Aplikasi *Text Mining* untuk Automasi Penentuan Tren Topik Skripsi dengan Metode *K-Means Clustering*

Kestrilia Rega Prilianti
Program Studi Teknik Informatika Universitas MaChung
kestrilia@machung.ac.id

Hendra Wijaya
Program Studi Teknik Informatika Universitas MaChung
310910022@student.machung.ac.id

ABSTRAK

Pengetahuan tentang tren topik skripsi mahasiswa di sebuah Universitas pada umumnya maupun di program studi tertentu pada khususnya dapat membawa manfaat yang sangat positif bagi pengembangan kurikulum maupun perencanaan *roadmap* penelitian skala institusi. Namun, teknologi untuk dapat secara cepat mendapatkan pengetahuan secara menyeluruh dari penelitian-penelitian yang telah dilakukan mahasiswa melalui proyek skripsinya sangatlah terbatas jika dibandingkan dengan teknologi penyimpanan dokumen yang tersedia. Melalui penelitian ini dikembangkan sebuah aplikasi untuk menggali pengetahuan dari topik-topik skripsi mahasiswa yang biasanya terkumpul melalui *repository* digital perpustakaan Universitas. Proses tersebut dilakukan secara semi-otomatis dengan memanfaatkan teknologi *text mining* dan algoritma *k-means clustering* terhadap kumpulan dokumen digital abstrak dari buku skripsi. Uji coba dilakukan dengan melibatkan beberapa kepala progran studi dan dosen penanggungjawab skripsi. Dari uji coba tersebut diperoleh hasil yang baik yaitu 89% responden manyatakan tren topik skripsi yang dihasilkan oleh aplikasi telah sesuai dengan kondisi yang sesungguhnya.

Kata Kunci

Skripsi, Dokumen, Text Mining, K-Means Clustering.

1. PENDAHULUAN

Semakin pesat dan murahnya perkembangan teknologi media penyimpanan digital telah mendorong terjadinya ledakan jumlah dokumen elektronik yang tersimpan dalam *repository* perpustakaan Universitas. Berbagai karya ilmiah dari sivitas akademika seperti skripsi, laporan penelitian, laporan kerja praktek dan lain sebagainya telah tersedia dalam versi digital. Namun, pada umumnya fenomena ini tidak disertai dengan pertumbuhan jumlah informasi atau pengetahuan yang dapat disarikan dari dokumen-dokumen elektronik tersebut (Gupta, 2011). Metode *text mining* merupakan pengembangan dari metode *data mining* yang dapat diterapkan untuk mengatasi masalah tersebut. Algoritma-algoritma dalam *text mining* dibuat untuk dapat mengenali data yang sifatnya semi terstruktur misalnya sinopsis, abstrak maupun isi dari dokumen-dokumen (Gupta & Lehal, 2009).

Beberapa aplikasi *text mining* telah diterapkan di perpustakaan terutamanya untuk pencarian bahan pustaka berbasis teks (Yuwono, 2005; Santoso, 2012; Sari, 2012). Meskipun demikian belum banyak aplikasi dikembangkan untuk tujuan analisis. Sehingga sangatlah sulit untuk dapat dengan segera mengetahui topik penelitian populer pada tahun tertentu ataupun kecenderungan minat penelitian mahasiswa program studi tertentu misalnya. Oleh karena itu pada penelitian ini dikembangkan aplikasi berbasis *text mining* dengan menerapkan algoritma *k-means clustering* yang dapat membantu para pengambil kebijakan seperti kepala program studi, dekan ataupun kepala perpustakaan untuk dapat dengan cepat menganalisis tren topik skripsi pada program studi dan tahun tertentu yang dikehendaki. Adapun beberapa kemanfaatan strategis yang didapat melalui analisis tentang tren topik skripsi adalah: (1) Sebagai petunjuk untuk menyusun rencana penelitian jangka panjang Universitas; (2) Sebagai bahan evaluasi untuk menentukan strategi pengembangan teknologi tepat guna yang dapat dipatenkan; (3) Sebagai bahan kajian untuk pengembangan kurikulum.

Pembahasan pada artikel ini dimulai dengan pendahuluan yang disusul dengan kajian yang terkait dengan konsep dasar dari metode yang digunakan. Berikutnya adalah pemaparan desain dari sistem yang dibangun yang kemudian didiskusikan pada bagian hasil dan pembahasan dan akhirnya ditutup dengan kesimpulan.

