Available *online* at : http://bit.ly/InfoTekJar

InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan

ISSN (Print) 2540-7597 | ISSN (Online) 2540-7600

Optimalisasi Algoritma Rabin Karp menggunakan TF-IDF Dalam Pencocokan Text Pada Penilaian Ujian Essay Otomatis

Saeful Bahri[1], *Rusda Wajhillah*[2*]

[1]*STMIK Nusamandiri, Jl. Damai No. 8, Warung Jati Barat (Margasatwa), Pasar Minggu Jakarta Selatan, Indonesia*
[2]*Universitas Bina Sarana Informatika, Jl. Cemerlang No. 8, Sukakarya Sukabumi, Indonesia*

KEYWORDS

Rabin Karp, TF-IDF, Penilaian Otomatis.TF-IDF

CORRESPONDENCE

Phone: +62 (0751) 12345678

E-mail: rusda.rwh@bsi.ac.id

A B S T R A C T

Penilaian capaian belajar merupakan salah satu tolak ukur dalam keberhasilan proses belajar mengajar, salah satu metode pengukuran capaian tersebut adalah dengan penilaian essay, namun pada prosesnya penilaian dengan essay terdapat beberapa kekurangan diantaranya objektifitas penilai dalam memberikan hasil penilaian tersebut, beberapa peneliti telah melakukan penelitian tentang sebuah sistem penilaian essay secara otomatis diantaranya menggunakan beberapa algoritma seperti LSA (*Latent Semantic analysis*) dan *Neural Network*, algoritma tersebut memiliki beberapa kekurangan seperti pada LSA yang memiliki kekurangan dalam penanganan vector dalam mencocokan teks, sedangkan NN perlu data yang besar dalam mencocokan teks, pada penelitian ini akan diterapkan algoritma rabin karp, yang bekerja secara langsung dalam mencocokan teks berdasarkan Hash yang ditambahakan TF-IDF yang berguna untuk melakukan pengindekan dan menghitung frekuensi kemunculan teks pda sebuah dokumen, kedua metode ini terbukti mampu meningkatkan hasil pencocokan sebesar 11,81%

PENDAHULUAN

Penilaian essay merupakan sebuah proses pengukuran dan penilaian hasil capaian belajar, dalam proses penilaian esai memakan banyak waktu [1], penilaian essay perlu dilakukan menggunakan sebuah sistem yang mampu melakukan penilaian secara otomatis, karena jika penilaian dilakukan secara manual hasil penilaian akan subjektif [2][3], Metode penilaian essay otomatis menggunakan beberapa metode seperti algoritma rabin karp [3], Latent Semantic Analysis [4] dan Neural Network[5] telah banyak diusulkan oleh banyak peneliti dalam penilaian otomatis essay.

Latent semantic analysis kuat dalam pengindexan teks dengan kata kunci sedangkan memiliki kelemahan dalam menangani dimensi vektor yang berhubungan dengan pengukuran kemiripan teks [6], neural network memiliki kelebihan dalam mentolerir kesalahan dalam pengukuran kemiripan, akan tetapi salah satu kekurangan yang dimiliki oleh neural network yaitu memerlukan keyword data yang besar [7].

Rabin karp mampu memecahkan masalah yang ada pada LSA yaitu penanganan dimensi vector karena rabin karp bekerja langsung dengan cara mencocokan teks dengan menggunakan hash sebagai parameter untuk pengukuran tingkat kemiripan teks [3]. Tetapi rabin karp memiliki kelemahan dalam menangani teks dengan nilai hash yang sama dan kata yang kurang bermakna [8][9].

TF-IDF merupakan sebuah metode pembobotan teks dengan memanfaatkan index dari teks, metode ini menghitung kemunculan kata yang ada dalam sebuah teks[10], sehingga kata dasar dari teks dapat ditemukan dan nilai hash akan menjadi lebih spesifik didapat. Pada penelitian ini TF-IDF akan diterapkan untuk meringkas jawaban essay, sehingga hasil perbandingan antara jawaban dan kunci jawaban menjadi lebih akurat.

