

**PERINGKAS MULTI-DOKUMEN UNTUK BAHASA INDONESIA
MENGUNAKAN TEKNIK CENTROID-BASED SUMMARIZATION DAN
TEKNIK K-MEANS-BASED SUMMARIZATION**



Wisnu Linggacusuma Wardhana

120400092Y

Fakultas Ilmu Komputer

Universitas Indonesia

Depok 2008

HALAMAN PERSETUJUAN

Judul Tugas Akhir

**PERINGKAS MULTI-DOKUMEN UNTUK BAHASA INDONESIA
MENGUNAKAN TEKNIK CENTROID-BASED SUMMARIZATION DAN
TEKNIK K-MEANS-BASED SUMMARIZATION**

Nama: Wisnu Linggacusuma Wardhana

NPM: 120400092Y

Laporan tugas akhir ini telah diperiksa dan disetujui.

Depok, 1 Agustus 2008

Menyetujui,

Mirna Adriani, Ph.D.

Pembimbing Tugas Akhir

KATA PENGANTAR

Penulis mengucapkan puji syukur kepada Allah SWT atas rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan tugas akhir dengan baik. Tugas akhir yang dilakukan penulis tidak mungkin dapat berjalan dengan lancar tanpa bantuan dari berbagai pihak, yang telah memberikan bantuan dan dukungan kepada penulis selama pengerjaan tugas akhir berlangsung. Oleh karena itu, penulis mengucapkan banyak terima kasih kepada:

1. Seluruh anggota keluarga penulis yang selalu mendokan penulis dan memberikan dorongan untuk menyelesaikan tugas akhir ini.
2. Ibu Mirna Adriani, selaku pembimbing tugas akhir penulis, yang telah memberikan arahan dan dorongan untuk melakukan yang terbaik dalam penelitian ini, serta memberikan kesempatan yang luar biasa bagi penulis untuk pertama kalinya mengikuti konferensi internasional.
3. Teman-teman seperjuangan TA yang *subang* dan merupakan *bspkers* sejati yang selalu menjadi pengganggu dan penghambat pengerjaan tugas akhir. Bacub Baspak, Adoen Itam, Jojon, Ikhi.
4. Spesial terima kasih untuk Ibu Ratih Amalia, S.Kom, dan adiknya untuk bantuannya yang sangat-sangat luar biasa di saat-saat kritis akhir pengerjaan tugas akhir ini sampai rela begadang hingga jam 3 subuh.
5. Para teman-teman seperguruan penghuni Lab IR lainnya, Rora, Desmond, Franky, Eliza, Rahmat, Femphy, Ame, Tuti, Lia yang telah menjadi teman sehidup semati selama pengerjaan tugas akhir ini.
6. Teman-teman *subang suke* lainnya Alibun, Piratud, Unyil, Pandu, Kresdut, Makidor, Kemon yang telah memberi kenangan indah dan buruk selama ini.
7. Teman-teman angkatan 2004 tercinta dan tersayang. Aji, Adolf, Moja, Aria, Aryo, Wamir, Gita, Gagang, Botem, Pongo, Angky, Daniel, Richard, Rado, Arfan, Adrianus, Ardi, Hendra, Adit, Hadi, Siheq, Mika, Riza, Jani, Rissa, Cybill, Michael, Reza, Jere, RAP, Wahyu Sulis, Smel, Sawie, Mea, Martin San Wa, dll yang belum disebutkan. Hidup 2004!
8. *Thanks to Ice Frog* dan para pahlawan dota yang di dalam kepalanya hanya ada kata-kata *kriyet*, *join*, dan *Gege*, dan para pengendali *topway* dan *welcom*. Asa,

Kura, Idur, Dede, Fernan, Mu, Abe, Zaki, Baski, Kusut. Juga para penghuni Ristek lainnya dan MIC Aria, Teddy, Pray, Mahdi, Wence, Ical, Herman, Wisnu senior, Fu, Sagi.

9. Para pemain klub futsal paling jago sejagad Fasilkom Raya 3309 FC. Fu, Renggo, Andra, Yewe, Pras.
10. Mang Acep, Pak Wiryo, Pak Sanin, Pak Macho, Mbak Leha dan staff Janitor dan Satpam lainnya, yang sudah membantu penulis selama kuliah di Fasilkom.
11. Laptop Toshiba canggih dengan OS-nya yang *bspk* yang telah dipekerjakan secara paksa dan romusha sehingga penulis dapat menyelesaikan tugas akhir ini.
12. Semua pihak lain yang merasa pantas untuk disebutkan namun belum disebutkan baik sengaja maupun tidak disengaja, karena tidak mungkin disebutkan satu per satu, semoga Allah SWT dapat membalas budi baik Anda sekalian dengan berlipat ganda.

