Identifikasi Plagiasi Karya Ilmiah berbasis Temu Kembali Informasi Menggunakan Algoritam Edit Distance Melalui Peringkasan Teks Otomatis

Iyan Mulyana*[1], Andi Chairunnas[2], Aries Maesya[3]
[1,2,3]Jurusan Ilmu Komputer, FMIPA Universitas Pakuan, Bogor
e-mail: *[1]iyandelon@yahoo.com, [2]andi.chairunnas@gmail.com, [3]a.maesya@unpac.ac.id

Abstrak

Karya ilmiah yang baik selain mengandung kebenaran ilmiah juga merupakan karya asli penulis atau tim penulis yang dapat dipertangungjawabkan. Karya ilmiah yang baik bukan merupakan hasil menjiplak, meniru atau yang lebih dikenal dengan plagiat dari karya orang lain. Tujuan dari penelitian ini adalah mengembangkan sistem yang dapat mengidentifikasi plagiasi karya ilmiah secara on line berbasis Temu Kembali Informasi menggunakan algoritma edit distance dan melalui peringkasan teks secara otomatis. Sistem yang dikembangkan terdiri dari tiga modul utama yaitu modul temu kembali informasi menggunakan metode pembobotan dan pengukuran kemiripan Vector Space Model, Modul pengukuran kemiripan dokumen mengunakan algoritam edit distance dan modul peringkasan teks otomatis menggunakan Terms Frequency-inverse document frequency (TF-IDF). Berdasakan hasil evaluasi diperoleh tingkat akurasi sistem antara 80-90 %. Untuk dokumen yang berisi lebih dari 100 kata diperoleh peningkatan kecepatan pengukuran kemiripan rata-rata 50 % setelah dilakukan peringkasan teks secara otomtis terlebih dahulu.

Kata kunci*— karya ilmiah, plagiat, temu kembali informasi, TF-IDF, edit distance*

Abstract

Good scientific work contains scientific truth and author idea without any plagiarism. Plagiarism is taking essay (opinions, etc.) of other people and make it as if the essay (opinions, etc.) themselves, eg published writings of others in the name of himself. The purpose of this research is to develop a system that can identify plagiarism scientific papers based information retreival using edit distance algorithm and through automated summarizing text. The system consists of three main modules, namely Module Information Retrieval, Automated Summarizing Text Module and Document Similarity Measurement Module. In search of Information Retrieval used weighting method and similarity measurement of Vector Space Model. In the process of summarizing text can be usedn Terms frequency-inverse document frequency Based on the assessments of the accuracy of the system is obtained between 80-90%. As for the level of document similarity measurement speed with a minimum amount of 100 words an increase in the average processing rate of 50% after summarizing text.

Keywords— Plagiarism, Scientific words, information retrieval, Tf-IDF, Edit Distance.

1. PENDAHULUAN

Karya ilmiah merupakan kegiatan intelektual yang terdiridari proses penelitian, pengamatan, peninjauan dan pemikiran yang mendalam terhadap suatu pokok permasalahan tertentu. Di kalangan akademis karya ilmiah tersebut dijadikan sebagai sarana untuk menyampaikan gagasannya kepada masyarakat mengenai suatu topik yang diteliti dan diamatinya secara mendalam. Karya ilmiah yang dipublikasikan melalui media teknologi informasi memiliki

dampak positif untuk penelitian-penelitian selanjutnya. Tetapi dampak negatifnya adalah adanya kemungkinan tindakan plagiarisme yaitu perbuatan secara sengaja atau tidak sengaja dalam memperoleh nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya ilmiah pihak lain[1]. Karya ilmiah yang baik selain mengandung kebenaran ilmiah juga merupakan karya asli penulis atau tim penulis yang dapat dipertangungjawabkan. Karya ilmiah yang baik bukan merupakan hasil menjiplak, meniru atau yang lebih dikenal dengan plagiat dari karya orang lain.

