IMPLEMENTASI VECTOR SPACE MODEL DAN BEBERAPA NOTASI METODE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) PADA SISTEM TEMU KEMBALI INFORMASI

Oka Karmayasa dan Ida Bagus Mahendra

Program Studi Teknik Informatika,

Jurusan Ilmu Komputer,

Fakultas Matematika Dan Ilmu Pengetahuan Alam,

Universitas Udayana

Email: oka.karmayasa@cs.unud.ac.id

ABSTRAK

Sistem temu kembali informasi merupakan sistem yang digunakan untuk menemukan informasi yang relevan dengan kebutuhan dari penggunanya secara otomatis berdasarkan kesesuaian dengan *query* dari suatu koleksi informasi. Penelitian ini bertujuan untuk mengenal karakteristik beberapa notasi pembobotan TF-IDF serta mengimplementasikan model ruang vektor menggunakan beberapa notasi pada metode pembobotan TF-IDF pada sistem temu kembali informasi. Notasi TF-IDF yang digunakan pada penelitian ini adalah *anc.ntc, lnc.ltc* dan *ltc.ltc*. Pada Model Ruang Vektor, pembobotan *term* dilakukan disisi dokumen dan *query*. Pembobotan yang dihasilkan pada algoritma TF-IDF akan menjadi *variabel* dalam perhitungan *cosine similarity*. Hasil dari *cosine similarity* pada masing-masing dokumen terhadap *query* akan diurutkan secara *descending*, sehingga hasil pencarian akan menampilkan dokumen yang paling mendekati kata kunci. Sistem ini dikembangkan menggunakan bahasa pemrograman PHP dan dokumen yang digunakan sebagai data uji sebanyak 50 artikel berita yang penulis kutip dari beberapa situs di internet. Penelitian ini telah berhasil mengimplementasikan *vector space model* dan tiga notasi pembobotan TF-IDF. Hasil dari penelitian ini menunjukkan bahwa tiap notasi pembobotan TF-IDF memiliki karekteristik yang berbeda-beda dan menghasilkan urutan dokumen relevan yang berbeda, antara notasi satu dengan notasi lainnya.

***Kata Kunci** : Model Ruang Vektor, TF-IDF, Sistem temu kembali informasi*

ABSTRACT

Information retrieval system is a system used to automatically find some relevant information based on query, from the information collections. This study aims to identify some of the characteristics of the TF-IDF weighting notation and implement a vector space model using some notation on the TF-IDF weighting method in information retrieval. Notation used in this study is anc.ntc, lnc.ltc and ltc.ltc. In the Vector Space Model, term weighting calculated on the documents and queries side. Weighting generated on TF-IDF algorithm will become a variable in the calculation of cosine similarity. The results of the cosine similarity of each document to the query will be sorted in descending order, so the search results will display the most relevant documents first. The system was developed using the PHP programming language and used 50 news articles as test data, that quoted from several sites on the internet. This study has successfully implemented the vector space model and three TF-IDF weighting notation. The results of this study indicate that each notation TF-IDF weighting has different characteristics and generate a sequence of relevant documents, which different between each notation.

Keywords: *Vector Space Model,* TF-IDF*, Information Retrieval*

PENDAHULUAN

Sistem temu kembali informasi (*information retrieval system*) merupakan sistem yang dapat digunakan untuk menemukan informasi yang relevan dengan kebutuhan dari penggunanya secara otomatis dari suatu koleksi informasi (Mandala dan Setiawan, 2002). Sistem temu kembali menerima masukan (*input)* berupa kata-kata kunci dari informasi yang dicari, dan dalam waktu yang relatif singkat sistem akan menampilkan daftar dokumen yang sesuai dengan kebutuhan informasi pengguna.

Metode Ruang Vektor adalah suatu metode untuk merepresentasikan sistem temu kembali informasi ke dalam vektor dan memperhitungkan fungsi *similarity* dalam proses pencocokan beberapa vektor. Suatu sistem temu kembali informasi terdiri atas dua bagian, yaitu penyimpanan dokumen dan pemrosesan *query*. Baik query maupun dokumen-dokumen yang disimpan, dinyatakan dalam bentuk vektor (Zafikri, 2008). Elemen vektor tersebut adalah hasil dari pembotan kata (*term*) pada dokumen dan *query*.

