

DETEKSI KESESUAIAN BIDANG MINAT TERHADAP PROPOSAL TUGAS AKHIR MAHASISWA STUDI KASUS : MAHASISWA SI UKDW

Nia Meliana
Umi Proboyekti, Jong Jek Siang

Abstrak

Pembuatan tugas akhir mahasiswa diharapkan sesuai dengan bidang minat yang diambil mahasiswa sehingga ilmu yang diperoleh dalam matakuliah bidang minat dapat diterapkan dalam tugas akhir mahasiswa. Untuk dapat mendeteksi kesesuaian tugas akhir mahasiswa dengan bidang minat digunakan metode pembobotan TF-IDF dan perhitungan probabilitas. Proses yang dilakukan dalam sistem meliputi proses menghilangkan stopword, stemming, pembobotan dengan metode TF-IDF, perhitungan probabilitas dan analisis deteksi kesesuaian. Hasil output sistem berupa hasil deteksi kesesuaian probabilitas kata yang termasuk SIB dan SIMM dan analisis hasil deteksi tugas akhir.

Dari hasil deteksi yang telah dilakukan dengan menggunakan 15 proposal skripsi angkatan 2008 menghasilkan 93,3333 % proposal skripsi yang dikatakan sesuai dengan bidang minat yang diambil mahasiswa dan 6,6667% proposal skripsi yang dikatakan tidak sesuai dengan bidang minat mahasiswa. Dari 93,3333 % proposal skripsi yang dikatakan sesuai dengan bidang minat terlihat 66,6666 % yang dikatakan mempunyai tingkat kesesuaian rendah dan 26,6667 % yang dikatakan mempunyai tingkat kesesuaian tinggi dengan bidang minat yang diambil oleh mahasiswa.

Kata Kunci : *probabilitas, stopword, stemming, TF-IDF*

1. Pendahuluan

Dalam Program Studi Sistem Informasi (SI) terdapat bidang minat yaitu Sistem Informasi Bisnis (SIB) dan Sistem Informasi Multimedia (SIMM). Dengan adanya pembagian bidang minat dalam Program Studi Sistem Informasi, diharapkan tugas akhir yang dilakukan oleh mahasiswa dapat mencerminkan bidang minat yang diambil. Akan tetapi, berdasarkan pengamatan yang dilakukan terhadap tugas akhir untuk Program Studi Sistem Informasi tidak sedikit tugas akhir mahasiswa yang kurang mencerminkan bidang minat yang diambilnya.

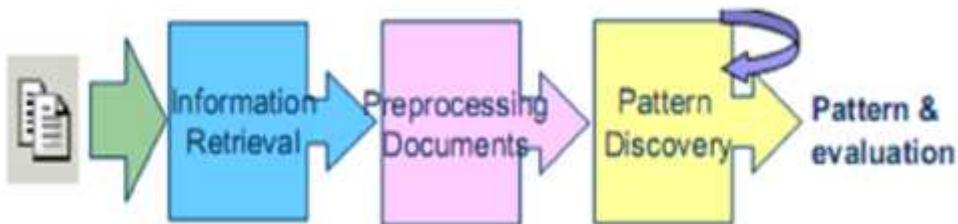
Permasalahan yang terjadi yaitu bagaimana mendeteksi apakah tugas akhir mahasiswa sesuai dengan bidang minat yang diambil oleh mahasiswa. Dengan permasalahan yang terjadi dibutuhkan metode yang dapat digunakan untuk mengatasi permasalahan tersebut. Metode yang digunakan dalam penelitian ini adalah metode pembobotan kata TF-IDF dan perhitungan probabilitas.

Penelitian ini hanya terbatas pada mahasiswa Program Studi Sistem Informasi. Melalui penelitian ini, diharapkan dapat membantu mahasiswa dalam mendeteksi apakah tugas akhir yang akan dilakukan sesuai dengan bidang minat yang ambil dan membantu dosen pengarah dalam mengarahkan tugas akhir mahasiswa agar sesuai dengan bidang minat yang diambil oleh mahasiswa yang bersangkutan. Penelitian ini menggunakan ringkasan dari proposal tugas akhir mahasiswa Program Studi Sistem Informasi sehingga mahasiswa yang sedang mengambil tugas akhir dapat mengetahui kesesuaian tugas akhir dengan bidang minat yang diambil sebelum penelitian untuk tugas akhir dilakukan.

