

**MESIN PENCARI DOKUMEN BAHASA INDONESIA, STUDI
EFEKTIVITAS PENCARIAN PADA VEKTOR SPACE MODEL
ALGORITMA PORTER, PEMBOBOTAN FREKUENSI TERM
BERBANDING JUMLAH TERM SKEMA KESAMAAN DICE**

TESIS



Oleh :
ZAFRULLAH
1311600660

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

JAKARTA
2015

**MESIN PENCARI DOKUMEN BAHASA INDONESIA, STUDI
EFEKTIVITAS PENCARIAN PADA VEKTOR SPACE MODEL
ALGORITMA PORTER, PEMBOBOTAN FREKUENSI TERM
BERBANDING JUMLAH TERM SKEMA KESAMAAN DICE**

TESIS

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



Oleh :
ZAFRULLAH
1311600660

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

JAKARTA
2015



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

LEMBAR PENGESAHAN

Nama Mahasiswa : Zafrullah
Nomor Induk Mahasiswa : 1311600660
Konsentrasi : Teknologi Sistem Informasi
Jenjang Studi : Strata 2
Judul Tesis : Mesin Pencari Dokumen Bahasa Indonesia, Studi Efektivitas Pencarian Pada Vektor Space Model Algoritma Porter, Pembobotan Frekuensi Term Berbanding Jumlah Term Skema Kesamaan Dice

Jakarta, Senin 10 Agustus 2015

Tim Penguji:

Ketua,

Prof. Dr. Moedjiono, Msc

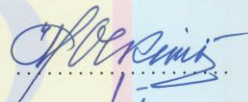
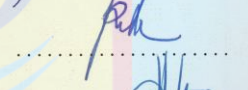
Anggota,

Dr., Ir. Nazori Az, M.T

Pembimbing Utama,

Dr. Krisna Adivarta M.

Tanda Tangan:


Ketua Program Studi



Dr., Ir. Nazori Az, M.T



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

LEMBAR PERNYATAAN

Nama Mahasiswa : Zafrullah
Nomor Induk Mahasiswa : 1311600660
Program Studi : Magister Ilmu Komputer
Konsentrasi : Teknologi Sistem Informasi
Fakultas/Program : Pascasarjana

Menyatakan bahwa Tesis yang berjudul :

Mesin Pencari Dokumen Bahasa Indonesia, Studi Efektivitas Pencarian Pada Vector Space Model Algoritma Porter, Pembobotan Frekuensi Term Berbanding Jumlah Term 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 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 dikemudian hari pernyataan ini tidak benar.

Jakarta, 10 Agustus 2015



(Zafrullah)

