DETEKSI PLAGIAT DOKUMEN MENGGUNAKAN ALGORITMA RABIN-KARP

GILLAN FATAHILLAH

175100008

Universitas Mitra Indonesia, Sistem Informasi

ABSTRAK

Plagiat atau biasa disebut penjiplakan adalah sebuah masalah yang cukup signifikan pada banyak sekolah dan universitas. Hal plagiat yang biasanya dilakukan terhadap konten digital adalah melakukan copy-paste, quote, dan revisi terhadap dokumen asli. Untuk mengantisipasinya, dibutuhkan suatu cara yang dapat menganalisis teknik-teknik plagiat yang dilakukan. Ada beberapa pendekatan yang bisa diambil, salah satunya dengan mempergunakan algoritma pencarian string Rabin-Karp. Makalah ini hanya membahas secara skematis bagaimana algoritma Rabin-Karp bekerja dalam mendeteksi plagiat pada suatu dokumen, bukan implementasinya dalam sebuah program atau aplikasi. Kata kunci: Plagiat, Rabin-Karp, pencarian string.

A. PENDAHULUAN

Dokumen digital memang dapat dijiplak dengan sangat mudah. Praktik plagiat atau penjiplakan ini sudah sering terjadi khususnya pada kalangan akademisi baik lingkungan sekolah maupun perguruan tinggi. Tindakan plagiat yang dilakukan siswa atau mahasiswa ini sangat tidak mencerminkan sikap kreatif dan terpelajar sebagai kaum intelektual. Bentuk plagiat ini dapat bermacam macam seperti misalnya yang paling sering dilakukan yaitu copy-paste-edit suatu dokumen baik merupakan pekerjaan teman maupun dokumen yang berasal dari website, sehingga menjadi suatu hasil yang berbeda dari sebelumnya.

B. TINJAUAN PUSTAKA

Hashing Hashing adalah suatu cara untuk mentransformasi sebuah string menjadi suatu nilai yang unik dengan panjang tertentu (fixed-length) yang berfungsi sebagai JURNAL ILMU KOMPUTER DAN TEKNOLOGI INFORMASI, VOL III NO.2, OKTOBER 2003 penanda string tersebut. Fungsi untuk menghasilkan nilai ini disebut fungsi hash, sedangkan nilai yang dihasilkan disebut nilai hash. Contoh sederhana hashing adalah: Firdaus, Hari Munir, Rinaldi Rabin, Michael Karp, Richard menjadi 7864 Firdaus, Hari 9802 Munir, Rinaldi 1990 Rabin, Michael 8822 Karp, Richard Perbandingan kedua Hasil perbandingan juga tidak sama, maka dilakukan pergeseran. Begitu pula dengan perbandingan ketiga. Pada perbandingan keempat, didapatkan nilai hash yang sama. Gambar 4. Perbandingan keempat (nilai hash sama) Karena nilai hash sama, maka dilakukan perbandingan string karakter per karakter antara "bca" dan "cab". Didapatkan hasil bahwa kedua string tidak sama. Kembali substring bergeser ke kanan.

C. DISKUSI

JURNAL ILMU KOMPUTER DAN TEKNOLOGI INFORMASI, VOL III NO.2,

OKTOBER 2003 Pada perbandingan yang kelima algoritma RabinKarp dapat disesuaikan dengan tambahan penggunaan filter atau set data structure, untuk mengecek apakah hasil hashing string masukan ada pada kumpulan nilai hash dari pola yang dicari. Filter digunakan untuk mengeliminasi tanda baca (punctuation) dan beberapa kata dan kata sambung yang kurang signifikan untuk diberikan nilai hash, sedangkan set data structure adalah sekumpulan struktur data yang digunakan untuk membantu pencarian. Secara garis besar, pseudocode untuk algoritma RabinKarp untuk pencarian kumpulan string berpola banyak adalah: (diasumsikan semua string masukan pada himpunan s memiliki panjang yang sama m) Bila algoritma lain dapat mencari string berpola tunggal dalam waktu $O(n)$, jika digunakan untuk mencari pola sebanyak k , maka akan membutuhkan waktu selama $O(nk)$. Sedangkan varian Rabin-Karp diatas lebih mangkus karena diharapkan dapat mencari dengan kompleksitas waktu $O(n+k)$.

