

**MESIN PENCARI DOKUMEN STUDY EFEKTIVITAS PADA VEKTOR
SPACE MODEL, ALGORITMA PORTER, PEMBOBOTAN FREKUENSI
TERM DIBANDING JUMLAH KATA DALAM DOKUMEN DAN FUNGSI
COSINE**

TESIS



Oleh :

HARRY TANSIL

1311600181

PROGRAM STUDI: MAGISTER ILMU KOMPUTER (MKOM)

PROGRAM PASCA SARJANA ILMU KOMPUTER

JAKARTA

2015



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

LEMBAR PERNYATAAN

Nama Mahasiswa : Harry Tansil
Nomor Induk Mahasiswa : 1311600181
Program Study : Magister Ilmu Komputer
Konsetrasi : Sistem Informasi

Meyatakan bahwa Tesis yang berjudul:

MESIN PENCARI DOKUMEN STUDY EFEKTIVITAS PADA VEKTOR SPACE MODEL, ALGORITMA PORTER, PEMBOBOTAN FREKUENSI TERM DIBANDING JUMLAH KATA DALAM DOKUMEN DAN FUNGSI 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 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, 31 Agustus 2015


Harry Tansil



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

LEMBAR PENGESAHAN

Nama Mahasiswa : **HARRY TANSIL**
Nomor Induk Mahasiswa : **1311600181**
Konsentrasi : Teknologi Sistem Informasi
Judul Tesis : Mesin pencari dokumen study efektivitas pada vektor space model, algoritma porter, pembobotan frekuensi term dibanding jumlah kata dalam dokumen dan fungsi cosine

Telah diperiksa, diuji dan dipertahankan dalam sidang ujian Tesis pada Hari Senin Tanggal 10 Agustus 2015 dan disetujui oleh Tim Penguji Tesis untuk menyusun Naskah Akhir Tesis.

Jakarta, 31 Agustus 2015

Tim Penguji:

Ketua,

Prof. Dr. Moedjiono, M.Sc

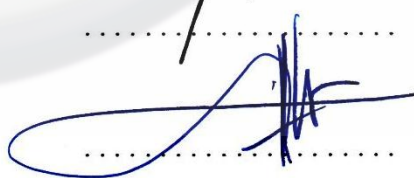
Anggota,

Dr. Ir. Nazori Az, M.T

Pembimbing Utama,

Dr. Krisna Ardiyarta Musodo

Tanda Tangan:


.....

.....

Ketua Program Studi


Dr. Ir. Nazori Az, M.T

ABSTRAK

Teknologi informasi yang berkembang pesat dewasa ini dan semakin 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 berbanding jumlah kata dalam dokumen frekuensi kata dalam pencarian serta fungsi kesamaan. *Information Retrieval* (Perolehan Informasi) merupakan suatu pencarian informasi (biasanya berupa dokumen) yang didasarkan pada suatu *query* (hasil *input user*) dan menjadi keinginan *user* dari dokumen yang ada. Algoritma *stemming porter* digunakan untuk melakukan *stemming* menjadikan kebentuk kata dasar dan pemprofilan menggunakan frekuensi kata dan kesamaan *cosine*. 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) pengguna dapat menggunakan 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*, *Information Retrieval*, *Vector Space Model*, *Stemming*, *Cosine*.

ABSTRACT

This time, technology development increase rapidly. 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 directly multiplying the number of words in a document in a word frequency as well 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) and to become 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 cosine 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) the user can use application to be the 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, Information Retrieval, Vector Space Model, Stemming, Cosine.

