PENGELOMPOKAN ABSTRAK SKRIPSI MENGGUNAKAN METODE *SUFFIX TREE CLUSTERING* DAN *SINGULAR VALUE DECOMPOSITION*

Lina Tri Andaru[1)], Bambang Soedijono W[2)], Armadyah Amborowati[3)]
[1) 2)]MTI Program Pasca Sarjana STMIK AMIKOM Yogyakarta
[3)]Teknik Informatika STMIK AMIKOM Yogyakarta
Jl Ring road Utara, Condongcatur, Sleman, Yogyakarta 55281
Email : arlin21@gmail.com[1)], armadyah.a@amikom.ac.id[3)]

Abstrak

Perkembangan teknologi informasi mulai merambah ke berbagai bidang pada suatu Negara, salah satunya dunia pendidikan.Mahasiswa tidak dapat memahami setiap informasi yang ada pada skripsi karena keterbatasan waktu dan banyaknya skripsi menjadi salah satu masalah. Selain itu juga STMIK Sinar Nusantara juga kesulitan dalam mempergunakan skripsi sebagai bahan pengukur dalam memberikan kebijakan atau pengarahan kepada mahasiswa untuk melakukan penelitian skripsi. Oleh karena itu, diperlukan sebuah metode untuk mengelompokkan skripsi melalui abstrak khususnya kata kunci agar memudahkan dalam pengambilan informasi atau data sesuai kebutuhan mahasiswa dan instansi tersebut .

Permasalahan tersebut dapat diatasi dengan menggunakan model clustering untuk mengelompokkan skripsi yang sesuai dengan kompetensi bidang penelitian yang terkait sehingga memudahkan pengguna dalam memilih dokumen skripsi yang relevan dengan penelitian skripsi. Penelitian ini menggunakan metode Suffix Tree Clustering (STC) untuk mengelompokkan skripsi. Selain itu juga digunakan metode Singular Value Dekomposition (SVD) sebagai pendukung dalam perhitungan similiary pada kombinasi base cluster.

Hasil pengujian adalah metode Suffix Tree Clustering dan Singular Value Decomposite dapat digunakan dalam pengelompokan abstrak skripsi menggunakan kata kunci menjadi tiga kelompok yaitu Komputer Akutansi, Ebusines dan Sistem Informasi Perusahaan/Instansi pada STMIK Sinar Nusantara Surakarta dengan nilai keakurasian 87,67%.

Kata kunci:*STC, SDV, pengelompokkan.*

1. Pendahuluan

STMIK Sinar Nusantara merupakan salah satu Sekolah Tinggi Manajemen dan Komputer di Surakarta yang memiliki banyak mahasiswa. Hal ini didukung dengan adanya ISO dan akreditasi B pada salah satu jurusannya. Salah satu jurusan di STMIK Sinar Nusantara adalah S1 Sistem Informasi. Jurusan sistem Infomasi memiliki bidang kompetensi yang harus dikuasai yaitu Komputerisasi Akuntansi, E-Business dan Sistem Informasi Perusahaan/Instansi.

Data berukuran besar yang sudah disimpan jarang digunakan secara optimal karena manusia seringkali tidak memiliki waktu dan kemampuan yang cukup untuk mengelolanya [1]. Skripsi merupakan salah satu data yang disimpan dan jarang digunakan secara optimal. Skripsi hanya diletakkan di rak perpustakaan tanpa adanya informasi yang mengelompokkan skripsi ke dalam kelompok kelompok tertentu. Hal ini menyebabkan, mahasiswa yang ingin mencari referensi terkadang kesulitan dalam memilih referensi.Mahasiswa tidak memahami setiap informasi yang ada pada skripsi karena keterbatasan waktu dan banyaknya skripsi menjadi salah satu masalah. Selain itu juga STMIK Sinar Nusantara juga kesulitan dalam mempergunakan skripsi sebagai bahan pengukur dalam memberikan kebijakan atau pengarahan kepada mahasiswa untuk melakukan penelitian skripsi. Oleh karena itu, diperlukan sebuah metode untuk mengelompokkan skripsi melalui abstrak khususnya kata kunci agar memudahkan dalam pengambilan informasi atau data sesuai kebutuhan mahasiswa dan instansi tersebut.

