

***SOURCE DETECTION PADA KASUS PLAGIARISME
DOKUMEN MENGGUNAKAN METODE BIWORD
WINNOWER DAN RETRIEVAL BERBASIS OKAPI BM25***

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Teknik
Pada Jurusan Teknik Informatika

Oleh

SYARIF HIDAYATULLAH

11051101864



**FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU
PEKANBARU**

2014

LEMBAR PERSETUJUAN

***SOURCE DETECTION* PADA KASUS PLAGIARISME
DOKUMEN MENGGUNAKAN METODE *BIWORD*
WINNOWER DAN *RETRIEVAL* BERBASIS OKAPI BM25**

TUGAS AKHIR

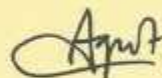
Oleh

SYARIF HIDAYATULLAH

11051101864

Telah diperiksa dan disetujui sebagai laporan tugas akhir
di Pekanbaru, pada tanggal 18 Februari 2014

Pembimbing,



Surya Agustian, ST, M.Kom

NIP. 19760830 201101 1 003

LEMBAR PENGESAHAN

**SOURCE DETECTION PADA KASUS PLAGIARISME
DOKUMEN MENGGUNAKAN METODE BIWORD
WINNOWER DAN RETRIEVAL BERBASIS OKAPI BM25**

TUGAS AKHIR

Oleh

SYARIF HIDAYATULLAH

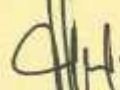
11051101864

Telah dipertahankan di depan sidang dewan penguji
sebagai salah satu syarat untuk memperoleh gelar sarjana Teknik Informatika
Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau
di Pekanbaru, pada tanggal 18 Februari 2014

Pekanbaru, 18 Februari 2014

Mengesahkan,

Ketua Jurusan,



Elin Haerani, S.T. M. Kom

NIP. 19810323 200710 2 003



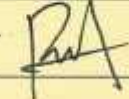
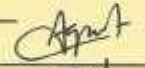
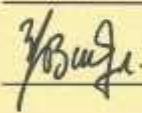
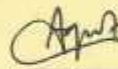
Dekan,

Dra. Hj. Yenita Morena, M.Si

NIP. 19601125 198503 2 002

DEWAN PENGUJI

Ketua : Surya Agustian, S.T, M.Kom
Sekretaris : Surya Agustian, S.T, M.Kom
Penguji I : Elvia Budianita, S.T, M.Cs
Penguji II : Rahmad Abdillah, M.T



***SOURCE DETECTION PADA KASUS PLAGIARISME
DOKUMEN MENGGUNAKAN METODE BIWORD
WINNOWER DAN RETRIEVAL BERBASIS OKAPI BM25***

