ANALYSIS WINNOWING ALGORITHM FOR TEXT PLAGIARISM DETECTION USING THREE METHOD SIMILARITY

Luke Michael Febriansyah[1], Shinta Estri Wahyuningrum[2]

[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer, Universitas Katholik Soegijapranata

[1]14k10040@student.unika.ac.id, [2]shinta@unika.ac.id

Abstract

Cases of plagiarism in recent years has been an issues. Based on that issues, this research will create a system to detect similarity in a text. There is an aspect as reference of the research that is analyze the plagiarism algorithm. This research will analyze the accuracy one of plagiarism check algorithm, winnowing algorithm. Winnowing algorithm is a plagiarism detection algorithm based on document fingerprinting. To calculate percentage similarity of document fingerprinting in text, there are 3 methods to measure similarity that will be used in this research, which is jaccard similarity coefficient, sorensen dice similarity coefficient, and berg similarity coefficient.

Keywords: *plagiarism text, winnowing algorithm, document fingerprint, jaccard similarity,sorensen dice similarity, andberg similarity*

Pendahuluan

Kasus plagiasi yang terjadi pada beberapa tahun belakangan ini cukup menjadi pembahasan. Tingkat tindakan plagiasi meningkat dari tahun ke tahun (news.okezone.com-2014). Dokumen dan teks telah ditiru tanpa adanya sumber yang tertera. Salah satu alasan yang mendasari suatu tindakan plagiasi pada sekarang ini adalah karena mudah nya akses dan akses bebas terhadap pekerjaan orang lain di internet. Karena permasalahan plagiasi tersebut maka beberapa metode dan algoritma telah dibuat untuk memecahkan kasus plagiasi seperti algoritma rabin-karp, algoritma manber dan algoritma winnowing. Beberapa algoritma tersebut memberikan persentase nilai kemiripan dari dokumen dan teks dengan cara yang berbeda. Permasalahannya yaitu akurasi dari beberapa algoritma tersebut untuk melakukan pendeteksian plagiasi pada suatu teks.

Pada penelitian ini dilakukan analisa akurasi dari salah satu algoritma plagiasi yaitu algoritma winnowing dengan mengimplementasikan 3 metode kemiripan yang berbeda. Algoritma winnowing adalah algoritma untuk mendeteksi plagiasi dengan berdasarkan dokumen fingerprint. Langkah dasar dari algoritma winnowing yaitu memasukkan nilai k-gram dan nilai window untuk menemukan nilai dokumen fingerprint dan menghitung kemiripan dokumen fingerprint dari teks dengan jaccard similarity coefficient, sorensen dice similarity coefficient, dan andberg similarity

coefficient. Analisa dilakukan berdasarkan uji coba dengan nilai k-gram yang berbeda, nilai window yang berbeda dan membandingkan persentase hasil dari tiap metode kemiripan.

Landasan Teori

Berdasarkan dari Jurnal Teknik Informatika Universitas Gunadarma dengan judul "Perbandingan Pendekatan Deteksi Plagiarism Dokumen Dalam Bahasa Inggris", lebih membahas terhadap arti dan makna dari plagiasi dan beberapa pendekatan yang bisa digunakan untuk mendeteksi plagiasi pada teks berbahasa inggris. Ada dua pendekatan yang digunakan untuk mengatasi masalah tersebut yaitu algoritma manber dan algoritma winnowing. Algoritma manber memilih fingerprint berdasar nilai hash yang memenuhi kriteria dari 0 mod P dan algoritma winnowing memilih fingerprint berdasar nilai terkecil dari setiap window yang terbentuk. Dan kesimpulan yang didapat adalah algoritma winnowing lebih cocok digunakan daripada manber untuk mendeteksi kemiripan pada teks berbahasa inggris, karena algoritma winnowing memberikan informasi posisi dari fingerprint.

Berdasarkan dari Jurnal Teknik Informatika Universitas Trunojoyo Madura dengan judul "Sistem Penilaian Esai Otomatis Pada E-Learning Dengan Algoritma Winnowing", penilaian otomatis untuk esai dapat dilakukan dengan mengimplementasikan algoritma winnowing. Dibahas bahwa jawaban dari setiap siswa diolah menggunakan algoritma winnowing dan dibandingkan dengan kunci jawaban dari guru. Kesimpulannya adalah semakin besar nilai k-gram akan memberikan nilai kemiripan jawaban yang akan lebih besar dan semakin kecil nilai k-gram akan memberikan nilai kemiripan jawaban yang lebih kecil.

