Similaritas Dokumen Tugas Akhir Menggunakan Metode Rocchio

Abdul Najib, Textianis Grandis
Program Studi Teknologi Informasi, Politeknik Negeri Samarinda
Jl. DR Ciptomangunkusumo Kampus Gunung Lipan, Samarinda, 75131
E-Mail: abdulnajib@polnes.ac.id[1)], textianis.grandis@gmail.com[2)]

***Abstract*—Berkembangnya teknologi informasi yang sangat pesat memungkinkan pengolahan data semakin mudah dilakukan untuk dijadikan informasi yang bermanfaat bagi mahasiswa terutama dalam proses pembuatan tugas akhir. Sangat penting bagi mahasiswa untuk menghindari adanya kemiripan dokumen tugas akhir antara satu dengan lainnya. Penelitian ini menggunakan metode Rocchio karena memiliki konsep umum pada dokumen yang relevan dan non-relevan sebagai sarana meningkatkan pencarian untuk mengetahui kemiripan dokumen abstrak tugas akhir dengan dokumen abstrak lainnya. Tahapan dari penelitian ini terdiri dari proses case folding, filtering, tokenizing dan term weighting. Pada tahap term weighting atau pembobotan menggunakan metode TF-IDF. Setelah dilakukan pembobotan kata tiap dokumen abstrak, selanjutnya menghitung tingkat kemiripan dokumen abstrak dengan perhitungan modifikasi query berdasarkan persamaan rocchio relevance feedback. Berdasarkan hasil perhitungan kemiripan diperoleh ranking dokumen abstrak tugas akhir dengan nilai kemiripan tertinggi yaitu pada Dokumen 4 (D4) dengan nilai 48,1514. Dengan metode ini kedepannya dapat di implementasikan ke dalam pembuatan aplikasi.**

Keywords— Dokumen Abstrak, Rocchio Relevance Feedback, Similaritas, TF-IDF

I. PENDAHULUAN

Tugas akhir merupakan karya tulis yang disusun oleh mahasiswa setiap jenjang Pendidikan Tinggi. Dalam pembuatan penulisan Tugas Akhir (TA) mahasiswa dituntut untuk menggunakan pengetahuan dan keterampilan sehingga mampu menyelesaikan permasalahan yang ada dan tidak membuat penulisan yang sama dengan penulisan yang telah dibuat sebelumnya. Jika penulisan tersebut banyak memiliki kemiripan maka penulisan tersebut mengandung unsur document similarity (kemiripan dokumen).

Pentingnya penulisan laporan tugas akhir maka diharapkan sehingga dapat dihindari adanya kemiripan yang ada terutama pada dokumen abstrak. Dalam menentukan kemiripan dokumen abstrak dapat dilakukan dengan perhitungan secara baik dan relevan, hal ini dapat dilakukannya dengan adanya sistem temu-kembali informasi (*Information Retrieval System*) yang termasuk di dalam *text mining*. *Text mining* merupakan salah satu cara atau metode yang digunakan dalam mengatasi kemiripan dokumen, dengan proses pengambilan data berupa teks dari sebuah sumber dokumen. Dengan *text mining* dapat dicari kata kunci yang mewakili isi dari suatu dokumen, kemudian dokumen tersebut dianalisa dan dilakukan pencocokan dengan dokumen lain. Salah satu metode yang digunakan di dalam *text mining* terkait dengan similaritas dokumen adalah metode *rocchio*.