Dengan adanya publikasi karya ilmiah diharapkan kontrol dan pemeriksaan terhadap plagiasi karya ilmiah dapat dilakukan baik oleh penulis, pengelola media publikasi maupun oleh pembaca. Kondisi yang selama ini berjalan untuk mengontrol dan memeriksa suatu karya ilmiah termasuk kategori plagiat atau tidak adalah dengan cara membaca dan melihat langsung karya ilmiah yang di periksa. Dengan cara tersebut akan menjadi kendala apabila jumlah karya ilmiah yang akan diperiksa cukup banyak karena akan memerlukan waktu yang cukup lama.

Selain hal di atas permasalahan yang muncul adalah bagaimana seorang penulis atau pembaca dapat mengidentifikasi plagiasi karya ilmiah dengan cepat dan mudah. Untuk karya ilmiah yang dipublikasikan secara *on line* salah satu solusi yang dapat dilaksanakan adalah dengan menggunakan sistem temu kembali informasi yaitu sistem yang berfungsi untuk menemukan informasi yang terkandung dalam sebuah dokumen[2]. Sistem temu kembali informasi di atas akan lebih cepat prosesnya dengan menggunakan peringkasan teks secara otomatis dan akan memberikan ketepatan yang tinggi jika menggunakan agoritma *edit distance* untuk identifikasi plagiasi karya ilmiahnya .

2. METODE PENELITIAN

Tahapan metode penelitian yang dilakukan meliputi: analisis Sistem, Perancanagn Sistem, Implementasi sistem dan evaluasi Sistem.

2.1. Analisis Sistem

Pada tahap ini dilakukan analisis terhadap kebutuhan sistem yang akan dikembangkan antara lain menganalisis dan mempelajari penggabungan Modul Sistem Temu Kembali Informasi dengan Modul Pengukuran kemiripan tanpa melalui Peringkasan Teks Otomatis maupun melalui Peringkasan Teks Otomatis.

2.2. Perancangan Sistem

Perancangan Sistem yang dikembangkan antara lain :

- Modul Sistem Temu Kembali Informasi untuk pengukuran Kemiripan Dokumen berbahasa Indonesia tanpa melalui Peringkasan Teks Otomatis.
- Modul Sistem Temu Kembali Informasi untuk pengukuran Kemiripan Dokumen berbahasa Indonesia tanpa melalui Peringkasan Teks Otomatis .

2.3. Implementasi Sistem

Implementasi untuk kedua modul sistem dikembangkan berbasis web menggunakan Bahasa Pemrogaram PHP. Adapun tahap pembuatan aplikasinya sebagai berikut:

- Pembuatan Modul Aplikasi Temu Kembali Informasi menggunakan metode *Vector Space Model*
- Pembuatan Modul Pengukuran Kemiripan Dokumen menggunakan Algoritma *Edit Distance*
- Pembuatan Modul Peringkasan teks otomatis menggunakan metode TF-IDF

- Mengintegrasikan ketiga modul mejadi Sistem Temu Kembali Informasi untuk identifikasi plagiasi karya ilmiah tanpa Peringkasan Teks Otomatis atau melalui Peringkasan Teks Otomatis

2.4. Evaluasi Sistem

Evaluasi sistem yang dilaksanakan terdiri dari tiga rancangan percobaan antara lain :

- Pengujian tingkat akurasi Sistem Temu Kembali Informasi untuk identifikasi plagiasi karya ilmiah tanpa Peringkasan Teks Otomatis dan melalui Peringkasan Teks Otomatis menggunakan dokumen yang isinya tidak sama
- Pengujian tingkat akurasi Sistem Temu Kembali Informasi untuk identifikasi plagiasi karya ilmiah tanpa Peringkasatan Teks Otomatis dan melalui Peringkasan Teks Otomatis menggunakan dokumen isinya semuanya sama.
- Pengujian tingkat kecepatan proses Sistem Temu Kembali Informasi untuk identifikasi plagiasi karya ilmiah tanpa Peringkasatan Teks Otomatis dan melalui Peringkasan Teks.

