KOMPARASI KINERJA ALGORITMA SIMILARITAS *INNER PRODUCT FAMILY* PADA *RULE BASE STEMMER*STUDI KASUS DOKUMEN TEKS BAHASA JAWA

Fatkhul Amin[1], Sugiyamta[2], Arif Jananto[3]
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Stikubank
e-mail: [1]fatkhulamin@edu.unisbank.ac.id, [2]sugiyamta@edu.unisbank.ac.id, [3]ajananto09@gmail.com

ABSTRAK

Komparasi Algoritma Similaritas Inner Product Family dilakukan untuk mengetahui efektifitas algoritma dalam menemukan dokumen teks pada studi kasus dokumen teks bahasa jawa. Dokumen bahasa jawa yang menjadi obyek sejumlah 48.753 kata yang didapatkan dari majalah bahasa jawa penjebar semangad, Joko lodang dan Jaya Baya. Hasil dari komparasi antara metode Harmonic Mean, Dice, Kumar hassebrook, dan Cosine; Dokumen Teks Bahasa Jawa dengan No Dokumen ARI163, JL112014BD dan LEO31 pada metode Harmonic Mean, Dice, Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.822 (ARI163), 0.411 (JL112014BD) dan 0.111 (LEO31). Dokumen Teks Bahasa Jawa dengan No Dokumen LEO63 pada metode Harmonic Mean dan Dice menghasilkan bobot dokumen yang sama yaitu 0.263. pada metode Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.065. Metode Harmonic mean dan Dice menghasilkan bobot lebih tinggi yaitu 0.263 dibandingkan dengan metode Kumar-Hassebrook dan Cosine yang menghasilkan bobot 0.065. Dokumen Teks Bahasa Jawa dengan No Dokumen LEO14 pada metode Harmonic Mean, Dice dan Kumar-Hassebrook menghasilkan bobot dokumen yang sama yaitu 0.111. pada metode Cosine menghasilkan bobot dokumen 0.104. Metode Harmonic mean, Dice dan Kumar-Hassebrook menghasilkan bobot lebih tinggi yaitu 0.111 dibandingkan dengan metode Cosine yang menghasilkan bobot 0.104. Dokumen terambil paling sedikit oleh Metode Cosine yaitu 5 dokumen, Kumar-hassebrook 19 dokumen, Dice 20 dokumen dan Harmonic Mean 29 dokumen

***Kata Kunci**: Komparasi similaritas, Harmonic Mean,Dice, Kumar hassebrook, Cosine.*

1. PENDAHULUAN

Informasi yang cepat dan akurat dari hasil pencarian pada Sistem Temu Kembali Informasi (STKI) atau Mesin pencari (search engine) menjadi hal utama bagi pengguna. Saat ini hasil pencarian informasi pada setiap mesin pencari memiliki model dan hasil yang berbeda-beda dan menghasilkan hasil pencarian dengan hasil dokumen teks dalam jumlah yang besar. Hasil pencarian dengan tingkat akurasi tinggi merupakan harapan dari pengguna STKI ketika mencari informasi. Sistem temu kembali informasi akan memberikan nilai tambah dalam pecarian informasi jika keinginan *user* bisa terpenuhi. Ketika informasi bisa kita dapatkan dengan cepat dengan menghasilkan hasil pencarian yang sangat banyak membuat pengguna memerlukan waktu lebih dalam memilih dokumen yang sesuai dengan keinginannya. STKI dengan hasil pencarian akurat dan dokumen yang dihasilkan tidak banyak bisa digunakan metode similaritas dari kelompok *Inner Product Family.*

Pada kelompok *Inner Product Family* ada 5 metode similaritas (Sung-Hyuk Cha , 2007) yang digunakan dalam pembuatan STKI. Adapun metode dalam kelompok *Inner Product Family* diantaranya Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice. Implementasi tiap metode menghasilkan tingkat akurasi yang tinggi dengan hasil bobot berbeda-beda. Metode-metode yang ada digunakan dalam pembuatan STKI dengan model stemmer yang sama akan menghasilkan hal yang berbeda untuk hasil pencarian, namun demikian hasil dari implementasi sudah memiliki akurasi tinggi dengan dasar hasil uji recall tinggi dan hasil precission rendah. Stemmer yang digunakan untuk mendukung komparasi adalah stemmer bahasa jawa yang telah memperhatikan semantik bahasa jawa dalam proses pembuatannya (amin, 2017). Stemmer bahasa jawa digunakan juga karena data yang digunakan adalah dokumen teks bahasa jawa.