Berdasarkan dari Jurnal Teknik Informatika Universitas Malikussaleh dengan judul "Sistem Pendeteksian Kemiripan Judul Skripsi Menggunakan Algoritma Winnowing", Algortima winnowing dapat membantu untuk mendeteksi kemiripan dari judul skripsi. Adapun skema sistem dari projek ini yaitu melakukan input judul skripsi, pengolahan awal teks judul skripsi, dan kemudian pengolahan dengan algoritma winnowing dan menghitung nilai kemiripan judul skripsi dengan dice similarity coefficient. Dari semua proses tersebut, hasil kemiripan dari algoritma winnowing dibandingkan dengan semua data judul skripsi yang sudah ada di database sistem. Dan kesimpulan yang didapat adalah keputusan persentase tingkat plagiasi dari judul skripsi menjadi lebih cepat dan akurat.

Metodologi Penelitian

Beberapa tahapan metode yang harus dilewati dan digunakan dalam penelitian

1. Menemukan teks sampel dan teks testing

Teks sampel adalah teks yang digunakan sebagai acuan testing untuk mengetahui nilai optimal dari k-gram dan window dari 1 sampai 10. Teks sample yang digunakan diambil dari wikipedia dengan topik sejarah komputer yang ditulis ulang dalam bentuk format txt. Sedangkan teks testing adalah teks yang digunakan sebagai acuan testing untuk mengetahui tingkat akurasi dari 3 metode kemiripan yang dipakai. Teks testing yang digunakan diambil dari salah satu kasus plagiasi yang terjadi di Indonesia yang ditulis di salah satu web berita dan ditulis ulang dalam bentuk format txt.

2. Memberikan nilai k-gram dan nilai window

Nilai k-gram adalah bobot karakter yang digunakan sebagai referensi untuk memecah kata. Pemecahan kata dengan k-gram adalah langkah awal dari algoritma winnowing. Sedangkan nilai window adalah nilai yang digunakan sebagai referensi untuk memecah nilai hash dari proses rolling hash ke dalam bentuk window. Jarak k-gram dan window yang digunakan dalam penelitian yaitu antara 1 sampai dengan 10.

3. Implementasi dengan algoritma winnowing

   a. Proses awal *whitespace insensitivity* dengan cara menghilangkan karakter yang tidak relevan seperti simbol, spasi dan juga mengubah huruf besar menjadi huruf kecil.

   b. Proses kedua adalah memecah kata berdasarkan input nilai k-gram.

   c. Proses ketiga adalah proses rolling hash, mengubah setiap kata yang dihasilkan dari proses kedua menjadi nilai hash.

   d. Proses keempat adalah memecah nilai hash yang dihasilkan ke dalam bentuk index window berdasarkan input nilai window.

   e. Proses kelima adalah memilih nilai hash terkecil dari setiap index window untuk dijadikan dokumen fingerprint.

   f. Proses keenam adalah menghitung persentase kemiripan dengan jaccard similarity coefficients, sorensen dice similarity coefficient, dan andberg similarity coefficients.

   g. Berikut rumus dari 3 metode kemiripan yang digunakan :

      - Jaccard Similarity Coefficients

        Rumus : D(A,B) = $A \cap B$ / $A \cup B$ * 100

      - Sorensen Dice Similarity Coefficients

        Rumus : D(A,B) = (2 * $A \cap B$) / (A + B) * 100

      - Andberg Similarity Coefficients

        Rumus : D(A,B) = $A \cap B$ / ($A \cup B$ + A $\triangle$ B) * 100

4. Proses uji coba dan analisa

Uji coba dengan memberikan nilai k-gram dan nilai window dari 1 sampai dengan 10 kemudian analisis terhadap hasil persentase kemiripan yang didapat dari 3 metode kemiripan. Setelah dilakukan analisa dari tiap metode kemiripan kemudian dibandingkan satu sama lain.

