Implementasi Algoritma *Merging Context Seeds* untuk *Plagiarism Detection*

Yusuf Anugrah Putra Aditama[1], Ir. Moch. Arif Bijaksana[2], Mohamad Syahrul Mubarok[3]

[1,2,3]Teknik Informatika, Universitas Telkom; Jl. Telekomunikasi, Dayeuhkolot, Sukapura, Dayeuhkolot, Bandung, Jawa Barat 40257

Abstrak

Plagiat merupakan masalah yang sering ditemukan di masyarakat, bahkan menurut survey 89% responden sering menemukan kasus plagiat pada bidangnya masing-masing. Tindak plagiat ini dapat berupa mengambil tulisan orang lain yang digunakan untuk kepentingan diri sendiri. Adapun salah satu pendekatan yang dapat dilakukan untuk mendeteksi tindak plagiat ini adalah dengan *Text Alignment.* Sehingga pada penelitian ini diusung salah satu metode yaitu *Merging Context Seeds* yang bekerja dengan cara menggabungkan ciri yang ada pada *suspicious document* dan *source-document* dengan metode ekstraksi ciri *n-skip-k-grams*. Dengan diimplementasikannya metode *Merging Context Seeds*.

Kata kunci*— merging context seeds, seeds, merge.*

1. PENDAHULUAN

1.1 Latar Belakang

Menurut hasil survey yang dilakukan oleh iThenticate[3], sebanyak 89% res-ponden yang ditemui menjawab sering menemui kasus plagiat pada bidangnya masing-masing. Dan lebih dari 25% responden menyatakan bahwa plagiat merupakan masalah serius yang harus diselesaikan. Namun dengan makin banyaknya dokumen yang terkumpul akan semakin sulit mendeteksi tindak plagiat secara manual. Text alignment adalah solusi yang ada untuk menye-lesaikan masalah plagiat yang ada, dengan cara membangkitkan bagian pada dua buah dokumen yang terindikasi plagiat.

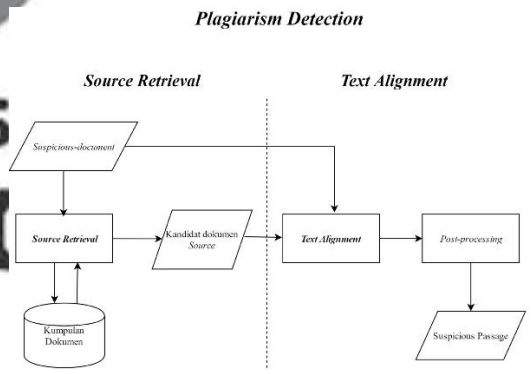
Paragraf atau kalimat yang serupa yang digunakan pada dua buah dokumen dapat dideteksi dari penggunaan kata dan penataannya. Hal ini yang menjadi alasan pendekatan text alignment dapat dilakukan, yaitu dengan menjajarkan seluruh tur yang ada pada dua buah dokumen dan mencari irisan antara dua buah dokumen tersebut. Dari seluruh metode text alignment yang ada, dipilih Merging Context Seeds yang merupakan salah satu metode yang diajukan pada PAN[4] yang memiliki nilai perfomansi plagdet 0.826. Merging context seeds terfokus kepada seeds yang merupakan tur yang beririsan antara dua buah dokumen, dan melakukan clustering atau pengelompokan data pada seeds yang ada. Dari kelompok data yang didapat, dipilih yang merupakan tindak plagiat.

Permasalahan yang akan diselesaikan pada penilitian ini adalah, bagaimana mengimplementasikan algoritma merging context seeds dan membangun sistem yang mampu mendeteksi tindak plagiat dari berbagai tipe tindak plagiat yang ada secara akurat.

2. LANDASAN TEORI DAN METODE PENELITIAN

2.1 *Plagiarism Detection*

Plagiarism Detection adalah solusi untuk masalah plagiarism. Tujuan utama dari Plagiarism Detection ini adalah mengidenti kasi suatu dokumen yang disebut dengan suspicious-document, apakah dokumen tersebut mempunyai teks atau bagian yang diambil dari satu atau beberapa dokumen lain (source-document). Dan membuktikan apabila terdapat teks yang diambil dari source-document di suspicious-document. Gambar 2.1 menunjukan alur Plagiarism Detection secara umum[4, 1].



Gambar 2.1: Alur keseluruhanPlagiarism Detection[1]

2.2 Merging Context Seeds

Merging Context Seeds merupakan salah satu metode dengan pendekatan Text Alignment. Metode ini memiliki tahapan yang serupa pada pendekatan Text Alignment[8, 1] pada umumnya.

