

**STUDI KOMPARATIF ATAS VSM YANG MENERAPKAN KOEFISIEN
RUSSELL/RAO PADA TEKNIK KNN TERHADAP TF-IDF DAN WIDF:
KASUS PROTOTYPE APLIKASI KATEGORISASI TEKS BERBAHASA INDONESIA
DENGAN STEMMING NAZIEF ADRIANI**



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

JAKARTA

2016

**STUDI KOMPARATIF ATAS VSM YANG MENERAPKAN KOEFISIEN
RUSSELL/RAO PADA TEKNIK KNN TERHADAP TF-IDF DAN WIDF:
KASUS PROTOTYPE APLIKASI KATEGORISASI TEKS BERBAHASA INDONESIA
DENGAN STEMMING NAZIEF ADRIANI**

TESIS

Diajukan untuk memenuhi salah satu persyaratan memperoleh gelar Magister
Ilmu Komputer (MKOM)



Oleh
BAMBANG PRIAMBODO
1411600420

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

JAKARTA

2016



**PROGRAM STUDI: MAGISTER ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR**

LEMBAR PERNYATAAN

Nama Mahasiswa : Bambang Priambodo
Nomor Induk Mahasiswa : 1411600420
Program Studi : Magister Ilmu Komputer
Konsentrasi : Rekayasa Komputer Terapan
Fakultas / Program : Program Pasca Sarjana

menyatakan bahwa Tesis yang berjudul :

Studi komparatif atas VSM yang menerapkan koefisien Russell/4ao
pada teknik KAN Terhadap TF-IDF dan W-IDF : Kasus prototipe
aplikasi kategorisasi teks berbahasa Indonesia dgn stemming
Nazief Adrian.

1. merupakan hasil karya tulis ilmiah sendiri dan bukan merupakan karya yang pernah diajukan untuk memperoleh gelar akademik oleh pihak lain,
2. saya ijin 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 di kemudian hari pernyataan ini tidak benar.

Jakarta, 29 Januari 2016



Bambang Priambodo



**PROGRAM STUDI: MAGISTER ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR**

LEMBAR PENGESAHAN

Nama Mahasiswa : Bambang Priambodo
Nomor Induk Mahasiswa : 1411600420
Konsentrasi : Rekayasa Komputasi Terapan
Jenjang Studi : Strata 2 / S2
Judul Tesis : Studi Komparatif atas VSM yang Menerapkan
Koefisien Russell/Rao pada Teknik KNN terhadap
TF-IDF dan W-IDF: Kasus Prototipe Aplikasi
Kategori Teks Berbahasa Indonesia dengan Stemming
Nazief Adriani

Jakarta, 29 Januari 2016

Tim Penguji:

Tanda Tangan:

Ketua,

(Prof. DR. Moedjiono, M.Sc)

Anggota,

(DR. Ir. Nazori Az, M.T)

Pembimbing,

(Dr. Krisna Adiyarta, M.Sc)

Ketua Program Studi

(Dr. Ir. Nazori Az, M.T)

ABSTRAK

Proses kategorisasi secara otomatis terhadap dokumen hukum yang dalam format PDF melibatkan pembobotan teks dalam dokumen untuk menentukan profil dari sebuah dokumen dalam kumpulan dokumen. Pembobotan teks dengan TF-IDF mempunyai kelemahan yaitu semua dokumen yang berisi kata tertentu diperlakukan sama, dengan kata lain jika istilah "laut" terdapat di dalam 4 dokumen dengan frekuensi yang berbeda di masing-masing dokumen-dokumen ini, TF-IDF tidak mempertimbangkan berapa kali dan di mana saja kata "laut" disebutkan di dalam 4 dokumen tersebut. Penelitian ini akan membandingkan antara teknik pembobotan TF-IDF dengan WIDF dalam VSM yang menggunakan koefisien Russel/Rao pada *Machine Learning* yang menggunakan *K-Nearest Neighbor*. Berdasar tinjauan terhadap penelitian lain yang menjadi rujukan, teknik pembobotan WIDF menunjukkan hasil yang lebih baik daripada TF-IDF. Penelitian ini akan fokus kepada dokumen yang berbahasa Indonesia, terutama dokumen produk hukum. Penelitian ini juga menggunakan *stemming* yang khusus untuk bahasa Indonesia yaitu *stemming* Nazief Adriani yang merupakan varian dari teknik *stemming* Porter yang khusus untuk teks berbahasa Inggris. Dari hasil rangkaian percobaan dalam penelitian ini disimpulkan bahwa untuk teknik pemrofilan kata W-IDF menunjukkan performansi lebih baik dari teknik pemrofilan TF-IDF. Nilai rata-rata dari Presisi tinggi dihasilkan dari nilai k kemudian nilainya menurun pada nilai k yang semakin besar. Hal sebaliknya terjadi pada rata-rata nilai Recall, dimana nilai yang besar dihasilkan dari nilai k yang besar dan menurun pada nilai k yang semakin kecil. Selain itu hasil penelitian juga menunjukkan bahwa teknik TF-IDF maupun W-IDF dapat mempengaruhi dalam menentukan profil sebuah dokumen atas kata-kata dasar pembentukannya.

