

**MODEL MESIN PENCARI DOKUMEN:
Studi Efektifitas Pada Penerapan *Vector Space Model* dengan
Algoritma Porter, Pembobotan Berdasarkan Kata Berbanding
Jumlah Kata Dalam Dokumen Serta Fungsi Kesamaan Jaccard**

TESIS



Oleh:

WIDYA RATRI

1211601974

PROGRAM STUDI: MAGISTER ILMU KOMPUTER (MKOM)

PROGRAM PASCASARJANA

UNIVERSITAS BUDI LUHUR

JAKARTA

2015

**MODEL MESIN PENCARI DOKUMEN:
Studi Efektifitas Pada Penerapan *Vector Space Model* dengan
Algoritma Porter, Pembobotan Berdasarkan Kata Berbanding
Jumlah Kata Dalam Dokumen Serta Fungsi Kesamaan Jaccard**

TESIS

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



Oleh:

WIDYA RATRI

1211601974

PROGRAM STUDI: MAGISTER ILMU KOMPUTER (MKOM)

PROGRAM PASCASARJANA

UNIVERSITAS BUDI LUHUR

JAKARTA

2015



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

LEMBAR PERNYATAAN

Saya yang bertanda tangan dibawah ini,

Nama : Widya Ratri
NIM : 1211601974
Program Studi : Teknologi Sistem Informasi
Fakultas : Magister Ilmu Komputer

Menyatakan bahwa TESIS yang berjudul:

MODEL MESIN PENCARI DOKUMEN: Studi Efektifitas Pada Penempatan *Vector Space Model* dengan Algoritma Porter, Pembobotan Berdasarkan Kata Berbanding Jumlah Kata Dalam Dokumen Serta Fungsi Kesamaan Jaccard

1. Menupakan hasil karya tulis ilmiah sendiri dan bukan merupakan karya yang pernah diajukan untuk memperoleh gelar akademik oleh pihak lain,
2. Saya ijinan 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, 10 September 2015


6000
(Widya Ratri)



PROGRAM STUDI: MAGISTER ILMU KOMPUTER
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR

LEMBAR PENGESAHAN

Nama : Widya Ratri
NIM : 1211601974
Bidang Peminatan : Teknologi Sistem Informasi
Jenjang Studi : Strata 2
Judul : MODEL MESIN PENCARI DOKUMEN: Studi Efektifitas Pada Penerapan *Vector Space Model* dengan Algoritma Porter, Pembobotan Berdasarkan Kata Berbanding Jumlah Kata Dalam Dokumen Serta Fungsi Kesamaan Jaccard

Jakarta, September 2015

Tim Penguji:

Ketua Penguji I,
Prof. Dr. Moedjiono, M.Sc

Penguji II,
Dr. Ir. Nazori AZ, MT

Pembimbing Utama,
Dr. Krisna Adiyarta M.

Tanda Tangan

Ketua Program Studi,

Dr. Ir. Nazori AZ, M.T

ABSTRAK

Perkembangan teknologi saat ini membawa pengaruh besar dalam proses penyimpanan maupun pencarian informasi di berbagai belahan dunia. Pencarian informasi berdasarkan query oleh pengguna, yang diharapkan dapat menemukan koleksi dokumen berdasarkan kebutuhan pengguna, dikenal dengan *Information Retrieval System* atau sistem temu balik informasi. Penelitian ini membahas tentang implementasi sistem temu balik informasi untuk mencari dan menemukan dokumen teks berbahasa Indonesia menggunakan *Vector Space Retrieval Model*. Penelitian ini menggunakan algoritma *Stemming Porter* untuk memecah sebuah kata menjadi kata dasar. Tujuan penelitian ini untuk menyediakan solusi pada mesin pencarian agar mampu menyediakan informasi dokumen teks pada database yang tepat menggunakan kata kunci tertentu dengan tingkat kemiripan yang relevan. Hasil dari pencarian direpresentasikan dengan urutan/ranking kemiripan dokumen dengan *query*. Dapat disimpulkan bahwa tingkat keefektifitasan hasil pencarian lebih tepat jika dilakukan dengan menggunakan *query* berdasarkan *key word* pada dokumen. Pembobotan dengan menghitung frekuensi term yang muncul mempengaruhi persamaan dokumen. Semakin baik nilai suatu *query* yang dimasukkan maka semakin tinggi hasil nilai suatu *similarity* yang dikeluarkan maka semakin baik dan semakin banyak pula data yang relevan.

