Klastering Dokumen Berita dari Web menggunakan Algoritma Single Pass Clustering

Herny Februariyanti dan Eri Zuliarso
Fakultas Teknologi Informasi, Universitas Stikubank
Email: herny@unisbank.ac.id, eri299@gmail.com

Abstrak

Dalam proses penelusuran informasi melalui internet sering diperoleh informasi yang sangat banyak, tetapi sebagian besar diantaranya adalah informasi yang tidak dibutuhkan. Dari sudut pandang temu kembali informasi (information retrieval), semakin banyaknya informasi yang tersedia di internet justru semakin mempersulit untuk menemukan kembali informasi yang relevan, yaitu informasi yang sesuai dengan kebutuhan.

Dalam suatu sistem temu kembali informasi, kemampuan untuk menemukan informasi yang tersedia diukur dengan recall dan kemampuan untuk menemukan informasi yang relevan diukur dengan ketelitian, maka proses penelusuran dalam situasi seperti tersebut di atas akan menghasilkan recall yang tinggi tetapi ketelitian rendah.

Penelitian ini berusaha untuk mengklaster dokumen dengan menggunakan Algoritma Single Pass Clustering. Klastering ini ditekankan untuk dokumen berbahasa Indonesia. Keterkaitan antar dokumen diukur berdasarkan kemiripan antar dokumen (similarity).

Algoritma ini diuji coba dengan menggunakan dokumen berita arsip berita online Kompas http://www.kompas.com/archive dalam format HTML Hasil uji coba menunjukkan bahwa algoritma ini dapat diaplikasikan untuk pengelompokan dokumen berbahasa Indonesia. Pemilihan kata kunci yang tepat akan meningkatkan kualitas temu kembali informasi (information retrieval) pada dokumen.

Keywords: information retrieval, simmilaritas, singgle pass clustering, recall, precision

PENDAHULUAN

Kemajuan yang pesat di bidang teknologi informasi terutama internet, telah menimbulkan lonjakan informasi yang hebat. Hal ini terjadi karena internet memungkinkan banyak orang untuk memproduksi, memanipulasi, mengakses dan menyebarluaskan informasi dengan "mudah".

Salah satu cara untuk memperoleh informasi yang seimbang seperti apa yang diinginkan adalah dengan membaca beberapa dokumen yang membahas topik yang sama. Akan tetapi cara ini menyulitkan pembaca untuk menangkap topik bahasan utama dari dokumen–dokumen tersebut karena harus mengingat–ingat isi dokumen yang telah dibaca sebelumnya.

Dalam proses penelusuran informasi melalui internet sering diperoleh informasi yang sangat banyak, tetapi sebagian besar diantaranya adalah informasi yang tidak dibutuhkan. Oleh karena itu, dari sudut pandang temu kembali informasi (*information retrieval*), semakin banyaknya informasi yang tersedia di internet justru semakin mempersulit untuk menemukan kembali informasi yang relevan, yaitu informasi yang sesuai dengan kebutuhan. Dalam suatu sistem temu kembali informasi, kemampuan untuk menemukan informasi yang tersedia diukur dengan *recall* dan kemampuan untuk menemukan informasi yang relevan diukur

dengan ketelitian, maka proses penelusuran dalam situasi seperti tersebut di atas akan menghasilkan *recall* yang tinggi tetapi ketelitian rendah.

Sistem yang tepat untuk masalah tersebut adalah Sistem Temu Kembali Informasi yang dapat menghasilkan integrasi dari beberapa dokumen elektronik yang berbeda dengan topik bahasan yang sama secara otomatis. Proses integrasi akan menghasilkan dokumen baru yang mengandung semua bagian dari dokumen–dokumen awal, namun memiliki susunan antar kalimat serta antar paragraf yang berbeda. Perbedaan ini karena saat proses integrasi topik–topik bahasan yang serupa (*similar*) dari semua dokumen dikumpulkan menjadi satu paragraf dan disusun ulang kalimat per kalimat sesuai dengan besarnya kesamaan (*similiarity*) antar kata (*term*). Dengan membaca hasil integrasi diharapkan pembaca dapat terbantu dalam menyerap informasi penting yang ada dalam kumpulan dokumen yang berbeda dan tidak perlu lagi membaca sekumpulan dokumen satu per satu.

METODE PENELITIAN

Metode yang akan digunakan dalam penelitian ini terdiri dari langkah-langkah sebagai berikut:

1. Obyek Penelitian

   Obyek penelitian dari penelitian ini adalah dokumen teks berupa halaman web http://www.kompas.com.

