SISTEM AUTO MATCHING DOKUMEN PROPOSAL TUGAS AKHIR BERBAHASA INDONESIA MENGGUNAKAN JACCARD MEASURE

Reny Octaviana (1210651082)
Email : renyocta93@gmail.com

Program Studi Teknik Informatika, Fakultas Teknik,
Universitas Muhammadiyah Jember

ABSTRAK

Setiap tahun akademis banyak mahasiswayang membuat proposal tugas akhir untuk diajukan sebagai Tugas Akhir mahasiswa sebagai salah satu persyaratan untuk memperoleh gelar Sarjana (S1) . Seringkali ditemukan tingkat kemiripan topik atau isi dokumen, antara satu dokumen proposal dengan dokumen proposal mahasiswa lainnya.

Text Mining bisa dianggap subyek riset yang tergolong baru. *Text mining* dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokan dan menganalisa *unstruktured* teks dalam jumlah besar. Algoritma *jaccard measure* merupakan metode yang memanfaatkan nilai kemiripan dokumen. Penggunaan metode *jaccard measure* dikarenakan nilai akurasi kemiripan dari metode *jaccard measure* yang mendekati nilai keakuratan sangat tinggi.

1.1.Latar Belakang

Dalam dokumen entri sering terjadi duplikasi, maka dari itu dapat menggunakan metode jaccard meausre agar tahu tingkat kemiripan dokumen satu dengan dokumen lainnya. Perlu adanya sistem Auto Maching untuk mengetahui tingkat kemiripan dalam proposal salah satu metode kemiripan dokumen yang banyak digunakan adalah Jaccard Measure.

Algoritma *jaccard mesure* merupakan metode yang memanfaatkan nilai kemiripan dokumen. Penggunaan metode *jaccard measure* dikarenakan nilai akurasi kemiripan dari metode *jaccard measure* yang mendekati nilai keakuratan sangat tinggi. *Jaccard measure* juga merupakan salah satu teknik yang dapat dipergunakan untuk melakukan analisis dalam menghitung tingkat kemiripan suatu dokumen dengan tujuan menghasilkan perolehan yang optimal. *Jaccard measure* yaitu dengan cara memberi pembobotan pada suatu document proposal tugas akhir menghitung *Term Frequency/Inverse Document Frequency (TF/IDF)*. Metode Jaccard Measure merupakan salah satu metode dengan konsep similarity yang digunakan untuk menghitung kesamaan

dari beberapa dokumen. Dengan menggunakan metode ini, dapat memaksimalkan penilaian dengan mengukur tingkat kesamaan antar dokumen – dokumen.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang diuraikan sebelumnya, terdapat beberapa permasalahan yang akan diangkat dalam proposal tugas akhir ini, antara lain :

1. Bagaimana mengetahui jumlah tingkat kemiripan antara dokumen satu dengan dokumen lainnya.
2. Bagaimana melakukan pra-proses menggunakan *text mining* dan perhitungan TF/IDF.
3. Bagaimana penerapan metode jaccard measure untuk mendapatkan derajat kemiripan antara dokumen.

1.3. Batasan Masalah

Agar tidak menyimpang jauh dari permasalahan, maka Tugas Akhir ini mempunyai batasan masalah sebagai berikut:

1. Dataset berupa 34 dokumen dari bab 1, 2, dan 3 Proposal Tugas Akhir prodi Teknik Informatika pada tahun 2008-2013 berbahasa Indonesia.
2. Text mining digunakan pada dokumen berbahasa Indonesia.
3. Metode yang digunakan adalah *jaccard measure*.
4. Kemiripan dokumen berdasarkan lampiran antara bab dan keselarasan dalam data semua bab.

1.4. Tujuan Penelitian

Tujuan dari Tugas Akhir ini antara lain :

1. Untuk mengetahui jumlah tingkat kemiripan antara dokumen satu dengan dokumen lainnya.
2. Mengolah data atau dokumen dengan *text mining* dan menghitung bobot dengan TF/IDF.
3. Untuk menerapkan metode jaccard measure agar mendapatkan derajat kemiripan antara dokumen.

1.5. Manfaat

Tujuan dari Tugas Akhir ini antara lain :

1. Untuk mengetahui jumlah tingkat kemiripan antara dokumen satu dengan dokumen lainnya.

2. Mengolah data atau dokumen dengan *text mining* dan menghitung bobot dengan TF/IDF.
3. Untuk menerapkan metode jaccard measure agar mendapatkan derajat kemiripan antara dokumen.