2. Landasan Teori

2.1. Text Mining

Pada intinya text mining merupakan proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen Word, PDF dan kutipan teks. Proses kerja text mining sama dengan proses kerja data mining pada umumnya hanya saja yang ditambah adalah text databases. Tahapan dalam *text mining* digambarkan pada Gambar 1. Jika ada asumsi bahwa dokumen sudah relevan, maka tahap information retrieval dihilangkan.



Gambar 1. Tahapan dalam Text Mining

2.2. Term Frequency-Invers Document Frequency (TF-IDF)

TF-IDF merupakan suatu metode untuk memberikan bobot hubungan suatu kata/term terhadap dokumen. Metode TF-IDF ini menggabungkan dua konsep untuk perhitungan bobot yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen teks tertentu dan inverse frekuensi dokumen teks yang mengandung kata tersebut. Menurut Sholom M. Weiss (2005 : 30), persamaan yang dapat digunakan yaitu :

$$IDF_{(w)} = \log \frac{D}{DF_{(w)}} \dots \dots \dots [2.1]$$

Dalam penelitian ini, Persamaan 2.1 yang digunakan untuk memberi bobot pada kata yang terdapat dalam data pembobotan kata di database. Pada persamaan tersebut, D merupakan jumlah dokumen yang terdapat dalam corpus/koleksi dan $DF_{(w)}$ merupakan jumlah dokumen yang mengandung kata w. Setelah didapat nilai $IDF_{(w)}$ kemudian dihitung TF-IDF(d,w) dengan rumus :

$$TFIDF_{(d,w)} = TF_{(d,w)} * IDF_{(w)} \dots \dots \dots [2.2]$$

Pada persamaan 2.2 dapat terlihat bobot TF-IDF merupakan frekuensi kemunculan kata W dimodifikasi dengan sebuah faktor skala $IDF_{(w)}$. Persamaan $IDF_{(w)}$ secara sederhana menghitung jumlah dokumen yang berisi kata W($DF_{(w)}$) dan membalik skalanya. Ketika suatu kata muncul di beberapa dokumen, maka kata tersebut dipertimbangkan sebagai kata yang tidak penting dan nilai faktor skala akan rendah, bahkan mendekati nol. Demikian juga sebaliknya, jika kata bersifat unik dan muncul hanya di beberapa dokumen, faktor skala akan membesarkan karena kata tersebut bersifat penting.

2.3. Perhitungan Probabilitas

Persamaan perhitungan probabilitas yang digunakan sebagai berikut :

Persamaan dengan N1 menyatakan jumlah bobot kumpulan kata yang mengandung SI Bisnis.

Persamaan dengan N2 menyatakan jumlah bobot kumpulan kata yang mengandung SI Multimedia.

Sedangkan N adalah jumlah bobot kumpulan kata yang mengandung SI Bisnis dan SI Multimedia.

3. Perancangan Sistem

Secara garis besar, kerja sistem terbagi dalam tiga tahap, yaitu :

