ANALISA *CLUSTERING* MENGGUNAKAN METODE K-MEANS DAN *HIERARCHICAL CLUSTERING* (STUDI KASUS : DOKUMEN SKRIPSI JURUSAN KIMIA, FMIPA, UNIVERSITAS SEBELAS MARET)

Lynda Rahmawati
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami 36A Kentingan
Surakarta 57126
lynda.rahmawati@gmail.com

Sari Widya Sihwi
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami 36A Kentingan
Surakarta 57126
sari.widya.sihwi@gmail.com

Esti Suryani
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami 36A Kentingan
Surakarta 57126
suryapalapa@yahoo.com

ABSTRAK

Data penelitian dapat dikelompokkan berdasarkan kemiripan tema, objek maupun metode penelitian. Hasil pengelompokkan data penelitian dapat memperlihatkan bagaimana pola kemiripan penelitian dan variasi tema penelitian dari waktu ke waktu. Hasil pengelompokan juga dapat memperlihatkan tema yang banyak diambil mahasiswa dan yang jarang diambil mahasiswa pada waktu tertentu. Informasi tersebut diharapkan dapat membantu dosen dalam mengevaluasi metode pembelajaran yang telah dilakukan. Penelitian ini mengelompokkan dokumen skripsi Jurusan Kimia, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret. Jurusan Kimia dipilih karena jumlah data penelitiannya cukup banyak.

Pengelompokan data penelitian yang umumnya berbentuk teks dapat dilakukan dengan text mining dengan metode clustering. Metode clustering yang digunakan pada penelitian ini adalah kombinasi antara metode Hierarchical Clustering dan K-Means Clustering. Data penelitian dipilih dokumen skripsi. Bagian dari dokumen yang diolah adalah bagian abstrak.

Clustering dokeman menghasilkan 16 cluster. Hasil cluster dianalisa keterkaitan antar dokumennya dan diperkirakan tema dari tiap cluster. Hasil cluster dilihat pula keterkaitannya dengan dosen yang mengajar Jurusan Kimia. Hasil analisa cluster memperlihatkan bahwa keahlian dosen mempengaruhi variasi tema penelitian yang dilakukan oleh mahasiswa. Diketahui pula bahwa banyaknya penelitian di suatu tema berkaitan dengan minat siswa dan proyek dosen di Jurusan Kimia.

Kata kunci : *Abstrak Skripsi, Clustering, Hierarchical Clustering, K-Means Clustering, Text Mining.*

1. PENDAHULUAN

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. Hasil pengelompokan juga dapat memperlihatkan materi yang banyak diambil mahasiswa dan yang jarang diambil mahasiswa pada waktu tertentu. Informasi tersebut diharapkan dapat membantu dosen dalam mengevaluasi metode pembelajaran yang telah dilakukan pada materi yang banyak ataupun sedikit diambil sebagai bahan penelitian mahasiswa.

Penelitian ini mengelompokkan dokumen skripsi Jurusan Kimia, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sebelas Maret. Jurusan Kimia dipilih karena jumlah data penelitiannya cukup banyak. Setiap tahunnya Jurusan Kimia 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 [2]. Salah satu metode *clustering* adalah K-Means. K-Means mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien [3]. Akan tetapi, hasil *clustering* dengan K-Means sangat bergantung pada pusat awal *cluster*. Hasil *clustering* dengan metode K-Means baik jika penentuan pusat *cluster* tepat. Metode *Hierarchical Clustering* dapat digunakan untuk mengatasi masalah penentuan pusat *cluster* pada K-Means. Penelitian ini mengkombinasikan K-Means dengan *Hierarchical Clustering*. Hasil dari *Hierarchical Clustering* akan digunakan dalam penentuan pusat awal *cluster K-means clustering*. Kombinasi antara metode *Hierarchical Clustering* dan *K-Means Clustering* ini telah diuji oleh Alfina [4] dan terbukti bahwa kombinasi ini lebih baik dibandingkan K-Means.

2. LANDASAN TEORI

2.1 *Text mining*

Text mining adalah proses menemukan hal baru, yang sebelumnya tidak diketahui, mengenai informasi yang berpotensi untuk diambil manfaatnya dari sumber data yang tidak terstruktur mencakup dokumen bisnis, komentar customer, halaman web dan file XML [5].

Text mining hampir sama dengan data mining dalam hal tujuan dan proses, tapi pada *text mining* inputnya adalah file data tidak terstruktur seperti dokumen dalam bentuk word, PDF, text, XML dan sebagainya [6]. *Text mining* dapat digunakan dalam beberapa hal yaitu ekstraksi informasi, *topic tracking*, *summarization*, kategorisasi dan *clustering*.

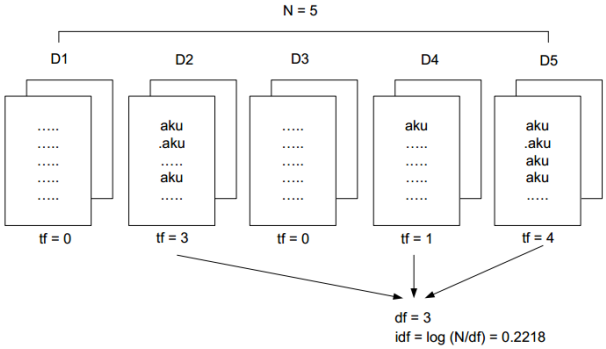
2.2 Text *Preprocessing*

Text processing berfungsi mengubah data tekstual yang tidak terstruktur ke dalam data terstruktur dan disimpan dalam basis data [2]. Tahap *preprocessing* terdiri dari beberapa langkah yaitu : *case floding*, *tokenisasi*, *filtering* dan *stemming*.