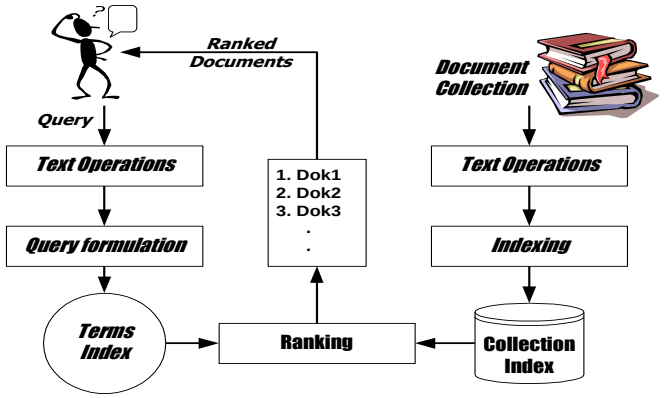
Metode *rocchio* dapat digunakan untuk membandingkan dokumen terhadap kesamaan isi antara data dengan merepresentasikan semua data dalam sebuah vector suatu *term* (istilah kata). *Rocchio* memiliki konsep umum pada dokumen yang relevan dan non-relevan sebagai saranan meningkatkan pencarian untuk mengetahui kemiripan dokumen dengan dokumen lainnya

II. METODE

A. *Information Retrieval*

Sistem temu kembali informasi (*Information retrieval system*) digunakan untuk menemukan kembali (retrieval) informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Sistem ini terutama berhubungan dengan pencarian informasi yang isinya tidak memiliki struktur, hal ini yang membedakannya dengan basis data. Dokumen merupakan salah satu contoh dari informasi yang tidak terstruktur. Hal ini terjadi karena isi dari dokumen tergantung pada pembuat dokumen tersebut.

Menurut Mandala, dkk. (2006) *Information Retrievel* dapat digambarkan sebagai berikut :



Gambar. 1.Sistem Temu Kembali Informasi

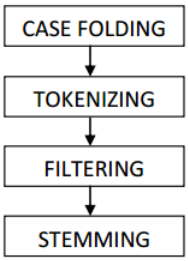
Sistem temu kembali informasi menerima query dari pengguna, kemudian dilakukan *text operation* dengan menggunakan formula tertentu untuk selanjutnya dilakukan perangkingan terhadap dokumen pada koleksi berdasarkan kesesuaiannya dengan query. Hasil perangkingan (ranking) yang diberikan kepada pengguna merupakan dokumen yang menurut sistem relevan dengan query. Namun relevan dokumen terhadap suatu query merupakan penilaian pengguna yang subjektif dan dipengaruhi banyak faktor seperti topik, dan perwakilan, sumber informasi maupun tujuan pengguna.

B. *Text Mining*

Text mining merupakan perluasan dari *data mining* atau *knowledge discovery in database* untuk menemukan pola-pola menarik dari basisdata berskala besar.

Perbedaan yang mendasar dari *text mining* dan *data mining* adalah sumber data yang diproses, dalam *data mining* menggunakan data dari basisdata yang memiliki format terstruktur sedangkan *text mining* menggunakan data dalam format teks yang tidak terstruktur karena berdasarkan tata bahasa manusia (*natural language*).

Untuk mengetahui lebih rinci dari sebuah teks yang tidak terstruktur ini diperlukan langkah-langkah dalam pemrosesan teks (*text processing*) yang digunakan untuk perubahan suatu teks menjadi *term index* dari data yang belum terstruktur menjadi data terstruktur sesuai dengan kebutuhan proses dari *mining* lebih lanjut. Tahapan secara umum yang dilakukan sebagai berikut:



Gambar. 2.Tahapan Umum Text Processing

Berdasarkan pada Gambar 2 berkenaan dengan tahapan yang secara umum dilakukan pada pemrosesan teks sebagai berikut: 1) *Case Folding* dilakukan dalam rangka perubahan setiap karakter yang ada menjadi huruf kecil serta menghilangkan karakter lain selain dari huruf; 2) *Tokenizing* merupakan tahap pemotongan *string input* pada kalimat berdasarkan tiap kata / *token* yang menyusunnya; 3) *Filtering* adalah tahap untuk mengambil kata-kata penting dari hasil tokenizing. Proses ini dilakukan dengan membuang kata-kata yang kurang penting (*stoplist*) atau menyimpan kata yang penting (*wordlist*). *Stoplist* / *stopward* merupakan kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. *Wordlist* merupakan kata-kata yang penting (deskriptif) yang harus disimpan dan tidak dibuang dengan pendekatan *bag-of-words*; 4) *Stemming* merupakan pencarian akhir kata dari tiap kata hasil *filtering*. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Misalnya kata "menghasilkan" akan menjadi "hasil".