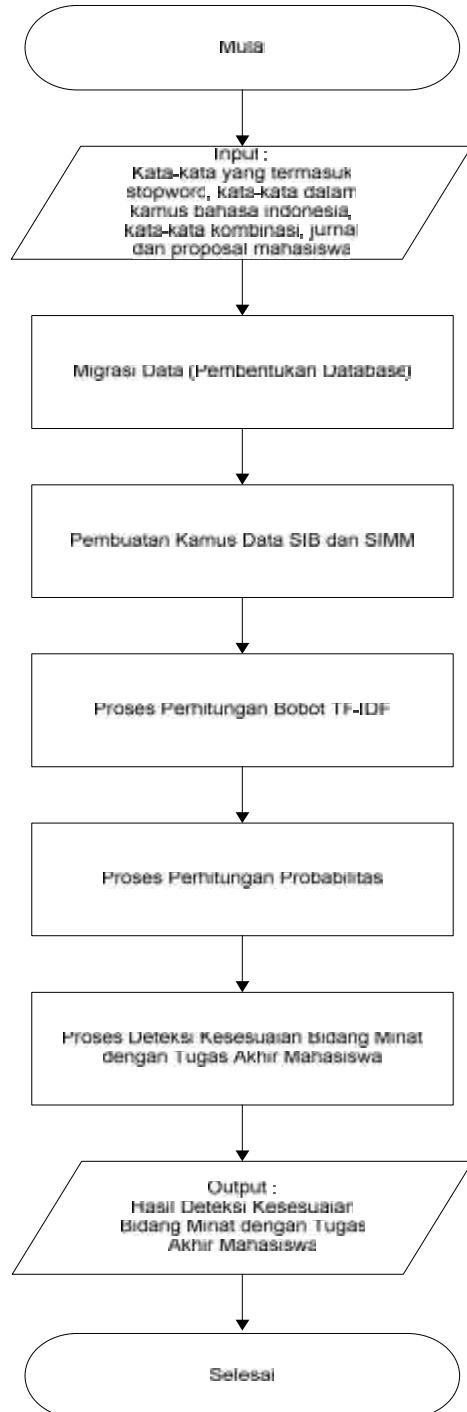
- a. Tahap *preprocessing* : meliputi proses migrasi data dan pembuatan Kamus Data SIB dan Kamus Data SIMM. Kata-kata dalam Kamus Data SIB dan SIMM berasal dari buku-buku dan jurnal-jurnal yang berkaitan dengan SIB dan SIMM. Kata-kata Kamus Data SIB dari jurnal sebagian besar berasal dari jurnal Manajemen Bisnis, jurnal Bisnis & Manajemen, jurnal Akuntansi FE Unsil, jurnal Akuntansi & Keuangan dan jurnal Akuntansi. Kata-kata dalam Kamus Data SIMM yang berasal dari jurnal sebagian besar berasal dari jurnal Multimedia.
 - b. Tahap Perhitungan Probabilitas SIB dan Probabilitas SIMM : meliputi perhitungan bobot kata yang termasuk SIB atau SIMM menggunakan metode TF-IDF dan perhitungan probabilitas. Hasil pembobotan kata dengan TF-IDF untuk setiap dokumen proposal mahasiswa akan disimpan dalam tabel yang berbeda sesuai dengan kategori kata SIB dan SIMM maka akan terbentuk 2 tabel yaitu hasil_tf_sib dan hasil_tf_simm. Hasil pembobotan dan nilai untuk setiap dokumen dari kedua tabel itulah yang akan digunakan dalam proses perhitungan probabilitas sehingga menghasilkan nilai probabilitas untuk SIB dan SIMM untuk setiap dokumen. Proses perhitungan probabilitas SIB dan SIMM yang dilakukan dapat diuraikan sebagai berikut :

Pengguna melakukan *input* jumlah dokumen yang akan dideteksi.

- i. Pengguna melakukan *input* data berupa NIM, nama mahasiswa, bidang minat yang diambil mahasiswa, judul tugas akhir, latar belakang masalah proposal dan rumusan masalah proposal sebanyak jumlah dokumen yang dimasukkan ke dalam sistem.
 - ii. Bila proses *input* berhasil, sistem melakukan proses pembersihan *stopword*. Setelah proses pembersihan selesai, kata-kata yang lolos dalam pembersihan disimpan dalam database kemudian dilakukan proses *stemming*. Setelah proses *stemming* selesai, dilanjutkan dengan proses pembobotan TF-IDF. Kemudian diperoleh bobot untuk kata-kata yang mengandung SIB dan SIMM. Dari hasil pembobotan kata SIB dan SIMM, dapat dihitung probabilitas kata yang mengandung SIB dan SIMM dengan perhitungan probabilitas.
 - c. Tahap Deteksi Kesesuaian : meliputi proses deteksi kesesuaian berdasarkan bidang minat dan analisis kesesuaian. Pada proses deteksi, bila probabilitas kata yang mengandung salah satu bidang minat yang diambil lebih besar dari probabilitas kata dari bidang minat yang lain maka dinyatakan sesuai dengan bidang minat tersebut. Suatu proposal dapat dikatakan memiliki tingkat kesesuaian tinggi jika selisih probabilitasnya lebih besar dari 0,4 dan jumlah kata yang termasuk dalam bidang minat yang diambil mahasiswa lebih dari 10 kata maka dikatakan

proposal mahasiswa memiliki tingkat kecocokan yang tinggi dalam bidang minat tersebut. *Output* yang dihasilkan oleh sistem adalah hasil probabilitas kata yang termasuk SIB, probabilitas kata yang termasuk SIMM dan hasil analisis deteksi kesesuaian.

d.