3. DETEKSI PLAGIAT DOKUMEN

Algoritma Rabin-Karp digunakan dalam mendeteksi plagiat sebab memungkinkan untuk mencari pola tulisan yang didapat dari substring-substring pada sebuah teks dalam dokumen, dimana algoritma pencarian fungsi hash memadai untuk setiap string pada k -gram yang dihasilkan. Berikut contoh hashing untuk string-string pada set s (agar memudahkan, nilai hash string yang diberikan acak dan berharga kecil). 12 22 13 15 35 26 18

Dengan cara pengulangan iteratif sampai mencapai akhir teks, nilai hash string-string pada set s dicocokkan dengan nilai hash penggalan string sepanjang 4 karakter pada teks (benar apabila nilai hash penggalan teks asli terdapat pada himpunan nilai hash string-string masukan). Kita sebut saja hs bagi himpunan nilai hash untuk set s . $hash(even) \in hs$ $hash(veni) \in hs$ $hash(enif) \in hs$ $hash(nifn) \in hs$ $hash(ifno) \in hs$ $hash(fnon) \in hs$... $hash(youl) \notin hs$ (bukan elemen hs) $hash(ould) \notin hs$ (bukan elemen hs) $hash(uldb) \notin hs$ (bukan elemen hs) $hash(ldbe) \notin hs$ (bukan elemen hs) $hash(dbse) \notin hs$ (bukan elemen hs) ... $hash(pute) \in hs$ $hash(uter) \in hs$ Pada setiap penggalan teks asli sepanjang 4 karakter, dilakukan hashing dengan fungsi hash yang sama seperti yang dilakukan terhadap string-string set s . Pencarian nilai hash penggalan teks asli pada hs dapat dilakukan dengan algoritma pencarian biasa, sesuai kehendak. Hasil akhir yang didapat untuk perbandingan dokumen pada Tabel 1 adalah: Tabel 2 Hasil perbandingan dokumen mahasiswa dan dokumen asli dari website Dokumen mahasiswa

Even if none of these things happen to you, youd be surprised at what you might find lurking on your computer! What are malware, adware, and spyware? Malware is a short for malicious software and is a catch all phrase to describe any software that is designed to damage a computer. Dokumen asli dari <http://www.jellico.com/spyware.html> Even if none of these things happen to you, you'd be surprised at what you might find lurking on your computer! What is it? Malware (short for "malicious software") is any software developed for the purpose of doing harm to a computer.

JURNAL ILMU KOMPUTER DAN TEKNOLOGI INFORMASI, VOL III NO.2, OKTOBER 2003 Dari

hasil perbandingan kedua potongan dokumen, terlihat sangat jelas bahwa mahasiswa mengopi sebagian besar dokumen asli. Pola tulisan mahasiswa menunjukkan kemiripan dengan pola penulisan pada dokumen asli, sehingga dapat disimpulkan ia telah melakukan tindakan plagiat. 4. PENINGKATAN AKURASI LANJUTAN Untuk peningkatan akurasi pencarian pola dokumen plagiat, dapat dikembangkan varian dari algoritma RabinKarp untuk menyesuaikan dengan jenis dokumen. Salah satu contohnya menyesuaikan filtering dengan menambah atau mengubah kata-kata apa saja yang dapat diabaikan dan yang tidak. Selain itu dapat pula melakukan teknik pengelompokan nilai hash. Teknik ini digunakan untuk mengelompokkan nilai hash string sesuai keterurutan pada teks asli sehingga perbandingan yang dilakukan adalah perbandingan per kelompok dengan ketentuan tertentu seperti, kelompok $0 \bmod 5$ (hasil modulo dengan 5 = 0) dan sebagainya, tentunya disesuaikan dengan fungsi hash.

D. KESIMPULAN

Kemiripan pola antar dua buah dokumen dapat dicari dengan menerapkan prinsip algoritma pencarian string Rabin-Karp. Algoritma Rabin-Karp menghasilkan efisiensi waktu yang baik dalam mendeteksi string yang memiliki lebih dari satu pola. Hal ini membuat algoritma Rabin-Karp dimanfaatkan dalam melakukan pendeteksian terhadap tindak plagiat dokumen.

E. REFERENSI

- [1] Noynaert, J.Evan, "Plagiarsim Detection Software", [http://www.cs.uwec.edu/mics/CD%20image/papers/paper9 7.pdf](http://www.cs.uwec.edu/mics/CD%20image/papers/paper9%207.pdf), waktu akses: 16 Mei 2008 pukul 15.00. [2] <http://www.sparknotes.com/cs/searching/hashtables/>, waktu akses: 16 Mei 2008 pukul 16.00. [3] Aiken, Alex, Saul Scheimer, dan Daniel S. Wilkerson, "Winnowing: Local Algorithms for Document Fingerprinting", <http://www.theory.stanford.edu/~aiken/publications/papers/sigmod03.pdf>, waktu akses: 16 Mei 2008 pukul 15.00. [4] Munir, Rinaldi, "Diktat Kuliah IF2251 Strategi Algoritmik", Departemen Teknik Informatika ITB, 2007. [5] <http://en.wikipedia.org/wiki/>, waktu akses: 16 Mei 2008