

**MODDEL MESIN PENCARI DOKUMEN BAHASA INDONESIA, STUDI
EFEKTIFITAS PADA VECTOR SPACE MODEL ALGORITMA
STEMMING POTER PEMBOBOTAN FREKUENSI TERM
BERBANDING FREKUENSI TERM DALAM PENCARIAN
DAN FUNGSI KESAMAAN COSINE**

TESIS



Oleh:

LASARUS PELIPUS MALESE

1311601692

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

JAKARTA

2015

**MODDEL MESIN PENCARI DOKUMEN BAHASA INDONESIA, STUDI
EFEKTIFITAS PADA VECTOR SPACE MODEL ALGORITMA
STEMMING POTER PEMBOBOTAN FREKUENSI TERM
BERBANDING FREKUENSI TERM DALAM PENCARIAN
DAN FUNGSI KESAMAAN COSINE**

TESIS

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



Oleh:

LASARUS PELIPUS MALESE

1311601692

PROGRAM STUDI : MAGISTER ILMU KOMPUTER (MKOM)

FAKULTAS PASCASARJANA

UNIVERSITAS BUDI LUHUR

JAKARTA

2015



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

LEMBAR PERNYATAAN

Saya yang bertanda tangan dibawah ini :

Nama Mahasiswa : LASARUS PELIPUS MALESE
Nomor Induk Mahasiswa : 1311601692
Program Studi : MAGISTER ILMU KOMPUTER
Konsentrasi : TEKNOLOGI SISTEM INFORMASI
Jenjang Studi : STRATA 2 (SE)

Menyatakan bahwa Tesis yang berjudul :

Model Mesin Pencari Dokumen Bahasa Indonesia Studi
Efektivitas pada Vector Space Model Algoritma Stemming
Pola Pembobotan Frekuensi Term Berbanding Frekuensi
Term Dalam Pencarian dan Fungsi Kesamaan Cosine

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 etik 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.

Jakarta, 08 Agustus 2015



(Lasarus Pelipus Malese)



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

LEMBAR PENGESAHAN

Nama Mahasiswa : Lasarus Pelipus Malese
Nomor Induk Mahasiswa : 1311601692
Konsentrasi : Teknologi Sistem Informasi
Judul Proposal Tesis : Moddel Mesin Pencari Dokumen Bahasa Indonesia,
Studi Efektifitas Pada Vector Space Model Algoritma
Stemming Poter Pembobotan Frekuensi Term
Berbanding Frekuensi Term Dalam Pencarian Dan
Fungsi Kesamaan Cosine

Jakarta, 8 Agustus 2015

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

Seiring dengan semakin berharganya nilai sebuah informasi dan semakin banyak sumber-sumber informasi, maka semakin banyak pula kebutuhan manusia untuk dapat menemukan informasi yang sesuai keperluannya dengan cepat. *Information Retrieval* (Perolehan Informasi) merupakan suatu pencarian informasi(biasanya berupa dokumen) yang didasarkan pada suatu *query* (*inputan user*) yang diharapkan dapat memenuhi keinginan *user* dari dokumen yang ada. Dua aspek penting dalam konsep *Information Retrieval* yang diterapkan dalam sebuah perancangan mesin pencari yaitu representasi dari informasi dan pengukuran yang akan mengukur nilai kesamaan antara dua obyek. Informasi yang dapat direpresentasikan menjadi sebuah obyek yang direpresentasikan dalam berbagai bentuk dan model (*heterogeneous*). Keadaan ini mengakibatkan bahwa pencarian untuk suatu obyek informasi yang diinginkan akan dapat dipetakan kepada beberapa obyek informati yang dinilai relevan. Relevansi dua informasi di-ukur dari keberadaan kata kunci (*keyword*) dan bobotnya. Konsekuensi logis atas keadaan ini adalah bahwa dalam melakukan pencarian obyek yang diinginkan, ditemukan terdapat ketidakpastian (*uncertainly*) terhadap penggunaan *keyword* pada *query* oleh pengguna dengan keberadaan *keyword* pada dokumen. Pada penelitian ini akan difokuskan pada studi efektifitas vektor model dengan menggunakan algoritma *steming poter* untuk membentuk kata-kata menjadi sebuah kata baku serta pembobotan frekuwensi term untuk menentukan tingkat kepentingan setiap indeks *term* dalam sebuah dokumen dan fungsi kesamaan cosine dalam mengukur kemiripan queri dengan dokumen. Hasil pengujian nilai kualitas rata-rata *precesion* dan *recall* bahwa untuk semua bentuk *query* pencarian dokumen mempunyai nilai *precesion* 100% artinya baik pencarian berdasarkan bentuk *query* isi, judul dan dokumen mempunyai *precesion* yang baik sedangkan hasil pengujian *recall* menunjukkan hasil yang berbeda, dimana pencarian berdasarkan bentuk queri isi dokumen menempati *recall* tertinggi sebesar 90%, sedangkan pencarian berdasarkan bentuk queri judul dokumen dengan nilai *recall* 70% dan pencarian berdasarkan bentuk *query* dokumen menempati posisi terendah dengan *recall* 33%.

