

KOMPARASI ALGORITMA WINNOWING DAN RABIN KARP MENDETEKSI KEMIRIPAN TUGAS MAHASISWA

Ida Bagus Ketut Surya Arnawa

Program Studi Sistem Informasi

Institut Teknologi Dan Bisnis STIKOM Bali, Denpasar, Bali

arnawa@stikom-bali.ac.id

ABSTRACT

Learning is a process of improving one's self-quality both in terms of knowledge, attitudes and behavior. In the learning process in higher education, a lecturer often gives assignments to students. The purpose of giving assignments to students is to train them to want to learn, to be responsible and as material for evaluating the ability of students to understand a material that has been given. In making assignments, students often commit acts of plagiarism either intentionally or unintentionally. Many factors encourage students to take plagiarism, one of which is arrogance in making assignments whose goal is to finish assignments quickly and the lack of a sense of responsibility for students in doing assignments. One way to prevent and minimize plagiarism is to use document matching techniques with one another. with the aim of knowing the level of similarity of a text document. There are many algorithms in Text Mining that can be used such as the Jaro-Winkler algorithm, Rabin-Karp, Winnowing and so on. Referring to the description above, the author is interested in conducting a comparison analysis of the performance of the Jaro-Winkler, Rabin Karp and Winnowing algorithms in measuring the level of similarity of a document. text. In the trial process, it was found that the Rabin Karp algorithm has better performance than the Winnowing algorithm in detecting the similarity of student assignments.

Keywords: comparison, winnowing, Rabin Karp.

ABSTRAK

Belajar merupakan sebuah proses peningkatan kualitas diri seseorang baik dari segi pengetahuan, sikap maupun tingkah laku. Dalam proses belajar di perguruan tinggi sering kali seorang dosen memberikan tugas kepada mahasiswa. Adapun tujuan dari pemberian tugas kepada peserta didik yaitu melatih agar mau belajar, bisa bertanggung jawab serta sebagai bahan untuk evaluasi terhadap kemampuan peserta didik dalam memahami suatu materi yang sudah diberikan. Dalam pembuatan tugas mahasiswa sering melakukan tindakan plagiarisme baik dilakukan secara sengaja maupun tidak sengaja. Banyak faktor yang mendorong mahasiswa melakukan tindakan plagiarisme salah satunya sikap arogansi dalam pembuatan tugas yang tujuannya supaya tugas cepat selesai serta kurangnya rasa tanggung jawab mahasiswa dalam mengerjakan tugas. Salah satu cara untuk mencegah serta meminimalisir tindakan plagiarisme yaitu dengan menggunakan teknik pencocokan dokumen satu dengan yang lainnya dengan tujuan untuk mengetahui tingkat kemiripan suatu dokumen text. Ada banyak algoritma dalam Text Mining yang dapat digunakan seperti algoritma Jaro-Winkler, Rabin-Karp, Winnowing dan lain sebagainya. Merujuk dari uraian diatas penulis tertarik untuk melakukan analisis perbandingan kinerja algoritma Jaro-Winkler, Rabin Karp serta Winnowing dalam mengukur tingkat kemiripan suatu dokumen text. Dalam proses uji coba ditemukan bahwa algoritma Rabin Karp memiliki unjuk kerja yang lebih baik dibandingkan algoritma Winnowing dalam kasus mendeteksi kemiripan tugas mahasiswa.

Kata Kunci : komparasi, winnowing, Rabin-Karp.

PENDAHULUAN

Belajar merupakan sebuah proses peningkatan kualitas diri seseorang baik dari segi pengetahuan, sikap maupun tingkah laku. Banyak cara yang dapat ditempuh oleh seseorang untuk dapat belajar salah satunya dengan masuk perguruan tinggi. Dalam proses belajar di perguruan tinggi sering kali seorang dosen memberikan tugas kepada mahasiswa. Tugas merupakan suatu kewajiban yang harus dikerjakan oleh mahasiswa. Ada beberapa macam tugas diantaranya tugas perseorangan dan tugas kelompok. Adapun tujuan dari pemberian tugas kepada peserta didik yaitu melatih agar mau belajar, bisa bertanggung jawab serta sebagai bahan untuk evaluasi terhadap kemampuan peserta didik dalam memahami suatu materi yang sudah diberikan.