2.. Data yang diperlukan

   Merupakan data yang mendukung dalam penelitian ini meliputi data primer dan data sekunder.

   a. Data primer

      Data yang diperoleh dari arsip berita online Kompas http://www.kompas.com/archive dalam format HTML oleh penulis disimpan dalam format teks.

   b. Data Sekunder

      Data yang diperoleh dengan membaca dan mempelajari referensi mengenai *stemming, text mining, clustering, indexing, term weighting, simmilarity, query expand*.

3. Teknik Pengumpulan Data

   Pengumpulan data mempunyai tujuan mendapatkan materi – materi yang mempunyai keterkaitan dengan topik penelitian. Pengumpulan data dimaksudkan agar mendapatkan bahan-bahan yang relevan, akurat dan reliable. Maka teknik pengumpulan data yang dilakukan dalam penelitian ini adalah dengan metode Observasi, Studi Pustaka dan Metode pengembagan dengan menggunakan model prototyping.

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 (Gordon, 1991; Ellis, 1996).

Tujuan klastering dokumen adalah untuk memisahkan dokumen yang relevan dari dokumen yang tidak relevan (Jian Zhang, dkk., 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 (Jian Zhang, dkk., 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.

Dalam Sistem Temu Kembali Informasi, klastering dokumen memberikan beberapa manfaat, antara lain:

1. Mempercepat pemrosesan *query* dengan menelusur hanya pada sejumlah kecil anggota atau wakil klaster, sehingga dapat mempercepat proses temu kembali informasi
2. Membantu melokalisasi dokumen yang relevan
3. Membentuk kelas-kelas dokumen sehingga mempermudah penjelajahan dan pemberian interpretasi terhadap hasil penelusuran
4. Meningkatkan efektivitas dan efisiensi temu kembali informasi dan memberikan alternatif metode penelusuran

Selain itu, penggabungan antara penelusuran secara menyeluruh (*full search*) dengan penelusuran berbasis klaster (*cluster-based retrieval*) dapat meningkatkan ketelitian sampai dengan 25%. Hal yang sama dikemukakan oleh (Jian Zhang, dkk., 2001) bahwa penggabungan antara metode pengklasteran dengan *fusion* (pemberian peringkat terdadap dokumen secara keseluruhan) akan meningkatkan efektivitas temu kembali informasi.

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

Metode Clustering Single Pass

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 yang sering digunakan dalam Single Pass Clustering adalah sebagai berikut:

1. for each data d loop
   a. find a cluster c that maximizes an objective function
   b. if the value of the objective function > a threshold value then include d in c
   c. else create a new cluster whose only data is d

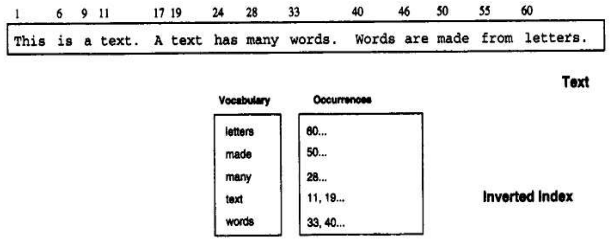
2. end loop

Dalam menggunakan algoritma single pass, dua hal yang perlu menjadi perhatian adalah penentuan objective function dan penentuan threshold value. Objective function yang ditentukan haruslah sebisa mungkin mencerminkan keadaan data yang dimodel dan dapat memberikan nilai tingkat kesamaan atau perbedaan yang terkandung di dalam data tersebut. Penentuan threshold value juga merupakan hal yang subjektif, makin besar nilai threshold, makin mudah suatu data untuk bergabung ke dalam suatu cluster, dan demikian juga sebaliknya.

INDEX INVERTED

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.

Dan tiap kata terdapat juga informasi mengenai semua posisi kemunculannya (*occurences*) secara rinci. Posisi dapat merefer kepada posisi kata ataupun karakter. Hal ini dapat dilihat dengan jelas dengan memperhatikan Gambar 1.



