



### **Hak cipta dan penggunaan kembali:**

Lisensi ini mengizinkan setiap orang untuk mengubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

### **Copyright and reuse:**

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

**IMPLEMENTASI ALGORITMA RABIN-KARP DAN  
STEMMING CONFIX-STRIPPING PADA SISTEM  
PENDETEKSI KEMIRIPAN DOKUMEN TEKS**

**SKRIPSI**

**Diajukan sebagai salah satu syarat untuk memperoleh gelar  
Sarjana Komputer (S.Kom.)**



**UMN**  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

**Deardo Dibrianto Sinaga**

**13110110084**

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNIK DAN INFORMATIKA  
UNIVERSITAS MULTIMEDIA NUSANTARA**

**TANGERANG**

**2017**

## **LEMBAR PENGESAHAN SKRIPSI**

### **IMPLEMENTASI ALGORITMA RABIN-KARP DAN STEMMING CONFIX-STRIPPING PADA SISTEM PENDETEKSI KEMIRIPAN DOKUMEN TEKS**

Oleh

Nama : Deardo Dibrianto Sinaga

NIM : 13110110084

Program Studi : Teknik Informatika

Fakultas : Teknik dan Informatika

Tangerang, 14 Agustus 2017

Ketua Sidang

Ni Made Satvika Iswari, S.T.,M.T. Arya Wicaksana, S.Kom., M.Eng.Sc.,  
OCA, CEH

Dosen Penguji

Dosen Pembimbing

Seng Hansun, S.Si., M.Cs.

Mengetahui,

Ketua Program Studi

Teknik Informatika

Maria Irmina Prasetyowati, S.Kom., M.T.

## **PERNYATAAN TIDAK MELAKUKAN PLAGIAT**

Dengan ini saya,

Nama : Deardo Dibrianto Sinaga

NIM : 13110110084

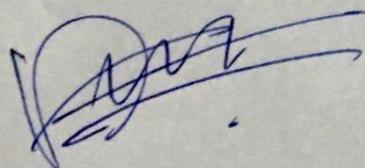
Program Studi : Teknik Informatika

Fakultas : Teknik dan Informatika

menyatakan bahwa skripsi yang berjudul "**Implementasi Algoritma Rabin-Karp dan Stemming Confix-Stripping pada Sistem Pendekripsi Kemiripan Dokumen Teks**" ini adalah karya ilmiah saya sendiri, bukan plagiat dari karya ilmiah yang ditulis oleh orang lain atau lembaga lain, dan semua karya ilmiah orang lain atau lembaga lain yang dirujuk dalam skripsi ini telah disebutkan sumber kutipannya serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/ penyimpangan, baik dalam pelaksanaan skripsi maupun dalam penulisan laporan skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah Skripsi yang telah saya tempuh.

Tangerang, 14 Agustus 2017



(Deardo Dibrianto Sinaga)

## KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa atas berkat dan karunia-Nya sehingga penulis dapat menyelesaikan penyusunan skripsi yang berjudul “Implementasi Algoritma Rabin-Karp dan Stemming Confix-Stripping pada Sistem Pendekripsi Kemiripan Dokumen Teks” ini dengan baik.

Penulis juga menyadari penulisan skripsi ini dapat terselesaikan dengan baik berkat dukungan serta bantuan dari berbagai pihak, oleh karena itu penulis ingin mengucapkan terima kasih kepada:

1. Dr. Ninok Leksono, selaku Rektor Universitas Multimedia Nusantara.
2. Kanisius Karyono, S.T., M.T., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Maria Irmina Prasetyowati, S.Kom, M.T., selaku Ketua Program Studi Teknik Informatika Universitas Multimedia Nusantara.
4. Seng Hansun, S.Si., M.Cs., selaku dosen pembimbing yang telah membimbing penulis selama proses penulisan skripsi sehingga skripsi ini dapat terselesaikan dengan baik.
5. Orang tua dan keluarga yang memberikan dukungan serta bantuan kepada penulis.
6. Olivia Kristi, selaku orang terdekat yang selalu memberikan bantuan dan dukungan selama penulisan skripsi berlangsung.
7. Rekan MDL yang telah mendukung, membantu dan bersama-sama dengan penulis selama masa kuliah dan penulisan skripsi.

8. Rekan dan sahabat penulis yang saling mendukung dan membantu dalam menyelesaikan skripsi.
9. Dosen dan staf Universitas Multimedia Nusantara yang telah membantu dan memberi pengetahuan kepada penulis.
10. Pihak lain yang turut membantu penulis dalam menyelesaikan skripsi ini yang tidak dapat disebutkan satu per satu.

Semoga skripsi ini dapat bermanfaat dan menjadi sumber informasi ataupun sumber inspirasi bagi para pembaca.