Kata Kunci : *query*, *Information Retrieval*, studi efektifitas vektor model, algoritma *steming poter*, fungsi kesamaan cosine.

ABSTRACT

Along with the precious value of an information and more sources of information, the more the human need to find information as appropriate quickly. Information Retrieval (Information Retrieval) is an information retrieval (usually in the form of a document), which is based on a query (user input) that is expected to meet the wishes of the user of an existing document. Two important aspects of the concept of Information Retrieval applied in a design of a search engine that is representasi of information and measurement will measure the value of the similarity between two objects. The information can be represented into an object that is represented in various forms and models (heterogeneous). This situation resulted in a search for an object that the desired information will be mapped to multiple objects informati considered relevant. Relevance two information-measuring of the existence of keywords (keywords) and weight. Logical consequence of this situation is that in searching the desired object, found there is uncertainty (uncertainly) on the use of keywords in the query by the user to the presence of keywords in the document. This research will be focused on the study of the effectiveness of vector models using algorithms stemming poter to form words into a standard word and weighting frekuwensi term to determine the level of importance of each index term in a document and cosine similarity function to measure the similarity of the query with the document. Results of testing the value of the average quality precesion and recall that for all forms of search query documents have value precesion 100% means that both search based query form contents, title and documents have precesion good while test results recall showed different results, which search by shape query the contents of the document occupy the highest recall by 90%, while a search based on the query form document title with recall value of 70% and a search based on the query form document occupies the lowest position with 33% recall.

Keywords: query, Information Retrieval, study the effectiveness of vector models, algorithms stemming poter, cosine similarity function.

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yesus Kristus, karena berkat dan hikmat-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul Model Mesin Pencari Dokumen Bahasa Indonesia, Studi Efektifitas Pada Vector Space Model Algoritma Stemming Poter Pembobotan Frekuensi Term Berbanding Frekuensi Term Dalam Pencarian Dan Fungsi Kesamaan Cosine. Tujuan dari penyusunan tesis ini dalam rangka memenuhi salah satu syarat untuk menyelesaikan studi Strata-2 pada program studi Magister Ilmu Komputer (MKOM) Fakultas Pascasarjana Universitas Budi Luhur, Jakarta.

Penulis sudah berusaha secara maksimal untuk meneliti dan menyusun tesis ini dan bila masih terdapat kekurangan mohon dimaklumi. Penulis juga menyadari bahwa tanpa kesempatan, bimbingan, bantuan, serta dukungan, tesis ini tidak berarti sama sekali. Maka pada kesempatan ini, penulis ingin menyampaikan terima kasih yang sebesar-besarnya terutama kepada :

1. Tuhan Yesus Kristus yang telah memberikan kesehatan serta kesempatan untuk menyelesaikan proposal tesis ini.
2. Isteri dan anak tercinta yang senantiasa menjadi inspirasi, memberi doa, dan dukungan selama berlangsungnya masa perkuliahan hingga memasuki masa penyelesaian perkuliahan;
3. Orang tua, keluarga dan rekan-rekan seperjuangan yang telah banyak memberikan doa serta motivasi untuk menyelesaikan tesis ini.
4. Dr. Krisna Adiyarta, M.Sc selaku dosen 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.
5. Dr. Ir. Nazori AZ, MT selaku Kaprodi Magister Komputer yang telah memberikan kepercayaan dan kesempatan kepada Penulis untuk menyusun proposal tesis ini.

