



UNIVERSITAS INDONESIA

**SEGMENTASI DOKUMEN BAHASA INDONESIA
MENGUNAKAN METODE *GENETIC ALGORITHM***

SKRIPSI

VINKY HALIM

1205000916

**FAKULTAS ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER**

DEPOK

JULI 2009



UNIVERSITAS INDONESIA

**SEGMENTASI DOKUMEN BAHASA INDONESIA
MENGUNAKAN METODE *GENETIC ALGORITHM***

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar S.Kom

VINKY HALIM

1205000916

**FAKULTAS ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER
DEPOK
JULI 2009**

HALAMAN PERNYATAAN ORISINALITAS

Skripsi ini adalah hasil karya sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar.

Nama : Vinky Halim

NPM : 1205000916

Tanda Tangan :

Tanggal : 19 Juli 2009

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama : Vinky Halim

NPM : 1205000916

Program Studi : Ilmu Komputer

Judul Skripsi : Segmentasi Dokumen Bahasa Indonesia Menggunakan Metode
Genetic Algorithm

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia

DEWAN PENGUJI

Pembimbing : Dr. Hisar Maruli Manurung, S.Kom. (.....)

Penguji : Adila Alfa Krisnadhi, S.Kom., M.Sc. (.....)

Penguji : Dr. Ade Azurat, S.Kom. (.....)

Ditetapkan di : Fakultas Ilmu Komputer

Tanggal : 16 Juli 2009

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas berkat, karunia dan penyertaan-Nya sehingga penulis dapat menyelesaikan pembuatan tugas akhir ini dengan sebaik-baiknya dalam jangka waktu yang telah ditentukan. Selama proses penyelesaian tugas akhir ini, penulis banyak memperoleh bantuan dan masukan dari berbagai pihak yang sangat berarti bagi penulis. Pada kesempatan ini penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya terutama kepada:

1. Orang tua, kakak, dan seluruh anggota keluarga lainnya yang telah memberikan doa, semangat, dukungan, dan perhatian selama penulis menyelesaikan tugas akhir ini.
2. Bapak Hisar Maruli Manurung selaku dosen pembimbing tugas akhir.
3. Bapak Achmad Nizar Hidayanto selaku dosen pembimbing akademis.
4. Herliani Iskandar Setiawan dan keluarga yang telah banyak memberikan dukungan.
5. Rekan-rekan di Lab IR yang telah banyak memberikan masukan-masukan yang sangat membantu selama penulis menyelesaikan pembuatan tugas akhir ini.
6. Adhitya N.R dan Teddy Wijaya yang telah banyak membantu memberikan ide-ide bermanfaat selama pengerjaan tugas akhir.
7. Ananta D.P, Bambang Adhi, Darwin Cuputra, Hansel Tanuwijaya, dan Refly Harold Hadiwijaya sebagai kelompok bermain DOTA yang telah menghentikan kegiatan secara sukarela selama pengerjaan tugas akhir ini.
8. Bayu Distiawan, Suryanto Ang, Prajna Wira, dan Armando Yonathan sebagai kelompok bermain bulutangkis yang telah membantu memberikan penyegaran selama pengerjaan tugas akhir ini.
9. Bernadia Puspasari, Clara Vania Kitt, Evi D.J, Yenni Novriani, Mursal Rais, Haryadi Herdian, Nur Aisyah, dan Octo Alexandro sebagai kelompok lab 09 yang banyak memberikan inspirasi dan semangat selama pengerjaan tugas akhir ini.

10. Angkatan 2005, yang selama 4 tahun ini sudah berjuang bersama penulis untuk menyelesaikan studi di Fakultas Ilmu Komputer ini.
11. Anggota KUKSA CSUI, yang telah menemani dalam suka dan duka.
12. Seluruh teman-teman yang tidak dapat disebutkan satu persatu, terima kasih atas segala bantuannya.

Penulis sangat sadar bahwa dalam pengerjaan tugas akhir ini, masih banyak kekurangan dan kesalahan yang penulis lakukan. Oleh karena itu, penulis sangat terbuka untuk menerima setiap kritik dan saran yang membangun yang berkaitan dengan tugas akhir ini. Semoga laporan tugas akhir ini dapat bermanfaat bagi para pembaca dimasa yang akan datang.



