

**PROTOTIPE APLIKASI KATEGORISASI TEKS BERBAHASA
INDONESIA BERDASARKAN PEMANFAATAN VECTOR SPACE
MODEL PADA PENERAPAN ALGORITMA TF-IDF DAN WIDF
DENGAN FUNGSI BRAY CURTIS DAN ALGORITMA NAZIEF
ANDRIANI**

TESIS



Oleh :

DYAH SUKOWATI

1111600837

PROGRAM STUDI: MAGISTER ILMU KOMPUTER (MKOM)

PROGRAM PASCASARJANA

BUDI LUHUR JAKARTA

JAKARTA

2016

**PROTOTYPE APLIKASI KATEGORISASI TEKS BERBAHASA
INDONESIA BERDASARKAN PEMANFAATAN VECTOR SPACE
MODEL PADA PENERAPAN ALGORITMA TF-IDF DAN WIDF
DENGAN FUNGSI BRAY/CURTIS DAN ALGORITMA NAZIEF
ANDRIANI**

TESIS

**Diajukan sebagai salah satu persyaratan
untuk mendapatkan gelar Magister Ilmu Komputer (M.Kom)**



Oleh :

DYAH SUKOWATI

1111600837

**PROGRAM STUDI: MAGISTER ILMU KOMPUTER (MKOM)
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR JAKARTA
JAKARTA
2016**



**PROGRAM STUDI : MAGISTER ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS BUDILUHUR**

LEMBAR PERNYATAAN

Saya yang bertanda tangan dibawah ini :

Nama Mahasiswa : Dyah Sukowati
Nomor Induk Mahasiswa : 1111600837
Program Studi : Magister Ilmu Komputer
Konsentrasi : Teknologi Sistem Informasi
Fakultas/ Program : Pasca Sarjana

Menyatakan bahwa Tesis yang berjudul :

Prototipe Aplikasi Kategorisasi Teks Berbahasa Indonesia Berdasarkan Pemanfaatan Vector Space Model TF-IDF dan WIDF dengan Fungsi Bray Curtis dan Algoritma Naïve Andriani

- 1) Merupakan hasil karya tulis ilmiah sendiri dan bukan merupakan karya yang pernah diajukan untuk memperoleh gelar akademik pihak lain.
- 2) Saya izinkan untuk dikelola oleh Universitas Budi Luhur sesuai dengan norma hukum dan etika yang berlaku

Pernyataan ini saya buat dengan penuh tanggung jawab dan saya bersedia menerima konsekuensi apapun sesuai aturan yang berlaku apabila dikemudian hari pernyataan ini tidak benar.

Takerta 26 Februari 2016

(...Dyah Sukowati...)



PROGRAM STUDI MAGISTER ILMU KOMPUTER
PROGRAM PASCA SARJANA
UNIVERSITAS BUDI LUHUR

LEMBAR PENGESAHAN

Nama Mahasiswa : Dyah Sukowati
NIM : 1111600837
Konsentrasi : Teknologi Sistem Informasi
Jenjang Studi : Strata 2
Topik/Judul Tesis : Prototipe Aplikasi Kategorisasi Teks Berbahasa
Indonesia Berdasarkan Pemanfaatan *Vector*
Space Model pada Penerapan Algoritma TF-IDF
dan WIDF dengan Fungsi *Bray Curtis* dan
Algoritma Nazief Andriani

Jakarta, 26 Februari 2016

Tim Penguji :

Ketua,
Prof. Moedjiono, M.Sc

Anggota,
Dr. Ir. Nazori AZ, M.T

Pembimbing,
Dr. Krisna, M.Sc, M.M

Tanda Tangan :

Ketua Program Studi

Dr. Ir. Nazori AZ, M.T

ABSTRAK

Information Retrieval (IR) merupakan sistem yang digunakan untuk menemukan informasi yang relevan dengan kebutuhan dari pengguna secara otomatis berdasarkan kesesuaian dengan query dari suatu koleksi informasi. Dalam penelitian kali ini bertujuan untuk mengenal karakteristik beberapa notasi pembobotan TF-IDF dan WIDF serta mengimplementasikan model ruang vektor menggunakan beberapa notasi pada metode pembobotan TF-IDF dan WIDF pada sistem Information Retrieval (IR). Fungsi yang digunakan untuk pembobotan TF-IDF dan WIDF pada penelitian ini adalah fungsi Bray/ Curtis. Pada Model Ruang Vektor, pembobotan term dilakukan disisi dokumen dan query. Pembobotan yang dihasilkan pada algoritma TF-IDF dan WIDF 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 atau relevansi. System ini dikembangkan menggunakan bahasa pemrograman PHP dan dokumen yang digunakan sebagai data uji sebanyak 40 artikel berita yang penulis kutip dari beberapa situs di internet. Penelitian ini telah berhasil mengimplementasikan vector space model dan pembobotan TF-IDF dan WIDF dengan fungsi Bray/ Curtis dan algoritma Nazief Andriani. Dimana hasil perbandingan tersebut akan menghasilkan angka 0 sampai dengan 1, yang artinya adalah untuk menyatakan tingkat relevansi dokumen dengan term query.