Tangerang, 14 Agustus 2017



Penulis

# **IMPLEMENTASI ALGORITMA RABIN-KARP DAN STEMMING CONFIG-STRIPPING PADA SISTEM PENDETEKSI KEMIRIPAN DOKUMEN TEKS**

## **ABSTRAK**

Seiring dengan perkembangan teknologi dan informasi yang begitu pesat, pencarian informasi menjadi semakin mudah dan cepat. Hal tersebut memberikan dampak positif dalam memperluas wawasan dan pengetahuan, tak lepas dari itu, hal tersebut juga membawa dampak negatif seperti plagiarisme. Berdasarkan penelitian yang dilakukan oleh Wibowo, telah banyak *software* dan *website* yang dapat digunakan untuk memeriksa terjadinya plagiarisme, namun *software* tersebut kurang cocok bagi penulisan karya ilmiah dalam Bahasa Indonesia karena dirancang untuk teks dalam Bahasa Inggris. Oleh karena itu dibutuhkan sistem pendekripsi kesamaan pada dokumen yang lebih cocok untuk penulisan dalam Bahasa Indonesia. Rabin-Karp merupakan salah satu algoritma yang dapat digunakan dalam pengecekan kemiripan dokumen, sedangkan *Config-Stripping* merupakan algoritma yang dapat melakukan pencarian kata dasar dalam Bahasa Indonesia. Penelitian ini telah berhasil mengimplementasikan algoritma Rabin-Karp dan *Config-Stripping* dengan baik. Pengujian yang dilakukan dengan berbagai skenario dokumen maupun algoritma telah memberikan beberapa hasil performa sistem dalam segi waktu maupun tingkat *similarity*. Sistem dengan Rabin-Karp murni dapat memberikan performa sistem terbaik, baik dalam segi waktu maupun tingkat *similarity* dengan rata-rata kecepatan total waktu proses sebesar 0.0123 detik dan rata-rata tingkat *similarity* sebesar 89.1967%. Tingkat akurasi yang diberikan oleh sistem adalah sebesar 0.7. Sistem dengan skenario algoritma yang ditambahkan proses *stemming* ataupun N-Gram juga dapat meningkatkan beberapa hasil pengujian dalam segi waktu pemrosesan maupun tingkat *similarity*.

Kata Kunci: Rabin-Karp, *Config-Stripping*, kemiripan, plagiarisme.

**UNIVERSITAS  
MULTIMEDIA  
NUSANTARA**

# **THE IMPLEMENTATION OF RABIN-KARP AND CONFIX-STRIPPING STEMMING ALGORITHM IN SIMILARITY DETECTION SYSTEM FOR TEXT DOCUMENT**

## **ABSTRACT**

Along with the rapid development of information and technology, information search becomes easier and faster. It has some positive impacts in expanding insight and knowledge, yet it also has negative impacts, such as plagiarism. According to Wibowo's research, there are so many softwares and websites that can be used to check the occurrence of plagiarism, but unfortunately, those softwares are not really suitable for scientific papers or research which are written in Bahasa Indonesia because it is designed for text in English. Therefore, people especially for Indonesian researchers, need a document similarity detection system that is more suitable for papers written in Bahasa Indonesia. Rabin-Karp is one of algorithms that can be used in checking the similarity between documents, while confix-stripping is an algorithm that can perform basic word search in Bahasa Indonesia. This research has successfully implemented Rabin-Karp and Confix-Stripping algorithms very well. Tests performed with various document scenarios as well as algorithms have given some performance results of the system in terms of time and similarity level. The system with the pure Rabin-Karp can provide the best system performance, both in terms of time and accuracy with an average total processing time speed of 0.0123 seconds and the average similarity rate of 89.1967%. The level of accuracy given by the system is 0.7. System with algorithm scenarios that has been added with stemming process or N-Gram can also improve some test results in terms of processing time and similarity level.

Keywords: Rabin-Karp, Confix-Stripping, similarity, plagiarism.

