PERBANDINGAN PENDEKATAN DETEKSI PLAGIARISM DOKUMEN DALAM BAHASA INGGRIS

[1]Ana Kurniawati
[2]I Wayan Simri Wicaksana

[1,2]Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma
({ana,iwayan}@staff.gunadarma.ac.id)

ABSTRAK

Praktik plagiarisme (penjiplakan) dalam penulisan penelitian cukup sering terjadi di kalangan akademisi, khususnya mahasiswa. Plagiarisme adalah tindakan penyalahgunaan, pencurian/perampasan, penerbitan, pernyataan, atau menyatakan sebagai milik sendiri sebuah pikiran, ide, tulisan, atau ciptaan yang sebenarnya milik orang lain. Di kalangan mahasiswa yang selalu berinteraksi dengan komputer yang mempermudah praktik plagiat mengingat adanya fasilitas untuk menyalin dan mengubah teks (copy and paste) dan fasilitas koneksi yang memungkinkan untuk mengakses hasil karya orang lain secara bebas melalui internet, praktik plagiarisme ini sering dilakukan. Untuk meminimalisasi praktik plagiarisme, diperlukan pendeteksian terhadap penulisan. Pada makalah ini akan dipaparkan hasil analisis dua metode untuk mendeteksi plagiarisme dokumen. Aspek kelebihan dan kekurangan dari pendekatan-pendekatan tersebut digunakan sebagai tolak ukur untuk membangun pendekatan yang lebih optimal untuk mendeteksi plagiarisme dokumen.

***Kata Kunci** : deteksi, dokumen, dokumen fingerprinting, perbandingan, plagiarism,*

1. PENDAHULUAN

Praktek plagiarisme (penjiplakan) dalam penulisan penelitian cukup sering terjadi di kalangan akademisi, khususnya mahasiswa. Plagiarisme adalah tindakan penyalahgunaan, pencurian/perampasan, penerbitan, pernyataan, atau menyatakan sebagai milik sendiri sebuah pikiran, ide, tulisan, atau ciptaan yang sebenarnya milik orang lain. Di kalangan mahasiswa yang selalu berinteraksi dengan komputer yang mempermudah praktik plagiat mengingat adanya fasilitas untuk menyalin dan mengubah teks (*copy and paste*) dan fasilitas koneksi yang memungkinkan untuk mengakses hasil karya orang lain secara bebas melalui internet, praktik plagiarisme ini sering dilakukan. Untuk meminimalisasi praktik plagiarisme, diperlukan pendeteksian terhadap penulisan.

Untuk mengatasi praktik plagiarisme, tidaklah cukup hanya mengingatkan kepada mahasiswa bahwa tindakan plagiarisme tidak baik dilakukan. Pendeteksian praktik plagiarisme merupakan solusi yang sebaiknya dilakukan sehingga tindakan curang tersebut dapat diminimalisasi. Untuk meminimalisasi praktik plagiarisme, diperlukan pendeteksian terhadap penulisan. Namun, proses pendeteksian secara manual sulit untuk dilakukan karena jumlah penulisan yang banyak. Sehingga diperlukan sistem untuk mendeteksi plagiarisme.

Metode untuk mendeteksi plagiarisme dapat di klasifikasikan

menjadi tiga metode [2] yaitu metode perbandingan teks lengkap, metode dokumen fingerprinting dan metode kesamaan kata kunci. Dalam paper ini akan diuraikan untuk metode dokumen fingerprinting. Penelitian-penelitian yang telah dilakukan untuk dokumen fingerprinting menggunakan algoritma yaitu Running Kap Robin Matching and Greedy String Tiling (RKR-GST)[7], pendekatan Manber [5], pendekatan Heintze dan algoritma winnowing [8,11]

Pada makalah ini akan dipaparkan hasil analisis pendekatan atau metode yang ada untuk mendeteksi plagiarisme dokumen. Pendekatan atau metode yang dipaparkan adalah pendekatan Manber dan algoritma winnowing. Analisis yang dilakukan adalah dengan melihat aspek kelebihan dan kekurangan dari pendekatan-pendekatan atau metode-metode tersebut.