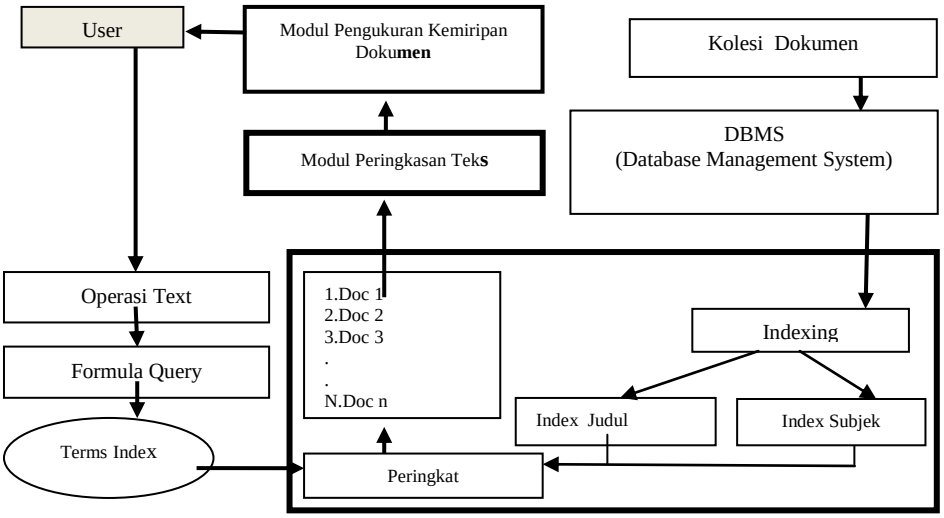
3. HASIL DAN PEMBAHASAN

3.1. Rancangan Sistem

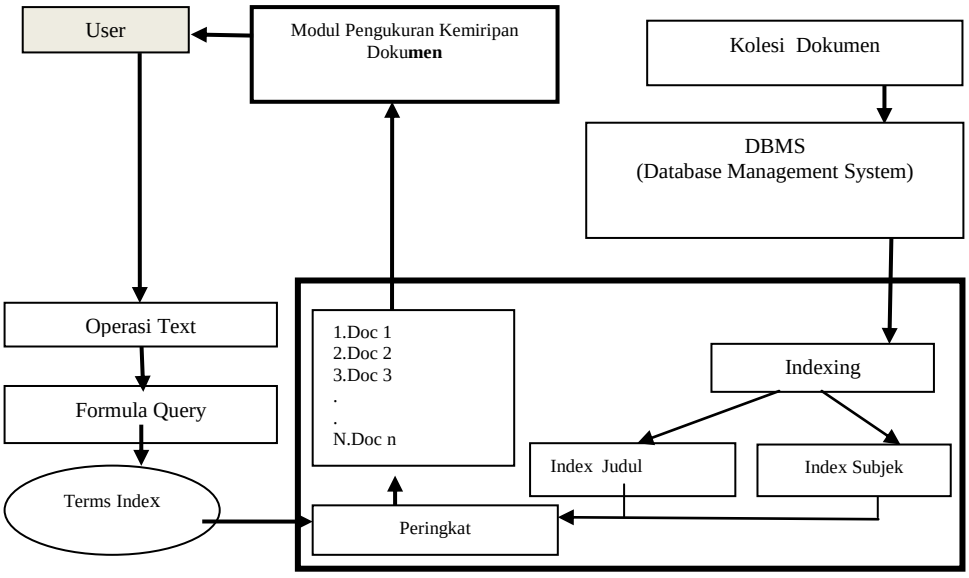
Pada penelitian ini dikembangankan Sistem Temu Kembali Informasi menggunakan teknik peringkasan teks otomatis dan algoritma *edit distance* untuk identifikasi plagiasi karya ilmiah secara *on line*. Secara umum sistem yang dikembangkan terdapat tiga modul utama yang terintegrasi yaitu modul temu kembali informasi, modul peringkasan teks otomatis dan modul kemiripan dokumen. Adapun tahapannya adalah dokumen hasil temu kembali Informasi diproses peringkasan teks. Selanjutnya dokumen hasil peringkasan diukur kemiripannya dengan dokumen query .