2.1. Proposal Tugas Akhir

Proposal Tugas Akhir adalah sebuah dokument dimana masih sebagai perencanaan untuk meneruskan pada tugas akhir. Proposal Tugas Akhir dibuat untuk melangkah menuju tugas akhir untuk syarat kelulusan S1. Tugas akhir adalah kegiatan mandiri yang dilakukan dalam rangka untuk memenuhi sebagai persyaratan kelulusan mahasiswa. Dalam tugas akhir atau skripsi berisi tentang penelitian atau perancangan yang berdasarkan rasional tertentu yang dinilai penting dan bermanfaat ditinjau dari beberapa segi. Skripsi adalah karya tulis ilmiah resmi akhir seorang mahasiswa dalam menyelesaikan Program Sarjana (S1). Skripsi merupakan bukti kemampuan akademik mahasiswa dalam penelitian yang berhubungan dengan masalah yang dibahas.

2.2. *Text Mining*

Text mining bisa dianggap subyek riset yang tergolong baru. *Text mining* dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokkan dan menganalisa *unstructured* teks dalam jumlah besar. Dalam memberikan solusi, *text mining* mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *Data mining, Information retrieval, Statistik dan Matematik, Machine Learning, Linguistic, Natural Languange Processing, dan Visualization.* Kegiatan riset untuk *text mining* antara lain ekstraksi dan penyimpanan teks, *preprocessing* akan konten teks, pengumpulan data statistik dan *indexing* dan analisa konten.

2.3. TF/IDF

Metode TF/IDF (Robertson, 2005) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan *inverse document frequency* yang mengandung kata tersebut. Frekuensi kemunculan kata didalam dokumen yang diberikan menunjukkan seberapa penting kata tersebut

didalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi didalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen (*database*).

2.4. *Jaccard Measure*

Jaccard Coeficient adalah salah satu metode yang dipakai untuk menghitung *similarity* antara dua objek *(items)*. Seperti halnya *cosine distance* dan *matching coefficient,* secara umum perhitungan metode ini didasarkan pada *vector space similarity measure*. *Jaccard similarity* atau *Jaccard Coefficient* (Tan et.all, 2005) menghitung *similarity* antara dua objek, X dan Y yang dinyatakan dalam dua buah vektor, sebagai berikut:

$$Sim(D_1, Q_1) = \frac{\sum_{k=1}^{n}(D_{1,k} \times Q_{1,k})}{\sum_{k=1}^{n}(D_{1,k}) + \sum_{k=1}^{n} Q_{1,k} - \sum_{k=1}^{n}(D_{1,k} \times Q_{1,k})}$$

2.5. *Precision* Dan *Recall*

Precision merupakan salah satu parameter pengukuran hasil *retrival* terhadap dokumen. Dengan kata lain, *precision* dapat diartikan sebagai kecocokan antara permintaan informasi dengan respon dari permintaan tersebut. *Precision* dapat dihitung dengan :

$$Precision = \frac{\text{jumlah dokumen relevan yang } terretrieve}{\text{jumlah seluruh dokumen}}$$

Sedangkan *recall* merupakan parameter yang didapat dari jumlah dokumen terambil yang relevan dibagi dengan keseluruhan jumlah dokument yang relevan. *Recall* digunakan untuk melakukan pengukuran terhadap tingkat keberhasilan sistem dalam mengenali suatu dokumen yang relevan terhadap kueri.

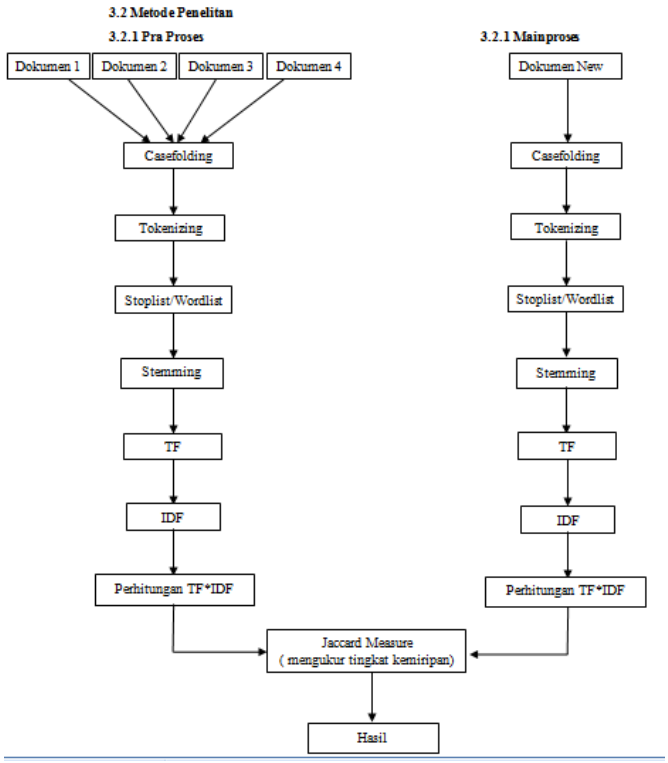
$$recall = \frac{\text{jumlah dokumen relevan yang teretrieve}}{\text{jumlah seluruh dokumen relevan dalam koleksi}}$$

