TEMU KEMBALI DOKUMEN SUMBER RUJUKAN DALAM KASUS DAUR ULANG TEKS

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

NATHANIEL CLARENCE HARYANTO
71150003

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2019

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI
TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS SECARA ONLINE
UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA

Saya yang bertanda tangan di bawah ini:

NIM	: 71150003
Nama	: NATHANIEL CLARENCE HARYANTO
Prodi / Fakultas	: INFORMATIKA / TEKNOLOGI INFORMASI
Judul Tugas Akhir	: TEMU KEMBALI DOKUMEN SUMBER RUJUKAN DALAM KASUS DAUR ULANG TEKS

bersedia menyerahkan Tugas Akhir kepada Universitas melalui Perpustakaan untuk keperluan akademis dan memberikan **Hak Bebas Royalti Non Ekslusif (*Non-exclusive Royalty-free Right*)** serta bersedia Tugas Akhirnya dipublikasikan secara online dan dapat diakses secara lengkap (full access).

Dengan Hak Bebas Royalti Nonekslusif ini Perpustakaan Universitas Kristen Duta Wacana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk 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 sebenar-benarnya.

Yogyakarta, 18 Juni 2019
Yang menyatakan,

Nathaniel Clarence Haryanto
71150003

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

TEMU KEMBALI DOKUMEN SUMBER RUJUKAN DALAM KASUS DAUR ULANG TEKS

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 18 Juni 2019

NATHANIEL CLARENCE HARYANTO
71150003

HALAMAN PERSETUJUAN

Judul Skripsi	:	TEMU KEMBALI DOKUMEN SUMBER RUJUKAN DALAM KASUS DAUR ULANG TEKS
Nama Mahasiswa	:	NATHANIEL CLARENCE HARYANTO
N I M	:	71150003
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Genap
Tahun Akademik	:	2018/2019

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 18 Juni 2019

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, Dr. Phil.	Antonius Rachmat C., S.Kom.,M.Cs.

HALAMAN PENGESAHAN

TEMU KEMBALI DOKUMEN SUMBER RUJUKAN DALAM KASUS DAUR ULANG TEKS

Oleh: NATHANIEL CLARENCE HARYANTO / 71150003

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 13 Juni 2019

Yogyakarta, 18 Juni 2019
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Antonius Rachmat C., S.Kom.,M.Cs.
3. Gloria Virginia, S.Kom., MAI, Ph.D.
4. Willy Sudiarto Raharjo, S.Kom.,M.Cs.

Dekan

(Budi Susanto, S.Kom., M.T.)

Ketua Program Studi

(Gloria Virginia, Ph.D.)

Halaman Ucapan Terimakasih

Penulis mengucapkan puji dan syukur kepada Tuhan yang Mahaesa atas karuniaNya sehingga laporan tugas akhir berjudul "Temu Kembali Dokumen Sumber Rujukan Dalam Kasus Daur Ulang Teks" dapat selesai tanpa kurang suatu apapun. Laporan ini bertujuan untuk menyelesaikan studi tingkat Strata-1 di program studi Informatika Universitas Kristen Duta Wacana. Penulis berharap laporan ini dapat berguna baik bagi pembaca maupun penulis.

Dalam menyelesaikan penelitian dan penulisan laporan ini, penulis mendapat banyak bantuan baik secara langsung maupun tidak langsung. Oleh karena itu, pada kesempatan ini penulis menyampaikan terimakasih kepada:

1. Ibu Dr. Phil. Lucia Dwi Krisnawati selaku Dosen Pembimbing I yang telah membimbing dan mengarahkan penulis dalam menyelesaikan tugas akhir.
2. Bapak Antonius Rachmat C., S.Kom., M.Cs selaku Dosen Pembimbing II yang telah membimbing dan mengarahkan penulis dalam menyelesaikan tugas akhir.
3. Keluarga penulis yang telah mendukung baik dalam doa maupun dukungan secara moril dan materiil.
4. Teman-teman penulis yang tidak dapat disebutkan satu per satu dalam tulisan ini.

Penulis menyadari laporan ini masih jauh dari kata sempurna. Oleh karena itu, segala saran dan kritik yang disampaikan akan dipertimbangkan dan diharapkan dapat menjadi koreksi baik bagi penulis maupun pihak lain yang terlibat. Akhir kata, penulis sampaikan terimakasih dan semoga laporan ini dapat digunakan sebagaimana mestinya.

Yogyakarta, Mei 2019

Nathaniel Clarence H.