Dalam modul Temu Kembali Informasi diterapkan metode pembobotan dan pengukuran kemiripan *Vector Space Model* yaitu suatu metode untuk merepresentasikan sistem temu kembali informasi ke dalam vektor dan memperhitungkan fungsi *similarity* dalam proses pencocokan beberapa vektor [3].

Dalam proses peringkasan teks otomatis dapat diterapkan *Terms Frequency-Inverse Document Frequency* (TF-IDF)[4][5]. Sedangkan dalam Pengukuran kemiripan dokumen dapat diterapkan Algoritam *Edit Distance* [6][7]. Gambaran umum sistem yang dikembangkan ditunjukan pada Gambar 1 dan Gambar 2.



Gambar 1. Rancangan Sistem yang dikembangkan melalui peringkasan teks otomatis



Gambar 2. Rancangan Sistem yang dikembangkan tanpa peringkasan teks otomatis

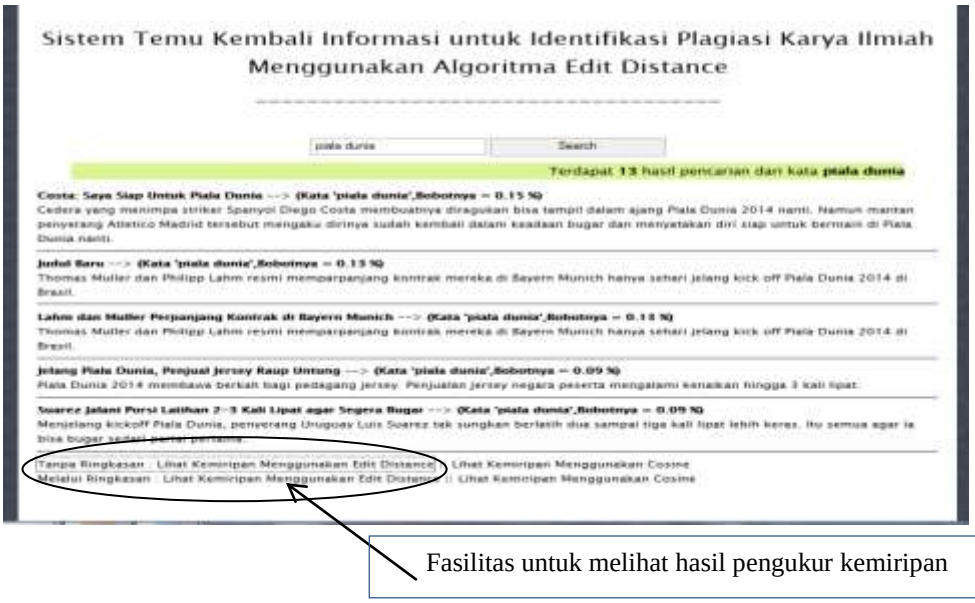
3.2. Hasil Implementasi sistem

Adapun tampilan halaman utama dari sistem yang dikembangkan ditunjukan pada Gambar 3.

Gambar 3. Halaman utama Sistem Temu Kembali Informasi untuk identifikasi plagiasi karya ilmiah menggunakan Algoritma Edit Distance

Halaman utama di atas berisi fasilitas pencarian temu kembali infomasi berdasarkan inputan query *keyword* yang dimasukan oleh user. Setelah melalui *preprocessing* dan proses

pembobotan setiap kata dari *corpus* menggunakan metode *Vector Space Model***.** Tampilan hasil pencarian dokumen Temu Kembali Informasi ditunjukan pada Gambar 4.



Gambar 4. Tampilan Halaman Hasil Pencarian Modul Temu Kembali Informasi

Setelah hasil pencarian Temu Kembali Informasi diperoleh seperti ditunjukan pada Gambar 5, Tahap selanjutnya adalah masing- masing dokumen dibandingkan dan di ukur kemiripannya menggunakan Algoritma *Edit Distance.* Pada aplikasi hasil pencarian tersedia dua fasilitas untuk melihat hasil kemiripan dokumen yaitu pengukuran kemiripan tanpa melalui peringkasan teks dan melalui peringkasan teks yang terdapat pada bagian bawah halaman Temu Kembali Informasi.