C. *Term Weighting*

Term Weighting (Mandala R, 2000) merupakan proses penghitungan bobot tiap *term* yang dicari pada setiap dokumen sehingga dapat diketahui ketersediaan dan kemiripan suatu *term* di dalam dokumen. Secara umum dalam menentukan kemiripan suatu *term* digunakan skema pembobotan " *term frequency* * *inverse document frequency* " disebut sebagai nilai bobot *term* atau β. *Term frequency* (tf) adalah frekuensi dari kemunculan sebuah *term* dalam dokumen yang bersangkutan. *Idf* merupakan perhitungan dari bagaimana *term* didistribusikan secara luas pada koleksi dokumen yang bersangkutan. *Inverse document frequency* menunjukkan hubungan ketersediaan sebuah *term* dalam seluruh dokumen. Semakin sedikit jumlah dokumen yang mengandung *term* yang dimaksud, maka nilai idf semakin besar. Nilai idf sebuah term dirumuskan dalam persamaan berikut :

$$idf = log \frac{n}{df} \quad (1)$$

Dimana:
idf : Nilai inverse document frequency
n : Jumlah dokumen
df : Nilai document frequency (frekuensi dari sebuat term)

Perhitungan bobot dari term tertentu dalam sebuah dokumen dapat menggunakan perkalian tf dan idf yang menunjukkan bahwa deskripsi terbaik dari dokumen adalah term yang banyak muncul dalam dokumen tersebut dan sangat sedikit muncul pada dokumen yang lain. perhitungan bobot term adalah sebagai berikut :

$$\beta = (tf) * (idf) \quad (2)$$

Dimana:
β : Nilai bobot *term*
tf : Nilai term frequency
idf : Nilai inverse document frequency

D. *Rocchio Relevance Feedback*

Teknik relevance feedback ditemukan pertama kali oleh Rocchio. Rocchio memandang feedback sebagai permasalah dalam mencari sebuah query optimal, yaitu query yang memaksimalkan selisih antara dokumen relevan dengan dokumen tak relevan. Relevance feedback berguna untuk mendekatkan query ke rataan dokumen relevan dan menjauhkan ke rataan dokumen tak relevan. Hal ini dapat dilakukan melalui penambahan istilah query dan penyesuaian bobot istilah query sehingga sesuai dengan kegunaan istilah tersebut dalam fungsinya membedakan dokumen relevan dan tak relevan (Salton G and Buckley C, 1990).

$$R = N + \beta \left(\left(\frac{Dp}{Np}\right) - \left(\frac{Dn}{Nn}\right)\right) \quad (3)$$

Dimana :
R : Tingkat kemiripan
N : Jumlah term tiap dokumen
β : Nilai bobot term
Dp : Term dari dokumen relevan
Np : Jumlah dokumen relevan
Dn : Term dari dokumen tak relevan
Nn : Jumlah dokumen tak relevan.

Perhitungan dengan menggunakan rumus di atas yaitu dokumen yang memiliki nilai R terbesar adalah dokumen yang paling sesuai dengan input term (*query*) dari user.

III. HASIL DAN PEMBAHASAN

A. Data dokumen abstrak Tugas Akhir

Dokumen abstrak yang dipergunakan dalam perhitungan dengan metode Rocchio ini berasal dari abstrak dokumen tugas akhir mahasiswa Jurusan Teknologi Informasi tahun 2015-2016 yang dijadikan contoh sebanyak 36 dokumen dengan rincian 19 dokumen dijadikan sebagai data latih dan 17 dokumen dijadikan sebagai data uji. Dari data tersebut memiliki 4 kategori yaitu tentang *text mining*, SPK, Jaringan dan Citra Digital.

