PENERAPAN ALGORITMA *TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY* (TF-IDF) UNTUK *TEXT MINING*

Musfiroh Nurjannah [1)], Hamdani[2)], Inda Fitri Astuti[3)]

[1)]Mahasiswa S1 Program Studi Ilmu Komputer FMIPA Universitas Mulawarman
[2,3)]Dosen Program Studi Ilmu Komputer FMIPA Universitas Mulawarman
Email : fifi.fadli@gmail.com

ABSTRAK

Algoritma *Term Frequency Inverse-Document Frequency* merupakan suatu algoritma yang menggalikan antara *Term frequency* dengan *Inverse Document Frequency*. *Term frequency* yaitu jumlah kemunculan sebuah *term* pada sebuah dokumen. *Inverse Document Frequency* yaitu pengurangan dominasi *term* yang sering muncul diberbagai dokumen, dengan memperhitungkan kebalikan frekuensi dokumen yang mengandung suatu kata.

Text Mining pada umumnya adalah *unstructured data*, atau minimal *semistructured*. Maka merupakan tantangan tambahan pada *text mining* yaitu struktur teks yang kompleks dan tidak lengkap, arti yang tidak jelas dan tidak standard, dan bahasa yang berbeda ditambah translasi yang tidak akurat.

Hasil dari penelitian menunjukan bahwa, penerapkan algoritma *term frequency inverse-document frequency* untuk *text mining* sangat membantu pengguna. untuk mendapatkan informasi pada kumpulan dokumen. Dengan format *file* txt berdasarkan kata kunci yang dimasukan oleh pengguna pada sistem. Dengan koleksi uji kata 'upaya' pada *query* maka didapatkan keluaran dengan bobot nilai 8.65441 yang merupakan jumlah kata terbanyak sesuai dengan *query*.

Kata Kunci *: TF-IDF, Text Mining, Ruang Vektor.*

.

PENDAHULUAN

Seiring dengan perkembangan informasi banyak pihak menyadari bahwa masalah utama telah bergeser dari cara mengakses informasi menjadi memilih informasi yang berguna secara selektif. Menemukan informasi berdasarkan kesesuaian dengan *query* (masukan berupa ekspresi kebutuhan informasi oleh pengguna) dari suatu kumpulan informasi yang relevan dengan kebutuhan dari penggunanya secara otomatis tidak mungkin dilakukan secara manual, karena kumpulan informasi yang sangat besar dan terus bertambah besar. Maka diperlukannya penambangan kata (*Text Mining*) yaitu banyaknya data yang berupa teks yang terdapat pada dokumen kemudian mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen.

Berdasarkan penelitian sebelumnya pada skripsi dengan judul "Implementasi Metode *Term Frequency Inverse Document Frequency* (TF-IDF) Pada Sistem Temu Kembali Informasi" (Zafikri,2008), penulis mencoba menggunakan objek penelitian dan kriteria kumpulan dokumen, begitu juga dengan *programming* yang digunakan.

Dalam perancangan aplikasi ini, metode yang digunakan juga menggunakan algoritma *Term Frequency - Inverse Document Frequency* (TF-IDF). Metode ini merupakan algoritma yang melakukan penggabungan dua metode yaitu konsep frekuensi kemunculan *term* dalam sebuah dokumen dan *inverse* frekuensi dokumen yang mengandung kata tersebut, akan mampu meningkatkan proporsi jumlah dokumen yang dapat ditemukan kembali dan yang dianggap relevan secara sekaligus. Sehingga kriteria *term* yang paling tepat adalah *term* yang sering muncul dalam dokumen secara individu, namun jarang dijumpai pada dokumen lainnya.