2. KAJIAN TERKAIT

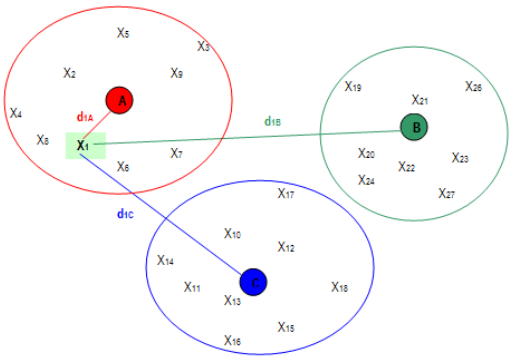
2.1 Text Mining

Text mining adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen (Han & Kamber, 2006). Prosedur utama dalam metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik tertentu seperti analisis kelompok, klasifikasi dan asosiasi. Tahapan dalam *text mining* secara umum adalah *tokenizing, filtering, stemming, tagging,* dan *analyzing* (Michael, 2004). *Tokenizing* merupakan tahapan untuk memisah-misahkan setiap kata (*token*) pada dokumen *input*. *Filtering* merupakan proses seleksi terhadap kata-kata yang dihasilkan dari proses *tokenizing*, dapat dilakukan dengan algoritma *stop list* maupun *word list*. Algoritma *stop list* akan membuang kata-kata yang tidak penting seperti kata ganti, kata keterangan, kata sambung, kata depan dan kata sandang. Sebaliknya, algoritma *word list* akan menyimpan kata-kata yang penting. Proses *stemming* kemudian dilakukan untuk mencari kata dasar dari setiap kata yang telah lolos proses *filtering*. Terdapat 4 varian algoritma untuk proses *stemming* ini, yaitu: (1) *Table lookup, s*eluruh kata dasar disimpan dalam memori untuk selanjutnya dijadikan acuan dalam pemeriksaan dokumen *input*. Kelemahan metode ini adalah membutuhkan ruang penyimpanan yang besar; (2) *Successor variety*, setiap kata dalam dokumen *input* yang akan diperiksa dipecah secara bertahap menjadi awalan-awalan (prefiks). Untuk setiap awalan kemudian dicari kemungkinan bentuk lainnya (variasinya) didalam *corpus*, pencarian dihentikan jika jumlah temuan telah melampaui nilai batas tertentu; (3) *N-gram*, pemeriksaan setiap kata dalam dokumen *input* dilakukan dengan menerapkan

konsep *clustering*. Setiap kata dicari nilai kedekatannya dengan kata-kata yang lain dan disimpan dalam sebuah matriks. Matriks tersebut kemudian dijadikan acuan untuk melakukan pengelompokan kata-kata; (4) *Affix removal*, untuk setiap kata pada dokumen *input* dihilangkan awalan dan akhirannya dengan mengacu kepada *action rules*. Pada dokumen yang berbahasa Indonesia, proses *filtering* secara sederhana dilakukan dengan menghilangkan awalan dan akhiran dari setiap kata. Jika dokumen berbahasa Inggris, maka diperlukan proses lanjutan yang disebut sebagai *tagging*. Proses *tagging* dilakukan untuk mencari bentuk awal dari setiap kata lampau. Setelah semua kata penting berhasil dikoleksi dari rangkaian proses tersebut, maka tahap berikutnya adalah *analyzing* yaitu menentukan keterhubungan antar dokumen dengan mengamati frekuensi kemunculan tiap kata yang ada pada tiap dokumen.

2.2 K-Means Clustering

K-means clustering merupakan metode yang populer digunakan untuk mendapatkan dekripsi dari sekumpulan data dengan cara mengungkapkan kecenderungan setiap individu data untuk berkelompok dengan individu-individu data lainnya. Kecenderungan pengelompokan tersebut didasarkan pada kemiripan karakteristik individu-individu data yang ada. Ide dasar dari teknik ini adalah menemukan pusat dari setiap kelompok data yang mungkin ada untuk kemudian mengelompokkan setiap data individul kedalam salah satu dari kelompok-kelompok tersebut berdasarkan jaraknya (Turban dkk, 2005). Semakin dekat jarak data individual, sebut saja X_1 dengan salah satu pusat dari kelompok yang ada , sebut saja A, maka semakin jelas bahwa X_1 tersebut merupakan anggota dari kelompok yang berpusat di A dan semakin jelas pula bahwa X_1 bukan anggota dari kelompok-kelompok yang lainnya (ilustrasi dapat dilihat pada gambar 1). Secara kuantitatif hal ini ditunjukkan melalui fakta bahwa d_{1A} yaitu jarak dari X_1 ke A mempunyai nilai yang paling kecil jika dibandingankan dengan