ABSTRAK

Dalam kehidupan, manusia tidak bisa lepas dari kebutuhan dan permasalahan. Banyak sekali kebutuhan yang dimiliki setiap manusia. Salah satunya kebutuhan akan informasi berupa dokumen. Kebutuhan tersebut dapat diperoleh dari berbagai sumber seperti media elektronik, media cetak, ataupun melalui media internet. Dan model pencarian informasinya pun dapat berupa pencarian secara manual dengan bantuan pengguna sendiri ataupun secara otomatis, yaitu dengan bantuan suatu program pencari. Ini merupakan salah satu cara yang dapat dilakukan untuk mempermudah pengguna menemukan dokumen yang sesuai dengan keinginannya. Dan program tersebut adalah dengan system *Information Retrival* atau sistem perolehan informasi. Sistem perolehan informasi merupakan system yang berfungsi untuk menemukan informasi yang relevan dengan kebutuhan pemakai. Pengguna sistem informasi akan mengetikkan kata-kata kunci (*query*) yang mewakili hal yang ingin dicarinya. *Query-query* tersebut dalam penelitian ini berupa isi dari suatu dokumen yang berasal dari kumpulan dokumen yang ada. Kemudian sistem akan mencari dokumen-dokumen yang mengandung kata -kata kunci yang terdapat dalam *query* dan menampilkan dokumen-dokumen tersebut, dan terturut menurut tingkat relevansinya. Pada penelitian kali ini peneliti meneliti suatu berkas identitas abstraksi jurnal berbahasa indonesia dengan menggunakan metode VSM (*Vector Space Model*) dan melalui proses pengindeksan parsing (*tokenization*), *filtering* (*stopword removal*) dan *stemming* dalam kali ini peneliti menggunakan *stemming* porter untuk mendapatkan kata dasar dari bahasa indonesia, dengan pembobotan frekuensi kemunculan *term* berbanding banyaknya jumlah *term* dokumen dari dokumen koleksi yang muncul dan akan dihitung kemiripan dari vektor tersebut dengan perhitungan kesamaan *Dice*, untuk data yang diuji peneliti pengkoleksi 3 kelompok penelitian jurnal dan setiap kelompok memiliki jumlah string judul beserta isi dokumen abstraksi yang berbeda. Dan dilakukan pengujian pada proses tipe pemberian *keyword* antara lain *base on query*, *base on title* dan *base on* dokumen dengan masing masing *keyword* memiliki topik yang berbeda, dari hasil pengujian tersebut akan diambil nilai *similarity* dan diolah untuk mendapatkan hasil yang relevan sesuai topik dimasing-masing *keyword* dan dicari nilai efektif dengan menggunakan perhitungan *precision* dan *recall*, sehingga dapat direpresentasikan melalui grafik dimasing-masing pengujian *keyword*. Dan hasil dari pengujian yang dilakukan, nilai pencarian *keyword* berdasarkan isi dianggap memiliki hasil yang paling baik dan efektif dibandingkan dengan pencarian berdasarkan judul dan pencarian berdasarkan dokumen. Karena nilai *precision* dan *recall* yang mendekati nilai 1.

Kata kunci : Sistem Perolehan Informasi, *Vektor Space Model*, *Stemming* Porter, Pembobotan *term*, *Similarity Dice*, *Precision* dan *Recall*.

ABSTRACT

In life, man can not be separated from the needs and problems. Many needs of every human being. One was the need for information in the form of documents. These needs can be obtained from various sources such as electronic media, print media, or via the Internet. And information search model can also be a manual search with the help of the user himself or automatically, with the help of a search program. This is one way that can be done to simplify the user find documents in accordance with her wishes. And the program is the Information retrieval system or information retrieval system. Information acquisition system is a system that serves to find information that is relevant to the needs of users. Information system users would type keywords (query) that represent things you want looking for. The queries in this study of the contents of a document originating from the collection of documents. Then the system will search for documents containing the word - keyword contained in the query and displays the documents, and terturut according to level of relevance. In the present study the researchers examined a file identity abstraction journals Indonesian language by using VSM (Vector Space Model) and through the process of indexing parsing (tokenization), filtering (stopword removal) and stemming in this time researchers use stemming porter to obtain basic words of Indonesian, weighted frequency of occurrence of term versus the large number of terms a document from a document collection that arise and will be calculated similarity of vectors by calculating the similarity Dice, for the data tested researchers pengkoleksi three groups of research journals and each group has a number of strings titles along with the contents of the document different abstraction. And conducted tests on the type of administration of the keyword among others base on the query, base on title and base on documents with each keyword has a different topic, from the results of such testing would have taken the value of similarity and processed to get the results to the relevant topic in the respective keyword and sought effective value using the calculation precision and recall, so it can be represented by a graph in the respective test keyword. And the results of tests performed, the value of keyword search by content is considered to have the best results and effective than search by title and search based on paperwork. Because the value of precision and recall are approaching a value of 1.

