa bumi dapat d...	199
http://health.kompas.com/read/2011/10/26/11414996/...	GempaKagetkanWargaMentawai.htm	SIKAKAP, KOMPAS.com — Sebagian warga di Kecamatan ...	200
http://internasional.kompas.com/read/2011/03/15/09...	FOTOTSUNAMIInilahNerakadiBumi.htm	KOMPAS.com — "Hal yang paling nyata terasa adalah ...	201
http://edukasi.kompas.com/read/2011/10/10/19153993...	PameranPendidikanJepang.htm	JAKARTA, KOMPAS.com – Sebanyak 17 institusi pendi...	202
http://foto.detik.com/readfoto/2011/04/10/155239/1...	simulasi-penanganan-gempa-bumi.htm	International Organization for Migration (IOM) dan...	203
http://news.detik.com/read/2012/02/15/155950	latihan-gempa-bumi-digelar-di-new-delhi.htm	You are redirected to FacebookYou are	204

Gambar 2 Gambar Tabel Hasil Crawler

Modul Stemming

Proses *stemming* adalah proses pembentukan kata dasar. Term yang diperoleh dari tahap pembuangan stop word akan dilakukan proses stemming. Algoritma stemming yang digunakan adalah modifikasi Porter stemmer dari (Tala FZ., 2003) Stemming digunakan untuk mereduksi bentuk term untuk menghindari ketidak cocokan yang dapat mengurangi recall, di mana term-term yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk. Setelah dihasilkan kata dasar dilanjutkan dengan proses hitung similaritas antar dokumen. Masing-masing dokumen akan dihitung cacah term yang sama antara dokumen yang satu dengan dokumen yang lain. Hasil dari hitung cacah akan dihasilkan dokumen dengan nilai similaritas dokumen. Nilai similaritas dokumen yang

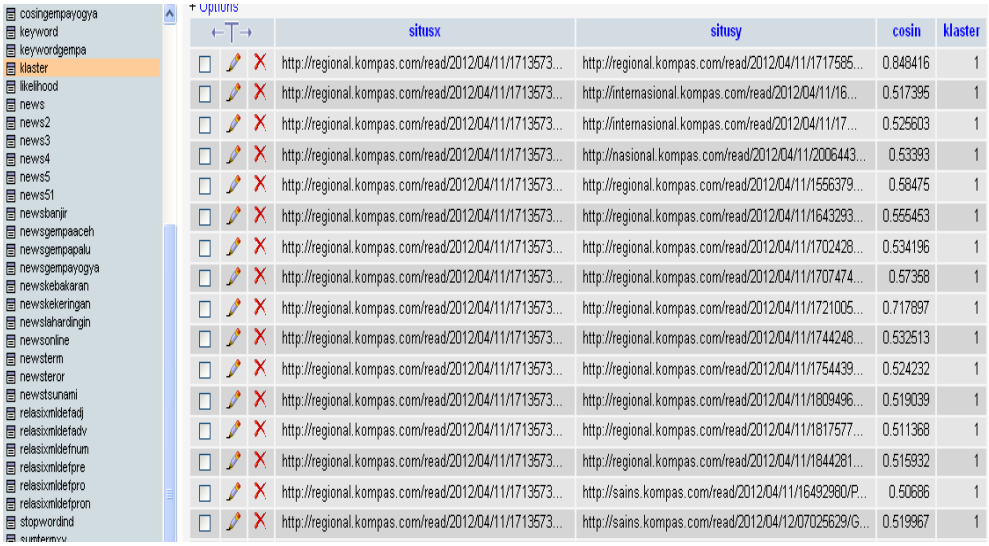
tertinggi dapat dianggap bahwa dokumen tersebut paling simmilar, yaitu memiliki banyak kesamaan. Hasil proses perhitungan similaritas dokumen akan diindeks dan disimpan dalam tabel cosin similaritas yang dapat dilihat pada Gambar 3