Gambar 1 Contoh Teks dan I*nverted File*-nya

STEMMING

Penelitian pencarian tentang efek *stemming* bahasa Indonesia dalam proses temu kembali dilakukan oleh Tala (2003). Algoritma *stemmer* untuk bahasa Indonesia yang dikembangkan adalah algoritma *purely ruled-based stemmer*. Algoritma ini adalah mengadopsi dari algoritma *English Porter Stemmer* yang dikembangkan oleh Frakes (1992). Dipilihnya algoritma Porter untuk dikembangkan sebagai algoritma stemmer untuk bahasa Indonesia karena pemikiran dasar dari algoritma stemmer Porter cocok dengan struktur morfologi kata-kata di dalam bahasa Indonesia. Perbedaan algoritma ini dengan algoritma yang telah dikembangkan oleh Nazief dan Adriani yaitu tidak adanya *dictionary* sehingga algoritma dapat dikatakan murni berbasis *rule* (*purely rule-based stemmer*).

Morphologi bahasa Indonesia dapat terdiri dari turunan dan imbuhan kata. Imbuhan yang sederhana digunakan akhiran dimana tidak akan merubah makna dari kata dasar.

Dalam proses *stemming* bahasa Indonesia ini terdapat beberapa tahap. Sebuah kata akan dites dengan menggunakan *rule* yang dibuat pada setiap tahap. Pada setiap tahap, sebuah kata yang memenuhi kondisi untuk *rule* pada tahap itu maka kata tersebut akan diganti dengan kata baru yang dibentuk dengan *substitution rule* (aturan pengganti).

COSIN SIMMILARITY DISTANCE

Metode cosine distance merupakan metode yang digunakan untuk menghitung similarity (tingkat kesamaan) antar dua buah obyek. Untuk tujuan klastering dokumen fungsi yang baik adalah fungsi Cosine Similaritas. Berikut adalah persamaan dari metode *Cosine Distance* :

Untuk notasi himpunan digunakan rumus :

$$Similarity(X,Y) = \frac{X \cap Y}{|X|^{\frac{1}{2}} . |Y|^{\frac{1}{2}}}$$

Dimana :

$X \cap Y$ adalah jumlah term yang ada di dokumen X dan yang ada di dokumenY

| X | adalah jumlah term yang ada di dokumen X

| Y | adalah jumlah term yang ada di dokumen Y

DESKRIPSI SISTEM

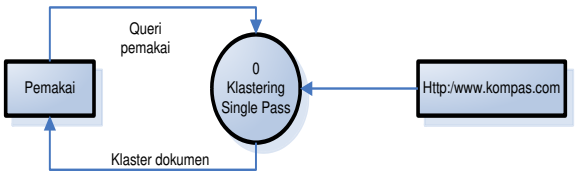
Program aplikasi adalah membangun sistem klastering dokumen dengan menggunakan algoritma *Clustering Single Pass*. Program aplikasi digunakan user untuk membantu mendapatkan dokumen yang berada dalam klaster yang sama.

Dengan menggunakan program aplikasi ini user akan dengan mudah mendapatkan informasi dokumen yang sejenis tanpa harus membaca beberapa dokumen. User tidak perlu menyimpulkan sendiri dari dokumen yang dibacanya untuk mendapatkan informasi yang diinginkan.

Diagram Alir Dokumen

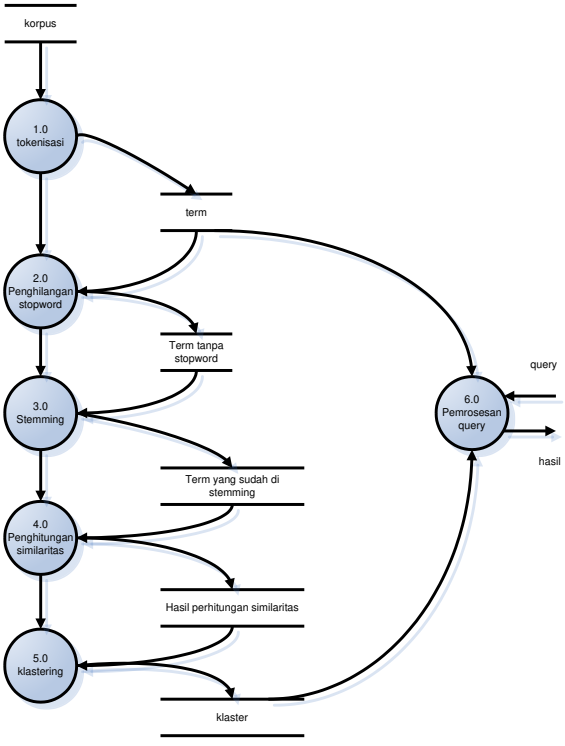
Diagram alir data adalah representasi grafis dari aliran data yang melewati sistem. Dalam penelitian ini diagram alir data dirancang diagram konteks dan diagram level 1.

