

PENENTUAN KEMIRIPAN TOPIK PROYEK AKHIR BERDASARKAN ABSTRAK PADA JURUSAN TEKNIK INFORMATIKA MENGGUNAKAN METODE SINGLE LINKAGE HIERARCHICAL

Nur Rosyid M, Entin Martiana, Damitha Vidyastana,
Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember Kampus ITS Keputih Sukolilo Surabaya 60111, Indonesia
Tel:+62-31-5947280 Fax:+62-31-5946114
Email: rosyid@eepis-its.edu, entin@eepis-its.edu, damitha_it05@yahoo.com

Abstrak

Menentukan kemiripan topik proyek akhir berdasarkan deskripsi pada jurusan teknik informatika menggunakan metode single linkage hierarchical adalah suatu metode untuk mengetahui kemiripan atau kedekatan judul proyek akhir antara input yang dimasukkan user dengan judul proyek akhir yang sudah dikerjakan atau sudah ada pada jurusan teknik informatika dengan teknik clustering dan proses inner product.

Pada data – data abstrak proyek akhir yang sudah dikerjakan atau sudah ada dilakukan proses clustering dengan menggunakan Single Linkage Hierarchical Method (SLHM) sampai terbentuk sembilan cluster sesuai dengan bidang yang ada pada jurusan teknik informatika. Kemudian dilakukan proses inner product antara input yang sudah melewati proses text mining dengan sembilan cluster yang terbentuk. Proses selanjutnya dilakukan proses perkalian transpose antara input dengan anggota – anggota dari cluster terpilih yang hasil dari nilai proses inner productnya memiliki nilai tertinggi, proses perkalian transpose ini bertujuan untuk menunjukkan judul manakah yang memiliki nilai tertinggi, semakin tinggi nilainya maka semakin memiliki kedekatan atau kemiripan dengan judul dan abstrak yang diinputkan. Data – data yang digunakan untuk membentuk metadata clustering adalah data abstrak proyek akhir teknik informatika tahun 2006 – 2009, sedangkan judul dan abstrak yang diinputkan merupakan judul untuk pengajuan proyek akhir jurusan teknik informatika.

Dari empat puluh kali percobaan dengan inputan yang berbeda di setiap percobaan hasilnya memberikan kesimpulan bahwa pada umumnya penentuan kemiripan topik proyek akhir berdasarkan abstrak pada jurusan teknik informatika dengan metode single linkage hierarchical dapat digunakan untuk mengetahui kemiripan atau kedekatan judul proyek akhir sesuai dengan inputan. Semakin atas urutan/ranking dari output judul yang dihasilkan maka semakin mendekati dengan inputan.

Kata kunci : *text mining, Centroid Linkage Hierarchical Method, inner product, transpose..*

1. PENDAHULUAN

Perkembangan teknologi dewasa ini telah mengalami perkembangan yang semakin pesat dan semakin banyak pula orang yang dapat beradaptasi dengan kemajuan teknologi sekarang ini, teknologi dapat digunakan oleh semua kalangan yang dapat memanfaatkannya untuk berbagai macam keperluan sehingga memudahkan mereka dalam memenuhi kebutuhan dengan lebih cepat, efisien dan akurat sehingga tidak menyita waktu. Sama halnya dengan perkembangan Teknik Informatika yang muncul pada cabang ilmu baru yaitu Pencarian Informasi (Information Retrieval).

Pencarian informasi berdasarkan keyword/inputan ini sangat berguna untuk pencarian terarah dan berguna ketika user ingin mengetahui informasi yang berhubungan dengan keyword yang dicari. Begitu juga dengan pencarian informasi saat mahasiswa ingin mengetahui judul-judul yang mendekati

proyek akhir yang akan diajukan . Karena setiap mahasiswa diharuskan untuk mengajukan judul proyek akhir, dimana judul tersebut harus merupakan judul yang baru dan belum pernah ada sebelumnya. Untuk itu dibutuhkan suatu sistem yang dapat memberikan informasi kepada mahasiswa yang ingin melakukan pencarian informasi untuk mengetahui informasi judul-judul yang mendekati proyek akhir yang diajukan . Dalam hal ini proyek akhir pada jurusan Teknik Informatika. Dengan demikian diharapkan dapat memperoleh hasil akhir informasi yang benar-benar relevan atau sesuai dengan inputan yang dicari.