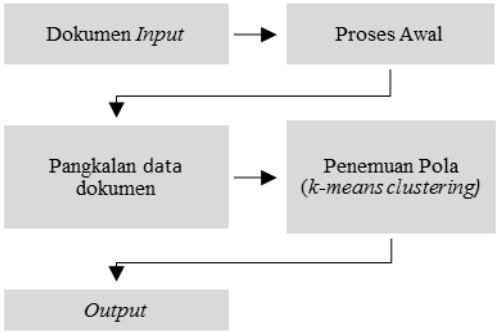
d_{1B} dan d_{1C}.

Gambar 1. Ilustrasi penentuan keanggotaan kelompok berdasarkan jarak

Untuk menemukan pusat yang paling sesuai sebagai upaya merepresentasikan posisi dari sebuah kelompok data terhadap kelompok data yang lainnya dilakukan sebuah proses perulangan. Proses perulangan ini dimulai dengan menentukan secara sembarang posisi dari pusat-pusat kelompok yang telah ditetapkan. Selanjutnya ditentukan keanggotaan setiap individu data berdasarkan jarak terpendek terhadap pusat-pusat tersebut. Pada iterasi kedua dan seterusnya dilakukan pembaharuan posisi pusat untuk semua kelompok. Selanjutnya dilakukan pembaharuan keanggotaan untuk setiap kelompok. Sebuah data yang semula adalah anggota kelompok C misalnya, dapat menjadi anggota kelompok B pada akhirnya. Proses perulangan ini akan berhenti setelah tidak terjadi lagi perubahan anggota kelompok, yang artinya jarak dari setiap anggota kelompok terhadap pusatnya masing-masing telah mencapai minimal dan jarak antara kelompok yang satu dengan kelompok yang lainnya telah mencapai maksimal. Pembaharuan pusat kelompok dilakukan dengan menghitung nilai rata-rata yang baru setelah adanya penambahan ataupun pengurangan anggota kelompok dari proses sebelumnya.

3. DESAIN SISTEM

Desain dari sistem automasi penentuan tren topik skripsi dapat dilihat pada gambar 2. Desain tersebut meliputi pemilihan dokumen *input*, proses awal, pangkalan data dokumen, penemuan pola dengan algoritma *k-means clustering* dan bentuk *output* dari sistem.



Gambar 2. Bagan desain sistem automasi penentuan tren topik skripsi

3.1 Dokumen *Input*

Input dari sistem adalah bagian abstrak dari buku skripsi. Tipe *file* yang digunakan dapat berupa .txt, .doc maupun .pdf.

3.2 Proses Awal

Proses awal meliputi tahapan *tokenizing*, *filtering*, *stemming* dan *tagging*. Algoritma yang dipilih untuk tahap *filtering* adalah *stop list*. Algoritma *word list* tidak dipilih mengingat tujuan dari aplikasi ini adalah untuk mengungkapkan kecenderungan topik dari berbagai macam skripsi tanpa memberi batasan apapun pada isi dari skripsinya. Dengan menerapkan *word list* akan banyak informasi baru yang tidak dapat direkam oleh sistem. Pada tahap *filtering*, atas pertimbangan efisiensi memori dan kecepatan eksekusi dipilihlah algoritma *porter* yang berbasikan pada konsep *affix removal*. Seluruh kata yang telah lolos proses *filtering*, *stemming* dan *tagging* selanjutnya disebut sebagai *term*. Untuk setiap *term*, dihitung kemunculannya pada setiap dokumen *input* dan hasilnya disimpan dalam pangkalan data dokumen untuk kemudian digunakan pada proses penemuan pola. Dengan demikian hasil dari tahap proses awal ini adalah sebuah matriks dua dimensi yang memuat informasi distribusi *term-term* dalam setiap dokumen yang menjadi target analisis.