Diagram konteks sistem temu kembali informasi diperlihatkan pada Gambar 2. Sistem temu kembali informasi memiliki satu entitas luar yaitu user Entitas luar User memberikan input *query* (teks bahasa Indonesia). Dan sistem akan mengeluarkan output klaster dokumen, kelompok dokumen yang berada dalam klaster yang sama dengan *query* yang diinput user.



Gambar 2. Diagram Konteks Sistem Temu Kembali Informasi

Dalam Gambar 3. diperlihatkan Gambar Diagram alir dokumen level 1 Sistem Temu Kembali Informasi terdiri dari proses baca file, proses preprosesing, proses indexing, proses hitung similaritas, proses pembentukan klaster dan pemrosesan *query*. Proses diawali dengan baca file abstrak dalam format teks. Hasil proses baca file akan disimpan dalam penyimpanan data corpus. Dari tabel corpus sistem akan melakukan proses tokenisasi, proses pembuangan stop word dan proses stemming.



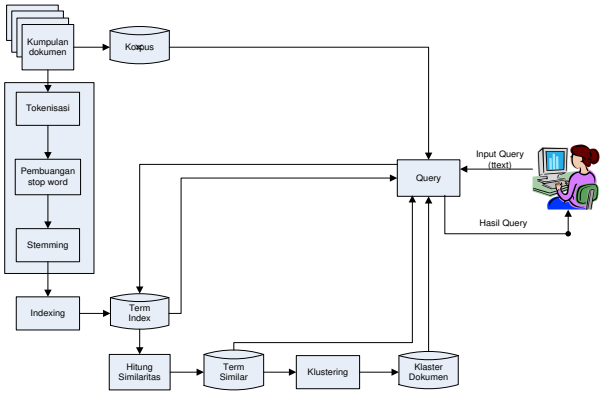
Gambar 3. Diagram Alir Data Level 1 Sistem Temu Kembali Informasi

Kemudian sistem akan melakukan proses indexing. Hasil proses indexing adalah term yang telah diindeks dan disimpan dalam penyimpanan data koleksi. Penyimpanan data koleksi akan digunakan sistem untuk proses hitung similaritas, dari proses akan dihasilkan dokumen similar yang selanjutnya akan digunakan untuk proses pembentukkan klaster dokumen.

User akan melakukan input *query* yang akan diproses oleh sistem, sistem akan menghasilkan output klaster dokumen dan dokumen similar.

ARSITEKTUR SITEM TEMU KEMBALI INFORMASI

Sistem Temu Kembali Informasi dengan algoritma *Clustering Single Pass* sebagai suatu sistem memiliki beberapa proses (modul) yang membangun sistem secara keseluruhan. Modul Sistem Temu Kembali Informasi terdiri dari : modul tokenizations (tokenisasi), modul stop word removal (pembuangan stop word), modul stemming (pengubahan kata dasar), modul term *indexing* (pengindeksan kata), term similarity (kesamaan kata) dan modul clustering (pengelompokan). Secara lengkap arsitektur dari modul Sistem Temu Kembali Informasi dapat dilihat pada



Gambar 4. Arsitektur Sistem Temu Kembali Informasi

Masing-masing modul Sistem Temu Kembali Informasi dapat dijelaskan sebagai berikut :

1. Modul tokenisasi

Sebelum kata dipisahkan dari kalimatnya, terlebih dahulu dibersihkan dari tanda baca, tag html dan angka. Pada penelitian ini untuk membersihkan tanda baca dapat digunakan perintah yang disediakan oleh Java. Pembersihan dilakukan sebelum proses tokenisasi (*tokenizations)* dimaksudkan untuk memperkecil hasil dari tokenisasi. Pada proses tokenisasi akan dibaca dokumen abstrak dalam format teks akan dilakukan proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Pada umumnya setiap kata terindentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses tokenisasi mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan kata.