STATE OF THE ART

Algortima rabinkarp juga telah digunakan pada penelitian Anan Bahrul Khoir dkk (2019) , dalam penelitian nya yang berjudul *Implementation of rabin-karp algorithm to determine the similarity of synoptic gospels*, pada penelitian ini algoritma rabinkarp diterapkan dalam mendeteksi kesamaan pada kitab injil sinoptik, yang menghasilkan injil matius dan lukas memiliki bungan yang erat dari segi isi dibanding lukas dan markus.

Saeful bahri (2014) dalam penelitiannya yang berjudul penilaian ujian essay otomatis berbasis algoritma rabin karp, pada penelitian tersebut membahaas tentang penerapaan algoritma rabinkarp dalam penilaian ujian essay, proses penilaian hanya

https://doi.org/10.30743/infotekjar.v4i2.2294

dilakukan untuk ujian essay yang dikategorikan objektiv question. Dan jawaban hanya terbatas pada 150 karakter.

Satia suhada (2017) melakukan penelitian tentang rabin karp dengan judul Implementasi algoritma rabinkarp dan stemming najief andriani untuk deteksi plagiarisme dokumen, dalam penelitian ini dibahas tentang menggabungkan dua algoritma yaitu najief andriani dan rabinkarp, najief andriani digunakan untuk melakukan stemming kata kemudian rabin karp digunakan untuk mendeteksi kemiripan kata dengan persamaan dice similarity coefisien, hasil yang dihailkan dari pennelitian tersebut adalah meningkatnya akurasi karena corpus dataa yaang dilatih menjadi relatif lebih bersih.

Rabin Karp

Algoritm rabinkarp pertamakali ditemukan oleh Michael O rabin dan Richar M Karp pada tahun 1987 mereka menciptakan sebuah algoritma string matching. Prinsip dasar algotima ini adalah dengan cara mecari persamaan pola antara text input dan text pembanding, dengan nilai kompleksitas dari algoritma ini adalah O((n – m + 1)m) to O(nm + 1), Rabin karp merupakan sebuah algoritma pencarian yang memanfaatkan pola berupa substring dalam sebuah teks menggunakan hashing[9][8], untuk teks dangan panjang n dan panjang m, memiliki waktu komputasi terbaik adalah O(n), yang terburuk adalah O((n-m+1)m). Terdapat lima langkah dalam pencocokan string menggunakan algoritma rabin karp diantaranya:

1. Menghilangkan tanda baca, merubah teks kedalam kalimat dasar dan merubah huruf kapital jadi non kapital.
2. Membagi teks kedalam gram-gram dengan nilai gram ditentukan secara acak.
   Contoh : sebuah kalimat memiliki gram= 4
   Kalimat : saya pergi ke pasar
   K-Gram : 4
   Hasil : |saya|pergi|kepa|sar
3. Menemukan nilai hash dengan fungsi rolling hash dari masing-masing gram yang terbentuk.
4. Menemukan persamaan nilai hash dari 2 teks.

Menentukan persamaan 2 buah teks dengan persamaan *Dice's Similarity Coeficient.*

Dice Similarity Coefisien

Istilah dice Similarity Coefisien digunakan untuk menentukan kesamaan antara kedua dokumen. Dalam Rabin-Karp, teks-teks akan dibagi menjadi beberapa substring. Setelah itu, setiap karakter dalam substring secar simutan akan diubah menjadi angka biner fungsi yang disebut hashing. Proses ini akan menghasilkan nilai hash untuk masing-masing dokumen. Kemudian, algoritma akan mencari nilai hash yang mirip dalam dokumen selanjutnya disebut sebagai Finger Print.