	situsx	situsy	cosin	status
	http://edukasi.kompas.com/read/2012/04/11/18291331...	http://edukasi.kompas.com/read/2012/04/11/19260470...	0.716834	true
	http://edukasi.kompas.com/read/2012/04/11/19260470...	http://edukasi.kompas.com/read/2012/04/11/18291331...	0.716834	true
	http://edukasi.kompas.com/read/2012/04/12/10595087...	http://edukasi.kompas.com/read/2012/04/11/18291331...	0.662677	true
	http://edukasi.kompas.com/read/2012/04/12/10595087...	http://edukasi.kompas.com/read/2012/04/11/19260470...	0.637109	true
	http://entertainment.kompas.com/read/2012/04/11/17...	http://edukasi.kompas.com/read/2012/04/11/18291331...	0.476882	false
	http://entertainment.kompas.com/read/2012/04/11/17...	http://edukasi.kompas.com/read/2012/04/11/19260470...	0.377661	false
	http://entertainment.kompas.com/read/2012/04/11/18...	http://edukasi.kompas.com/read/2012/04/11/18291331...	0.293291	false
	http://entertainment.kompas.com/read/2012/04/11/18...	http://edukasi.kompas.com/read/2012/04/11/19260470...	0.273724	false
	http://entertainment.kompas.com/read/2012/04/11/19...	http://edukasi.kompas.com/read/2012/04/11/18291331...	0.216873	false
	http://entertainment.kompas.com/read/2012/04/11/19...	http://edukasi.kompas.com/read/2012/04/11/19260470...	0.226637	false
	http://entertainment.kompas.com/read/2012/04/11/19...	http://edukasi.kompas.com/read/2012/04/11/18291331...	0.309361	false
	http://entertainment.kompas.com/read/2012/04/11/19...	http://edukasi.kompas.com/read/2012/04/11/19260470...	0.245699	false