Kata kunci: *Information Retrieval System*, *search engine*, *Stemming Porter*, *Vector Space Retrieval Model*, *similarity*.

ABSTRACT

Development of technology today's brings great impact in either storing process or search information in around the world. Not only used as automation process of access the information, but create accuracy and speed in data search as well. Searching information based on query by user as known as Information Retrieval System. This research discusses about the implementation of Information Retrieval System to search and find text document in Bahasa Indonesia using Vector Space Retrieval Model. This research used Stemming Porter algorithm to split words to be a basic word. The goal of this research is to provide a solution to search engine to be able to serve text document on correct database using specific keyword with a relevant Similarity level. And the result of search will represented by ranking of document Similarity with query. It can be concluded that the degree of effectiveness of the search results more precise if it is done by using queries based on key word in the document. Weighting by counting the frequency of terms that appear affect the equation documents. The better the value of a query is entered, the higher the result the value of a similarity issued the better and the data is more relevant.

Keyword: Information Retrieval System, search engine, Stemming Porter, Vector Space Retrieval Model, Similarity

KATA PENGANTAR

Alhamdulillah puji dan syukur penulis panjatkan kehadirat Allah SWT, karena atas rahmat, taufiq, dan hidayah serta karuniaNya, tesis dengan Judul :

“MODEL MESIN PENCARI DOKUMEN: Studi Efektifitas Pada Penerapan *Vector Space Model* dengan Algoritma Porter, Pembobotan Berdasarkan Kata Dalam Dokumen Berbanding Jumlah Kata Dalam Dokumen Serta Fungsi Kesamaan Jaccard” ini dapat diselesaikan oleh penulis.

Tujuan pembuatan tesis ini adalah untuk memenuhi salah satu syarat menyelesaikan jenjang Strata Dua (S-2) pada Program studi Magister Ilmu Komputer (M.KOM) Program Sarjana Universitas Budi Luhur. Pada kesempatan ini penulis menyampaikan terima kasih dan penghargaan yang setinggi-tingginya kepada :

1. Kedua orang tua penulis, Joko Mulyarjo dan Endang Giri Tirtawati atas doa dan dukungan yang tidak pernah putus.
2. Suami tercinta, Audi Arief Pramadya yang tanpa lelah terus mendampingi dan menyemangati penulis serta memberi dukungan moral maupun materil.
3. Bima Yhoga Batara, Damar Ragil, Mita Oktawiranti, Riski Purwo Handariningsih, Vennynova Lianty dan Ardi Fajar Sucianto, sahabat dan keluarga penulis yang telah banyak membantu untuk kelancaran penelitian ini.
4. Prof. Dr. Moedjiono, M.Sc., selaku Direktur Program Pascasarjana Universitas Budi Luhur, yang telah melayani serta memberikan fasilitas bagi kelancaran studi.
5. Dr. Ir Nazori AZ, MT selaku ketua Program Studi Pascasarjana Magister Ilmu Komputer Universitas Budi Luhur.
6. Dr. Krisna Adiyarta Musodo selaku Pembimbing yang telah berusaha dengan sabar dan cermat membimbing dan mengarahkan penyusun untuk menyelesaikan penelitian ini.

7. Rekan-rekan seperjuangan di Program Pascasarjana Magister Ilmu Komputer Universitas Budi Luhur yang telah bantu-membanu agar tesis ini dapat diselesaikan dengan baik dan tepat waktu.
8. Seluruh Dosen Magister Ilmu Komputer Universitas Budi Luhur yang telah berkenan mentransfer dan membuka cakrawala ilmu pengetahuan kepada penyusun.

Serta semua pihak yang terlalu banyak untuk penulis sebutkan satu persatu sehingga terwujudnya penulisan tesis ini. Penulis menyadari bahwa penulisan tesis ini masih jauh sekali dari sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan karya ilmiah yang penulis hasilkan untuk yang akan datang.