Gambar 2. *Flowchart* sistem yang akan dibangun

Pada flowchart Gambar 2 terlihat proses yang terjadi dalam sistem secara umum. *Input* sistem yang berupa kata-kata *stopword*, kata-kata kamus bahasa Indonesia dan kata-kata kombinasi disimpan dalam file excel selanjutnya ditransfer ke dalam SQL Server 2008 untuk dapat digunakan dalam sistem. *Input* sistem yang berupa jurnal digunakan untuk pembuatan Kamus Data SIB dan SIMM. Dalam proses pembuatan Kamus Data SIB dan SIMM dilakukan proses penyaringan dari kata-kata yang termasuk dalam *stopword*. Kemudian kata-kata yang tidak termasuk dalam *stopword* dianalisis sehingga menghasilkan kata-kata yang termasuk dalam SIB dan SIMM. *Input* sistem yang berupa proposal tugas akhir mahasiswa digunakan dalam proses proses deteksi kesesuaian. Dalam proses deteksi kesesuaian meliputi proses *stopword* untuk menghilangkan kata-kata yang tidak digunakan dan proses *stemming* untuk memperoleh kata dasar dari kata-kata yang tidak termasuk dalam *stopword*. Setelah proses *stemming* selesai dilanjutkan dengan proses pemisahan kata yang termasuk dalam SIB atau SIMM berdasarkan Kamus Data SIB dan SIMM. Setelah proses pemisahan kata dilanjutkan proses pembobotan kata TF-IDF yang kemudian dapat dihitung probabilitas SIB dan SIMM. Selanjutnya melakukan proses analisis deteksi kesesuaian dari hasil probabilitas SIB dan SIMM.

4. Implementasi Sistem

4.1.Tahap *Preprocessing*

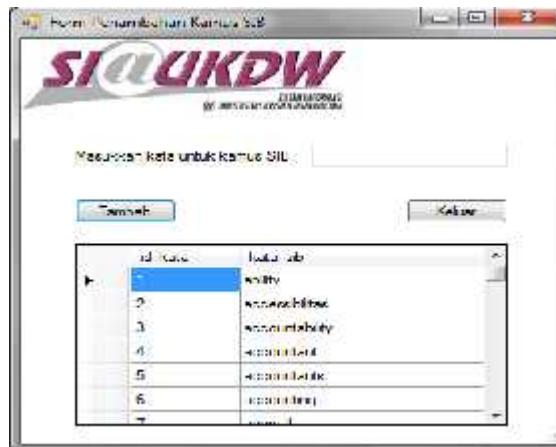
Tahap *Preprocessing* dimulai dengan proses awal yaitu pembuatan daftar kata untuk *stopword*. Untuk proses awal pembuatan daftar kata untuk *stopword* dilakukan dengan proses transfer dari file excel dengan ekstensi .csv yang didapat dari blog Bapak Budi Susanto ke dalam SQL SERVER 2008. Sistem juga menyediakan fasilitas untuk menambah kata dalam *stopword*. Tampilan yang disediakan untuk menambah kata dalam *stopword* terlihat pada Gambar 3.



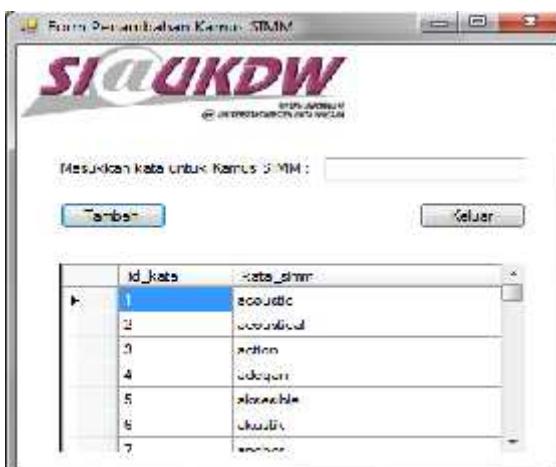
Gambar 3. Tampilan Menu *Stopword*

Setelah proses pembuatan *stopword* selesai, dilanjutkan dengan proses pembuatan Kamus Data SIB dan SIMM. Untuk pembuatan Kamus Data SIB dan SIMM dibutuhkan buku-buku dan jurnal-jurnal yang berkaitan dengan SIB dan SIMM. Pembuatan Kamus Data SIB dan SIMM yang melalui jurnal dilakukan dengan pembuangan kata-kata yang termasuk dalam *stopword* kemudian dapat dilihat kata-kata yang muncul berjumlah dibawah 10 dan mempunyai arti dalam SIB dan

