HIERARCHICAL AGGLOMERATIVE CLUSTERING UNTUK PENGELOMPOKAN SKRIPSI MAHASISWA

Herny Februariyanti[1], Dwi Budi Santoso[2]
[1,2]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Stikubank,
e-mail: [1]hernyfeb@edu.unisbank.ac.id, [2]dbs@ edu.unisbank.ac.id

ABSTRAK

Data penelitian dapat dikelompokkan berdasarkan kemiripan tema, objek maupun metode penelitian. Hasil pengelompokkan data penelitian dapat memperlihatkan bagaimana pola kemiripan penelitian dari waktu ke waktu. Hasil pengelompokan dapat menunjukkan kapan waktu penelitian mahasiswa banyak mengambil materi yang sama dan kapan waktu penelitian mahasiswa beragam. Pengelompokan data penelitian yang umumnya berbentuk teks dapat dilakukan dengan text mining. Penelitian ini mengelompokkan dokumen skripsi Program Studi Sistem Informasi Fakultas Teknologi Informasi Universitas Stikubank Semarang. Semakin bertambahnya penelitian skripsi dengan mata kuliah terbatas menyebabkan semakin banyak pula mahasiswa yang mengambil penelitian yang mirip tema, objek, atau metode penelitian dengan penelitian sebelumnya.

Penelitian ini melakukan clustering Judul Skripsi mahasiswa Program Studi Sistem Informasi dengan menggunakan algoritma Hierarchical Agglomerative Clustering. Untuk mengetahui kemiripan antar judul dilakukan proses menghitung jarak antar objek dalam hal ini kemiripan antara masing-masing judul skripsi menggunakan algoritma dice coefficient. Objek judul yang memiliki jarak paling pendek maka merupakan objek judul yang paling mirip. Dari hasil proses perhitungan jarak kemiripan antar dokumen dihasilkan dokumen paling mirip dengan jarak sebesar 0.2. Selanjutnya hasil clustering dilakukan pemotongan pada 2 titik yaitu titik 0.25 yang menghasilkan dan pemotongan pada titik 0.5 dengan menghasilkan 5 buah cluster.

***Kata Kunci**: Text Mining, dice coefficient, Hierarchical Agglomerative*

1. PENDAHULUAN

Setiap tahunnya Fakultas Teknologi Informasi meluluskan mahasiswa dengan penelitian skripsi yang beragam. Setiap tahunnya jumlah data skripsi selalu bertambah. Semakin bertambahnya penelitian skripsi dengan mata kuliah terbatas menyebabkan semakin banyak pula mahasiswa yang mengambil penelitian yang mirip tema, objek, atau metode penelitian dengan penelitian sebelumnya. Pengelompokan data penelitian yang umumnya berbentuk teks dapat dilakukan dengan text mining. Tujuan dari *text mining* adalah untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen [1]. Terdapat beberapa metode *text mining* salah satunya adalah *clustering*. *Clustering* adalah suatu metode analisa data untuk memecahkan masalah pengelompokan data. Kategori algoritma *clustering* yang banyak dikenal adalah *Hierarchical Agglomerative*.

Pada algoritma *clustering,* data akan dikelompokkan menjadi cluster-cluster berdasarkan kemiripan satu data dengan yang lain. Prinsip dari clustering adalah memaksimalkan kesamaan antar anggota satu cluster dan meminimumkan kesamaan antar anggota cluster yang berbeda. *Hierarchical Clustering* adalah salah satu algoritma clustering yang dapat digunakan untuk meng-cluster dokumen (*document clustering*).

Hasil keseluruhan dari algoritma *hierarchical clustering* secara grafik dapat digambarkan sebagai tree, yang disebut dengan dendogram. Tree ini secara grafik menggambarkan proses penggabungan dari cluster – cluster yang ada, sehingga menghasilkan cluster dengan level yang lebih tinggi.

2. TINJAUAN PUSTAKA

2.1 Text Mining

Text mining, yang juga disebut sebagai *Teks Data Mining* (TDM) atau *Knowledge Discovery in Text* (KDT), secara umum engacu ada proses ekstraksi informasi dari dokumen-dokumen teks tak terstruktur (*unstructured*). Text mining dapat didefinisikan sebagai penemuan informasi baru dan tidak diketahui sebelumnya oleh komputer, yang secara otomatis mengekstrak informasi dari sumber -sumber teks tak terstruktur yang berbeda. Kunci dari proses ini adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber[2].

Tujuan utama *text mining* adalah mendukung proses *knowledge discovery* pada koleksi dokumen yang besar. Pada prinsipnya, *text mining* adalah bidang ilmu multidisipliner, melibatkan *information retrieval* (IR), text analysis, *information extraction* (IE), *clustering, categorization, visualization, database technology, natural language processing* (NLP), *machinelearning*, dan data mining. Dapat pula dikatakan bahwa text mining merupakan salah satu bentuk aplikasi kecerdasan buatan (*artificial intelligence* / AI).

