

**MESIN PENCARI DOKUMEN TEKS BAHASA INDONESIA, STUDI
EFEKTIFITAS PENCARIAN PADA VECTOR SPACE MODEL,
ALGORITMA STEMMING PORTER, PEMBOBOTAN FREKUENSI
TERM SERTA FUNGSI JACCARD**

TESIS



Oleh :
ZUMHUR ALAMIN
1311600900

PROGRAM STUDI : MAGISTER ILMU KOMPUTER (M. KOM)
PROGRAM PASCASARJANA
UNIVERSITAS BUDI LUHUR

JAKARTA
2015

**MESIN PENCARI DOKUMEN TEKS BAHASA INDONESIA, STUDI
EFEKTIFITAS PENCARIAN PADA VECTOR SPACE MODEL,
ALGORITMA STEMMING PORTER, PEMBOBOTAN FREKUENSI
TERM SERTA FUNGSI JACCARD**

TESIS

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



Oleh :
ZUMHUR ALAMIN
1311600900

PROGRAM STUDI : MAGISTER ILMU KOMPUTER (M. KOM)
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 Mahasiswa : Zumhur Alamin
Nomor Induk Mahasiswa : 1311600900
Program Studi : Magister Ilmu Komputer
Konsentrasi : Teknologi Sistem Informasi
Fakultas/Program : Pascasarjana

Menyatakan bahwa Tesis yang berjudul :

Mesin Pencari Dokumen Teks Bahasa Indonesia, Studi Efektifitas Pencarian Pada *Vector Space Model*, *Algoritma Stemming Porter*, Pembobotan Frekuensi *Term* Serta Fungsi Jaccard

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, 10 Agustus 2015

METERAI
TEMPEL
6000
ENAM RIBU RUPIAH
(Zumhur Alamin)



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

LEMBAR PENGESAHAN TESIS

Nama Mahasiswa : Zumhur Alamin
Nomor Induk Mahasiswa : 1311600900
Konsentrasi : Teknologi Sistem Informasi
Jenjang Studi : Pascasarjana
Judul Tesis : Mesin Pencari Dokumen Teks Bahasa Indonesia,
Studi Efektifitas Pencarian Pada *Vector Space Model*,
Algoritma Stemming Porter, Pembobotan Frekuensi
Term Serta Fungsi Jaccard

Jakarta, 10 Agustus 2015
Tim Penguji:

Tanda Tangan:

Ketua,
(Prof. Dr. Moedjiono, M.Sc)

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

Pembimbing Utama,
(Dr. Krisna Adiyarta M.)

Ketua Program Studi

Dr., Ir. Nazori Az, M.T

ABSTRAK

Sebuah informasi berupa data teks tersimpan pada berbagai macam dokumen salah satunya adalah dokumen pdf. Tiap dokumen terdiri dari beberapa kalimat yang merupakan sekumpulan kata. Dalam sebuah proses pencarian dokumen, terjadi perbandingan kemiripan antara kata kunci (*keyword*) dan kata yang terdapat dalam sebuah dokumen. Untuk dapat menemukan dokumen sesuai kebutuhan, diperlukan sebuah teknik pencarian yang tepat. Dalam proses pencarian dokumen sesuai kebutuhan dapat memanfaatkan sistem temu kembali informasi (*information retrieval system*). Salah satu bentuk aplikasi dari *information retrieval system* (IRS) adalah mesin pencari (*search engine*). Penelitian menggunakan pembobotan *vector space model* menekankan kepada teknik pembobotan berdasarkan kata-kata (*term*). Elemen pembobotan *vector space model* yaitu pembobotan *term* terdapat pada dokumen dan *query* yang dinyatakan dalam bentuk vektor. Pembobotan yang didasarkan pada *term* selalu dikaitkan dengan teknik *stemming* untuk mendapatkan bentuk kata dasar dari *term* yang bersangkutan. Pada proses pencarian, imbuhan merupakan bagian dari informasi yang perlu dihilangkan untuk mencapai efektifitas hasil pencarian. Beberapa pengembangan teknik *stemming* kata seperti penggunaan aturan imbuhan bahasa Indonesia dan penggunaan kamus kata dasar. Selain itu, perlu dilakukan proses menghilangkan *term* yang tidak bermakna dari informasi yang merupakan proses filtrasi dikenal dengan istilah *stopword removal*. Penelitian ini membahas model mesin pencari yang mengimplementasikan *algoritma stemming* Porter pada proses *stemming* kata berbahasa Indonesia dan penerapan *vector space model* dengan skema pembobotan berdasarkan frekuensi kemuculan kata (*term*) serta fungsi kesamaan Jaccard. Hasil dari penelitian ini, diperoleh teknik pencarian yang menunjukkan tingkat efektifitas tinggi dengan teknik pengukuran *precision* dan *recall* dari hasil pencarian dokumen pdf berbahasa Indonesia pada proses pencarian.