Gambar 3 Gambar Tabel Hasil Proses Cosine Coefficient

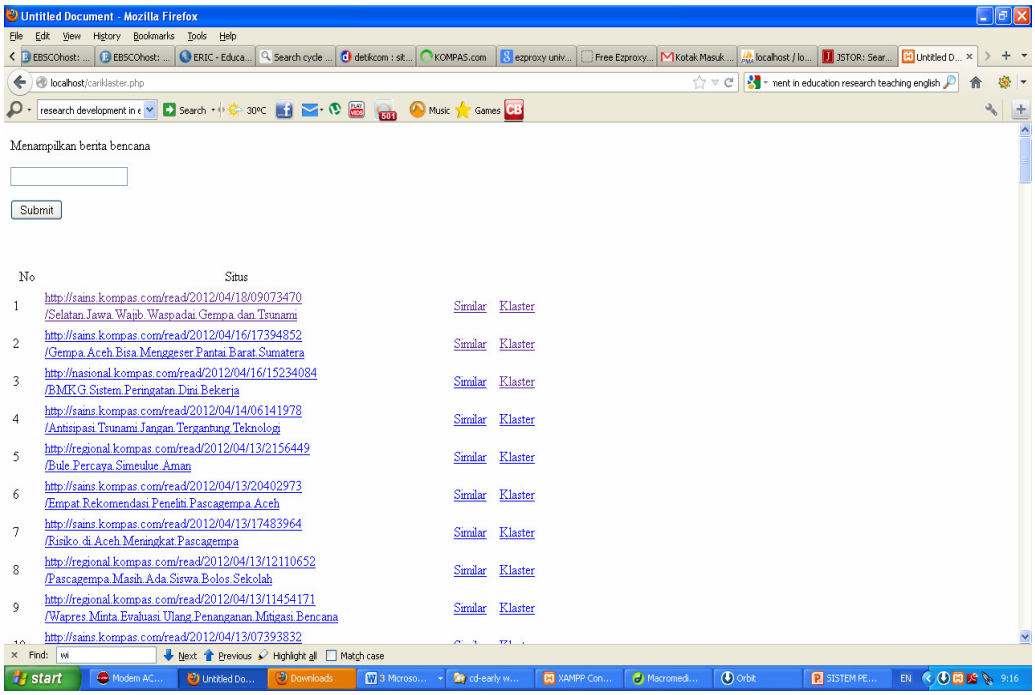
Modul Klastering

Proses klastering pada penelitian ini digunakan algoritma Clustering Single Pass. Proses klastering akan dilakukan dari hasil output proses hitung similaritas, yaitu nilai similaritas antar dokumen. Proses pertama adalah mencari similaritas tertinggi (maksimal). Dokumen dengan similaritas tertinggi akan menjadi klaster *C1*. Selanjutnya dicari dokumen yang memiliki similaritas diatas threshold. Threshold yang dipakai dalam penelitian ini adalah lebih besar dari 0.5. Hasil proses klastering disimpan dalam tabel klaster dapat dilihat pada Gambar 4.



Gambar 4. Gambar Tabel Hasil Proses Klastering

Untuk kemudahan user dibuat user interface dimana user dapat memasukkan kata kunci tertentu dan sistem akan menampilkan klaster dokumen berita sesuai dengan kata kunci yang diinputkan user. Tampilan interface sistem dapat dilihat pada Gambar 5



Gambar 5. User Interface Proses Klastering

IX. KESIMPULAN

Dari hasil penelitian yang telah dilakukan dapat disimpulkan hal-hal sebagai berikut:

1. Pembobotan term frekuensi dan cosine similaritas digunakan untuk menunjukkan kemiripan antar dokumen.
2. Sistem dapat menampilkan dokumen yang mempunyai kedekatan similaritas dari query yang diinputkan user.
3. Dokumen yang membahas topik yang sama cenderung untuk mengelompok menjadi satu klaster.
4. *Single Pass clustering* cukup handal digunakan sebagai algoritma untuk klasifikasi *event*
5. Klaster dapat membantu menemukan dokumen yang ada dalam satu klaster dengan query yang diinputkan user.
6. Klaster dapat membantu mendapatkan dokumen yang relevan.

X. DAFTAR PUSTAKA

[1.] Allan et al ,1998, *Topic Detection and Tracking Pilot Study : Final Report.*, Proc. DARPA Broadcast News Transcription & Understanding Workshop, Morgan Kaufman, San Francisco, pp194-218

[2.] Arifin, AZ., Setiono, AN,2002, *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*, SITIA, Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh Nopember

[3.] Februariyanti, H, 2010, *Prototipe Mesin Pencari Dokumen Teks,* Penelitian Universitas Stikubank,

[4.] Februariyanti, H., Winarko, E., 2010, *Klastering Dokumen Menggunakan Hierarchical Agglomerative Clustering*, Seminar Nasional Teknologi Informasi, STIKOM, Surabaya.

[5.] Frakes WB. And Baeza Yates, R., 1992, Information Retrieval Data Structures And Algorithm, Prentice-Hall International Edition, 1992.

[6.] Tala FZ., 2003, A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia Institut for Logic, Language and Computation Universiteit van Amsterdam The Netherlands.

[7.] Utami, E., Cahyanto, AD., 2008, *Sistem Peringatan Dini Pada Bencana Banjir Berbasis Sms Gateway Di Gnu/Linux Merupakan Alternatif Yang Sederhana Dan Menarik Dalam Meningkatan Pelayanan Badan Meteorologi Dan Geofisika Dengan Alokasi Dana Yang Rendah*, Seminar Nasional Aplikasi Teknologi Informasi 2008 (SNATI 2008) , Yogyakarta.

[8.] Zhang J., Jianfeng G., Ming Z., Jiaxing W., 2001, Improving the Effectiveness of Information Retrieval with Clustering and Fusion, Computational Linguistics and Chinese Language Processing, vol. 6, No.1.

[9.] Zuliarso,E., Mustofa,K., 2009a, *Crawling Web Berbasis Konten*, Dinamik, Jurnal Teknologi Informasi, Universitas Stikubank Semarang, Vol XIV, Juli 2009

[10.] Zuliarso,E.,Mustofa,K., 2009b, *Crawling Web Berdasarkan Ontologi*, Seminar Nasional V, Jurusan Matematika, FMIPA Universitas Negeri Semarang,Ontober 2009