6. Prof. Moedjiono, M.Sc. selaku Direktur Program Pascasarjana sekaligus selaku dosen yang telah banyak memberikan pengarahan dan saran mengenai tata cara penulisan proposal tesis ini.
7. Para Staf-staf dosen Pascasarjana Magister Komputer Universitas Budiluhur yang telah banyak memberikan ilmunya kepada Penulis, sehingga dapat menyelesaikan proposal tesis ini.
8. Terima kasih juga kepada teman-teman kuliah MKOM Angkatan 2015 yang telah memberikan semangat untuk menyelesaikan tesis ini.

Akhir kata, penulis berharap agar proposal tesis ini dapat bermanfaat bagi pihak yang memerlukannya.

Jakarta, 08 Agustus 2015



(Lasarus Pelipus Malese)

HALAMAN PERSEMBAHAN

“Untuk segala sesuatu ada waktunya dan Tuhan membuat segala sesuatu indah pada waktunya” (Pengkhotbah 3:11a)

*Kupersembahkan untuk istri (Eunike) dan anak (Jethro) tercinta,
keluarga besar ku
(Malese - Molebila)*



DAFTAR ISI

	Halaman
LEMBAR JUDUL	i
LEMBAR PERNYATAAN	ii
LEMBAR PENGESAHAN	iii
ABSTRAK	iv
ABSTRAC	v
KATA PENGANTAR	vi
HALAMAN PERSEMBAHAN	vii
DAFTAR ISI	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL	xiv
DAFTAR LAMPIRAN	xiv
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	7
1.5. Daftar Pengertian	8
BAB II LANDASAN TEORI DAN KERANGKA KONSEP	10
2.1. Tinjauan Pustaka	10
2.1.1. <i>Search Engine</i>	10
2.1.2. PHP	11
2.1.3. <i>Information Retrieval (IR)</i>	11
2.1.3.1. Tujuan dan Fungsi <i>Information Retrieval System</i> ...	14
2.1.3.2. Arsitektural <i>Information Retrieval System</i>	16
2.1.4. Morfologi Bahasa Indonesia	17
2.1.4.1. Morfologi	17
2.1.4.2. Proses Morfologi	18
2.1.5. Stemming	21
2.1.6. Definisi <i>Stemming</i>	22
2.1.7. Algoritma <i>Stemming Porter</i>	23
2.1.8. Model Representasi Dokumen	25
2.1.9. <i>Tokenisasi</i>	28

2.1.9.1. <i>Stopword</i>	28
2.1.9.2. <i>Proses Stemming</i>	29
2.1.9.3. <i>Pembobotan (Indeksing)</i>	30
2.1.9.4. <i>Kesamaan (Similarity)</i>	30
2.2. Tinjauan Studi.....	30
2.3. Kerangka Konsep.....	34
2.4. Hipotesis	35
BAB III METODOLOGI DAN RANCANGAN DESAIN PENELITIAN.....	36
3.1. Metode Penelitian	36
3.2. Sampling/Metode Pemilihan Sampel	37
3.3. Instrumentasi Penelitian	38
3.4. Teknik Analisis Prototipe Model	38
3.4.1. Teknik Analisis	38
3.4.2. <i>Prototipe Model</i>	39
3.4.3. Teknik Pengindeksan	43
3.4.4. Teknik Pengujian	45
3.4.5. Teknik Pengukuran Efektifitas	46
3.4.5.1. <i>Recall & Precision</i>	46
3.5. Langkah-langkah Penelitian	48
3.6. Jadwal Penelitian	51
BAB IV PEMBAHASAN HASIL PENELITIAN	53
4.1. Pengelompokan Data	53
4.2. Prototipe Model Mesin Pencari Dokumen Bahasa Indonesia	55
4.3. Hasil Pengukuran Berdasarkan Isi Dokumen	58
4.3.1. Perhitungan Nilai Rata-rata Isi Dokumen	70
4.4. Hasil Pengukuran Berdasarkan Judul Dokumen	71
4.4.1. Perhitungan Nilai Rata-rata Judul Dokumen.....	80
4.5. Hasil Pengukuran Berdasarkan Dokumen	81
4.5.1. Perhitungan Nilai Rata-rata Dokumen	101
4.6. Performa Waktu Tanggap	102
4.7. Hasil dan Pembahasan	103
BAB V PENUTUP.....	104
5.1. Kesimpulan	104
5.2. Saran	104
DAFTAR PUSTAKA	106
LAMPIRAN-LAMPIRAN	109
DAFTAR RIWAYAT HIDUP	126