Kata kunci : *Information Retrieval System, Vector Space Model, Stemming Porter, kesamaan Jaccard, Precision, Recall.*

ABSTRACT

The data text stored on the various kinds of documents, one of them is a pdf document. Each document consists of several sentences which is a set of words. In a document searching process, there are similarity comparisons between keywords (keywords) and words contained in a document. To be able to find the document, needed a proper searching techniques. Information retrieval (information retrieval system) is used in the document searching process. One of the applications of information retrieval system (IRS) is a search engine (search engine). Research using the weighting vector space model (vector space model) which emphasizes the weighting technique based on words (terms). Weighting element vector space model is weighted terms contained in the documents and queries are expressed in vector form. The weightings are based on the term has always been connected with the stemming technique to get the basis word forms of the relevant term. In the search process, affix is a part of a information that needs to be eliminated to achieve the effectiveness of the search results. Some development techniques such as the use of the word Indonesian's rule of stemming affix and basic word dictionary. In addition, there is should be a process to eliminate non-significant terms of the information that a filtration process known as stopword removal. This study discusses the search engine model that applying the Porter stemming algorithm in the process of stemming words in Indonesian language and applying of vector space model with a weighting scheme based on appear of word frequency (term) and Jaccard similarity function. Results from this study, obtained a search technique that shows a high level of effectiveness measurement techniques precision and recall of search results pdf document in Indonesian language in the search process.

Keywords: Information Retrieval System, Vector Space Model, Porter Stemming, Jaccard Similarity, Precision, Recall.

KATA PENGANTAR

Puja dan puji syukur penulis panjatkan kepada Allah SWT, shalawat serta salam semoga tetap tercurahkan kepada nabi Muhammad SAW. Berkat rahmat, hidayah serta bimbingan Allah SWT penulis dapat menyelesaikan Tesis yang berjudul *Mesin Pencari Dokumen Teks Bahasa Indonesia, Studi Efektifitas Pencarian Pada Vector Space Model, Algoritma Stemming Porter, Pembobotan Frekuensi Term Serta Fungsi Jaccard*. Tujuan dari penulisan tesis ini adalah sebagai salah satu syarat untuk penyusunan tesis pada Program Studi Magister Ilmu Komputer, Program Pascasarjan Universitas Budi Luhur Jakarta.

Pada kesempatan ini penulis menyampaikan rasa hormat dan ucapan terima kasih yang sebesar besarnya kepada :

1. Bapak H. Ir. Zainuddin dan Ibu Hj. Nurjanah, S.Sos selaku orang tua yang selalu memberikan kasih sayang dan doanya tanpa henti kepada penulis, serta Kakak saya (Nurmayni) dan Adik-adik saya (Nurlatifah, Awaluddin, Ihya, dan Hakim) yang selalu memberikan dukungan dan semangat serta doanya.
2. Bapak Prof. Dr. Moedjiono, M.Sc. selaku Direktur Program Pascasarjana Universitas Budi Luhur Jakarta yang telah memberikan pengetahuan dalam penelitian ilmiah dan Bapak Dr. Krisna, M.Sc., selaku dosen pembimbing tesis yang telah membimbing dan memotivasi penulis dalam menyelesaikan tesis ini.
3. Randitha Missouri, S.Pd yang selalu memberikan semangat serta dukungan hingga penulis dapat menyelesaikan penyusunan tesis ini.
4. Rekan-rekan mahasiswa MKOM Universitas Budi Luhur terima kasih atas kebersamaan, kerja keras dan dukungannya.
5. Bapak dan Ibu Dosen serta para staf di Program Studi Magister Ilmu Komputer Universitas Budi Luhur yang telah dengan sabar memberikan ilmu pengetahuan, pencerahan, dan bimbingan dalam belajar.

Dalam penyusunan Tesis ini tidak lepas dari kekurangan dan kekhilafan yang disebabkan oleh keterbatasan dari penulis. Harapan penulis dalam penyusunan Tesis ini dapat memberikan manfaat serta tambahan ilmu pengetahuan bagi pembaca dan penulis sendiri khususnya. Kesempurnaan hanyalah milik Allah SWT, oleh karena itu saran dan kritik bersifat membangun senantiasa penulis harapkan.