Dalam pembuatan tugas mahasiswa sering melakukan tindakan plagiarisme baik dilakukan secara sengaja maupun tidak sengaja. plagiarisme merupakan suatu tindakan dimana ide atau gagasan orang lain diambil begitu saja tanpa menyebutkan sumbernya atau tanpa seijin dari yang memiliki ide atau gagasan tersebut. Dalam dunia pendidikan plagiarisme merupakan suatu tindakan yang sangat tidak terpuji. UU No. 9 tahun 2012 pasal 2 ayat (1), pasal 3 ayat (1),(2), pasal 12, pasal, pasal 15 dan pasal 26 ayat (1) menyebutkan tentang pelanggaran terhadap tindakan plagiarisme. Banyak faktor yang mendorong mahasiswa melakukan tindakan plagiarisme salah satunya sikap arogansi dalam pembuatan tugas yang tujuannya supaya tugas cepat selesai serta kurangnya rasa tanggung jawab mahasiswa dalam mengerjakan tugas. Hal ini perlu dicegah supaya tidak menjadi kebiasaan bagi mahasiswa. Banyak kendala yang dialami oleh seorang dosen dalam mengetahui tugas mahasiswa tersebut termasuk melakukan tindakan plagiarisme atau tidak karena banyak tugas yang perlu dikoreksi serta banyak memerlukan

waktu dalam mencocokkan kemiripan suatu tugas.

Salah satu cara untuk mencegah serta meminimalisir tindakan plagiarisme yaitu dengan menggunakan teknik pencocokan dokumen satu dengan yang lainnya dengan tujuan untuk mengetahui tingkat kemiripan suatu dokumen text. Text Mining merupakan salah satu solusi yang bisa digunakan untuk mengetahui tingkat kemiripan suatu dokumen text. Ada banyak algoritma dalam Text Mining yang dapat digunakan seperti algoritma Jaro-Winkler, Rabin-Karp, Winnowing dan lain sebagainya. Semua algoritma pencocokan string memiliki kekurangan dan kelebihan masing-masing. Merujuk dari uraian diatas penulis tertarik untuk melakukan analisis perbandingan kinerja algoritma Winnowing dan Rabin Karp dalam mengukur tingkat kemiripan suatu dokumen text.

TINJAUAN PUSTAKA

Algoritma Winnowing

Algoritma winnowing merupakan salah satu algoritma yang dapat digunakan untuk mengukur tingkat kesamaan kata (document fingerprinting) untuk mendeteksi plagiarism. Rolling hash merupakan algoritma yang digunakan mencari nilai hash dalam winnowing. Nilai hash merupakan nilai numerik yang terbentuk dari perhitungan ASCII setiap karakter. Langkah-langkah dalam penerapan Algoritma Winnowing adalah sebagai berikut [1][5]:

1. Langkah Pertama

Penghapusan karakter yang tidak relevan yaitu berupa spasi, tanda baca serta spesial karakter seperti !, @, #, \$, %, ^, &, *, (,), _ , - ?

Contoh :

Belajar Pemrograman C#

Akan dirubah menjadi

belajarpemrogramanc

2. Langkah Kedua

Pembentukan Rangkaian n-gram yaitu dengan cara membentuk rangkaian karakter sepanjang n dari hasil pembuangan karakter yang tidak relevan pada langkang pertama. Dari text diatas telah dibersihkan dengan ukuran k=5

belaj elaja lajar ajarp jarpe arpem
rpemr pemro emrog mrogr rogra
ogram grama raman amanc

3. Langkah Ketiga

Menghitung fungsi hash untuk setiap n-gram yaitu dengan cara menghitung nilai - nilai hash dari setiap gram, algoritma holling hash digunakan untuk menghitung nilai hash dalam algoritma winnowing. Rolling hash merupakan suatu teknik untuk mentransformasikan sebuah string menjadi nilai yang unik dengan panjang tertentu yang berfungsi sebagai penanda string tersebut. Fungsi tersebut menghasilkan nilai ini disebut fungsi hash, sedangkan nilai yang dihasilkan disebut nilai hash. Fungsi hash $H(c1...ck)$ didefinisikan sebagai berikut :

$$H(ck) = c1*b(k-1) + c2*b(k-2) + ... + ck*b(k-k)$$

Keterangan :

c = nilai ascii karakter

b = basis (bilangan prima)

k = banyak karakter

hasil rolling hash dari kalimat diatas yaitu

1583490 1635489 1724342 1574366
1696180 1586683 1831680 1788777
1638938 1762281 1830629 1777214
1672790 1812547 1578302

4. Langkah Keempat

Pembentukan window dari nilai hash dari window dengan ukuran 3 yaitu sebagai berikut :