Proses *case floding* menghilangkan karakter selain huruf dan mengubah semua huruf menjadi *lowercase*. Proses *tokenisasi* memotong data awal yang berupa kalimat menjadi kata. Data hasil dari proses *tokenisasi* dilanjutkan dengan proses *filtering*. Proses *filtering* mengambil kata-kata penting dari hasil proses *tokenisasi*. Langkah proses ini bisa dilakukan dengan dua teknik yaitu *stop list* (membuang kata yang kurang penting) *dan word list* (menyimpan kata yang penting). Data hasil *filtering* kemudian diolah dengan *stemming*. Tahap *stemming* adalah tahap mencari root kata dari tiap kata hasil *filtering*. Abstrak penelitian yang diolah dalam penelitian ini adalah abstrak yang berbahasa Indonesia. Algoritma stemming untuk bahasa Indonesia yang digunakan adalah algoritma Nazief-Adriani [7]

2.3 *Term Weighting* dengan *Term Frequency* (tf) – Inverse Document Frequency (idf)

Pembobotan *Term* (*Term Weighting*) bertujuan untuk menentukan bobot setiap *term*. Perhitungan bobot *term* memerlukan dua hal yaitu *Term Frequency* (tf) dan *Inverse Document Frequency* (idf). *Term Frequency* (tf) merupakan frekuensi kemunculan suatu kata (*term*) dalam suatu dokumen. Nilai tf bervariasi ditiap dokumen bergantung pada kemunculan kata di suatu dokumen. Besar nilai tf sebanding dengan tingkat kemunculan *term* di dokumen. Semakin sering *term* muncul pada suatu dokumen, semakin besar pula nilai tf pada dokumen tersebut dan semakin jarang *term* muncul semakin kecil pula nilai tf. Selain *Term Frequency* diperlukan pula *Inverse Document Frequency* (idf) pada pembobotan *term*. *Inverse Document Frequency* (idf) merupakan frekuensi kemunculan *term* pada keseluruhan dokumen. Nilai idf berkaitan dengan distribusi *term* di berbagai dokumen. Nilai idf berbanding terbalik dengan jumlah dokumen yang mengandung. *Term* yang jarang muncul pada keseluruhan dokumen memiliki nilai idf lebih besar dibanding dengan *term* yang sering muncul. Jika setiap dokumen dalam koleksi mengandung *term* yang bersangkutan, maka nilai idf dari *term* tersebut adalah nol (0). Hal ini menunjukkan bahwa setiap *term* yang muncul pada dokumen dalam koleksi tidak berguna untuk membedakan dokumen berdasarkan topik tertentu. Ilustrasi algoritma tf-idf ditunjukkan pada Gambar 1 :



Gambar 1. Ilustrasi algoritma tf-idf [8]

Keterangan :
D1, …, D5 = dokumen
tf = banyaknya *term* yang dicari pada setiap dokumen
N = total dokumen
df = banyaknya dokumen yang mengandung *term* yang dicari

Persamaan dalam menghitung nilai tf-idf adalah [9]:

$$W_{i,j} = tf_{ij} \times idf_j = tf_{i,j} \times \log\left(\frac{N}{df_j}\right) \quad \text{.....(1)}$$

Dimana :

$W_{i,j}$ = bobot *term* ke-j terhadap dokumen ke-i
tf_{ij} = jumlah kemunculan *term* j ke dalam dokumen i
N = jumlah dokumen secara keseluruhan
df_j = jumlah dokumen yang mengandung *term* j

Perhitungan bobot dari *term* tertentu dalam sebuah dokumen dengan menggunakan tf x idf menunjukkan bahwa deskripsi terbaik dari dokumen adalah *term* yang banyak muncul dalam dokumen tersebut dan sangat sedikit muncul pada dokumen lain [9].

2.4 *Clustering*

Clustering adalah proses mengelompokkan atau penggolongan objek berdasarkan informasi yang diperoleh dari data yang menjelaskan hubungan antar objek dengan prinsip untuk memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas/*cluster* [10]. *Clustering* membagi data ke dalam grup-grup yang mempunyai obyek yang karakteristiknya sama.

2.4.1 *Hierarchical clustering*

Hierarchical clustering adalah metode analisis kelompok yang berusaha untuk membangun sebuah hierarki kelompok. *Hierarchical clustering* dibagi menjadi dua yaitu *Agglomeratif Clustering* dan *Difisive Clustering*. *Agglomeratif Clustering* mengelompokkan data dengan pendekatan bawah atas (*bottom up*), sedangkan *Difisif Clustering* menggunakan pendekatan atas bawah (*top-bottom*).

Metode hierarchical *agglomeratif clustering*, mengasumsikan setiap data yang ada sebagai *cluster* di awal proses. Jika jumlah data adalah n, dan jumlah *cluster* adalah k, maka besarnya n = k. Kemudian dihitung jarak antar *cluster*-nya dengan menggunakan *Euclidean distance* berdasarkan jarak rata-rata antar objek. Selanjutnya, dari hasil perhitungan jarak dipilih jarak yang paling minimal dan digabungkan

sehingga besarnya n = n -1. Ketika dua *cluster* digabungkan, jarak antara dua *cluster* yang digabungkan dengan *cluster* yang lain di-update. Penggabungan *cluster* akan terus dilakukan dan akan berhenti jika memenuhi kondisi jumlah k = 1. Pada akhir tahap hierarchical *clustering* diperoleh dendrogram yang menunjukkan urutan pengelompokan masing-masing anggota dalam *cluster*.

Penelitian ini menggunakan metode ward sebagai metode *update* jarak. Metode Ward dapat membentuk *cluster* berdasarkan jumlah total kuadrat deviasi tiap pengamatan dari rata-rata *cluster* yang menjadi anggotanya [11]. Metode Ward berusaha untuk meminimalkan variasi antar objek dalam satu *cluster* dan memaksimalkan variasi dengan objek yang ada di *cluster* lainnya. Jarak antara dua *cluster* yang terbentuk pada metode Ward adalah sum of squares diantara dua *cluster* tersebut. Metode Ward didasarkan pada kriteria *sum square error* (SSE) dengan ukuran kehomogenan antara dua objek berdasarkan jumlah kuadrat kesalahan minimal. Perhitungan pada metode ward menggunakan rumus berikut :

$$I_{(uv)w} = \frac{n_u+n_w}{n_{uv}+n_w} I_{uw} + \frac{n_v+n_w}{n_{uv}+n_w} I_{vw} - \frac{n_w}{n_{uv}+n_w} I_{uv} \text{.....(2)}$$

Dengan u dan v *cluster* yang digabung, w *cluster* lain yang dicari jaraknya dengan *cluster* gabungan uv, $I_{(uv)w}$ jarak antara *cluster* uv dan *cluster* w, I_{uw} jarak antara *cluster* u dan *cluster* w, I_{vw} jarak antara *cluster* v dan *cluster* w, I_{uv} jarak antara *cluster* u dan *cluster* v, n_u, n_v, n_w dan adalah banyaknya objek pada *cluster* ke-u, ke-v dan ke-w.