KATA PENGANTAR

Puji dan syukur kepada Tuhan Yang Maha Esa atas semua berkat dan kasih sayangnya dalam penulis bisa menyelesaikan Proposal tesis yang berjudul mesin pencari dokumen study efektivitas pada vektor space model, algoritma porter, pembobotan frekuensi term dibanding jumlah kata dalam dokumen dan fungsi cosine. Tujuan dari penulisan tesis ini adalah sebagai salah satu syarat untuk menyusun tesis 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. Tuhan Yesus Kristus, pencipta alam semesta yang selalu memudahkan segala urusan dan selalu memberi semangat dan menemani di saat apa pun.
2. Kedua orang tua, Koh Erni dan Koh Erna dan Oma ku tercinta dan keluarga besar Tansil 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. Terima kasih atas kasih serta motivasi kepada sahabat-sahabat tim larjo, (Ghilman, Wahyu, Listra, Dessy, Arul) yang selalu memberikan dukungan hingga saya dapat menyelesaikan penyusunan tesis ini.
5. Para Mahasiswa S1 dan S2 bimbingan Pak Krisna, Mba Rosa, Mba Widya, Mas Zum, Pak Kardi, Mas Lazarus dan yang belum tersebut namanya dan yang telah berkerja sama dalam setiap halnya, berjuang bersama, memberikan motivasi, dan waktu untuk mengerjakan tesis tersebut dengan berbagai keterbatasan yang dimiliki.
6. Keluarga XL angkatan 2013 S2 Universitas Budi Luhur , memberikan inspirasi penulis dalam penelitian dan membantu memberkan informasi yang diperlukan terkait penelitian ini.
7. Keluarga besar Pelkat PT Bahtera Iman yang selalu memberikan doa dan dukungan dalam setiap perjuangan penulisan tesis ini.

8. Terima kasih keluarga besar PB Merlion yang selalu memberikan motivasi kepada saya dan membantu menemani dalam setiap mengerjakan tesis ini.
9. 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 Proposal 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 penelitian Tesis ini dapat memberikan manfaat tambahan pengetahuan bagi pembaca pada umumnya dan saya pada khususnya. Kesempurnaan hanyalah milik Tuhan semata dan ketidak sempurnaan ada pada saya, oleh karena itu saran dan kritik bersifat membangun senantiasa saya harapkan.

Jakarta, 24 Agustus 2015

Harry Tansil

DAFTAR ISI

	Halaman
ABSTRAK.....	i
ABSTRACT.....	ii
KATA PENGANTAR	iii
DAFTAR ISI.....	v
DAFTAR GAMBAR	vii
DAFTAR TABEL.....	viii
DAFTAR LAMPIRAN.....	xi
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Masalah Penelitian.....	2
1.2.1 Identifikasi Masalah	2
1.2.2 Batasan dan Ruang Lingkup 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 Penulisan	6
1.5 Daftar Pengertian.....	6
BAB II LANDASAN TEORI DAN PEMIKIRAN.....	9
2.1 Tinjauan Pustaka	9
2.1.1 Konsep dan Pengertian Information Retrival System	9
2.1.2 Tujuan dan fungsi Information Retrival System	13
2.1.3 Arsitektur Information Retrival System	15
2.1.4 Jenis-jenis metode Information Retrival System.....	16
2.1.5 Model Representasi dokumen	17
2.1.6 Metodologi Pengindeksan teks	21
2.1.6.1. Parsing.....	23
2.1.6.2. Filtering.....	23
2.1.6.3. Stemming.....	24
2.1.6.4. Dokument term weightin.....	30
2.1.6.5. Similarity fuction	31