Jakarta, 1 Juli 2015

Zumhur Alamin



DAFTAR ISI

Halaman

LEMBAR PERNYATAAN.....	i
LEMBAR PENGESAHAN TESIS.....	ii
ABSTRAK.....	iii
ABSTRACT.....	iv
KATA PENGANTAR	v
DAFTAR ISI.....	vii
DAFTAR GAMBAR	ix
DAFTAR TABEL.....	x
DAFTAR LAMPIRAN.....	xi
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Masalah Penelitian	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.....	5
1.4 Tata Urut Penulisan.....	6
1.5 Daftar Pengertian.....	7
BAB II LANDASAN TEORI DAN PEMIKIRAN.....	9
2.1 Tinjauan Pustaka	9
2.1.1 <i>Information Retrieval</i>	9
2.1.2 <i>Information Retrieval System (IRS)</i>	10
2.1.3 Morfologi Bahasa Indonesia.....	15
2.1.4 <i>Vector Space Model</i>	24
2.1.5 Pengindeksan Teks	27
2.1.6 Tokenisasi.....	28
2.1.7 Filtrasi.....	29
2.1.8 Stemming.....	30
2.1.9 Koefisien Jaccard	36
2.1.10 Pembobotan Kata (<i>Term Weighting</i>).....	37
2.1.11 Pengukuran Efektifitas Pencarian.....	39
2.2 Tinjauan Studi	41
2.3 Tinjauan Obyek Penelitian	46
2.3.1 Model Mesin Pencari.....	46
2.3.2 Koleksi Pengujian.....	46
2.4 Kerangka Konsep	46
2.5 Hipotesis.....	49
BAB III METODOLOGI DAN RANCANGAN PENELITIAN	50
3.1 Metode Penelitian.....	50
3.2 Metode Pemilihan Sampel	51
3.3 Instrumentasi	52
3.4 Teknik Analisis dan Prototipe Model.....	52
3.4.1 Teknik Analisis.....	52
3.4.2 Prototipe Model	53

3.5	Teknik Pengindeksan Teks.....	55
3.6	Teknik Pembobotan Kata.....	56
3.7	Teknik Pengujian.....	58
3.8	Teknik Pengukuran Efektifitas.....	59
3.9	Langkah-Langkah Penelitian.....	59
3.10	Jadwal Penelitian.....	62
BAB IV	PEMBAHASAN HASIL PENELITIAN.....	64
4.1	Analisis Koleksi Pengujian	64
4.2	Analisis Model Mesin Pencari	67
4.3	Pengujian Efektifitas Hasil Pencarian	71
4.3.1	Bentuk Pengujian Efektifitas Hasil Pencarian.....	71
4.3.2	Rancangan Pengujian Efektifitas Hasil Pencarian.....	71
4.4	Hasil Pengujian Efektifitas.....	73
4.4.1	Hasil Pengujian Query Q1 dengan Teknik Pencarian TP1.....	73
4.4.2	Hasil Pengujian Query Q2 dengan Teknik Pencarian TP1.....	74
4.4.3	Hasil Pengujian Query Q3 dengan Teknik Pencarian TP1.....	75
4.4.4	Hasil Pengujian Query Q4 dengan Teknik Pencarian TP1.....	76
4.4.5	Hasil Pengujian Query Q1 dengan Teknik Pencarian TP2.....	77
4.4.6	Hasil Pengujian Query Q2 dengan Teknik Pencarian TP2.....	77
4.4.7	Hasil Pengujian Query Q3 dengan Teknik Pencarian TP2.....	78
4.4.8	Hasil Pengujian Query Q4 dengan Teknik Pencarian TP2.....	79
4.4.9	Hasil Pengujian Dokumen DOK1 dengan Teknik Pencarian TP3	79
4.4.10	Hasil Pengujian Dokumen DOK2 dengan Teknik Pencarian TP3	81
4.4.11	Hasil Pengujian Dokumen DOK3 dengan Teknik Pencarian TP3	82
4.4.12	Hasil Pengujian Dokumen DOK4 dengan Teknik Pencarian TP3	84
4.5	Analisis Hasil Pengujian Efektifitas.....	85
4.6	Implikasi Penelitian.....	90
4.6.1	Aspek Sistem.....	90
4.6.2	Aspek Manajerial	91
4.6.3	Aspek Penelitian Lanjut	92
BAB V	PENUTUP	93
5.1	Kesimpulan	93
5.2	Saran.....	94
	DAFTAR PUSTAKA	95
	LAMPIRAN-LAMPIRAN	98
	RIWAYAT HIDUP SINGKAT	114