Penelitian-penelitian yang membahas tentang perbandingan atau evaluasi dari berbagai metode mendeteksi plagiarisme telah dilakukan. Mengacu kepada [10], perbandingan yang dilakukan terhadap tool deteksi plagiarisme dengan melihat beberapa atribut seperti tipe dari dokumen, tipe dari kumpulan dokumen, dan atribut yang lain seperti pengguna yang menggunakan tool tersebut. Penelitian yang lain tentang perbandingan metode mendeteksi plagiarisme dilakukan oleh J. Evan, [4] dipaparkan bahwa metode mendeteksi plagiarisme di klasifikasikan menjadi dua kategori yaitu manual dan atomatis. Untuk kategori otomatis dapat di bagi menjadi tiga macam yaitu metode kuis, metode tipe penulisan dan perbandingan dengan sumber asli. Ketiga metode inilah yang di bandingkan.

2. METODE TERKINI PADA DETEKSI PLAGARISME

Pengertian Plagiarisme

Mendahului pembahasan lebih mendalam dari topik yang diangkat, penulis menjabarkan definisi yang digunakan dalam menyatakan tindakan plagiarisme. Plagiarisme adalah tindakan penyalahgunaan, pencurian/perampasan, penerbitan, pernyataan, atau menyatakan sebagai milik sendiri sebuah pikiran, ide, tulisan, atau ciptaan yang sebenarnya milik orang lain. [1]

Sistem pendeteksi plagiarisme dapat di kembangkan untuk :

1. Data teks seperti essay, artikel, jurnal, penelitian dan sebagainya.
2. Dokumen teks yang lebih terstruktur seperti bahasa pemrograman.

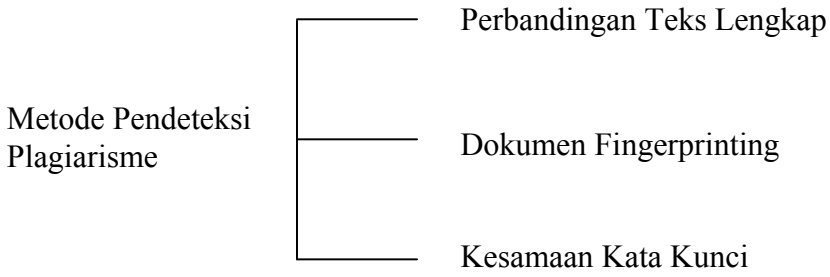
Tipe-Tipe Plagiarisme

Beberapa tipe plagiarisme yaitu : [9]

1. Word-for-word plagiarism
   Menyalin setiap kata secara langsung tanpa diubah sedikitpun.
2. Plagirism of authorship
   Mengakui hasil karya orang lain sebagai hasil karya sendiri dengan cara mencantumkan nama sendiri menggantikan nama pengarang yang sebenarnya.
3. Plagiarism of ideas
   Mengakui hasil pemikiran atau ide orang lain.
4. Plagiarism of sources
   Jika seorang penulis menggunakan kutipan dari penulis lainnya tanpa mencantumkan sumbernya.

Metode Pendeteksi Plagiarisme

Metode Pendeteksi Plagiarisme di bagi menjadi tiga bagian yaitu metode perbandingan teks lengkap, metode dokumen fingerprinting, dan metode kesamaan kata kunci. Metode pendeteksi plagiarisme dapat di gambarkan sebagai berikut : [2]



Gambar 1. Klasifikasi Metode Pendeteksi Plagiarisme

Berikut ini penjelasan dari masing-masing metode dan algoritma pendeteksi plagiarisme. Ketiga metode tersebut adalah :

1. Perbandingan Teks Lengkap
   Metode ini di terapkan dengan membandingkan semua isi dokumen. Dapat diterapkan untuk dokumen yang besar. Pendekatan ini membutuhkan waktu yang lama tetapi cukup efektif, karena kumpulan dokumen yang diperbandingkan adalah dokumen yang di simpan pada penyimpanan lokal. Metode perbandingan teks lengkap tidak dapat diterapkan untuk kumpulan dokumen yang tidak terdapat pada dokumen lokal. Algoritma yang digunakan pada metode ini adalah algoritma brute force, algoritma edit distance, algoritma boyer moore dan algoritma lavenshtein distance

2. Dokumen Fingerprinting
   Dokumen fingerprinting merupakan metode yang digunakan untuk mendeteksi keakuratan salinan antar dokumen, baik semua teks yang terdapat di dalam dokumen atau hanya sebagian teks saja. Prinsip kerja dari metode dokumen fingerprinting ini adalah dengan menggunakan teknik hashing. Teknik hashing adalah sebuah fungsi yang mengkonversi setiap string menjadi bilangan.

