ABSTRAK

Sistem pengecekan kemiripan dokumen tugas akhir pada Fakultas Teknologi Informasi Universitas Kristen Maranatha masih dilakukan secara manual sehingga mahasiswa dapat meniru dokumen tugas akhir orang lain. Untuk membantu para dosen mendeteksi peniruan dokumen tugas akhir, dibutuhkan sebuah sistem pendeteksian kemiripan judul dan isi dokumen. Sistem ini dibuat menggunakan konsep *data mining*, dimana pengecilan dimensi, penghilangkan gangguan, dan perestrukturisasi masukan akan dilakukan. Sistem akan memroses masukan yang telah diekstrak untuk memperhitungkan nilai kemiripannya dengan masukan yang lain. Pertama-tama, sistem mengambil semua data dokumen tugas akhir dari basis data dan satu dokumen pilihan pengguna. Kemudian, sistem melakukan proses ekstraksi dokumen terhadap masukan berupa dokumen melalui *case folding* dan *tokenisasi*, *filtering*, dan *stemming* (menggunakan algoritma Nazief-Adriani) pada setiap dokumen. Setelah proses ekstraksi selesai, sistem dapat memroses dokumen untuk memperhitungkan nilai kemiripannya dengan dua cara, yaitu *cosine similarity* dan algoritma Smith-Waterman. Dengan *cosine similarity*, sistem menghitung jumlah kata-kata pada dokumen pilihan pengguna dan dokumen-dokumen lainnya, kemudian mengalkulasi nilai kemiripannya dengan hukum kosinus. Dengan algoritma Smith-Waterman, sistem membandingkan urutan kata-kata dari dokumen pilihan pengguna dan dokumen-dokumen lainnya, kemudian menghasilkan nilai kemiripannya. Pada akhirnya, sistem dapat menghasilkan angka dari 0% hingga 100% untuk setiap dokumen. Sistem menampilkan dokumen-dokumen yang terurut dari nilai kemiripan terbesar hingga terkecil.

Kata kunci: algoritma Smith-Waterman, algoritma Nazief-Adriani, *cosine similarity*, *data mining*, dokumen tugas akhir, nilai kemiripan

ABSTRACT

The Information Technology of Maranatha Christian University's documents of final term paper checking system is still run manually so that students can copy others' document of final term paper. To help lecturers detecting document of final term paper copying, a similarity detecting documents of final term paper is needed. This system is made using data mining concept, where dimension reduction, noise removal, and input restructuring will be implemented. This system will process extracted input to count its similarity value compared to other inputs. First off, system fetches all documents of final term paper from database and one user-chosen document. Then, system does document extraction process toward the input in form of a document through case folding and tokenization, filtering, and stemming (using Nazief-Adriani algorithm) to every document. After the extraction process completed, system can process the user-chosen document to count its similarity value in two ways: cosine similarity and Smith-Waterman algorithm. With cosine similarity, system counts the number of words of user-chose document and other documents, and then calculates its similarity value with cosines law. Using Smith-Waterman algorithm, system compares words sequence from user-chose document and other documents, and then outputs its similarity value. In the end, system will produce percentage from 0% to 100% for every document. System shows sorted documents descendingly based on its similarity value.

Keywords: cosine similarity, data mining, document of final term, Nazief-Adriani algorithm, similarity value, Smith-Waterman algorithm

DAFTAR ISI

LEMBAR PENGESAHAN i
PERNYATAAN ORISINALITAS LAPORAN PENELITIAN ii
PERNYATAAN PUBLIKASI LAPORAN PENELITIAN iii
PRAKATA iv
ABSTRAK v
ABSTRACT vi
DAFTAR ISI vii
DAFTAR GAMBAR x
DAFTAR TABEL xi
DAFTAR LAMPIRAN xii
BAB I PENDAHULUAN 1
1.1 Latar Belakang Masalah 1
1.2 Rumusan Masalah 1
1.3 Tujuan Pembahasan 2
1.4 Ruang Lingkup Kajian 2
1.5 Sumber Data 2
1.6 Sistematika Penyajian 3
BAB II KAJIAN TEORI 4
2.1 Kajian Teori Sistem Informasi 4
2.2 *Entity Relationship Diagram* (ERD) 5
2.3 *Unified Modelling Language* (UML) 9
2.3.1 *Use Case Diagram* 9
2.3.2 *Class Diagram* 9
2.3.3 *Activity Diagram* 10
2.4 Bagan Alir (*Flowchart*) 11
2.4.1 Bagan Alir Sistem 11
2.4.2 Bagan Alir Dokumen 14
2.4.3 Bagan Alir Skematik 14
2.4.4 Bagan Alir Program 14
2.4.5 Bagan Alir Proses 14
2.5 Basis Data 15
2.6 *Structured Query Language* (SQL) 15
2.7 PHP 16
2.8 Pemrograman Berorientasi Objek dalam PHP 17
2.9 MySQL 19
2.10 PHP Designer 2007 20
2.11 XAMPP 20
2.12 Plagiarisme 20
2.13 *Text Mining* 21
2.13.1 *Infomation Retrieval* (IR) 21
2.13.2 *Natural Language Processing* (NLP) 22
2.13.3 Fungsi *Text Mining* 22
2.14 Vektor 23
2.14.1 Istilah-istilah Vektor 24
2.14.1.1 *Vector Length* (Panjang Vektor) 24
2.14.1.2 *Vector Addition* (Penambahan Vektor) 24
2.14.1.3 *Scalar Multiplication* (Perkalian Skalar) 25
2.14.1.4 *Inner Product* 25