B. Pemrosesan Teks

Berdasarkan pada data abstrak dokumen tugas akhir, *case folding* dilakukan untuk penghapusan *delimiter* dan spasi yang ada pada dokumen.

TABLE I. *CASE FOLDING* PADA ABSTRAK TUGAS AKHIR

Dok	Menghilangkan Delimiter
D1	kaya dan luasnya informasi meningkatkan akan kebutuhan data dalam kehidupan yang sehari-hari kebutuhan akan data tersebut mendorong berkembangnya *information retrieval* atau sistem temu kembali informasi. pengembangan *information retrieval* sering ....
D2	setiap tahun jurusan teknologi informasi polnes merupakan salah satu jurusan yang dapat menghasilkan lulusan mahasiswa yang berkompeten khususnya di dunia informatika dan sebagian besar lulusan dari jurusan teknologi informasi polnes telah dikatakan siap untuk berada di dunia kerja ....
D3	dalam bahasa indonesia terdapat kata dasar dan kata imbuhan sebuah kata berimbuhan dapat terbentuk dengan adanya penambahan imbuhan awalan sisipan ataupun akhiran pada sebuah kata dasar sehingga ....
D4	prediksi adalah suatu proses memperkirakan secara sistematis tentang sesuatu yang paling mungkin terjadi dimasa depan berdasarkan informasi masa lalu dan sekarang yang dimiliki agar kesalahannya selisih antara suatu yang terjadi dengan hasil perkiraan dapat diperkecil. prediksi ....

Catatan : contoh data di dalam tabel tersebut di diatas disederhanakan

Proses selanjutnya adalah *filtering* sehingga mendapatkan hasil berikut sebagai berikut:

TABLE II. PROSES FILTERING

No	Data Latih					
1	kaya dan luarnya informasi meningkatkan akan kebutuhan data dalam kehidupan			luasnya	informasi	Meningkatkan
2	setiap tahun jurusan teknologi informasi polnes merupakan salah satu jurusan yang ..			jurusan	teknologi	informasi
3	dalam bahasa indonesia terdapat kata dasar dan kata imbuhan sebuah kata		bahasa	indonesia		
4	prediksi adalah suatu proses memperkirakan secara sistematis tentang	prediksi		proses		

Setelah *filtering* dilakukan, proses selanjutnya adalah *tokenizing* yaitu tahap pemotongan string input pada kalimat berdasarkan tiap kata/token yang menyusunnya serta menghilangkan kata yang ganda atau duplikat.

TABLE III. PROSES TOKENIZING PADA DOKUMEN

D1	D2	D3	D4
luasnya	jurusan	bahasa	prediksi
informasi	teknologi	indonesia	proses
Meningkat-kan	informasi	dasar	sistematis
kebutuhan	polnes	imbuhan	informasi
data	Menghasilkan	berimbuhan	Kesalahan-nya
kehidupan	lulusan	awalan	selisih
mendorong	mahasiswa	sisipan	hasil
berkembangnya	berkompeten	akhiran	perkiraan
information	dunia	menghilangkan	diperkecil
retrieval	Informati-ka	menghapus	kejadian

Dari hasil *tokenizing* yang dilakukan selanjutnya dapat dilakukan proses *stemming* dimana proses ini akan mengambil kata dasar dari teks yang didapatkan jika ada awala, sisipan atau akhiran. Sehingga akan di dapatkan kata dasar yang siap digunakan untuk proses perhitungan selanjutnya.