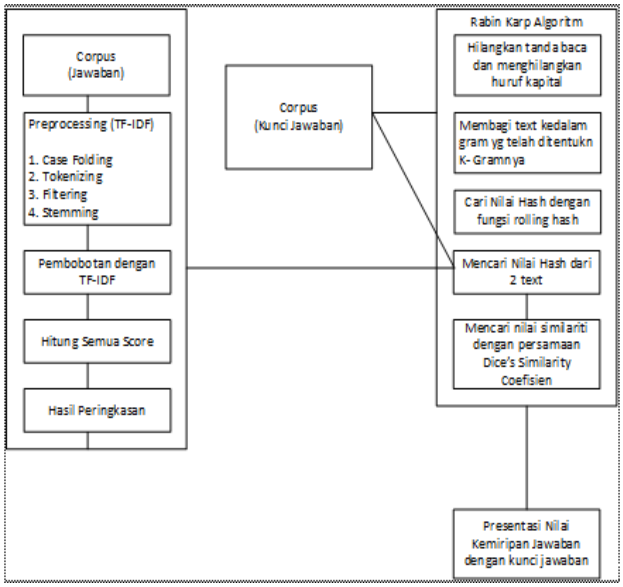
Setelah menemukan nilai hash unik dari kedua dokumen, kemudian mencari nilai hash yang ditemukan di keduanya sidik jari. Perhitungan kesamaan antara dua dokumen berdasarkan rumus di bawah ini.

$$SC(X,Y)=2\frac{(x \cap y)}{|x|+|y|}$$

Di mana SC (X, Y) mewakili tingkat kesamaan berkisar dari 0 hingga 1. X mewakili jumlah sidik jari dalam dokumen X, dan Y sendiri mewakili jumlah sidik jari dalam dokumen Y

METODE PENELITIAN

Dalam penelitian ini kami melakukan penambahan preprosesing dalam tahapan pencarian bobot kalimat yang paling banyak muncul proses penelitian dapan dilihat pada gambar berikut



(Sumber: Penelitian, 2019)

Gambar 1. Diagram Alir penelitian

Tahap tahap dalaam penelitian akan dijelaskan dibawaha ini.

Preprocessing(Rabin Karp)

1. Menghilangkan tanda baca dan merubah kalimat menjadi lower case (Cleaning), Pada tahap ini, corpus jawaban harus dibersihkan dengan cara menghilangkan tanda baca seperti ".", ",", "?","!". Tujuannya adalah agar corpus jawaban yang akan dijadikan sebagai acuan terbebas dari kata-kata atau simbol lain yang tidak memiliki manfaat substansi dari kalimat. Masih pada tahap ini juga, dilakukan prose cleaning corpus jawaban berupa penyetaraan jenis huruf menjadi huruf kecil semua.
2. Membagi teks kedalam gram dengan k-gram yang telah ditentukan, Langkah ini dilakukan dengan cara membagi rata teks jawaban ke dalam k-gram sesuai dengan ketentuan awal.
3. Mencari nilai hash dengan fungsi rolling hash dari tiap gram yang terbentuk, Setelah gram terbentuk, berikutnya adalah mencari nilai hash. Hashing dibutuhkan pada langkah ini sebagai langkah untuk mengubah jenis data yang ada menjadi bilangan bulat yang sederhana.
4. Mencari nilai hash yang sama antara 2 teks, Teks yang sama dari corpus jawaban dan corpus kunci jawaban kemudian dicari nilai hashnya
5. Menentukan persamaan 2 teks dengan persamaan dice similarity Coefisien penentuan persamaan teks dengan dice similarity coefisien, dengan cara menghitung jumlah K-Gram yang digunakan pada jawaban dan kunci jawaban yang akan menghasilkan dokumen fingerprint yang didapat dari jumlah nilai K-Gram yang sama. Dice similarity Coeficient dapat dihitung dengan persamaan berikut

$$S = \frac{2C}{A + B}$$

Dimana S adalalah nilai similarity dan C jumlah K-Gram yang diambil dari jawaban dan kunci jawaban yang dibandingkan sedagkan A dan B merupakan nilai K dari masing masing jawaban dan kunci jawaban.