3.3 Pangkalan Data Dokumen

Gambar 3 menunjukkan bentuk dari matriks distribusi *term* yang diperoleh dari tahap proses awal. Setiap sel pada matriks berisi frekuensi kemunculan *term* (direpresentasikan sebagai baris) pada dokumen tertentu (direpresentasikan sebagai kolom). Untuk efisiensi memori dan akurasi penemuan pola,

selanjutnya ditetapkan batasan (*threshold*) jumlah *term* yang akan dipertahankan di dalam matriks. Semua *term* yang frekuensi kemunculannya kurang dari 6 kali di semua dokumen yang diolah dianggap tidak representatif. *Term-term* tersebut selanjutnya dikeluarkan dari koleksi *term* yang ada. Dengan demikian hanya *term-term* yang berpotensi kuat untuk mengkarakterisasi kelompok saja yang akan dijadikan bahan pada tahap selanjutnya yaitu penemuan pola.

Term	310...	310...	310...	310...	310...	310...	310...	310...	310...	310...	310...	310...	310...	310...	310...	310...	31
analy...	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0		
artifi...	0	0	2	0	0	0	0	0	0	2	0	2	0	0	0		
avera...	0	0	2	0	0	0	0	0	0	0	0	2	2	2	0		
back...	0	0	2	0	0	0	0	0	0	0	0	2	0	0	2		
band...	0	0	0	0	7	0	0	7	0	0	0	0	0	0	0		
base6	1	1	0	1	0	0	1	0	1	0	1	0	0	1	0		
cellul...	0	0	0	0	0	5	5	0	0	0	0	0	0	0	0		
client8	0	0	0	6	6	0	0	0	0	0	0	0	0	0	0		
comp...	0	0	4	0	0	0	0	0	0	4	0	0	0	0	0		

Gambar 3. Cuplikan matriks distribusi *term* hasil dari proses awal seluruh dokumen

3.4 Penemuan Pola

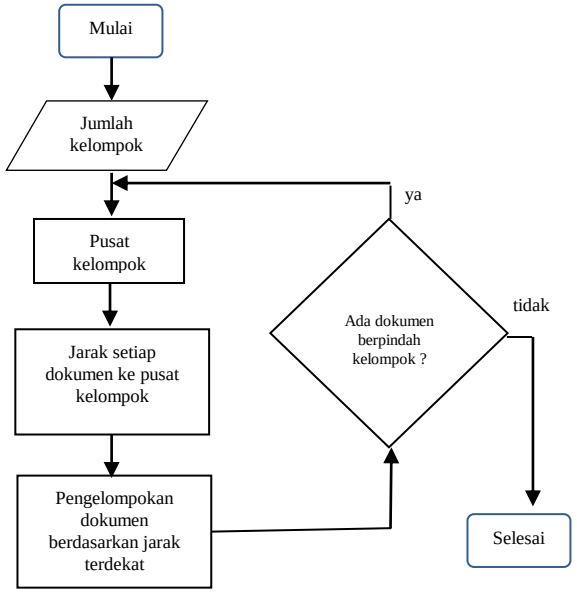
Dalam upaya mendapatkan kecenderungan topik skripsi, maka salah satu cara yang dapat ditempuh adalah dengan melakukan eksplorasi terhadap kecenderungan kemiripan dari tiap topik skripsi terhadap topik-topik skripsi lainnya. Kecenderungan semacam ini dapat dilihat dengan melakukan pemetaan kedekatan dari sebuah topik skripsi dengan topik-topik skripsi lainnya sehingga pada akhirnya akan terungkap kelompok-kelompok topik skripsi. Analisis kelompok adalah teknik yang paling sesuai untuk mendapatkan kelompok-kelompok besar dari topik-topik skripsi yang ada. Algoritma *k-means clustering* kemudian digunakan dalam penelitian ini untuk memfasilitasi analisis kelompok tersebut. Algoritma *k-means clustering* dimulai dengan penentuan jumlah kelompok (*cluster*) yang akan dibentuk (lihat gambar 4). Tahap selanjutnya adalah penentuan pusat dari setiap kelompok yang akan dibentuk. Untuk pertama kalinya pusat kelompok ini ditentukan secara acak. Namun, untuk mengurangi jumlah iterasi dan sekaligus mendapatkan hasil yang konvergen, maka dapat dilakukan modifikasi pada teknik inisialisasi pusat kelompok tersebut. Dengan menggunakan persamaan(1) penentuan pusat kelompok untuk yang pertama kalinya tidaklah sepenuhnya acak melainkan memperhatikan distribusi kemunculan *term* yang ada.