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

## DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI .....	ii
PERNYATAAN TIDAK MELAKUKAN PLAGIAT .....	iii
KATA PENGANTAR .....	iv
ABSTRAK .....	vi
ABSTRACT .....	vii
DAFTAR ISI .....	viii
DAFTAR TABEL .....	ix
DAFTAR GAMBAR .....	x
DAFTAR RUMUS .....	xii
BAB I PENDAHULUAN .....	1
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah .....	3
1.3 Batasan Masalah .....	3
1.4 Tujuan Penelitian .....	4
1.5 Manfaat Penelitian .....	4
1.6 Sistematika Penulisan Laporan Penelitian .....	4
BAB II LANDASAN TEORI .....	6
2.1 Plagiarisme .....	6
2.2 Algoritma Rabin-Karp .....	7
2.3 Hash .....	10
2.4 N-Gram .....	10
2.5 <i>Dice's Similarity Coefficient</i> .....	11
2.6 Algoritma <i>Config-Stripping Stemmer</i> .....	11
2.7 <i>Confusion Matrix</i> .....	15
BAB III METODOLOGI DAN PERANCANGAN SISTEM .....	17
3.1 Metodologi Penelitian .....	17
3.2 Perancangan Aplikasi .....	18
3.2.1 Flowchart .....	18
3.2.2 Struktur Tabel .....	35
3.2.3 Desain Antarmuka .....	36
BAB IV IMPLEMENTASI DAN UJI COBA .....	43
4.1 Spesifikasi Sistem .....	43
4.2 Implementasi .....	44
4.2.1 Program dan Antarmuka .....	44
4.2.2 Implementasi Algoritma .....	50
4.3 Pengumpulan Data dan Skenario Dokumen .....	59
4.4 Uji Coba .....	60
4.4.1 Uji Coba Aplikasi .....	60
4.4.2 Analisis Hasil Uji Coba Dokumen .....	65
4.4.3 Uji Akurasi Sistem .....	77
BAB V SIMPULAN DAN SARAN .....	80
5.1 Simpulan .....	80
5.2 Saran .....	81
DAFTAR PUSTAKA .....	82
DAFTAR LAMPIRAN .....	85

## DAFTAR TABEL

Tabel 2.1 Contoh Pemotongan N-Gram Pada Kalimat.....	10
Tabel 2.2 Kombinasi Prefiks dan Sufiks yang Tidak Diperbolehkan.....	13
Tabel 2.3 Aturan Peluruhan Kata Dasar .....	13
Tabel 2.4 Modifikasi dan Tambahan Aturan pada Tabel 2.3.....	14
Tabel 2.5 <i>Confusion Matrix</i> .....	15
Tabel 3.1 Struktur Tabel <i>user</i> .....	35
Tabel 3.2 Struktur Tabel <i>file_uploaded</i> .....	36
Tabel 3.3 Struktur Tabel <i>kata_dasar</i> .....	36
Tabel 3.4 Struktur Tabel <i>history</i> .....	36
Tabel 4.1 Skenario Modifikasi Dokumen .....	59
Tabel 4.2 Teks Uji Perhitungan Rabin-Karp .....	61
Tabel 4.3 <i>Parsing</i> N-Gram Manual .....	63
Tabel 4.4 Kalimat Uji <i>Stemming Configx-Stripping</i> .....	64
Tabel 4.5 Hasil Uji dengan Plagiarism Checker X .....	78
Tabel 4.6 Hasil Uji dengan Sistem Penelitian.....	78
Tabel 4.7 <i>Confusion Matrix</i> Klasifikasi Berat .....	79
Tabel 4.8 <i>Confusion Matrix</i> Klasifikasi Sedang .....	79



## DAFTAR GAMBAR

Gambar 2.1	Fingerprint Awal .....	8
Gambar 2.2	Menggeser <i>Fingerprint</i> .....	8
Gambar 2.3	Perbandingan Kedua .....	8
Gambar 2.4	Perbandingan Keempat (Nilai <i>Hash</i> Sama) .....	9
Gambar 2.5	Perbandingan Kelima ( <i>String</i> Ditemukan).....	9
Gambar 3.1	Flowchart Keseluruhan Sistem .....	19
Gambar 3.2	Flowchart <i>Login Admin</i> .....	20
Gambar 3.3	Flowchart <i>Upload</i> .....	21
Gambar 3.4	Flowchart <i>Match File</i> .....	22
Gambar 3.5	Flowchart <i>Testing Rabin-Karp</i> .....	23
Gambar 3.6	Flowchart <i>Testing Confix-Stripping</i> .....	23
Gambar 3.7	Flowchart Model Rabin-Karp .....	24
Gambar 3.8	Flowchart <i>Parse N-Gram</i> .....	25
Gambar 3.9	Flowchart <i>Hash Teks</i> .....	26
Gambar 3.10	Flowchart <i>Fingerprint</i> .....	27
Gambar 3.11	Flowchart Model Algoritma <i>Confix-Stripping</i> .....	29
Gambar 3.12	Flowchart Langkah 1.....	30
Gambar 3.13	Flowchart Langkah 2.....	30
Gambar 3.14	Flowchart Langkah 3.....	31
Gambar 3.15	Flowchart Langkah 4.....	32
Gambar 3.16	Flowchart Langkah 5.....	33
Gambar 3.17	Flowchart Kombinasi Terlarang .....	34
Gambar 3.18	Flowchart Langkah 6.....	35
Gambar 3.19	Desain Halaman Home Guest .....	37
Gambar 3.20	Desain Halaman Login Admin.....	37
Gambar 3.21	Desain Halaman Home Admin .....	38
Gambar 3.22	Desain Halaman Match .....	38
Gambar 3.23	Desain Halaman Confirm.....	39
Gambar 3.24	Desain Halaman Result .....	39
Gambar 3.25	Desain Halaman History .....	40
Gambar 3.26	Desain Halaman About .....	40
Gambar 3.27	Desain Halaman Help.....	41
Gambar 3.28	Desain Halaman <i>Testing Rabin-Karp</i> .....	41
Gambar 3.29	Desain Halaman <i>Testing Confix-Stripping</i> .....	42
Gambar 4.1	Halaman Home Guest .....	44
Gambar 4.2	Halaman Login Admin.....	45
Gambar 4.3	Halaman Home Admin .....	45
Gambar 4.4	Halaman About .....	46
Gambar 4.5	Halaman Help.....	46
Gambar 4.6	Halaman Match .....	47
Gambar 4.7	Halaman Confirm.....	47
Gambar 4.8	Halaman Result .....	48
Gambar 4.9	Halaman History .....	48
Gambar 4.10	Halaman Testing Rabin-Karp.....	49
Gambar 4.11	Halaman Testing <i>Confix-Stripping</i> .....	49

