

**MESIN PENCARI DOKUMEN, STUDI EFEKTIVITAS PADA
VEKTOR SPACE MODEL, ALGORITMA PORTER,
PEMBOBOTAN FREKUENSI KEMUNCULAN TERM DALAM
DOKUMEN DENGAN SKEMA KESAMAAN DICE**

TESIS



Oleh :
Rosdiana
1111601561

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

JAKARTA
2015

**MESIN PENCARI DOKUMEN, STUDI EFEKTIVITAS PADA
VEKTOR SPACE MODEL, ALGORITMA PORTER,
PEMBOBOTAN FREKUENSI KEMUNCULAN TERM DALAM
DOKUMEN DENGAN SKEMA KESAMAAN DICE**

TESIS

Diajukan sebagai salah satu persyaratan
untuk mendapatkan gelar Magister Ilmu Komputer (MKOM)



Oleh :
Rosdiana
1111601561

**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 bertandatangan dibawah ini :

Nama Mahasiswa : **ROSDIANA**
Nomor Induk Mahasiswa : **1111601561**
Program Studi : **Magister Ilmu Komputer**
Konsentrasi : **Teknologi Sistem Informasi**

menyatakan bahwa Tesis yang berjudul :

**MESIN PENCARI DOKUMEN, STUDI EFEKTIVITAS PADA VEKTOR
SPACE MODEL, ALGORITMA PORTER, PEMBOBOTAN FREKUENSI
KEMUNCULAN TERM DALAM DOKUMEN DENGAN SKEMA
KESAMAAN DICE**

1. Merupakan hasil karya tulis ilmiah sendiri dan bukan merupakan karya yang pernah diajukan untuk memperoleh gelar akademik oleh 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.

Jakarta 8 Agustus 2015

ROSDIANA



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

LEMBAR PERSETUJUAN TESIS

Nama Mahasiswa : **ROSDIANA**
Nomor Induk Mahasiswa : 1111601561
Jenjang Studi : Strata Dua (S2)
Konsentrasi : Teknologi Sistem Informasi
Judul Tesis : **MESIN PENCARI DOKUMEN, STUDI
EFEKTIVITAS PADA VEKTOR SPACE
MODEL, ALGORITMA PORTER,
PEMBOBOTAN FREKUENSI
KEMUNCULAN TERM DALAM DOKUMEN
DENGAN SKEMA KESAMAAN DICE**

Telah diperiksa, diuji dan dipertahankan dalam sidang ujian Tesis pada Hari Sabtu, Tanggal Delapan Agustus Tahun Dua Ribu Lima Belas, dan disetujui oleh Tim Penguji Tesis untuk menyusun Naskah Akhir Tesis.

Jakarta, 1 September 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

Bertambahnya kebutuhan manusia akan informasi, maka semakin banyak pula informasi yang ada dan perlu disimpan apabila akan dipergunakan kembali sewaktu-waktu. Penelitian ini bertujuan untuk memudahkan *user* dalam mendapatkan informasi sesuai tanpa harus mengetikkan persis seperti dokumen yang dicari karena sistem pencarian dilakukan dengan mengacu kepada isi dokumen. Untuk menemukan dokumen yang dicari sesuai dengan kebutuhan *user*, sistem pencarian dokumen ini dibuat dengan konsep *information retrieval* (perolehan informasi) dengan menggunakan algoritma *stemming porter* dan pemprofilan teks berbasis *vector space model* dengan pembobotan frekuensi kata dalam dokumen, serta fungsi kesamaan Dice. *Information Retrieval* (Perolehan Informasi) merupakan suatu pencarian informasi (biasanya berupa dokumen) yang didasarkan pada suatu *query* (hasil *input user*) yang diharapkan dapat memenuhi keinginan *user* dari dokumen yang ada. Algoritma *stemming porter* digunakan untuk melakukan *stemming* menjadikan kebentuk kata dasar dan pemprofilan menggunakan frekuensi kata dan kesamaan *dice*. Algoritma *stemming porter* digunakan untuk teks berbahasa Indonesia karena memiliki keakuratan yang tinggi dalam pencarian dokumen teks berbahasa Indonesia. Dengan terbentuknya *Information Retrieval* (perolehan informasi) maka dapat terbentuknya aplikasi yang dapat melakukan pencarian dokumen yang lebih efektif dan relevan sesuai dengan pencarian dan dapat diterapkan sebagai sistem pencarian dokumen elektronik untuk membantu pengguna untuk melakukan pencarian kata pada sebuah artikel pada sebuah situs portal *website* berita.