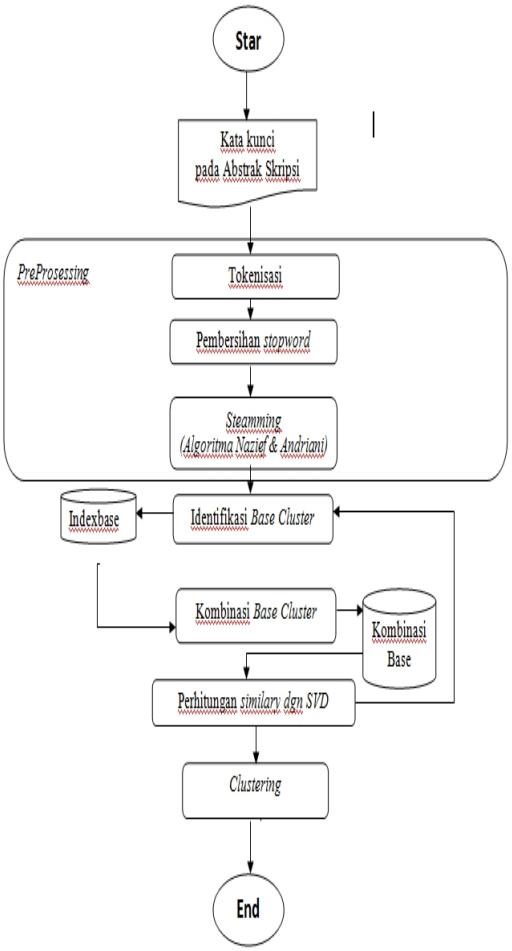
Peningkatan jumlah dokumen dalam format teks yang cukup signifikan membuat proses pengelompokkan atau klasterisasi dokumen (*document clustering*) menjadi penting. Zhao dan Karypis [2]mendefinisikan klasterisasi sebagai proses yang membagi suatu set objek menjadibeberapa jumlah kelompok (klaster) secara spesifik. Klusterisasi merupakan salah satu metode yang dapat digunakan untuk menemukan keterkaitan antar dokumen. Klasterisasi dokumen bertujuan membagi dokumen dalam beberapa kelompok sedemikian hingga dokumen-dokumen dalam klaster yang sama (*intra*-klaster) memiliki kesamaan yang tinggi, sementara dokumen-dokumen dalam klaster yang berbeda (*inter*-klaster) memiliki kesamaan yang rendah [3].

Permasalahan tersebut dapat diatasi dengan menggunakan model *clustering* untuk mengelompokkan skripsi yang sesuai dengan kompetensi bidang penelitian yang terkait sehingga memudahkan pengguna dalam memilih dokumen skripsi yang relevan dengan penelitian skripsi. Penelitian ini menggunakan *metode Suffix Tree*

Clustering (STC) untuk mengelompokkan skripsi. Selain itu juga digunakan metode *Singular Value Dekomposition (SVD)*sebagai pendukung dalam perhitungan *similiary*pada kombinasi base cluster.

Tujuan yang ingin dicapai dalam penelitian ini adalah Untuk mengetahui apakah penelitian dengan metode *Suffix Tree Clustering* dan*Singular Value Dekomposition* dapat digunakan untuk mengelompokkan skripsi pada jurusan Sistem Informasi di STMIK Sinar Nusantara Surakarta menjadi 3 bidang kompetensi yaitu Komputerisasi Akuntansi, E-Business dan Sistem Informasi Perusahaan/Instansi.

Gambar 1 , menjelaskan tahapan yang dilakukan dalam penelitian. Tahapan mulia dari proses *preprocessing,* identifikasi *base cluster,* kombinasi *base cluster* sampai dengan tahap perhitungan *SVD.*



Gambar 1. Alur Penelitian

2. Tinjauan Pustaka

Suffix Tree Clustering

Penggunaan algoritma *clustering.* Algoritma *Suffix treeClustering* (STC) memiliki dua kunci utama, yaitu :1). Menggunakan *phrase* sebagai dasar pembentukan *cluster*nya.2). Menggunakan suatu definisi *cluster* sederhana.