Keywords : *Information Retrieval Systems, Vektor Space Model, Stemming Porter, Frekuensi term weighting, Similarity Dice, Precision dan Recall*

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa, karena berkat rahmat-Nya penulis bisa menyelesaikan penelitian tesis yang berjudul *Model Mesin Pencari Dokumen Teks Dalam Bahasa Indonesia, Study Efektivitas Pencarian Pada Penerapan Vektor Space Model Dengan Algoritma Stemming Porter, Pembobotan Frekuensi Kemunculan Term Berbanding Dengan Banyaknya Jumlah Term Dalam Sebuah Dokumen Serta Skema Fungsi Kesamaan Dice*. 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:

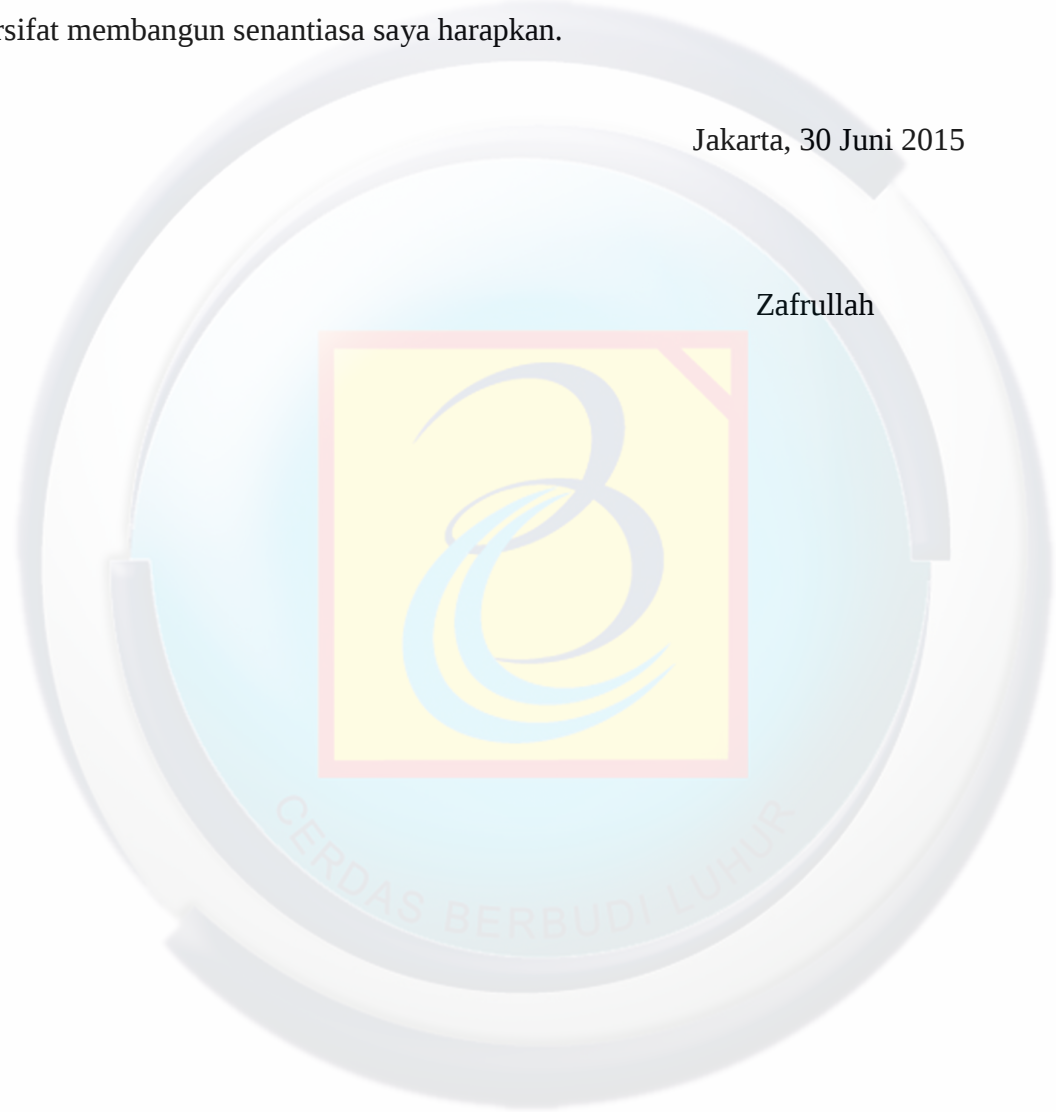
1. Allah SWT, pencipta alam semesta yang selalu memudahkan segala urusan dan selalu memberi semangat dan menemani di saat apa pun.
2. Kedua Orang Tua dan Keluarga yang selalu memberikan dukungan, kasih sayang dan perhatian sehingga saya dapat menyelesaikan penyusunan tesis ini.
3. Bapak Ir. Krisna Adiyarta, M.Sc. selaku dosen pembimbing tesis yang telah membimbing dan memotivasi penulis dalam menyelesaikan tesis ini.
4. Venty Fitriany Nurunisa yang selalu memberikan dukungan hingga saya dapat menyelesaikan penyusunan tesis ini “gembul”.
5. Rekan-rekan mahasiswa MKOM Universitas Budi Luhur kelas XL Semester 1 dan 2 dan MKOM Semester 3 konsentrasi Teknologi Sistem Informasi, terima kasih atas kebersamaan, kerja keras dan dukungan semangatnya.
6. Rekan-rekan seperjuangan untuk program thesis dalam pengambilan IRS
7. Rekan-rekan “Larjo” (Ghilman, Wahyu, Listra Dessy, Harry) yang selalu memberikan dukungan hingga saya dapat menyelesaikan penyusunan tesis ini.
8. 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 kekhilafan 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 SWT dan ketidak sempurnaan ada pada saya, oleh karena itu saran dan kritik bersifat membangun senantiasa saya harapkan.