2.4.2 *K-means clustering*

K-Means adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu *cluster* dan memaksimalkan variasi dengan data yang ada di *cluster* lainnya [12].

Menurut Sarwono [13], Berikut adalah langkah-langkah dari algoritma K-Means:

1. Menentukan banyak k-*cluster* yang ingin dibentuk.
2. Membangkitkan nilai random untuk pusat *cluster* awal (*centroid*) sebanyak k-*cluster*.
3. Menghitung jarak setiap data input terhadap masing-masing *centroid* menggunakan rumus jarak Eucledian (Eucledian Distance) hingga ditemukan jarak yang paling dekat dari setiap data dengan *centroid*. Berikut adalah persamaan Eucledian Distance:

   $$d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \text{.....(3)}$$

   dengan $d(x_i, \mu_i)$ adalah jarak antara *cluster* x dengan pusat *cluster* μ pada kata ke $_i$, x_i adalah bobot kata ke $_i$ pada *cluster* yang ingin dicari jaraknya, μ_i bobot kata ke $_i$ pada pusat *cluster*.
4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan *centroid* (jarak terkecil).
5. Mengupdate nilai *centroid*. Nilai *centroid* baru diperoleh dari rata-rata *cluster* yang bersangkutan dengan menggunakan rumus:

   $$C_k = \frac{1}{n_k} \sum d_i \text{.....(4)}$$

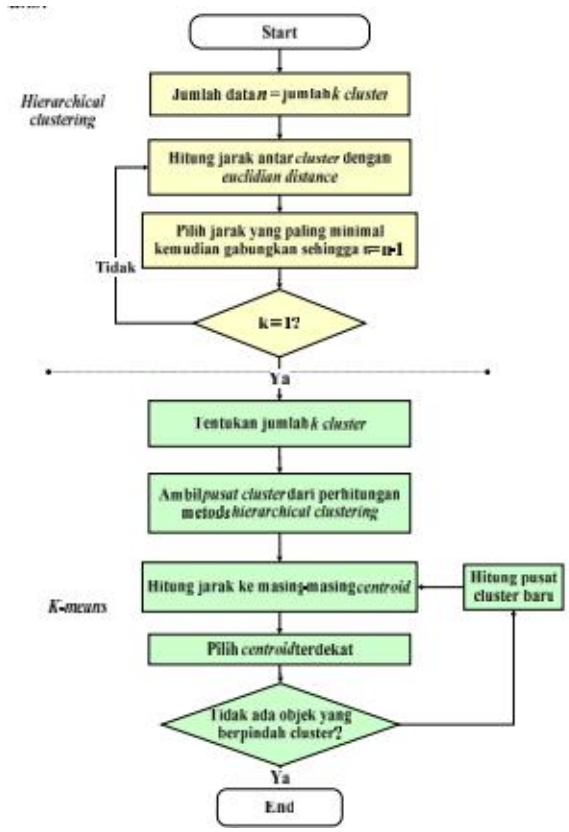
   dimana:

   n_k = jumlah data dalam *cluster*

   d_i = jumlah dari nilai jarak yang masuk dalam masing-masing *cluster*
6. Melakukan perulangan dari langkah 2 hingga 5 hingga anggota tiap *cluster* tidak ada yang berubah.
7. Jika langkah 6 telah terpenuhi, maka nilai rata-rata pusat *cluster* (μj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

2.4.3 Metode Gabungan *Hierarchical Clustering* dengan *K-means Clustering*

Penggabungan metode *Hierarchical clustering* dengan *K-means clustering* dimaksudkan agar hasil *clustering* lebih baik. Hasil dari metode *Hierarchical clustering* digunakan untuk menentukan pusat *cluster*. Pusat *cluster* yang dihasilkan *Hierarchical clustering* selanjutnya digunakan sebagai pusat *cluster* awal pada perhitungan *K-means clustering*. Gambar 2 menggambarkan proses *clustering* menggunakan kombinasi antara metode *Hierarchical clustering* dengan *K-means clustering*.



Gambar 2. Kombinasi metode *Hierarchical clustering* dan *K-means clustering* [5]

3. METODOLOGI

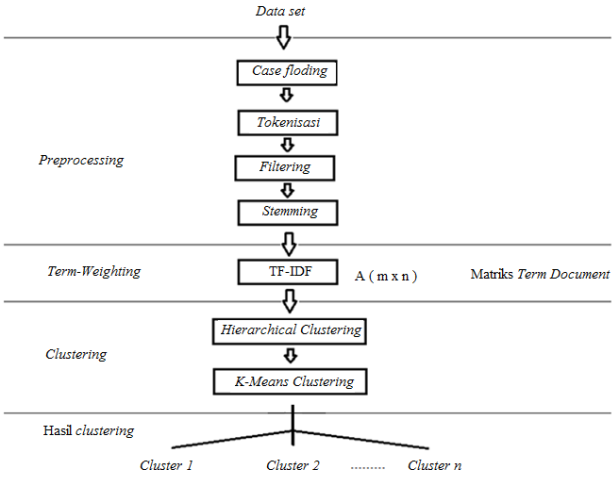
Objek pada penelitian ini adalah data skripsi mahasiswa strata-1 Jurusan Kimia, Fakultas Matematika dan Ilmu Pengetahuan, Universitas Sebelas Maret. Data dikelompokkan dengan *clustering*. Setelah *cluster* terbentuk dilakukan analisa hasil *clustering*

3.1 Data Set

Data penelitian yang digunakan bersumber dari portal UNS yaitu http://www.digilib.uns.ac.id. Dipilih data skripsi dari Jurusan Kimia, Fakultas MIPA, UNS. Data yang dipilih adalah data yang terbit pada periode tahun 2009-2013. Data yang diolah dalam *clustering* adalah bagian abstrak yang berbahasa Indonesia.