Text mining mencoba memecahkan masalah *information overload* dengan menggunakan teknik-teknik dari bidang ilmu yang terkait. *Text mining* dapat dipandang sebagai suatu perluasan dari data mining atau *knowledge -discovery in database* (KDD), yang mencoba untuk menemukan pola-pola menarik dari basis data berskala besar. Namun *text mining* memiliki potensi komersil yang lebih tinggi dibandingkan dengan data

mining, karena kebanyakan format alami dari penyimpanan informasi adalah berupa teks. *Text mining* menggunakan informasi teks tak terstruktur dan mengujinya dalam upaya mengungkap struktur dan arti yang tersembunyi di dalam teks. Perbedaan mendasar antara *text mining* dan data mining terletak pada sumber data yang digunakan. Pada data mining, pola-pola diekstrak dari basis data yang terstruktur, sedangkan di *text mining*, pola -pola diekstrak dari data tekstual (*natural language*). Secara umum, basis data didesain untuk program dengan tujuan melakukan pemrosesan secara otomatis, sedangkan teks ditulis untuk dibaca langsung oleh manusia[3].

Ada empat tahap proses pokok dalam *text mining* , yaitu pemrosesan awal terhadap teks (*text preprocessing*), transformasi teks (*text transformation*), pemilihan fitur (*feature selection*), dan penemuan pola (*pattern discovery*) [4].

2.2 Clustering Dokumen

Clustering didefinisikan sebagai upaya mengelompokkan data ke dalam cluster sedemikian sehingga data-data di dalam cluster yang sama memiliki lebih kesamaan dibandingkan dengan data-data pada cluster yang berbeda. Bisa juga diartikan sebagai proses untuk mendefinisikan pemetaan / m*apping* $f : D \rightarrow C$ dari beberapa data $D = \{t_1, t_{2, \ldots\ldots} t_n\}$ kedalam beberapa cluster $C=\{ c_{1,} c_{2,\ldots,} c_n\}$ berdasarkan kesamaan antar t_i . Sebuah cluster adalah sekumpulan obyek yang digabung bersama karena persamaan atau kedekatannya. Clustering 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 clustering adalah untuk menentukan pengelompokan dari suatu set data. Akan tetapi tidak ada ”ukuran terbaik” untuk pengelompokan data. Untuk pengelompokkan data tergantung tujuan akhir dari clustering, maka diperlukan suatu kriteria sehingga hasil clustering seperti yang diinginkan. Penelitian tentang *clustering document* telah banyak dilakukan. Secara umum clustering dokumen adalah proses mengelompokkan dokumen berdasarkan kemiripan antara satu dengan yang lain dalam satu cluster[5] [6].

Tujuan clustering dokumen adalah untuk memisahkan dokumen yang relevan dari dokumen yang tidak *relevan* [7]. 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 cluster. Clustering dokumen dapat dilakukan sebelum atau sesudah proses temu kembali[7]. Pada clustering dokumen yang dilakukan sebelum proses temu kembali informasi, koleksi dokumen dikelompokkan ke dalam cluster berdasarkan kemiripan (*similarity*) antar dokumen. Selanjutnya dalam proses temu kembali informasi, apabila suatu dokumen ditemukan maka seluruh dokumen yang berada dalam cluster yang sama dengan dokumen tersebut juga dapat ditemukan.

Penggabungan antara penelusuran secara menyeluruh (*full search*) dengan penelusuran berbasis cluster (*cluster-based retrieval*) dapat meningkatkan ketelitian sampai dengan 25%. Penggabungan antara metode pengclusteran dengan *fusion* (pemberian peringkat terdadap dokumen secara keseluruhan) akan meningkatkan efektivitas temu kembali informasi. Pada algoritma clustering, dokumen akan dikelompokkan menjadi *cluster-cluster* berdasarkan kemiripan satu data dengan yang lain.[7] Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu cluster dan meminimumkan kesamaan antar anggota *cluster* yang berbeda.

2.3 Clustering Hierarchical

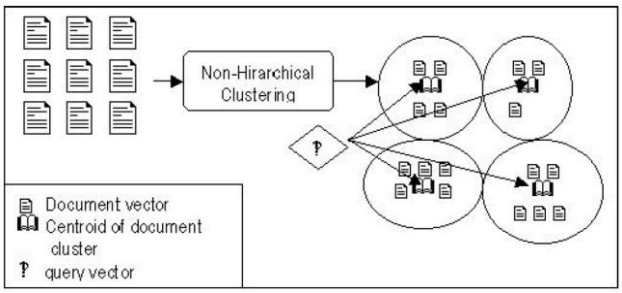
Metode pembentukan cluster biasanya dikategorikan menurut tipe dari struktur cluster yang dihasilkan. Secara umum metode cluster terbagi menjadi dua, yaitu metode *Non-Hierarchical Clustering* (klastering non-hirarkhis) dan metode *Hierarchical Clustering* (klastering hirarkhis).