Metode TF-IDF (*Term Frequncy Inverse Document Frequency*) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (*term*) terhadap dokumen (Robertson, 2005). Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan *inverse* frekuensi dokumen yang mengandung kata tersebut.

Terdapat beberapa cara atau metode dalam melakukan pembobotan kata pada metode TF-IDF, yaitu melalui skema pembobotan *query* dan dokumen. Skema pembobotan *query* dan dokumen merupakan salah satu jenis pembobotan kata pada *System for the Mechanical Analysis and Retrieval of Text* (SMART) atau sering disebut sebagai SMART *notation* (notasi SMART). Pada notasi SMART, merepresentasikan pembobotan ke dalam bentuk *ddd.qqq* (Manning *et al*, 2009).

Berdasarkan penelitian-penelitian sebelumnya, yang membahas tentang penerapan metode TF-IDF. Penulis menemukan banyak terdapat variasi formula dalam mengimplementasikan metode TF-IDF pada pembobotan kata. Jika varian formula tersebut direpresentasikan ke dalam bentuk *ddd.qqq*, secara umum terdapat beberapa jenis notasi yang dikembangkan antara lain, *nnc.nnc*, *anc.ntc*, *ltc.ltc* dan *lnc.ltc*.

Penelitian ini bertujuan untuk mengenal karakteristik beberapa notasi pembobotan TF-IDF serta meng-implementasikan model ruang vektor menggunakan beberapa notasi pada metode pembobotan TF-IDF pada sistem temu kembali informasi.

MATERI DAN METODE

Sistem temu kembali informasi secara umum terdiri dari dua tahapan besar, yaitu melakukan *preprocessing* terhadap *database* dan kemudian menerapkan metode tertentu untuk menghitung kedekatan (relevansi atau *similarity*) antara dokumen di dalam *database* yang telah dipreprocess dengan query pengguna. Sebagai hasilnya, sistem mengembalikan suatu daftar dokumen terurut *descending* (*ranking*) sesuai nilai kemiripannya dengan query pengguna.

Proses *preprocessing* meliputi tokenisasi, *stop-word removal*, *stemming*, dan *term weighting*.

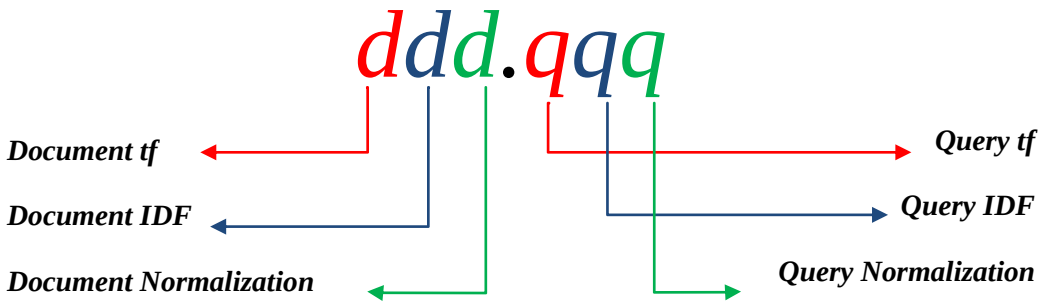
Pembobotan kata (*term weighting*) adalah proses pembobotan pada kata. Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan *term* dalam dokumen. Frekuensi kemunculan (*term frequency*) merupakan petunjuk sejauh mana *term* tersebut mewakili isi dokumen. Semakin besar kemunculan suatu *term* dalam dokumen akan memberikan nilai kesesuian yang semakin besar.

Faktor lain yang diperhatikan dalam pemberian bobot adalah kejarangmunculan kata (*term scarcity*) dalam koleksi. Kata yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting (*uncommon tems*) daripada kata yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kata (*inverse document frequency*) (Mandala dan Setiawan, 2002).

Faktor terakhirnya adalah faktor normalisasi terhadap panjang dokumen. Dokumen dalam suatu koleksi memiliki karakteristik panjang yang beragam. Ketimpangan terjadi karena dokumen yang panjang akan cenderung mempunyai frekuensi kemunculan kata yang besar. Sehingga untuk mengurangi ketimpangan tersebut diperlukan faktor normalisasi dalam pembobotan (Mandala dan Setiawan, 2002).