DAFTAR GAMBAR

Gambar	Halaman
II-1 <i>Flowcart Information Retrieval System (IRS)</i>	12
II-2 Beberapa Bagian <i>Information Retrieval System (IRS)</i>	13
II-3 Beberapa Dasar Kata Bahasa Indonesia	22
II-4 Komponen <i>Vector Space Model</i>	24
II-5 Representasi dokumen dan <i>query</i> pada ruang vektor.....	26
II-6 Representasi matriks kata dalam dokumen	26
II-7 Tahapan Pengideksan Teks	28
II-8 Jenis-Jenis Metode Stemming	31
II-9 Desain Porter Stemmer untuk Bahasa Indonesia	32
II-10 Koefisien Jaccard	37
II-11 <i>Recall</i> dan <i>Precision</i> pada contoh hail IRS	40
II-12 Kerangka Konsep	48
III-1 Contoh Proses Tokenisasi <i>Query</i>	54
III-2 Contoh Hasil Tokenisasi <i>Query</i> dan Dokumen.....	56
III-3 Contoh Hasil <i>Stemming</i>	57
III-4 Langkah-Langkah Penelitian.....	61
IV-1 <i>Flowchart</i> Pembuatan Koleksi Dokumen	67
IV-2 <i>Flowchart</i> Pencarian Berdasarkan <i>Query</i>	68
IV-3 <i>Flowchart</i> Pencarian Berdasarkan Dokumen.....	69
IV-4 Grafik <i>Precision</i> dan <i>Recall</i> Hasil Pencarian Membandingkan Query dengan Isi Koleksi Pengujian	86
IV-5 Grafik <i>Precision</i> dan <i>Recall</i> Hasil Pencarian Membandingkan Query dengan Judul Koleksi Pengujian	87
IV-6 Grafik <i>Precision</i> dan <i>Recall</i> Hasil Pencarian Membandingkan Dokumen dengan Isi Koleksi Pengujian	88
IV-7 Grafik <i>Average Precision</i> dan <i>Recall</i> Hasil Pencarian dari 3 Teknik Pencarian.....	89

DAFTAR TABEL

Tabel	Halaman
I-1	Daftar Pengertian 7
II-1	Komponen Utama <i>Information Retrieval System</i> (IRS) 14
II-2	Definisi morf, morfem dan alomorf 18
II-3	Pola Kanonik Otto von Dempwolff 20
II-4	Contoh Penggabungan dari Tiga Struktur Suku Kata 21
II-5	Contoh Imbuhan (Afiks) 23
II-6	Proses Porter Stemmer Bahasa Indonesia 33
II-7	Tabel Akhiran (Kata Sandang) 33
II-8	Tabel Akhiran (Kata Ganti Kepunyaan) 34
II-9	Tabel Awalan Pertama pada Tahap Ketiga <i>Stemming</i> 34
II-10	Tabel Awalan Kedua pada Tahap Keempat <i>Stemming</i> 35
II-11	Tabel akhiran pada Tahap Kelima <i>Stemming</i> 35
II-12	Tabel Pasangan Awalan dan Akhiran Tidak Sah 36
II-13	Ringkasan Tinjauan Studi 42
III-1	Contoh <i>Query</i> dan Dokumen Perhitungan Bobot 56
III-2	Pembobotan Kata <i>Query</i> dan Dokumen 57
III-3	Hasil Perangkingan Dokumen 58
III-4	Jadwal Penelitian 62
IV-1	Daftar Koleksi Pengujian 64
IV-2	Jumlah Koleksi Pengujian Berdasarkan Sumber 66
IV-3	Jumlah Koleksi Pengujian Berdasarkan Kategori 66
IV-4	Teknik Pencarian 72
IV-5	<i>Query</i> yang digunakan dalam pengujian 72
IV-6	Dokumen yang digunakan dalam pengujian 72
IV-7	Hasil Pengujian <i>Query</i> Q1 dengan Teknik Pencarian TP1 73
IV-8	Hasil Pengujian <i>Query</i> Q2 dengan Teknik Pencarian TP1 74
IV-9	Hasil Pengujian <i>Query</i> Q3 dengan Teknik Pencarian TP1 75
IV-10	Hasil Pengujian <i>Query</i> Q4 dengan Teknik Pencarian TP1 76
IV-11	Hasil Pengujian <i>Query</i> Q1 dengan Teknik Pencarian TP2 77
IV-12	Hasil Pengujian <i>Query</i> Q2 dengan Teknik Pencarian TP2 78
IV-13	Hasil Pengujian <i>Query</i> Q3 dengan Teknik Pencarian TP2 78
IV-14	Hasil Pengujian <i>Query</i> Q4 dengan Teknik Pencarian TP2 79
IV-15	Hasil Pengujian Dokumen DOK1 dengan Teknik Pencarian TP3 80
IV-16	Hasil Pengujian Dokumen DOK2 dengan Teknik Pencarian TP3 81
IV-17	Hasil Pengujian Dokumen DOK3 dengan Teknik Pencarian TP3 82
IV-18	Hasil Pengujian Dokumen DOK4 dengan Teknik Pencarian TP3 84
IV-19	<i>Precision Recall</i> dari Teknik Pencarian TP1 86
IV-20	<i>Precision Recall</i> dari Teknik Pencarian TP2 86
IV-21	<i>Precision Recall</i> dari Teknik Pencarian TP3 87
IV-22	<i>Average Precision Recall</i> dari Teknik Pencarian (TP1, TP2, TP3) 88