Kata Kunci: Kategorisasi Teks, Pembobotan Kata, TF-IDF, WIDF, Vektor Space Model, Russel-Rao, KNN

ABSTRACT

Automatic categorization process of the legal documents in PDF format involves text weighting in the document to determine the profile of a document in a group of documents. Profiling using TF-IDF has disadvantage where all documents which contain certain words are treated as the same, in other words if the term "sea" contained in four documents with different frequencies in each of these documents, TF-IDF does not consider how many times and where the word "sea" is mentioned in four documents. This study will compare profiling technique TF-IDF with WIDF in VSM using the coefficient Russell / Rao on Machine Learning by using K-Nearest Neighbor. Based on a review of other similar studies, WIDF profiling techniques showed better results than TF-IDF. But so far the previous similar studies have not been tested by the Indonesian-language text, this study will use the Indonesian language text and use Nazief Andriani's stemming algorithms. From the results of a series of experiments in this study concluded that profililing techniques W-IDF showed better performance than profiling techniques TF-IDF. The average value of high precision resulting from the value of k then its value decreases in the value of k increases. The opposite occurs in the average value of Recall, where great value is generated from large k value and decreases at smaller values of k . In addition the results also showed that the technique TF-IDF and W-IDF can influence in determining the profile of a document on the basis of word formation.

Keywords: Text Categoization, Term Weighting, TF-IDF, WIDF, Vector Space Model, Russell-Rao, KNN

KATA PENGANTAR

Puji syukur saya panjatkan ke hadirat Allah SWT karena atas rahmat-Nyalah saya dapat menyelesaikan penyusunan laporan penelitian yang berjudul **“STUDI KOMPARATIF ATAS VSM YANG MENERAPKAN KOEFISIEN RUSSELL/RAO PADA TEKNIK KNN TERHADAP TF-IDF DAN WIDF: KASUS PROTOTYPE APLIKASI KATEGORISASI TEKS BERBAHASA INDONESIA DENGAN STEMMING NAZIEF ADRIANI”**. Penyusunan laporan penelitian ini merupakan salah satu persyaratan untuk menyelesaikan kuliah di Program Pascasarjana, dengan Program Studi Magister Ilmu Komputer di Universitas Budi Luhur.

Dalam penyusunan laporan penelitian ini, saya menyampaikan ucapan terima kasih yang tidak terhingga kepada pihak-pihak yang membantu dalam menyelesaikan laporan ini, khususnya kepada:

1. Allah SWT yang telah memberikan waktu dan petunjuk-Nya dalam menyelesaikan laporan penelitian ini.
2. Istri, anak-anak dan keluarga besar tercinta yang telah memberikan dorongan dan bantuan serta pengertian yang besar kepada kami dalam menyelesaikan proposal ini.
3. Bapak Dr. Krisna Adiyarta, selaku Pembimbing yang telah menyediakan waktu, saran, bantuan, bimbingan, pengarahan, dorongan, kepercayaan dan kesempatan kepada Penulis dalam menyelesaikan masalah-masalah yang dihadapi oleh Penulis selama menyelesaikan proposal tesis ini.
4. Bapak Dr. Dwi Handoko, M.Eng, yang telah membantu dalam pengujian algoritma dan data.
5. Dosen serta staff dari Universitas Budi Luhur terutama dari Prodi Magister Ilmu Komputer yang telah membantu banyak selama ini.
6. Rekan-rekan mahasiswa mulai dari Semester I, II dan III.

7. Rekan-rekan dari Paralabs.
8. Semua pihak yang tidak dapat disebutkan satu persatu, yang telah memberikan bantuan dalam penyusunan proposal ini.

Dalam penyusunan proposal ini, saya merasa masih banyak kekurangan baik pada teknis penulisan maupun materi, mengingat akan kemampuan yang dimiliki. Untuk itu, kritik dan saran dari semua pihak sangat kami harapkan demi penyempurnaan pembuatan laporan ini.

Akhir kata, penulis berharap semoga Allah SWT memberikan imbalan yang setimpal pada mereka yang telah memberikan bantuan, dan dapat menjadikan semua bantuan ini sebagai ibadah, Aamiin Yaa Robbal ‘Alamiin.