Depok, Agustus 2008

Wisnu Linggokusuma
Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN.....	ii
ABSTRAK.....	iii
KATA PENGANTAR.....	v
DAFTAR ISI.....	vii
DAFTAR GAMBAR.....	ix
DAFTAR GRAFIK.....	x
DAFTAR TABEL.....	xi
BAB 1 PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Tujuan.....	3
1.4 Ruang Lingkup Penelitian.....	3
1.5 Metodologi Penelitian.....	4
1.6 Sistematika Penulisan.....	4
BAB 2 LANDASAN TEORI.....	6
2.1 Perolehan Informasi.....	6
2.2 Sistem Perolehan Informasi.....	7
2.1.1 Pembobotan Kata.....	8
2.1.2 <i>Stopwords</i>	9
2.1.3 Pemotongan Kata Berimbuhan (<i>Stemming</i>).....	10
2.3 Peringkasan Dokumen.....	11
2.3.1 Ringkasan.....	11
2.3.2 <i>Compression Rate</i>	12
2.3.3 Peringkasan Multi-dokumen.....	13
2.3.4 Proses Peringkasan Multi-dokumen.....	14
2.4 Penelitian-penelitian Sebelumnya.....	15
2.4.1 Summarizing Online News Articles (SUMMONS).....	15
2.4.2 MEAD.....	16
2.4.3 Peringkasan Multi-dokumen Multi-bahasa.....	17
2.5 Pengelompokan Dokumen.....	19
2.6 Teknik Peringkasan Multi-dokumen.....	21
2.6.1 Teknik Peringkasan <i>Centroid-based Summarization</i>	21
2.6.2 Teknik Peringkasan <i>K-means-based Summarization</i>	28
2.7 Peringkasan Multi-dokumen untuk Dokumen Berbahasa Indonesia.....	35
2.7.1 Proses Peringkasan Multi-dokumen untuk Dokumen Berbahasa Indonesia.....	35
2.7.2 Evaluasi Hasil Ringkasan.....	37
BAB 3 EKSPERIMEN.....	41
3.1 Jenis Data.....	41
3.1.1 Koleksi dokumen.....	41

3.1.2	Daftar <i>Stopwords</i> Bahasa Indonesia	43
3.1.3	Ringkasan Referensi	43
3.1.4	Penilaian Juri.....	44
3.2	Aplikasi yang Digunakan dalam Eksperimen	45
3.2.1	Pemotong Kata Berimbuhan Indonesia	45
3.2.2	Lemur <i>Toolkit</i>	45
3.3	Skenario Eksperimen.....	46
3.3.1	Eksperimen Peringkasan Multi-dokumen untuk Dokumen Berbahasa Indonesia dengan Teknik <i>Centroid-based Summarization</i>	47
3.3.2	Eksperimen Peringkasan Multi-dokumen untuk Dokumen Berbahasa Indonesia dengan Teknik <i>K-means-based Summarization</i>	50
BAB 4	HASIL DAN ANALISIS	53
4.1	Hasil dan Analisis Eksperimen Teknik Peringkasan <i>Centroid-based Summarization</i>	53
4.1.1	Hasil Peringkasan Teknik <i>Centroid-based Summarization</i> dengan Metode Pengevaluasian Perbandingan terhadap Ringkasan Referensi ..	53
4.1.2	Analisis Hasil Peringkasan Teknik <i>Centroid-based Summarization</i> dengan Metode Pengevaluasian Perbandingan terhadap Ringkasan Referensi	55
4.1.3	Hasil Peringkasan Teknik <i>Centroid-based Summarization</i> dengan Metode Pengevaluasian <i>Interjudge Agreement</i>	58
4.1.4	Analisis Hasil Peringkasan Teknik <i>Centroid-based Summarization</i> dengan Metode Pengevaluasian <i>Interjudge Agreement</i>	59
4.2	Hasil dan Analisis Eksperimen Teknik Peringkasan <i>K-means-based Summarization</i>	61
4.2.1	Hasil Peringkasan Teknik <i>K-means-based Summarization</i> dengan Metode Pengevaluasian Perbandingan terhadap Ringkasan Referensi ..	62
4.2.2	Analisis Hasil Peringkasan Teknik <i>K-means-based Summarization</i> dengan Metode Pengevaluasian Perbandingan terhadap Ringkasan Referensi	63
4.2.3	Hasil Peringkasan Teknik <i>K-means-based Summarization</i> dengan Metode Pengevaluasian <i>Interjudge Agreement</i>	65
4.2.4	Analisis Hasil Peringkasan Teknik <i>Centroid-based Summarization</i> dengan Metode Pengevaluasian <i>Interjudge Agreement</i>	66
4.3	Perbandingan Hasil Peringkasan Teknik <i>Centroid-based Summarization</i> dan <i>K-means-based Summarization</i>	68
BAB 5	PENUTUP.....	72
5.1	Kesimpulan.....	72
5.2	Saran.....	74
	DAFTAR PUSTAKA	75
	LAMPIRAN A Hasil Ringkasan Teknik <i>Centroid-based Summarization</i>	77
	LAMPIRAN B Hasil Ringkasan Teknik <i>K-means-based Summarization</i>	84