Akhir kata semoga tesis ini dapat bermanfaat bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Jakarta, September 2015

Penulis

DAFTAR ISI

	Halaman
LEMBAR PERNYATAAN.....	i
LEMBAR PENGESAHAN	ii
ABSTRAK.....	v
ABSTRACT	iv
KATA PENGANTAR	v
DAFTAR ISI	vii
DAFTAR GAMBAR.....	x
DAFTAR TABEL.....	xi
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 Pembatasan Masalah.....	3
1.2.3 Rumusan Masalah	3
1.3 Tujuan dan Manfaat Penelitian	4
1.3.1 Tujuan Penelitian.....	4
1.3.2 Manfaat Penelitian.....	4
1.4 Tata Urut/Sistematika Penulisan.....	4
BAB II LANDASAN TEORI.....	6
2.1 Tinjauan Pustaka.....	6
2.1.1 IR (<i>Information Retrieval</i>).....	6
2.1.2 Pemprofilan Teks.....	8
2.1.3 <i>Vector space model</i>	9
2.1.4 Tokenisasi.....	9
2.1.5 Filtrasi.....	10
2.1.6 <i>Stop Word</i>	10
2.1.7 Stemming.....	11

2.1.8 Algoritma <i>Stemming</i> Porter	11
2.1.9 Pencocokan Pemprofilan Teks	15
2.1.10 Teknik Pembobotan Kata	16
2.1.11 Pengukuran Keefektifitasan Perolehan Informasi	19
2.2 Tinjauan Studi	19
2.3 Kerangka Konseptual	23
2.4 Hipotesis	24
BAB III METODOLOGI	25
3.1 Metode Penelitian	25
3.2 Metode Pemilihan Sampel	25
3.3 Metode Pengumpulan Data	25
3.4 Instrumentasi	26
3.5 Metode Pengukuran Efektifitas	26
3.6 Langkah-langkah Penelitian	27
3.7 Jadwal Penelitian	29
BAB IV PEMBAHASAN HASIL PENELITIAN	30
4.1 Pengelompokan dan Analisis Data	30
4.2 Pengujian Penelitian	33
4.2.1 Bentuk Pengujian	33
4.2.2 Cara Pengujian	34
4.3 Hasil Pengujian	35
4.4 Analisa Hasil Pengujian	47
4.5 Implikasi Penelitian	49
4.5.1 Aspek Sistem	49
4.5.2 Aspek Manajerial	49
4.5.3 Aspek Penelitian Lanjut	50
4.6 Rencana Implementasi	50
4.6.1 Tahapan Implementasi	50
BAB V PENUTUP	51
5.1 Kesimpulan	51

5.2	Saran	51
DAFTAR PUSTAKA		53
LAMPIRAN – LAMPIRAN.....		55
RIWAYAT HIDUP SINGKAT		96



DAFTAR GAMBAR

Gambar	Halaman
2.1 Proses <i>Information Retrieval System</i>	7
2.2 <i>Stemming</i> Porter Bahasa Indonesia	12
2.3 Kerangka Konseptual.....	24
4.1 Grafik perbandingan rata-rata <i>precision</i> dan <i>recall</i>	48



CERDAS BERBUDI LUHUR

DAFTAR TABEL

Tabel	Halaman
2.1 Aturan Untuk <i>Inflectional Particle</i>	13
2.2 Aturan Untuk <i>Inflectional Possesive Pronoun</i>	13
2.3 Aturan Untuk <i>First Order Derivational Prefix</i>	13
2.4 Aturan Untuk <i>Second order Derivational PreviX</i>	14
2.5 Aturan Untuk <i>Derivational Suffix</i>	14
2.6 Contoh dokumen perhitungan	16
2.7 Proses Tokenisasi.....	17
2.8 Proses <i>Stemming</i>	17
2.9 Perhitungan pembobotan (<i>fin</i>)	17
2.10 Perhitungan pembobotan <i>term</i> berbanding jumlah <i>term</i>	18
2.11 Hasil perangkingan dokumen.....	18
2.12 Tabel Penelitian Terdahulu	21
3.1 Jadwal Penelitian	29
4.1 Tabel Judul Dokumen Kategori Teknologi	31
4.2 Tabel Judul Dokumen Kategori Otomotif	31
4.3 Tabel Judul Dokumen Kategori Sepak Bola.....	32
4.4 Tabel Judul Dokumen Kategori Ekonomi	32
4.5 Tabel Judul Dokumen Kategori Berita	33
4.6 Q1 <i>Query</i> berdasarkan judul dokumen kategori Teknologi.....	35
4.7 Q2 <i>Query</i> berdasarkan judul dokumen kategori Otomotif.....	36
4.8 Q3 <i>Query</i> berdasarkan judul dokumen kategori Sepak Bola.....	36
4.9 Q4 <i>Query</i> berdasarkan judul dokumen kategori Ekonomi.....	37
4.10 Q5 <i>Query</i> berdasarkan judul dokumen kategori Berita	38