Metode non-hirarkhis disebut juga metode partisi, yaitu membagi serangkaian data yang terdiri dari n obyek ke dalam k cluster (k<n) yang tidak saling tumpang-tindih (*overlap*), dimana nilai *k* telah ditentukan sebelumnya. Salah satu prosedur pengelompokkan pada non-hirarkhis adalah dengan menggunakan metode *k-means*. Metode ini merupakan metode pengelompokkan yang bertujuan untuk mengelompokkan objek sedemikian hingga jarak tiap-tiap objek kepusat kelompok didalam suatu kelompok adalah minimum.

Pembentukan cluster dokumen dalam Sistem Temu Kembali Informasi dengan metode non-hirarkhis adalah sebagi berikut[8] :

a. Membandingkan ciri-ciri identifikasi (*identifier*) suatu dokumen dengan dokumen lain yang ada dalam koleksi dan mengelompokkan dokumen-dokumen yang memiliki serangkaian ciri-ciri identifikasi yang serupa ke dalam satu cluster.
b. Pada setiap cluster dokumen yang dihasilkan, dipilih sebuah unsur yang dapat mewakili seluruh dokumen yang ada dalam cluster yang bersangkutan yang disebut *centroid. Centroid* atau perwakilan cluster adalah sebuah *record* yang dapat mewakili ciri-ciri atau karakteristik dokumen dalam sebuah cluster.
c. Proses penelusuran dilakukan dalam dua tahap, yaitu: 1) membandingkan *query* dengan *centroid* pada masing-masing cluster dokumen; 2) mencocokkan *query* dengan masing-masing dokumen dalam cluster yang mengandung *centroid* yang paling sesuai.

Proses pembentukan cluster dokumen dan penelusuran tersebut dapat diilustrasikan seperti pada Gambar 1 di bawah ini :



Gambar 1 Cluster Dokumen dengan Metode Non-hirarkhis

Metode cluster yang kedua adalah metode *Hierarchical Clustering* (klastering hirarkhis). Metode pengelompokkan hirarkhis biasanya digunakan apabila belum ada informasi jumlah kelompok yang akan dipilih. Arah pengelompokkan bisa bersifat *divisive* (*top to down*) artinya dari 1 cluster sampai menjadi k buah cluster atau bersifat *agglomerative* (bottom up) artinya dari n cluster (dari n-buah data yang ada) menjadi k buah cluster. Teknik hirarkhis (hierarchical methods) adalah teknik clustering membentuk kontruksi hirarki atau berdasarkan tingkatan tertentu seperti struktur pohon. Dengan demikian proses pengelompokkannya dilakukan secara bertingkat atau bertahap.

Hierarchical Clusstering adalah salah satu algoritma clustering yang dapat digunakan untuk meng-*cluster* dokumen *(document clustering*). Dari teknik *Hierarchical Clustering*, dapat dihasilkan suatu kumpulan partisi yang berurutan, dimana dalam kumpulan tersebut terdapat:

a. Cluster-cluster yang mempunyai poin – poin individu. *Cluster-cluster* ini berada di level yang paling bawah.
b. Sebuah cluster yang didalamnya terdapat poin – poin yang dipunyai semua *cluster* didalamnya. *Single cluster* ini berada di *level* yang paling atas.

Pembentukan cluster dokumen dalam Sistem Temu Kembali Informasi dengan metode hirarkhis adalah sebagi berikut:

a. Mengidentifikasi dua dokumen yang paling mirip dan menggabungkannya menjadi sebuah cluster.
b. Mengidentifikasi dan menggabungkan dua dokumen yang paling mirip berikutnya menjadi sebuah cluster sampai semua dokumen tergabung dalam cluster-cluster yang terbentuk.
b. Proses penelusuran dokumen dilakukan dengan cara mencocokkan *query* dengan *centroid. Centroid* merupakan dokumen parent pada masing-masing cluster dokumen. Berikutnya dokumen yang berada dalam satu cluster dengan *centroid* akan ditampilkan sebagai hasil *query*.