Dalam penelitian Novan [4], Suffix tree Clustering memiliki dua langkah utama. Dalam langkah pertama, pencarian *shared phrase* untuk semua dokumen berita yang dikoleksi. Disebut *shared phrase* sebagai *phrase cluster* atau *base cluster,* yang ditemukan dengan menggunakan suatu struktur data yang dinamakan *suffix tree.* Dalam langkah kedua, mengkombinasikan *base cluster-basecluster* ke dalam suatu *cluster.* Penggabungan antar dua *base cluster* didasarkan pada jumlah dokumen yang melakukan *overlap* diantara kedua *base cluster* tersebut . Suatu *phrase* yang dimaksud dalam konteks algoritma ini adalah urutan satu atau lebih kata-kata. STC memiliki tiga langkah utama, yaitu :

1. *Cleaning* Dokumen.
2. Identifikasi *Base Cluster* menggunakan *Suffix tree.*
3. Mengkombinasikan *Base Cluster* ke dalam suatu *cluster*.

Singular Value Decomposition

Algoritma *Singular Value Decomposition* (SVD) pertama kali diusulkan oleh Eckartand Young [5] termasuk metode eksplorasi statistik multidimensi dengan latar belakang metematika aljabar linier. Algoritma *Singular Value Decomposition* (SVD) mempunyai kelebihan pada efisiensi waktu proses [6] dapat digunakan untuk memaksimalkan perhitungan similary.

Singular Value Decomposition adalah metode aljabar linier [7] yang memecah matriks *A* (*terms-documents*) berdimensi *t* x *d* menjadi tiga matriks *TSD*. *T* adalah matriks kata (*terms*) berukuran *t* x *r*, *S* adalah matriks diagonal berisi nilai skalar (*eigen values*) berdimensi *r* x *r*, dan *r* ditentukan sebelumnya, dan *D* adalah matriks dokumen berukuran *r* x *d*. Dekomposisi nilai singular dari matriks *A* dinyatakan sebagai *A = TSDT*.

SVD dapat mereduksi dimensi dari matriks *A* dengan cara mengurangi ukuran *r* dari matriksdiagonal *S*. Pengurangan dimensi dari matriks *S* dilakukan dengan cara mengubah semua nilaidiagonal matriks *S* menjadi nol, kecuali untuk nilai diagonal dari dimensi yang tersisa.Pengalian ketiga matriks *TSDT* akan membentuk matriks *A* awal dengan nilai setiap elemennyamendekati nilai sebenarnya [8]

Pengelompokkan (*Clustering*)

Metode clustering adalah metode yang memiliki kemampuan untuk menganalisis serta mengelompokkan secara otomatis dokumen-dokumen. Teknik *Clustering* pada umumnya menggunakan kata dan dokumen yang bisasnya dianggap sebagai kumpulan kata-kata tanpa adanya urutan atau disebut dengan bag of word. *Suffix Tree Clustering (STC)* adalah algoritma pertama yang menggunakan frasa (*multi word term*) sehingga

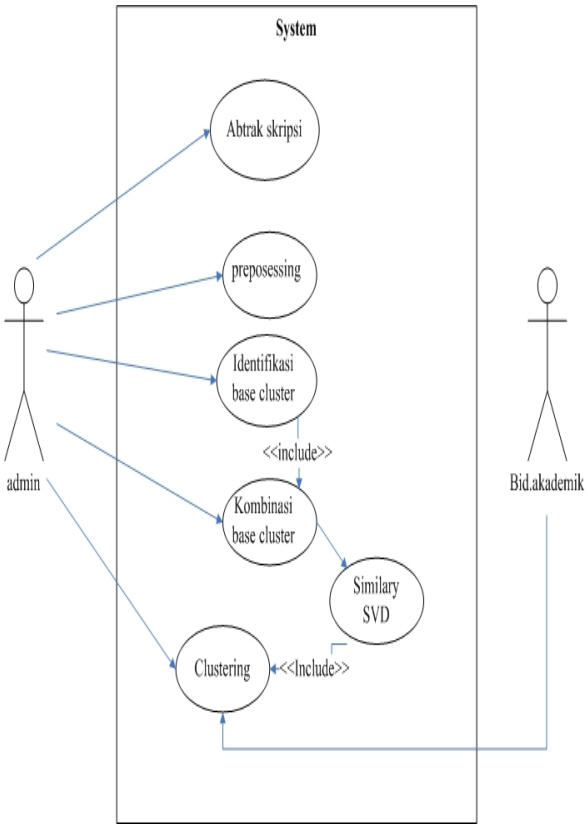
prosesnya lebih sederhana dibandingkan dengan algoritma yang lain [9].