2. Modul pembuangan stop word

Proses pembuangan stop word dimaksudkan untuk mengetahui suatu kata masuk ke dalam stop word atau tidak. Pembuangan stopword adalah proses pembuangan term yang tidak memiliki arti atau tidak relevan. Term yang diperoleh dari tahap tokenisasi dicek dalam suatu daftar *stopword*, apabila sebuah kata masuk di dalam daftar stopword maka kata tersebut tidak akan diproses lebih lanjut. Sebaliknya apabila sebuah kata tidak termasuk di dalam daftar stopword maka kata tersebut akan masuk keproses berikutnya. Daftar stop word tersimpan dalam suatu tabel, dalam penelitian ini menggunakan daftar stop word yang digunakan oleh Tala (2003), yang merupakan stop word Bahasa Indonesia yang berisi kata-kata seperti ; ini, itu, yang, ke, di, dalam, kepada, dan seterusnya sebanyak 780 kata.

3. 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, 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.

Proses stemming adalah bagian dari proses filtering, tujuan utama dari proses stemming adalah mengembalikan kata dalam bentuk dasarnya. Dengan kata dasar dapat 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.

4. Modul indexing

Proses *indexing* merupakan tahapan preprocessing yang sangat penting dalam sistem temu kembali informasi sebelum pemrosesan query. Pada proses ini seluruh dokumen dalam koleksi disimpan dalam suatu file dengan format sedemikian sehingga dokumen satu dengan dokumen yang lain dapat dibedakan. Setelah kata telah dikembalikan dalam bentuk asal (kata dasar), kata-kata tersebut disimpan kedalam tabel basis data. Penelitian ini menggunakan metode *Inverted Index*, dengan struktur terdiri dari: kata (*term*) dan kemunculan. Kata-kata tersebut adalah himpunan dari kata-kata yang ada pada dokumen, merupakan ekstraksi dari kumpulan dokumen yang ada. Setiap term akan ditunjukkan informasi mengenai semua posisi kemunculannya secara rinci.

5. Modul hitung similaritas

Relevansi sebuah dokumen ke sebuah *query* didasarkan pada *similarity* (similaritas) diantara vektor dokumen dan vektor *query*. Koordinat dari bobot istilah secara dasarnya diturunkan dari frekuensi kemunculan dari istilah. Pada modul ini akan dihitung presentase kemunculan tiap kata (term) dan presentase kesamaan antar dua term. Metode yang digunakan untuk menghitung adalah metode *cosine simmilarity* dengan menggunakan rumus seperti diuraikan pada persamaan (1).

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.

6. Modul klastering

Pada penelitian ini dokumen akan dibuat klaster dengan menggunakan metode *Clustering Single Pass*. Metode ini berawal dari objek-objek individual. Jadi pada awalnya banyaknya klaster sama dengan banyaknya objek. Pertama-tama objek-objek yang paling mirip dikelompokkan, dan kelompok-kelompok awal ini digabungkan sesuai dengan kemiripannya (similaritas). Akhirnya, sewaktu kemiripan berkurang, semua subkelompok digabungkan menjadi satu klaster tunggal. Begitu seterusnya dari hasil similaritas yang tertinggi akan dibandingkan dengan dokumen yang satu dengan dokumen yang lain, sehingga didapat similaritas terendah. Hasil similaritas terendah menyatakan bahwa dokumen tersebut merupakan klaster yang berbeda.

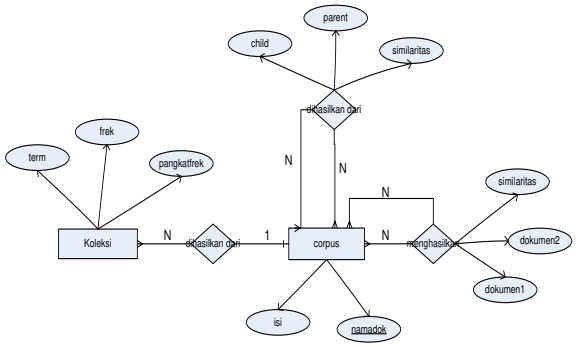
PERANCANGAN DATABASE

Untuk menjawab pertanyaan tentang pemrosesan data, metode pemodelan data menggunakan ERD (*Entity Relationshyp Diagram*) atau Diagram Hubungan Entitas yang memungkinkan perekayasaan perangkat lunak untuk mengidentifikasi objek data dan hubungannya dengan menggunakan notasi grafis.

Entity Relationshyp Diagram digunakan untuk memudahkan struktur data dan hubungan antar data, karena hal ini relatif kompleks. Dengan *Entity Relationshyp Diagram* dapat melakukan pengujian model dengan mengabaikan proses yang harus dilakukan.

Dalam rancangan sistem basis data untuk Sistem Temu Kembali Informasi Bahasa Indonesia, digunakan *Entity Relationshyp Diagram* atau Diagram Hubungan Entitas dan desain tabel untuk menggambarkan atribut-atributnya yang ditunjukkan pada Gambar 5.

