Penerapan Algoritma *Cosine Similarty* Dan Pembobotan TF-ID Pada System Klasifikasi Dokumen Skripsi

Forman Sahala Tua Saragih G [*], Marsono [**], Jaka Prayudha [*]
[#1]Program Studi Sistem Informasi, STMIK Triguna Dharma
[#2,3]Program Studi Sistem Informasi, STMIK Triguna Dharma

Article Info
Article history:
Receivedxxxx xx[th], 2020
Revised xxxx xx[th], 2020
Accepted xxxx xx[th], 2020

Keyword:
Cosine Similarty
Dokumen Skripsi

ABSTRACT
Algoritma Cosine Smilerty digunakan untuk mendapatkan periode dari sebuah Oracle yang diberikan. Algoritma ini menentukan proses yang harus dilakukan komputer kuantum sehingga hasil yangdiharapkan dapat ditemukan secara efisien dan dengan biaya komputasi seminimal mungkin. Skripsi adalah karya tulis ilmiah yang mengemukakan pendapat penulis berdasarkan pendapat orang lain. Pendapat yang diajukan harus didukung oleh data dan fakta empiris-objektif, baik berdasarkan penelitian langsung (observasi lapangan). Dalam penerapannya Algoritma Cosine Smilerty sangat sesuai untuk mengamankan dokumen skripsi sehingga menghasilkam tingkat keamanan yang baik.

First Author
Nama: Forman Sahala Tua Saragih G
Kantor : STMIK Triguna Dharma
Program Studi : Sistem Informasi
E-Mail : formansalahatuaragihg@gmail.com

1. PENDAHULUAN

Pemanfaatan komputer sebagai alat bantu untuk menyelesaikan pekerjaan manusia semakin berkembang bahkan sudah hampir mencapai taraf kecerdasan manusia. Di mana, berbagai macam aplikasi dikembangkan dan dibangun sesuai dengan kebutuhan sehingga dapat membantu pengguna dalam mengatasi masalahnya ketika ahli yang dibutuhkan tidak berada di tempat, misalnya untuk menentukan plagiarisme suatu karya ilmiah sehingga peneliti tidak dapat melakukan klaim atas penelitian orang lain karena dapat diketahui tingkat kemiripan hasil penelitiannya dengan penelitian yang sudah ada sebelumnya terutama dalam penulisan skripsi pada perguruan tinggi.

Universitas maupun perguruan tinggi swasta yang ada di kota Medan memiliki beberapa program studi yang mewajibkan mahasiswa untuk melakukan penelitian ilmiah dalam bentuk skripsi sebagai syarat untuk memperoleh gelar sarjana di program studi tersebut, salah satunya program studi Teknik Informatika dan Sistem Informatika Komputer. Program studi ini dalam menentukan judul penelitian yang akan dilakukan oleh mahasiswa memberikan syarat tidak boleh sama dengan peneliti lainnya baik yang masih melakukan penelitian atau sudah selesai. oleh karena itu, program studi ini harus memiliki indikator tertentu yang digunakan untuk menentukan judul penelitian tersebut disetujui atau tidak disetujui untuk dilakukan oleh mahasiswa yang mengajukannya. Indikator yang digunakan program studi masih menggunakan cara manual berdasarkan objek yang diteliti serta metode yang digunakan. Maka program studi ini membutuhkan suatu aplikasi yang dapat digunakan untuk menentukan kemiripan judul skripsi berdasarkan judul skripsi sesuai dengan abstrak yang digunakan.

Penentuan kemiripan judul isi skripsi ini dibangun menggunakan algoritma Cosine Smilerty. Algoritma ini digunakan untuk memecahkan masalah *black-box* secara eksponensial lebih cepat daripada algoritma klasik, termasuk dibatasi-kesalahan algoritma probabilistik. Algoritma ini, yang akan menghasilkan percepatan eksponensial atas semua algoritma klasik yang kita anggap efisien.

Berdasarkan uraian di atas, maka penulis tertarik untuk melakukan penelitian tentang: "**Penerapan Algoritma *Cosine Smilerty* dan Pembobotan TF-ID Pada System Klasifikasi Dokumen Skripsi"** sebagai judul skripsi ini.

2. Kajian Pustaka

2.1 Algoritma

Kata algoritma, di dalam dunia literatur barat lebih dikenal dengan sebutan *Algorizm*. Panggilan inilah yang kemudian dipakai untuk menyebut konsep algorithm yang ditemukannya. Dalam bahasa Indonesia, kemudian menyebutnya sebagai algoritma [3].

2.2 Cosine Similarty

Cosine similarity adalah ukuran kesamaan yang lebih umum digunakan dalam *information retrieval* dan merupakan ukuran sudut antara vektor dokumen *Da* (titik (ax,bx)) dan *Db*(titik (ay,by).