3. Kesamaan Kata Kunci.
   Prinsip dari metode ini adalah mengekstrak kata kunci dari dokumen dan kemudian di bandingkan dengan kata kunci pada dokumen yang lain. Pendekatan yang digunakan pada metode ini adalah teknik dot.

Pendekatan Metode Dokumen Fingerprinting[8,11]

Seperti yang telah diuraikan sebelumnya, prinsip kerja dari metode dokumen fingerprinting ini adalah dengan menggunakan teknik hashing. Teknik hashing adalah sebuah fungsi yang mengkonversi setiap string menjadi bilangan kemudian menyimpannya dalam sebuah skema atau bagan. Ide dasar metode dokumen fingerprinting adalah menyimpan skema atau bagan kecil yang berisi kumpulan angka atau bilangan yang akan dibandingkan dengan skema atau bagan antar dua dokumen. Skema digital dokumen fingerprinting terdiri dari sejumlah posisi yang diberi tanda di dalam dokumen, algoritma fingerprinting yang akan memilih tanda yang akan di tambahkan untuk setiap posisi tergantung pada jumlah salinan.

Secara umum prinsip kerja dari metode dokumen fingerprinting adalah dengan tahapan sebagai berikut :

1. Asumsikan teks adalah string s yang panjangnya t.
2. Hilangkan tanda baca dan spasi
3. Sebelum melakukan fungsi hash dengan menggunakan notasi k-gram. k-gram merupakan substring yang berdampingan dari panjang k. Membagi dokumen menjadi k-gram, dimana k merupakan parameter yang di pilih pengguna.
4. Lakukan fungsi hash untuk setiap k-grams
5. Memilih beberapa hasil hash menjadi dokumen fingerprinting.

Permasalahan yang muncul adalah bagaimana memilih fingerprint dari hasil hash. Terdapat beberapa pendekatan untuk menangani masalah tersebut. Pada makalah ini akan di bahas dua pendekatan yaitu pendekatan Manber dan algoritma Winnowing.

Pendekatan Manber [5]

Pendekatan Manber merupakan salah satu pendekatan pada metode dokumen fingerprinting. Seperti yang telah diuraikan sebelumnya bahwa prinsip kerja dari metode dokumen finger printing ada lima langkah. Yang menjadi permasalahan adalah pada langkah yang ke-lima yaitu bagaimana memilih hasil dari proses hashing. Pendekatan Manber memilih hasil dari proses hashing dengan cara memilih semua hasil hashing dengan yang memenuhi kriteria 0 mod p. Dengan cara ini fingerprints yang terpilih tidak tergantung dari posisinya. Pendekatan ini mudah untuk diimplementasikan.

Pendekatan Metoda Algoritma Winnowing [8]

Algoritma winnowing merupakan algoritma dokumen fingerprinting yang digunakan untuk mendeteksi salinan dokumen dengan menggunakan teknik hashing. Untuk meng-hash dokumen dengan menggunakan k-gram, panjang substring k dimana k merupakan nilai yang dipilih oleh pengguna. Dokumen akan dibagi ke dalam k-gram yang mungkin dan kemudian k-gram tersebut akan di hash. Untuk memilih fingerprint dari hasil yang di hash, dilakukan pembagian dengan menggunakan window w, dan dipilih nilai yang paling kecil.

Difinisi Winnowing :[11]
Dari setiap window dipilih nilai hash yang paling minimum atau kecil. Jika terdapat nilai minimum lebih dari satu nilai, maka pilih dari window sebelah kanan. Kemudian simpan semua hasil hash yang telah dipilih yang merupakan fingerprint dokumen.

Diberikan kumpulan dokumen, ingin menemukan substring yang sama diantara dokumen-dokumen tersebut, propertis yang dilakukan adalah :

1. Jika terdapat string yang sama yang panjangnya sama dengan panjang t, dimana t merupakan jaminan ambang nilai yang ditentukan, maka pencocokan terdeteksi.
2. Tidak dapat mendeteksi beberapa pencocokan jika lebih pendek dari gangguan nilai ambang, k.