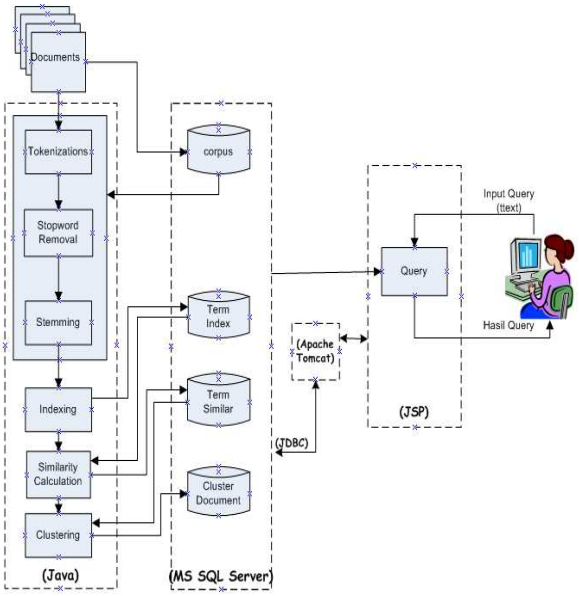
Terlihat pada Gambar 5 bahwa entity corpus memiliki hubungan *one to many* (satu ke banyak) dengan entity koleksi, karena satu corpus dapat menghasilkan banyak koleksi dengan atribut namadok sebagai *primary key* untuk relasi. Hal ini juga sama bahwa entity corpus berelasi *recursive,* dengan relasi *many to many* (banyak ke banyak) yang akan menghasilkan nilai similaritas sebagai hasil relasi. Nilai similaritas akan disimpan dalam tabel cosin. Entity corpus juga berelasi *recursive* dengan relasi *many to many* (banyak ke banyak) dengan nilai similaritas akan menghasilkan klaster dokumen. Klaster dokumen akan disimpan dalam tabel klaster.



Gambar 5. ERD Sistem Temu Kembali Informasi

ARSITEKTUR SISTEM TEMU KEMBALI INFORMASI

Perangkat lunak pada penelitian ini, dibuat dengan menggunakan bahasa pemrograman Java pada platform Java Development Kit (JDK) 6.0. Pemrograman Java digunakan untuk implementasi proses-proses dalam Sistem Temu Kembali Informasi. Database yang digunakan untuk menyimpan data adalah MS SQL Server 2008. User dapat memasukkan *query* melalui interface yang dibangun dengan aplikasi JSP (Java Server Page) dengan Apache Tomcat 6.0 sebagai web server. Untuk mengkoneksikan web server dengan database MS SQL server digunakan aplikasi JDBC. Implementasi untuk perangkat lunak masing-masing proses diperlihatkan pada Gambar 5.



Gambar 6. Implementasi Arsitektur Sistem Temu Kembali Informasi

HASIL DAN PEMBAHASAN

Setelah Sistem Informasi Temu Kembali Informasi Klastering Berita on Line dapat diimplementasikan sesuai dengan desain yang telah dibuat. Tahap selanjutnya adalah tahap melakukan percobaan atau testing dan evaluasi terhadap sistem yang dibuat. Pada tahap pengetesan ini penulis tidak menemukan kesalahan pada program baik secara logika maupun sintaks pada kode program.

Pengujian yang penulis lakukan dalam Sistem Informasi Temu Kembali Informasi Klastering Berita on Line yang berjumlah 60 file dalam format html, telah mampu untuk tidak melakukan indeks-indeks kata umum (stop word) dan telah membentuk kata dasar dari tiap *term* yang ada dalam dokumen abstrak tersebut. Selanjutnya setiap term telah dihitung frekuensinya dan diberikan pembobotan menggunakan cosine similaritas dan selanjutnya term tersebut disimpan pada database korpus.

Selanjutnya dalam pengujian terhadap sistem penulis melakukan pengujian input string *query* dan kemudian hasil pengujian input *query* dilakukan pengukuran hasil retrieval (temu kembali informasi) hasil dengan pengujian *recall precision*.

Pengujian Input Query

Pada tahap pengujian input *query* dilakukan dengan cara memasukkan *query* "pendidikan", "jaringan", "bisnis", dan "manajemen bisnis". Terlihat pada Gambar 6 adalah salah satu contoh hasil tampilan dari input *query* "pendidikan". Hasil proses dari *query* akan ditampilkan dokumen-dokumen yang berada dalam klaster yang sama.