Dalam penelitian ini akan dibahas tentang bagaimana cara untuk mengetahui judul-judul yang mendekati proyek akhir yang akan diajukan, dengan membandingkan berdasarkan dari abstrak yang sudah ada dengan tahapan yang dilakukan secara umum dalam text mining seperti tokenizing, filtering (stoplist dan

wordlist), yang kemudian akan diklasifikasi atau dikelompokkan. Metode yang digunakan dalam penelitian ini adalah metode clustering Single Linkage Hierarchical. Setelah terbentuknya sembilan cluster sesuai dengan bidang yang ada, maka akan dilakukan proses inner product dan perkalian transpose untuk mengetahui judul apa saja yang memiliki kemiripan/kedekatan dengan yang di inputkan.

2. TEORI PENUNJANG

2.1. TEXT MINING

Text mining adalah proses penemuan akan informasi atau trend baru yang sebelumnya tidak terungkap dengan memproses dan menganalisa data dalam jumlah besar.

Proses *text mining* meliputi proses *tokenizing* dan *filtering*.

2.1.1 Tokenizing

Tokenizing adalah proses penghilangan tanda baca pada kalimat yang ada dalam dokumen sehingga menghasilkan kata-kata yang berdiri sendiri-sendiri.

2.1.2 Filtering

Tahap *filtering* adalah tahap pengambilan kata-kata yang penting dari hasil *tokenizing*. Tahap *filtering* ini dapat menggunakan algoritma *stoplist* atau *wordlist*. *Stoplist* yaitu *filtering* terhadap kata-kata yang tidak layak untuk dijadikan sebagai keyword sehingga kata-kata tersebut dapat dihilangkan. Sedangkan *wordlist* adalah daftar kata-kata yang mungkin digunakan sebagai keyword, dengan demikian maka tentu jumlah kata yang termasuk dalam *wordlist* akan lebih banyak daripada *stoplist*, sehingga dalam penelitian ini digunakan daftar *wordlist* dan *stoplist*. Oleh karena belum tersedia, maka pada penelitian ini juga akan berusaha mencari *stoplist* tersebut secara manual.

2.2. CLUSTERING SINGLE LINKAGE HIERARCHICAL METHOD

Single Linkage adalah proses clustering yang didasarkan pada jarak terdekat antar obyeknya (minimum distance).

Algoritma Single Linkage Hierarchical Method pada Clustering dapat dilakukan dengan langkah-langkah sebagai berikut :

1. Diasumsikan setiap data dianggap sebagai cluster. Jika

n =jumlah data dan c =jumlah cluster, berarti ada $c=n$.

2. Menghitung jarak antar cluster dengan Euclidian distance.
3. Mencari 2 cluster yang mempunyai jarak antar cluster yang paling minimal dan digabungkan (merge) kedalam cluster baru (sehingga $c=c-1$)
4. Kembali ke langkah 3, dan diulangi sampai dicapai cluster yang diinginkan.