Terdapat beberapa cara atau metode dalam melakukan pembobotan kata pada metode TF-IDF, yaitu melalui skema pembobotan *query* dan dokumen. Skema pembobotan *query* dan dokumen merupakan salah satu jenis pembobotan kata pada *System for the Mechanical Analysis and Retrieval of Text* (SMART) atau sering disebut sebagai SMART *notation* (notasi SMART). Pada notasi SMART, merepresentasikan pembobotan ke dalam bentuk *ddd.qqq* (Manning *et al*, 2009).

Tiga huruf pertama pada *ddd.qqq* yaitu *ddd* merupakan pembobotan kata pada vektor dokumen dan tiga huruf selanjutnya yaitu *qqq* menunjukkan pembobotan pada vektor *query*. Masing-masing dari tiga huruf pada tiap kelompok menunjukkan kode untuk penggunaan *term frequency* (tf), *inverse document frequency* (IDF), dan jenis normalisasi yang digunakan.



Gambar 1. Penjelasan Notasi SMART

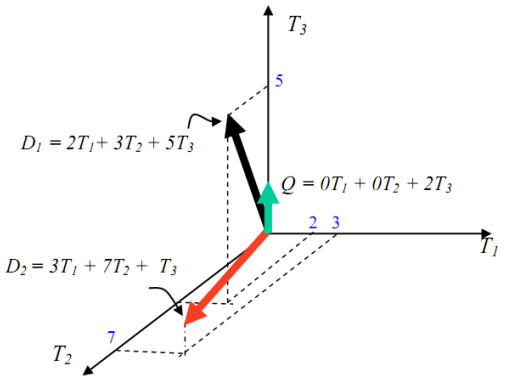
Tabel 1. Notasi pada TF-IDF (Sumber : Yogatama, 2008)

Term Frequency		
Abjad pertama	**Persamaan**	**Deskripsi**
n	tf	*Raw term frequency*
l	$tf = 1 + \log(tf)$	*Logarithm term frequency*
b	1 / 0	*Binary term frequency*
a	$tf = 0{,}5 + 0{,}5 \times \frac{tf}{\max(tf)}$	*Augmented term frequency*
Inverse Document Frequency		
Abjad pertama	**Persamaan**	**Deskripsi**
n	1	IDF tidak diperhitungkan
t	$idf = (log(D/df) + 1)$	Nilai logaritmik dari IDF
Normalisasi		
Abjad pertama	**Persamaan**	**Deskripsi**
n	1	Normalisasi tidak diperhitungkan
c	$\sqrt{w_1^2 + w_2^2 + \ldots + w_n^2}$	Normalisasi terhadap panjang dokumen

Pada tabel 2.1 terdapat jenis-jenis notasi yang dapat diguakan untuk menyusun metode pembobotan TF-IDF, banyak kombinasi pembobotan dapat dikembangkan. Secara umum terdapat beberapa jenis notasi yang dapat dikembangkan antara lain, *ntc.ntc*, *ltc.ltc* dan *lnc.ltc*.

Metode Ruang Vektor adalah suatu metode untuk merepresentasikan sistem

temu kembali informasi. 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*.



Gambar 2. Representasi Dokumen dan *Query* pada Ruang Vektor

(Sumber : Mandala dan Setiawan, 2002)

Perhitungan kesamaan antara vektor *query* dan vektor dokumen dilihat dari sudut yang paling kecil. Sudut yang dibentuk oleh dua buah vektor dapat dihitung dengan melakukan perkalian dalam (inner product), sehingga rumus relevansinya, adalah:

$$R\ (Q.D) = \cos\theta = \frac{Q \cdot D}{|Q||D|} \quad \text{..... (1)}$$

dimana:

Q = bobot *query* |Q| = panjang *query*

D = bobot dok |D| = panjang dok

Proses perangkingan dari dokumen dapat dianggap sebagai proses pemilihan (vektor) dokumen yang dekat dengan (vektor) *query*, kedekatan ini diindikasikan dengan sudut yang dibentuk. Nilai *cosinus* yang cenderung besar mengindikasikan bahwa dokumen cenderung sesuai *query*. Nilai cosinus sama dengan 1 mengindikasikan bahwa dokumen sesuai dengan *query*.