**SYARIF HIDAYATULLAH
11051101864**

Tanggal Sidang: 18 Februari 2014

Periode Wisuda: Juni 2014

Jurusan Teknik Informatika

Fakultas Sains dan Teknologi

Universitas Islam Negeri Sultan Syarif Kasim Riau

ABSTRAK

Tindakan plagiarisme merupakan salah satu dampak negatif kemajuan teknologi informasi dan komunikasi. Untuk mengatasi plagiarisme maka perlu adanya suatu aplikasi yang dapat mendeteksi plagiarisme. Saat ini sudah banyak penelitian yang membangun aplikasi pendeteksi plagiarisme, namun aplikasi tersebut hanya sebatas membandingkan kemiripan antara dua dokumen saja, tidak termasuk mendeteksi kemiripan dokumen dengan banyak dokumen dan mendeteksi dimana dokumen sumbernya. Pada penelitian ini akan dibangun sebuah aplikasi yang dapat mendeteksi dokumen sumber dari sebuah dokumen uji dan dapat membandingkan kemiripannya terhadap dokumen sumber yang diperoleh. Dalam mendeteksi dokumen sumber, isi dokumen uji akan dibentuk menjadi *query* dengan metode *fingerprint biword winnowing*. *Query* yang dibentuk terbagi menjadi dua jenis yakni *query* dengan *stemming* dan tanpa *stemming*. Setiap jenis *query* terbagi lagi menjadi 3 yakni *query* dengan n frekuensi *fingerprint* tertinggi, tengah dan terendah. Kemudian *query* tersebut akan dimasukkan kedalam sistem *information retrieval* (IR) Model Okapi BM25 untuk dicari dokumen sumbernya. Dokumen sumber yang diperoleh kemudian akan dibandingkan tingkat kemiripannya terhadap dokumen dengan menggunakan algoritma *biword winnowing*. Pada penelitian ini dilakukan pengujian hasil pembentukan *query* terhadap dokumen sumber yang diperoleh dan pengujian kemiripan dokumen uji dengan dokumen sumber yang diperoleh. *Output* pengujian hasil pembentukan *query* dimana n adalah 5, rekomendasi jenis *query* yang terbaik adalah *query stemming* dan tanpa *stemming* 5 frekuensi *fingerprint* tengah. Sedangkan pengujian kemiripan dokumen disimpulkan bahwa dokumen sumber yang diperoleh benar terdapat kemiripan isi dengan dokumen uji dengan kemiripan 7,36-65,32%.

Kata Kunci: *biword winnowing, information retrieval, jaccard coefficient, Okapi BM25, query.*

***SOURCE DETECTION OF PLAGIARISM CASE DOCUMENT
USING BIWORD WINNOWER AND RETRIEVAL BASED ON
OKAPI BM25 MODEL***

**SYARIF HIDAYATULLAH
11051101864**

Date of Final Exam: February 18th, 2014

Graduation Ceremony Period: June 2014

*Informatics Engineering Departement
Faculty of Sciences and Technology
State Islamic University of Sultan Syarif Kasim Riau*

ABSTRACT

Plagiarism is one of the negative impact of information and communication technology advances. To overcome the plagiarism, an application that can detect plagiarism is needed. Currently, there are many studies that establish plagiarism detection application, but the application is limited only to compare the similarity between two documents, it is not including detecting similarity of a document with many documents and detect where the source documents. This research will build an application that can detect the source documents of a test document and be able to compare the similarity of the source documents were obtained. In detecting the source document, the contents of the test document will be formed into a query with Winnower fingerprint biword method. Formed query is divided into two types, they are query by stemming and no stemming. Each type of query is divided into three namely queries with highest fingerprint frequency, middle and low. Then the query will be incorporated into the system of information retrieval (IR) models Okapi BM25 to look for the source document. Similarity level of source documents obtained will be compared to the document by using biword Winnower algorithm. In this research, the test queries formation result to source document obtained and the similarity of test document to source document will be done. . Output test results query formation where n is 5, the best recommendation is query by stemming and query without stemming with 5 fingerprint middle frequency. While the similarity test concluded that the source documents obtained correctly has similarities with the contents of the test document from 7.36 to 65.32%.

Keywords : *biword winnower, information retrieval, jaccard coefficient, Okapi BM25, query.*

KATA PENGANTAR

Assalammu'alaikum wa rahmatullahi wa barakatuh.

Alhamdulillah rabbil'alamin, tak henti-hentinya penulis ucapkan kehadiran Tuhan yang tiada Tuhan selain Dia, Allah SWT, yang dengan rahmat dan hidayahNya penulis mampu menyelesaikan Tugas Akhir ini dengan baik. Tidak lupa dan tak akan pernah lupa bershalawat kepada Nabi dan RasulNya, Muhammad SAW yang hanya menginginkan keimanan dan keselamatan bagi umatnya dan sangat belas kasihan lagi penyayang kepada orang-orang mukmin.