{ 1583490	1635489	1724342}
{ 1635489	1724342	1574366}
{ 1724342	1574366	1696180}
{ 1574366	1696180	1586683}
{ 1696180	1586683	1831680}
{ 1586683	1831680	1788777}
{ 1831680	1788777	1638938}
{ 1788777	1638938	1762281}
{ 1638938	1762281	1830629}
{ 1762281	1830629	1777214}
{ 1830629	1777214	1672790}
{ 1777214	1672790	1812547}
{ 1672790	1812547	1578302 }

5. Langkah Kelima

Langkah terakhir yaitu memilih nilai terkecil dari setiap window untuk dijadikan fingerprint, hasil dari nilai fingerprintnya sebagai berikut:

[1583490,0]	[1574366,3]	[1586683,5]
[1638938,8]	[1762281,9]	[1672790,12]
[1578302,14]		

Nilai fingerprint yang dibentuk dari algoritma winnowing digunakan untuk mengukur prosentase kemiripan teks pada persamaan Jaccard Coefficient. Persamaan Jaccard Coefficient digunakan untuk menghitung kemiripan (similarity) dari kumpulan kata-kata yang telah dihitung nilai hash nya. Berikut ini rumus persamaan Jaccard Coefficient.

$$\text{Similarity} = \frac{\text{Jumlah fingerprint sama}}{\text{Total seluruh fingerprint}} \times 100$$

Algoritma Rabin-Karp

Algoritma Rabin-Karp diciptakan oleh Michael O. Rabin dan Richard M. Karp pada tahun 1987 yang menggunakan fungsi hashing untuk menemukan pattern di dalam

string teks. Algoritma Rabin-Karp digunakan sebagai algoritma deteksi kesamaan pada dokumen tugas akhir. Algoritma ini memiliki beberapa keunggulan dalam penerapannya, salah satunya adalah algoritma ini sangat cocok digunakan untuk string yang panjang. Karakteristik algoritma Rabin-Karp [2][4], yaitu :

- Menggunakan sebuah fungsi hashing.
- Tahap preprocessing menggunakan kompleksitas waktu $O(m)$
- Untuk tahap pencarian kompleksitasnya : $O(mn)$
- Waktu yang diperlukan $O(n+m)$

Dalam algoritma Rabin-Karp, ada beberapa tahap yang harus dilelalui dalam implementasi algoritma tersebut.

1. Preprocessing

Tahap ini melakukan analisis semantik (kebenaran arti) dan sintaktik (kebenaran susunan) teks. Tujuan dari pemrosesan awal adalah untuk mempersiapkan teks menjadi data yang mengalami pengolahan lebih lanjut. Pembersihan yang dilakukan adalah pembersihan suatu teks dari karakter-karakter yang bersifat sebagai stoplist, dimana karakter tersebut banyak dan pasti muncul dalam setiap teks.

2. Rolling Hash

Fungsi yang digunakan untuk menghasilkan nilai hash dari rangkaian gram dalam Algoritma Rabin-Karp adalah dengan menggunakan fungsi rolling hash. Fungsi hash ($C1...Ck$) didefinisikan sebagai berikut,

$$c1*(k-1)+c2*b(k-2)+...+c(k-1)*b+k+c_k... (1)$$

Keterangan :

c : nilai ASCII karakter

b : basis (bilangan prima)

k : banyak karakter

Metode hashing digunakan untuk mempercepat pencarian atau pencocokan suatu string.

Apabila tidak di-hash, pencarian akan dilakukan karakter per karakter. Namun pencarian akan menjadi lebih mangkus setelah di-hash karena hanya akan membandingkan empat digit angka untuk mengetahui kesamaan suatu substring.

Berikut ini merupakan contoh penghitungan rolling hash dengan k-grams = 6 dan basis = 101:

Diketahui

Teks : abracadabra

ASCII : a = 97, b = 98, r = 114

Maka

$\text{hash}(\text{"abr"}) = (97 \times 101^2) + (98 \times 101^1) + (114 \times 101^0) = 999.509$ Untuk menghitung nilai hash dari substring selanjutnya, dapat dilakukan dengan mengurangi hash dengan nilai dari karakter yang hilang kemudian mengalikannya dengan basis dan menambahkannya dengan nilai karakter yang ditambahkan. $\text{hash}(\text{"bra"}) = [101 \times (999.509 - (97 \times 101^2))] + (97 \times 101^0) = 1.011.309$.