SIMM yang dapat masuk dalam Kamus Data SIB dan SIMM. Kemudian kata-kata yang berkaitan dengan SIB dan SIMM dimasukkan ke dalam file excel dengan ekstensi .csv selanjutnya ditransfer ke dalam tabel untuk Kamus Data SIB dan SIMM yang terdapat pada SQL SERVER 2008. Kata-kata yang berhasil diperoleh dalam Kamus Data yaitu 476 kata untuk SIB dan 221 kata untuk SIMM. Kata-kata dalam Kamus Data SIB dan SIMM dapat ditambah dengan menggunakan tampilan antarmuka pada Gambar 4 dan Gambar 5 yang disediakan oleh sistem.



Gambar 4. Tampilan Menu Kamus Data SIB



Gambar 5. Tampilan Menu Kamus Data SIMM

4.2 Tahap Perhitungan Probabilitas dan Analisis Kesesuaian

Pada tahap perhitungan probabilitas dan analisis kesesuaian disediakan menu Deteksi Kesesuaian dalam sistem. Menu Deteksi Kesesuaian digunakan untuk melakukan proses deteksi terhadap proposal tugas akhir mahasiswa. Dalam menu Deteksi Kesesuaian diperlukan *input* berupa jumlah dokumen yang dimasukkan ke dalam sistem. Tampilan awal untuk proses deteksi terlihat pada Gambar 6.



Gambar 6. Tampilan untuk *input* jumlah dokumen

Pada tampilan Gambar 6 terlihat bahwa jumlah dokumen harus mengalami uji validasi sehingga *input* yang dimasukkan sesuai dengan yang diperlukan sistem. Selanjutnya memasukkan NIM, nama, bidang minat, judul tugas akhir, latar belakang dan rumusan masalah dari proposal tugas akhir mahasiswa sebanyak jumlah dokumen yang dimasukkan ke dalam sistem. Dalam penelitian ini, digunakan 15 proposal tugas akhir mahasiswa SI angkatan 2008. Proses yang dilakukan terhadap 15 proposal meliputi proses stopword, stemming, pemisahan kata yang termasuk SIB atau SIMM, pembobotan TF-IDF untuk setiap kata yang termasuk SIB atau SIMM, perhitungan probabilitas SIB dan SIMM dan analisis kesesuaian. Dari proses yang telah dilakukan terhadap 15 proposal mahasiswa dihasilkan hasil kesesuaian pada Tabel 1.