Tugas Akhir ini disusun sebagai salah satu syarat untuk mendapatkan gelar kesarjanaan pada jurusan Teknik Informatika Universitas Islam Negeri Sultan Syarif Kasim Riau. Banyak sekali pihak yang telah membantu penulis dalam penyusunan laporan ini, baik berupa bantuan materi ataupun berupa motivasi dan dukungan kepada penulis. Semua itu tentu terlalu banyak bagi penulis untuk membalasnya, namun pada kesempatan ini penulis hanya dapat mengucapkan terima kasih kepada :

1. ALLAH SWT yang karena hidayah dan rahmat nya, sehingga semua kebaikan mendekati saya dan saya bisa melakukan hal yang terbaik yang dapat saya lakukan.
2. Rasulullah SAW, seorang pimpinan umat islam sekaligus panutan umat islam yang tiada henti-hentinya berjuang demi menyebarkan agama islam.
3. Terima kasih kepada Kedua orang Tua Penulis, Ibu dan Bapak, serta adek yang tiada hentinya memanjatkan doa, memberikan dukungan dan semangat untuk kesuksesan penulis.
4. Bapak Prof. Dr. H. M. Nazir, selaku Rektor Universitas Islam Negeri Sultan Syarif Kasim Riau.
5. Ibu Dra. Hj. Yenita Morena, M.Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau.
6. Ibu Elin Haerani, S.T, M. Kom selaku Ketua Jurusan Teknik Informatika Fakultas Sains dan Teknologi UIN SUSKA RIAU.

7. Surya Agustian, S.T, M. Kom Selaku dosen pembimbing tugas akhir. Terimakasih pak untuk waktu yang selalu bapak luangkan untuk penulis, ilmu, semangat, dan motivasinya yang luar biasa. Terimakasih banyak pak.
8. Elvia Budianita, S.T,M.Cs, selaku dosen penguji 1 yang banyak membantu dan memberi masukan penulis dalam penyempurnaan Laporan Tugas Akhir ini, untuk ilmu-ilmunya.
9. Rahmad Abdillah, M.T, selaku dosen penguji 2, terimakasih pak untuk ilmu-ilmunya, saran-sarannya, perbaikan-perbaikannya, dan masukannya.
10. Muhammad Affandes, M.T, sebagai koordinator tugas akhir yang telah memberi masukan-masukan untuk penyelesaian tugas akhir ini, dan sangat sabar membantu penulis dalam mempersiapkan semua kebutuhan penulis dalam penyelesaian Tugas Akhir ini.
11. Bapak dan Ibu Dosen Jurusan Teknik Informatika Fakultas Sains dan Teknologi UIN SUSKA RIAU yang telah banyak memberikan ilmunya kepada penulis.
12. Teman-teman seperjuangan TIF B angkatan 10, Agung, Rizky, Denanda, Ardy, Alif, Rolli, Zul, Yessi, Fe, Nuraisyah dan yang lainnya yang tidak bisa disebutkan satu persatu. Tetap semangat dan tetap berjaya.

Penulis menyadari bahwa dalam penulisan laporan ini masih banyak kesalahan dan kekurangan, oleh karena itu kritik dan saran yang sifatnya membangun sangat penulis harapkan untuk kesempurnaan laporan ini. Akhirnya penulis berharap semoga laporan ini dapat memberikan sesuatu yang bermanfaat bagi siapa saja yang membacanya. Amin.