Kata kunci : Algoritma *Stemming Porter*, *Vector Space Model*, *Information Retrieval*, *Stemming*, *dice*.

ABSTRACT

As long as Information technology develops, it also increase the human needed of information. So, the more information we get it, there also more information needs to be saved any time we needed. This study aims to facilitate the user, to get appropriate information without having to type the document exactly that want to search because the searching system is done based on the content of the document. To find out the appropriate document with the user needs, the system searches the document was created with the concept of information retrieval (information acquisition) using porter stemming algorithms and text-based profiling vector space model with weighted frequency of words in a document as similarity search function. Information Retrieval is search information (usually in the form of a document) that is based on a query (the result of user input) that is expected to meet the needs of the user from an existing document. Porter stemming algorithm used to perform basic word stemming make to forms and profiling using word frequency and dice similarity. Porter stemming algorithm used for the text in Indonesian language because it has a high accuracy in the search text documents in Indonesian language. With the formation of Information Retrieval (information acquisition) is can be formation of an application that can perform document searches more effective and relevant in appropriate with the search and can be applied an electronic document search system to help users to perform a search word in an article on a website portal site news.

Keywords : Algorithm Stemming Porter, Vector Space Model, Information Retrieval, Stemming, Dice.

KATA PENGANTAR

Puja dan puji syukur kepada Allah, S.W.T, atas semua limpahan kasih sayang-Nya, penulis dapat menyelesaikan penelitian tesis yang berjudul Mesin Pencari Dokumen, Studi Efektifitas Pada Vektor *Space* Model, Algoritma Porter, Pembobotan Frekuensi Kemunculan Term Dalam Dokumen Dengan Skema Fungsi Kesamaan Dice.

Tujuan dari penulisan tesis ini adalah sebagai salah satu syarat untuk kelulusan pada Program Studi Magister Ilmu Komputer, Universitas Budi Luhur Jakarta.

Rasa dan ucapan terima kasih penulis persembahkan kepada semua pihak yang telah membantu penulis dalam menyusun tesis ini:

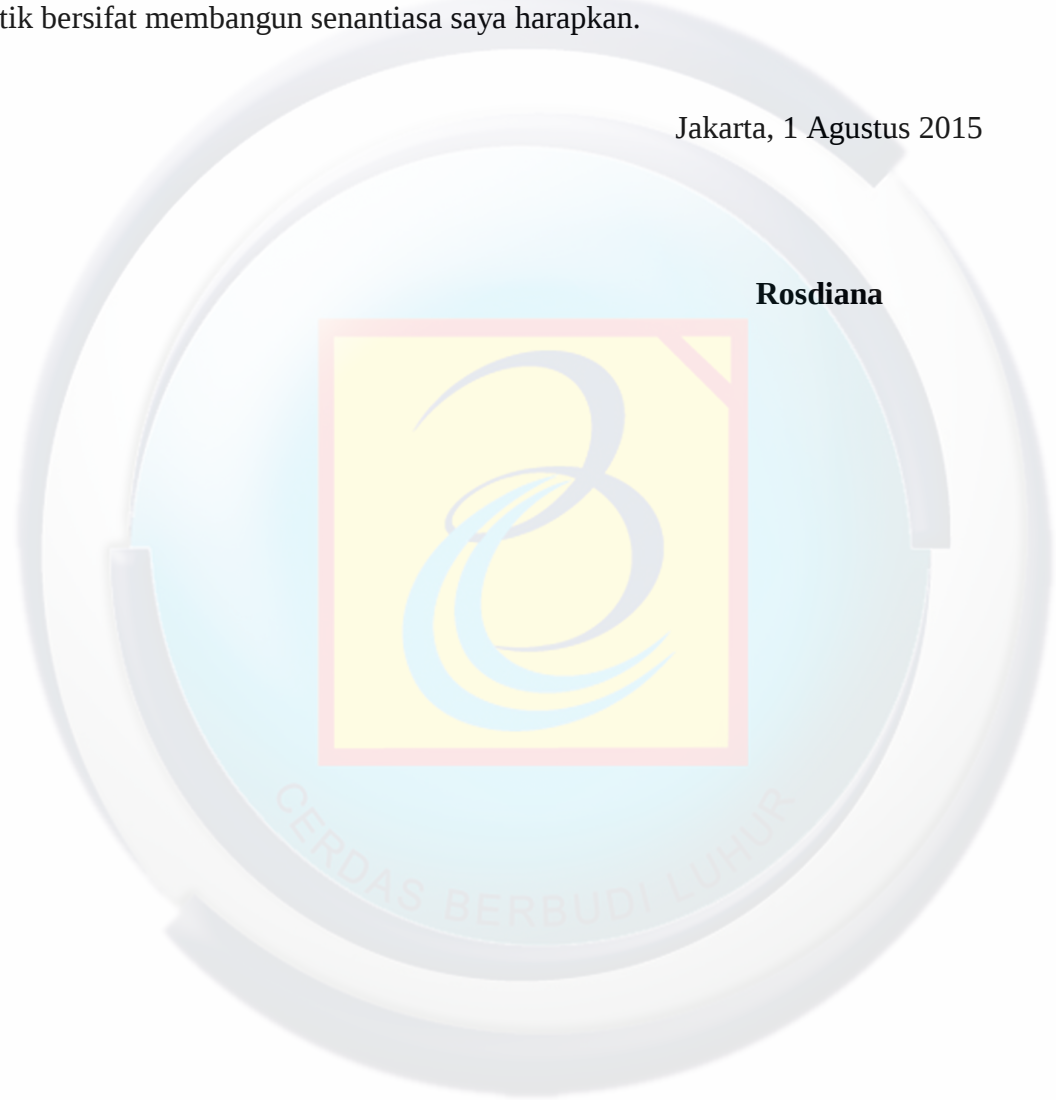
1. Allah, S.W.T, Tuhan pencipta makhluk, bumi, alam semesta dan kehidupan, sebagai motivasi paling utama menyelesaikan magister ini, karena Allah mencintai manusia yang senantiasa menimba ilmu dan mengamalkannya.
2. Suami dan keluarga tercinta yang selalu memberikan dukungan, kasih sayang dan perhatian, sehingga saya dapat menyelesaikan penyusunan tesis ini.
3. Bapak Dr. Krisna Adiyarta Musodo, selaku dosen pembimbing tesis yang telah membimbing dan memotivasi penulis dalam menyelesaikan tesis ini.
4. Para mahasiswa Strata Satu (S1) dan Strata Dua (S2) bimbingan Dr. Krisna yang telah bekerjasama dalam setiap halnya, berjuang bersama, memberikan motivasi, dan waktu untuk mengerjakan tesis ini dengan berbagai keterbatasan yang dimiliki.
5. Bapak dan Ibu Dosen pengampu mata kuliah di Program Studi Magister Ilmu Komputer Universitas Budi Luhur yang telah dengan sabar memberikan ilmu pengetahuan, pencerahan, dan bimbingan dalam belajar.

Saya menyadari, bahwa dalam penyusunan Tesis ini tidak lepas dari kekurangan dan kesalahan yang disebabkan oleh beberapa keterbatasan yang saya miliki baik keterbatasan waktu, pengalaman, pengetahuan dan kemampuan yang ada pada diri saya.

Harapan saya agar segala daya dan upaya yang saya lakukan dalam menyusun Tesis ini dapat memberikan manfaat tambahan pengetahuan bagi pembaca pada umumnya dan saya pada khususnya. Kesempurnaan hanyalah milik Allah S.W.T dan ketidaksempurnaan ada pada saya, oleh karena itu saran dan kritik bersifat membangun senantiasa saya harapkan.