DAFTAR GAMBAR

Gambar	Halaman
II-1 Ilustrasi <i>information retrieval system</i>	12
II-2 Skema <i>information retrieval (IR) system</i>	13
II-3 Proses Sistem Temu-Kembali Informasi	16
II-4 Representasi dokumen dan <i>vektor</i> pada ruang <i>vektor</i>	26
II-5 Representasi matriks kata-dokumen	27
II-6 Desain dari Porter Stemmer untuk Bahasa Indonesia	29
II-7 Kerangka Pemikiran	34
III-1 Contoh Proses <i>Tokenisasi Query</i>	39
III-2 Tahapan Pengideksan Teks	44
III-3 Relasi antara <i>relevant</i> dan <i>retrieved</i> dokumen	46
III-4 Trade-off antara <i>precision</i> dan <i>recall</i>	48
III-5 Langkah-langkah Penelitian	50
IV-1 Tampilan Halaman <i>Home</i>	54
IV-2 Tampilan Search By Document	55
IV-3 Tampilan Halaman Menu Login	55
IV-4 Tampilan Halaman <i>Dashboard</i>	56
IV-5 Tampilan Halaman <i>Setting Akun</i>	56
IV-6 Tampilan Halaman Tambah <i>User</i>	61
IV-7 Tampilan Halaman <i>List Data User</i>	57
IV-8 Tampilan Halaman <i>Upload Document PDF Admin</i>	58

DAFTAR TABEL

Tabel	Halaman
II-1 Aturan Untuk Inflectional Particle	23
II-2 Aturan Untuk Inflectional Possesive Pronoun	24
II-3 Aturan Untuk First Order Derivational Prefix	24
II-4 Aturan Untuk Second order Derivational Previx	24
II-5 Aturan Untuk Derivational Suffix	25
II-6 Ringkasan tinjauan studi	32
III-1 Komponen Penghapusan <i>Particle</i> dan <i>Possesive Pronoun</i>	40
III-2 Komponen Penghapusan Awalan Pertama	41
III-3 Komponen Penghapusan Awalan Kedua	41
III-4 Komponen Penghapusan Akhiran	41
III-5 Jadwal Penelitian	51
IV-1 Proses perhitungan bobot pada query Q1 dan dokumen berdasarkan isi	60
IV-2 Similarity dokumen dari query Q1 dan dokumen berdasarkan isi	61
IV-3 Proses perhitungan bobot pada query Q2 dan dokumen berdasarkan isi	64
IV-4 Similarity dokumen dari query Q2 dan dokumen berdasarkan isi	64
IV-5 Proses perhitungan bobot pada query Q3 dan dokumen berdasarkan isi	68
IV-6 Similarity dokumen dari query Q3 dan dokumen berdasarkan isi	69
IV-7 Nilai rata-rata <i>precision</i> dan <i>recall</i> berdasarkan isi dokumen	71
IV-8 Proses perhitungan bobot pada query Q1 dan dokumen berdasarkan judul	73
IV-9 Similarity dokumen dari Q1 dan dokumen berdasarkan judul	74
IV-10 Proses perhitungan bobot pada query Q2 dan dokumen berdasarkan judul	76
IV-11 Similarity dokumen dari query Q2 dan dokumen berdasarkan judul	76
IV-12 Proses perhitungan bobot pada query Q3 dan dokumen berdasarkan judul	79
IV-13 Similarity dokumen dari query Q3 dan dokumen berdasarkan judul	79
IV-14 Nilai rata-rata <i>precision</i> dan <i>recall</i> berdasarkan judul dokumen	81
IV-15 Proses perhitungan bobot pada query Q1 dan dokumen berdasarkan dokumen	85
IV-16 Similarity dokumen dari query Q1 dan dokumen berdasarkan dokumen	86