Jakarta, 30 Juni 2015

Zafrullah



DAFTAR ISI

Isi	Halaman
ABSTRAK	i
KATA PENGANTAR	iii
DAFTAR ISI	v
DAFTAR GAMBAR	vii
DAFTAR TABEL	viii
DAFTAR LAMPIRAN	ix
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	7
BAB II LANDASAN TEORI DAN PEMIKIRAN	9
2.1 Tinjauan Pustaka	9
2.1.1 Konsep dan Pengertian Information Retrieval System	9
2.1.2 Tujuan dan Fungsi Information Retrieval System	11
2.1.3 Arsitektural Information Retrieval System	12
2.1.4 Jenis - Jenis Metode Information Retrieval System	14
2.1.5 Model Representasi Dokumen	15
2.1.6 Metodologi Pengindeksan Teks	18
2.1.6.1 Parsing	20
2.1.6.2 Filtering	20
2.1.6.3 Stemming	21
2.1.7 Document Term Weighting f_{in}/k_n	26
2.1.8 Similarity Function	28
2.1.9 Pengukuran Relevansi dan Efektifitas Information Retrieval System	29
2.1.10 Jurnal Elektronik atau e-journal	31
2.1.10.1 Pengertian e-journal	31
2.1.10.2 Aplikasi Pemanfaatan e-journal	32
2.2 Tinjauan Studi	33