Algoritma Nazief & Adriani

Algoritma *stemming* untuk bahasa yang satu berbeda dengan algoritma *stemming* untuk bahasa lainnya. Sebagai contoh bahasa Inggris memiliki morfologi yang berbeda dengan bahasa Indonesia sehingga algoritma *stemming* untuk kedua bahasa tersebut juga berbeda. Proses *stemming* pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan *root* word (kata dasar) dari sebuah kata. Pada umumnya kata dasar pada bahasa Indonesia terdiri dari kombinasi[10]

Perancangan Sistem

Langkah awal dalam merancang sistem ini dengan menggunakan UML (Unified Modelling Language). Rancangan pertama dengan membuat Use Case diagram yang merupakan penggambaran sistem dapat dilihat dari sudut pandang pengguna (user/ actor) sistem tersebut. Gambar 2, menjelaskan *usecase* yang memiliki dua aktor. Aktor pertama adalah admin yang dapat melakukan semua proses yang ada, sedangkan aktor kedua adalah bidang akademik yang hanya dapat melakukan satu proses.



Gambar 2. *Use case diagram*

3. Pembahasan

a. *Preprosessing*

1)*Tokenisasi*

Tahap ini tidak terlalu digunakan, karena data yang dibaca berupa kata kunci yang sudah ada pemecahan tiap kata dan tidak menggunakan simbol yang tidak bernilai. Tahap ini dugunakan jika data yang kita iputkan banyak dan memiliki beberapa simbol.

2)*Stopword* dan *stoplist removal*

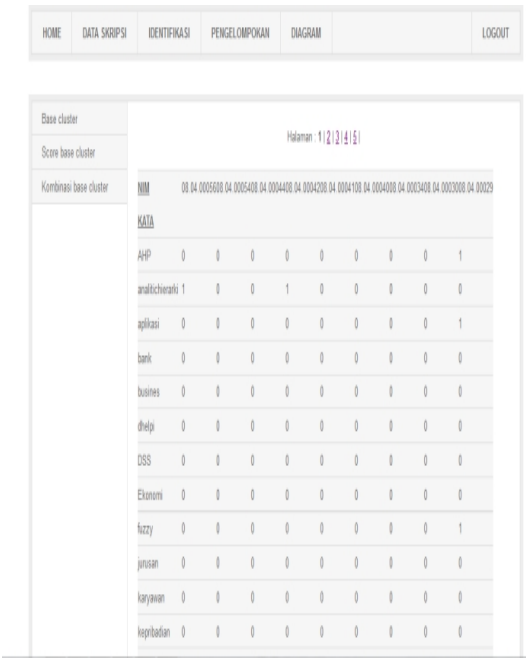
Kata yang dihasilkan bermacam macam, mulai dari kata yang bernilai sampai dengan kata yang tidak dapat digunakan atau tidak bernilai, seperti kata sambung, kata depan dan lain-lain. Kata-kata tersebut langsung dihilangkan tanpa proses penyimpanan terlebih dahulu.

3)*Steaming*

Pengambilan kata dasar menggunakan algoritma nazief dan andriani. Proses ini akan dilakukan pengambilan kata dasar setelah proses *stopword removal*.

b. Identifikasi *base cluster*

Pada proses ini dilakukan identifikasi *base cluster* tiap data. Dari 100 abstrak yang ada, dilakukan proses *prepossessing* dan merubahnya menjadi *base cluster*. Gambar 3, menjelaskan tentang hasil dari identifikasi *base cluster.*