3. METODE PENELITIAN

1. Obyek Penelitian
   Obyek penelitian dari penelitian ini dokumen abstrak skripsi mahasiswa Fakultas Teknologi Informasi Universitas Stikubank Semarang.
2. Teknik Pengumpulan Data
   Pengumpulan data dimaksudkan agar mendapatkan bahan-bahan yang relevan, akurat dan reliable. Maka teknik pengumpulan data yang dilakukan dalam penelitian ini adalah sebagai berikut :
   a. Observasi
      Dengan melakukan pengamatan dan pencatatan secara sistematis tentang hal-hal yang berhubungan dengan basis data dokumen abstrak skripsi mahasiswa Fakultas Teknologi Informasi Universitas Stikubank Semarang
   b. Studi Pustaka
      Dengan pengumpulan data dari bahan-bahan referensi, arsip, dan dokumen yang berhubungan dengan permasalan dalam penelitian ini.
3. Metode Pengembangan
   Penelitian ini menggunakan model *prototyping*. Di dalam model ini sistem dirancang dan dibangun secara bertahap dan untuk setiap tahap pengembangan dilakukan percobaan-percobaan untuk melihat apakah sistem sudah bekerja sesuai dengan yang diinginkan. Tahap-tahap prototyping adalah :
   a. Analisa
      Pada tahap ini dilakukan analisa tentang masalah penelitian dan menentukan pemecahan masalah yang tepat untuk menyelesaikannya.
   b. Disain
      Pada tahap ini dibangun rancangan sistem dengan beberapa diagram bantu seperti Data Flow Diagram.
   c. Prototype
      Pada tahap ini dibangun aplikasi berbasis web yang sesuai dengan disain dan kebutuhan sistem.

d. Pengujian
Pada tahap ini dilakukan pengujian pada pustaka fungsi yang sudah dibangun.
e. Evaluasi
Pada tahap ini dilakukan evaluasi apakah performa aplikasi sudah sesuai dengan yang diharapkan, apabila belum maka dilakukan penyesuaian-penyesuaian secukupnya.
f. Penyesuaian
Tahap ini dilakukan apabila pada evaluasi performa aplikasi kurang memadai dan dibutuhkan perbaikan, tahap ini melakukan penyesuaian dan perbaikan pada aplikasi sesuia dengan kebutuhan.

4. HASIL DAN PEMBAHASAN

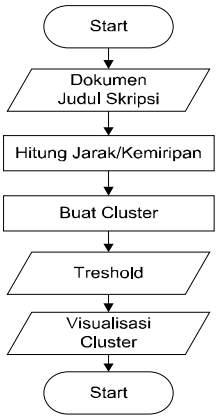
4.1 Asitektur Sistem

Arsitektur merupakan gambaran menyeluruh struktur sistem beserta seluruh variable atau *passing* parameter sehingga siap ditulis sebagai modul program. Desain arsitektur dapat dilihat pada gambar 4.

a. Upload judul skripsi
Pengguna melakukan *upload judul skripsi* dan disimpan dalam *document root* sistem. *File judul skripsi difilter*, diambil khusus baris yang mengandung *judul skripsi* dan disimpan dalam *data storekeyphrase*. *Penggunaakan* mendapatkan notifikasi bahwa *upload file judul skripsi* berhasil.
b. Pre-process
Pre process adalah tahapan pembersihan teks dari judul skripsi yang dimulai dari memecah kalimat menjadi kata(tokenisasi), membuang kata yang tidak mempunyai makna(Stopword removal) dan merubahnya menjadi kata dasar(stemming). Kata dasar ini kemudian disimpan di tabel indeks. Pengguna mendapatkan notifikasi bahwa tahapan *pre process* berhasil.
c. Hitung Jarak/Kemiripan
Setelah tahapan *pre* processselesai, berikutnya adalah menghitung jarak antar objek dalam hal ini kemiripan antara masing-masing *judul skripsi*. Sebagai contoh objek pada langkah pertama objek nomor 1 dihitung jaraknya dengan objek nomor 2 sampai nomor 10, setelah itu objek nomor 2 dihitung jaraknya dengan objek nomor 3 sampai nomor 10, dan seterusnya sampai yang terakhir objek nomor 9 dibandingkan dengan objek nomor 10.
Algoritma yang digunakan untuk menghitung jarak adalah *dice coefficient* dengan persamaan sebagai berikut:

$$\text{sim}(A,B) = \frac{2\,A \cap B}{A + B} \qquad (1)$$

Sebagai contoh jika kita terapkan rumus *dice coefficient* pada objek nomor 1 dan 2 maka perhitungannya adalah :
A = {sistem, pakar, diagnosa, penyakit, tanaman, padi}
B = { sistem, pakar, diagnosa, penyakit, tanaman, tebu}
sim (A,B) = 2 * 5 / (6 + 6) = 10/12 = 0,83
dice coefficient = 1 – 0.83 = 0.17
d. *Buat Cluster*
Pembentukan*cluster* pada daftar *search keyword* menggunakan algoritma *hac* dimana pada awalnya setiap *judul skripsi*dianggap sebagai satu *cluster*, kemudian dicari jarak terpendek (yang paling mirip) dengan *cluster* lainnya, kedua *judul skripsi* tersebut digabung menjadi satu. Kemudian *cluster* yang baru dihitung lagi jaraknya dengan *cluster-cluster* yang lain, dicari lagi yang terpendek, kemudian digabungkan, sampai akhirnya menjadi satu *cluste.*
e. Visualisasi *Cluster*
Proses *terakhir* adalah visualisasi *cluster*. *Hac* clustering termasuk dalam *hierarchical clustering* dimana jumlah *cluster* tidak ditentukan sebelumnya. Supaya dalam visualisasinya pengguna lebih mudah untuk menganalisa, hirarki harus dipotong dibeberapa titik, dalam hal ini pemotongan menggunakan *threshold* yang merupakan angka kemiripan minimal dua buah objek dapat digabungkan menjadi satu *cluster*. Cluster ditampilkan kepada pengguna berdasarkan data dari data store cluster dan data store *keyphrase*.