IV-17	Proses perhitungan bobot pada query Q2 dan dokumen berdasarkan dokumen	91
IV-18	Similarity dokumen dari query Q2 dan dokumen berdasarkan dokumen	92
IV-19	Proses perhitungan bobot pada query Q3 dan dokumen berdasarkan dokumen	98
IV-20	Similarity dokumen dari query Q3 dan dokumen berdasarkan dokumen	100
IV-21	Nilai rata-rata <i>precision</i> dan <i>recall</i> berdasarkan dokumen	102
IV-22	perhitungan nilai rata-rata precision, recall dan performa dari pencarian dokumen	103



DAFTAR LAMPIRAN

Lampiran	Halaman
1 Tampilan isi dokumen Q1 menggunakan mesin pencari	108
2 Tampilan isi dokumen Q2 menggunakan mesin pencari	108
3 Tampilan isi dokumen Q3 menggunakan mesin pencari	109
4 Tampilan judul dokumen Q1 menggunakan mesin pencari	109
5 Tampilan judul dokumen Q2 menggunakan mesin pencari	110
6 Tampilan judul dokumen Q3 menggunakan mesin pencari	110
7 Tampilan dokumen Q1 menggunakan mesin pencari	111
8 Tampilan dokumen Q2 menggunakan mesin pencari	111
9 Tampilan dokumen Q3 menggunakan mesin pencari	112
10 Tampilan List Dokumen	112
11 Scrip Pembobotan	113
12 Scrip Stemming	121

DAFTAR PUSTAKA

- [Al-Bahra. 2005]. Analisis dan Desain Sistem Informasi. Yogyakarta: Graha Ilmu.
- [Agus Zainal Arifin, 2002]. *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*. Surabaya: Institut Teknologi Sepuluh Nopember (ITS).
- [Alwi, M. Y. 2013]. *Efektifitas Sistem Temu Kembali Informasi Dalam Meningkatkan Pemanfaatan Koleksi :(Studi Kasus Pemustaka Pada Perpustakaan STMIK Potensi Utama Medan)*. Medan: Usu Perss.
- [Baeza-Yates, R. And Ribiero-Neto, B., *Modern Information retrieval*, New York. Addison-Wesley, 1999
- [Christopher, D. Manning, Prabhakar Raghavan, Hinrich Schütze. 2009]. *An Introduction to Information Retrieval*. Cambrigde: Cambridge UP.
- [Cios, Krzysztof J. Etc.2007]. *Data Mining A Knowledge Discovery Approach*, Springer.(online).(http://www.4shared.com/document/FyVdn5pm/Data_Mining_A_Knowledge_Discov.html, diakses 8 maret 2011).
- [Dipa, 2015]. *Sistem Temu Kembali Informasi atau Information Retrieval*. Retrieved May 10, 2015, from http://baliinnetwork.blogspot.com/http://baliinnetwork.blogspot.com/2015/03/sistem-temu-kembali-informasi-atau_3.html
- [Fatmawati, 2013]. *Studi Komparatif Kecepatan Temu Kembali Informasi di Depo Arsip Koran Suara Merdeka Antara Sistem Simpan Manual dengan Foto Repro*. Diponegoro University.
- [Guritno, S., Sudaryono., dan Rahardja, U., 2011]. *Theory And Application Of IT Reearch*. Andi: Yogyakarta.
- [Liddy. E, 2001], How a search engine works **URL:** <http://www.infoday.com/searcher/may01/liddy.htm>
- [Literacy, T. S. 2011]. *The SCONUL Seven Pillars of Information Literacy*. The Sconul.
- [Lu, 1999]. *Multimedia Database Management Systems*. Boston: Artech House,INC