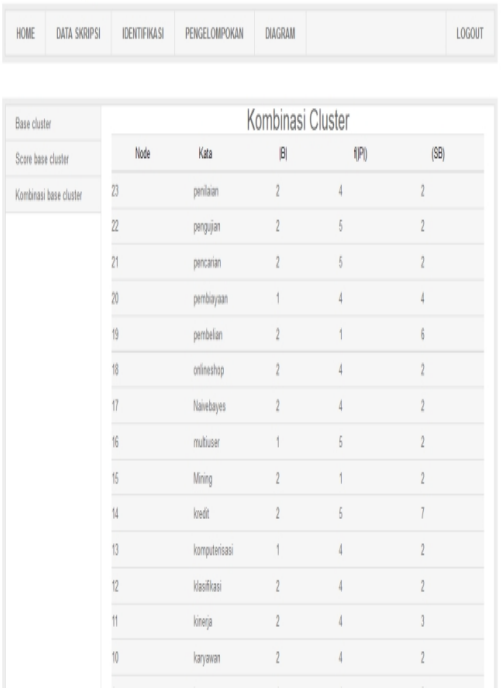


Gambar 3. Identifikasi *base cluster*

c. Kombinasi *base cluster*

Pada tahap kombinasi *base cluster*, didapat total kombinasi data yang diperoleh dari perhitungan

tiap *base cluster*. Gambar 4, menampilkan kombinasi *base cluster*.



Gambar 4. Kombinasi *base cluster*

d. Perhitungan *SVD*

Pada tahap ini dilakukan perhitungan *SVD* yang digunakan untuk mengelompokkan skripsi. Pada 100 skripsi yang telah diproses ada beberapa kata yang hanya muncul di satu atau dua abtrak saja.Oleh karena itu, perlu dilakukan penghilangan kata kata tersebut untuk mempercepat proses perhitungan. Proses reduksi yang dihasilkan memiliki nilai akurasi yang berbeda beda.

Pada reduksi ke 30% nilai akurasi yang dicapai lebih baik dibandingkan reduksi lainnya yaitu 87.67%. Berikut tampilan nilai reduksi dan hasil akurasi pada tabel 1 :

Tabel 1. Hasil Perhitungan SVD

Reduksi Dimensi	Nilai SVD	Jumlah *basecluster*	Akurasi
10%	10	3	73.45%
20%	13	8	76.67%
30%	15	11	87.67%
40%	17	13	72.56%
50%	18	15	68.67%
60%	20	20	56.33%
70%	23	24	53.33%
80%	24	28	33.33%
90%	25	30	33.33%

Gambar 5, merupakan tampilan laporan berupa grafik hasil pengelompokkan abstrak skripsi berdasarkan kata kunci adalah sebagai berikut :

Gambar 5. Hasil Pengelompokkan

Pengujian Sistem

Pengujian dilakukan dua cara seperti berikut :

a). *Recall*, yakni tingkat keberhasilan mengenali suatu kompetensi dari seluruh kompetensi yang seharusnya dikenali. Rumusnya r = a/(a+c) untuk a+c > 0. Selain itu tidak didefinisikan

b). *Precision*(presisi), yakni tingkat ketepatan hasil klusteringterhadap suatu kompetensi.Artinya,dari seluruh dokumen hasil klustering, berapa persenkah yangdinyatakan benar. Rumusnya adalah p = a/(a+b) jika a+b > 0. Hasil pengujian sebagai berikut :

Tabel 2, menampilkan hasil pengujian recall dan pengujian presisi, dengan akurasi 87,67%. Tabel tersebut merupakan hasil dari pengelompokkan yang dikenali dan tidak dikenali.Ada sekitar 6 Abstrak yang tidak dikenali dengan nilai akurasi 87.67%.

Tabel 2. Hasil Pengujian

Jml data	Tidak dikenali	Dikenali	Kelomp	Akurasi	*recall*	*presisi*
100	6	54	SI	87.67 %	0.33	0.37
		26	Akutansi			
		14	Bisnis			

4. Penutup

Kesimpulan

1. Metode *Suffix Tree Clustering* dan *Singular Value Decomposite* dapat digunakan dalam pengelompokan abstrak skripsi menggunakan kata kunci menjadi tiga kelompok yaitu Komputer Akutansi, Ebusines dan Sistem Informasi Perusahaan/Instansi pada STMIK Sinar Nusantara Surakarta.
2. Akurasi yang sesuai diperoleh pada 100 data abstrak dengan menggunakan reduksi dimensi 30% yaitu akurasi 87.67% dengan hasil 54 data pada kelompok Sistem Informasi Perusahaan/ Instansi , 14 data terkelompok dalam Ebusines dan 26 data terkelompok dalam Komputer akutansi (pengujian *recall* adalah 0.33 dan *presisi* sebesar 0.37).