Hasil dan Pembahasan

Pada tahap uji coba, dibagi menjadi 3 bagian uji coba. Uji coba pertama adalah analisa terhadap nilai k-gram dan window dalam jarak 1 sampai 10 untuk mengetahui nilai optimal dari k-gram dan window dari jarak tersebut. Setelah mendapatkan nilai optimal dari k-gram dan window yang bisa digunakan. Kemudian dilanjutkan dengan uji coba kedua yaitu analisa terhadap basis bilangan prima pada rolling hash untuk mengetahui nilai optimal dari basis bilangan prima yang bisa digunakan. Dan uji coba ketiga adalah membandingkan 3 metode kemiripan yaitu Jaccard Similarity Coefficient, Sorensen Dice Similarity Coefficient, Andberg Similarity Coefficient.

Teks sampel yang digunakan pada uji coba pertama adalah sejarah komputer yang diambil dari wikipedia dan ditulis ulang menjadi file txt dan dibagi menjadi 4 teks sample.

1.1 Uji Coba Pertama

Uji coba pertama dilakukan dengan menginput nilai k-gram dan window dari 1 sampai dengan 10 ke dalam sistem. Uji coba dilakukan secara berulang dengan jarak k-gram dan window tersebut. Jadi hasil dari uji coba pertama yaitu 100 kali percobaan pada tiap metode kemiripan. Uji coba pertama dicatat ke dalam sebuah tabel, berikut salah satu pencatatan yang dilakukan pada uji coba pertama :

Jaccard Similarity Table Test

Tabel 1: JaccardTableTest1

kgram	**window**	**text 1 = text 2**	**text 1 = text 3**	**text 1 = text 4**
1	1	95,46%	95,45%	86,36%
1	2	95%	90%	85%
1	3	94,74%	84,21%	73,68%
1	4	94,12%	82,35%	70,59%
1	5	86,67%	86,67%	60%
1	6	91,67%	83,33%	58,33%
1	7	81,82%	81,82%	63,64%

1	8	88,89%	88,89%	77,78%
1	9	87,5%	100%	75%
1	10	85,71%	100%	71,43%
Mean		**90,16%**	**89,27%**	**72,18%**

Sorensen Dice Similarity Table Test

Tabel 2: SorensenDiceTableTest1

kgram	**window**	**text 1 = text 2**	**text 1 = text 3**	**text 1 = text 4**
1	1	97,67%	97,67%	92,68%
1	2	97,43%	94,74%	91,89%
1	3	97,30%	91,43%	84,85%
1	4	96,97%	90,32%	82,76%
1	5	92,86%	92,86%	75%
1	6	95,65%	90,91%	73,68%
1	7	90%	90%	77,78%
1	8	94,12%	94,12%	87,5%
1	9	93,33%	100%	85,71%
1	10	92,31%	100%	83,33%
Mean		**94,76%**	**94,20%**	**83,52%**

Andberg Similarity Table Test

Tabel 3: AndbergTableTest1

kgram	**window**	**text 1 = text 2**	**text 1 = text 3**	**text 1 = text 4**
1	1	91,30%	91,30%	76%
1	2	90,48%	81,82%	73,91%
1	3	90%	72,73%	58,33%
1	4	88.89%	70%	54,54%
1	5	76,47%	76,47%	42,86%
1	6	84,61%	71,43%	41,18%
1	7	69,23%	69,23%	46,67%

1	8	80%	80%	63,64%
1	9	77,78%	100%	60%
1	10	75%	100%	55,56%
Mean		**82,38%**	**81,30%**	**57,27%**

1.2 Uji coba kedua

Uji coba kedua adalah analisa terhadap basis bilangan prima pada rolling hash. Teks sampel yang digunakan masih sama dengan teks sampel pada uji coba pertama. Nilai k-gram dan window yang digunakan pada uji coba kedua yaitu nilai k-gram 3 dan nilai window 1.

Jaccard Similarity Table Test

Kgram : 3 , Window : 1

Tabel 4: JaccardTableTest2

Basic Prime Number	Text 1 = Text 2	Text 1 = Text 3	Text 1 = Text 4
3	94,79%	72,51%	55,92%
5	91,46%	63,72%	44,51%
7	86,77%	61,64%	40,74%
11	86,47%	57,34%	38,07%
19	86,55%	56,50%	37,44%
29	86,56%	55,95%	37%
47	86,56%	55,95%	37%
109	86,56%	55,95%	37%
199	86,56%	55,95%	37%

Sorensen Dice Similarity Table Test

Kgram : 3, Window : 1

Tabel 5: SorensenDiceTableTest2

Basic Prime Number	Text 1 = Text 2	Text 1 = Text 3	Text 1 = Text 4
3	97,32%	84,06%	71,73%
5	95,54%	77,84%	61,60%