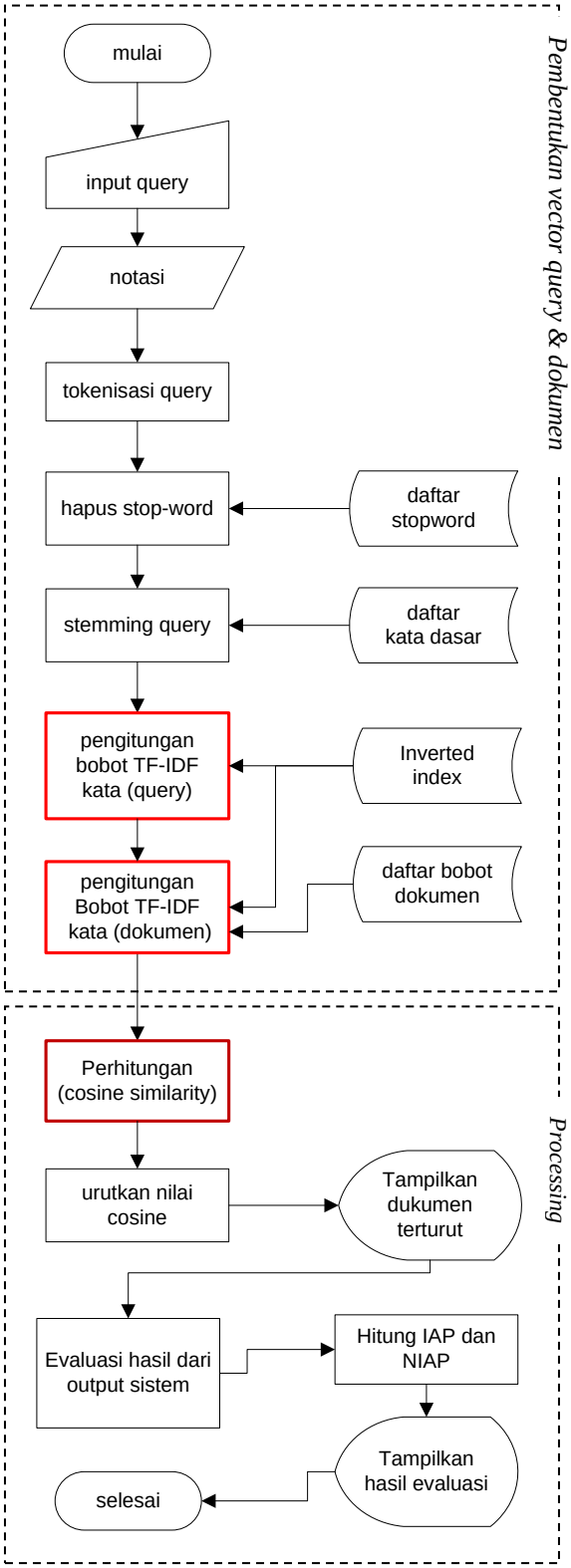
HASIL DAN PERANCANGAN

Pada penelitian ini, penulis melakukan implementasi *vector space* model terhadap masing-masing notasi pembobotan TF-IDF. Sistem ini dikembangkan menggunakan tiga buah notasi pembobotan TF-IDF, yaitu *anc.ntc*, *lnc.ltc* dan *ltc.ltc*.

Data yang digunakan adalah data dalam bentuk dokumen. Dokumen yang akan dijadikan sumber data adalah dokumen berbahasa Indonesia yang berformat teks / HTML dan berupa koleksi data uji. Data uji yang dipilih adalah artikel berita berbahasa Indonesia dari situs berita di media internet

Adapun rancangan penelitian yang dilakukan dalam melakukan implementasi *vector space model* terhadap metode TF-IDF ditunjukkan oleh gambar . Dari gambar tersebut dapat dilihat bahwa sistem terdiri dari dua tahapan utama yaitu:

- subsistem pembentukan vektor *query* dan dokumen, yaitu meliputi proses *preprocessing* dan pembobotan
- subsistem *processing*, disini dilakukan impmentasi *vector space model*.



Gambar 3. Skema Rancangan Penelitian

PEMBAHASAN

Dalam melakukan implementasi *vector space model* terhadap metode pembobotan TF-IDF pada suatu sistem temu kembali informasi, penulis menggunakan bahasa pemrograman web yaitu PHP (*PHP Hypertext Pre-Processor*). Hal ini bertujuan untuk menciptakan sistem yang informatif dan *familiar* dengan penggunanya, karena sebagian besar sistem temu kembali informasi, berbasis web. Serta dapat diakses dari berbagai *platform*. Penggunaan bahasa pemrograman PHP dikombinasikan dengan bahasa pemrograman CSS dan *Javascript*, untuk memberikan tampilan yang menarik kepada pengguna agar dapat lebih mudah menganalisis performa dari beberapa notasi pembobotan yang digunakan.