2.3 Skripsi

Skripsi adalah karya tulis ilmiah yang mengemukakan pendapat penulis berdasarkan pendapat orang lain. Pendapat yang diajukan harus didukung oleh data dan fakta empiris-objektif, baik berdasarkan penelitian langsung (observasi lapangan).

3. Metode Penelitian

Penelitian yang akan dilakukan nantinya direncanakan kedalam langkah-langkah secara sistematis. Penelitian ini dilakukan dengan beberapa langkah atau metode, yaitu:

1. Studi Pustaka (*Library Research*)
   Studi Pustaka dilakukan dengan cara mempelajari teori literatur dan buku-buku yang berhubungan dengan objek kajian sebagai dasar dalam penelitian, dengan tujuan memperoleh dasar teoritis gambaran dari apa yang dilakukan untuk mendapatkan referensi dari beberapa sumber baik dari jurnal maupun buku yang membahas tentang metode *Cosine Smilerty*.
2. Pengamatan (*observation)*
   Pengamatan dilakukan langsung di universitas yang ada di kota medan untuk mengetahui bagaimana dalam menentukan judul skripsi mahasiswa agar terhindar dari *Plagiat*.
3. Pengumpulan Data
   Pada tahap ini dilakukan pengumpulan data dengan cara menanyakan langsung kepada bagian kapalah bidang study atau Kaprodi untuk mengatahui skripsi yang dibuat oleh mahasiswa tidak pernah di buat oleh mahasiswa, khusunya pada universitas tersebut.

4. Algoritma Sistem

Berikut ini adalah Langkah-langkah dari algoritma cosine similarty :

1. Pembuangan karakter yang tidak relevan
2. Pembentukan rangkaian n-gram
3. Perhitungan fungsi hash untuk setiap n-gram
4. Pembentukan window dari nilai hash
5. Pemilihan fingerprint dari setiap window
6. Persamaan jaccard coeficient

4.1 Pembuangan Karakter Yang Tidak Relevan

Sebagai contoh, akan dilakukan proses deteksi kemiripan judul 1 “Sistem Pakar Diagnosa Kanker Serviks Menggunakan Metode Bayes” terhadap judul 2 “Sistem Pakar Diagnosa Penyakit Dyspepsia Menggunakan Metode Bayes” maka penyelesaianya sebagai berikut :

Pembuangan Karakter yang Tidak Relevan Pembuangan karakter yang tidak relevan dengan teks Judul 1 “Sistem Pakar Diagnosa Kanker Serviks Menggunakan Metode Bayes” maka akan terbentuk seperti dibawah ini

sistempakardiagnosakankerserviks menggunakanmetodebayes

Gambar 1 Hasil pembuangan karakter tidak relevan

4.2 Pembentukan rangkaian n-gram

Pembentukan Rangkaian n-gram Pembentukan rangkaian n-gram, di mana jika digunakan n = 5, maka pada teks judul 1 “sistempakardiagnosakankerserviksmenggunakanmetodebayes” akan terbentuk 50 rangkaian n-gram yaitu :

siste istem stemp tempa empak mpaka pakar akard kardi ardia rdiag diagn iagno agnos gnosa nosak osaka sakan akank kanke anker nkers kerse erser rserv servi ervik rviks viksm iksme ksmen smeng mengg enggu nggun gguna gunak unaka nakan akanm kanme anmet nmeto metod etode todeb odeba debay

Gambar 2 Hasil proses pembentukan n-gram

4.3 Perhitungan Fungsi Hash untuk Setiap N-Gram

Perhitungan nilai hash pada rangkaian ngram pertama “siste” dengan nilai basis (b) = 2, panjang rangkaian ngram(n) = 5.

$$H_{(siste)} = asci(s)*2^4 + asci(i)*2^3 + asci(s)*2^2 + asci(t)*2^1 + asci(e)*2^0$$
$$= 115*16 + 105*8 + 115*4 + 116*2 + 101$$
$$= 3473$$

Hasil semua perhitungan nilai hash yaitu:

3473	3375	3502	3421	3237	3339	3304	3124	3249
3171	3341	3144	3199	3153	3299	3409	3395	3348
3123	3243	3176	3363	3307	3304	3494	3445	3317
3517	3495	3315	3380	3439	3301	3231	3340	3257
3325	3451	3268	3125	3247	3186	3379	3338	3289
3444	3273	3115	3131	3145				

Gambar 3 Hasil perhitungan nilai hash

4.4 Pembentukan Window Dari Nilai Hash

Pembentukan *window* dari hasil perhitungan nilai *hash* pada tahap sebelumnya dengan ukuran lebar *window* (w) = 7

4.5 Pemilihan Fingerprint Dari Setiap Window

Pemilihan nilai *fingerprint* dari hasil pembentukan *window* pada tahap sebelumnya adalah sebagai berikut:
Fingerprint yang terbentuk yaitu :
Fingerprint judul 1 adalah 3237, 3124, 3144, 3123, 3176, 3304, 3315, 3301, 3231, 3125, 3186, 3115.
Fingerprint judul 2 dengan langkah proses sama dengan judul 1 adalah 3237, 3124, 3144, 3123, 3176, 3304, 3315, 3301, 3231, 3125, 3186, 3115.