Jakarta, Januari 2016

Bambang Priambodo

Halaman Persembahan

*Kupersembahkan untuk istri (Ika) dan anak-anakku (Lathiifah dan Nadia)
tercinta, serta keluarga besarku
(Bambang - Ika)*



DAFTAR ISI

ABSTRAK	IV
ABSTRACT	V
KATA PENGANTAR	VI
HALAMAN PERSEMBAHAN	VIII
<i>KUPERSEMBAHKAN UNTUK ISTRI (IKA) DAN ANAK-ANAKKU (LATHIIFAH DAN NADIA) TERCINTA, SERTA KELUARGA BESARKU</i>	VIII
DAFTAR ISI	IX
DAFTAR TABEL.....	XI
DAFTAR GAMBAR.....	XII
DAFTAR LAMPIRAN	XIII
BAB I PENDAHULUAN	1
1.1. LATAR BELAKANG.....	1
1.2. MASALAH PENELITIAN.....	2
1.2.1 Identifikasi Masalah.....	2
1.2.2. Batasan Masalah.....	3
1.2.3. Rumusan Masalah	3
1.3. TUJUAN DAN MANFAAT PENELITIAN	3
1.4. RUANG LINGKUP	3
1.5. HASIL YANG DIHARAPKAN	3
1.6. SISTEMATIKA PENULISAN	4
1.7. DAFTAR PENGERTIAN/ISTILAH.....	4
BAB II LANDASAN TEORI DAN KERANGKA KONSEP	6
2.1. TEORI PENDUKUNG.....	6
2.1.1. Kategorisasi Teks.....	6
2.1.3. Stemming Nazief Adriani	8
2.1.4. TF-IDF.....	10
2.1.5. WIDF	11
2.1.6. Vektor Space Model	12
2.1.7. K-Nearest Neighbors.....	14
2.1.8. Pengukuran F.....	15
2.2. TINJAUAN STUDI.....	16
2.2.1. Penelitian yang dilakukan oleh Musa Hadi et al.....	16
2.2.2. Penelitian yang dilakukan oleh C. Deisy et al.....	16
2.2.3. Perbandingan	16
2.3. TINJAUAN OBYEK PENELITIAN.....	17
2.4. KERANGKA PEMIKIRAN	18
2.5. HIPOTESIS.....	19
BAB III METODOLOGI DAN PERANCANGAN PENELITIAN	20
3.1. METODE YANG DIAJUKAN	20
3.2. PEMILIHAN SAMPEL	21
3.3. METODE PENGUMPULAN DATA	22

3.4. PENGUJIAN HIPOTESA.....	22
3.5. INSTRUMENTASI.....	23
3.6. TEKNIK ANALISIS PROTOTIPE MODEL	24
3.6.1. Teknik Analisis	24
3.6.2. Prototipe Model.....	24
3.6.3. Teknik Pengindeksan.....	26
3.6.4. Teknik Pengujian.....	28
3.6.4. Teknik Pengukuran Performansi	28
3.7. LANGKAH-LANGKAH PENELITIAN.....	30
3.8. JADWAL PENELITIAN.....	30
BAB IV PEMBAHASAN HASIL PENELITIAN	33
4.1. DOKUMEN DAN PENGELOMPOKAN DOKUMEN.....	33
4.2. PROTOTIPE MODEL MESIN KLASIFIKASI.....	33
4.2.1. Prototipe CLI.....	33
4.2.2. Prototipe Web.....	34
4.3. HASIL PENGUKURAN BERDASAR ISI DOKUMEN LATIH	35
4.4. HASIL PENGUJIAN	37
4.4.1. KNN, $K=5$	38
4.4.2. KNN dengan $K=6$, $K=10$ dan $K=15$	39
4.5. IMPLIKASI PENELITIAN.....	42
4.5.1 Aspek Sistem	42
4.5.2. Aspek Manajerial.....	44
4.5.3. Aspek Penelitian Lanjut.....	45
4.6. RENCANA IMPLEMENTASI SISTEM.....	46
BAB V PENUTUP.....	49
5.1. KESIMPULAN.....	49
5.2. SARAN	49
DAFTAR PUSTAKA.....	51
LAMPIRAN-LAMPIRAN	54
DAFTAR RIWAYAT HIDUP.....	66