Nilai konstan t dan $k \leq t$ dipilih oleh pengguna. Menghindari pencocokan string yang sama dibawah nilai gangguan nilai ambang dengan mempertimbangkan hash k-grams.

3. DISKUSI

Pada bagian ini akan di paparkan contoh dan cara penyelesaiannya dengan penggunakan pendekatan Manber dan winnowing. Diberikan contoh teks sebagai berikut :

"A do run run run, a do run run"

Penyelesaian dengan menggunakan metode dokumen fingerprinting adalah sebagai berikut :

1. Teks yang akan di deteksi yaitu ***A do run run run, a do run run***
2. Hilangkan tanda baca dan spasi.
   Pada langkah 1 diberikan contoh teks yaitu ***A do run run run***, a ***do run run***. Kemudian pada langkah kedua adalah menghilangkan tanda baca, huruf besar diganti huruf kecil dan spasi pada teks. Hasil dari langkah kedua dapat dilihat pada gambar 2 berikut ini :

adorunrunrunadorunrun

Gambar 2. Teks tanpa tanda baca dan spasi.

3. Kemudian dari hasil langkah kedua, teks tersebut di bagi menjadi 5-grams. Hasil dari langkah ketiga dapat dilihat pada gambar 3 berikut ini.

adoru dorun orunr runru unrun nrunu runru
unrun nruna runad unado nador adoru dorun
orunr runru unrun

Gambar 3. Teks dengan 5-grams

4. Lakukan hashing
   Pada langkah keempat ini, hasil dari langkah ketiga akan di hash. Hasil dari langkah keempat ini dapat dilihat pada gambar 4 berikut ini :

77 74 42 17 98 50 17 98 8 88 67 39 77 74 42 17 98

Gambar 4. Hasil Hashing

5. Memilih hasil hash.
   Untuk memilih hasil hash akan di selesaikan dengan 2 pendekatan yaitu pendekatan Manber dan winnowing. Pendekatan pertama yang akan di bahas adalah pendekatan monber.

   Pendekatan Manber

Dari hasil pada langkah keempat atau dari hasil hashing, akan dipilih mana yang menjadi finger print. Pemilihan dilakukan dengan cara 0 mod p, dimana p adalah 4 sehingga 0 mod 4. Maka hasilnya adalah :

72 8 88 72

Gambar 5. Hasil hash yang dipilih dengan menggunakan 0 mod 4.

Algoritma Winnowing
Pada algoritma winnowing, untuk menghasilkan fingerprint, terdapat 3 langkah yaitu :

1. Untuk memilih hasil yang telah di hash, dilakukan dengan menggunakan membagi ke window w dengan panjang yang ditentukan oleh pengguna. Kemudian pilih nilai yang minimum, dan beri tanda dengan menebalkan. Hasilnya dapat dilihat pada gambar 6 berikut ini.

(77, 74, 42, **17**)	(74, 42, 17, 98)
(42, 17, 98, 50)	(17, 98, 50, **17**)
(98, 50, 17, 98)	(50, 17, 98, **8**)
(17, 98, 8, 88)	(98, 8, 88, 67)
(8, 88, 67, 39)	(88, 67, **39**, 77)
(67, 39, 77, 74)	(39, 77, 74, 72)
(77, 74, 42, **17**)	(74,42, 17, 98)

Gambar 6. Window hash dengan panjang 4

2. Setelah itu memilih hasil yang telah di bagi menjadi window. Hasilnya dapatdilihat pada gambar 7 berikut ini :

17 17 8 39 17

Gambar 7. Fingerprint yang dipilih dengan menggunakan winnowing

3. Setelah itu ditambahkan informasi posisi fingerprint di dalam dokumen. Hasilnya dapat dilihat pada gambar 8 berikut ini. Gambar 8 menampilkan kumpulan pasangan fingerprint dan posisi untuk contoh ini.

[17,3] [17,6] [8,8] [39,11] [17,15]

Gambar 8. Hasil fingerprint dengan informasi posisi

Perbandingan Pendekatan Manber dan Algoritma Winnowing

Dari penyelesaian yang diuraikan diatas dapat dilihat perbedaan dari kedua pendekatan tersebut. Perbedaan-perbedaan tersebut adalah sebagai berikut :

1. Jumlah Langkah
   Jika dilihat dari jumlah langkah atau tahapan penyelesaian dari

kedua pendekatan tersebut, maka pendekatan pertama yaitu pendekatan Manber lebih sedikit yaitu satu langkah atau satu tahap. Sedangkan pada pendekatan kedua yaitu algoritma winnowing, lebih banyak yaitu tiga langkah atau 3 tahap.