Depok, 19 Juli 2009

Vinky Halim

**HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS
AKHIR UNTUK KEPENTINGAN AKADEMIS**

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan dibawah ini:

Nama : Vinky Halim

NPM : 1205000916

Program Studi : Ilmu Komputer

Departemen : Ilmu Komputer

Fakultas : Ilmu Komputer

Jenis Karya : Skripsi

demikian pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (*Non-Exclusive Royalty Free Right*)** atas karya ilmiah saya yang berjudul:

Segmentasi Dokumen Bahasa Indonesia Menggunakan Metode *Genetic Algorithm*

berserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya:

Dibuat di : Depok

Pada Tanggal : 19 Juli 2009

Yang menyatakan

(Vinky Halim)

DAFTAR ISI

HALAMAN JUDUL.....	i
HALAMAN PERNYATAAN ORISINALITAS.....	ii
HALAMAN PENGESAHAN.....	iii
KATA PENGANTAR	iv
HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS	vi
ABSTRAK	vii
ABSTRACT.....	viii
DAFTAR ISI.....	ix
DAFTAR TABEL.....	xii
DAFTAR GAMBAR	xiii
BAB 1 PENDAHULUAN	1
1.1 LATAR BELAKANG PENELITIAN	1
1.2 PERUMUSAN MASALAH.....	2
1.3 TUJUAN PENELITIAN	2
1.4 RUANG LINGKUP PENELITIAN	2
1.5 METODOLOGI PENELITIAN	3
1.6 SISTEMATIKA PENULISAN	3
BAB 2 LANDASAN TEORI.....	6
2.1 SEGMENTASI DOKUMEN	6
2.1.1 TOPIC DETECTION AND TRACKING (TDT).....	8
2.1.2 TEXTTILING.....	9
2.1.3 PENELITIAN SEGMENTASI DOKUMEN TEKS BAHASA INDONESIA MENGGUNAKAN METODE TEXTTILING	11
2.2 GENETIC ALGORITHM.....	12
2.2.1 PENELITIAN SEGGEN	15
2.2.2 MULTIOBJECTIVE PROBLEM AND OPTIMIZATION	20
2.2.3 STRENGTH PARETO EVOLUTIONARY ALGORITHM (SPEA) ..	22
2.2.4 STRENGTH PARETO EVOLUTIONARY ALGORITHM 2 (SPEA 2)	25
BAB 3 PERANCANGAN	28

3.1	GAMBARAN UMUM PROSES SEGMENTASI DOKUMEN	28
3.2	DATA.....	30
3.2.1	PENGOLAHAN DATA	31
3.2.2	PEMBUATAN TEST CASE.....	32
3.3	SEGMENTASI DOKUMEN DENGAN GENETIC ALGORITHM	33
3.3.1	PARAMETER GENETIC	34
3.3.2	VARIABEL EKSPERIMENTASI	36
3.4	PERBANDINGAN ANTARA METODE GENETIC ALGORITHM DAN METODE TEXTTILING.....	38
3.5	METODE PENGUKURAN TINGKAT AKURASI	38
BAB 4	IMPLEMENTASI.....	40
4.1	PENGOLAHAN DATA.....	40
4.1.1	PENGGABUNGAN DOKUMEN.....	40
4.1.2	PEMBUANGAN <i>STOPWORD</i> , <i>STEMMING</i> , DAN PEMISAHAN ANTAR KALIMAT	41
4.2	ECJ 18	46
4.2.1	REPRESENTASI INDIVIDU/ <i>CHROMOSOME</i> DAN <i>GENETIC OPERATOR</i>	46
4.2.2	<i>FITNESS FUNCTION</i>	48
BAB 5	HASIL DAN PEMBAHASAN.....	52
5.1	RAGAM ASPEK PEMBAHASAN HASIL SEGMENTASI DOKUMEN	52
5.2	HASIL SEGMENTASI DOKUMEN DITINJAU DARI ASPEK <i>FITNESS FUNCTION</i>	55
5.3	HASIL SEGMENTASI DOKUMEN DITINJAU DARI ASPEK METODE PENGHITUNGAN <i>SIMILARITY</i>	58
5.4	HASIL SEGMENTASI DOKUMEN DITINJAU DARI ASPEK UKURAN POPULASI DAN JUMLAH ITERASI YANG DIGUNAKAN.....	60
5.5	HASIL SEGMENTASI DOKUMEN UNTUK MEMBANDINGKAN <i>FITNESS FUNCTION</i> KOMBINASI LINEAR 1:8 DAN SPEA 2 DENGAN AGGREGASI 1:8.....	63
5.6	HASIL SEGMENTASI DOKUMEN DITINJAU DARI ASPEK <i>GENETIC OPERATOR</i> YANG DIGUNAKAN	65