Penelitian ini menggunakan tiga notasi pembobotan, yaitu *anc.ntc*, *lnc.ltc* dan *ltc.ltc*. Sehingga dalam implementasi program dikembangkan beberapa fungsi untuk menghitung beberapa pembotan TF, yaitu pembobotan TF notasi *a*, *n*, dan *l*. Dan untuk pembobotan IDF dikembangkan fungsi untuk menghitung pembobotan IDF notasi *n* dan *t*. Untuk normalisasi, hanya dikembangkan untuk notasi *c*. Berikut merupakan *pseudocode* fungsi untuk mengghitung notasi TF-IDF.

```
function
notasi_tf(nilai,notasi,freq)
{
  if(notasi=="l")
      return log(nilai)+1
  end if
  elseif(notasi=="a")
      return
0.5+(0.5*(nilai/freq))
  end elseif
  elseif($notasi=="n")
      return nilai
  end elseif
  else
      return 0
  end else
}

function
notasi_idf(freq,nilai,notasi)
{
  if(notasi=="t")
      return log(freq/nilai)+1
  end if
  elseif(notasi=="n")
      return 1
  end elseif
  else
      return 0
  end else
}
function notasi_c(nilai,norm)
{
  return nilai/norm;
}
```

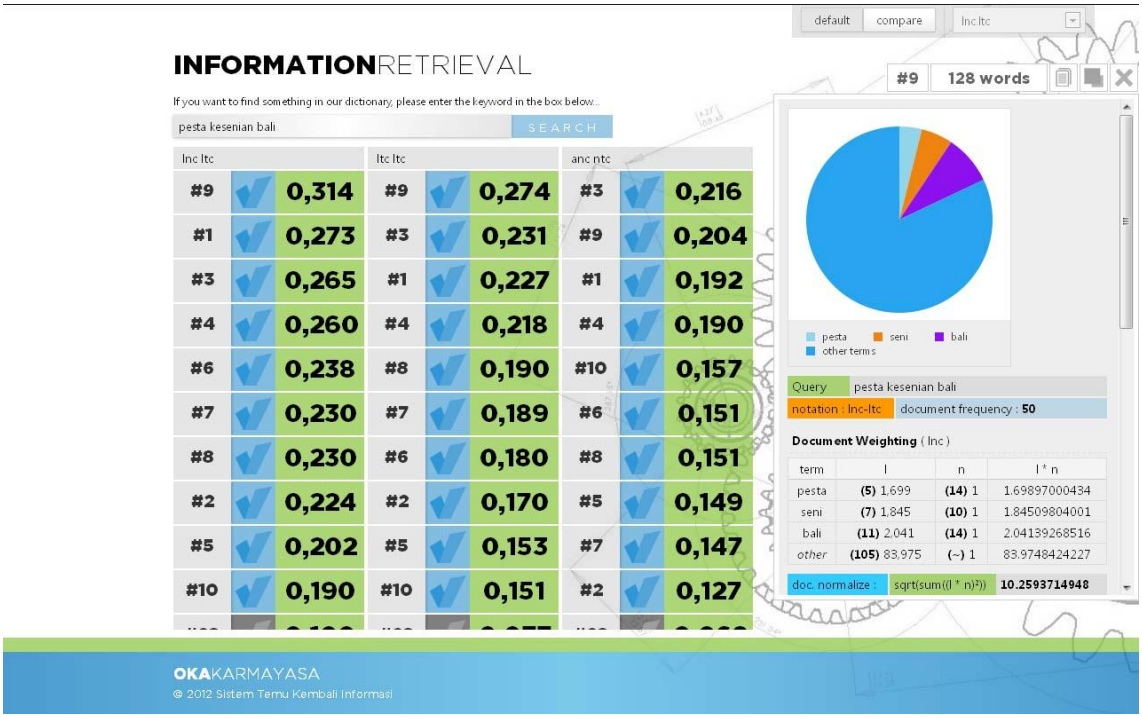
Pada Model Ruang Vektor, pembobotan *term* dilakukan disisi dokumen dan *query*. Berdasarkan rancangan penelitian, pembobotan TF-IDF terdapat pada subsistem pembentukan vektor *query* dan dokumen.