Gambar 7. Hasil Dokumen Input Query "pendidikan"

Pengujian Recall Presisi

Untuk mengevaluasi secara manual kesamaan diantara dokumen dalam cluster-cluster yang telah dikelompokkan digunakan standar sebagaimana Tabel 1. Tabel tersebut berisi berbagai kemungkinan hasil klasifikasi pada tiap *event* (*Per Event contengency table*).

Tabel 1. kategori hasil klasifikasi

	In Event	Not In event
In cluster	a	b
Not In Cluster	c	d

Tabel 1 menunjukkan bahwa hasil klasifikasi adakalanya memang termasuk event (a) yang dimaksud dan adakalanya tidak (b). Sedangkan dokumen yang tidak termasuk dalam hasil klasifikasi suatu *event*, adakalanya memang bukan anggota *event* itu (d) dan adakalanya ternyata seharusnya menjadi anggota event tersebut (c). Dalam hal ini, kempat

parameter di atas digunakan untuk menghitung 2 parameter evaluasi, yakni :

1. *Recall*, yakni tingkat keberhasilan mengenali suatu event dari seluruh event yang seharusnya dikenali. Rumusnya adalah r = a/(a+c) untuk a+c > 0. Selain itu tidak didefinisikan
2. *Precision*, yakni tingkat ketepatan hasil klasifikasi terhadap suatu event.Artinya,dari seluruh dokumen hasil klasifikasi, berapa persenkah yang dinyatakan benar.

   Rumusnya adalah p = a/(a+b) jika a+b > 0.

Selain itu tidak didefinisikan Dari hasil evaluasi yang dilakukan terhadap data training yang diambil dari suara pembaruan *online m*ulai tanggal 1 Agustus 2001 sampai dengan 31 Agustus 2001 dengan perincian sebagai berikut :

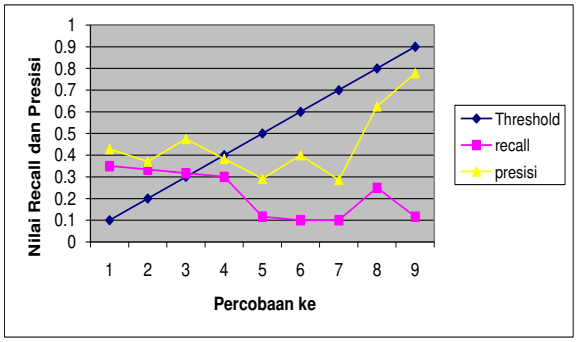
- Diambil 60 Dokumen sebagai Restrospective/training set yang dikalsifikasikan kedalam 23 event
- Setelah melalui proses *stemming* maka dapat dilakukan penghitungan frekuensi kata dalam dokumen dengan menggunakan kamus sejumlah 29.349 kata dasar bahasa Indonesia
- Dari matrik yang dibentuk dihasilkan sebanyak 13.000 *record* untuk kaitan dokumen dengan kata tf(t,d) dengan frekuensi di atas 0

Didapatkan hasil nilai Recall dan Precision sebagaimana Tabel 2.

Tabel 2. Tabel Hasil Uji Coba

Threshold	**recall**	**presisi**
0.1	0.35	0.428571429
0.2	0.333333333	0.37037037
0.3	0.316666667	0.475
0.4	0.3	0.382978723
0.5	0.116666667	0.291666667
0.6	0.1	0.4
0.7	0.1	0.285714286
0.8	0.25	0.625
0.9	0.116666667	0.777777778

Distribusi nilai kedua parameter dapat digambarkan dengan grafik pada Gambar 7



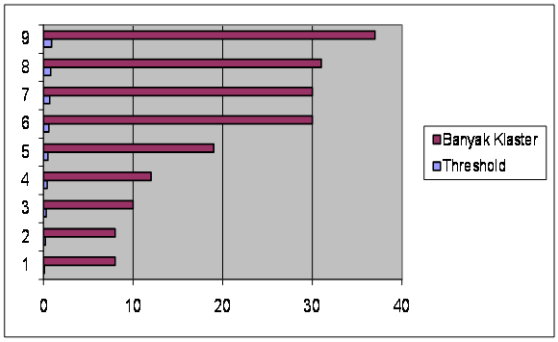
Gambar 8. Grafik Recall dan Precision untuk *Algoritma Single Pass*