Metode ini akan mengolah dokumen pada level karakter. Diketahui sebuah dokumen terdiri dari bagian-bagian (huruf, kata, atau kalimat) yaitu P , di-mana P = fx_i : 0 a i < b ng dimana x_i = (c; i); c 2 C. C merupakan kumpulan

simbol, dan i merupakan letak kemunculan karakter. P juga dapat dinotasikan dengan P = [x_{ai} ; x_{bi}].

2.2.1 Seed generation

Terdapat dokumen yang terindikasi bernama dokumen X (suspicious-document) dan dokumen sumbernya bernama dokumen Y (source-document). Setiap karakter yang ada pada tiap dokumen dipetakan kedalam index map untuk meng-etahui letak kemunculan karakter dan tur nantinya.

Preprocessing

Kedua dokumen akan melalui proses preprocessing dengan tahapan yaitu :

1. Menghapus white space dan enter space.
2. Mengubah seluruh karakter yang ada menjadi huruf kecil.
3. Menghapus seluruh karakter yang tidak termasuk kedalam alphanumeric character.
4. Menghapus seluruh *stopwords.*

Ekstraksi Ciri

Dokumen yang telah melalui tahap preprocessing kemudian diekstraksi cirinya dengan menggunakan k-skip-n-grams dengan nilai k = 1; 2; 3; 4 dan n = 2, atau dapat disebut dengan 1 4skip bigram.

Feature Relevance Filtering

Pada tahap ini tur yang ada akan dihitung jumlah kemunculannya pada do-kumennya. Apabila jumlah kemunculan tur sesuai dengan threshold, maka fitur akan disimpan, apabila melebihi threshold maka fitur akan dihapus.

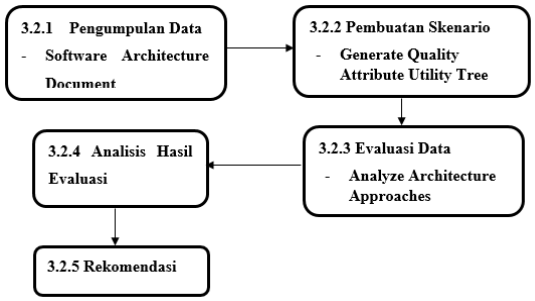
Seed Generation

Seluruh tur yang ada pada dokumen X dan dokumen Y dipetakan menjadi index map dari XY. Lalu dari index map XY karakter yang mempunyai tur yang sama pada kedua dokumen disebut sebagai passage reference, sedangkan kumpulan dari passage reference ini dapat disebut sebagai seed set. Passage reference ini merupakan titik awal untuk proses dari pendekatan text alignment pada metode Merging Context Seeds.

3. HASIL DAN PEMBAHASAN

3.1 Gambaran Umum

Pada penelitian ini, penulis akan melakukan analisis terhadap hasil evaluasi dan sensitivitas Sistem RFID Universitas Telkom dengan menggunakan ATAM. Penggunaan ATAM lebih lanjut dilakukan pada bagian pengujian dan analisis. Pada analisis ini dilakukan dengan beberapa tahap, yaitu: pengumpulan data, pembuatan skenario, evaluasi data, analisis hasil evaluasi, dan rekomendasi. Dalam analisis ini tidak semua langkah dalam ATAM dilakukan, hal ini dikarenakan perlu adanya team evaluasi yang banyak untuk melakukan evaluasi dengan ATAM. Sehingga dalam tugas akhir ini hanya sebagian dari langkah ATAM yang dilakukan.



Gambar 3-1 Tahapan Proses

Penjelasan mengenai gambaran umum yang terdapat pada gambar 3-1, aktifitas yang dilakukan adalah sebagai berikut:

1. Data yang digunakan pada tugas akhir ini adalah *Software Architecture Document* Sistem RFID Universitas Telkom yang diperoleh dari Direktorat Sistem Informasi Telkom University. Dalam dokumen ini, penggambaran sistem RFID Universitas Telkom digambarkan dengan konsep '4+1' model view. Architectural view yang digambarkan pada dokumen ini adalah:
   - *Logical view*
   - *Process view*
   - *Deployment view*
   - *Implementation view*
   - *Use Case view*
2. Dari data yang didapatkan, dibuat pendekatan arsitektur sistem untuk menentukan faktor kualitas yang akan dilakukan evaluasi. Dari arsitektur sistem yang ada dapat ditentukan utility tree untuk faktor kualitas sistem. Faktor kualitas yang akan dilakukan analisis adalah reliability, performance, dan usability. Dari masing-masing faktor kualitas tersebut kemudian dibuat skenario sebagai acuan untuk melakukan analisis dan evaluasi arsitektur sistem. Tiap skenario kualitas sistem terdiri dari source, stimulus, artifact, environment, response dan response measure.
3. Dari skenario yang dibuat akan dibandingkan dengan arsitektur sistem, apakah arsitektur tersebut sudah dapat menangani permasalahan kualitas sistem.
4. Setelah dilakukan evaluasi sistem, hasil evaluasi akan dianalisis dengan arsitektur sistem yang sudah dibuat. Analisis ini dilakukan untuk