Jakarta, 1 Agustus 2015

Rosdiana



DAFTAR ISI

	Halaman
ABSTRAK	i
ABSTRACT	ii
KATA PENGANTAR	iii
DAFTAR ISI	v
DAFTAR GAMBAR	vii
DAFTAR TABEL	viii
BAB I PENDAHULUAN	1
1.1 Latar Belakang.....	1
1.2 Masalah Penelitian.....	3
1.2.1 Identifikasi Masalah	3
1.2.2 Batasan dan Ruang Lingkup Masalah	3
1.2.3 Rumusan Masalah.....	4
1.3 Tujuan dan Manfaat Penelitian.....	5
1.3.1 Tujuan Penelitian.....	5
1.3.2 Manfaat Penelitian.....	5
1.4 Tata Urut Penulisan	7
1.5 Daftar Pengertian.....	8
BAB II LANDASAN TEORI DAN PEMIKIRAN.....	10
2.1 Tinjauan Pustaka	10
2.1.1 Konsep dan Pengertian Information Retrival System ...	10
2.1.2 Tujuan dan fungsi Information Retrival System	15
2.1.3 Arsitektur Information Retrival System	17
2.1.4 Jenis-jenis metode Information Retrival System	18
2.1.5 Model Representasi dokumen	19
2.1.6 Metodologi Pengindeksan teks.....	23
2.1.6.1. Parsing.....	24
2.1.6.2. Filtering.....	26
2.1.6.3. Stemming.....	26
2.1.6.4. Document term weightin.....	34
2.1.6.5. Similarity fuction	35
2.1.7 Pengukuran relvansi dan efektivitas Information Retrival System	36
2.1.8 Sejarah dari media elektronik koran Kompas.com.....	39
2.2 Tinjauan Studi.....	43
2.3 Kerangka Konseptual dan Hipotesis.....	47
2.3.1 Kerangka Konsep/Pola Pikir Pemecahan Masalah.....	47
2.3.2 Hipotesis	48

BAB III	METODE PENELITIAN	50
3.1	Metode Penelitian.....	50
3.2	Metode Pengujian.....	53
3.3	Metode Mengumpulkan Data	53
3.4	Instrumentasi	56
3.5	Teknis Analisis dan Rancangan Sistem.....	57
3.5.1	Teknis Analisis	57
3.5.2	Langkah – langkah Penelitian	57
3.5.3	Bentuk dan Sumber Data.....	60
3.5.4	Prototype Model.....	61
3.5.5	Jadwal Penelitian.....	63
BAB IV	PEMBAHASAN.....	66
4.1	Pengelompokan Data.....	66
4.2	Prototipe Model	66
4.3	Implikasi Penelitian	69
4.4	Analisis Hasil.....	88
4.5	Grafik Hasil Analisis	92
4.6	Rencana Implikasi Penelitian	93
4.6.1	Aspek Sistem	93
4.6.2	Aspek Manajerial.....	93
4.6.3	Aspek Penelitian Lanjut	93
4.6.4	Rencana Implementasi.....	94
4.6.5	Tahapan Implementasi.....	94
4.6.6	Penjelasan Kegiatan Implementasi.....	95
BAB V	PENUTUP	96
5.1	Kesimpulan.....	96
5.2	Saran	66

DAFTAR PUSTAKA
LAMPIRAN-LAMPIRAN
RIWAYAT HIDUP SINGKAT

DAFTAR GAMBAR

Gambar	Halaman
II-1 Proses Information Retrieval.....	12
II-2 Proses Information Retrieval Sistem	17
II-3 Représentasi dokument dan vektor pada ruang vektor	21
II-4 Représentasi matriks kata dokumen	22
II-5 Représentasi grafis sudut vektor dan query.....	22
II-6 Tahapan didalam pemrosesan teks.....	25
II-7 Jenis-jenis metode stemming.....	27
II-8 Desain dari Porter Stemmer untuk bahasa indonesia	29
II-9 Recall dan Precision pada contoh hasil IRS.....	38
II-10 Pola representasi relevance ()	39
III-1 Langkah-langkah penelitian.....	58
III-2 Contoh Proses Tokenisasi Query	62
IV-1 Grafik rata – rata Precision dan Recall per-topik untuk Isi.....	89
IV-2 Grafik rata – rata Precision dan Recall per-topik untuk Judul.....	90
IV-3 Grafik rata – rata Precision dan Recall per-topik untuk Keyword.....	91
IV-4 Grafik Uji Efektifitas.....	92