2.14.1.5 *Orthogonality* (Ortogonalitas) 25
2.14.1.6 *Normal Vector* (Vektor Normal) 26
2.14.1.7 *Orthonormal Vector* (Vektor Ortonormal) 26
2.14.1.8 Proses *Gram-Schmidt Orthonormalization* 27
2.15 Matriks 28
2.15.1 Notasi Matriks 29
2.15.2 Istilah Matriks 30
2.15.2.1 *Square Matrix* 30
2.15.2.2 *Transpose* 30
2.15.2.3 *Matrix Multiplication* (Perkalian Matriks) 31
2.15.2.4 *Identity Matrix* (Matriks Identitas) 32
2.15.2.5 *Orthogonal Matrix* (Matriks Ortogonal) 32
2.15.2.6 *Diagonal Matrix* (Matriks Diagonal) 33
2.15.2.7 *Determinant* (Determinan) 33
2.15.2.8 *Eigenvector* dan *Eigenvalues* 34
2.16 Ekstraksi Dokumen 36
2.16.1 *Case Folding* dan *Tokenizing* 37
2.16.2 *Filtering* 37
2.16.3 *Stemming* 38
2.17 Algoritma Nazief-Adriani 38
2.18 *Latent Semantic Indexing* (LSI) 42
2.18.1 *Singular Value Decomposition* (SVD) 43
2.18.1.1 Contoh SVD Penuh 43
2.18.1.2 SVD Tereduksi (*Reduced* SVD) 49
2.18.1.3 Contoh SVD Tereduksi 50
2.19 Algoritma Smith-Waterman 52
BAB III ANALISIS DAN RANCANGAN SISTEM 55
3.1 Proses Bisnis 55
3.1.1 Pengecekan Judul Dokumen 55
3.1.2 Pengecekan Isi Dua Dokumen 56
3.1.3 Ekstraksi Dokumen 58
3.1.4 *Case Folding* 59
3.1.5 *Tokenizing* 60
3.1.6 *Filtering* 60
3.1.7 *Stemming* dengan Algoritma Nazief-Adriani 62
3.1.8 Cek Kamus 64
3.1.9 *Delete Inflection Suffixes* 65
3.1.10 Cek *Rule Precedence* 67
3.1.11 Cek *Prefix Disallowed Suffixes* 68
3.1.12 *Delete Derivation Suffixes* 69
3.1.13 *Delete Derivation Prefixes* 71
3.1.14 Proses Pemotongan Kata 73
3.2 *Entity Relational Diagram* (ERD) 75
3.2.1 Transformasi ERD 75
3.3 *Use Case Diagram* 78
3.3.1 Use Case Scenario 78
3.4 *Class Diagram* 80
3.5 *Activity Diagram* 80
3.6 Perancangan Sketsa *User Interface* 81
BAB IV HASIL PENELITIAN 86

4.1 Tampilan untuk Perhitungan Kemiripan Judul Dokumen 86
4.1.1 Halaman Beranda (Dosen) 86
4.1.2 Halaman Pengecekan Kelayakan Topik 87
4.1.3 Halaman Pengecekan Topik.....88
4.2 Tampilan untuk Perhitungan Kemiripan Isi Dua Dokumen.....89
4.2.1 Halaman Perbandingan Dua Dokumen.....89
4.2.2 Halaman Hasil Perbandingan Dua Dokumen.....90
4.3 Tampilan untuk Perhitungan Kemiripan Semua Judul 91
4.3.1 Halaman Perhitungan Kemiripan Semua Judul.....91
4.3.2 Halaman Hasil Perhitungan Kemiripan Judul.....92
BAB V PEMBAHASAN DAN UJI COBA HASIL PENELITIAN.....93
BAB VI SIMPULAN DAN SARAN 95
6.1 Simpulan.....95
6.2 Saran 95
DAFTAR PUSTAKA.....xi
RIWAYAT HIDUP PENULIS xiii