$$pusat = \left((n_t - n_r) * (R(1,2) - R.ND)\right) \quad (1)$$

n_r dan n_t berturut-turut adalah frekuensi terendah dan tertinggi yang terdapat pada matriks distribusi kemunculan *term.* R(1,2) adalah bilangan acak yang nilainya terletak diantara 1 sampai dengan 2 dan R.ND adalah sembarang bilangan acak. Selanjutnya dihitung jarak dari semua dokumen yang ada ke setiap pusat kelompok. Dokumen akan dikelompokkan ke pusat yang terdekat dengan menggunakan formula *cosine similarity* sebagaimana pada persamaan(2), $d1$ adalah vektor yang merepresentasikan kemunculan *term* pada dokumen 1 dan $d2$ adalah vektor yang merepresentasikan kemunculan *term* pada dokumen 2.

$$cosine(d1, d2) = \frac{(d1.d2)}{\|d1\|.\|d2\|} \quad (2)$$

Untuk setiap kelompok yang terbentuk, pusat disesuaikan. Pusat yang baru diperoleh dengan menghitung rata-rata frekuensi setiap *term* dari seluruh dokumen yang ada pada kelompok tersebut. Proses berulang terus sampai tidak satupun dokumen berpindah keanggotaan.



Gambar 4. Diagram alir proses *k-means clustering*

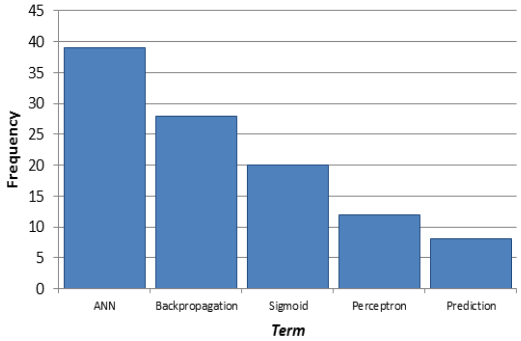
Untuk melakukan evaluasi terhadap hasil pengelompokan, maka dilakukan pengukuran *purity*. *Purity* menghitung rasio antara jumlah dokumen yang pengelompokannya benar dengan total dokumen yang dianalisis. Nilainya berkisar dari 0 sampai dengan 1. Semakin mendekati 1 mengindikasikan bahwa semakin banyak dokumen yang telah sesuai atau benar pengelompokannya. Sebaliknya, semakin mendekati 0 menunjukkan semakin sedikit dokumen yang benar pengelompokannya. Nilai *purity* diukur dengan menggunakan persamaan(3), Ω adalah himpunan dari dokumen-dokumen ω_k sedangkan C adalah himpunan dari kelompok-kelompok c_j.

$$purity(\Omega, C) = \frac{1}{N}\sum_{k} \max_{j} |\omega_k \cap c_j| \quad (3)$$

3.5 *Output*

Hasil pengelompokan dari tahap *k-means clustering* selanjutnya dilaporkan dengan menggunakan teknik pareto. Untuk setiap kelompok dibuat satu diagram pareto. Konsep dari diagram ini adalah melaporkan urutan kemunculan *term* secara menurun. Artinya *term* yang paling banyak muncul pada suatu kelompok tertentu akan diletakkan paling awal. Untuk mempermudah penarikan kesimpulan, jumlah *term* yang dimunculkan dalam diagram pareto dibatasi sesuai dengan kehendak pengguna. Pada gambar 5. ditampilkan contoh diagram pareto untuk sebuah kelompok yang *term*-nya dibatasi 5 saja. Selain laporan kemunculan *term*, ditampilkan juga data statistik dari proses *k*-

means clustering seperti jumlah dokumen dalam tiap kelompok dan banyak iterasi yang diperlukan.