DAFTAR TABEL

Tabel	Halaman
II-1 Tabel akhiran –lah, -kah, -pun.....	30
II-2 Tabel akhiran –ku, -mu, -nya	31
II-3 Tabel awalan 1/cluster 3.....	31
II-4 Tabel awalan 2/cluster 4.....	32
II-5 Tabel akhiran /cluster 5.....	33
II-6 Pasangan awalan dan akhiran yang tidak sah	33
II-7 Ringkasan tinjauan studi.....	44
III-1 Tabel Strategi Penelitian 1.....	50
III-2 Tabel Strategi Penelitian 2.....	51
III-3 Jadwal Penelitian.....	63
IV-1 Precision dan Recall Topik “Tekno” Berdasarkan Isi Dokumen.....	70
IV-2 Precision dan Recall Topik “Female” Berdasarkan Isi Dokumen.....	70
IV-3 Precision dan Recall Topik “Edukasi” Berdasarkan Isi Dokumen.....	71
IV-4 Precision dan Recall Topik “Travel” Berdasarkan Isi Dokumen.....	71
IV-5 Precision dan Recall Topik “Ekonomi” Berdasarkan Isi Dokumen.....	72
IV-6 Precision dan Recall Topik “Tekno” Berdasarkan Judul.....	73
IV-7 Precision dan Recall Topik “Female” Berdasarkan Judul.....	73
IV-8 Precision dan Recall Topik “Edukasi” Berdasarkan Judul.....	74
IV-9 Precision dan Recall Topik “Travel” Berdasarkan Judul.....	74
IV-10 Precision dan Recall Topik “Ekonomi” Berdasarkan Judul.....	74
IV-11 Precision dan Recall Topik “Tekno” Berdasarkan Keyword.....	75
IV-12 Precision dan Recall Topik “Female” Berdasarkan Keyword.....	78
IV-13 Precision dan Recall Topik “Female” Berdasarkan Keyword.....	81
IV-14 Precision dan Recall Topik “Travel” Berdasarkan Keyword.....	84
IV-15 Precision dan Recall Topik “Ekonomi” Berdasarkan Ekonomi.....	86
IV-16 Rata – rata Precision dan Recall per-topik untuk Isi.....	89
IV-17 Rata – rata Precision dan Recall per-topik untuk Judul.....	90
IV-18 Rata – rata Precision dan Recall per-topik untuk Keyword.....	91
IV-19 Rata – rata Precision dan Recall per kategori.....	92
IV-20 Rencana Implementasi.....	94

DAFTAR LAMPIRAN

Lampiran

1. Lampiran 1. Contoh Artikel Edukasi
2. Lampiran 2. Contoh Artikel Female
3. Lampiran 2. Interface Aplikasi Pengujian Tingkat Relevansi



DAFTAR PUSTAKA

- [Agus 2002] Agus Zainal Arifin, A. N. *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering*. Surabaya: Institut Teknologi Sepuluh Nopember (ITS), 2002.
- [Alwi 2013] Alwi, M. Y. *Efektifitas Information Retrieval System Dalam Meningkatkan Pemanfaatan Koleksi : (Studi Kasus Pemustaka Pada Perpustakaan STMIK Potensi Utama Medan)*. Medan: Usu Perss, 2013.
- [Barasa 2009] Barasa, B. E. *Efektivitas STBI Rekam Medis*. Medan: USU Press, 2009.
- [Christopher 2009] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge: Cambridge UP, 2009.
- [David 1993] David Ellis, J. F.-H., & Willett, P. *Measuring The Degree of Similarity Between Object in Text Retrieval Systems. Perspective in Information Management*, 128 – 149, 1993.
- [Dipa 2015] Dipa, I. W. (2015). *Information Retrieval System atau Information Retrieval*. Dipetik May 10, 2015, dari http://baliinnetwork.blogspot.com:2015/03/sistem-temu-kembali-informasi-atau_3.html, 2015.
- [Fatmawati 2013] Fatmawati, E. *Studi Komparatif Kecepatan Temu Kembali Informasi di Depo Arsip Koran Suara Merdeka Antara Sistem Simpan Manual dengan Foto Repro*. Diponegoro University, 2013.
- [Hasugian 2006] Hasugian, J. *Penggunaan Bahasa Alamiah dan Kosa Kata Terkendali dalam Sistem Temu Balik Informasi Berbasis Teks*. 74, 2006
- [Hasugian 2009] Hasugian, J. *Dasar - Dasar Ilmu Perpustakaan dan Informasi*. Medan: USU Press, 2009.