3.2 Alur Proses *Clustering*

Alur proses *clustering* yang digunakan pada penelitian ini digambarkan oleh Gambar 3.



Gambar 3. Alur proses pengolahan data

- Tahap *Preprocessing*

Proses pengolahan data dimulai dengan *preprocessing*. Pada tahap *preprocessing* terdapat empat proses yaitu *case floding*, *tokenisasi*, *filtering*, dan *stemming*. Input pada tahap *preprocessing* adalah bagian abstrak pada skipsi yang berupa paragraf-paragraf. *Preprocessing* menghasilkan output berupa *bag-of-word* yaitu matriks berisi kata-kata yang diolah dalam penelitian.

- Tahap *Term Weighting*

Setelah bag-of-word diperoleh proses dilanjutkan dengan *term weighting*. Tahap *term weighting* menghitung nilai bobotnya setiap kata dengan TF-IDF. Hasil dari proses ini adalah *term-weight-matrix* yaitu matriks yang berisi bobot-bobot kata pada dokumen-dokumen.

- Tahap *Clustering*

Term-weight-matrix diolah dengan *Hierarchical clustering*. Dua data dengan jarak terkecil digabung menjadi satu *cluster*, dan proses terus diulang sampai jumlah *cluster* = 1. Hasil dari *hierarchical clustering* berupa dendogram.

K-Means mengolah *term-weight-matrix* dengan mengelompokkan data berupa dokumen ke dalam *cluster-cluster*. Penentuan pusat *cluster* menggunakan hasil dari *Hierarchical clustering*. Sejumlah *cluster* yang memuat beberapa dokumen diambil dari dendogram hasil *Hierarchical clustering* dan dihitung nilai rata-rata dokumen di tiap *cluster* yang diambil. Nilai rata-rata dokumen ini digunakan sebagai pusat *cluster* dalam perhitungan *K-means clustering*. Hasil dari *K-means clustering* adalah daftar *cluster* dari data-data yang diolah.

3.3 Tahap Analisa dan Validasi Hasil *Clustering*

Tahap analisa dan validasi melakukan analisa dan validasi hasil *clustering* dengan mengamati hasil dari *cluster-cluster* yang terbentuk. Setiap dokumen dilihat keterkaitannya dengan dokumen lain dalam satu *cluster* dan ditentukan tema pada tiap *cluster*. Hasil *clustering* dibandingkan terhadap variabel tahun dan dibandingkan jumlah dokumen pada tiap *cluster*. Hasil *clustering* dianalisa dan divalidasi oleh dua pakar yaitu Pakar 1, Komisi Tugas Akhir Jurusan Kimia, FMIPA, UNS, Teguh Endah Saraswati, M.Sc., Ph.D dan Pakar 2, Ketua Laboratorium Kimia, FMIPA, UNS, Dr. Sayekti Wahyuningsih, M.Si.

4. HASIL DAN PEMBAHASAN

4.1. Data Set

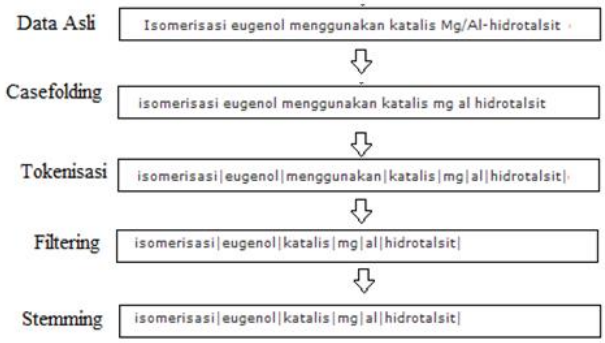
Data penelitian diperoleh dari database digilib UNS. Data skripsi Jurusan Kimia berjumlah 223 dengan rentang tahun 2003-2013 disortir sehingga diperoleh data berjumlah 161 dengan rentang waktu tahun 2009-2013. Selain dari segi tahun penyortiran data juga dilakukan jika ada ada yang tidak lengkap abstraknya ataupun jika ada data dengan abstrak bahasa Inggris.

4.2 Proses *Clustering* Dokumen

Proses *clustering* dokumen dilakukan dengan melalui *preprocessing data*, *term-weighting*, dan *clustering data*.

- *Preprocessing*

Proses *preprocessing* pada tahap ini dilakukan dengan empat bagian tahapan yaitu *case floding*, *tokenisasi*, *filtering*, dan *stemming*. Gambaran dari proses tahapan *preprocessing* ditunjukkan oleh Gambar 4.



Gambar 4. Tahap *Preprocessing*

Gambar 4 menunjukkan tahapan *preprocessing* data dari data skripsi berjudul "Isomerisasi eugenol menggunakan mg/al-hidrotalsit dengan radiasi gelombang mikro". Hasil proses *stemming* kadang-kadang masih memuat kata yang *term*asuk *stopword*. Oleh karena itu setelah *stemming* dilakukan *filtering* lagi sehingga tidak ada *stopword* pada *bag-of-word*.

- *Term-Weighting*

Proses *term-weighting* menggunakan tf-idf. Proses tahap *term-weighting* dimulai dengan mendaftar seluruh *term* pada seluruh dokumen. Kemudian proses dilanjutkan dengan menghitung frekuensi setiap *term* pada setiap dokumen (tf). Selanjutnya proses dilanjutkan dengan menghitung jumlah dokumen yang memuat *term* (df). Proses dilanjutkan dengan

menghitung *Inverse Document Frequency (idf)* dan bobot *term* (w) dengan rumus (1). Perhitungan *term-weighting* pada sistem ini ditunjukkan pada Tabel 1.