Penghitungan jarak antar obyek, maupun antar clusternya dilakukan dengan Euclidian distance, khususnya untuk data numerik. Untuk data 2 dimensi, digunakan persamaan sebagai berikut :

$$d(x, y) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2} \quad (1)$$

2.3. METADATA

Definisi sederhana dari metadata adalah data mengenai data. Metadata ini mengandung informasi mengenai isi dari suatu data yang dipakai untuk keperluan manajemen file/data itu nantinya dalam suatu basis data. Jika data tersebut dalam bentuk teks, metadatanya biasanya berupa keterangan mengenai nama ruas (field), panjang field, dan tipe fieldnya: integer, character, date, dll. Satu lagi untuk jenis data berupa kumpulan file, metadatanya adalah nama-nama file, tipe file, dan nama pengelola (administrator) dari file-file tersebut. Tabel 1 merupakan contoh tabel metadata :

Tabel 2 : tabel Metadata data TA

Atribut Judul	Face	Sketsa	Proyek	Informasi	Database	...
1	5	12	0	14	0	...
2	16	20	7	0	0	...
3	0	14	0	6	2	...
4	10	3	0	8	13	...
5	10	10	2	8	16	...
6	0	0	7	3	15	...
...

2.4. INNER PRODUCT

Inner product yang merupakan istilah lain dari dot product atau dapat juga disebut sebagai scalar product adalah perkalian antara dua vektor yang dapat berupa vektor baris atau vektor kolom, sehingga hasilnya berupa scalar.

$$a \cdot b = \sum_{i=1}^n a_i b_i = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

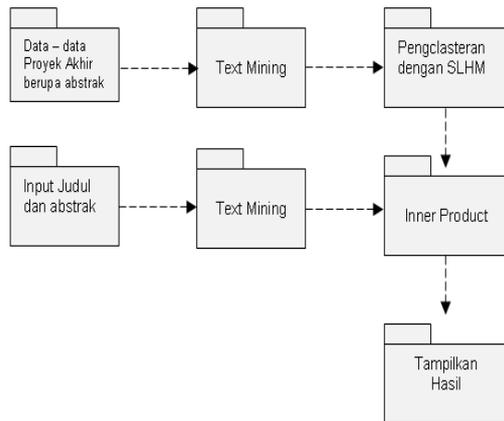
Contoh perhitungan dari persamaan di atas dapat dilihat pada contoh berikut :

$$\begin{bmatrix} 3 & 1 & 4 \end{bmatrix} \cdot \begin{bmatrix} 2 & 5 & 1 \end{bmatrix} = (3)(2) + (1)(5) + (4)(1) = 15$$

$$\begin{bmatrix} 1 & 3 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix} = (1)(3) + (3)(5) + (2)(2) = 22$$

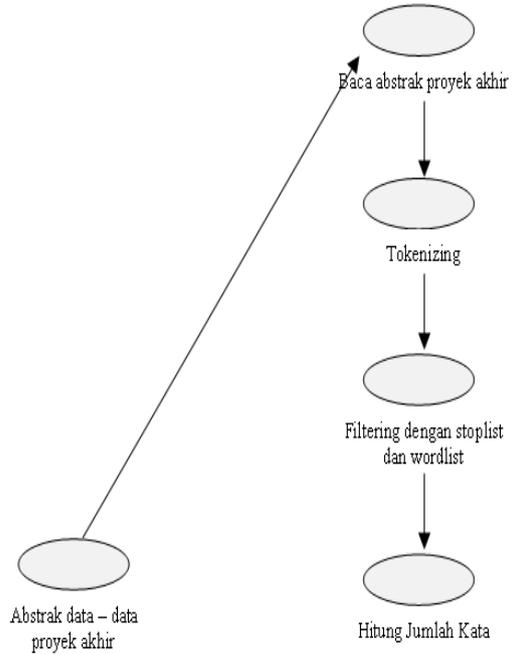
3. PERANCANGAN SISTEM

Sistem perangkat lunak ini pada dasarnya merupakan alat bantu untuk sedikit memberikan solusi pada kebutuhan kita untuk mengetahui judul proyek akhir yang mempunyai kedekatan atau kemiripan dengan judul dan abstrak yang diinputkan. Diagram sistemnya adalah seperti gambar 1 berikut:

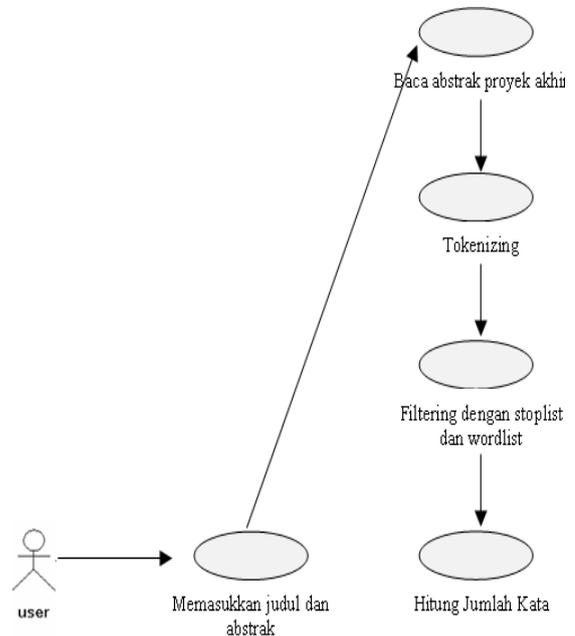


Gambar 1. Garis besar diagram sistem

Langkah awal dari sistem pada penelitian ini adalah pengimplementasian fungsi *text mining* yang meliputi proses *tokenizing* dan *filtering* yang dioptimalkan sebagai sarana untuk memproses judul dan abstrak dari inputan dan untuk memproses data – data abstrak proyek akhir. Diagram alur dari proses *text mining* adalah seperti gambar 2 dan gambar 3 berikut :

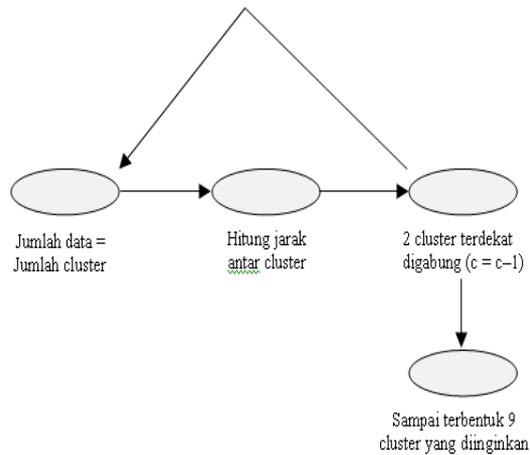


Gambar 2. diagram alur proses text mining data proyek akhir berupa abstrak



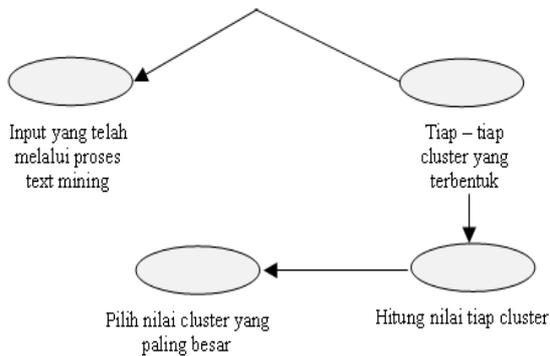
Gambar 3. diagram alur proses text mining data proyek akhir pada input

Langkah selanjutnya adalah proses pengclasteran terhadap data – data abstrak proyek akhir yang telah mengalami proses text mining menggunakan metode Single Linkage Hierarchical Method (SLHM) seperti yang ditunjukkan pada gambar 4 berikut :



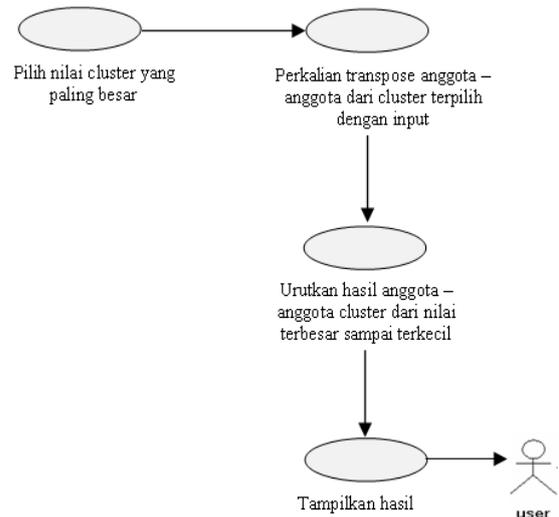
Gambar 4. Digram alur proses pengklasteran dengan SLHM