Wassalamu'alaikum wa rahmatullahi wa barakatuh

Pekanbaru, 18 Februari 2014

Penulis

DAFTAR ISI

LEMBAR PERSETUJUAN.....	ii
LEMBAR PENGESAHAN	iii
LEMBAR HAK ATAS KEKAYAAN INTELEKTUAL.....	iv
LEMBAR PERNYATAAN	v
ABSTRAK	vii
ABSTRACT.....	viii
KATA PENGANTAR	ix
DAFTAR ISI.....	xi
DAFTAR GAMBAR	xiii
DAFTAR TABEL.....	xv
DAFTAR RUMUS	xvii
BAB I PENDAHULUAN.....	I-1
1.1 Latar Belakang	I-1
1.2 Rumusan Masalah	I-4
1.3 Batasan Masalah.....	I-4
1.4 Tujuan	I-4
1.5 Sistematika Pembahasan	I-5
BAB II LANDASAN TEORI	II-1
2.1 Pengertian Plagiarisme.....	II-1
2.2 Metode Mendeteksi Plagiarisme	II-1
2.3 Information Retrieval	II-2
2.3.1 Arsitektur Sistem Information Retrieval.....	II-4
2.3.2 Pembuatan Index.....	II-4
2.3.3 Algoritma Nazief dan Adriani	II-6
2.3.4 Pembobotan Kata	II-9
2.4 Model Okapi Best Match (BM) 25	II-10
2.5 Algoritma Winnowing	II-11
2.5.1 Konsep Biword pada Winnowing.....	II-15
2.5.2 Jaccard Coefficient.....	II-16
BAB III METODOLOGI PENELITIAN.....	III-1
3.1 Identifikasi Masalah	III-1
3.2 Merumuskan Masalah	III-1
3.3 Study Literatur	III-2

3.4 Analisa Aplikasi	III-2
3.5 Perancangan Aplikasi.....	III-5
3.6 Implementasi	III-5
3.7 Pengujian.....	III-6
3.8 Kesimpulan dan Saran.....	III-6
BAB IV ANALISA DAN PERANCANGAN	IV-1
4.1 Analisa Source Detection Dokumen	IV-2
4.1.1 Analisa Pembuatan Query.....	IV-2
4.1.2 Analisa Sistem IR Model Okapi BM25	IV-15
4.1.3 Analisa Deteksi kemiripan dokumen dengan Algoritma Biword Winnowing	IV-24
4.2 Perancangan Aplikasi.....	IV-39
4.2.1 Perancangan File Teks (Flat File)	IV-40
4.2.2 Perancangan Struktur Menu.....	IV-40
4.2.3 Perancangan Interface	IV-41
4.2.3.1 Rancangan Interface Menu Beranda.....	IV-41
4.2.3.2 Rancangan Interface Menu Koleksi Dokumen.....	IV-42
4.2.3.3 Rancangan Interface Menu Source Detection	IV-43
4.2.3.4 Rancangan Interface Menu Bantuan.....	IV-45
BAB V IMPLEMENTASI DAN PENGUJIAN	V-1
5.1 Tahapan Implementasi	V-1
5.1.1 Batasan Implementasi	V-1
5.1.2 Lingkungan Implementasi	V-2
5.1.3 Implementasi Interface Aplikasi	V-2
5.2 Pengujian aplikasi	V-6
5.2.1 Rencana Pengujian.....	V-6
5.2.1.1 Pengujian Hasil Pembuatan Query dari Dokumen yang Diduga Plagiarisme Terhadap Hasil Pencarian Dokumen Sumber	V-8
5.2.1.2 Pengujian Kemiripan Dokumen Uji Terhadap Dokumen Sumber yang Diperoleh dari Proses Pencarian dengan Model Okapi BM25.	V-32
5.2.2 Kesimpulan Pengujian	V-35
BAB VI PENUTUP	VI-1
6.1 Kesimpulan	VI-1
6.2 Saran.....	VI-2
DAFTAR PUSTAKA	
DAFTAR RIWAYAT HIDUP	