Tabel 1. Tahap *Term-Weighting*

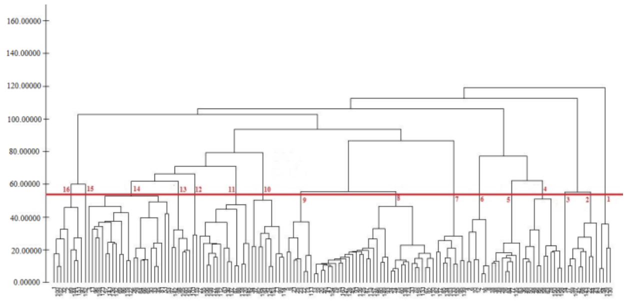
Keyword	tf	df	idf	tf-idf
stabil	3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 5 0	11	1.17	3.51 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2.34 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.17 0 0 0 0 0 0 0 0 1.17 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.17 2.34 0 0 0 0 0 0 0 0 0 1.17 0 0 0 0 0 0 0 0 2.34 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1.17 1.17 0 0 0 0 0 5.85 0

Tabel 1 menunjukkan hasil perhitungan tf-idf untuk kata "stabil" pada 161 dokumen. Kolom tf menunjukkan array yang berisi frekuensi kata "stabil" pada 161 dokumen. Kolom df menunjukkan jumlah dokumen yang memuat kata "stabil". Kolom idf menunjukkan hasil perhitungan *inverse* dari df. Kolom tf-idf menunjukkan hasil perkalian dari tf dengan idf. Kolom tf-idf menunjukkan array yang berisi bobot kata "stabil" pada 161 dokumen.

Proses *feature selection* dengan *df feature selection* dilakukan setelah diperoleh *term-weight-matrix*. *Df feature selection* ini berfungsi untuk membatasi *term* yang diolah dalam proses *clustering*. Df adalah jumlah dokumen yang mengandung suatu *term*. Nilai df yang kecil menunjukkan *term* jarang muncul sedangkan nilai df yang besar menunjukkan *term* yang sering muncul. Pembatasan nilai df dilakukan dengan adanya asumsi bahwa *term* yang muncul dalam sedikit dokumen tidak memiliki pengaruh yang besar dalam proses *clustering* dokumen. Sebaliknya jika suatu *term* muncul dalam banyak dokumen, maka *term* tersebut mempunyai tingkat kepentingan yang lebih kecil karena *term* tersebut dapat dianggap sebagai *term* umum. *Df feature selection* pada penelitian ini menggunakan *min threshold* 3 dan *max threshold* 40.

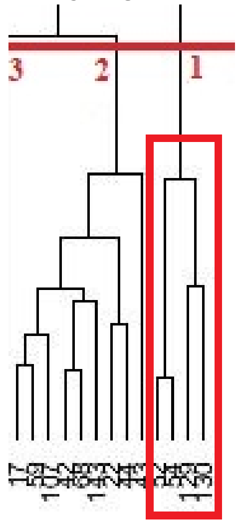
- *Clustering*

Hasil dari *tem-weighting* berupa *term-weight matrix* selanjutnya diolah dengan *hierarchical clustering* dan *k-means clustering*. Tahap awal *hierarchical clustering* menganggap setiap dokumen sebagai *cluster*. Proses *clustering* pada *hierarchical clustering* dimulai dengan menghitung jarak antar dokumen. Hasil perhitungan jarak kemudian digunakan oleh sistem untuk melakukan proses *clustering* dengan *hierarchical clustering*. Hasil dari proses *hierarchical clustering* diGambarkan dengan dendrogram yang. Gambar dendrogram dipotong, sehingga menghasilkan 16 *cluster* seperti ditunjukkan pada Gambar 5.



Gambar 5. Dengrogram Hasil Hierarchical *Clustering*

Proses penghitungan pusat *cluster* diperoleh dari hasil pemotongan dendrogram. Sebagai contoh, dari Gambar 6 pusat *cluster* 1 diperoleh dari perhitungan rata-rata bobot *term* dokumen ber-id 52, 54, 129, dan 130. Pusat *cluster* untuk 16 *cluster* dihitung. Pusat *cluster* ini digunakan sebagai pusat *cluster* awal pada *clustering* dengan metode *k-means clustering.*



Gambar 6. Penentuan Pusat *Cluster*

Setelah pusat *cluster* diperoleh, sistem melanjutkan proses *clustering* dengan kmeans *clustering*. Hasil dari proses *kmeans clustering* diperoleh *cluster-cluster* yang terdiri dari dokumen-dokumen. Gambar 7 memperlihatkan tampilan hasil dari proses *clustering*. Kolom 'Id Dokumen' menunjukkan nomor id dari dokumen. Kolom 'Tahun' menunjukkan tahun terbit dokumen. Kolom 'Judul' menunjukkan judul dokumen. Kolom 'Cluster' menunjukkan *cluster* dokumen.

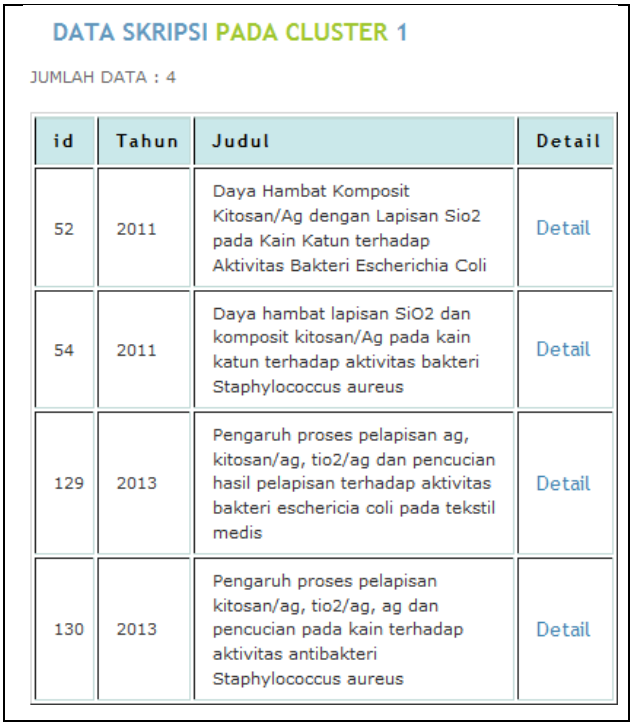
Id Dokumen	Tahun	Judul	Cluster
1	2009	Perubahan Sifat Fisik dan Kimia Pada Uji Kestabilan Panas serta Biodegradasi Biokomposit Polipropilena Daur Ulang Dengan Serbuk Sekam Padi	16
2	2009	Sintesis biokomposit polipropilena daur ulang termodifikasi secara reaktif dengan penguat serbuk bambu ukuran partikel 80 dan 150 mesh	16
3	2009	Perlakuan NH4CL dan gelombang mikro terhadap karakter keasaman montmorillonit	14