C. Tahap Perhitungan Bobot

Berdasarkan pada hasil yang didapatkan pada saat pemrosesan teks (*text processing*) selanjutnya dihitung bobot untuk masing-masing *term* yang dicari sehingga dapat diketahui kemiripan dan ketersediaan suatu *term* dalam dokumen. Penentuan bobot ini dilakukan dengan menghitung nilai *tf*, *idf* dalam tabel berikut:

TABLE IV. PERHITUNGAN TF, IDF

Skema Pembobotan							
	TF						DF
TERM	KK	D1	D2	D3	D4	DF	Log(N/NF)
akhir	1	1			1	2	0,30103
akhiran				1		1	0,60206
algoritma			1	1		2	0.30103
analisis					1	1	0,60206
apriori			1			1	0,60206
asosiasi			1			1	0,60206

awalan				1		1	0,60206
bahasa				1		1	0,60206
bayes	1				1	1	0,60206
berbahasa	1			1		1	0,60206
berhasil	1			1		1	0,60206
berimbuhan				1		1	0,60206
berkembangnya		1				1	0,60206
berkompeten			1			1	0,60206
....	...	...	...	...	...	...	....

Catatan : contoh data di dalam tabel tersebut di diatas disederhanakan

Berdasarkan pada Tabel 1, terdapat warna biru, warna terebut merupakan *term* yang terdapat dalam kata kunci dalam data uji. *Idf* menunjukkan nilai *inverse* dari *df* terhadap tiap kata yang diperoleh dari hasil logaritma (*log*) dari N sebagai jumlah keseluruhan dokumen yang dibagi dengan *df*. Perhitungan *idf* untuk kata "*penelitian*" adalah sebagai berikut dengan persamaan (1):

$$\text{IDF}_{(\text{penelitian})} = \log\left(\frac{4}{2}\right)$$

$$\text{IDF}_{(\text{penelitian})} = \log(2)$$

$$\text{IDF}_{(\text{penelitian})} = 0{,}30103$$

Setelah mendapatkan nilai *tf* dan *idf*, selanjutnya adalah menghitung bobot tiap term yang mempunyai ketersediaan dalam tiap dokumen. Bobot (W) berupa hasil kali antar *tf* dan *idf* untuk tiap kata pada persamaan (2)

$$\text{D1: } \beta_{(\text{penelitian})} = 0 \times 0{,}30103 = 0$$

$$\text{D2: } \beta_{(\text{penelitian})} = 0 \times 0{,}30103 = 0$$

$$\text{D3: } \beta_{(\text{penelitian})} = 1 \times 0{,}30103 = 0{,}30103$$

$$\text{D4: } \beta_{(\text{penelitian})} = 1 \times 0{,}30103 = 0{,}30103$$

TABLE V. PEMBOBOTAN KATA

Skema Pembobotan					
TERM	W (bobot) = TF.IDF				
	WKK	WD1	WD2	WD3	WD4
akhir	0,30103	0,30103	0	0	0,30103
akhiran	0	0	0	0,60206	0
algortima	0	0	0,30103	0,30103	0
analisis	0	0	0	0	0,602059
apriori	0	0	0,60206	0	0
asosiasi	0	0	0,60206	0	0
awalan	0	0	0	0,60206	0
bahasa	0	0	0	0,60206	0
bayes	0,60206	0	0	0	0,602059
berbahasa	0	0	0	0,60206	0
berhasil	0,60206	0	0	0,60206	0
....	....	....	....	....	....
		2,430874	2,305935	1,6300887	2,73190365

Catatan : contoh data di dalam tabel tersebut di diatas disederhanakan

Informasi yang dapat disajikan sesuai dengan Tabel 2 menunjukkan bobot W atau hasil kali *tf* dan *idf* untuk tiap kata dalam tiap dokumen. Didalam tabel terdapat term berwarna abu-abu, warna terebut merupakan bobot tiap term yang terdapat dalam kata kunci dalam data uji. Bobot kata kunci untuk masing-masing dokumen dijumlahkan untuk mendapatkan total bobot kata kunci tiap dokumen