Gambar 2. Arsitektur Sistem

Penelitian ini menggunakan data *file access.log* yang diunduh dari *server website* www.unisbank.ac.id dari tanggal 1 Januari 2014 sampai dengan 31 Januari 2014. Hasil dari unduhan kemudian akan dilakukan pengolahan sampai terbentuk *cluster judul skripsi* yang terdapat pada *file* tersebut. Data *judul skripsi*kemudian akan dilakukan pembersihan dan hanya diambil *field judul skripsi*sebagaimana terlihat pada tabel 1

Tabel 1. Contoh Data Judul skripsi

No	Judul
1	sistem informasi rawat inap rumah sakit islam sultan agung semarang
2	rancang bangun sistem informasi akademik pada man kendal berbasis web
3	pengaplikasian e-busines pada produk cv. amanah dengan asp.net
4	rancang bangun aplikasi perpustakaan pada smk cut nya' dien semarang dengan menggunakan java
5	pengelolaan advertising pada radio 101.6 ibc semarang

Metode *hac* diimplementasikan dengan menggunakan matriks n x n dari jarak antara semua pasangan objek. Langkah pertama yang dilakukan adalah membandingkan jarak masing-masing obyek data menggunakan *dice* similarity. Nilai kemiripan yang dihasilkan memiliki *range* antara 0 sampai dengan 1. Semakin mendekati 0 maka dapat dikatakan dua obyek yang dibandingkan semakin mirip, dan sebaliknya semakin mendekati angka 1 semakin tidak mirip. Data hasil perhitungan *similaritas* adalah sebagai berikut :

id	r	s	drs
2	1	2	0.8
3	1	3	1
4	1	4	0.8824
5	1	5	0.7895
6	1	6	0.4444

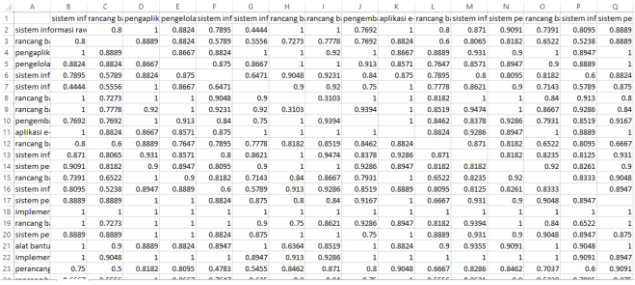
Gambar 3. Jarak/Kemiripan antar Obyek Data

Penghitungan jarak dihitung dengan cara membandingkan keseluruhan obyek data. Obyek data nomor 1, dihitung jaraknya dengan obyek data nomor 2, kemudian hasilnya disimpan. Berikutnya obyek data nomor 1 dibandingkan dengan obyek nomor 3, seterusnya sampai obyek nomor 1 dibandingkan dengan obyek ke n, dimana n adalah jumlah maksimal data, sehingga untuk membandingkan obyek nomor 1 sampai dengan obyek nomor 44 . Berikutnya proses dimulai dari obyek nomor 2 dibandingkan dengan obyek nomor 3, sampai dengan obyek nomor 2 dibandingkan dengan obyek nomor 44, dan seterusnya sampai obyek nomor 43 dibandingkan dengan obyek nomor 44.

Perhitungan jarak menggunakan *dice similarity* dimana jarak antara dua obyek dihitung dengan cara irisan antara dua obyek dikalikan 2 dan dibagi dengan total dari obyek A ditambah obyek B seperti yang dituliskan pada persamaan 2.1. Contoh perhitungan untuk obyek nomor 1 dan nomor 2 adalah sebagai berikut :

a. Sistem informasi rawat inap rumah sakit islam sultan agung semarang
b. Rancang bangun sistem informasi akademik pada man kendal berbasis web