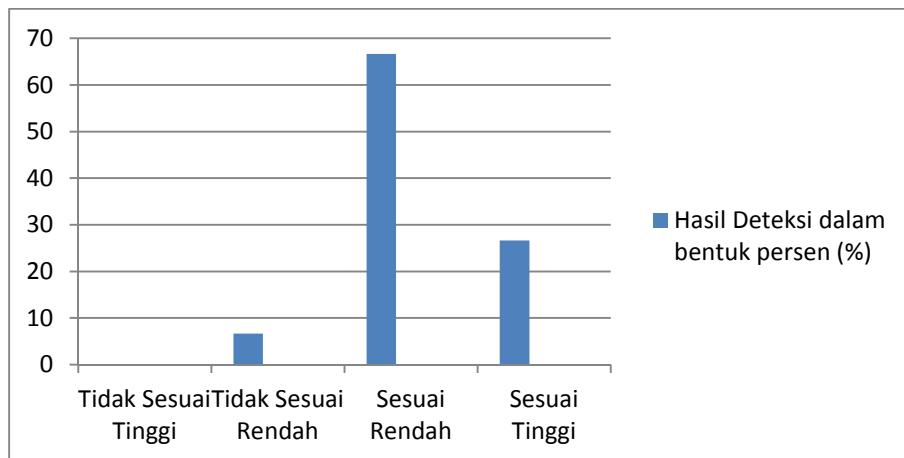
Tabel 1. Hasil Deteksi Kesesuaian

No	NIM	Nama	Skor	Uji Validasi	Pada Tgl	Pada Jam	Sesi	Deteksi	Hasil	Total Kesesuaian
1	200801	Adi Surya	SE	Hasilnya adalah jumlah dokumen yang dimasukkan	10/11/2012	07:00	09:00	Siap	Tidak	
2	200802	ABDUL RAHMAN	SE	SEMANTIK DAN KONSEP TRANSFORMASI DENGAN MENGGUNAKAN FIZIK	10/11/2012	07:00	10:00	Siap	Tidak	
3	200803	Devi Puspita	SE	Universitas Islam Negeri Syarif Hidayah Jakarta	10/11/2012	07:00	10:00	Siap	Tidak	
4	200804	Fitri Rahayu	SE	Pengaruh Kultur dan Pengembangan Sosial Masyarakat di Desa	10/11/2012	07:00	10:00	Siap	Tidak	
5	200805	Uzairi	SE	Penyebarluasan Sistem Sistem Pendukung Keputusan di Australia	10/11/2012	07:02	07:06	Siap	Tidak	
6	200806	Widya	SE	PERAMALAN PENGETAHUAN PENGETAHUAN DENGAN METODE	10/11/2012	07:00	10:00	Siap	Tidak	
7	200807	Andi Dwiyan	SE	Hasilnya jumlah file pada folder yang dimasukkan	10/11/2012	07:03	07:04	Siap	Tidak	
8	200808	Amrik Syah	SEW	IMPLEMENTASI PENERAPAN SISTEM INFORMASI UNIVERSITAS TIKARAH	10/11/2012	07:00	07:45	Siap	Tidak	
9	200809	Sari Mayati	SE	SI SISTEM PENGETAHUAN DENGAN METODE NEURONAL	10/11/2012	07:00	10:00	Siap	Tidak	
10	200810	Yeni Puspitasari	SE	Siaran Radio Pemuda Pekajar Cibinong Bogor	10/11/2012	07:00	10:00	Siap	Tidak	
11	200811	Imam Mulyadi	SE	Sistem Penerapan Sistem Pengetahuan Dalam Penerapan Pendekar	10/11/2012	07:00	10:00	Siap	Tidak	
12	200812	THEOPHILE ENVISON	SEW	IMPLEMENTASI SISTEM PENGETAHUAN DENGAN METODE SENSATION	10/11/2012	07:00	10:00	Siap	Tidak	
13	200813	Zainul Syaiful	SE	SI SISTEM PENGETAHUAN DENGAN METODE NEURONAL	10/11/2012	07:00	07:30	Siap	Tidak	
14	200814	W. Wicak	SE	Deteksi kesesuaian yang dimasukkan dengan yang diambil	10/11/2012	07:00	07:00	Tidak Siap	Tidak	
15	200815	Iskandar Puspita	SEW	IMPLEMENTASI SISTEM PENGETAHUAN DENGAN METODE SENSATION	10/11/2012	07:00	07:30	Siap	Tidak	

Hasil proses deteksi kesesuaian dari setiap dokumen pada Tabel 1 dapat menjelaskan bahwa sistem dapat memproses deteksi dengan baik. Hal ini terlihat dari sebagian besar dokumen memiliki dinyatakan sesuai dengan bidang minat yang diambil oleh mahasiswa. Dari hasil proses deteksi

pada Tabel 1 menyatakan bahwa 14 proposal skripsi yang dinyatakan sesuai dengan bidang minat yang diambil oleh mahasiswa dan 1 proposal skripsi yang dinyatakan tidak sesuai dengan bidang minat.

Berdasarkan hasil keseluruhan dari proses deteksi kesesuaian terhadap 15 dokumen proposal skripsi mahasiswa angkatan 2008 dapat diambil kesimpulan bahwa menghasilkan 93,3333 % proposal skripsi yang dikatakan sesuai dengan bidang minat yang diambil mahasiswa dan 6,6667% proposal skripsi yang dikatakan tidak sesuai dengan bidang minat mahasiswa. Dari 93,3333% proposal skripsi yang dikatakan sesuai dengan bidang minat terlihat 66,6666% yang dikatakan mempunyai tingkat kesesuaian rendah dan 26,6667% yang dikatakan mempunyai tingkat kesesuaian tinggi dengan bidang minat yang diambil oleh mahasiswa. Jadi dapat disimpulkan faktor-faktor yang dapat mempengaruhi kesesuaian proposal skripsi meliputi jumlah kata yang mengandung SIB dan SIMM, jumlah dokumen yang mengandung kata SIB dan SIMM, kata-kata yang terdapat dalam Kamus Data SIB dan SIMM, probabilitas SIB, probabilitas SIMM dan selisih probabilitas SIB dan SIMM. Hasil proses deteksi dalam bentuk grafik dapat terlihat pada Gambar 7.