Kata Kunci: Information Retrieval (IR), Model Ruang Vektor, TF-IDF, WIDF, Bray/ Curtis

ABSTRACT

Information Retrieval (IR) System is used to find information that is relevant to the needs of its users automatically based on conformity with the query of a collection of information. In my research this time aims to get to know some of the characteristics of the TF-IDF weighting notation and WIDF and implements a vector space model using some notation on the TF-IDF weighting method and Information Retrieval (IR) System at the WIDF. The function used for TF-IDF weighting and WIDF on research this is a function of the Bray/Curtis. On the vector space Model, the term weighting is done on document and query. The resulting weighting algorithm TF-IDF and WIDF will become variable in the calculation of the cosine similarity. The result of the cosine similarity on each document against a query will be sorted in descending, so the search results will display the document closest keyword or relevance. The system was developed using the PHP programming language and documents used as test data as much as 40 news article author quoted from several sites on the internet. This research has been successfully implementing the vector space model and the TF-IDF weighting and WIDF with Bray/Curtis and Nazief Andriani Algorithm. Where the results of the comparison would yield the number 0 to 1, which means it is to declare the level of relevance of the documents with the term query.

Keywords: Information Retrieval (IR), The Vector Space Model, TF-IDF, WIDF, Bray/Curtis

KATA PENGANTAR

Dengan mengucapkan puji dan syukur kepada Allah SWT atas berkat, rahmat dan taufik-Nya, penulis dapat menyelesaikan penulisan tesis ini sebagai salah satu syarat kelulusan dari Pascasarjana Magister Ilmu Komputer Universitas Budi Luhur.

Penulis sadar bahwa masih banyak kekurangan yang terdapat dalam penulisan ini yang harus diperbaiki. Penulis mengharapkan kritik dan saran untuk memperbaiki penulisan tesis sehingga nantinya akan menghasilkan penelitian yang lebih baik. Penulis banyak sekali mendapatkan bantuan dan bimbingan dalam penulisan tesis ini dan dengan segala kerendahan hati saya haturkan ucapan terima kasih sebesar - besarnya kepada:

1. Orang Tua dan keluarga tercinta, yang telah memberikan dukungan, semangat dan yang tak henti-hentinya mendo'akan.
2. Suamiku yang selalu mendukung serta membantu penulis dalam menyelesaikan tesis ini sehingga berjalan dengan lancar.
3. Dr. Krisna, M.Sc, M.M selaku pembimbing yang telah banyak memberikan pengarahan dan meluangkan waktu dalam proses penulisan.
4. Prof. Moedjiono, M.Sc, selaku Direktur Program Pascasarjana, Universitas Budi Luhur.
5. Dr. Ir. Nazori, M.T, selaku Dekan Program Pascasarjana Magister Ilmu Komputer, Universitas Budi Luhur.
6. Semua pihak yang telah membantu selesainya penulisan tesis ini.

Jakarta, 16 Januari 2016

Penulis

DAFTAR ISI

	Halaman
LEMBAR JUDUL	i
LEMBAR PERNYATAAN	ii
LEMBAR PENGESAHAN TESIS.....	iii
ABSTRAK	iv
ABSTRACT	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR GAMBAR	xii
DAFTAR TABEL	xiv
DAFTAR LAMPIRAN.....	xv
BAB I PENDAHULUAN	
1.1. Latar Belakang.....	1
1.2. Latar Belakang.....	3
1.2.1. Identifikasi Masalah	3
1.2.2. Batasan Masalah.....	3
1.2.3. Rumusan Masalah.....	4
1.3. Tujuan dan Manfaat Penelitian.....	4
1.3.1. Tujuan Penelitian.....	4
1.3.2. Manfaat Penelitian.....	4
1.4. Tata Urut Penulisan	4
1.5. Daftar Pengertian	5