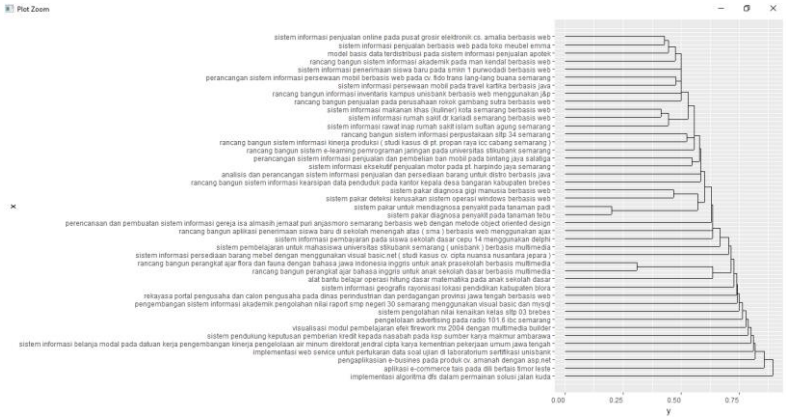
maka nilai *dice similarity* dari kedua obyek tersebut adalah 2 * 2 / 20 = 0,2. Sehingga nilai nol koma dua disini dapat diartikan bahwa kedua obyek hanya sedikit yang mirip atau jaraknya jauh. Untuk memudahkan dengan perhitungan *hac*dimana penggabungan dua obyek *cluster* berdasarkan jarak terdekat maka jarak dihitung dengan cara 1 – nilai *dice similarity*, sehingga untuk kedua obyek tersebut, nilai jarak yang tersimpan di tabel adalah 1-0,2 = 0,8. Berdasarkan cara penghitungan diatas, maka matriks jarak yang didapatkan untuk m=0 dapat dilihat pada gambar 4.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		sistem inf	rancang bi	pengaplik	pengelola	sistem inf	sistem inf	rancang bi	rancang bi	pengemb	aplikasi e-	rancang bi	sistem inf	sistem pe	rancang bi	sistem inf	sistem pe
2	sistem informasi	rancang bi	0.8	1	0.8824	0.7895	0.4444	1	1	0.7692	1	0.8	0.871	0.9091	0.7391	0.8095	0.8889
3	rancang bi	0.8	1	0.8889	0.8824	0.5789	0.5556	0.7273	0.7778	0.7692	0.8824	0.6	0.8065	0.8182	0.6522	0.5238	0.8889
4	pengaplik	1	0.8889		0.8667	0.8824	1	1	0.92	1	0.8667	0.8889	0.931	0.9	1	0.8947	1
5	pengelola	0.8824	0.8824	0.8667		0.875	0.8667	1	1	0.913	0.8571	0.7647	0.8571	0.8947	0.9	0.8889	1
6	sistem inf	0.7895	0.5789	0.8824	0.875		0.6471	0.9048	0.9231	0.84	0.875	0.7895	0.8	0.8095	0.8182	0.6	0.8824
7	sistem inf	0.4444	0.5556	1	0.8667	0.6471		0.9	0.92	0.75	1	0.7778	0.8621	0.9	0.7143	0.5789	0.875
8	rancang bi	1	0.7273	1	1	0.9048	0.9		0.3103	1	1	0.8182	1	1	0.84	0.913	0.8
9	rancang bi	1	0.7778	0.92	1	0.9231	0.92	0.3103		0.9394	1	0.8519	0.9474	1	0.8667	0.9286	0.84
10	pengemba	0.7692	0.7692	1	0.913	0.84	0.75	1	0.9394		1	0.8462	0.8378	0.9286	0.7931	0.8519	0.9167
11	aplikasi e-	1	0.8824	0.8667	0.8571	0.875	1	1	1	1		0.8824	0.9286	0.8947	1	0.8889	1
12	rancang bi	0.8	0.6	0.8889	0.7647	0.7895	0.7778	0.8182	0.8519	0.8462	0.8824		0.871	0.8182	0.6522	0.8095	0.6667
13	sistem inf	0.871	0.8065	0.931	0.8571	0.8	0.8621	1	0.9474	0.8378	0.9286	0.871		0.8182	0.8235	0.8125	0.931
14	sistem pe	0.9091	0.8182	0.9	0.8947	0.8095	0.9	1	1	0.9286	0.8947	0.8182	0.8182		0.92	0.8261	0.9
15	rancang bi	0.7391	0.6522	1	0.9	0.8182	0.7143	0.84	0.8667	0.7931	1	0.6522	0.8235	0.92		0.8333	0.9048
16	sistem inf	0.8095	0.5238	0.8947	0.8889	0.6	0.5789	0.913	0.9286	0.8519	0.8889	0.8095	0.8125	0.8261	0.8333		0.8947
17	sistem pe	0.8889	0.8889	1	1	0.8824	0.875	0.8	0.84	0.9167	1	0.6667	0.931	0.9	0.9048	0.8947	
18	implemer	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	rancang bi	1	0.7273	1	1	1	0.9	0.75	0.8621	0.9286	0.8947	0.8182	0.9394	1	0.84	0.6522	1
20	sistem pe	0.8889	0.8889	1	1	0.8824	0.875	1	1	0.75	1	0.8889	0.931	0.9	0.9048	0.8947	0.875
21	alat bantu	1	0.9	0.8889	0.8824	0.8947	1	0.6364	0.8519	1	0.8824	0.9	0.9355	0.9091	1	0.9048	1
22	implemer	1	0.9048	1	1	1	0.8947	0.913	0.9286	1	1	1	1	1	1	0.9091	0.8947
23	perancang	0.75	0.5	0.8182	0.8095	0.4783	0.5455	0.8462	0.871	0.8	0.9048	0.6667	0.8286	0.8462	0.7037	0.6	0.9091