Penulis

Intisari

Temu Kembali Dokumen Sumber Rujukan Dalam Kasus Daur Ulang Teks

Deteksi daur ulang teks biasanya dikhususkan untuk deteksi plagiarisme yang terjadi, terutama pada tulisan akademik. Dokumen kandidat yang diperoleh dari proses temu kembali kurang akurat sehingga proses *text alignment* dan pascapemrosesan menjadi terlalu berat. Diperlukan pengembangan pada proses temu kembali sehingga dapat mengimbangi perkembangan pada proses *text alignment* dan pascapemrosesan.

Dalam tulisan ini, proses temu kembali dikembangkan dengan memanfaatkan pemilihan kata kunci pada algoritma pembuatan ringkasan. Setelah kata kunci ditemukan, dokumen kandidat akan dipilih dengan membandingkan *query* yang didapatkan dengan dokumen sumber menggunakan tf-idf dan kemiripan kosinus. Dokumen yang dipilih kemudian akan diuji dengan *filtering* sebanyak satu hingga sepuluh dokumen dan tanpa *filtering*.

Nilai presisi dan *recall* terbaik dokumen artifisial didapatkan dari pengujian dengan *filtering* sebanyak satu dokumen dengan nilai presisi 0,967 dan *recall* 0,967. Nilai *recall* 0,66 didapatkan dari dokumen tes tersimulasi dengan pengujian tanpa *filtering*. Diperlukan pengujian lebih lanjut untuk data artifisial dokumen dengan menggunakan sinonim. Selain itu, diperlukan pula pengujian lebih lanjut untuk data dokumen tes tersimulasi dengan penyamaran bentuk ringkasan dan pengujian dengan mempertimbangkan semantik dari *query* pencarian.

Kata kunci: pembobotan lokal kata, tf-idf, *recall*, temu kembali

Daftar Isi

Daftar Gambar

Daftar Tabel

Daftar Singkatan

GB	= *gigabyte*
IDE	= *Integrated Development Environment*
NLTK	= *Natural Language Toolkit*
RAM	= *Random Access Memory*
tf-idf	= *term frequency-inverse document frequency*
VSM	= *Vector Space Model*
WLScore	= *word local score* (pembobotan lokal kata)

Intisari

Temu Kembali Dokumen Sumber Rujukan Dalam Kasus Daur Ulang Teks

Deteksi daur ulang teks biasanya dikhususkan untuk deteksi plagiarisme yang terjadi, terutama pada tulisan akademik. Dokumen kandidat yang diperoleh dari proses temu kembali kurang akurat sehingga proses *text alignment* dan pascapemrosesan menjadi terlalu berat. Diperlukan pengembangan pada proses temu kembali sehingga dapat mengimbangi perkembangan pada proses *text alignment* dan pascapemrosesan.

Dalam tulisan ini, proses temu kembali dikembangkan dengan memanfaatkan pemilihan kata kunci pada algoritma pembuatan ringkasan. Setelah kata kunci ditemukan, dokumen kandidat akan dipilih dengan membandingkan *query* yang didapatkan dengan dokumen sumber menggunakan tf-idf dan kemiripan kosinus. Dokumen yang dipilih kemudian akan diuji dengan *filtering* sebanyak satu hingga sepuluh dokumen dan tanpa *filtering*.

Nilai presisi dan *recall* terbaik dokumen artifisial didapatkan dari pengujian dengan *filtering* sebanyak satu dokumen dengan nilai presisi 0,967 dan *recall* 0,967. Nilai *recall* 0,66 didapatkan dari dokumen tes tersimulasi dengan pengujian tanpa *filtering*. Diperlukan pengujian lebih lanjut untuk data artifisial dokumen dengan menggunakan sinonim. Selain itu, diperlukan pula pengujian lebih lanjut untuk data dokumen tes tersimulasi dengan penyamaran bentuk ringkasan dan pengujian dengan mempertimbangkan semantik dari *query* pencarian.

Kata kunci: pembobotan lokal kata, tf-idf, *recall*, temu kembali

Bab I
Pendahuluan

1.1. Latar Belakang

Kasus daur ulang teks merujuk pada plagiarisme sekaligus referensi yang sering ditemukan pada tulisan akademik. Deteksi daur ulang teks biasanya dikhususkan untuk deteksi plagiarisme yang terjadi, terutama pada tulisan jenis ini. Penelitian yang dikhususkan untuk mencari informasi pada pascapemrosesan dalam kasus deteksi daur ulang teks sudah jauh lebih maju dibanding temu kembali. Sayangnya, algoritma *text alignment* dan pascapemrosesan menjadi kurang optimal karena jumlah dokumen kandidat yang dihasilkan proses sebelumnya terlalu banyak.