2.1.7	Pengukuran relevansi dan efektivitas Information Retrieval System	32
2.1.8	Artikel makanan atau Detik.com	36
2.1.8.1.	Sejarah dan situs dari detik.com.....	36
2.2	Tinjauan Studi.....	37
2.3	Kerangka Konseptual dan Hipotesis.....	41
BAB III	METODOLOGI DAN RANCANGAN PENELITIAN	43
3.1	Metode Penelitian	43
3.2	Metode Pengujian	43
3.3	Metode Mengumpulkan Data	44
3.4	Instrumentasi	45
3.5	Teknis Analisis dan Rancangan Sistem.....	46
3.5.1	Teknis Analisis	46
3.5.2	Rancangan Sistem	46
3.6	Teknis Implementasi Sistem.....	47
3.7	Langkah – Langkah Penelitian	47
3.8	Teknik Pengujian.....	49
3.9	Jadwal Penelitian	50
BAB IV	PEMBAHASAN HASIL PENELITIAN.....	52
4.1	Pengelompokan Data.....	52
4.2	Prototipe Model	52
4.3	Pengujian Hasil Model	55
4.4	Analisis Hasil.....	62
4.5	Grafik Hasil Analisis	63
4.6	Implikasi Penelitian	64
4.6.1	Aspek sistem	64
4.6.2	Aspek manajerial.....	64
4.6.3	Aspek penelitian lanjutan	64
4.6.4	Rencana Implementasi	65
4.6.5	Tahapan Implementasi	65
4.6.6	Penjelasan kegiatan implementasi	65
BAB V	PENUTUP	67
5.1	Kesimpulan.....	67
5.2	Saran	67
DAFTAR PUSTAKA		68
LAMPIRAN-LAMPIRAN.....		01
RIWAYAT HIDUP SINGKAT.....		05

DAFTAR GAMBAR

Gambar	Halaman
II-1 Bagian-bagian <i>information sistem retrieval</i> (IR)	10
II-2 Proses Sistem Temu-Kembali Informasi.....	15
II-3 Représentasi dokument dan vektor pada ruang vektor	18
II-4 Représentasi matriks kata dokumen	19
II-5 reprintsntasi grafis sudut vektor dan query	20
II-6 Arsitektur sistem pencari dokumen teks	22
II-7 Jenis-jenis metode stemming.....	24
II-8 Desain dari Porter Stemmer untuk bahasa indonesia	26
II-9 Recall dan Precision pada contoh hasil IRS.....	34
II-10 Pola representasi revance ()	35
III-1 Langkah-langkah penelitian.....	48
III-2 Jadwal Penelitaian.....	53
IV-1 Grafik hasil analisa penelitian.....	53

DAFTAR TABEL

Tabel	Halaman
II-1 Tabel akhiran –lah, -kah, -pun.....	27
II-2 Tabel akhiran –ku, -mu, -nya	27
II-3 Tabel awalan 1/cluster 3.....	28
II-4 Tabel awalan 2/cluster 4.....	29
II-5 Tabel akhiran /cluster 5	29
II-6 Pasangan awalan dan akhiran yang tidak sah	30
II-7 Ringkasan tinjauan studi	38
III-1 Jadwal Penelitian.....	50
IV-4.3.1.1 Precision dan recall topik makanan	55
IV-4.3.1.2 Precision dan recall topik olahraga	56
IV-4.3.1.3 Precision dan recall topik kesehatan.....	56
IV-4.3.2.1 Precision dan recall topik makanan	57
IV-4.3.2.2 Precision dan recall topik olahraga	57
IV-4.3.2.3 Precision dan recall topik kesehatan.....	58
IV-4.3.3.1 Precision dan recall topik makanan	58
IV-4.3.3.2 Precision dan recall topik olahraga	59
IV-4.3.3.3 Precision dan recall topik kesehatan.....	61
IV-4.4.1 Rata-rata precision dan recall pertopik untuk Isi	62
IV-4.4.2 Rata-rata precision dan recall pertopik untuk Judul	62
IV-4.4.3 Rata-rata precision dan recall pertopik untuk Kata Kunci.....	62
IV-4.4.4 Rencana Implementasi	65

DAFTAR LAMPIRAN

Lampiran	Halaman
1. <u>Lampiran Contoh Artikel</u>	1
2. <u>Lampiran Interface dari software yang digunakan</u>	2
3. <u>Lampiran CV</u>	5



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 Sistem Temu Kembali Informasi 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). *Sistem Temu Kembali Informasi 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.