DAFTAR GAMBAR

Gambar 2.1 Proses perolehan informasi sederhana	8
Gambar 2.2 Proses peringkasan multi-dokumen	14
Gambar 2.3 Pengelompokan <i>K-means</i> (sumber: http://en.wikipedia.org/wiki/K-means_algorithm)	20
Gambar 2.4 Peringkasan multi-dokumen menggunakan teknik <i>centroid-based summarization</i>	22
Gambar 2.5 Pengurutan dokumen dan kalimat	28
Gambar 2.6 Peringkasan multi-dokumen menggunakan teknik <i>k-means-based summarization</i>	29
Gambar 2.7 Proses peringkasan multi-dokumen untuk dokumen berbahasa Indonesia	36
Gambar 3.1 Contoh format standar dokumen	41
Gambar 3.2 Skenario eksperimen	47
Gambar 3.3 Proses peringkasan <i>centroid-based summarization</i> untuk dokumen berbahasa Indonesia	48
Gambar 3.4 Proses peringkasan <i>k-means-based summarization</i> untuk dokumen berbahasa Indonesia	52

DAFTAR GRAFIK

Grafik 4.1 Hasil evaluasi teknik peringkasan <i>centroid-based summarization</i> dengan metode perbandingan terhadap ringkasan referensi.....	55
Grafik 4.2 Nilai rata-rata penilaian ketiga orang juri terhadap ringkasan yang dihasilkan oleh teknik <i>centroid-based summarization</i>	61
Grafik 4.3 Hasil evaluasi teknik peringkasan <i>k-means-based summarization</i> dengan metode perbandingan terhadap ringkasan referensi.....	63
Grafik 4.4 Pengaruh besar kelompok dokumen terhadap nilai kualitas ringkasan.....	64
Grafik 4.5 Nilai rata-rata penilaian juri terhadap ringkasan yang dihasilkan oleh teknik <i>k-means-based summarization</i>	67
Grafik 4.6 Perbandingan hasil evaluasi dengan metode perbandingan ringkasan referensi terhadap peringkasan teknik <i>centroid-based summarization</i> dan <i>k-means-based summarization</i>	69
Grafik 4.7 Perbandingan hasil evaluasi dengan metode <i>interjudge agreement</i> terhadap peringkasan teknik <i>centroid-based summarization</i> dan <i>k-means-based summarization</i>	70

DAFTAR TABEL

Tabel 4.1 Hasil evaluasi peringkasan <i>centroid-based summarization</i> dengan metode perbandingan terhadap ringkasan referensi.....	54
Tabel 4.2 <i>Interjudge agreement</i> pada hasil peringkasan <i>centroid-based summarization</i> pada 10% <i>compression rate</i>	58
Tabel 4.3 <i>Interjudge agreement</i> pada hasil peringkasan <i>centroid-based summarization</i> pada 20% <i>compression rate</i>	59
Tabel 4.4 Hasil evaluasi peringkasan <i>k-means-based summarization</i> dengan metode perbandingan terhadap ringkasan referensi.....	62
Tabel 4.5 <i>Interjudge agreement</i> pada hasil peringkasan <i>k-means-based summarization</i> pada 20% <i>compression rate</i>	66
Tabel 4.6 Perbandingan hasil evaluasi dengan metode perbandingan ringkasan referensi terhadap peringkasan teknik <i>centroid-based summarization</i> dan <i>k-means-based summarization</i>	69
Tabel 4.7 Perbandingan hasil evaluasi dengan metode <i>interjudge agreement</i> terhadap peringkasan teknik <i>centroid-based summarization</i> dan <i>k-means-based summarization</i>	70
Tabel A.1 Kelompok dokumen A: Flu Burung di Indonesia.....	77
Tabel A.2 Kelompok dokumen B: Italia Juara Dunia 2006.....	78
Tabel A.3 Kelompok dokumen C: Manchester United Mengalahkan AS Roma 7-1 di Liga Champions	78
Tabel A.4 Kelompok dokumen D: Pencemaran Merkuri di Teluk Buyat Manado	79
Tabel A.5 Kelompok dokumen E: Mantan Presiden Soeharto Meninggal.....	80
Tabel A.6 Kelompok dokumen F: Tsunami di Aceh.....	80
Tabel A.7 Kelompok dokumen G: Manchester United Juara Liga Champions	82
Tabel B.1 Kelompok dokumen A: Flu Burung di Indonesia.....	84
Tabel B.2 Kelompok dokumen B: Italia Juara Dunia 2006.....	85
Tabel B.3 Kelompok dokumen C: Manchester United Mengalahkan AS Roma 7-1 di Liga Champions	85
Tabel B.4 Kelompok dokumen D: Pencemaran Merkuri di Teluk Buyat Manado	86
Tabel B.5 Kelompok dokumen E: Mantan Presiden Soeharto Meninggal.....	87
Tabel B.6 Kelompok dokumen F: Tsunami di Aceh	88
Tabel B.7 Kelompok dokumen G: Manchester United Juara Liga Champions	89