Penelitian terdahulu tentang komparasi dilakukan oleh Khuat Thanh Tung, dkk (2016) dengan topik *A Comparison of Algorithms used to measure the Similarity between two documents.* Mengukur kesamaan dokumen memainkan peran penting dalam penelitian dan aplikasi terkait teks seperti pengelompokkan dokumen, deteksi plagiarisme, pencarian informasi, terjemahan mesin dan penilaian esai otomatis. Banyak penelitian telah diajukan untuk memecahkan masalah ini. Dokumen dapat dikelompokkan menjadi tiga pendekatan utama: Berbasis String, berbasis Corpus dan berbasis Kesamaan Pengetahuan. Pada riset ini kesamaan dua dokumen diukur dengan menggunakan dua ukuran berbasis string yang berbasis karakter dan algoritma berbasis-kata. Dalam metode berbasis karakter, n-gram digunakan untuk mencari sidik jari untuk algoritma sidik jari dan penanda, maka koefisien Dice digunakan untuk mencocokkan dua sidik jari yang ditemukan. Dalam pengukuran berbasis istilah, algoritma kesamaan Cosinus digunakan. Riset ini membandingkan keefektifan algoritma yang digunakan untuk mengukur

kesamaan antara dua dokumen. Diperoleh hasil bahwa kinerja sidik jari dan penampakan lebih baik daripada kesamaan kosinus. Selain itu, algoritma penampakan lebih stabil daripada yang lain.

Penelitian tentang komparasi sebelumnya dilakukan oleh Vikas Thada, Dkk, (2015). Dengan tema *Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web.*Sebuah koefisien kemiripan mewakili kesamaan antara dua dokumen, dua pertanyaan, atau satu dokumen dan satu query. Dokumen yang diambil juga dapat peringkat di urutan dianggap penting. Sebuah koefisien kemiripan adalah fungsi yang menghitung tingkat kesamaan antara sepasang objek teks. Ada sejumlah besar koefisien kesamaan diusulkan dalam literatur, karena yang terbaik ukuran kesamaan tidak ada (belum!). Dalam makalah ini kami melakukan analisis perbandingan untuk mencari tahu dokumen yang paling relevan untuk himpunan kata kunci dengan menggunakan tiga koefisien kesamaan yaitu Jaccard, Dice dan Cosine koefisien. performa menggunakan pendekatan algoritma genetika. Karena sifat acak dari algoritma genetika yang terbaik nilai fitness adalah rata-rata 10 berjalan dari kode yang sama untuk tetap jumlah koefisien kemiripan.Satu set dokumen diambil untuk query yang diberikan dari Google yang mengetahui kemudian rata-rata relevansi dalam hal nilai-nilai fitness menggunakan koefisien kesamaan dihitung. Dalam makalah ini kami memiliki rata-rata 10 generasi yang berbeda untuk setiap query dengan menjalankan program 10 kali untuk nilai tetap Probabilitas Crossover Pc = 0,7 dan Probabilitas Mutasi Pm = 0,01. Percobaan yang sama dilakukan untuk 10

Penelitan tentang metode similaritas juga dilakukan oleh Mingyang, dkk.(2005) dengan topik *Comparing Similarity Calculation Methods in Conversational CBR*. *Conversational Case-Based-Reasoning* (CCBR) menyediakan sebuah dialog inisiatif campuran untuk membimbing pengguna dalam menyusun deskripsi masalah mereka secara bertahap melalui rangkaian tanya jawab. Perhitungan kesamaan dalam CCBR, seperti pada CBR tradisional, memainkan peran penting dalam proses pencarian karena menentukan kualitas kasus yang diambil. Dalam makalah ini, dianalisis karakteristik yang berbeda dari query (kasus baru) antara CCBR dan CBR tradisional, dan berpendapat bahwa metode perhitungan kesamaan yang hanya memperhitungkan fitur yang muncul dalam query, disebut query-bias, lebih sesuai untuk CCBR Percobaan dirancang dan dijalankan pada 36 dataset. Hasilnya menunjukkan bahwa pada 31 dataset dari total 36, sistem CCBR menggunakan metode perhitungan kesamaan query-bias menghasilkan kinerja yang lebih efektif daripada metode perhitungan kesamaan kasus-bias dan sama-sama bias.

Adapun solusi untuk mendapatkan informasi kinerja tentang metode similaritas yang ada pada kelompok *Inner Product Family* adalah dengan membuat komparasi *software* Sistem Temu Kembali Informasi (STKI) menggunakan Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice *yang* didukung oleh *Stemmer* Bahasa Jawa.