Preprocessing (TF-IDF)

Pada Proses pengoptimalisasi Rabin karp dan TF -IDF

Corpus yang digunakan didapat dari:

1. Jawaban
   Jawaban kemudian akan diprosess menggunakan TF-IDF untuk diringkas kalimatnya, setelah kalimat ringkasan didapat selanjutnya, diproses kembali menggunakan algoritma rabin karp untuk mengetahui tingkat kemiripan dari jawaban dan kunci jawaban.
2. Kunci jawaban
   kunci jawaban didapat dari isian jawaban yang diinput oleh penilai, kunci jawaban yang dibuat harus memiliki beberapa keyword, kunci jawaban akan melalui fase preprocessing untuk selanjutnya dibandingkan dengan jawaban

HASIL DAN PEMBAHASAN

Hasil penelitian pada pembahasan sebelumnya, dilakukan dengan cara mengimplentasikan algoritma rabinkarp kedalam sebuah bahasa pemrograman, kemudian hasil implementasi tersebut dibandingkan untuk mencari perbedaan hasil analisis kemiripannya dengan cara menghitung nilai hash berdasarkan *dice similarity*.

1. Studi kasus
   Dalam penelitian ini diambil sebuah studi kasus ujian mata kuliah yang bersifat essay. Hasil pencocokan yang diteliti adalah hasil pencarian kemiripan menggunakan algoritma rabin karp dan hasil pencarian kemiripan dengan rabin karp ditambah dengan algoritma TF-IDF, dengan metode penentuan kesamaan teks yang menggunakan *dice similarity coefisien.*

2. Analisis Hasil
   Dari hasil perhitungan kemiripan dengan *dice similarity coefisien* dengan kunci jawaban: “Metode untuk melakukan pengukuran system informasi” dapat dilihat dalam tabel hasil penelitian sebagai berikut:

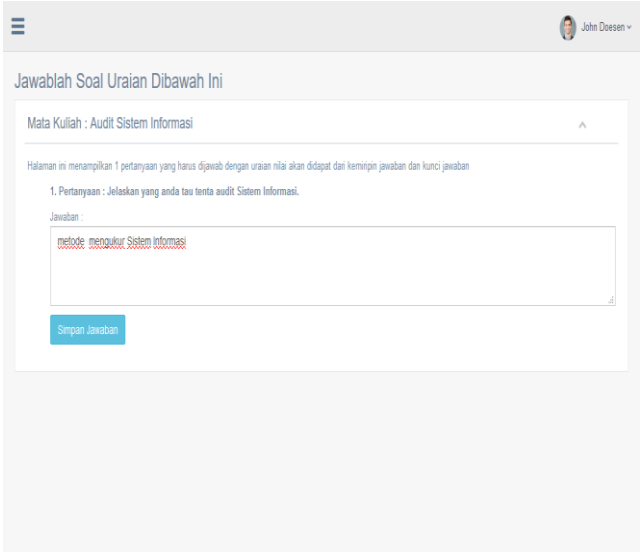
Tabel 2. Hasil Penelitian

Jawaban	Rabin karp	Rabin Karp +TDF-IDF	Hasil
Metode untuk melakukan pengukuran Sistem Informasi	100%	100%	0
cara untuk melakukan pengukuran Sistem Informasi	89,13%	99,25%	10,12%
cara untuk mengukur Sistem Informasi	60,00%	78,12%	18,12%
Metode untuk mengukur Sistem Informasi	73,17%	80,12%	6,95%
Metode untuk mengukur Sistem Informasi	57,89%	70,02%	12,13%

(Sumber: Hasil Penelitian, 2019)

Berdasarkan 4 sample jawaban yang berbeda dengan kunci jawaban yang sama diperoleh rata-rata peningkatan hasil kemiripan sebesar 11.81%.

3. Hasil Implementasi Algoritma.
   Algoritma diimplentasikan menggunakan bahasa pemrograman berbasis server berikut hasil dari implementasi algoritma rabin-karp tersebut.