4.6 Persamaan Jaccard Coeficient

Perhitungan kesamaan dengan menggunakan persamaan *jaccard coefficient* yaitu:
Fingerprint Judul 1: 3237, 3124, 3144, 3123, 3176, 3304, 3315, 3301, 3231, 3125, 3186, 3115
Fingerprint Judul 2: 3237, 3124, 3144, 3153, 3193, 3160, 3337, 3279, 3204, 3151, 3231, 3125, 3186 3115
Jumlah *fingerprint* yang sama (3237, 3124, 3144, 3231, 3125, 3186,3115) = 7
Keseluruhan *fingerprint* = (3237, 3124, 3144, 3153, 3193, 3160, 3337, 3279, 3204, 3151, 3231, 3125, 3186, 3115, 3123, 3176, 3304, 3315, 3301) = 19

$$Similarity(Kemiripan) = \frac{7}{19}*100\%$$
$$= 36\%$$

Berdasarkan hasil kesamaan-kesamaan kedua fingerprint, maka prosentase kemiripan teks antara judul 1 dan judul 2 yang dibentuk yaitu 36%.

5. Pembentukan rangkaian n-gram

1. Form Login

Form login merupakan form untuk dapat mengoperasikan program yang telah dirancang, terlebih dahulu user memasukan user name dan password setelah diinput maka sistem akan mengvalidasa data tersebut, jika sesuai maka akan muncil tampilan menu utama, form login dapat dilihat pada gambar berikut ini.

Gambar 4 Form Login

2. Form Menu Utama

Form utama merupakan tampilan utama dari isi program, dimana dengan tampilan menu utama yang ada di user dapat melakukan pengoperasian program secara maksimal dan juga dapat menggunakan fasilitas yang ada pada program tersebut.

Gambar 5 Form Menu Utama

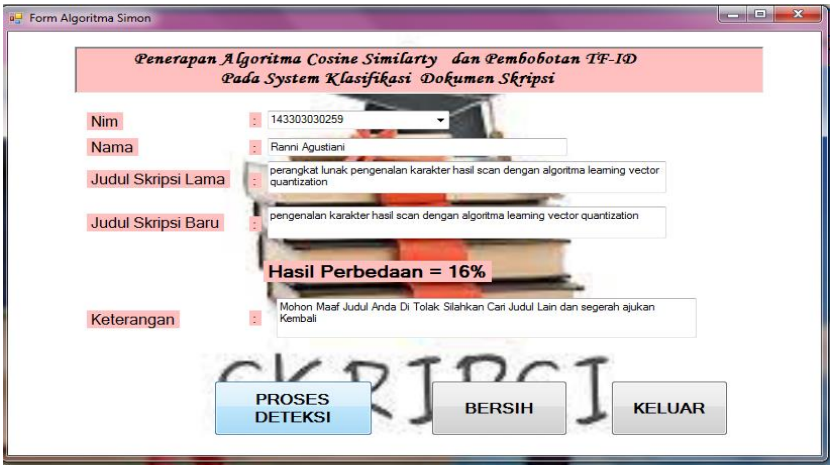
3. Form Data Skripsi

Form data skripsi berfungsi untuk menginput data mahasiswa yang sedang mengajukan judul skripsi.