5.7 HASIL SEGMENTASI DOKUMEN DITINJAU DARI ASPEK BANYAKNYA JUMLAH SEGMENT.....	68
5.8 HASIL SEGMENTASI DOKUMEN DITINJAU DARI ASPEK KEMIRIPAN ANTAR DOKUMEN PENYUSUN	70
5.9 PERBANDINGAN HASIL SEGMENTASI DENGAN METODE <i>GENETIC ALGORITHM</i> DAN HASIL SEGMENTASI DENGAN METODE <i>TEXTTILING</i>	73
5.10 ANALISA HASIL.....	76
5.11 RANGKUMAN HASIL.....	79
BAB 6 PENUTUP	82
6.1 KESIMPULAN	82
6.2 KENDALA.....	83
6.3 SARAN	83
DAFTAR PUSTAKA	85
LAMPIRAN 1: DAFTAR STOPWORD.....	87
LAMPIRAN 2: CONTOH TEST CASE TIDAK MIRIP.....	88
LAMPIRAN 3 : CONTOH TEST CASE MIRIP	89
LAMPIRAN 4 : CONTOH TEST CASE SANGAT MIRIP.....	90

DAFTAR TABEL

Tabel 5.1 Tabel rangkuman keseluruhan variabel percobaan.....	53
Tabel 5.2 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek <i>fitness function</i>	55
Tabel 5.3 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek metode penghitungan <i>similarity</i>	58
Tabel 5.4 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek jumlah iterasi	60
Tabel 5.5 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek ukuran populasi	61
Tabel 5.6 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan perbandingan <i>fitness function</i> antara kombinasi linear 1:8 dan SPEA 2	64
Tabel 5.7 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek probabilitas mutasi	66
Tabel 5.8 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek tipe <i>crossover</i>	66
Tabel 5.9 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek banyaknya jumlah segmen	69
Tabel 5.10 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan ditinjau dari aspek kemiripan antar dokumen penyusun	71
Tabel 5.11 Tabel nilai rata-rata <i>precision</i> dan <i>recall</i> hasil percobaan perbandingan antara metode <i>genetic algorithm</i> dengan metode Textiling.....	74

DAFTAR GAMBAR

Gambar 2.1	Contoh dokumen yang belum tersegmen	7
Gambar 2.2	Contoh dokumen yang sudah tersegmen.....	8
Gambar 2.3	Contoh proses <i>crossover</i>	14
Gambar 2.4	Contoh proses mutasi	14
Gambar 2.5	<i>Pseudocode genetic algorithm</i>	15
Gambar 2.6	Ilustrasi <i>multiobjective problem</i> dengan <i>pareto optimality</i> pada domain fungsi objektif	21
Gambar 2.7	<i>Pseudocode</i> SPEA (Zitzler, 1999)	23
Gambar 2.8	Alur proses SPEA (Zitzler, 1999)	24
Gambar 2.9	<i>Pseudocode</i> penghitungan <i>fitness score</i> (Zitzler, 1999).....	25
Gambar 2.10	<i>Pseudocode</i> SPEA 2 (Zitzler, 2004)	27
Gambar 3.1	Gambaran umum proses segmentasi yang dilakukan	30
Gambar 3.2	Contoh representasi individu/ <i>chromosome</i> yang digunakan.....	35
Gambar 4.1	<i>Pseudocode</i> untuk proses <i>append</i> dokumen.....	41
Gambar 4.2	Contoh kata-kata yang termasuk <i>stopword</i>	42
Gambar 4.3	Contoh <i>output</i> proses <i>stemming</i>	43
Gambar 4.4	Contoh kegunaan proses <i>stemming</i> pada penghitungan nilai <i>similarity</i>	44
Gambar 4.5	Contoh dokumen setelah dilakukan proses pemisahan antar kalimat	44
Gambar 4.6	<i>Pseudocode</i> untuk proses pengolahan dokumen.....	45
Gambar 4.7	Contoh representasi individu pada penelitian	47
Gambar 4.8	<i>Single point crossover</i>	48