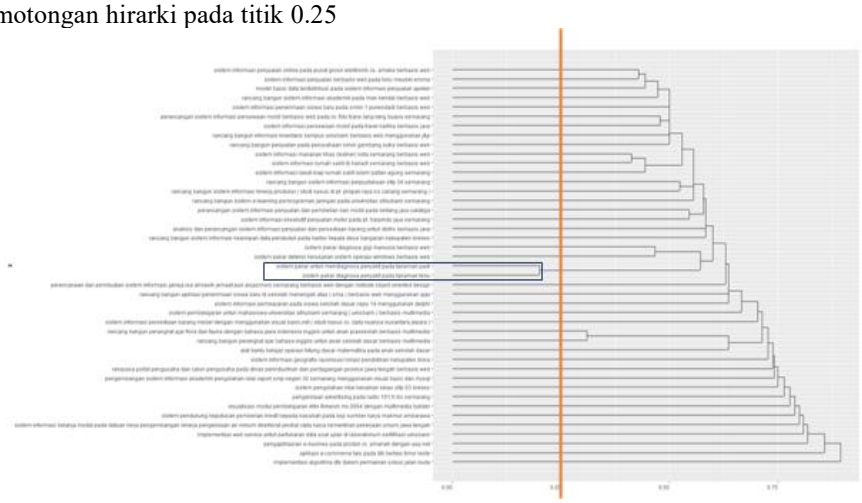
Gambar 4. Matriks Jarak

Dengan menggunakan *hac* maka didapatkan *cluster* yang digambarkan dengan dendrogram seperti terlihat pada gambar 5.



Gambar 5. Dendrogram Cluster

Gambar 5 memperlihatkan 1 *cluster* secara keseluruhan obyek data. Hal ini menyebabkan pengguna sistem kesulitan untuk menganalisa topik secara terpisah apalagi jika jumlah obyek data sangat besar, sehingga *cluster* perlu dipecah, supaya dapat dilihat topik-topik yang spesifik pada daftar *judul skripsi* dalam obyek data. Penggunaan *threshold* sebagai ambang batas penggabungan obyek data menjadi sebuah *cluster* dapat dijadikan sebagai solusi untuk memecah *cluster* besar menjadi beberapa *cluster*. Penerapan *threshold* dalam *dendrogram* pada gambar 5 dilakukan dengan cara pemotongan hirarki pada titik tertentu. Gambar 6 memperlihatkan pemotongan hirarki pada titik 0.25



Gambar 6. pemotongan *dendrogram* pada titik 0.25

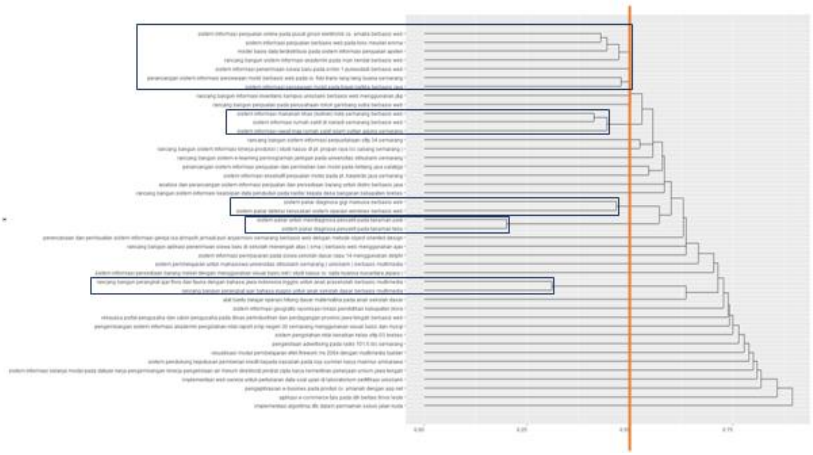
Pemotongan *dendrogram* pada titik 0.25 menghasilkan 1 buah cluster seperti terlihat pada gambar yang ditandai dengan kotak warna biru. Mengacu pada data pada tabel , maka 1*cluster* tersebut mempunyai obyek data sebagai berikut :

- cluster 1 dengan obyek data {34. sistem pakar untuk mendiagnosa penyakit pada tanaman padi, 28. sistem pakar diagnosa penyakit pada tanaman tebu}

Dari 1cluster yang terbentuk, maka pengguna mendapatkan 1 topik yang bisa dianggap sebagai topik yang paling banyak diambil mahasiswa untuk skripsi diasumsikan topiknya adalah irisan masing-masing obyek data pada setiap *cluster* sehingga topik tersebut adalah :

- Sistem pakar diagnosa penyakit pada tanaman

Percobaan berikutnya, pemotongan dapat dilakukan pada titik yang lebih kecil atau lebih besar. Gambar 7 memperlihatkan pemotongan *dendrogram* pada level <=0.5