Untuk mengembangkan lebih lanjut deteksi daur ulang teks yang ada, pengembangan pada proses *text alignment* dan pascapemrosesan perlu diimbangi dengan proses sebelumnya, yaitu temu kembali. Pemilihan kata kunci yang tepat dalam proses ini diperlukan untuk dapat menemukan dokumen kandidat. Selain itu, membatasi jumlah dokumen kandidat yang diambil juga dapat meringankan kerja sistem pada proses *text alignment* dan pascapemrosesan.

Penelitian ini bertujuan untuk mengoptimalkan kerja sistem pada deteksi daur ulang dengan mengembangkan proses temu kembali. Sistem akan menemukan dokumen kandidat dengan memanfaatkan pemilihan kata kunci pada algoritma pembentukan ringkasan. Luaran dari sistem berupa dokumen kandidat yang nantinya akan digunakan dalam proses *text alignment* dan pascapemrosesan untuk deteksi daur ulang teks lebih lanjut.

1.2. Rumusan Masalah

Optimasi proses temu kembali pada deteksi daur ulang teks memiliki beberapa kendala, salah satunya adalah jenis penyamaran teks. Daur ulang teks sendiri dapat dijabarkan dalam empat cara, yaitu *copy and paste*, *copy and shake*, parafrase, serta ringkasan. Untuk dapat menemukan dokumen kandidat yang tepat,

maka diperlukan sistem yang dapat mengatasi teknik daur ulang teks tersebut. Oleh karena itu, permasalahan dapat dijabarkan sebagai berikut:

a. Mengatasi teknik penyamaran dan panjang teks yang berbeda pada dokumen uji.
b. Berapa nilai presisi dan *recall* paling memuaskan yang dapat diperoleh dalam pemilihan dokumen kandidat

1.3. Batasan Masalah

Dalam penelitian ini, kasus penyamaran daur ulang teks yang akan diteliti adalah salin dan tempel, salin dan kocok, serta parafrase. Dokumen sumber yang digunakan berasal dari korpus dalam penelitian (Krisnawati, 2017) dan (Krisnawati & Schülz, 2017) secara luring. Pemilihan *query* akan menggunakan metode pembobotan lokal kata dan kemiripan diukur menggunakan metode *Vector Space Model* (VSM) dengan kemiripan kosinus. Proses seleksi token dari *query* tidak memperhitungkan bobot lokal token. Pencarian dokumen dari *query* yang telah dibentuk tidak memperhitungkan semantik. Aplikasi dibangun menggunakan bahasa pemrograman *python*.

1.4. Tujuan Penelitian

Menemukan kandidat dokumen sumber dengan nilai presisi dan *recall* yang memuaskan. Dokumen uji disamarkan dengan panjang dan tipe daur ulang yang berbeda.

1.5. Manfaat Penelitian

Dokumen kandidat yang ditemukan dengan metode ini dapat digunakan untuk mengembangkan deteksi daur ulang teks yang lebih akurat.

1.6. Metode Penelitian

Penelitian ini akan dijalankan dengan langkah-langkah sebagai berikut:

1. Mencari dokumen sumber yang akan dijadikan rujukan untuk dokumen uji.
2. Melakukan proses prapemrosesan dan *indexing* untuk dokumen sumber dengan menggunakan tf-idf.
3. Melakukan prapemrosesan untuk dokumen uji dan menyusun *query* dengan menggunakan *Word Local Score* oleh Kiabod, Dehkordi, dan Sharafi (Kiabod, Dehkordi, & Sharafi, 2012).
4. Dokumen kandidat akan didapatkan dari melakukan kemiripan kosinus antara *query* yang telah dibentuk dengan hasil *indexing* dari dokumen sumber.
5. Mengambil lima dokumen kandidat dengan nilai kemiripan tertinggi dari hasil kemiripan kosinus yang telah dibuat.
6. Menghitung presisi dan *recall* dari dokumen kandidat yang telah dihasilkan dari proses sebelumnya.

1.7. Sistematika Penelitian

Bagian awal dari penulisan laporan dimulai dengan bab Pendahuluan, Tinjauan Pustaka dan Landasan Teori, serta Metodologi Penelitian. Bab Pendahuluan akan membahas tentang latar belakang dari penelitian, masalah yang ditemui, serta tujuan dan manfaat dari penelitian. Tinjauan Pustaka dan Landasan Teori akan membahas tentang dasar-dasar teori yang digunakan sebagai acuan dalam penelitian ini. Bab ketiga, Metodologi Penelitian, akan menjelaskan secara detil langkah-langkah apa saja yang akan dikerjakan dalam penelitian dan mengapa langkah tersebut diperlukan.