Apabila yang dipilih adalah pengukuran kemiripan dokumen tanpa melalui peringkasan teks otomatis maka dokumen yang dihasilkan dari temu kembali informasi langsung diproses pada tahap *preprocessing*. Hasil *text preprocessing* kemudian dihitung jumlah karakter dan nilai *edit distance* untuk mendapatkan nilai *persentase* kemiripan antar dokumen. Tampilan hasil pengukuran kemiripan dokumen tanpa peringkasan teks otomatasi ditunjukan pada Gambar 5.

.

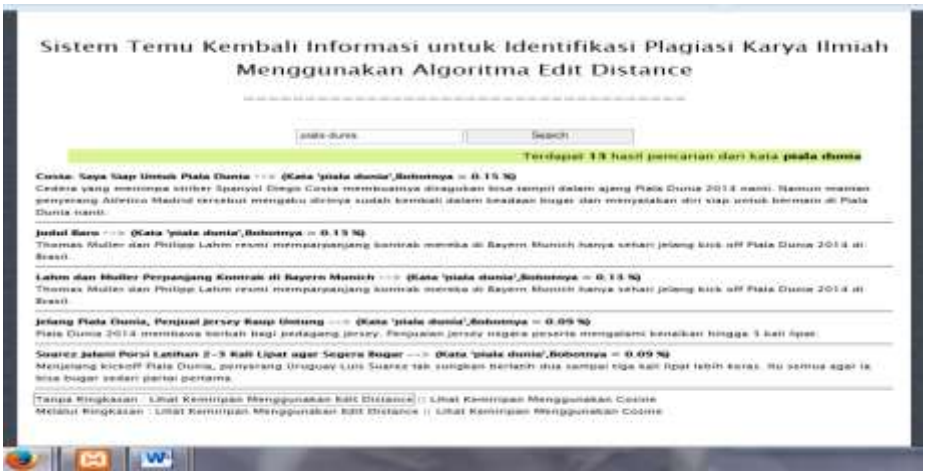
Gambar 5. Tampilan Hasil Pengukuran Kemiripan antar dokumen tanpa Peringkasan Teks Otomatis

Sedangkan apabila yang dipilih adalah pengukuran kemiripan dokumen melalui peringkasan teks otomatis maka dokumen yang dihasilkan dari temu kembali informasi diproses peringkasan teks otomatis. Hasil ringkasan yang diperolah selanjutnya diproses pada tahap *preprocessing*. Hasil *text preprocessing* kemudian dihitung jumlah karakter dan nilai *edit distance* untuk mendapatkan nilai *persentase* kemiripan antar dokumen. Tampilan hasil pengukuran kemiripan dokumen tanpa peringkasan teks otomatis ditunjukan pada Gambar 6.

Gambar 6. Tampilan Hasil Pengukuran Kemiripan antar dokumen melalui Peringkasan Teks Otomatis

3.3. Evaluasi Tngkat Akurasi Sistem

Evaluasi tingkat akurasi sistem dilakukan dengan membandingkan pengukuran kemiripan antar dokumen yang dihasilkan dari sistem yang dikembangkan dengan perhitungan secara manual. Gambar 7 menujukan dokumen hasil Temu Kembali Informasi yang akan di ukur kemiripan antar dokumennya. Sedangkan Gambar 8 menunjukan hasil pengukuran antar dokumen melalui sistem menggunakan algoritma *edit distanace.*



Gambar 7. Tampilan Hasil Temu Kembali Informasi menggunakan *Vector Space Model*

Gambar 8. Tampilan Hasil Pengukuran Kemiripan antar dokumen menggunakan Algoritma Edit Distance

Dari hasil pengukuran kemiripan dokumen C1 dan C2 seperti ditampilkan pada Gambar 8 diperoleh persentase kemiripannya adalah 21,70%.. Nilai tersebut dibandingkan dengan hasil perhitungan secara manual menggunakan algoritma *edit distance* dengan tahapan sebagai berikut :

1. Percobaan C1 dengan C2

a. Input teks dokumen:

C1 = Cedera yang menimpa striker Spanyol Diego Costa membuatnya diragukan bisa tampil dalam ajang Piala Dunia 2014 nanti. Namun mantan penyerang Atletico Madrid tersebut mengaku dirinya sudah kembali dalam keadaan bugar dan menyatakan diri siap untuk bermain di Piala Dunia nanti.