Gambar 7. Tampilan Hasil Clustering

4.3 Analisa dan Validasi Hasil *Clustering*

Proses *clustering* mengelompokkan dokumen yang mirip ke suatu kelompok berdasarkan kesamaan *term* yang muncul pada bagian abstrak. Di akhir proses *clustering* diperoleh *cluster-cluster* yang memuat dokumen yang mirip. Ada 16 *cluster* yang terbentuk dari proses *clustering*. Banyak *cluster* sejumlah 16 dipengaruhi oleh gambar dendrogram hasil hierarchical *clustering* yang ditunjukkan oleh Gambar 5. Pemisahan *cluster* dengan *threshold* yang ditunjukkan oleh Gambar 5 mengakibatkan jumlah *cluster* 16. Dipilihnya *threshold* tersebut melalui pertimbangan keterkaitan antar dokumen pada satu *cluster*. Ketika *threshold* dinaikkan, ada *cluster* yang bergabung dengan *cluster* lain. Penggabungan *cluster* ternyata mengakibatkan adanya dokumen-dokumen dengan tema berbeda masuk ke dalam satu *cluster*. Sedangkan, ketika *threshold* diturunkan, ada *cluster* dengan dokumen-dokumen yang memiliki tema yang sama terpisah. Karena itulah *threshold* tersebut dipilih.

Gambar 8 menunjukkan tampilan hasil *clustering* dokumen pada *cluster* 1. Kolom 'Id' menunjukkan id dokumen. Kolom 'Tahun' menunjukkan tahun skripsi. Kolom 'Judul' menunjukkan judul skripsi. Kolom 'Detail' berisi *link* menuju halaman detail data yang menunjukkan detail dokumen.



DATA SKRIPSI PADA CLUSTER 1

JUMLAH DATA : 4

id	Tahun	Judul	Detail
52	2011	Daya Hambat Komposit Kitosan/Ag dengan Lapisan Sio2 pada Kain Katun terhadap Aktivitas Bakteri Escherichia Coli	Detail
54	2011	Daya hambat lapisan SiO2 dan komposit kitosan/Ag pada kain katun terhadap aktivitas bakteri Staphylococcus aureus	Detail
129	2013	Pengaruh proses pelapisan ag, kitosan/ag, tio2/ag dan pencucian hasil pelapisan terhadap aktivitas bakteri eschericia coli pada tekstil medis	Detail
130	2013	Pengaruh proses pelapisan kitosan/ag, tio2/ag, ag dan pencucian pada kain terhadap aktivitas antibakteri Staphylococcus aureus	Detail

Gambar 8. Tampilan Hasil *Clustering*

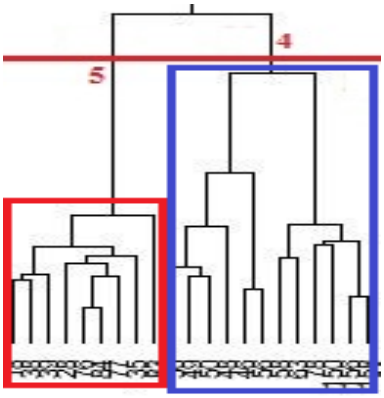
Setelah hasil *clustering* diperoleh, proses dilanjutkan dengan analisa hasil *clustering*. Proses analisa yang pertama dilakukan dengan meneliti pola keanekaragaman tema dibandingkan dengan dosen yang mengajar jurusan kimia. Berdasarkan judul dokumen di setiap *cluster*, Pakar 1 memperkirakan tema dan dosen yang membimbing mahasiswa yang meneliti tema. Tabel 2 memperlihatkan hasil analisa tema dan dosen pembimbing pada setiap *cluster*.

Tabel 2. Analisa Tema dan Dosen Pembimbing

No	Tema *Cluster*	Dosen Pembimbing
1	Analit	Candra Purnawan, M.Sc./ Dr. Desi Suci Handayani, M.Si.
2	Anorganik Kompleks	Prof. Drs Sentot Budi Rahardjo, Ph.D
3	Anorganik Kompleks	Prof. Drs Sentot Budi Rahardjo, Ph.D
4	Organik, Dengan fokus Isolasi, kromatografi, identifikasi dan antibakteri	Gabungan dosen dari jurusan kimia dan farmasi
5	Organik, Dengan fokus Isolasi, kromatografi, identifikasi dan antibakteri	Gabungan dosen dari jurusan kimia dan farmasi
6	Organic, Dengan fokus Antioksidan dan antijamur	Gabungan dosen dari jurusan kimia dan farmasi
7	Kimia Fisik, Dengan fokus pada preparasi material hydrotalcite	Dr. Eddy Heraldy, M.Si
8	Belum Spesifik	-
9	Kimia Fisik, dengan fokus aplikasi material	Dr. Eddy Heraldy, M.Si
10	Komputasi	Dr rer. nat. Fajar Rakhman W, M.Si.
11	Anorganik,	Dr. Sayekti Wahyuningsih, M.Si.
12	Organik, Fokus pada itakonat	Dr.rer.nat. Atmanto Heru Wibowo, M.Si.
13	Anorganik membran polimer	Dr. Eddy Heraldy, M.Si
14	Belum Spesifik	-
15	Anorganik polimer film	Prof. Dra Neng Sri Suharty, MS, PhD
16	Anorganik polimer film	Prof. Dra Neng Sri Suharty, MS, PhD