BAB II LANDASAN TEORI DAN KERANGKA KONSEP

2.1. Tinjauan Pustaka.....	7
2.1.1. Pengertian IR (<i>Information Retrieval</i>).....	7
2.1.2. Bagian-bagian IR.....	8
2.1.2.1 <i>Text Operations / Preprocessing</i>	8
2.1.2.2 <i>Indexing</i>	9
2.1.2.3 <i>Query Operations</i>	10
2.1.2.4 <i>Searching</i>	11
2.1.2.5 <i>Ranking</i>	11
2.1.3. Proses <i>Information Retrieval</i>	12
2.1.4. Aturan <i>stemming</i> dalam bahasa Indonesia.....	13
2.1.5. Kata.....	17
2.1.6. Kalimat	18
2.1.7. Paragraf	19
2.1.8. <i>Pemrosesan Teks</i>	29
2.1.8.1 Pemecahan Kalimat.....	20
2.1.8.2 <i>Case Folding</i>	20
2.1.8.3 <i>Filtering Kalimat</i>	21
2.1.8.4 <i>Tokenizing Kata</i>	21
2.1.8.5 <i>Automatic Text Summarization</i>	21
2.1.9. <i>Pemrosesan Teks Algoritma TF-IDF</i>	22
2.1.10. <i>Cosine Similarity</i>	23
2.1.11. <i>Algoritma k-Nearest Neighbor (k-NN atau KNN)</i> .	23
2.1.12. <i>Algoritma WIDF</i>	27
2.1.13. <i>Bray / Curtis</i>	27
2.1.14. <i>Unified Modeling Language</i>	28
2.1.15. <i>Usecase Diagram</i>	29

2.1.15.1	Aktor	29
2.1.15.1.1	<i>Use case</i>	29
2.1.15.1.2	Relasi.....	29
2.1.15.2.	Activity diagram	30
2.1.15.2.1.	<i>Action</i> dan <i>Action State</i>	30
2.1.15.2.2.	<i>Transition</i>	31
2.1.15.2.3.	Percabangan (<i>Branching</i>)... ..	31
2.1.15.2.4.	<i>Forking</i> dan <i>joining</i>	32
2.1.16.	Model OOAD	32
2.2	Tinjauan Studi.....	33
2.3	Tinjauan Obyek Penelitian.....	38
2.4	Kerangka Konsep.....	38
2.5	Hipotesis	40
 BAB III METODOLOGI DAN RANCANGAN PENELITIAN		
3.1.	Metode Penelitian	41
3.2.	Metode Pemilihan Sampel.....	41
3.3.	Metode Pengumpulan Data.....	41
3.4.	Instrumentasi.....	42
3.5.	Teknik Analisa, Perancangan dan Pengujian Sistem, Rancangan Strategis Analisa Kebutuhan.....	42
3.6.	Langkah-langkah Penelitian	49
3.6.1.	Studi Literatur.....	49
3.6.2.	Pengumpulan Data.....	50
3.6.3.	Analisis dan Perancangan.....	50
3.6.4.	Implementasi	50
3.6.5.	Testing	50