Gambar 4.9 <i>Two point crossover</i>	48
Gambar 4.10 <i>Pseudocode</i> untuk proses pembobotan kata dengan TF-IDF.....	49
Gambar 4.11 <i>Pseudocode</i> untuk proses pembobotan kata dengan TF-IDF (<i>Word Frequency</i>)	49
Gambar 4.12 <i>Pseudocode</i> penghitungan <i>similarity</i> dengan <i>cosine similarity</i>	50
Gambar 4.13 <i>Pseudocode</i> penghitungan <i>similarity</i> dengan <i>dice coefficient</i>	50
Gambar 4.14 <i>Pseudocode</i> penghitungan <i>internal cohesion</i> dan <i>dissimilarity</i>	51
Gambar 5.1 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek <i>fitness function</i>	56
Gambar 5.2 Gambar nilai rata-rata <i>recall</i> hasil percobaan ditinjau dari aspek <i>fitness function</i>	57
Gambar 5.3 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek metode penghitungan <i>similarity</i>	59
Gambar 5.4 Gambar nilai rata-rata <i>recall</i> hasil percobaan ditinjau dari aspek metode penghitungan <i>similarity</i>	59
Gambar 5.5 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek jumlah iterasi.....	61
Gambar 5.6 Gambar nilai rata-rata <i>recall</i> hasil percobaan ditinjau dari aspek jumlah iterasi.....	61
Gambar 5.7 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek ukuran populasi	62
Gambar 5.8 Gambar nilai rata-rata <i>recall</i> hasil percobaan ditinjau dari aspek ukuran populasi	62
Gambar 5.9 Gambar nilai rata-rata <i>precision</i> hasil percobaan perbandingan <i>fitness function</i> antara kombinasi linear 1:8 dan SPEA 2	64

Gambar 5.10 Gambar nilai rata-rata <i>recall</i> hasil percobaan perbandingan <i>fitness function</i> antara kombinasi linear 1:8 dan SPEA 2	64
Gambar 5.11 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek probabilitas mutasi	66
Gambar 5.12 Gambar nilai rata-rata <i>recall</i> hasil percobaan ditinjau dari aspek probabilitas mutasi	67
Gambar 5.13 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek tipe <i>crossover</i>	67
Gambar 5.14 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek tipe <i>crossover</i>	68
Gambar 5.15 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek banyaknya jumlah segmen	69
Gambar 5.16 Gambar nilai rata-rata <i>recall</i> hasil percobaan ditinjau dari aspek banyaknya jumlah segmen	70
Gambar 5.17 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek kemiripan antar dokumen penyusun	72
Gambar 5.18 Gambar nilai rata-rata <i>precision</i> hasil percobaan ditinjau dari aspek kemiripan antar dokumen penyusun	72
Gambar 5.19 Gambar nilai rata-rata <i>precision</i> hasil percobaan perbandingan antara metode <i>genetic algorithm</i> dengan metode Textiling	74
Gambar 5.20 Gambar nilai rata-rata <i>recall</i> hasil percobaan perbandingan antara metode <i>genetic algorithm</i> dengan metode Textiling	75
Gambar 5.21 <i>Input</i> yang digunakan	77
Gambar 5.22 Hasil segmentasi <i>input</i> dengan metode SPEA 2 agregasi 1:8	77
Gambar 5.23 Letak segmen sebenarnya dari <i>input</i> (memisahkan antara topik properti dengan topik olahraga tenis)	78