Tabel 2 menunjukkan perkiraan tema dan dosen pembimbing untuk tiap *cluster*. Tabel 2 memperlihatkan adanya *cluster-cluster* yang memiliki tema yang sama tapi terpisah menjadi *cluster* yang berbeda. *Cluster* 2 dan *cluster* 3 terpisah karena objek penelitian dari *cluster* 3 berbeda dengan objek pada penelitian dari *cluster* 2. *Cluster* 4 dan *cluster* 5 terpisah karena adanya perbedaan objek penelitian. Penelitian pada *cluster* 5 memiliki objek yang sama yaitu 'buah merah', sedangkan pada *cluster* 4 objek yang diteliti beragam. Oleh sebab itu jarak antara dokumen-dokumen pada *cluster* 5 lebih kecil dibandingan pada *cluster* 4. Perbedaan jarak pada *cluster* 4 dan 5 dapat dilihat pada Gambar 9. Besarnya jarak antar cluster ditunjukkan dengan tinggi garis antar dokumen pada *cluster* 4 dan *cluster* 5 pada Gambar 9. Semakin tinggi garis, semakin besar pula jarak antar dokumen. Gambar 9 menunjukkan bahwa jarak antar *cluster* pada *cluster* 5 lebih kecil daripada *cluster* 4. *Cluster* 15 terpisah dengan *cluster* 16 disebabkan oleh adanya perbedaan dalam metode penelitian. Pada *cluster* 15 terdapat

penambahan kaolin pada penelitian, sedangkan pada *cluster* 16 tidak terdapat penambahan kaolin.



Gambar 9. Perbedaan Jarak pada *Cluster* 4 dan *Cluster* 5

Tabel 2 menunjukkan keterkaitan antara tema dengan dosen pembimbing penelitian. Setiap tema umumnya dibimbing oleh dosen yang ahli di tema tersebut. Semakin beragam keahlian dosen, tema yang diteliti mahasiswa juga semakin beragam. Hal ini menyebabkan variasi tema penelitian berkaitan erat dengan keahlian dan minat dosen. Telah dilakukan konfirmasi dengan Pakar 1, dan ternyata memang ada pengaruh antara dosen dengan variasi penelitian. Keahlian dosen yang mengajar mempengaruhi variasi tema penelitian yang diambil mahasiswa.

Pola analisa hasil *clustering* yang kedua dilakukan dengan membandingkan banyaknya dokumen di suatu *cluster* dari tahun ke tahun.

Tabel 3. Representasi *Cluster* Pertahun

Cluster	**2009**	**2010**	**2011**	**2012**	**2013**	**Jumlah**
1	0	0	2	0	2	4
2	4	1	2	1	1	9
3	1	0	0	0	0	1
4	0	5	4	0	3	12
5	1	3	4	0	1	9
6	7	0	0	0	0	7
7	0	0	2	3	4	9
8	5	8	5	11	8	37
9	4	1	2	1	0	8
10	0	2	1	2	3	8
11	0	1	4	3	7	15
12	0	0	0	2	0	2
13	0	0	0	3	2	5
14	2	7	8	3	5	25
15	0	0	0	0	1	1
16	2	1	0	4	2	9

Tabel 3 menunjukkan naik turunnya jumlah penelitian di tiap *cluster* pada tahun 2009-2013. Setelah dilakukan validasi dengan Pakar 1 diketahui bahwa naik turunnya jumlah dokumen tersebut kemungkinan disebabkan oleh dua hal, yaitu :

- Minat mahasiswa.

Pada tahun 2009-2013 belum ada Komisi Tugas Akhir yang mengatur pemerataan bimbingan dosen. Mahasiswa mengambil tema penelitian berdasarkan keinginannya, sehingga banyaknya penelitian pada suatu tema tidak konsisten.

- Proyek dosen

Proyek dosen berpengaruh pada tema penelitian yang diambil mahasiswa. Hal ini disebabkan oleh adanya biaya yang disediakan pada proyek dosen, sedangkan penelitian di luar proyek dosen dibiayai secara mendiri oleh mahasiswa. Oleh karena itu, proyek dosen lebih banyak dipilih oleh mahasiswa.

Hal yang sama juga diungkapkan oleh Pakar 2. Pakar 2 menyatakan bahwa penelitian yang diambil oleh mahasiswa memang berkaitan dengan tema-tema yang didanai dan berkaitan pula dengan minat dosen pada suatu tema.

Tahun 2009 proyek dosen berkisar pada tema anorganik kompleks, organik dengan fokus antijamur dan antioksidan, dan kimia fisik. Proyek dosen tersebut mengakibatkan banyaknya mahasiswa yang mengambil penelitian dengan tema tersebut. Hal ini dapat dilihat dari Tabel 2 kolom ke-2 yang menunjukkan banyaknya dokumen yang terbit pada tahun 2009 di *cluster* 2 dengan tema anorganik kompleks, *cluster* 6 dengan tema organik dan *cluster* 9 dengan tema kimia fisik. Tahun 2010 proyek dosen bertema organik, sehingga dokumen pada *cluster* 4 dan *cluster* 5 dengan tema organik berjumlah banyak. Proyek dengan tema organik juga dilanjutkan di tahun 2011. Dokumen yang bertema organik yaitu pada *cluster* 4 dan *cluster* 5 berjumlah banyak pada tahun 2011. Selain tema organik, pada tahun 2011 proyek dosen juga ada yang mengambil tema anorganik. Jumlah dokumen pada *cluster* 11 dengan tema anorganik meningkat dibanding tahun 2009 dan 2010. Pada tahun 2012 dan 2013 ada banyak dosen yang memiliki proyek dengan tema yang beragam, sehingga tema penelitian mahasiswa menjadi lebih beragam. Keberagaman penelitian mahasiswa pada tahun 2013 dapat dilihat pada Tabel 2 kolom tahun 2012 dan 2013 yang menunjukkan banyaknya jumlah dokumen di beberapa *cluster*, antara lain *cluster* 7, *cluster* 10, *cluster* 11, *cluster* 13 dan *cluster* 15.