3. Pencocokan

Proses pencocokan dalam Algoritma Rabin-Karp dilakukan dengan menggunakan sebuah teorema yaitu:

Sebuah stringA identik dengan stringB, jika (syarat perlu) stringA memiliki hash key yang sama dengan hash key yang dimiliki oleh stringB

4. Pengukuran Nilai Kemiripan

Mengukur similarity (kemiripan) dan jarak antara dua entitas informasi adalah syarat inti pada semua kasus penemuan informasi, seperti pada Information Retrieval dan Data Mining yang kemudian dikembangkan dalam bentuk perangkat lunak, salah satunya adalah sistem deteksi kesamaan [3]. Penggunaan ukuran similarity yang tepat tidak hanya meningkatkan kualitas pilihan informasi tetapi juga membantu mengurangi waktu dan biaya proses sehingga

memperangkat lunak Dice's Similarity Coefficient dalam penghitungan nilai similarity yang menggunakan pendekatan k-gram.

$$S = K \times C / (A+B) \dots \dots \dots (2.2)$$

Dimana S adalah nilai similarity, A dan B adalah jumlah dari kumpulan k-grams dalam teks 1 dan teks 2. C adalah jumlah dari k-grams yang sama dari teks yang dibandingkan. Berikut adalah contoh dari penggunaan rumus tersebut. Terdapat dua buah dokumen teks (dok A dan dok B) dengan nilai k-gram masing-masing dokumen secara berturut-turut adalah 2608 dan 3040, sedangkan nilai k-gram yang sama adalah sebesar 1203. Maka hasil nilai dari kemiripan kedua dokumen tersebut adalah :

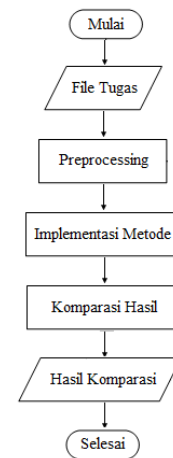
$$S = K \times C / (A+B) = 2 \times 1203 / (2608 + 3040)$$

$$= 2406 / 5648 = 0,4259 \times 100\% = 42,59\%$$

METODE PENELITIAN

Arsitektur Sistem

Perancangan sistem yang dibuat merupakan sistem untuk mendeteksi kemiripan tugas mahasiswa satu dengan yang lainnya. Sistem ini dapat menerima inputan berupa text yang merupakan tugas dari mahasiswa. Tahap awal sistem melakukan preprocessing terlebih dahulu terhadap tugas mahasiswa. Setelah melalui tahapan preprocessing tugas mahasiswa akan dicocokkan dengan tugas lainnya dengan menggunakan masing-masing metode yaitu Jaro-Winkler dan Winnowing. Masing-masing metode akan menghasilkan persentase kemiripan tugas mahasiswa yang satu dengan yang lainnya. Kemudian dari hasil tersebut dianalisis untuk mengetahui metode mana yang memiliki unjuk kerja yang paling baik. Gambar 1 merupakan gambaran arsitektur sistem komparasi algoritma text mining dalam mendeteksi kemiripan tugas mahasiswa.



Gambar 1. Arsitektur Sistem

HASIL DAN PEMBAHASAN

Implementasi Perbandingan Algoritma

Tahapan - tahapan untuk meng - implementasikan perbandingan algoritma Winnowing dengan algoritma rabin karp untuk mendeteksi presentase plagiarisme pada tugas mahasiswa adalah sebagai berikut :

1. Ambil 2 buah teks hasil tugas mahasiswa yang akan dibandingkan.

Tugas Mahasiswa 1

Harddisk adalah salah satu perangkat keras komputer yang menyediakan ruang penyimpanan yang digunakan untuk menyimpan atau membaca data pada komputer. Harddisk merupakan jenis memori komputer yang memiliki sifat non-volatile memory yang berarti data pada memori jenis ini tidak akan hilang meski perangkat tersebut telah dimatikan.

Tugas Mahasiswa 2

Haddisk merupakan perangkat keras komputer yang digunakan untuk penyimpanan. Harddisk merupakan salah satu jenis memory yang non-volatile yaitu tidak akan hilang meskipun perangkat dimatikan