Gambar 4.12	Potongan Kode Fungsi whiteInsensitive()	50
Gambar 4.13	Potongan Kode Fungsi nGram()	51
Gambar 4.14	Potongan Kode Fungi strHash()	51
Gambar 4.15	Potongan Kode Fungsi fingerPrint()	52
Gambar 4.16	Potongan Kode Fungsi similarity()	52
Gambar 4.17	Potongan Kode Fungsi rabinKarp()	53
Gambar 4.18	Potongan Kode Fungsi whiteInsensitive()	54
Gambar 4.19	Potongan Kode fungsi cek_db()	54
Gambar 4.20	Potongan Kode Fungsi forbid_form()	55
Gambar 4.21	Potongan Kode Fungsi step_2()	55
Gambar 4.22	Potongan Kode Fungsi step_3()	56
Gambar 4.23	Potongan Kode Fungsi step_4()	56
Gambar 4.24	Potongan Kode Fungsi step_5()	57
Gambar 4.25	Potongan Kode Fungsi confixStripping()	58
Gambar 4.26	Masukan Awal Sistem Uji Rabin-Karp.....	62
Gambar 4.27	Hasil Perhitungan Rabin-Karp dengan Sistem.....	62
Gambar 4.28	Hasil <i>Parsing</i> N-Gram Menggunakan Sistem.....	63
Gambar 4.29	Hasil <i>Stemming</i> Sistem.....	64
Gambar 4.30	Grafik Perbandingan Total Waktu Proses Pengujian Dokumen A dengan Algoritma Tipe 1 dan Tipe 2.....	66
Gambar 4.31	Grafik Perbandingan Total Waktu Proses Pengujian Dokumen A dengan Algoritma Tipe 3 dan Tipe 4.....	67
Gambar 4.32	Grafik Perbandingan Tingkat Simililarity Dokumen A .....	68
Gambar 4.33	Grafik Perbandingan Total Waktu Proses Pengujian Dokumen B dengan Tipe 1 dan Tipe 2 .....	69
Gambar 4.34	Grafik Perbandingan Total Waktu Proses Pengujian Dokumen B dengan Algoritma Tipe 3 dan Tipe 4.....	70
Gambar 4.35	Grafik Perbandingan Tingkat Simililarity Dokumen B.....	71
Gambar 4.36	Grafik Perbandingan Total Waktu Proses Pengujian Dokumen C dengan Algoritma Tipe 1 dan Tipe 2 .....	72
Gambar 4.37	Grafik Perbandingan Total Waktu Proses Pengujian Dokumen C dengan Algoritma Tipe 3 dan Tipe 4.....	73
Gambar 4.38	Grafik Perbandingan Tingkat Simililarity Dokumen C.....	74
Gambar 4.39	Grafik Rata-Rata Kecepatan Waktu Proses Rabin-Karp.....	75
Gambar 4.40	Grafik Rata-Rata Kecepatan Total Waktu Proses .....	76
Gambar 4.41	Grafik Rata-Rata Tingkat <i>Similarity</i> .....	77

**U N I V E R S I T A S**  
**M U L T I M E D I A**  
**N U S A N T A R A**

## DAFTAR RUMUS

Rumus 2.1 <i>Rolling Hash</i> .....	10
Rumus 2.2 <i>Dice's Similarity Coefficient</i> .....	11
Rumus 2.3 <i>Precision</i> .....	15
Rumus 2.4 <i>Recall</i> .....	16
Rumus 2.5 <i>F Measure</i> .....	16
Rumus 2.6 <i>Accuracy</i> .....	16