Berikutnya adalah proses inner product antara input dengan sembilan cluster yang terbentuk dengan langkah-langkah seperti pada gambar 5 berikut ini:



Gambar 5. Digram alur proses *inner product*

Tahap terakhir adalah proses *sorting* seperti pada gambar 6. Dimana hasil dari perkalian transpose dari anggota – anggota cluster yang terpilih dengan input akan dihitung dan hasilnya akan diurutkan kemudian hasilnya akan ditampilkan sebagai hasil yang paling mirip atau mendekati dari input.

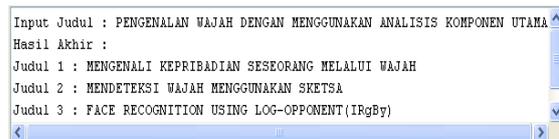


Gambar 6. Digram alur proses menampilkan hasil akhir

4. HASIL UJI COBA

Aplikasi Menentukan Kemiripan Topik Proyek Akhir Berdasarkan Abstrak Pada Jurusan Teknik Informatika Menggunakan Metode Single Linkage Hierarchical ini diterapkan pada uji coba berikut ini :

4.1 Uji coba digunakan untuk mengetahui ketepatan sistem dalam mengetahui kedekatan/kemiripan dari judul dan abstrak yang diinputkan, Judul dan abstrak yang diinputkan termasuk dalam bidang computer vision untuk pengenalan wajah. Hasilnya akhirnya seperti pada gambar 7 :



Gambar 7 Hasil Akhir Judul yang Paling Mendekati Inputan

Analisa percobaan 4.1

Dari hasil percobaan 4.1 dapat diketahui bahwa cluster yang memiliki nilai tertinggi saat proses inner product adalah cluster8, cluster ini beranggotakan tiga data abstrak proyek akhir, maka dari anggota – anggota cluster8 tersebut akan dilakukan proses perkalian transpose dengan input, dari proses tersebut didapatkan hasil akhir judul – judul yang memiliki kedekatan dengan input, semakin tinggi nilai dari proses perkalian transpose tersebut maka semakin memiliki kedekatan/kemiripan dengan input. Hasil urutan untuk tiga nilai tertinggi adalah sebagai berikut :

- Judul 1: Mengenali kepribadian seseorang melalui wajah (Nilai = 195)
- Judul 2: Mendeteksi wajah menggunakan sketsa (Nilai = 182)

Judul 3: Face recognition using log-opponent (IRgBy) (Nilai = 136)

Dari hasil akhir di atas untuk judul 1, judul 2, dan judul 3 dapat digolongkan termasuk dalam bidang computer vision untuk pengenalan / deteksi wajah sama seperti input judul dan abstrak pada percobaan 4.1 di atas yang juga termasuk dalam bidang computer vision untuk pengenalan wajah. Nilai hasil akhir dari proses perkalian transpose untuk judul 1, judul 2, dan judul 3 memiliki hasil yang cukup tinggi, dan perbedaan hasil nilai antar ketiga judul juga tidak begitu jauh berbeda.