C2 = Thomas Muller dan Philipp Lahm resmimemparpanjang kontrak mereka di Bayern Munich hanya sehari jelang kick off piala dunia 2014 di Brasil.

b. *Text preprocessing* dokumen *(case folding, tokenizing, filtering, stemming, dan sorting)*

C1 =

2014adaajangatleticobuatbugarcederacostaduniaduniaegomadridmainpialapialaraguserangsiapspanyolstrikertimpa

C2=

2014bayernbrasilduniaharijelangkickkontraklahmmemparpanjangmullermunichoffphilipppialaresmithomas

Jumlah Karakter C1= 106 Karakter

Jumlah Karakter C2= 97 Karakter

Hasil *text preprocessing* dari kedua dokumen mempunyai jumlah karakter yang berbeda yaitu S1 > S2.

c. Menentukan nilai *edit distance*

Dari hasil perhitungan nilai *edit distance* diperoleh 83, yang artinya kedua teks tersebut memiliki perbedaan teks berjumlah 83 karakter.

Perhitungan nilai kemiripan

$$\%\ \text{Nilai Kemiripan} = 1 - \frac{Distance}{Max(S1,S2)} \times 100$$

$$\%\ \text{Nilai Kemiripan} = 1 - \frac{83}{106} \times 100 = 21{,}698\%$$

Hasil dari perhitungan, kedua dokumen tersebut memperoleh nilai kemiripan 21,698%. Kedua teks tersebut memiliki kata-kata yang sangat berbeda, namun masih memiliki nilai kemiripan. Hal tersebut disebabkan oleh cara kerja *edit distance* yang mengukur kemiripan dengan mencocokkan perkarakter bukan perkata. Sehingga apabila ada karakter pada urutan yang sama, maka dihitung sebagai nilai kemiripan.

Berdasarkan hasil perhitungan manual dan hasil yang diperoleh dari sistem yang dikembangan diperoleh hasil prosentasi kemiripan yang sama, sehingga akurasi yang diperoleh 100 %

3.4. Evaluasi Tingkat Kecepatan Sistem

Evaluasi tingkat kecepatan sistem dilakukan terhadap 100 dokumen. Dengan jumlah kata per dokumennya antara 20 sampai 500 kata. Hasil evaluasi diperoleh rata-rata hasil tingkat kecepatan proses pengukuran kemiripan antar dokumen sebagaimana ditunjukan pada Tabel 1.

Tabel 1. Hasil Evaluasi Kecepatan Proses

Banyaknya Kata dalam dokumen	**Rata-rata waktu proses tanpa Peringkasan teks (detik)**	**Rata-rata waktu proses melalui Peringkasan teks (detik)**	**Selisih waktu Proses**	
20 - 40	27,75	27,55	0,20 detik	0,7 %
41 - 100	38,33	31,00	7,33 detik	19,1 %
101 – 200	71,04	38,50	32,54 detik	45,8 %
201- 500	153,14	79,03	74,11 detik	48,39 %

Pada Tabel 1 terlihat bahwa untuk dokumen dengan rata-rata 20 s.d. 40 kata kecepatan yang diperlukan untuk proses pengkurukan kemiripan dokumen adalah hampir sama. Artinya proses peringkasan teks tidak berpengaruh terhadap kecepatan proses pengukuran kemiripan.