Gambar 7. Tampilan Grafik Hasil Deteksi Kesesuaian

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan penelitian ini antara lain :

- Berdasarkan proses deteksi yang dilakukan terhadap 15 proposal skripsi mahasiswa angkatan 2008 menghasilkan menghasilkan 93,3333 % proposal skripsi yang dikatakan sesuai dengan bidang minat yang diambil mahasiswa dan 6,6667% proposal skripsi yang dikatakan tidak sesuai dengan bidang minat mahasiswa.
- Semakin sering satu kata muncul dalam sebuah dokumen maka akan membuat bobot TF-IDF dari kata tersebut bernilai besar. Sebaliknya semakin sering suatu kata muncul dalam kumpulan dokumen membuat bobot TF-IDF dari kata tersebut bernilai kecil.
- Kata-kata dalam Kamus Data SIB dan Kamus Data SIMM mempunyai pengaruh dalam menentukan kesesuaian dan tingkat kesesuaian tugas akhir mahasiswa dengan bidang minat yang diambil oleh mahasiswa dibuktikan dengan jika kata-kata dalam dokumen hanya sedikit yang sama dengan kata dalam Kamus Data SIB dan SIMM maka probabilitas SIB dan SIMM juga memiliki nilai yang kecil.

5.2 Saran

- Saran untuk mengembangkan sistem antara lain :
- Menambahkan jumlah kata dalam Kamus Data SIB dan SIMM sehingga hasil deteksi yang dihasilkan lebih akurat.
 - Menambah algoritma pembentukan frase kata dalam membuat Kamus Data SIB dan SIMM.

Daftar Pustaka

- Berry, Michael J. A. and Linoff, Gordon. (2000). *Mastering Data Mining : The Art and Science of Customer Relationship Management*. America : John Wiley & Sons.
- Dwi, Dewa Ayu. (2008). Sinyal Jual Beli Saham Menggunakan Metode Naïve Bayes. Skripsi S1. Yogyakarta : UKDW.
- Feldman, Ronen and Sanger, James. (2007). *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*. America : Cambridge.
- Hernawan. (2009). Sistem Klasifikasi Artikel Blog dengan Metode Naïve Bayes. Skripsi S1. Yogyakarta : UKDW.
- Kantsrdzic, Mehmed. (2003). *Data Mining : Concepts, Models, Methods, and Algorithms*. New Jersey : John Wiley & Sons.
- Konchady, Manu. (2006). *Text Mining Application Programming*, Canada : Thomson.
- Santosa, Budi. (2007). *Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta : Graha Ilmu.
- Kroenke, David. (1989). *Management Information Systems*, Singapore : McGraw-Hill, Inc.
- Kurniawan, Erick. (2011). *Membangun Aplikasi Mobile dengan Qt SDK*. Yogyakarta : Andy Offset.
- Madcoms. (2004). *Tutorial Teknik Memanipulasi Desain Grafis dan Photo dengan Corel PHOTO-PAINT Graphich Suite 11*. Yogyakarta : Andy Offset.
- Madcoms. (2005). *Manipulasi Dan Memperbaiki Foto Digital Dengan Adobe Photoshop CSI*. Yogyakarta : Andy Offset.
- Mann, Anthony T. (2003). *.Net Web Service For Dummies*, Canada : Wiley Publishing, Inc.
- Oetomo, Budi Sutedjo Dharma. (2002). *Perencanaan & Pembangunan Sistem Informasi*. Yogyakarta : Andy Offset.
- Steinmetz, Ralf and Nahrstedt, Klara. (1995). *Multimedia : Computing, Communications And Applications*. America : Prentice-Hall, Inc.
- Susanto, Budi. *Stopword*, <http://lecturer.ukdw.ac.id/budsus/pdf/stopwordID.csv> (Tanggal akses 30 September 2011).
- Tala, Fadillah. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, 6-10.
- Troy. (2004). Program Bantu Prediksi Missing Value dalam Data Nilai Huruf Mahasiswa menggunakan Metode Data Mining : Naïve Bayes dan Decision Tree. Skripsi S1. Yogyakarta : UKDW.
- Weiss, Sholom M. (2005). *Text Mining : Predictive Methods for Analyzing Unstructured Information*. America : Springer.