7	92,92%	76,27%	57,89%
11	92,74%	72,89%	55,15%
19	92,79%	72,21%	54,49%
29	92,80%	71,75%	54,02%
47	92,80%	71,75%	54,02%
109	92,80%	71,75%	54,02%
199	92,80%	71,75%	54,02%

Andberg Similarity Table Test

Kgram : 3 , Window : 1

Tabel 6: AndbergTableTest2

Basic Prime Number	Text 1 = Text 2	Text 1 = Text 3	Text 1 = Text 4
3	90,09%	56,88%	38,81%
5	84,27%	46,76%	28,63%
7	76,63%	44,55%	25,58%
11	76,16%	40,19%	23,51%
19	76,28%	39,37%	23,03%
29	76,31%	38,84%	22,70%
47	76,31%	38,84%	22,70%
109	76,31%	38,84%	22,70%
199	76,31%	38,84%	22,70%

Berdasarkan dari hasil uji coba pertama dan uji coba kedua, dapat disimpulkan bahwa penggunaan nilai k-gram 2 dan 3 pada jarak nilai window 1 sampai dengan 10 dan nilai basis bilangan prima 3 adalah nilai optimal yang bisa digunakan untuk uji coba ketiga. Karena pada uji coba pertama dan uji coba kedua dengan 3 metode kemiripan yang berbeda nilai tersebut menghasilkan nilai rata-rata tertinggi.

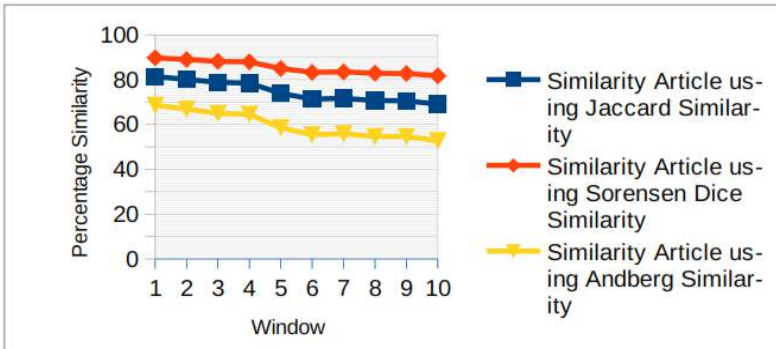
1.3 Uji coba ketiga

Uji coba ketiga adalah uji coba algoritma winnowing untuk mendeteksi plagiasi pada sebuah teks dan juga membandingkan 3 metode kemiripan yang digunakan. Teks testing yang digunakan yaitu salah satu kasus plagiasi di Indonesia seperti yang tertulis

di dalam berita (kabar24.bisnis.com/diduga-plagiat-ini-perbandingan-artikel-anggito-abimanyu-hotbonar-sinaga,2014).

Uji coba ketiga dilakukan dengan memasukkan nilai optimal k-gram 2 dan 3. Uji coba dilakukan secara berulang dengan nilai k-gram tersebut dan nilai window dengan jarak 1 sampai dengan 10. Jadi hasil dari uji coba ketiga adalah 20 kali percobaan pada tiap metode kemiripan. Berikut hasil dari percobaan ketiga dalam bentuk diagram :

kgram = 2

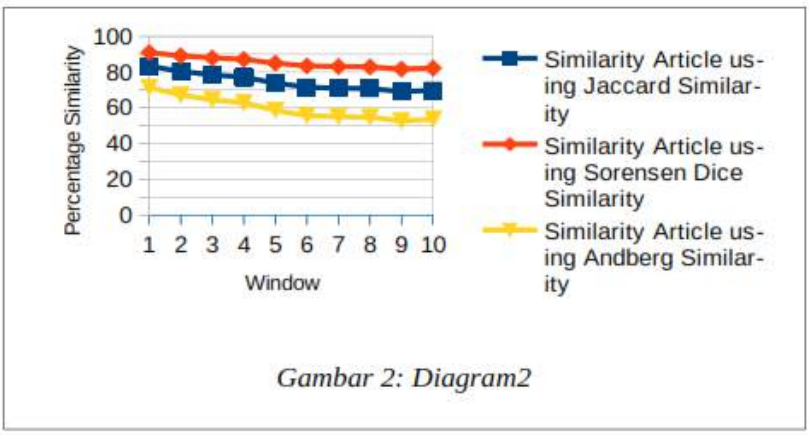


Gambar 1: Diagram1

Dengan nilai k-gram 2 dan nilai window dari 1 sampai dengan 10, analisa yang didapat :