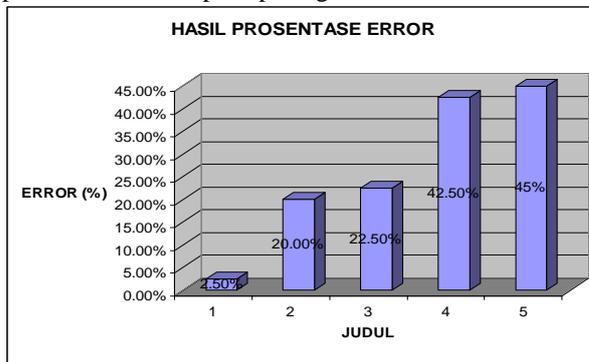
4.2 Mengetahui kedekatan judul yang didapatkan melalui beberapa kali percobaan yang diujikan pada Sistem :

Setelah dilakukan percobaan sebanyak 40 inputan pada judul dan abstrak yang berbeda di setiap percobaan, maka didapatkan prosentase error untuk hasil akhir yang kemiripan atau kedekatan judulnya berbeda bidang dengan inputan yang dimasukkan oleh user, percobaan ini hanya ditunjukkan pada lima judul yang paling mirip/dekat yang memiliki nilai tertinggi di hasil akhirnya, seperti yang ditunjukkan pada tabel 2 dan gambar 9 grafik di bawah ini :

Tabel 2 : tabel Hasil Prosentase Error

Judul 1	Judul 2	Judul 3	Judul 4	Judul 5
2.50%	20%	22.50%	42.50%	45%

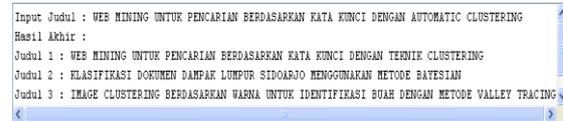
Dari hasil yang terlihat pada tabel 1 di atas dapat ditampilkan dalam bentuk grafik hasil prosentase error seperti pada gambar 9 :



Gambar 9 Hasil Prosentase Error pada lima judul tertinggi

Dari hasil yang terlihat pada gambar grafik di atas dapat dilihat bahwa semakin atas urutan/ranking dari output judul yang dihasilkan maka semakin memiliki kedekatan atau kemiripan dengan inputan, karena semakin bawah urutan/ranking dari judul yang dihasilkan semakin tidak mirip dengan input karena errornya yang semakin besar.

4.3 Uji coba digunakan untuk mengetahui ketepatan sistem dalam mengetahui kedekatan/kemiripan dari judul dan abstrak yang diinputkan, Judul dan abstrak yang diinputkan ini merupakan pengembangan dari judul proyek akhir yang sudah ada atau sudah dikerjakan. Hasilnya akhirnya seperti pada gambar 8 berikut ini :



Gambar 8 Hasil Akhir Judul yang Paling Mendekati Inputan

Analisa percobaan 4.2

Dari hasil percobaan 4.2 di atas dapat diketahui bahwa cluster yang memiliki nilai tertinggi saat proses inner product adalah cluster9, maka dari anggota – anggota cluster9 tersebut akan dilakukan proses perkalian transpose dengan input, dari proses tersebut didapatkan hasil akhir judul – judul yang memiliki kedekatan dengan input, semakin tinggi nilai dari proses perkalian transpose tersebut maka semakin memiliki kedekatan/kemiripan dengan input. Hasil urutan untuk tiga nilai tertinggi adalah sebagai berikut :

Judul 1: Web mining untuk pencarian berdasarkan kata kunci dengan teknik clustering (Nilai = 105)

Judul 2: Klasifikasi dokumen dampak lumpur sidoarjo menggunakan metode bayesian (Nilai = 69)

Judul 3: Image clustering berdasarkan warna untuk identifikasi buah dengan metode valley tracing (Nilai = 57)

Dari hasil akhir di atas dapat di lihat bahwa hasil dari judul 1 “Web mining untuk pencarian berdasarkan kata kunci dengan teknik clustering” merupakan judul proyek akhir yang dikembangkan oleh inputan user, jadi sistem ini dapat mengetahui kedekatan/kemiripan dari judul proyek akhir yang dikembangkan dari judul proyek akhir yang sudah ada atau sudah dikerjakan. Dari judul 1 dan judul 2 dapat digolongkan termasuk dalam bidang soft computing sama seperti bidang untuk inputan judul dan abstrak pada percobaan 4, sedangkan untuk judul 3 dapat digolongkan termasuk dalam bidang computer vision.

5. KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Dari percobaan dan hasil analisa yang dilakukan maka dapat diambil kesimpulan:

1. Hasil akhir dari sistem penentuan kemiripan topik proyek akhir berdasarkan abstrak pada jurusan teknik informatika dengan menggunakan metode single linkage hierarchical ini dapat digunakan untuk mengetahui kemiripan atau kedekatan judul dari input yang dimasukkan user sesuai dengan bidang yang diinputkan.
2. Hasil akhir dari sistem penentuan kemiripan topik proyek akhir berdasarkan abstrak pada jurusan teknik informatika ini dapat digunakan untuk mengetahui kemiripan atau kedekatan judul dari cluster yang hanya mempunyai satu data abstrak proyek akhir saja.
3. Hasil akhir dari sistem penentuan kemiripan topik proyek akhir berdasarkan abstrak pada jurusan teknik informatika ini dapat digunakan untuk mengetahui kemiripan atau kedekatan judul dari proyek akhir yang dikembangkan.
4. Semakin atas urutan/ranking dari output judul yang dihasilkan maka semakin memiliki kedekatan atau kemiripan dengan inputan.

5.2 SARAN

Dari percobaan dan hasil analisa yang dilakukan didapatkan saran-saran yang berguna untuk perbaikan kinerja sistem antara lain:

1. Untuk mendapatkan hasil yang lebih optimal, pada pengembangan sistem ini selanjutnya dapat menggunakan metode clustering yang lainnya seperti K-Means yang di optimasi, karena penggunaan metode single linkage hierarchical pada sistem penentuan kemiripan proyek akhir ini pembentukan clusternya kurang optimal, dikarenakan metode ini hanya mengenali jarak tanpa memperhatikan makna yang terkandung pada tiap – tiap atribut.
2. Karena pendataan kata untuk proses stoplist masih dilakukan secara manual, maka untuk pengembangan program ini selanjutnya diharapkan daftar kata dalam stoplist dapat dibakukan agar hasil text miningnya lebih optimal.

DAFTAR PUSTAKA

- [1] A.R. barakbah, *Text Mining*, In. Soft Computing Research Group 2006, Politeknik Elektronika Negeri Surabaya,ITS.
- [2] A.R. barakbah, *Clustering*, In. Workshop Data Mining 2006, Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya,ITS.
- [3] Permana Andri, *KLASIFIKASI TOPIK TA SESUAI DENGAN KEMAMPUAN MENGGUNAKAN METODA INNER PRODUCT*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya 2008
- [4] Aguseta Usmaida, *WEB MINING UNTUK PENCARIAN BERDASARKAN KATA KUNCI DENGAN TEKNIK CLUSTERING*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya 2007
- [5] Aguseta Usmaida, *WEB MINING UNTUK PENCARIAN BERDASARKAN KATA KUNCI DENGAN AUTOMATIC CLUSTERING*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya 2009
- [6] Zubin Mehta, *Penentuan Obat Alternatif untuk Anak dengan Inner Product*, Tugas Akhir Jurusan Teknologi Informasi Politeknik Elektronika Negeri Surabaya, Surabaya 2007
- [7] Perpustakaan Online ITS, <http://digilib.its.ac.id> , tanggal akses 9 Juni 2009
- [8] Perpustakaan Online ITB, <http://digilib.itb.ac.id> , tanggal akses 9 Juni 2009
- [9] Data Proyek Akhir PENS, <http://eepis-its.edu> , tanggal akses 22 Maret 2009
- [10] Wikimedia Foundation, Inc., "Clustering", <http://id.wikipedia.org/wiki/clustering>, tanggal akses 22 Maret 2009