Tabel 1 Algoritma Wnnowing

uji	n-gram (n)	window (w)	waktu proses (s)	kemiripan (%)
1	2	3	0.0172	64,4
	2	5	0.0170	56,0
	2	7	0.0188	43,7
	2	9	0.0238	54,5
2	3	3	0.0174	47
	3	5	0.0168	43,4
	3	7	0.0176	38,1
	3	9	0.0185	35
3	4	3	0.0200	39,5
	4	5	0.0173	38,8
	4	7	0.0194	35,2
	4	9	0.0184	34
4	5	3	0.0173	34,7
	5	5	0.0184	33,6
	5	7	0.0174	30,2
	5	9	0.0167	27,5
5	6	3	0.0176	30,4
	6	5	0.0202	29,4
	6	7	0.0190	25,9
	6	9	0.0179	24,1
6	7	3	0.0216	25,6
	7	5	0.0184	24
	7	7	0.0186	22,3
	7	9	0.0183	20
7	8	3	0.0180	20,7
	8	5	0.0173	21,6
	8	7	0.0178	20,6
	8	9	0.0187	18,1
8	9	3	0.0174	17,5
	9	5	0.0180	17,7
	9	7	0.0183	17,7
	9	9	0.0188	16,1
9	10	3	0.0222	13,4

	10	5	0.019	13,8
	10	7	0.0180	12,6
	10	9	0.0188	12,6

Tabel 2 Hasil window dan kemiripan terbaik algoritma winnowing

uji	n-gram (n)	window (w)	waktu proses (s)	kemiripan (%)
1	2	3	0.0172	64,4
2	3	3	0.0174	47
3	4	3	0.0200	39,5
4	5	3	0.0173	34,7
5	6	3	0.0176	30,4
6	7	5	0.0216	25,6
7	8	5	0.0173	21,6
8	9	5	0.0180	17,7
9	10	5	0.019	13,8

Tabel 3 Algoritma Rabin Karp

uji	n-gram (n)	waktu proses (s)	kemiripan
1	2	0.0177	86,87
2	3	0.0181	72,24
3	4	0.0188	64,63
4	5	0.0191	58,29
5	6	0.0189	53,39
6	7	0.0191	48,68
7	8	0.0192	43,44
8	9	0.0196	38,56
9	10	0.0210	33,96

Dari hasil perbandingan pendekatan algoritma winnowing dengan algoritma rabin karp dapat dilihat kemungkinan kemiripan tugas mahasiswa yang terbesar adalah dengan menggunakan pendekatan algoritma rabin karp yaitu pada uji coba ke 1 dengan nilai n-gram = 2, proses waktu 0.0177 dengan tingkat kemiripan terbesar yaitu 86,87 %.

SIMPULAN

Pendekatan Algoritma Rabin Karp lebih baik daripada pendekatan algoritma Winnowing karena menghasilkan tingkat presentase yang lebih besar. Terdapat perbedaan mendasar dari kedua algoritma yaitu pada algoritma winnowing memiliki window sedangkan algoritma rabin karp hanya memiliki k-gram Berdasarkan hasil Pengujian terhadap

perbandingan pendekatan algoritma winnowing dengan algoritma rabin karp dapat dilihat kemungkinan kemiripan teks Tugas Mahasiswa yang terbesar adalah dengan menggunakan pendekatan algoritma rabin karp yaitu pada ujicoba ke 1 dengan nilai $n\text{-gram} = 2$ proses waktu 0.0177 dengan tingkat kemiripan terkecil yaitu 86,87 %

DAFTAR PUSTAKA

- [1] Astutik, Sariyanti, et al. "Sistem Penilaian Esai Otomatis Pada E-learning Dengan Algoritma Winnowing." *Jurnal Informatika University Petra Kristian*, vol. 12, no. 2, Nov. 2014, pp. 47-52.
- [2] Putra, N. P., & Sularno, S. (2019). "Penerapan Algoritma Rabin-karp dengan Pendekatan Synonym Recognition sebagai Antisipasi Plagiarisme pada Penulisan Skripsi". *Jurnal Teknologi Dan Sistem Informasi Bisnis*, 1(2), 130-140.
- [3] Rahardjo, Andi Bekto. "Penerapan Data Mining Untuk Mengklasifikasi Penerima dan Bukan Penerima Kartu Identitas Miskin (KIM) Kelurahan Sumurrejo Gunungpati dengan Metode Naive Bayes Classifier." (2015): 1-8.
- [4] Putra, D. A., & Sujaini, H. (2015). Implementasi Algoritma Rabin-Karp untuk Membantu Pendeteksian Plagiat pada Karya Ilmiah. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 4(1), 66-74.
- [5] Setiawan, A. (2017). IMPLEMENTASI ALGORITMA WINNOWER UNTUK DETEKSI KEMIRIPAN JUDUL SKRIPSI STUDI KASUS STMIK BUDIDARMA. *Majalah Ilmiah INTI (Informasi dan Teknologi Ilmiah)*, 12(1).