2.3 Kerangka Konseptual dan Hipotesis	36
2.3.1 Kerangka Konsep/Pola Pikir Pemecahan Masalah.....	36
2.3.2 Hipotesis Penelitian	38
BAB III METODOLOGI DAN RANCANGAN PENELITIAN	39
3.1 Metode Penelitian.....	39
3.2 Metode Pemilihan Sampel	40
3.3 Asumsi dan Batasan	42
3.3.1 Bentuk dan Sumber Data	42
3.3.2 Asumsi – asumsi lain.....	43
3.4 Instrumentasi	44
3.5 Teknik Analisis dan Prototipe Model	45
3.5.1 Bentuk dan Sumber Data	45
3.5.2 <i>Prototype Model</i>	45
3.6 Teknik Pengujian	47
3.7 Teknik Pengukuran Efektifitas.....	48
3.8 Langkah-Langkah Penelitian	49
3.9 Jadwal Penelitian.....	52
BAB IV PEMBAHASAN IMPLEMENTASI HASIL PENELITIAN	54
4.1 Analisis Masalah Umum	54
4.2 Model Mesin Pencari	54
4.2.1 Representasi Data	54
4.2.1.1 Presisten Data	55
4.2.1.2 <i>Trasient Data</i>	57
4.2.2 Model Proses Koleksi Upload Dokumen	59
4.2.3 Model Proses Stemming.....	61
4.2.4 Model Proses <i>Search</i>	63
4.2.5 Implementasi Tampilan Layar.....	64
4.2.5.1 Tampilan Halaman <i>Search</i>	64
4.2.5.2 Tampilan Hasil <i>Search</i>	65
4.2.5.3 Tampilan Halaman <i>Dashboard Admin</i>	66
4.2.5.4 Tampilan Halaman <i>Upload</i> Dokumen Koleksi Admin.....	66
4.2.5.5 Tampilan Halaman List Dokumen Koleksi Admin.....	67
4.3 Perancangan Eksperimen	67
4.4 Hasil Eksperimen <i>Query</i> berdasarkan Judul Dokumen	69
4.5 Hasil Eksperimen <i>Query</i> berdasarkan Isi Dokumen	73
4.6 Hasil Eksperimen <i>Query</i> berdasarkan Dokumen dengan Dokumen.....	81
4.7 Analisa Hasil Experiment	90
4.8 Implikasi Penelitian.....	93
BAB V PENUTUP.....	95
5.1 Kesimpulan	98
5.2 Saran.....	99
DAFTAR PUSTAKA	101
LAMPIRAN	104
RIWAYAT HIDUP SINGKAT	109

DAFTAR GAMBAR

Gambar	Halaman
II - 1 Ilustrasi Sistem Temu-Kembali Informasi	9
II - 2 Proses Sistem Temu-Kembali Informasi.....	13
II - 3 Representasi dokumen dan vektor pada ruang vektor.....	16
II - 4 Representasi matriks kata-dokumen.....	17
II - 5 Representasi grafis sudut vektor dokumen dan <i>query</i>	17
II - 6 Tahapan di dalam pemrosesan teks.....	20
II - 7 Jenis-jenis Metode <i>Stemming</i>	22
II - 8 Desain dari <i>Porter Stemmer</i> untuk Bahasa Indonesia.....	23
II - 9 <i>Recall</i> dan <i>Precision</i> IRS	30
II - 10 Pola Representasi <i>Relevance</i>	31
II - 11 Pola Pikir	37
III - 1 Contoh Proses Tokenisasi <i>Query</i>	46
III - 2 Langkah – Langkah Penelitian.....	50
IV - 1 Pemodelan Data Konseptual	56
IV - 2 <i>Logical Record Structured</i>	56
IV - 3 <i>Flowchart</i> Proses <i>Upload</i> Dokumen Koleksi.....	59
IV - 4 <i>Flowchart</i> Proses <i>Stemming</i>	61
IV - 5 <i>Flowchart</i> Proses <i>Search</i>	63
IV - 6 Tampilan Halaman <i>Search</i>	65
IV - 7 Tampilan Halaman Hasil <i>Search</i>	65
IV - 8 Tampilan Halaman Dashboard Admin.....	66
IV - 9 Tampilan <i>Upload</i> Dokumen Koleksi	66
IV – 10 Tampilan List Dokumen Koleksi Admin.....	67
IV – 11 Tampilan Perhitungan <i>Precision Recall</i>	68
IV – 12 Grafik Hasil <i>Precision Recall</i> Q1, Q2, Q3 berdasarkan Judul	73
IV – 13 Grafik Hasil <i>Precision Recall</i> Q1, Q2, Q3 berdasarkan Isi	81
IV – 14 Grafik Hasil <i>Precision Recall</i> Q1, Q2, Q3 berdasarkan Dokumen	89
IV – 15 Grafik Hasil <i>Precision Recall</i> Q1 berdasarkan Judul, Isi, Dokumen	90
IV – 16 Grafik Hasil <i>Precision Recall</i> Q2 berdasarkan Judul, Isi, Dokumen	91
IV – 17 Grafik Hasil <i>Precision Recall</i> Q3 berdasarkan Judul, Isi, Dokumen	92