Gambar 2. Implementasi Hasil Penelitian

KESIMPULAN

Dari hasil penerapan TF-IDF pada proses pengelompokan kalimat yang akan dicari dan hitung tingkat kemiripan dengan algoritma rabinkarp mengalami rata-rata peningkatan sebesar 11,81%. Hal ini terjadi karena kalimat akar dikelompokan dan dihitung frekuensi kemunculan index oleh TF-IDF sehingga akar kata menjadi lebih mudah dan lebih akurat untuk dirubah kedalam fungsi hash.

Namun terdapat beberapa kekurangan dalam metode ini seperti tidak diketahuinya waktu eksekusi dari algoritma yang digunakan, sehingga perlu penelitian lebih lanjut tentang menghitung kompleksitas dari kedua algoritma yang digunakan.

UCAPAN TERIMAKASIH

Terimakasih kami ucapkan kepada semua pihak yang terlibat dalam penelitian ini. Mudah-mudahan penelitian dapat memberikan manfaat bagi semua pembaca.

https://doi.org/10.30743/infotekjar.v4i2.2294

REFERENCES

[1] K. Zupanc and Z. Bosnic, "Automated essay evaluation with semantic analysis," *Knowledge-Based Syst. Vol.*, vol. 120, pp. 118–132, 2017.

[2] C. Ramineni and D. M. Williamson, "Automated essay scoring: Psychometric guidelines and practices," *Assess. Writ.*, vol. 18, no. 1, pp. 25–39, Jan. 2013.

[3] E. Rasywir, Y. Pratama, Hendrawan, and M. Istoningtyas, "Removal of modulo as hashing modification process in essay scoring sistem using rabin-karp," *Proc. 2018 Int. Conf. Electr. Eng. Comput. Sci. ICECOS 2018*, vol. 2019-Janua, pp. 159–164, 2019.

[4] R. Setiadi Citawan, V. Christanti Mawardi, and B. Mulyawan, "Automatic Essay Scoring in E-learning Sistem Using LSA Method with N-Gram Feature for Bahasa Indonesia," *MATEC Web Conf.*, vol. 164, p. 01037, 2018.

[5] F. Dong, Y. Zhang, and J. Yang, "Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring," pp. 153–162, 2017.

[6] X. Hu, Z. Cai, P. Wiemer-Hastings, A. C. Graesser, and D. S. McNamara, "Strengths, Limitations, and Extensions of LSA," in *Handbook of Latent Semantic Analysis*, Routledge, 2015, pp. 1–40.

[7] H. Nguyen and L. Dery, "Neural Networks for Automated Essay Grading," pp. 1–11, 2016.

[8] S. Suhada and S. Bahri, "Implementasi Algoritma Rabin Karp Dan Stemming Najief Andriani Untuk Deteksi Plagiarisme Dokumen," *Swabumi*, vol. 5, no. 1, pp. 84–89, 2017.

[9] S. Bahri, "PENILAIAN OTOMATIS UJIAN ESSAY ONLINE BERBASIS ALGORITMA RABIN KARP," *Swabumi*, vol. 1, no. 1, 2014.

[10] M. E. Sulistyo, R. Saptono, and A. Asshidiq, "Penilaian Ujian Bertipe Essay Menggunakan Metode Teks Similarity," *Telematika*, vol. 12, no. 02, pp. 146–158, 2015.

[11] W. Zhang, T. Yoshida, and X. Tang, "Expert Sistems with Applications A comparative study of TF Ã IDF , LSI and multi-words for teks classification," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2758–2765, 2011.

BIODATA PENULIS

Saeful Bahri
Memiliki minat dibidang programming, Menyelesaikan Pendidikan pasca sarjana dibidang Ilmu Komputer pada tahun 2016. sejak tahun 2013 sampai saat ini masih aktif sebagai pengajar di salah satu perguruan tinggi swasta terkemuka di Indonesia.

Rusda Wajhillah
Anak ke-enam dari tujuh bersaudara Menyelesaikan Pendidikan formal Pasca Sarjana Ilmu Komputer pada tahun 2013. Aktif mengajar di salah satu Universitas di Indonesia. Selain mengajar, juga aktif dalam pembinaaan organisasi mahasiswa di lingkungan kampus.