DAFTAR GAMBAR

Gambar	Halaman
2.1 Arsitektur dasar sistem IR.....	II-4
3.1 Tahapan Penelitian	III-1
3.2 Analisa Sistem <i>Source Detection</i>	III-2
4.1 Analisa <i>Source Detection</i> Dokumen	IV-1
4.2 Tahapan Pembuatan <i>Query</i> dari <i>Fingerprint Biword Winnowing</i> dengan <i>Stemming</i>	IV-3
4.3 Flowchart Algoritma Nazief dan Adriani	IV-3
4.4 Tahapan Pembuatan <i>Query</i> dari <i>Fingerprint Biword Winnowing</i> tanpa <i>Stemming</i>	IV-9
4.5 Proses Pembuatan Inverted Index	IV-15
4.6 Proses Preprocessing Query	IV-21
4.7 Proses Pembuatan <i>Database Fingerprint</i> Dokumen.....	IV-24
4.8 Proses Pembentukan <i>Fingerprint</i> Dokumen Diduga Plagiarisme	IV-32
4.9 Rancangan Struktur Menu.....	IV-41
4.10 Rancangan <i>Interface</i>	IV-41
4.11 Rancangan <i>Interface</i> Menu Beranda.....	IV-42
4.12 Rancangan <i>Interface</i> Menu Koleksi Dokumen	IV-42
4.13 Rancangan <i>Interface</i> Tambah Koleksi Dokumen	IV-43
4.14 Rancangan <i>Interface</i> Lihat Isi Dokumen	IV-43
4.15 Rancangan <i>Interface</i> Pembuatan <i>Query</i>	IV-44
4.16 Rancangan <i>Interface</i> Pemilihan <i>Query</i>	IV-44
4.17 Rancangan <i>Interface</i> Dokumen Sumber	IV-44
4.18 Rancangan <i>Interface</i> Perbandingan Dokumen.....	IV-45
4.19 Rancangan <i>Interface</i> Halaman Bantuan.....	IV-45
5.1 <i>Interface</i> Menu Beranda.....	V-3
5.2 <i>Interface</i> Menu Koleksi Dokumen.....	V-3
5.3 <i>Interface</i> Tambah Koleksi Dokumen	V-3
5.4 <i>Interface</i> Lihat Isi Dokumen	V-4
5.5 <i>Interface</i> Pembuatan <i>Query</i>	V-4
5.6 <i>Interface</i> Pemilihan <i>Query</i>	V-5
5.7 <i>Interface</i> Dokumen Sumber	V-5
5.8 <i>Interface</i> Perbandingan Dokumen.....	V-5

5.9 <i>Interface</i> Menu Bantuan.....	V-6
5.10 Perbandingan Kemiripan Dokumen Pengujian I Perbandingan Kemiripan 1	V-32
5.11 Perbandingan Kemiripan Dokumen Pengujian I Perbandingan Kemiripan 2	V-33
5.12 Perbandingan Kemiripan Dokumen Pengujian I Perbandingan Kemiripan 3	V-33
5.13 Perbandingan Kemiripan Dokumen Pengujian II Perbandingan Kemiripan 1	V-34
5.14 Perbandingan Kemiripan Dokumen Pengujian II Perbandingan Kemiripan 2	V-34
5.15 Perbandingan Kemiripan Dokumen Pengujian II Perbandingan Kemiripan 3	V-35