3.1 Tahap Penelitian

Dalam tahap penelitian sistem Auto Matching Proposal Tugas Akhir menggunaka Jaccard Measure ini diperlukan langkah – langkah kegiatan penelitian untuk mendapatkan hasil yang maksimal. Untuk itu penulisan merencanakan beberapa langkah - langkah yang dapat memaksimalkan dalam pengerjaan sistem Auto Matching Proposal Tugas Akhir ini.

3.2 Metode Penelitian

Berikut ini adalah tahapan – tahapan dari proses penelitian. Dalam tahapan ini dilakukan beberapa tahapan dari proses penelitian dengan metode *Jaccard Measure.*



4.2 Skenario Pengujian

Skenario pengujian dilakukan dengan perlakukan pada aspek *query*. Perlakukan *query* tersebut menggunakan dengan Judul, *query* Bab I, *query* Bab II dan *query* Bab III.

- **Skenario 1 *Query* Judul** : Aplikasi Pemilihan Bidang Konsentrasi Di Ti Umj Berbasis Algoritma Id3

maka akan dihitung tingkat kemiripan *query* Judul dengan dokumen (D) untuk menghitung tingkat kesamaan penulis menggunakan persamaan metode *jaccard measure*, rumus metode *jaccard measure* seperti brikut:

$$Sim(D_i, Q) = \frac{\sum_{k=1}^{n}(D_{i,k} \times Q_{i,k})}{\sum_{k=1}^{n}(D_{i,k}) + \sum_{k=1}^{n} Q_{i,k} - \sum_{k=1}^{n}(D_{i,k} \times Q_{i,k})}$$

$$Sim(D_1, Q) = \frac{1{,}321}{2{,}078 + 2{,}078 - 1{,}321}$$

$$= \frac{1{,}321}{2{,}835} = 0{,}466$$

Dari langkah-langkah tersebut maka menghasilkan *similarity query* pada tiap dokumen (D) *similarity* selengkapnya dapat dilihat pada tabel 4.6 dibawah ini.

Tabel 4.6. Tabulasi Hasil Skenario 1

No	Dokumen	Kemiripan
1.	D13	0,466
2.	D3	0,385
3.	D21	0,381
4.	D2	0,156
5.	D6	0,156
6.	D17	0,156
7.	D24	0,156
8.	D32	0,156
9.	D31	0,152
10.	D20	0,148
11.	D4	0,034
12.	D7	0,034
13.	D9	0,034
14.	D12	0,034
15.	D14	0,034
16.	D19	0,034
17.	D33	0,034
18.	D18	0,028
19.	D29	0,028
20.	D1	0,022
21.	D5	0,022
22.	D8	0,022
23.	D11	0,022
24.	D22	0,022
25.	D23	0,022
26.	D34	0,022
27.	D16	0,016
28.	D30	0,016
29.	D28	0,015
30.	D15	0,009
31.	D25	0,008
32.	D10	0,001
33.	D27	0,001
34.	D26	0,000

4.3 Analisis Hasil Pengujian

Analisis hasil pengujian dilakukan dengan melakukan pencarian dengan mengevaluasi hasil perangkingan dokumen. Analisis pengujian dilakukan terhadapat sejumlah inputan *user*, mulai inputan ke-1 hingga ke-n. Dari masing-masing input tersebut dapat diperoleh nilai akurasi sebagai nilai evaluasi kemampuan sistem yang ditunjukkan pada tabel 4.10.