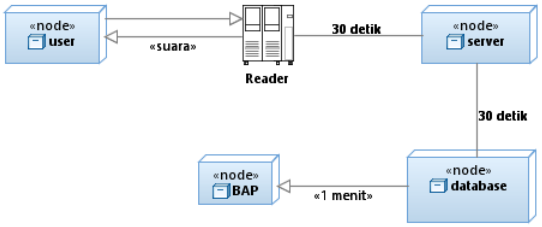
mengetahui poin tradeoff, resiko dan non resiko dari pendekatan arsitektur sistem

5. Dari hasi analisis evaluasi arsitektur sistem, didapat keputusan desain arsitektural untuk perbaikan sistem RFID universitas Telkom. Keputusan desain arsitektural bisa merupakan bentuk arsitektur baru untuk sistem ataupun perbaikan dari arsitektur sistem yang telah ada.

3.2 Analisis Tradeoff

Dari arsitektural sistem yang sudah ada, diberikan penggambaran arsitektur alternative untuk perbaikan sistem. Arsitektur alternative ini kemudiaan dilakukan analisis terhadap skenario dari masing-masing atribut kualitas.

a. Alternative 1



Gambar 3-2 Alternative Desain 1

Pada alternative desain yang pertama, notifikasi pada reader tetap menggunakan bunyi 1 kali atau 2 kali dan waktu pembacaan data dari reader ke database tetap sekitar 1 menit. Perbaikan dilakukan hanya pada waktu pembuatan BAP dari database pada igracias. Waktu untuk munculnya BAP pada igracias dibuat lebih cepat yakni selama 1 menit.

Desain arsitektur ini akan membuat waktu munculnya BAP pada igracias dosen menjadi lebih cepat meskipun dosen baru melakukkan tapping setelah semua mahasiswa melakukkan tapping. Akan tetapi notifikasi yang diberikan reader hanya bunyi 1 kali atau 2 kali, sehingga mahasiswa harus melakukan pengecekan kembali pada akun igracias masing-masing apakah presensi mereka diterima atau tidak. Selain itu waktu pembacaan data dari reader hingga ke database tetap sekitar 1 menit, sehingga jika waktu pembacaan data telah habis maka reader akan mengeluarkan bunyi 1 kali yang berarti data ditolak padahal presensi mahasiswa tersebut diterima.

b. Alternative 2

Gambar 3-3 Alternative Desain 2

Pada alternative yang kedua, perbaikan dilakukan pada reader yang memberikan notifikasi berupa tulisan apakah presensi diterima atau tidak dan waktu pembacaan data pada reader hingga database menjadi lebih lama sekitar 2 menit. Tapi waktu untuk munculnya BAP pada igracias dosen tetap sekitar maksimal 3 menit.

Dengan desain ini, mahasiswa tidak perlu melakukan pengecekan kembali ke akun igracias masing-masing karena sudah ada notifikasi pada reader bahwa presensi mahasiswa diterima atau tidak. Selain itu waktu pembacaan data dari reader hingga ke database menjadi lebih lama sekitar 2 menit agar reader tidak mengeluarkan notifikasi ambigu apakah presensi diterima atau tidak. Tapi waktu munculnya BAP pada igracias dosen apabila dosen melakukan tapping setelah semua mahasiswa melakukan tapping tetap lama sekitar 3 menit, karena dibutuhkan waktu untuk melakukan sinkronisasi data sesuai ruangan dan jadwal mata kuliah dosen tersebut.

1.3 Rekomendasi

Dari hasil analisis tradeoff kedua arsitektur alternative, maka dipilih arsitektur alternative kedua untuk dijadikan perbaikan arsitektur sistem RFID pada universitas Telkom. Hal ini dikarenakan pada arsitektur alternative kedua, mahasiswa tidak lagi harus melakukan pengecekan presensi lagi ke igracias karena reader sudah mengeluarkan notifikasi bahwa absensi mahasiswa tersebut diterima. Selain itu reader juga tidak akan memberikan notifikasi ambigu apakah data diterima atau tidak karena waktu pengecekan data dari reader hingga ke database lebih lama sekitar 2 menit. Dan waktu munculnya BAP selama 3 menit merupakan waktu normal setelah dosen melakukan tapping setelah semua atau beberapa mahasiswa melakukan tapping.

Berdasarkan hasil tradeoff tersebut maka rekomendasi untuk perbaikan sistem adalah dengan menambahkan notifikasi seperti tulisan pada reader bahwa presensi mahasiswa diterima dan menambahkan waktu untuk pengecekan data pada reader hingga ke database menjadi sekitar 2 menit agar reader tidak memberikan notifikasi ambigu apakah presensi mahasiswa ditolak atau diterima.