2. METODE PENELITIAN

Pada penelitian komparasi ini mengikuti langkah-langkah sebagai berikut ;

a. Identifikasi Masalah
   Identifikasi dilakukan dengan mengamati dan melihat hasil algoritma Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice *yang* didukung oleh *Stemmer* Bahasa Jawa.
b. Telaah Kepustakaan
   Telaah pustaka dilakukan dengan melihat hasil-hasil yang didapatkan tiap-tiap proses algoritma seperti; algoritma Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice
c. Rumuskan hipotesis
   Rumusan hipotesis menggunakan algoritma Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice
d. Rumuskan Asumsi
   Dengan menggunakan data-data yang sama dilakukan komparasi menggunakan algoritma Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice
e. Rancangan Pendekatan
   Pendekatan hasil dengan nilai mendekati 1 akan diperoleh hasil paling maksimal atau paling mirip. Implementasi dari algoritma Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice.
f. Validasi
   Dilakukan komparasi antara tiap algoritma Metode Harmonic mean, Metode Cosine, Metode Kumar Hassebrok, Metode Jaccard dan Metode Dice

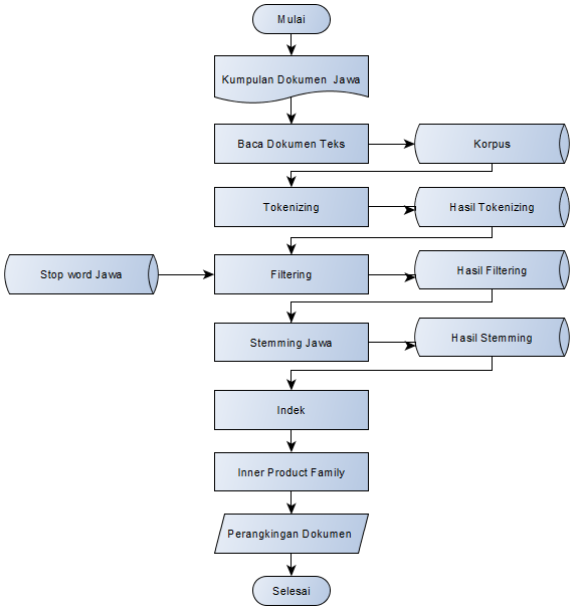
Gambar 1 menunjukkan bagan proses komparasi

Gambar 1. Proses Komparasi

3. HASIL DAN PEMBAHASAN

3.1. Flowchar STKI

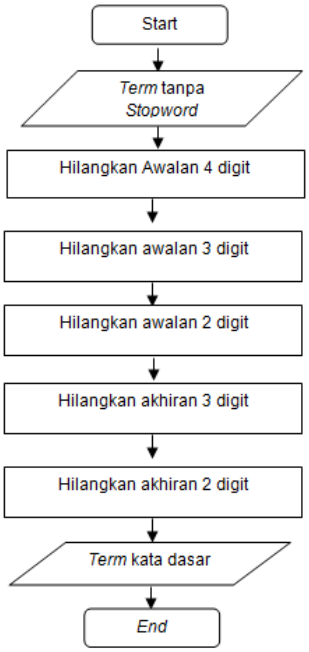
Flowchart STKI dibuat untuk memudahkan pengguna dan menampilkan model pencarian yang umum digunakan saat ini. STKI didesain untuk menemukan informasi secara akurat kepada pengguna (user). Proses STKI oleh sistem melalui proses-proses seperti gambar 2. *Flowchart*dimulai dengan *input* dokumen-dokumen teks bahasa jawa kedalam korpus. Selanjutnya dokumen teks bahasa jawa melalui proses preprosesing, dihitung bobotnya dan dihitungperingkatnya berdasarkan bobot dokumen yang tertinggi.



Gambar 2. flowchartSTKI Jawa

3.2. Stemmer Bahasa Jawa

Pada penelitian stemmer bahasa Jawa digunakan arsitektur *stemmer* bahasa Indonesia Tala. Hal ini dilakukan karena Tala melakukan penelitian tentang *stemmer* bahasa Indonesia, dan bahasa jawa adalah ibu dari bahasa Indonesia. Gambar 3 menunjukkan *stemmer* bahasa Indonesia



Gambar 3. Flowchart Proses Stemming

3.3. Hitung Harmonic-Mean, Dice, Kumar-Hassebrrok, dan Cosine

Proses selanjutnya adalah proses perhitungan pembobotan dan pemeringkatan menggunakan metode *Harmonic-Mean,* Dice, Kumar-Hassebrrok, dan Cosine. Proses ini dimulai dengan perhitungan tf, idf, tfidf, jarak dokumen dan query, similaritas dan *Harmonic-Mean,* Dice, Kumar-Hassebrrok, dan Cosine (gambar 5.3). Proses hitung Harmonic-Mean, dirancaDice, Kumar-Hassebrrok, dan Cosine ng menghasilkan dokumen hasil pencarian disertai dengan letak dokumen dan bobot dokumen.