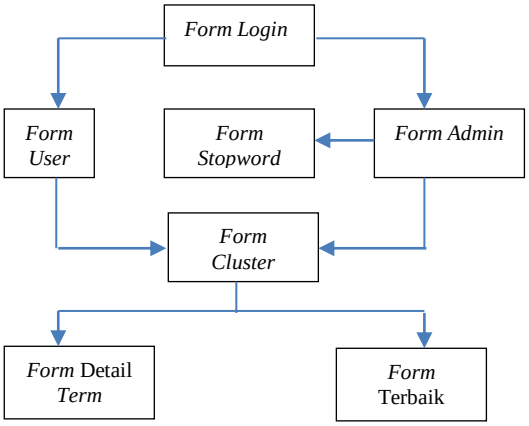


Gambar 5. Contoh diagram pareto untuk sebuah kelompok

4. HASIL DAN PEMBAHASAN

4.1 Perangkat Lunak

Untuk memfasilitasi pengguna berinteraksi dengan sistem automasi penentuan tren topik skripsi ini, dikembangkan sebuah antarmuka berupa aplikasi *desktop* yang dibangun dengan VB.Net. Secara umum fasilitas yang terdapat di dalam aplikasi ini dapat dilihat pada Gambar 6.



Gambar 6. Diagram fasilitas sistem automasi penentuan tren topik skripsi

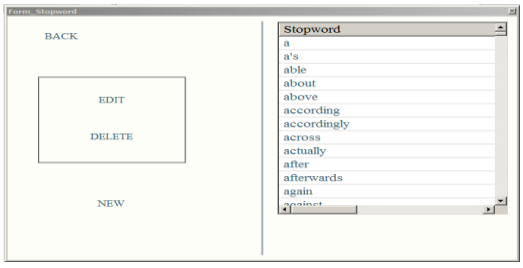
Pertama kali pengguna akan berinteraksi dengan *form login* (Gambar 7). Terdapat dua aturan pengguna yaitu sebagai *administrator* atau sebagai pengguna biasa (*user*). *Administrator* mempunyai hak untuk melakukan pengaturan hak akses pengguna biasa, mengunggah data kedalam pengkalan data, serta melakukan pengaturan untuk proses *filtering* dan *clustering*.

Gambar 7. Tampilan awal aplikasi

Gambar 8 menunjukkan halaman awal (*home*) aplikasi bagi *administrator* untuk dapat memulai pengaturan yang diperlukan. Melalui halaman tersebut *administrator* dapat menambahkan menu tahun masuk dari penulis skripsi (NEW ANGKATAN) dan juga jurusan atau program studinya (NEW JURUSAN). Fungsinya adalah untuk mempermudah pengguna menentukan batasan-batasan pada analisis tren yang akan dilakukannya. Misalkan ingin dilihat tren topik skripsi di program studi Teknik Informatika khusus mahasiswa angkatan 2007 saja, maka pengguna cukup memberi centang pada angkatan 2007 dan program studi Teknik Informatika. Melalui halaman ini pula *administrator* dapat menambahkan koleksi abstrak skripsi (INPUT ABSTRAK) kedalam pangkalan data yang akan menjadi bahan untuk analisis tren nantinya.

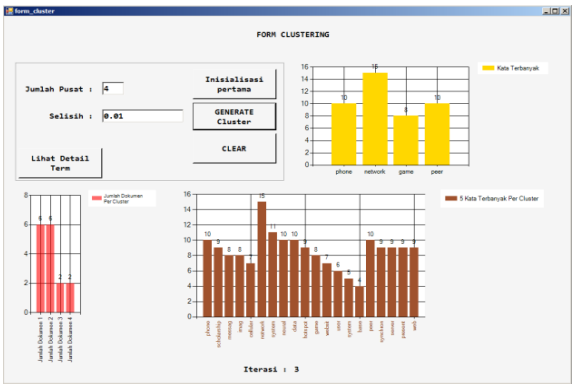
Gambar 8. Tampilan halaman awal saat *login* sebagai *administrator*

Pengaturan proses *filtering* dilakukan melalui sub menu StopWord (lihat gambar 9). Melalui halaman tersebut *administrator* dapat menambahkan maupun mengurangi kata-kata yang dijadikan *stop list* pada proses *filtering* dokumen. Proses *clustering* dilakukan dengan menekan tombol ANALISIS pada halaman awal. Pengguna kemudian akan diarahkan menuju halaman *cluster* sebagaimana yang dapat dilihat pada gambar 10.