Hasil penelitian memuat dua *cluster* dengan tema yang belum spesifik yaitu *cluster* 8 dan *cluster* 14. Menurut Pakar 2, hal ini kemungkinan disebabkan oleh diprosesnya seluruh kata pada bagian abstrak. Disarankan agar pada penelitian selanjutnya data yang diolah tidak hanya bagian abstrak, tetapi data bagian Bab 2 atau Bab 3 dokumen skripsi. Data pada bagian abstrak hanya gambaran kecil dari penelitian dan kata-kata yang ada di bagian abstrak kurang dapat merepresentasikan penelitian. Bab 2 dan Bab 3 dinilai lebih dapat merepresentasikan penelitian. Selain itu, kata yang diolah pada proses *clustering* sebaiknya dibatasi. Kata yang diolah seharusnya hanya kata-kata yang signifikan di Jurusan Kimia, sehingga hasil *clustering* bisa lebih baik lagi.

5. PENUTUP

5.1. Kesimpulan dan saran

Penelitian *clustering* dokumen dengan kombinasi metode *hierarchical clustering* dan *k-means clustering* ini sudah cukup menggambarkan keterkaitan antar dokumen. Hasil *clustering* menunjukkan adanya dokumen yang sejenis, yang

merepresentasikan kemiripan antar dokumen. Akan tetapi, penggunaan semua kata pada bagian abstrak kurang tepat. Penggunaan seluruh kata pada dokumen bisa mengakibatkan masuknya dokumen-dokumen yang berbeda tema ke dalam satu *cluster* karena yang diproses adalah kata yang tidak signifikan.

Hasil analisa *cluster* menunjukkan bahwa penelitian di Jurusan Kimia pada tahun 2009-2012 terbatas pada beberapa tema. Tahun 2009 tema yang banyak diteliti adalah anorganik kompleks, organik dengan fokus antijamur dan antioksidan, dan kimia fisik. Tahun 2010 tema yang banyak diteliti adalah tema organik. Tahun 2011 tema yang diteliti adalah arganik dan anorganik. Pada tahun 2012 dan 2013 tema penelitian di Jurusan Kimia lebih bervariasi daripada tahun-tahun sebelumnya. Hasil analisa *cluster* dokumen skripsi Jurusan Kimia, FMIPA, UNS memperlihatkan bahwa keahlian dosen sangat mempengaruhi variasi tema penelitian yang dilakukan oleh mahasiswa. Diketahui pula bahwa banyaknya penelitian di suatu tema berkaitan dengan minat siswa dan proyek dosen di Jurusan Kimia. Variasi tema proyek dosen mempengaruhi variasi tema penelitian mahasiswa.

Hasil analisa ini dapat dimanfaatkan oleh Jurusan Kimia sebagai pertimbangan apabila jurusan ingin mengembangkan variasi tema penelitian. Hasil analisa juga dapat dimanfaatkan sebagai tolak ukur minat mahasiswa. Selain itu, hasil analisa juga dapat dipakai sebagai kontrol proyek dosen sebagai masukan mengenai proyek yang perlu diteliti lebih lanjut dan proyek yang sudah terlalu banyak diteliti.

5.2. Saran

Untuk pengembangan lebih lanjut akan lebih baik apabila *clustering* dokumen menggunakan materi pada Bab 2 atau Bab 3 yang lebih dapat menggambarkan keseluruhan dokumen. Pemilihan kata sebaiknya terbatas pada kata-kata kunci yang signifikan pada penentuan jenis penelitian yang dilakukan.

5. REFERENCES

[1] Langgeni, D. P., Baizal, ZK. and Firdaus, A.W. 2010. *Clustering* Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. Seminar Nasional Informatika 2010 (semnasIF 2010) ISSN: 1979-2328. Yogyakarta

[2] Handoyo, R. 2013. Perbandingan Metode *Clustering* Menggunakan Metode Single Linkage dan K-Means pada Pengelompokan Dokumen. Proposal Tugas Akhir Institut Teknologi Telkom. Bandung

[3] Arai, K., Barakbah, A. R.. 2007. Hierarchical K-Means:an algorithm for centroids initialization for K-Means, the Faculty of Science and Engineering, Saga University, Vol. 36, No.1

[4] Alfina, T., Santosa, B. and Barakbah, A.R. 2010. Analisa Perbandingan Metode Hierarchical *clustering*, K-Means dan Gabungan Keduanya dalam *Cluster* Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS). Jurnal Teknik ITS Vol. 1, (Sept, 2012) ISSN: 2301-9271. Surabaya

[5] Delen, D., Crossland, M.D. 2008. Seeding the Survey and Analysis of Research Literature with *Text mining*

[6] Turban, E. Sharda, R. Dele, D. 2011.Decision Support and Business Intelligence Systems. New Jersey : Pearson Education Inc.

[7] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M., Williams, H.E. 2007. Stemming Indonesian : A Confix-Stripping Approach. Transaction on Asian Langeage Information Processing. Vol. 6, No. 4, Artikel 13. Association for Computing Machinery : New York

[8] Harlian, M. 2006. Machine Learning Text Categorization. University of Texas. Austin

[9] Lee, DL. 1997. Document Ranking and the Vector-Space Model. IEEE Software.

[10] Andayani, S. 2007. Pembentukan *Cluster* dalam Knowledge Discovery in Database dengan Algoritma K-Means. Seminar Nasional Matematika dan Pendidikan Matematika 2007. Universitas Negeri Yogyakarta. Yogyakarta.

[11] Oktavia, S., Mara, M. N., Satyahadewi, N. 2013. Pengelompokan Kinerja Dosen Jurusan Matematika FMIPA UNTAN Berdasarkan Penilaian Mahasiswa Menggunakan Metode Ward. Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster) Volume 02, No. 2 (2013), hal 93 – 100. Tanjungpura

[12] Agusta, Y. 2007. K-Means-Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika Vol.3 , 47-60.