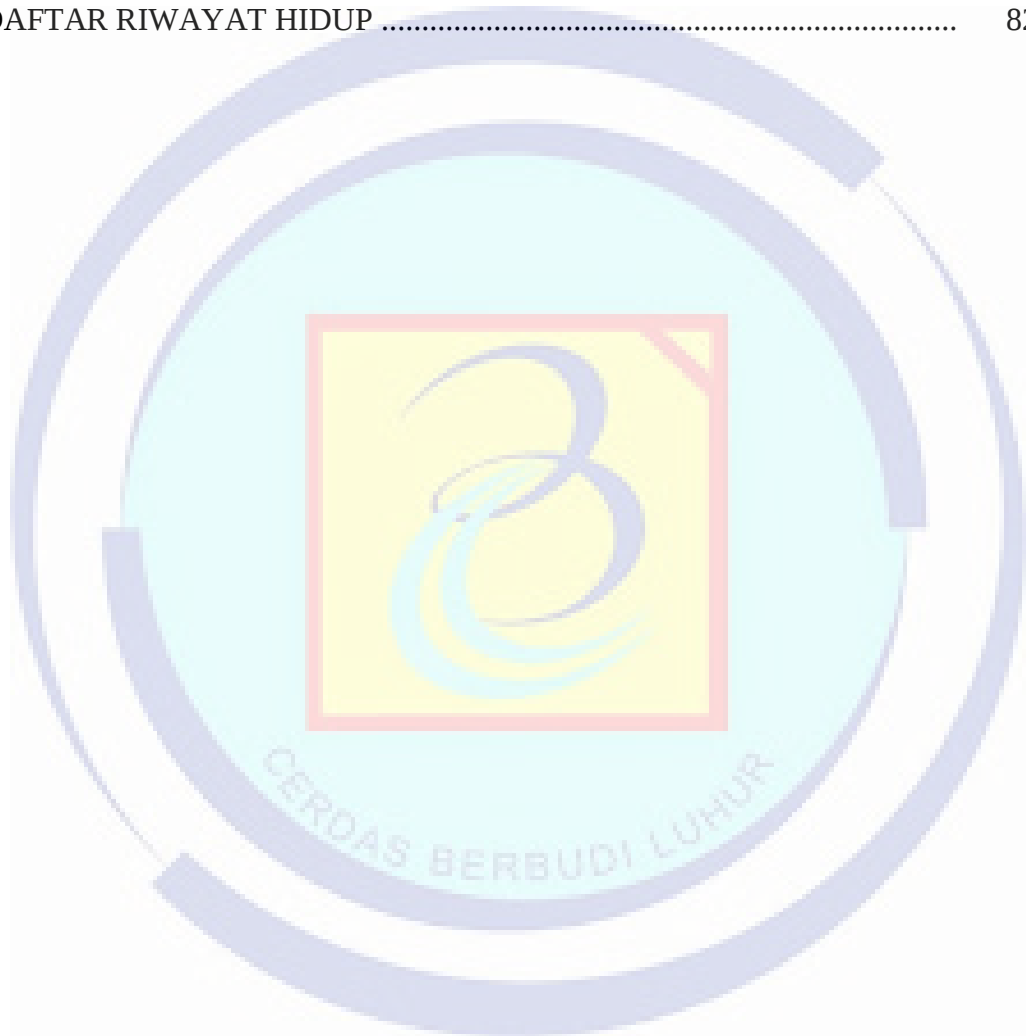
3.6.7.	Analisis Hasil Testing.....	50
3.6.8.	Pengambilan Kesimpulan.....	51
3.6.8.	Pembuatan Laporan	51
3.7.	Jadwal Penelitian	51

BAB IV PEMBAHASAN HASIL PENELITIAN

4.1.	Analisis Kebutuhan <i>Vector Space Model TF-IDF</i> dan <i>Widf</i>	52
4.2.	Perancangan Sistem/ <i>User Interface</i>	52
4.2.1.	<i>Use Case</i>	52
4.2.2.	<i>Activity Diagram</i>	54
4.2.3.	<i>Sequence Diagram</i>	55
4.2.4.	Implementasi Rancangan Sistem.....	54
	4.2.4.1 Tampilan Rancangan dari Sigi Pengguna..	57
	4.2.4.2 Tampilan Rancangan dari Sigi Admin	58
4.2.5.	Uji Kepuasan User.....	64
4.2.6.	Pengujian Sistem	66
4.3.	Implikasi Penelitian	68
4.3.1.	Aspek Sistem	68
4.3.1.	Aspek Manajerial.....	68
4.3.1.	Aspek Penelitian Lanjutan.....	69
4.4.	Rencana Implementasi Sistem	69
4.4.1.	Tahapan Implementasi Sistem.....	69
4.4.2.	Kegiatan Implementasi Sistem	69

BAB V PENUTUP

5.1. Kesimpulan	74
5.2. Saran.....	74
DAFTAR PUSTAKA	76
LAMPIRAN-LAMPIRAN	78
DAFTAR RIWAYAT HIDUP	82



DAFTAR GAMBAR

Gambar	Halaman
II.1 Interaksi user dengan sistem IR (Sumber : Baeza-Yates, 1999)	7
II.2 Tahap Preprocessing pada sistem IR (Sumber : Baeza-Yates, 1999)	8
II.3 Contoh teks dan pembangunan <i>inverted index</i> -nya (Sumber: Baeza-Yates, 1999)	10
II.4 Proses pencarian informasi (Sumber: Baeza-Yates, 1999)	13
II.5 <i>Graf Aturan Stemming</i>	17
II.6 <i>Proses text preprocessing</i>	19
II.7 <i>Algoritma KNN</i>	25
II.8.1 Delapan titik dalam satu dimensi dan estimasi densitas KNN dengan $k=3$ dan $k=5$	26
II.8.2 KNN mengestimasi densitas dua dimensi dengan $k=5$	26
II.9 <i>Usecase diagram</i>	30
II.10 Kerangka Pemikiran	39
III.1 Tampilan Menu Utama	43
III.2 Tampilan Search	43
III.3 Tampilan Liat Dokumen	44
III.4 Tampilan Antar Muka Sistem	44
III.5 Tampilan Menu Lihat Berita	45
III.6 Tampilan Menu Masukkan Berita	45
III.7 Tampilan Menu Search UBL QU	46
III.8 Tampilan Menu About	46
III.9 Tampilan Menu Hasil Tokenisasi	47
III.10 Tampilan Menu Hasil Stemming	47