DAFTAR TABEL

Tabel	Halaman
II - 1 Tabel akhiran -lah, -kah, -pun	24
II - 2 Tabel akhiran -ku, -mu, -nya	24
II - 3 Tabel awalan 1 / <i>cluster</i> 3	25
II - 4 Tabel awalan 2 / <i>cluster</i> 4	25
II - 5 Tabel akhiran / <i>cluster</i> 5.....	26
II - 6 Tabel pasangan awalan dan akhiran yang tidak sah.....	26
II - 7 Ringkasan Tinjauan Studi	34
III - 1 Tabel Strategi Penelitian 1	39
III - 2 Tabel Strategi Penelitian 2	40
III - 3 Jadwal Penelitian	52
IV - 1 Model Relasi pada table dok	57
IV - 2 Model Relasi pada table bobot judul	57
IV - 3 Model Relasi pada table bobot isi	57
IV - 4 Hasil Bentuk Array Satu Dimensi	58
IV - 5 Hasil <i>Similiarity</i> , <i>Precision</i> dan <i>Recall</i> Q1 Penelusuran berdasarkan Judul	69
IV - 6 Hasil <i>Similiarity</i> , <i>Precision</i> dan <i>Recall</i> Q2 Penelusuran berdasarkan Judul	71
IV - 7 Hasil <i>Similiarity</i> , <i>Precision</i> dan <i>Recall</i> Q3 Penelusuran berdasarkan Judul	72
IV - 8 Hasil <i>Similarity</i> , <i>Precision</i> dan <i>Recall</i> Q1 Penelusuran berdasarkan Isi	74
IV - 9 Hasil <i>Similarity</i> , <i>Precision</i> dan <i>Recall</i> Q2 Penelusuran berdasarkan Isi	76

DAFTAR LAMPIRAN

Lampiran	Halaman
I <i>Stopword</i>	104
II Kartu Bimbingan	108



DAFTAR PUSTAKA

- [Agus 2002] Agus Zainal Arifin, A. N. S. “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.
- [Brasa 2009] Barasa, B. E., “Efektivitas STBI Rekam Medis” Medan: USU Press, 2009.
- [Christopher D 2009] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. “*An Introduction to Information Retrieval*” Cambrigde: Cambridge UP, 2009.
- [David Ellis 1993] David Ellis, J. F.-H. & Willett, P. “*Measuring The Degree of Similarity Between Object in Text Retrieval Systems. Perspective in Information Management*” pp. 128 - 149, 1993.
- [Dipa 2015] Dipa, I. W. S. “Sistem Temu Kembali Informasi atau *Information Retrieval*” 2015 [Online] Available at: http://baliinnetwork.blogspot.com/2015/03/sistem-temu-kembali-informasi-atau_3.html (Diakses 10 May 2015).
- [Fatmawati 2013] Fatmawati, E. “Studi Komparatif Kecepatan Temu Kembali Informasi di Depo Arsip Koran Suara Merdeka Antara Sistem Simpan Manual dengan Foto Repro”, s.l.: s.n, 2013.
- [Hasugian 2006] Hasugian, J. “Penggunaan Bahasa Alamiah dan Kosa Kata Terkendali dalam Sistem Temu Balik Informasi Berbasis Teks” p. 74, 2006.
- [Hasugian 2009] Hasugian, J. “Dasar - Dasar Ilmu Perpustakaan dan Informasi” Medan: USU Press, 2009.