Gambar 6. pemotongan *dendrogram* pada titik <=0.5

Pemotongan *dendrogram* pada titik <=0.5 menghasilkan 5 buah *cluster* seperti terlihat pada gambar 4.8 yang ditandai dengan kotak warna biru. Mengacu pada data pada tabel, maka 5*cluster* tersebut mempunyai obyek data sebagai berikut :

a. cluster 1 dengan obyek data {35. sistem informasi penjualan online pada pusat grosir elektronik cs. amalia berbasis web, 30. sistem informasi penjualan berbasis web pada toko meubel emma, 27. model basis data terdistribusi pada sistem informasi penjualan apotek, 2. rancang bangun sistem informasi akademik pada man kendal berbasis web, 23. perancangan sistem informasi persewaan mobil berbasis web pada cv. fido trans lang-lang buana semarang, 6. sistem informasi persewaan mobil pada travel kartika berbasis java}
b. cluster 2 dengan obyek data {25. sistem informasi makanan khas (kuliner) kota semarang berbasis web, 7. sistem informasi rumah sakit dr.kariadi semarang berbasis web, 1. sistem informasi rawat inap rumah sakit islam sultan agung semarang }
c. cluster 3 dengan obyek data {44. sistem pakar diagnosa gigi manusia berbasis web, 33. sistem pakar deteksi kerusakan sistem operasi windows berbasis web}
d. cluster 4 dengan obyek data {36. sistem pakar untuk mendiagnosa penyakit pada tanaman padi, 28. sistem pakar diagnosa penyakit pada tanaman tebu}
e. cluster 5 dengan obyek data {9. rancang bangun perangkat ajar flora dan fauna dengan bahasa jawa indonesia inggris untuk anak prasekolah berbasis multimedia, 8. rancang bangun perangkat ajar bahasa inggris untuk anak sekolah dasar berbasis multimedia}

Dari 5 cluster yang terbentuk, maka pengguna mendapatkan 5 topik yang bisa dianggap sebagai topik yang paling banyak diambil mahasiswa untuk skripsi diasumsikan topiknya adalah irisan masing-masing obyek data pada setiap *cluster* sehingga topik tersebut adalah :

a. Sistem Informasi
b. Sistem Informasi Semarang
c. Sistem pakar berbasis web
d. Sistem pakar diagnosa penyakit pada tanaman
e. Rancang bangun perangkat ajar anak berbasis

5. KESIMPULAN

Berdasarkan pembahasan yang telah dilakukan sebelumnya, maka dapat diambil beberapa kesimpulan sebagai berikut:

a. Kedekatan atau kemiripan diukur dengan menggunakan algoritma *dice coefficient*, jarak terdekat yang dihasilkan adalah 0.2.
b. Semakin rendah nilai *threshold* yang digunakan semakin spesifik topik judul skripsi pada sebuah cluster dan semakin tinggi nilai *threshold* maka topik judul skripsi pada sebuah cluster semakin luas. Pada penelitian ini digunakan *threshold* 0.25 dihasilkan 1 cluster dan *threshold* 0.5 dihasilkan 5 cluster
c. Nilai *threshold* dijadikan acuan pengguna yang akan menganalisa cakupan topik judul skripsi, sehingga dalam hal ini tidak ada ketentuan berapa nilai *threshold* yang terbaik.
d. Cluster yang terbentuk, merepresentasikan topik-topik yang diminati oleh mahasiswa yaitu sistem pakar.

6. SARAN

Saran-saran yang dapat diberikan untuk penelitian berikutnya adalah :

a. Untuk membentuk *cluster* dengan 44 obyek data membutuhkan waktu sekitar 2 menit, maka perlu penelitian lebih lanjut untuk mengurangi waktu eksekusi.
b. Analisa topik *judul skripsi*pada *cluster* masih manual, penelitian berikutnya dapat diterapkan analisa topik secara otomatis pada sebuah *cluster* dengan kaidah tertentu

DAFTAR PUSTAKA

[1] Han, J., Kamber, M., 2006, *Data Mining Concept and Technique*, 2nd Ed, Elsevier.
[2] Tan, Ah-Hwee, 1999, *Text Mining*: *The state of the art and the challenges, Kent Ridge Digital Labs 21 Heng Mui Keng Terrace Singapore* 119613.
[3] Hearst M, 2003, *What is Text Mining* ? http://www.sims.berkeley.edu/~hearst/ text-mining.html
[4] Even-Zohar Y, 2002. *Introduction to Text Mining*, *Supercomputing*.
[5] Ellis, D., 1996, *Progress and Problems in Information Retrieval*, *2nd ed. London: Library AssociationI.*
[6] Michael D., Gordon, 1991, *User-Based Document Clustering by Redescribing Subject Descriptions with a Genetic Algorithm,. Journal of American Society for Information Science*, 311-322.
[7] 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.
[8] Salton, G., 1989, *Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer*, *Addison – Wesly Publishing Company, Inc. All rights reserved*.

.