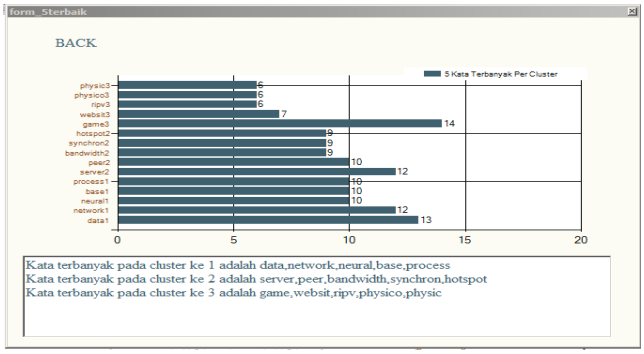
Gambar 9. Tampilan halaman untuk pengaturan *stop list*

Beberapa pengaturan yang dapat dilakukan adalah penentuan jumlah kelompok yang akan dibentuk dan nilai ambang pergeseran posisi pusat kelompok untuk menghentikan iterasi. Hasil proses *clustering* akan muncul di halaman yang sama dalam bentuk grafik batang. Statistik deskriptif yang ditampilkan untuk setiap kelompok adalah *term* terbanyak, jumlah dokumen, susunan lima *term* terbanyak dan jumlah iterasi.



Gambar 10. Tampilan halaman untuk proses *clustering*

Untuk mempermudah interpretasi, di akhir proses *clustering* ditampilkan halaman konklusi sebagaimana yang tampak pada Gambar 11. Setelah pengguna melihat informasi pada halaman konklusi dan dirasa kelompok yang dibentuk tidak representatif, maka proses *clustering* dapat diulangi kembali. Untuk kembali ke halaman *cluster* pengguna cukup menekan tombol BACK pada halaman konklusi.



Gambar 11. Tampilan halaman konklusi

Sampai dengan tahapan ini proses automasi dapat dikatakan berhenti, berikutnya diperlukan justifikasi dari pengguna untuk mempertimbangkan apakah konklusi yang diberikan oleh aplikasi telah cukup layak dan dapat digunakan untuk pengambilan kebijakan maupun analisis lanjutan. Oleh karena itu dapat dikatakan proses penentuan tren topik skripsi tidak sepenuhnya berjalan otomatis, namun semi-otomatis karena pada tahap akhir masih diperlukan justifikasi dari pengguna untuk menetapkan konklusinya.

4.2 Uji Coba

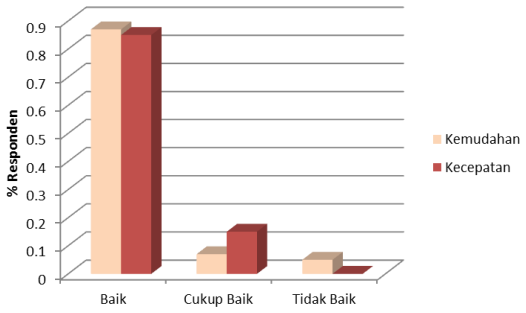
Uji coba pada aplikasi dilakukan baik dalam hal kinerja algoritma k-means clustering dalam melakukan pengelompokan, kemudahan dan kecepatan penggunaan aplikasi serata keakuratan konklusi yang diberikan. Uji coba dilakukan di Universitas Ma Chung. Bahan analisis adalah seluruh skripsi yang tersimpan di repository digital perpustakaan Universitas. Responden adalah beberapa kepala program studi dan dosen-dosen yang bertanggungjawab untuk skripsi di program studinya masing-masing.

Untuk memantau kinerja algoritma *k-means clustering* dalam melakukan pengelompokan terhadap dokumen-dokumen skripsi yang ada, telah dilakukan uji coba terhadap dokumen skripsi dari 6 program studi yang ada di Universitas Ma Chung. Hasilnya diperoleh rata-rata nilai *purity* sebesar 0.76 yang artinya sekitar 76% dokumen yang diolah telah berhasil dikelompokkan dengan benar oleh sistem. Ringkasan dari hasil uji coba tersebut dapat dilihat pada Tabel 1.