Berdasarkan uraian diatas, penulis merasa tertarik untuk meneliti lebih jauh mengenai metode *Term Frequency - Inverse Document Frequency* (TF-IDF) dengan mengambil konsep judul yaitu "Penerapan Algoritma *Term Frequency - Inverse Document Frequency* (TF-IDF) Untuk *Text Mining*"

METODE PENELITIAN

TF-IDF (*TERMS FREQUENCY-INVERSE DOCUMENT FREQUENCY*)

Metode TF-IDF 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 di dalam sebuah dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam 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 di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen.

Rumus untuk TF-IDF:

$$tf = 0{,}5 + 0{,}5\ x\ \frac{tf}{max(tf)} \qquad (1)$$

$$idf_t = log\left(\frac{D}{df_t}\right)$$

$$W_{d,t} = tf_{d,t}\ x\ IDF_{d,t}$$

Keterangan:

tf = banyaknya kata yang dicari pada sebuah dokumen

$max(tf)$ = jumlah kemunculan terbanyak *term* pada dokumen yang sama.

Nilai *D* = total dokumen

df_t = jumlah dokumen yang mengandung *term* t.

IDF = *Inversed Document Frequency (* log2 (D/df))

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

Rumus Relevansi merupakan penentuan relevansi dokumen dengan *query* dipandang sebagai pengukuran kesamaan (*similarity measure*) antara vektor dokumen dengan vektor *query*. Semakin sama suatu vektor dokumen dengan vektor *query* maka dokumen dapat dipandang semakin relevan dengan *query*

$$(Q, D) = cos\ \theta = \frac{Q \bullet D}{|Q||D|}$$

keterangan:

Q = bobot *query*

D = bobot dokumen

|*Q*| = panjang *query*

|*D*| = panjang dokumen

Rumus relevansi telah dinormalisasi

2. *Text Mining*

Tahapan yang dilakukan secara umum adalah *tokenizing*, *filtering*, *stemming*, *tagging*, dan *analyzing*.

1. *Tokenizing* adalah proses memecah dokumen menjadi kumpulan kata. *Tokenization* dapat dilakukan dengan menghilangkan tanda baca dan memisahkannya per spasi. Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca dan mengubah semua *token* ke bentuk huruf kecil (*lower case*).
2. *Filtering* adalah tahap mengambil kata-kata penting dari hasil token. *Filtering* dilakukan dengan menentukan *term* mana yang akan digunakan untuk merepresentasikan dokumen sehingga dapat mendeskripsikan isi dokumen dan membedakan dokumen tersebut dari dokumen lain di dalam koleksi.
3. *Stemming* adalah proses mengembalikan kata menjadi kata dasarnya. Hal ini bisa dilakukan dengan cara menghilangkan akhiran atau awalan dari sebuah kata.
4. *Tagging* adalah tahapan mencari bentuk awal atau *root* dari tiap kata lampau atau kata hasil *stemming.*
5. *Analyzing* adalah tahapan penentuan seberapa jauh keterhubungan antara kata-kata antar dokumen yang ada.

Untuk menerapkan perhitungan algoritma *term frequency - inverse document frequency* dengan kalimat yang lain dapat dilihat pada studi kasus :

Query (*Q*) = pengetahuan logistic
Dokumen 1 (D1) = Manajemen logistic
Dokumen 2 (D2) = Pengetahuan antar individu
Dokumen 3 (D3) = Dalam manajemen pengetahuan terdapat…… transfer pengetahuan logistik.

Maka, jumlah dokumen (D) = 3 dan perhitungan untuk *term frequency - inverse document frequency* dapat dilihat pada tabel 1.