Tabel 4.10. Tabulasi Tingkat *Similarity Treshold* 0,1

No	Proposal	Similarity									
1	Judul	D13	D3	D21	D2	D6	D17	D24	D32	D31	D20
		0.466	0.385	0.381	0.156	0.156	0.156	0.156	0.156	0.152	0.148
2	BAB I	D13	D31	D33	D17	D9	D12	D19	D31		
		0.317	0.265	0.157	0.140	0.139	0.134	0.129	0.107		
3	BAB II	D34	D18	D21	D17	D16	D12	D32	D9	D24	D22
		0.531	0.179	0.140	0.140	0.111	0.110	0.109	0.106	0.102	0.101
4	BAB III	D34	D17	D32	D18	D21					
		0.466	0.385	0.381	0.156	0.156					

Tabel 4.11. Tabulasi Tingkat *Similarity Treshold* 0,2

No	Proposal	Similarity		
1	Judul	D13	D3	D21
		0.466	0.385	0.381
2	BAB I	D13	D31	
		0.317	0.265	
3	BAB II	D34		
		0.531		
4	BAB III	D34	D17	D32
		0.466	0.385	0.381

Tabel 4.12. Tabulasi Tingkat *Similarity Treshold* 0,3

No	Proposal	Similarity		
1	Judul	D13	D3	D21
		0.466	0.385	0.381
2	BAB I	D13		
		0.317		
3	BAB II	D34		
		0.531		
4	BAB III	D34	D17	D32
		0.466	0.385	0.381

Tabel 4.13. Tabulasi Tingkat *Similarity Treshold* 0.4

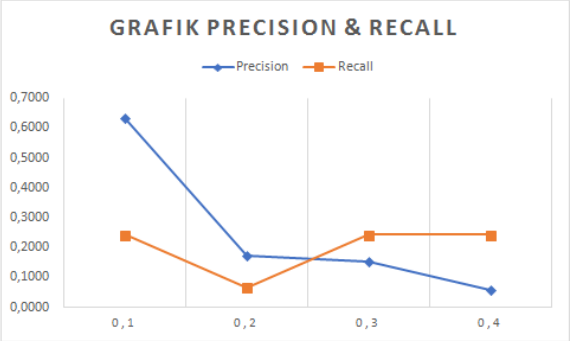
No	Proposal	Similarity	
1	Judul	D13	
		0.466	
2	BAB I		
3	BAB II		
4	BAB III	D34	
		0.466	

Berdasarkan hasil Analisis hasil pengujian dengan 4 kali percobaan dengan dokumen dataset berjumlah 34, maka akan terdapat hasil *precision*, dan *recall* sebagai berikut :

Tabel 4.14. Tingkat *Precision* dan *Recall*

No	Query	Threshold							
		0.1		0.2		0.3		0.4	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
1	Q1	0.76923	0.294	0.23077	0.088	0.23077	0.294	0.07692	0.294
2	Q2	0.61538	0.235	0.15385	0.059	0.07692	0.235	0	0.235
3	Q3	0.76923	0.294	0.07692	0.029	0.07692	0.294	0.07692	0.294
4	Q4	0.38462	0.147	0.23077	0.088	0.23077	0.147	0.07692	0.147
	rata - rata	0,6346	0,2425	0,1731	0,0660	0,1538	0,2425	0,0577	0,2425

Gambar 4.4. Grafik Tingkat *Precision* dan *Recall*



Dari hasil pengujian rata – rata yang telah dilakukan dengan 4 pengujian *query* yang berbeda terhadap dokumen berjumlah 34. Maka diperoleh nilai rata *precision* dan *recall* yang dihasilkan oleh sistem terletak pada threshold 0.1 yaitu dengan precision 0,6346 dan recall 0,2425. Pada penelitian ini nilai perhitungan tersebut akan menjadi acuan nilai akurasi yang dihasilkan sistem *auto matching* dokumen proposal tugas akhir berbahasa indonesia menggunakan *jaccard measure*.

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan bahwa :

1. Penerapan metode *TF-IDF* pada *auto matching* proposal tugas akhir dokumen berbahasa Indonesia diperoleh nilai *precision* dan *recall* yang

dihasilkan oleh sistem terbaik pada threshold 0.1 yaitu dengan precision 0,6346 dan recall 0,2425.

2. Metode *Jaccard Measure* dapat diterapkan untuk mengukur tingkat kemiripan proposal tugas akhir berbahasa Indonesia.

5.2 Saran

Berdasarkan hasil pengujian yang sudah dilakukan dapat dilihat bahwa sistem masih belum sempurna. Saran dari penulis untuk penelitian selanjutnya:

1. Sistem perlu penambahan daftar kata umum (*stemming*) berbahasa Indonesia yang lengkap.
2. Sistem mampu menangani dokumen yang berisi singkatan-singkatan alamat dan nama gelar.