Untuk dokumen dengan rata-rata 41 s.d. 100 kata proses pengukuran kemiripan dokumen antara sebelum menggunakan peringkasan teks dan sesudah menggunakan peringkasan teks mengalami peningkatan yaitu rata-rata 7 detik atau 19 % dari proses sebelum menggunakan peringkasan teks. Sedangkan untuk dokumen dengan rata-rata di atas 100 kata proses pengukuran kemiripan dokumen sebelum menggunakan peringkasan teks dan sesudah menggunakan peringkasan teks meghasilkan peningkatan yang cukup besar yaitu rata-rata 50 % dari proses sebelum menggunakan peringkasan teks.

4. KESIMPULAN

Berdasarkan hasil penelitian dapat disimpulkan bahwa :

- Dalam Sistem Temu Kembali Informasi Secara *On Line* untuk Identifikasi Plagiasi karya Ilmiah terdapat tiga modul utama yang terintegrasi yaitu Modul Temu Kembali Informasi, Modul Peringkasan Teks Otomatis dan Modul Pengukuran Kemiripan Dokumen.
- Hasil evaluasi tingkat akurasi Sistem Temu Kembali Informasi Secara *On Line* untuk Identifikasi Plagiasi karya Ilmiah menggunakan metode pembobotan *Vector Space Model dan* pengkuran kemiripan dengan algoritma *edit distance* baik menggunakan peringkasan teks maupun tanpa peringkasan teks memiliki nilai akurasi 80 % s.d 100 % tergantung jumlah kata dalam dokumennya.
- Hasil evaluasi kecepatan proses diperoleh bahwa peringkasan teks otomatis mempengaruhi tingkat kecepatan pengukuran kemripan dokumen untuk dokumen dengan jumlah 100 kata ke atas. Sedangkan untuk dokumen dibawah 100 kata pengaruhnya sangat kecil.

5. SARAN

Permasalahan yang sering muncul pada penelitian yang telah dilaksanakan adalah belum adanya arsitektur baku dan optimal yang dijadikan dasar pada saat *preprocessing* seperti *tokenizing*, *filtering* dan *stemming*. Oleh karena itu pada penelitian selanjutnya perlu disiapkan *preprocessing* menggunakan text mining yang optimal sehinga dapat membantu meningkatkan akurasi hasil identifikasi plagiat karya ilmiah.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada DIPA Kopertis Wilayah IV, Kementrian Pendidikan Nasional yang telah memberi dukungan financial terhadap penelitian ini, sesuai dengan Surat Perjanjian Penugasan Pelaksanaan Program Penelitian Hibah Bersaing Multi Tahun, tahun Anggaran 2014, Nomor : No.1042/K4/KL/2014 tanggal 5 Mei 2014

DAFTAR PUSTAKA

[1]. Kementrian Pendidikan Nasional. 2010. Peraturan Menteri Nasional Republik Indonesia Nomor 17 Tahun 2010 tentang Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi

[2]. Wisnani, 2004. Query Ganda Pada Sistem Temu Kembali Informasi berbasis Jaringan Inferensi : Makara Sains, Vol 8, No 2: 76-84.

[3]. Karmayasa,Oka. 2010. *Implementasi vector space model dan beberapa notasi metode term frequency inverse document frequency (TF-IDF) pada sitem kembali informasi*, Denpasar : Universitas Udayana.

[4]. Nengsih. 2006. Peringkasan Teks Otomatis untuk Dokumen Tunggal Berbahasa Indonesia Menggunakan Algoritma *Term Freqency* and *Lead Method*. Sekolah Tinggi Teknologi Bandung

[5]. Dinajan D, Martine AF, 2007. *A Survey on Automatic Text Summarization*. Language Technologies Institut Carnegie Mellon University

[6]. Dani, T.G. Limandra & L.R.E Adiseputra. 2006. *Deteksi Kemiripan Kode Program dengan Metode Preprocessing dan Perhitungan Leveinshtein* Distance, ISSN : 1411-6286. Prosiding Seminar Ilmiah Nasional Komputer dan Sistem Intelijen (KOMMIT 2006). Universitas Gunadarma, Depok.

[7]. Sempena S, 2009. Algoritma Program Dinamis *Edit Distance* untuk pengecekan Ejaan. Makalah IF 3051 Stategi Algoritma, ITB