Bagian akhir dari penulisan laporan skripsi akan dilanjutkan dengan bab Hasil dan Pembahasan dan ditutup dengan Kesimpulan dan Saran. Hasil dan Pembahasan akan membahas tentang hasil penelitian serta menjabarkan apa saja yang didapatkan dari penelitian tersebut. Pembahasan pada bab ini akan dijelaskan dengan dasar-dasar teori yang sudah menjadi acuan sebelumnya. Kesimpulan dan Saran akan menyimpulkan dari hasil penelitian serta memberikan saran untuk penelitian yang berhubungan selanjutnya.

Bab V
Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan implementasi dan pengujian yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut:

1. Sistem telah berhasil dibangun dan penyamaran dalam bentuk parafrase, salin dan tempel, serta salin dan kocok dapat diatasi dengan menggunakan metode pemilihan kata penting.
2. Nilai presisi dan *recall* yang memuaskan dapat diperoleh dengan menggunakan *filtering* satu dokumen pada dokumen uji artifisial, sedangkan pada dokumen tes tersimulasi nilai presisi yang memuaskan didapat dari *filtering* satu dokumen dan nilai *recall* dengan proses tanpa *filtering*.
3. Nilai presisi dan *recall* terbaik dokumen artifisial didapatkan dari pengujian dengan *filtering* sebanyak satu dokumen dengan nilai presisi 0,967 dan *recall* 0,967.
4. Nilai *recall* terbaik pada dokumen tes tersimulasi didapatkan dari pengujian tanpa *filtering*. Nilai *recall* yang didapatkan dari proses ini adalah 0,66.

5.2. Saran

Berikut adalah beberapa saran yang dapat digunakan untuk penelitian temu kembali dokumen sumber selanjutnya:

1. Diperlukan pengujian lebih lanjut untuk data artifisial dokumen dengan menggunakan sinonim.
2. Diperlukan pengujian lebih lanjut untuk data dokumen tes tersimulasi dengan penyamaran bentuk ringkasan.
3. Diperlukan penelitian lebih lanjut untuk pencarian dokumen dengan memperhatikan semantik dari *query* pencarian.

Daftar Pustaka

Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., Esposti, M.D. (2009). A Plagiarism Detection Procedure in Three-Steps: Selection, Matches, and "Squares".

Kiabod, M., Dehkordi, M. N., & Sharafi, M. (2012). A Novel Method of Significant Words Identification in Text Summarization. *Journal of Emerging Technologies in Web Intelligence,4*(3). doi:10x.4304/jetwi.4.3.252-258

Krisnawati, L. D., & Schülz, K. U. (2013). Plagiarism Detection for Indonesian Texts. *Proceedings of International Conference on Information Integration and Web-based Applications & Services - IIWAS 13*. doi:10.1145/2539150.2539213

Krisnawati, L. D. (2017). The Use of Phraseword and Local-weighted Terms as Features for Text Reuse and Plagiarism Detection. *Prosiding Seminar Hasil Penelitian Bagi Civitas Akademika UKDW,* 27-44.

Krisnawati, L. D., & Schülz, K. U. (2017). Significant Word-based Text Alignment for Text Reuse Detection. *SAHSS-2017, LEBCSR-17, LERIS-2017, Jan. 31-Feb. 1, 2017 Bali (Indonesia).* doi:10.17758/eirai.f0217102

Leilei, K., Haoliang, Q., Cuixia, D., Mingxing, W., & Zhongyuan, H. (2013). Approaches for Source Retrieval and Text Alignment of Plagiarism Detection—Notebook for PAN at CLEF 2013.

Leilei, K., Zhimao, L., Yong, H., Haoliang, Q., Zhongyuan, H., Qibo, W.,… & Jing, Z. (2015). Source Retrieval and Text Alignment Corpus Construction for Plagiarism Detection—Notebook for PAN at CLEF 2015.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. New York: Cambridge University Press.

Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P.,... & Stein, B. (2013). Overview of the 5th international competition on plagiarism detection. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* (pp. 301-331). CELCT.

Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., & Stein, B. (2015). Overview of the 6th international competition on plagiarism detection.

Potthast, M., Stein, B., Cedeno, A. B., and Rosso, P. An Evaluation Framework for Plagiarism Detection. In *Proceedings of 23th International Conference on Computational Linguistics (COLING 2010)* (August 2010), pp. 85–98. Beijing, China.

Stein, B., Eissen, S.M., Potthast, M. (2007). Strategies for Retrieving Plagiarized Documents