Analisis perhitungan algoritma TF-IDF pada tabel 1 yaitu :

1. Dari contoh studi kasus , dapat diketahui bahwa nilai bobot (w) dari D1 dan D2 adalah sama.
2. Apabila diurutkan maka proses sorting juga tidak akan dapat mrngurutkan secara tepat, karena nilai w keduanya sama.
3. Untuk mengatasi hal ini, digunakan algoritma dari *vector-space model.*

Tabel 1. Perhitungan TF-IDF

Token	*Q*	D1	D2	D3	*Q**D1	*Q**D2	*Q**D3
manajemen	0	0,031	0	0,031	0	0	0
transaksi	0	0,228	0	0	0	0	0
logistik	0	0,031	0	0,031	0,031	0	0,031
transfer	0,031	0	0	0,228	0	0	0
pengetahuan	0	0	0,031	0,124	0	0,031	0,062
individu	0	0	0,028	0	0	0	0
	Sqrt(*Q*)	Sqrt(Di)			Sum(*Q* dot Di)		
	0,249	0,539	0,643	0,031	0,031	0,031	0,093

Tabel 2. Perhitungan *Vector Space Model*

	tf				df	D/df	IDF	W			
Token	*Q*	D1	D2	D3			log(D/df)	*Q*	D1	D2	D3
manajemen	0	1	0	1	2	1,5	0,176	0	0,176	0	0,176
transaksi	0	1	0	0	1	3	0,477	0	0,477	0	0
logistik	1	1	0	1	2	1,5	0,176	0,176	0,176	0	0,176
transfer	0	0	0	1	1	3	0,477	0	0	0	0,477
pengetahuan	1	0	1	2	2	1,5	0,176	0,176	0	0,176	0,352
individu	0	0	1	0	1	3	0,477	0	0	0,477	0
	Total							**0,352**	**0,829**	**0,653**	**1,181**

Bobot (w) untuk D1 = 0,176 +0 = 0,176
Bobot (w) untuk D2 = 0+0,176 = 0,176
Bobot (w) untuk D3 = 0,176+0,352= 0,528

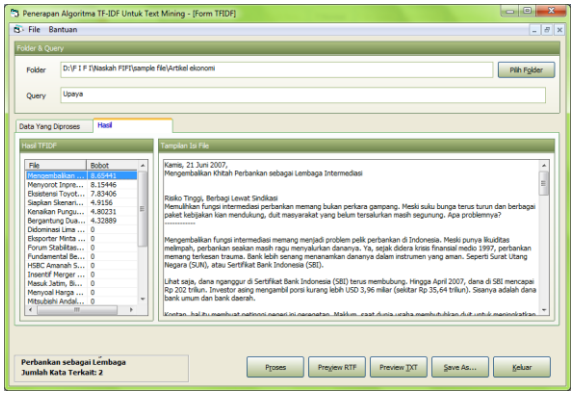
HASIL PENELITIAN

Sistem ini menerapkan algoritma *term frequency-inverse document frequency*, untuk mendapatkan nilai atau bobot dari suatu dokumen yang relevan dengan *query* yang telah diinputkan. Maka sistem ini akan mempermudah dari proses pencarian dari suatu kumpulan dokumen yang bertumpuk. Dengan menerpakan algoritama *term frequency-inverse document frequency* pada sistem ini , maka sistem ini diharapakan dapat menyelsaikan permasalahan yang relatif dalam pencarian ataupun menemukan informasi dari suatu dokumen.

Dengan menerapkan algoritma *term frequency-inverse document frequency* pada sistem ini dapat dihasilkan suatu bobot nilai dari tiap dokumen yang relevan dengan *query* yang diinputkan oleh pengguna. Bobot yang dihasilkan dari tiap dokumen pada sistem telah dihitung dengan mengkalikan antara *term frequency* dengan *inverse document frequency*, kemudian menghitung kemiripan antar dokumen yang dilakukan dengan cara menghitung cosine similarity antara vector dokumen koleksi dan vector dokumen.

Sistem ini memiliki bebrapa fungsi, antara lain :

1. Melakukan *maintenance* kata stop, seperti menambah, mengubah, dan menghapus.
2. Memberikan informasi bobot dari suatu dokumen yang relevan dengan *query*.
3. Menampilkan bobot dari semua dokumen yang ada pada folder yang sudah dipilih dan dapat menampilkan dokumen tersebut dan memyimpannya dengan ekstensi RTF yang dapat digunakan kembali oleh pengguna yang memerlukan.