III.11	Tampilan Menu Hasil Stop Word	48
III.12	Tampilan Hasil Vector Space Model TF-IDF	48
III.13	Langkah-Langkah Penelitian	49
IV.1	Use case user Search UBL QU	53
Gambar		Halaman
IV.2	Use case Admin Search UBL QU.....	53
IV.3	Activity Diagram User Search UBL QU	54
IV.4	Activity Diagram Admin Search UBL QU.....	55
IV.5	Sequance Diagram User Search UBL QU	56
IV.6	Sequance Diagram Admin Search UBL QU	56
IV.7	Menu Utama Pengguna.....	57
IV.8	Menu Search Pengguna.....	57
IV.9	Tampilan Link Dokumen Pengguna	58
IV.10	Menu Utama Admin	58
IV.11	Tampilan Masukan Berita.....	59
IV.12	Menu Lihat Berita.....	59
IV.13	Menu Search UBL QU.....	60
IV.14	Menu About	60
IV. 15	Menu Hasil Tokenisasi	61
IV. 16	Menu Hasil Stemming	61
IV. 17	Menu Hasil Stop Word	62
IV.18	Menu Hasil Vector Space Model (VSM) Tf-Idf.....	62
IV. 19	Menu Hasil Vector Space Model Widf.....	63
IV. 20	Menu Menu Hasil Bray/ Curtis.....	64
IV. 21	Menu Menu Hasil Relevansi.....	64
IV. 22	Grafik Rekapitulasi Hasil Kuisisioner.....	66
IV. 23	Flow chart Langkah – langkah aplikasi program.....	82

DAFTAR TABEL

Tabel		Halaman
II.1	Awalan, Sisipan, Leburan, dan Akhiran pada Bahasa Indonesia	14
II.2	Contoh pemecahan kalimat.....	20
II.3	Contoh <i>case folding</i>	20
II.4	Contoh <i>Filtering</i> kalimat.....	21
II.5	Contoh <i>Tokenizing</i> kata.....	21
II.6	Tabel Tinjauan Studi.....	36
III.1	Jadwal Penelitian.....	49
IV.1	Pembobotan Skala Likert.....	65
IV.2	Tabel Hasil Kuisioner	65
IV.3	Pengujian Black Box.....	67
IV.4	Tabel Rencana Implementasi.....	70

DAFTAR LAMPIRAN

Lampiran	Halaman
1 Kartu Bimbingan Tesis	78
2 Source Code	79
3 Hitungan Manual	80
4 Kuisisioner.....	81





DAFTAR PUSTAKA

- [Agusta 2009] Agusta, Ledy. *Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia*, 2009.
- [Ellis] Ellis, David, and Furner Hines Jonatham, and Willett Peter. *Measuring the Degree of Similarity Between Objects in Text Retrieval Sytems*
- [Gregorius] Gregorius, S. Budhi dan Gunawan, Ibnu dan Yuwono, Ferry. *Algoritma Porter Stemmer For Bahasa Indonesia untuk Pre-Processing Text Mining Berbasis Metode Market Basket Analysis*.
- [Herwansyah] Herwansyah, Adhit Herwansyah. *Aplikasi Pengkategorian Dokumen dan Pengukuran Tingkat Similaritas Dokumen Menggunakan Kata Kunci Pada Dokumen Penulisan Ilmiah Universitas Gunadarma*.
- [Jaya 2007] Jaya, Hendra. *Perbandingan Performansi Word Indexing dan Phrase Indexing dalam Sistem Temu Balik Informasi dengan Menggunakan Model Probabilistik*. Skripsi Terpublikasi. Bandung: Institut Teknologi Bandung, 2007.
- [Kurniawan] Kurniawan, Ernie dan Fransiska, Maria dan Tinaliah dan Rachmansyah. *Penerapan Algoritma K-Means untuk Clustering Dokumen E-Jurnal STMIK GI MDP*.
- [Mandala 2002] Mandala, R. dan Setiawan, H. *Peningkatan Performansi Sistem Temu- Kembali Informasi dengan Perluasan Query Secara Otomatis*. Bandung: Institut Teknologi Bandung, 2002.