Saran

Perlu dilakukan penelitian lebih lanjut dengan metode lain, agar didapat metode yang tepat dengan akurai yang lebih baik untuk semua bidang kompetensi keahlian dalam proses pengelompokkan dokumen abstrak dengan kata kunci.

Daftar Pustaka

[1] Hermadi, I. 2007. Clustering Menggunakan Self-Organizing Maps (Studi Kasus: Data PPMB IPB).*FMIPA Institut Pertanian Bogor*. 5: 2

[2] Zhao, Y. & Karypis, G. 2005, "Empirical and theoretical comparisons of selected criterion functions for document clustering", Machine Learning 55 (3), hal.50-62

[3] Jain, A.K., Murty, M.N. & Flynn, P.J. 1999, "Data Clustering : A Review", ACM Computing Survey Vol. 31, No. 3, hal.264-323

[4] Novan S. 2001. *Implementasi Aplikasi Information Retrieval Untuk Pendeteksian dan Klasifikasi Berita Kejadian Berbahasa Indonesia Berbasis Web*. Tugas Akhir, Jurusan Teknik Informatika Fakultas Teknologi Informasi ITS Surabaya

[5] Lipovetsky, Stan. 2009. "*PCA and SVD with nonnegative loadings,"GfK CustomResearch for excelence*, vol. 42, no. 1, pp. 1-30

[6] Sembiring, Rahmat Widia; Zain, Jasni Mohamad; Embong, Abdullah.2011. "Dimension Reduction of Health Data Clustering," *International Journal onNew Computer Architectures and Their Applications (IJNCAA)*, vol. 3, no. 1, pp. 1041-1050

[7] Bau III, David, Lloyd N. Trefethen, 1997. *Numerical Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics

[8] Ratna, Anak Agung Putri, Bagio Budiardjo, Djoko Hartanto, 2007. *Simple:Sistim Penilai Esei Otomatis untukMenilai Ujian dalam Bahasa Indonesia*. Departemen Elektro, Fakultas Teknik,Universitas Indonesia. Depok, Indonesia April 2007: 5-11: Makara

[9] Kusamaya. 2007.Pengembangan Suffix Tree Clustering untuk Comparative Text Mining. Tesis. Bandung: Institute Teknologi Bandung

[10] Nazief, Bobby dan Mirna Adriani. 1996. *Confix-Stripping: Approach to Stemming Algorithm for* Bahasa Indonesia, *Fakulty of Computer Science University of Indonesia.*

Biodata Penulis

Lina Tri Andaru Penulis Pertama , memperoleh gelar Ahli Madya (A.Md), Jurusan Manajemen InformatikaUniversitas Sebelas Maret Surakarta, lulus tahun 2010. Memperoleh gelar Sarjana Komputer(S.Kom), Jurusan Sistem Informasi STMIK Sinar Nusantara Surakarta, lulus tahun 2012. Saat ini sedang menempuh Program Pasca Sarjana Magister Teknik Informatika di STMIK AMIKOM Yogyakarta

Bambang Soedijono W, Penulis Kedua, memperoleh S1 Fak. MIPA, UGM,lulus tahun 1970. Memperoleh S3 Program Pascasarjana UGM,lulus tahun 1992. Saat ini sebagai dosen Magister Teknik Informatika di STMIK AMIKOM Yogyakarta

Armadyah Amborowati Penulis Ketiga, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi STMIK AMIKOM Yogyakart. Memperoleh gelar Master of Engineering(M.Eng) tahun 2009,di Program Pasca Sarjana Magister Teknologi Informasi Fakultas Teknik Elektro Universitas Gajah Mada Yogyakarta. Saat ini sedang menempuh S3 di fakultas Ilmu Komputer Universitas Gajah Mada Yogyakarta. Saat ini sebagai dosen Teknik Informatika di STMIK AMIKOM Yogyakarta