Gambar 1. Implementasi Hasil TF-IDF

Setelah proses pencarian dan perhitungan selesai maka akan tampil bobot nilai dari dokumen yang terkait dengan *query* yang telah diinputkan. Bobot nilai dari dokumen tersebut dapat dilihat pada tab Hasil, ketika *user* klik salah satu diantara dokuemen tersebut maka isi dari dokumen tersebut akan ditampilkan. Dibawah Hasil TF-IDF ditampilkan juga Jumlah Kata Terkait dengan *query*.

KESIMPULAN

Berdasarkan hasil penelitian penerapan algoritma *term frequency-inverse document frequency* (tf-idf), dapat diambil kesimpulan bahwa:

1. Sistem ini melakukan penerapan algoritma *term frequency-inverse document frequency* untuk *text mining* sehingga membantu pengguna mendapatkan dokumen terkait yang sesuai dengan query yang telah diinputkan.
2. Metode menggunakan Algoritma *term frequency-inverse document frequency* merupakan salah satu metode yang tepat untuk digunakan dalam pencarian kata ditiap dokumen.
3. Sistem ini dapat melakukan pencarian sesuai dengan *query* secara bersamaan untuk banyak file.
4. Metode pembobotan dokumen TF-IDF tidak selalu memberikan hasil performansi yang baik pada koleksi pengujian.
5. Dalam algoritma TF-IDF, frekuensi kemunculan sebuah kata (*term*) dalam dokumen tidak mempengaruhi hasil

perhitungan bobot dokumen oleh sistem (sifat monotonicity TF-IDF).

Keluaran sistem menampilkan urutan rangking pada setiap file berdasarkan *query* yang telah dihitung bobot dari tiap file dan akan menampilkan jumlah kata yang ditemukan dalam tiap file.

DAFTAR PUSTAKA

[1] Trunojoyo, H. *Sistem Temu Balik Informasi (Sebuah Contoh Implementasi).(http://husni.trunojoyo.ac.id/wp content/uploads/2010/03/Husni-IR-dan Klasifikasi.pdf).*

[2] Arifin, A. 2002. *Penggunaan Digital Tree Hibrida pada Aplikasi Information Retrieval untuk Dokumen Berita*. Surabaya : Institut Teknologi Sepuluh Nopember.

[3] Hendry. 2009. *Berbagai Aplikasi Databae dengan VB 6.0*. Jakarta : PT. Elex Media Komputindo.

[4] Ladjamudin, A. 2005. *Analisa dan Desain Sistem Informasi*. Yogyakarta : Penerbit Andi Yogyakarta.

[5] Mandala, R. dan Setiawan, H. 2002. *Peningkatan Performansi Sistem Temu-Kembali Informasi dengan Perluasan Query Secara Otomatis*. Bandung: Institut Teknologi Bandung.

[6] Munawar, 2005. *Permodelan Visual dengan UML*. Yogyakarta : GRAHA ILMU.

[7] Raymond, J. 2006. *Machine Learning Text Categorization*. Austin: University of Texas at Austin.

[8] Simarmata, J dan Paryudi, I. 2006. *Basis Data*. Yogyakarta : Andi.

[9] Naradhipa, R. 2009. *Pemilihan Kategori Artikel Berita dengan Text Mining*. Paper Terpublikasi. Bandung: Institut Teknologi Bandung.

[10] Ramadhany, T. 2008. *Implementasi Kombinasi Model Ruang Vektor dan Model Probabilistik Pada Sistem Temu Balik Informasi*. Skripsi Terpublikasi. Bandung: Institut Teknologi Bandung.

[11] http://lecturer.eepis-its.edu/~iwanarif/kuliah/dm/6Text%20Mining.pdf(Tanggal Akses 12 Maret 2011)

[12] http://papers.gunadarma.ac.id/index.php/computer/article/view/574/536(Tanggal Akses 13 Maret 2011)