DAFTAR TABEL

Tabel		Halaman
2.1	Tabel Variabel n pada Russell/Rao	14
2.2	Tabel Himpunan kemungkinan dalam IR	16
2.3	Tabel Perbandingan Penelitian	17
3.1	Template Tabel Hasil Pengukuran F	23
3.2	Jadwal Penelitian	31
4.1	25 Kata dasar yang paling sering muncul dari 90 dokumen latih	36
4.2	5 Kata dasar yang hanya sekali muncul dari 90 dokumen latih	36
4.3	25 df tertinggi dari kata dasar yang terdaftar dalam dokumen latih	37
4.4	5 df terendah dari kata yang terdaftar dalam dokumen latih	37
4.5	Rata-rata Nilai Precision dengan $k=5$	38
4.6	Rata-rata Nilai Recall dengan $k=5$	38
4.7	Rata-rata Nilai F dengan $k=5$	38
4.8	Rata-rata Nilai F dengan $k=6$	39
4.9	Rata-rata Nilai F dengan $k=10$	39
4.10	Rata-rata Nilai F dengan $k=15$	40
4.11	Perbandingan Nilai rata-rata F	40
4.12	Standar minimal spesifikasi server	43
4.13	Rencana Implementasi Sistem	46

DAFTAR GAMBAR

Gambar		Halaman
2.1	Proses Supervised TC	7
2.2	Preprocessing	7
2.3	Matriks Kata-Dokumen	12
2.4	Representasi dokumen dan vektor pada ruang vektor	13
2.5	KNN	15
2.6	Kerangka Pemikiran	18
3.1	Ekstraksi Teks dan Pembobotan Term	21
3.2	Proses Klasifikasi dengan KNN	21
3.3	Pengumpulan Data	22
3.4	Pengujian Hipotesis	23
3.5	Trade-off antara precision dan recall	29
4.1	Contoh penggunaan prototipe CLI untuk preprosesing data latih	34
4.2	Contoh penggunaan purwarupa web untuk Perhitungan Kueri dan Display Hasil Kueri	35
4.3	Grafik hasil perbandingan Nilai rata-rata F	40
4.4	Grafik Rata-rata nilai presisi	41
4.5	Grafik Rata-rata nilai recall	42

DAFTAR LAMPIRAN

Lampiran		Halaman
A	Daftar Stopwords	54
B	Stemming	60
C	Presisi, Recall dan F Untuk K=6	63
D	Presisi, Recall dan F Untuk K=10	64
E	Presisi, Recall dan F Untuk K=15	65



DAFTAR PUSTAKA

- [Agusta 2009] Agusta, Ledy. 2009. *Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia*, Konferensi Nasional Sistem Dan Informatika, Bali, 196–201.
- [Agus 2002]. Agus Zainal Arifin, 2002, *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*, Surabaya, Institut Teknologi Sepuluh Nopember (ITS).
- [Antonie 2002] Antonie, Maria-luiza. 2002. *Text Document Categorization by Term Association*, Proceedings of the IEEE International Conference on Data Mining (ICDM '2002), 19–26.
- [Manning ea 2009] D Manning, Christopher, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*, Information Retrieval, Cambridge UP. doi:10.1109/LPT.2009.2020494.
- [Deisy ea 2010] Deisy, C., M. Gowri, S. Baskar, S. M a Kalaiarasi, and N. Ramraj. 2010. *A Novel Term Weighting Scheme Midf for Text Categorization*, Journal of Engineering Science and Technology 5 (1): 94–107.
- [Dipa, 2015]. Dipa 2015, *Sistem Temu Kembali Informasi atau Information Retrieval*, http://baliinnetwork.blogspot.com/2015/03/sistem-temu-kembali-informasi-atau_3.html. (Diakses 1 Oktober 2015)
- [Ellis ea 1993] Ellis, David, Jonathan Furner-Hines, and Peter Willett. 1993. *Measuring the Degree of Similarity between Objects in Text Retrieval Systems*, Information Management 3.2 (January): 128–49.
- [Fatmawati, 2013]. Fatmawati, 2013, *Studi Komparatif Kecepatan Temu Kembali Informasi di Depo Arsip Koran Suara Merdeka Antara Sistem Simpan Manual dengan Foto Repro*. Diponegoro University.
- [Hadhiatma 2010] Hadhiatma, Agung, Jurusan Teknik Informatika, Universitas Sanata, and Dharma Yogyakarta. 2010, *Pencarian Dokumen Berdasarkan Kombinasi Antara Model Ruang Vektor Dan Model Domain Ontologi*, 2010 (semnasIF): 111–17.
- [Joachims 1998] Joachims, Thorsten. 1998, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Machine Learning: ECML-98 1398 (2): 137–42. doi:10.1007/BFb0026683.
- [Makhoul ea 1999] Makhoul, J, J Makhoul, F Kubala, F Kubala, R Schwartz, R Schwartz, R Weischedel, and R Weischedel. 1999, *Performance Measures for Information Extraction*. <http://citeseer.ist.psu.edu/makhoul99performance.html>. (Diakses 1 Oktober 2015)
- [Musa ea 2007] Musa Hadi, Wa'el, Fadi Thabtah, and Hussein Abdel-jaber. 2007. *A Comparative Study Using Vector Space Model with K-Nearest Neighbor*