Pada subsistem *processing*, merupakan proses penerapan model ruang vektor. Pembobotan yang dihasilkan pada algoritma TF-IDF akan menjadi *variabel* dalam perhitungan *cosine similarity*. Berikut merupakan *pseudocode* perhitungan *cosine similarity*

```
bobot_q[] = hasil pembobotan
term query
bobot_d[][] = hasil pembobotan
term dokumen

nomalisasi_q = normalisasi
query hasil pembobotan
nomalisasi_d[] = normalisasi
dokumen hasil pembobotan

hasil_bobot = 0
berita = SELECT * FROM
tb_berita ORDER BY Id

while hasil query berita
  dok = berita kolom id

  while x != end of array term
  do cacah bobot_q

vektor_d =
bobot_d[key(bobot_q)][dok]/noma
lisasi_d[dok]
vektor_q =
bobot_q[key(bobot_q)]/nomalisas
i_q
hasil_bobot=(vektor_q*vektor_d)
+ hasil_bobot

  end while
  nilai_tfidf[dok]=hasil_bobot

end while

descending_sort(nilai_tfidf)
```

Hasil dari *cosine similarity* pada masing-masing dokumen terhadap *query* akan diurutkan secara *descending*, sehingga hasil pencarian akan menampilkan dokumen yang paling mendekati kata kunci.

Gambar 4. Hasil pencarian query menggunakan notasi pembobotan *lnc.ltc*



Gambar 5. Perbandingan hasil pencarian query tiga notasi pembobotan

Sistem telah berhasil diimplementasikan menggunakan bahasa pemrograman PHP dan menggunakan database mySQL sebagai media penyimpanan data pada *web server* XAMPP. Dokumen yang digunakan

sebagai data uji sebanyak 50 artikel berita yang penulis kutip dari beberapa situs di internet.

Pengujian difokuskan pada proses-proses yang terjadi pada sistem. Mengecek apakah sistem telah memberikan hasil penghitungan bobot TF-IDF dan *Vektor Space Model* dengan benar dan akurat. Serta memastikan setiap fungsi-fungsi seperti *parsing, stopword removal, stemming,* dll dapat berjalan dengan baik. Dari hasil pengujian sistem, setiap elemen atau fungsi yang ada dalam proses temu kembali informasi dapat dijalankan dengan baik.

SIMPULAN

yang dapat diambil dari penelitian yang telah dilakukan adalah sebagai berikut:

1. Penelitian ini telah berhasil mengimplementasikan *vector space model* dan tiga notasi pembobotan TF-IDF , yaitu yaitu *anc.ntc*, *lnc.ltc* dan *ltc.ltc* pada sistem temu kembali informasi
2. Notasi pembobotan TF-IDF memiliki karekteristik yang berbeda-beda dan menghasilkan urutan dokumen relevan yang berbeda antara notasi satu dengan notasi lainnya.
3. Dalam menentukan notasi yang terbaik perlu adanya penelitan tentang analisis beberapa notasi pembobotan TF-IDF.

DAFTAR PUSTAKA

[1] Mandala, Rila dan Setiawan, Hendra. 2002. *Peningkatan Performansi Sistem Temu-Kembali Informasi dengan Perluasan Query Secara Otomatis*, Bandung : Departemen Teknik Informatika Institut Teknologi Bandung.

[3] Manning, Christopher D, Prabhakar Raghavan dan Hinrich Schutze. 2009. *An Introduction To Information Retreival*, England : Cambridge University Press.

[4] Robertson, Stephen. 2005. *Understanding Inverse Document Frequency: On theoretical arguments for IDF*, England : Journal of Documentation, Vol. 60, pp. 502–520

[4] Yogatama, Dani. 2008. *Studi Penggunaan Stemming untuk Meningkatkan Performansi Sistem Temu Balik Informasi*, Bandung : Departemen Teknik Informatika Institut Teknologi Bandung.

[5] Zafikri, A. 2008. *Implementasi Metode Term Frequency Inverse Document Frequency (TF-IDF) pada Sistem Temu Kembali Informasi*, Medan : Universitas Sumatra Utara.