Dari Gambar 7 terlihat bahwa nilai terbaik untuk recall didapat pada treshold pada angka 0.1. Sedanhkan nilai terbaik untuk presisi dihasilkan dari treshold pada angka 0.9. Treshold tersebut didapatkan dari percobaan secara linear terhadap 9 treshold yang berbeda. Percobaan dilakukan dengan menggunakan treshol mulai dari nilai treshold 0.1 sampai dengan mendapatkan jumlah kluster = jumlah dokumen atau nilai treshold di atas keseluruhan similarity maksimal. Untuk jumlah kluster yang dihasilkan dari 9 percobaan yang dilakukan dapat dilihat pada tabel 3

Tabel 3. Tabel Hasil Kluster

No	Threshold	Banyak Klaster
1	0.1	8
2	0.2	8
3	0.3	10
4	0.4	12
5	0.5	19
6	0.6	30
7	0.7	30
8	0.8	31
9	0.9	37

Sedangkan distribusi jumlah kluster yang dihasilkan dari percobaan yang dilakukan dapat dilihat pada Gambar 8



Gambar 9. Hasil Kluster

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. Klaster dapat membantu menemukan dokumen yang ada dalam satu klaster dengan query yang diinputkan user.
5. Klaster dapat membantu mendapatkan dokumen yang relevan.
6. Nilai treshold (nilai batas) yang paling bagus digunakan adalah 0.2 dengan nilai recall sebesar 0.33 dan precision 0.37

SARAN

Dengan keterbatasan kemampuan dan waktu yang tersedia penulis menyadari bahwa masih banyak terdapat kekurangan dalam sistem ini terutama metode klastering yang digunakan. Kedepan nantinya diharapkan dalam pengembangan Sistem Informasi berbasis web, penulis menyarankan beberapa hal:

1. Sistem yang dibuat dapat dikembangkan lebih lanjut dengan menerapkan pada file teks Bahasa Indonesia dengan melakukan modifikasi stop word dan algoritma stemming agar hasil stemming lebih optimal.
2. Bagi peneliti lain yang berniat mengembangkan sistem Informasi Temu Kembali Bahasa Indonesia ini disarankan untuk menggunakan metode klastering yang diperluas sehingga hasil klaster dokumen akan lebih baik.
3. Untuk term yang bernilai 0 (nol) dalam setiao dokumen tidak perlu dilibatkan dalam perhitungan, karena hanya akan menambah waktu perhitungan.

DAFTAR PUSTAKA

Baeza-Yates, R. & Ribeiro-Neto, B**.,** (1999). *Modern Information Retrieval*, Addison-Wesley.

Fadillah Z Tala*, (*2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands.

Ellis, David, (1996). *Progress and Problems in Information Retrieval*, 2nd ed. London: Library Association.

Gordon, Michael D., (1991). *User-Based Document Clustering by Redescribing Subject Descriptions with a Genetic Algorithm,*. Journal of American Society for Information Science, 311-322.

Issue. (n.d.). (2013). *Journal Portal | Islamic University of Indonesia | Universitas Islam Indonesia*. Retrieved July 31, 2013, from http://journal.uii.ac.id/index.php/Snati/issue/view/

Karypis G., Zhao Y., (2004). *Hierarchical Clustering Algorithms for Document Datasets*, *University of Minnesota, Department of Computer Science and Engineering and Digital Technology Center and Army HPC Research Center, Minneapolis, MN 55455.*

Pressman R, (1997). *Software Engineering,* Mc Graw Hill, USA.

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

Salton, G., (1989). *Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer*, Addison – Wesly Publishing Company, Inc. All rights reserved.

Salton, G. and Buckley, (1988). *Term Weigting Approaches in Automatic Text Retrieval, Department of Computer Science*, Cornell University, Ithaca, NY 14853, USA.

Salton, G., (1971). *Cluster Search Strategies and the Optimization of Retrieval Efectiveness*, dalam G. Salton, ed. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs: Prentice-Hall, 223-242

Steinbach, M., Xiong H., Ruslim A., Kumar V., (2007). *Characterizing Pattern Preserving Clustering, Department of Management Science and Information Systems Rutgers*, the State University of New Jersey, USA.

Wen Yue, (2005). *Using Query Expansion and Classification for Information Retrieval, College of Computer and Communication*, Hunan University ChangSha, Hunan Province, 410082,China.

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, February 2001, pp. 109-125.