1. Persentase kemiripan artikel menggunakan jaccard similarity mencapai 81,37%.
2. Persentase kemiripan artikel menggunakan sorensen dice similarity mencapai 89,73%.
3. Persentase kemiripan artikel menggunakan andberg similarity mencapai 68,59%.

kgram = 3



Gambar 2: Diagram2

Dengan nilai k-gram 3 dan nilai window dari 1 sampai dengan 10, analisa yang didapat :

1. Persentase kemiripan artikel menggunakan jaccard similarity mencapai 83,33%.
2. 2.Persentase kemiripan artikel menggunakan sorensen dice similarity mencapai 90,90%.
3. 3. Persentase kemiripan artikel menggunakan andberg similarity mencapai 71,43%.

Berdasarkan dari hasil uji coba ketiga maka jika 3 metode kemiripan yang digunakan diurutkan dengan nomor berikut hasilnya :

1. Sorensen Dice Similarity Coefficient
2. Jaccard Similarity Coefficient
3. Andberg Similarity Coefficient

Kesimpulan

Dari tiga tahap uji coba dan analisa terhadap algoritma winnowing untuk mendeteksi plagiasi pada sebuah teks, berikut kesimpulan yang diambil :

1. Semakin tinggi nilai window maka akurasi persentase kemiripan dari teks akan menurun.
2. Semakin kecil nilai k-gram maka akurasi persentase kemiripan dari teks akan meningkat.
3. Penggunaan nilai k-gram 1 akan menghasilkan persentase kemiripan yang tinggi tetapi tidak efektif karena teks hanya akan dipisah ke dalam bentuk abjad.
4. Nilai k-gram dan nilai window pada algoritma winnowing sangat mempengaruhi nilai persentase kemiripan pada teks.
5. Basis bilangan prima pada proses rolling hash di algoritma winnowing mempengaruhi hasil persentase kemiripan pada teks.
6. Perbedaan hasil persentase kemiripan yang dihasilkan dari 3 metode kemiripan disebabkan oleh perbedaan rumus yang digunakan pada tiap metode untuk menghitung dokumen fingerprint.

Daftar Pustaka

[1] Ana Kurniawati, Wayan Simri Wicaksana, “Perbandingan Pendekatan Deteksi Plagiarism Dokumen Dalam Bahasa Inggris”, 2008.

[2] Febrina Nafasati Prihantini, Dian Indudewi, “Kesadaran dan Perilaku Plagiarisme di Kalangan Mahasiswa”, Jurnal Dinamika Sosial Budaya, Volume 18, Nomor 1, 2016.

[3] Jeff M. Phillips, “Data Mining”, University of Utah, CS 6140, 2015.

[4] Jody, Agung Toto Wibowo, Anditya Arifianto, “Analsis dan Implementasi Algoritma Winnowing dengan Synonym Recognition pada Deteksi Plagiarisme untuk Dokumen Teks Berbahasa Indonesia”, Volume 2, ISSN:2355-9365, 2015.

[5] Khuat Thanh Tung, Nguyen Duc Hung, Le Thi My Hanh, “A Comparison of Algorithms used to measure the Similarity between two documents”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4, Issue 4, 2015.

[6] Nurdin, Amin Munthoha, “Sistem Pendeteksian Kemiripan Judul Skripsi Menggunakan Algoritma Winnowing”, ISSN:2540-7597.

[7] Pranajaya, “Analisis Dan Pencegahan Plagiarism Di Kalangan Mahasiswa : Studi Kasus Di Fakultas Teknologi Informasi Universitas Yarsi”, ISBN:9-789-7936-499-93, 2017.

[8] Reynald Karisma Wibowo, Khafiizh Hastuti, “Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Teks Pada Tugas Akhir Mahasiswa”, Techno.COM, Vol. 15, No. 4,303-311, 2016.

[9] Sariyanti Astutik, Andharini Dwi Cahyani, Mochammad Kautsar Sophan, “Sistem Penilaian Esai Otomatis Pada E-Learning Dengan Algoritma Winnowing”, Jurnal Informatika, Vol.12, No.2, 47-52, ISSN:1411-0105, 2014.