W_{D1} = 0,30103 + 0,124939+ 0,60206 + 0,124939 + 0 + 0,124939 + 0,60206 + 0,124939 + 0,124939 + 0,30103 = 2,430874

W_{D2} = 0,124939+ 0,124939 + 0 + 0,60206 + 0,124939 + 0,60206 + 0,60206 + 0,60206 = 2,305935

W_{D3} = 0,60206 + 0 + 0,124939 + 0,30103 + 0,60206 = 1,6300887

W_{D4} = 0,30103 + 0,60206 + 0.124939 + 0,124939 + 0 + 0,124939 + 0,124939 + 0,60206 + 0,30103 + 0,124939 + 0,30103 = 2,73190365

Setelah mendapat jumlah bobot tiap *term* dalam tiap dokumen, selanjutnya yaitu menghitung kemiripan dengan menggunakan metode *rocchio*.

D. Roccihio Relevance Feedback

Tingkat kemiripan dokumen selanjutnya dapat dihitung melakukan modifikasi *query* berdasarkan persamaan *rocchio relevance feedback* (3)

Dokumen 1:

$$48 + 2{,}4309\left(\left(\frac{10}{34}\right) - \left(\frac{38}{138}\right)\right) = 48{,}0456$$

Dokumen 2:

$$48 + 2{,}3059\left(\left(\frac{8}{34}\right) - \left(\frac{40}{138}\right)\right) = 47{,}8742$$

Dokumen 3:

$$28 + 1{,}6301\left(\left(\frac{5}{34}\right) - \left(\frac{23}{138}\right)\right) = 27{,}9680$$

Dokumen 4:

$$48 + 2{,}7319\left(\left(\frac{11}{34}\right) - \left(\frac{37}{138}\right)\right) = 48{,}1514$$

Dari hasil perhitungan berdasarkan persamaan *rocchio relevance feedback* didapat nilai *Dokumen 1* yaitu 48,0456, *Dokumen 2* yaitu 47,8742, *Dokumen 3* yaitu 27,9680 dan *Dokumen 4* yaitu 48,1514. Tahapan selanjutnya adalah melakukan proses perangkingan yang diurut dari terbesar sampai terkecil, sehingga dapat diketahui kemiripan antara dokumen abstrak tugas akhir

IV. KESIMPULAN

Hasil perhitungan diperoleh ranking dokumen sebagai berikut :

TABLE VI. HASIL PERANKINGAN

Dokumen	Total Persamaan	Rank
D4	48,1514	1
D1	48,0456	2
D2	47,8742	3
D3	27,9680	4

Berdasarkan kata kunci dari data uji dapat disimpulkan bahwa dokumen abstrak yang paling relevan memiliki total tertinggi diantara dokumen abstrak lainnya yaitu D4 dengan total 48,1514, diikuti D1 dengan total 48,0456, D2 dengan total 47,8742, D3 dengan total 27,9680. D4 merupakan dokumen abstrak yang paling relevan berdasarkan kata kunci dari data uji karena memiliki nilai tertinggi dengan dokumen abstrak lainnya.

REFERENSI

[1] Mandala, Rila dan Hendra Setiawan, "Peningkatan Performansi Sistem Temu-Kembali Informasi dengan Perluasan Query Secara Otomatis" Laboratorium Keahlian Informatika Teori Departemen Sistem Informasi, Institut Teknologi Bandung, 2006.

[2] Mandala Rila, The Exploration and Analysis of Using Multiple Types of Thesaurus for Query Expansion in Information Retrieval, Journal of Natural Language Processing, Volume 7, Number 2, pp.117-140, 2000

[3] Mandala Rila, Improving Information Retrieval System Performance by Combining Text-Mining Techniques, International Journal of Intelligent Data Analysis, Volume 4, Number 6, pp. 489-511, IOS Press, ISSN : 1088-467X, 2000.

[4] Salton G. and Buckley C., Improving Retrieval Performance by Relevance Feedback, Journal of American Society for Information Science 41(4) (1990), pp. 299-297.