DAFTAR TABEL

Tabel	Halaman
2.1 Kombinasi Awalan Akhiran yang Tidak Diizinkan.....	II-8
2.2 Jenis Awalan Kata yang Berawalan Te-	II-8
2.3 Jenis Awalan Berdasarkan Tipe Awalnya.....	II-9
4.1 Hasil Proses Tokenisasi Masing-masing Dokumen.....	IV-17
4.2 Hasil Proses <i>linguistic preprocessing</i> Masing-masing Dokumen.....	IV-18
4.3 Hasil <i>Indexing</i> dari Seluruh Token Dokumen	IV-18
4.4 Hasil Pembobotan Kata terhadap Kata Hasil <i>Indexing</i>	IV-20
4.5 Hasil Pemotongan Teks menjadi Kata Tunggal Masing-masing Dokumen.....	IV-25
4.6 Hasil Pembentukan <i>Biword</i> dari Kata Tunggal Masing-masing Dokumen.....	IV-27
4.7 Hasil Enkripsi <i>Biword</i> Masing-masing Dokumen dengan <i>MD5</i>	IV-28
4.8 Hasil <i>Rolling Hash Biword</i> Masing-masing dokumen yang telah Dienkripsi dengan <i>MD5</i>	IV-29
4.9 Hasil Pembuatan <i>Database Fingerprint</i>	IV-32
5.1 Hasil Pengujian I <i>Query</i> 1	V-8
5.2 Hasil Pengujian I <i>query</i> 2.....	V-9
5.3 Hasil Pengujian I <i>Query</i> 3.....	V-9
5.4 Hasil Pengujian I <i>Query</i> 4.....	V-10
5.5 Hasil Pengujian I <i>Query</i> 5	V-10
5.6 Hasil Pengujian I <i>Query</i> 6.....	V-11
5.7 Hasil Pengujian I keseluruhan <i>Query</i>	V-11
5.8 Hasil Pengujian II <i>Query</i> 1.....	V-13
5.9 Hasil Pengujian II <i>Query</i> 2.....	V-13
5.10 Hasil Pengujian II <i>Query</i> 3.....	V-14
5.11 Hasil Pengujian II <i>Query</i> 4.....	V-14
5.12 Hasil Pengujian II <i>Query</i> 5.....	V-15
5.13 Hasil Pengujian II <i>Query</i> 6.....	V-15
5.14 Hasil pengujian II keseluruhan <i>Query</i>	V-16
5.15 Hasil Pengujian III <i>Query</i> 1	V-17
5.16 Hasil Pengujian III <i>Query</i> 2	V-18

5.17 Hasil Pengujian III <i>Query</i> 3	V-18
5.18 Hasil Pengujian III <i>Query</i> 4	V-19
5.19 Hasil Pengujian III <i>Query</i> 5	V-19
5.20 Hasil Pengujian III <i>Query</i> 6	V-20
5.21 Hasil pengujian III keseluruhan <i>Query</i>	V-20
5.22 Hasil <i>Pengujian</i> IV <i>Query</i> 1	V-22
5.23 Hasil Pengujian IV <i>Query</i> 2	V-22
5.24 Hasil Pengujian IV <i>Query</i> 3	V-23
5.25 Hasil Pengujian IV <i>Query</i> 4	V-23
5.26 Hasil Pengujian IV <i>Query</i> 5	V-24
5.27 Hasil Pengujian IV <i>Query</i> 6	V-24
5.28 Hasil Pengujian IV Keseluruhan <i>Query</i>	V-25
5.29 Hasil Pengujian V <i>Query</i> 1	V-26
5.30 Hasil Pengujian V <i>Query</i> 2	V-27
5.31 Hasil Pengujian V <i>Query</i> 3	V-27
5.32 Hasil Pengujian V <i>Query</i> 4	V-28
5.33 Hasil pengujian V <i>Query</i> 5	V-28
5.34 Hasil Pengujian V <i>query</i> 6	V-29
5.35 Hasil Pengujian V keseluruhan <i>Query</i>	V-29
5.36 Jenis <i>Query</i> Terbaik Berdasarkan Hasil Masing-masing Pengujian	V-30
5.37 Dokumen Sumber yang Diperoleh Pengujian I	V-32
5.38 Dokumen Sumber yang Diperoleh Pengujian II	V-33

DAFTAR RUMUS

Rumus	Halaman
2.1 Persamaan <i>Logarithm term frequency</i>	II-9
2.2 Persamaan <i>Augmented term frequency</i>	II-10
2.3 Persamaan <i>Inverse document frequency (idf)</i>	II-10
2.4 RSV sederhana	II-10
2.5 RSV <i>query</i> panjang	II-10
2.6 RSV <i>query</i> pendek	II-11
2.7. Persamaan Metode <i>Hash</i>	II-13
2.8. Persamaan <i>Rolling Hash</i>	II-13
2.9. Persamaan <i>jaccard coefficient</i>	II-16