2. Informasi dari hasil atau output
   Jika dilihat dari output yang dihasilkan dari kedua pendekatan tersebut, maka pendekatan kedua yaitu algoritma winnowing lebih informatif karena selain menghasilkan hasil fingerprint juga terdapat informasi yang lain yaitu informasi posisi. Sedangkan pendekatan pertama yaitu pendekatan Manber tidak terdapat informasi posisi.
3. Kekurangan
   Kelemahan dari pendekatan Manber tidak memberikan jaminan bahwa kecocokan antar dokumen terdeteksi. Hal ini dikarenakan dokumen terdeteksi jika hanya hasil hash memenuhi nilai 0 mod p. Dari fungsi hash yang dipilih terdapat kemungkinan terjadinya benturan sangat kecil. Kelemahan dari algoritma winnowing adalah waktu prosesnya lebih lama.
4. Kelebihan
   Kelebihan dari pendekatan manber adalah proses penyelesaiannya sederhana, dengan waktu yang lebih cepat, dapat dengan mudah di implementasikan. Sedangkan pendekatan kedua yaitu winnowing mempunyai kelebihan yaitu hasilnya lebih informatif karena terdapat informasi posisi selain itu pendekatan ini memberikan jaminan terdeteksinya dokumen.

4. KESIMPULAN

Pada metode pendeteksi plagiarisme yaitu metode dokumen fingerprinting menggunakan teknik hashing. Dari kedua pendekatan yang telah diuraikan pada makalah ini, pendekatan atau algoritma winnowing lebih baik dari pendekatan Manber karena memberikan jaminan terdeteksinya dokumen sama dan mempunyai nilai tambah yang lain yaitu terdapatnya informasi posisi fingerprint pada dokumen.

Penelitian berikutnya adalah menguji apakah algoritma winnowing ini dapat memberikan hasil yang optimum juga jika diterapkan untuk dokumen berbahasa Indonesia.

4. DAFTAR PUSTAKA

Ardini Ridhatillah, *Dealing with Plagiarism in the Information System Research Community: A Look at Factors that Drive Plagiarism and Ways to Address Them,* MIS Quarterly; Vol. 27, No. 4, p. 511-532/December 2003

B. Stein, S. Meyer zu Eissen, *Near Similarity Search and Plagiarism Analysis,* 29th Annual Conference of the German Classification Society (GfKl), Magdeburg, ISDN 1431-8814, pp. 430 – 437, 2006.

George R.S Weir, *Work in Progress – Technology in plagiarism detection and management*, 34th ASEE/IEEE Frontiers in Education Conference, 2004.

J. Evan, *Plagiarism Detection Software*, Department of Computer Science, Mathematic and Physics, Missouri Western State Collage.

Manber Ubi, *Finding Similar files in a large file system*, In Proceedings of

the USENIX Winter 1994 Technical Conference, 1994.

Maxim Mozgovoy, *Fast and Reliable Plagiarism Detection System*, 37th ASEE/IEEE Frontiers in Education Conference, 2007.

Najib Baedlowi, Deka Aditia Adam, *String Matching dengan Menggunakan Algoritma Rabin Karp,* Laboratorium Ilmu dan Rekayasa Komputasi Departemen Teknik Informatika, Institut Teknologi Bandung, 2005.

Norzima Elbegbayan, *Winnowing, a Document Fingerprinting Algorithm,* Department of Computer Science, Linkoping University, TDDC03, Spring 2005.

Parvati Iyer and Abhipsita Singh, *Document Similarity Analysis for a Plagiarism Detection System,* 2nd Indian International Conference on Artificial Intelegence (IICAI-05), pp. 2534 – 2544, 2005.

Romans Lukashenko, *Computer Based Plagiarism Detection Methods and Tools : An Overview,* International Conference on Computer System and Technologies, 2007.

S. Schleimer, D. Wilkerson, and A. Aiken. *Winnowing: Local Algorithms for Document Fingerprinting*. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 76-85, June 2003.