4. KESIMPULAN

Sistem yang dibangun teralu sensitif dalam mengenali bagian yang ada pada dokumen untuk tipe plagiat tertentu. Hal ini terbukti dari nilai False Posi-tive yang tinggi pada pengujian untuk tipe plagiat No Obfuscation, Random Obfuscation, dan Translation Obfuscation. Selain itu hal ini didapatkan kare-na pada tipe plagiat No Plagiarism sistem masih menemukan adanya tindak plagiat pada pasangan

dokumen. Walaupun dari tipe plagiat No Plagiarism hanya 4%.

Nilai False Positive yang tinggi ini dikarenakan pada proses merge, seed yang ada pada bagian yang di plagiat dan bagian yang tidak di plagiat ikut terga-bung. Sehingga banyak bagian yang tidak plagiat, dianggap sebagai plagiat. Hal ini juga dapat dikarenakan masih banyak tur yang tidak relevan yang terbangun. Ataupun penggunaan parameter yang kurang cocok untuk seluruh dokumen yang di proses.

Pada tipe plagiat Random Obfuscation, dan Translation Obfuscation sistem juga tidak mampu menangani adanya perubahan pola kata pada bagian yang diplagiat sehingga teralu banyak bagian plagiat yang tidak terdeteksi oleh sis-tem.

Sedangkan pada tipe plagiat Summary Obfuscation, dimana tipe plagiat ini merangkum bagian pada dokumen source, sistem tidak dapat mengenali ada-nya tindak plagiat pada tipe plagiat ini secara baik.

5. SARAN

1. Menambahkan atau menggunakan parameter yang lebih baik daripada yang penulis gunakan pada penilitian ini. Seperti pada prosses gene-rasi tur, agar tur yang dihasilkan lebih relevan dibanding tur yang dihasilkan sistem saat ini.
2. Menggunakan ekstraksi ciri yang berbeda agar mengurangi seed yang tidak relevan.
3. Menambahkan metode lain untuk membantu menangani tipe plagiat Summary Obfuscation dimana parafrase sulit ditemukan oleh sistem yang dibangun saat ini.

DAFTAR PUSTAKA

[1] Ahsan, K., Shah, H., & Kingston, P. (2010). RFID Applications: An Introductory and Exploratory Study. *IJCSI International Journal of Computer Science Issues Vol. 7*.

[2] Anonim. (2014, April 8). *Architecture Tradeoff Analysis Method*. Retrieved from https://www.sei.cmu.edu/architecture/tools/evaluate/atam.cfm

[3] Babar, M. A., & Gorton, I. (n.d.). Comparison of Scenario-Based Software Architecture Evaluation Methods.

[4] Chao, C.-C., Yang, J.-M., & Jen, W.-Y. (2007). Determining technology trends and forecasts of RFID by a historical review and bibliometric analysis from 1991 to 2005. *Technovation 27*.

[5] Choi, H., & Yeom, K. (2002). An Approach to Software Architecture Evaluation with the 4+1 View Model of Architecture. *Proceedings of the Ninth Asia-Pacific Software Engineering Conference*. IEEE.

[6] Clements, P., Bergey, J., & Mason, D. (2005). Using the SEI Architecture Tradeoff Analysis Method to Evaluate WIN-T: A Case Study. *Software Architecture Technology Initiative*.

[7] Dobrica, L., & Niemela, E. (2002). A Survey on Software Architecture Analysis Methods. *IEEE Transactions on Software Engineering, Vol. 28, No. 7*.

[8] Domdouzis, K., Kumar, B., & Anumba, C. (2007). Radio-Frequency Identification (RFID) applications: A brief introduction. *Advanced Engineering Informatics 21*.

[9] Kazman, R., Bass, L., Lattanze, T., & Northrop, L. (2005). A Basis for Analyzing Software Architecture Analysis Method. *Software Quality Journal*, 329-355.

[10] Kazman, R., Clements, P., & Bass, L. (2012). *Software Architecture In Practice Third Edition.* Addison-Wesley Professional.

[11] Maréchaux, J.-L. (2005). Developing a J2EE Architecture with Rational Software Architect Using the Rational Unified Process.

[12] Smith, D., & Merson, P. (2003). Using Architecture Evaluation to Prepare a Large Web Based System for Evolution. *Proceedings of the Fifth IEEE International Workshop on Web Site Evolution.* IEEE.

[13] Sun, C., & Jiang, F. (2013). Research on RFID Applications in Construction Industry. *Journal of Networks, Vol. 8, No. 5*.

[14] Zhu, L., Aurum, A., Gorton, I., & Jeffery, R. (2005). Tradeoff and Sensitivity Analysis in Software Architecture Evaluation Using Analytic Hierarchy Process. *Software Quality Journal, 13*.