DAFTAR GAMBAR

Gambar 1 Entitas Kuat.....5
Gambar 2 Entitas Lemah.....5
Gambar 3 Atribut Komposit.....6
Gambar 4 Atribut Bernilai Banyak.....6
Gambar 5 Atribut Turunan.....6
Gambar 6 Relasi.....7
Gambar 7 Aktor.....9
Gambar 8 Tahap *Preprocessing*.....37
Gambar 9 *Tokenizing*.....37
Gambar 10 *Filtering*.....38
Gambar 11 *Stemming*.....38
Gambar 12 *Flowchart* Pengecekan Judul Dokumen.....56
Gambar 13 *Flowchart* Pengecekan Isi Dua Dokumen.....57
Gambar 14 *Flowchart* Ektraksi Dokumen.....58
Gambar 15 *Flowchart Case Folding*.....59
Gambar 16 *Flowchart Tokenizing*.....60
Gambar 17 *Flowchart Filtering*.....61
Gambar 18 *Flowchart Stemming* dengan Algoritma Nazief-Adriani.....63
Gambar 19 *Flowchart* Cek Kamus.....64
Gambar 20 *Flowchart Delete Inflection Suffixes*.....66
Gambar 21 *Flowchart* Cek *Rule Precedence*.....67
Gambar 22 *Flowchart* Cek *Prefix Disallowed Suffixes*.....68
Gambar 23 *Flowchart Delete Derivation Suffixes*.....70
Gambar 24 *Flowchart* Proses Pemotongan Kata.....74
Gambar 25 *Use Case Diagram*.....78
Gambar 26 *Activity Diagram* Pengecekan Kemiripan Judul Dokumen.....80
Gambar 27 *Activity Diagram* Pengecekan Kemiripan Isi Dua Dokumen.....81
Gambar 28 Sketsa Halaman Beranda Dosen.....82
Gambar 29 Sketsa Halaman Pengecekan Kelayakan Topik.....82
Gambar 30 Sketsa Halaman Pengecekan Topik.....83
Gambar 31 Sketsa Halaman Perbandingan Dua Dokumen.....84
Gambar 32 Sketsa Halaman Hasil Perbandingan Dua Dokumen.....84
Gambar 33 Sketsa Halaman Perhitungan Kemiripan Semua Judul.....85
Gambar 34 Sketsa Halaman Hasil Perhitungan Kemiripan Semua Judul.....85
Gambar 35 Halaman Beranda (Dosen).....86
Gambar 36 Halaman Pengecekan Kelayakan Topik.....87
Gambar 37 Halaman Pengecekan Topik.....88
Gambar 38 Halaman Perbandingan Dua Dokumen.....89
Gambar 39 Halaman Hasil Perbandingan Dua Dokumen.....90
Gambar 40 Halaman Perhitungan Kemiripan Semua Judul.....91
Gambar 41 Halaman Hasil Perhitungan Kemiripan Judul.....92

DAFTAR TABEL

Tabel I Simbol *Flowchart* 12
Tabel II 1997 *Fitness International Scorecard*. *Source*: Muscle & Fitness July 1997, p.139 28
Tabel III Kata × Dokumen Terhadap Beberapa Dokumen Buatan.....30
Tabel IV Kombinasi Prefiks dan Sufiks yang Tidak Diizinkan.....40
Tabel V Contoh Tabel Kata × Dokumen.....50
Tabel VI Tabel Pengguna 75
Tabel VII Tabel Mahasiswa.....75
Tabel VIII Tabel Dosen 75
Tabel IX Tabel Jabatan.....75
Tabel X Tabel Topik.....76
Tabel XI Tabel BelumLulus.....76
Tabel XII Tabel SudahLulus.....76
Tabel XIII Tabel PengumpulanTopik.....76
Tabel XIV Tabel Revisi 77
Tabel XV Tabel Semester.....77
Tabel XVI Tabel DosenMengurusTopik 77
Tabel XVII Tabel Stoplist 77
Tabel XVIII Tabel Katadasar.....77
Tabel XIX Uji Kasus Perhitungan Kemiripan Judul Dokumen.....93
Tabel XX Uji Kasus Perhitungan Kemiripan Isi Dua Dokumen 93
Tabel XXI Uji Kasus Kemiripan Semua Judul Dokumen 94

DAFTAR LAMPIRAN

LAMPIRAN A A-1
LAMPIRAN B B-1
LAMPIRAN C C-1
LAMPIRAN D D-1