Gambar 6 Form Data Skripsi

4. Form Algoritma Cosine Similarty

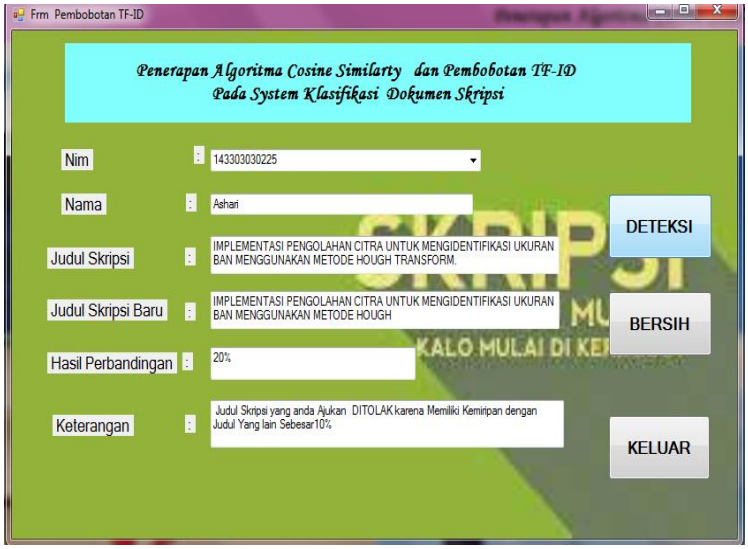
Form data cosine similarty berfungsi untuk menguji atau mengetahui tingkat kemiripan judul yang diajukan. Untuk menjalankan program ini terlebih dahulu menginput data skripsi yang baru diajukan. Setelah itu dilakukan analisa terhadap data yang telah diinput apakah pernah dibuat oleh mahasiswa lain atau tingkat kemiripan judul yang diajukan dengan judul sebelumnya, sehingga dari verifikasi judul tersebut dapat ditentukan apakah judul tersebut diterima atau ditolak.



Gambar 7 Form Algoritma Cosine Similarty

5. Form Pembobotan TF-ID

Form pembobotan TF-ID berfngsi untuk melakukan anlisa terhadap judul yang diajukan oleh mahasiswa, untuk menentukan apakah judul yang diajukan diterima atau ditolak. Penentuan apakah judul diterima atau ditolak dilakukan proses pembobotan TF-ID yang berfungsi untuk mengetahui judul yang diajukan apakah sudah pernah dibuat atau tidak, atau berapa besar kemiripan judul yang diajukan dengan judul yang sudah ada sebelumnya. Maka dari proses pembobotan TF-ID dapat ditentukan apakah judul skripsi diterima taua ditolak oleh pihak kaprodi maupun kampus yang bersangkutan.



Gambar 8 Form Pembobotan TF-ID

UCAPAN TERIMA KASIH

Terima Kasih diucapkan kepada semua pihak yang terlibat dalam pembuatan jurnal ini terlebih kepada teman saya Ronaldo Situmorang. Semoga Jurnal ini dapat dimanfaatkan dengan baik dan benar.

REFERENSI

[1] Afuan, L. (2013). *Stemming Dokumen Teks Bahasa Indonesia Menggunkan Algoritma Porter*. Jurnal Telematika, Vol. 6 No. 2.

[2] Alkautsar, A. (2012). *Perbandingan Efisiensi Model Ruang Vektor pada Sistem Temu Kembali Informasi.Bogor*: Institut Pertanian Bogor.

[3] Aziz, A. R. (2015). *Implementasi Vector Space Model dalam Pembangkitan Frequently Asked Question Otomatis dan Solusi yang Relevan untuk Keluhan Pelanggan.* Scientific Journal of Informatics, Vol.2, hlm.111-122.

[4] Baiti, N. A. (2017). *Query Answering System Hadis Muttafaqun 'Alaih Menggunakan Metode Dice Similarity dan Thesaurus Based Query Expansion.Malang*: UIN Maulana Malik Ibrahim Malang.

[5] Bunyamin, H. (2008). *Aplikasi Information Retrieval (IR) CATA Dengan Metode Generalized Vector Space Model*.Bandung: Universitas Kristen Maranatha.

[6] Chahal, M. (2016). *Information Retrieval using Dice Similarity Coefficient. International Journal of Advanced Research in Computer Sciencne and Software Engineering*, Vol.6, hlm.72-75.

[7] hada, V. (2013). *Comparison of Jaccard, Dice, Cosine Similarity Coefficient to Find Best Fitness Value for Web Retrieve Documents using Genetic Algorithm.* Interational Journal of Innovations in Engineering and Technology, Vol.2, hlm.202-205.

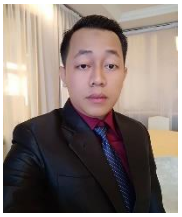
[8] Hakim, Rachmad, S. (2015). *Visual Basic 2010.* Jakarta: PT. Elex Media Komputindo.

[9] Munir, R. (2009). *Algoritma dan Pemrograman.* Bandung: Informatika.

[10] Suarna, Nana, ST. (2018). *Pedoman praktikum Microsoft Office Access 2010.* Bandung: CV. Yrama Widya.

[11] Rosa.A.S, (2014) *Pemodelan sistem Rekayasa perangkat lunak.* Jakarta: PT. Elex Media Komputindo

BIOGRAFI PENULIS

	**Forman Sahala Tua Saragih G**
	**Marsono, S.Kom., M.Kom**
	**Jaka Prayudah, S.Kom., M.Kom**