4.11 Q6 Query berdasarkan isi dokumen kategori Teknologi	38
4.12 Q7Query berdasarkan isi dokumen kategori Otomotif	39
4.13 Q8 Query berdasarkan isi dokumen kategori Sepak Bola	40
4.14 Q9 Query berdasarkan isi dokumen kategori Ekonomi	41
4.15 Q10 Query berdasarkan isi dokumen kategori Berita	42
4.16 Q11 Query berdasarkan <i>key word</i> dokumen kategori Teknologi	43
4.17 Q12 Query berdasarkan <i>key word</i> dokumen kategori Otomotif	43
4.18 Q13 Query berdasarkan <i>key word</i> dokumen kategori Sepak Bola.....	44
4.19 Q14 Query berdasarkan <i>key word</i> dokumen kategori Ekonomi	45
4.20 Q15 Query berdasarkan <i>key word</i> dokumen kategori Berita	46
4.21 Rata-rata <i>Precision</i> dan <i>Recall</i> dari <i>query</i> judul dokumen	47
4.22 Rata-rata <i>Precision</i> dan <i>Recall</i> dari <i>query</i> isi dokumen.....	47
4.23 Rata-rata <i>Precision</i> dan <i>Recall</i> dari <i>query key word</i> dokumen.....	48
4.24 Tahapan Implementasi	50

DAFTAR PUSTAKA

- [Agusta 2009] Agusta, Ledy. *Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk stemming Dokumen Teks Bahasa Indonesia*, Bali. Konferensi Nasional Sistem dan Informatika, KNS&I09-036. 2009.
- [Baeza 1999] Baeza R.Y. dan Neto R., 1999. Modern Information Retrieval, Addison Wesley-Pearson international edition, Boston. USA.
- [Doyle 1975] Doyle, Lauren B. *Information retrieval and processing*. Los Angeles: Melville Publishing, 1975.
- [Harjono 2005] Harjono, K. D. *Perluasan Vektor Pada Metode Search Vector Space*, Integral Vol. 10 No.2, Juli 2005 Jurusan Ilmu Komputer, Universitas Katolik Parahyangan, Bandung
- [Hasugian 2006] Hasugian, Jonner. 2006. *Penggunaan Bahasa Alamiah dan Kosa Kata Terkontrol dalam Sistem Temu Kembali Informasi Berbasis Teks. Dalam Jurnal Pustaka: Jurnal Studi Perpustakaan dan Informasi*, Vol.2, No.2, Desember 2006. USU Press.
- [Hemard 2000] Hémard, D. et.al., “From Access to Acceptability: Exploiting the Web to Design a New CALL Environment. Computer Assisted Language Learning”, Vol. 13, No. 2, 103–118, 2000
- [Kappel 2006] Kappel, G. B, *Web Enginerring The Discipline of Systematic Development of Web Applications*, West Sussex England: John Wiley & Sons Ltd, 2006
- [Karyono 2012] Karyono, Giat dan Utomo, S. F. Temu balik informasi pada dokumen teks berbahasa Indonesia dengan metode vector space retrieval model, Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012.
- [Lancaster 1979] Lancaster, F.W. *Information retrieval system : characteristic, testing and evaluation*. Second edition. New York: Willey: 1979.
- [Nadirman 2006] Nadirman, Firnas,. *Sistem Temu-Kembali Informasi Dengan Metode Vector Space Model Pada Pencarian File Dokumen Berbasis Teks*, Skripsi. Universitas Gajah Mada, 2006.
- [Ning 2004] Ning, Liu et.al., “Learning Similarity Measures in Non-orthogonal Space”, CIKM’04, November 8-13, 2004, Washington D.C., U.S.A.