DAFTAR LAMPIRAN

Lampiran	Halaman
1. Daftar Stopword Tala	98
2. Tampilan Layar Hasil Pencarian TP1 (<i>keyword</i> Q1).....	103
3. Tampilan Layar Hasil Pencarian TP1 (<i>keyword</i> Q2).....	104
4. Tampilan Layar Hasil Pencarian TP1 (<i>keyword</i> Q3).....	105
5. Tampilan Layar Hasil Pencarian TP1 (<i>keyword</i> Q4).....	106
6. Tampilan Layar Hasil Pencarian TP2 (<i>keyword</i> Q1).....	107
7. Tampilan Layar Hasil Pencarian TP2 (<i>keyword</i> Q2).....	108
8. Tampilan Layar Hasil Pencarian TP2 (<i>keyword</i> Q3).....	109
9. Tampilan Layar Hasil Pencarian TP2 (<i>keyword</i> Q4).....	109
10. Tampilan Layar Hasil Pencarian TP3 (<i>keyword</i> DOK1).....	110
11. Tampilan Layar Hasil Pencarian TP3 (<i>keyword</i> DOK2).....	111
12. Tampilan Layar Hasil Pencarian TP3 (<i>keyword</i> DOK3).....	112
13. Tampilan Layar Hasil Pencarian TP3 (<i>keyword</i> DOK4).....	113

DAFTAR PUSTAKA

- [Alwi 2013] Alwi, M. Y., *Efektifitas Sistem Temu Kembali Informasi Dalam Meningkatkan Pemanfaatan Koleksi Studi Kasus Pemustaka Pada Perpustakaan STMIK Potensi Utama Medan*, Medan: Usu Perss, 2013.
- [Asian 2007] Asian, Jelita., *Effective Techniques for Indonesian Text Retrieval*, Ph.D. diss., School of Computer Science and Information Technology, RMIT University, Australia, 2007.
<https://researchbank.rmit.edu.au/eserv/rmit:6312/Asian.pdf>. (Diakses 26 Mei 2015).
- [Bauer 1988] Bauer, Laurie., *Introducing Linguistic Morphology*, Edinburg University Press, 1988.
- [Christioko 2012] Christioko, Bernadus V., *Implementasi Sistem Temu Kembali Informasi (Studi Kasus: Dokumen Teks Berbahasa Indonesia)*, Jurnal Transformatika, Vol. 10, No. 1, (Juli, 2012): 1-10.
- [Guritno 2011] Guritno, Suryo., Sudaryono dan Rahardja, U., *Theory and Application of IT Research Metodologi Penelitian Teknologi Informasi*. Yogyakarta: CV. Andi Offset, 2011.
- [Hamzah 2009] Hamzah, Amir., *Temu Kembali Informasi Berbasis Kluster Untuk Sistem Temu Balik Informasi Teks Bahasa Indonesia*, Jurnal Teknologi, Vol. 2, (Juni, 2009): 1-7.
- [Hasugian 2009] Hasugian, J., *Dasar - Dasar Ilmu Perpustakaan dan Informasi*. Medan: USU Press, 2009.
- [Hersh 2008] Hersh, William., *Information Retrieval: A Health and Biomedical Perspective*, 3rd ed., New York: Springer Science & Business Media, 2008.
- [Karyono 2012] Karyono, Giat dan Utomo, Fandy Setyo., *Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia Dengan Metode Vector Space Retrieval Model*, Seminar Nasional Teknologi Informasi & Komunikasi Terapan (Semantik), Semarang, 2012.
- [Kent 1971] Kent, A., *Information Analysis and Retrieval*, 3rd Edition, Becker and Heys, New York, 1971.
- [Keraf 1982] Keraf, Gorys., *Eksposisi dan Deskripsi*, Ende-Flores:Nusa Indah, 1982.