Tabel 1. Ringkasan nilai *purity* dari uji coba kinerja algoritma *k-means clustering*

Program Studi Asal Dokumen	*Purity*
Teknik Informatika	0.83
Sistem Informasi	0.89
Teknik Industri	0.77
Manajemen	0.64
Akuntansi	0.82
Bahasa Inggris	0.66
Rata-Rata	**0.76**

Gambar 12 menunjukkan pendapat responden terhadap aspek kecepatan dan kemudahan penggunaan. Dari kedua aspek tersebut lebih dari 80% responden menyatakan baik. Dengan kata lain selama menggunakan aplikasi, pengguna merasa proses untuk menghasilkan konklusi relatif cepat dan penataan feature pada halaman-halaman aplikasi sangat mudah dipahami sehingga tidak diperlukan bantuan lebih lanjut dari pengembang aplikasi untuk memahaminya. Untuk uji coba aspek akurasi konklusi, responden diminta menilai apakah rekomendasi tren topik skripsi yang dihasilkan oleh aplikasi telah sesuai dengan kondisi yang sebenarnya ditemui di Universitas. Metode yang diterapkan adalah responden diminta membandingkan tren topik skripsi berdasarkan pengalaman dan pemantauannya selama bergabung dengan Universitas Ma Chung dengan tren topik skripsi yang diperoleh melalui aplikasi. Sekitar 89% responden menyatakan tren topik skripsi yang diperoleh melalui aplikasi sesuai dengan tren topik skripsi berdasarkan pengalaman dan pemantauannya.



Gambar 12. Grafik hasil uji coba aspek kecepatan dan kemudahan penggunaan

Beberapa masukan lain dari responden adalah hendaknya performa dari algoritma yang digunakan ditingkatkan sehingga aplikasi dapat secara otomatis mengetahui apakah dokumen *input* berbahasa Inggris atau Indonesia. Untuk saat ini, informasi tersebut harus diberikan oleh pengguna saat mengunggah dokumen kedalam pangkalan data yang terhubung dengan aplikasi. Adapun tujuan menginformasikan bahasa yang digunakan adalah untuk kepentingan pemilihan kelompok *stop list* yang sesuai dan penentuan algoritma yang akan diterapkan saat proses *filtering* dan *tagging*.

5. KESIMPULAN

Proses automasi penentuan tren topik skripsi dapat dilakukan dengan menerapkan metode *text mining*. Dengan metode tersebut analisis tren topik skripsi dapat dilakukan dengan cepat karena meminimalkan keterlibatan dari pengguna. Melalui penelitian ini ditunjukkan bahwa algoritma *k-means clustering* yang digunakan dalam proses penemuan pola tebukti dapat membantu proses pengelompokan berbagai topik skripsi yang ada sehingga diperoleh informasi yang bermakna dalam menentukan tren penelitian Universitas dari tahun ke tahun. Uji coba terhadap aplikasi yang dikembangkan juga menunjukkan hasil yang positif baik dalam aspek kinerja algoritma *k-means clustering*, kemudahan dan kecepatan penggunaan aplikasi maupun akurasi konklusi.

6. REFERENSI

Gupta, N., "Text Mining for Information Retreival," Ph.D. thesis, Jaypee Institute of Information Technology University, India, May 2011.

Gupta, V., Lehal, G.S., "A Survey of Text Mining Techniques and Application," *Journal of Emerging Technologies in Web Intelligence*, Vol. 1, pp. 60-75, 2009.

Han, J., Kamber, M., *Data Mining Concept and Technique*, 2nd Ed, Elsevier, 2006.

Michael, B., "Automatic Discovery of Similar Words," *Survey of Text Mining: Clustering, Classification and Retrieval*, LLC, pp. 24-43, 2004.

Santoso R., "Aplikasi Katalog Online untuk Pencarian Konten Buku dengan Metode Text mining pada Perpustakaan STIKOM Surabaya," Skripsi, STIKOM, Surabaya, 2012.

Sari, O.Y., "Aplikasi Text Mining untuk Pencarian Buku Menggunakan Metode Association Rules Analysis Guna Meningkatkan Pelayanan Perpustakaan pada SMP Negeri 1 Plaosan Magetan," Skripsi, Universitas Muhammadiyah Ponorogo, 2012.

Turban, E., Aronson, J.E., Liang, T.P., *Introduction to Data Mining*, Pearson, 2005.

Yuwono, F., "Pembuatan Aplikasi Text Mining untuk Pencarian Buku Koleksi Skripsi dengan Menggunakan Association Rules Analysis," Skripsi, Universitas Kristen Petra, Surabaya, 2005.