3.4. Similaritas Inner Product Family

Pada riset ini akan di bandingkan metode similaritas pada keluarga Inner Product Family. Adapun metode-metode similaritas yang akan dilakukan komparasi adalah persamaan (1), (2), (3), dan (4)

Metode Harmonic mean

$$s_{HM} = 2\sum_{i=1}^{d}\frac{P_i Q_i}{P_i + Q_i} \tag{1}$$

Metode Cosine

$$s_{Cos} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sqrt{\sum_{i=1}^{d} P_i^2}\sqrt{\sum_{i=1}^{d} Q_i^2}} \tag{2}$$

Metode Kumar Hassebrok

$$s_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} \tag{3}$$

Metode Dice

$$s_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} \quad (4)$$

dimana p dan q adalah dokumen yang berbeda. p_i adalah term i yang ada di dokurnen p q_i adalah term i yang ada di dokumen q.

3.5 Kinerja Algoritma Similaritas Inner Product Family

Kinerja similaritas didapatkan dari proses Pengolahan dokumen yang ditempatkan di korpus. Dokumen yang telah ditempatkan dikorpus dilakukan proses Preprosessing yang didalamnya terdapat 3 (tiga) proses yaitu; Tokenizing, Filtering dan Stemming. Setelah melalui proses stemmer menggunakan bahasa jawa metode rule based, selanjutnya tiap-tiap metode diuji cobakan dengan menggunakan beberapa keyword yang sama. Adapun keyword yang diujikan adalah keyword dengan 1 term, 2 term, 3 term, 4 term dan 5 term. Melalui ke-5 keyword tersebut, selanjutnya setiap hasil didata dan dilakukan analisa komparasi setelah sebelumnya dilakukan analisis persepsi untuk setiap keyword yang diujikan. Tahap akhir dari proses melihat kinerja adalah dengan dilakukan uji recall dan uji presisi

3.6 Komparasi Kinerja Studi Kasus Keyword Bahasa Jawa

Studi kasus pada komparasi beberapa metode ini menggunakan dokumen-dokumen teks Basa Jawa pada Majalah Penjebar Semangad, Jaya Baya dan Joko Lodang. *Query* yang dimasukkan pada STKI adalah *keyword* dengan 3 *term* " wong seneng ngalah"

Tabel 1. Bobot Dokumen semua metode pada *keyword* "Wong seneng ngalah"

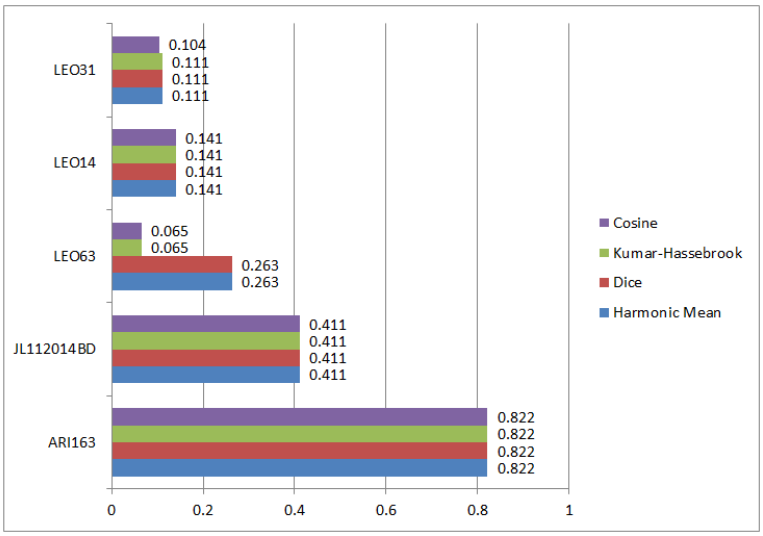
No	Dokumen	Harmonic Mean	Dice	Kumar-Hassebrook	Cosine
1	ARI163	0.822	0.822	0.822	0.822
2	JL112014BD	0.411	0.411	0.411	0.411
3	LEO63	0.263	0.263	0.065	0.065
4	LEO14	0.141	0.141	0.141	0.141
5	LEO31	0.111	0.111	0.111	0.104

Hasil pencarian dokumen dengan keyword "*wong seneng ngalah*", Tabel 1 menunjukkan Hasil:

a. Dokumen Teks Bahasa Jawa dengan No Dokumen ARI163 pada metode Harmonic Mean, Dice, Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.822.
b. Dokumen Teks Bahasa Jawa dengan No Dokumen JL112014BD pada metode Harmonic Mean, Dice, Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.411
c. Dokumen Teks Bahasa Jawa dengan No Dokumen LEO63 pada metode Harmonic Mean dan Dice menghasilkan bobot dokumen yang sama yaitu 0.263. pada metode Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.065. Metode Harmonic mean dan Dice menghasilkan bobot lebih tinggi yaitu 0.263 dibandingkan dengan metode Kumar-Hassebrook dan Cosine yang menghasilkan bobot 0.065.
d. Dokumen Teks Bahasa Jawa dengan No Dokumen LEO14 pada metode Harmonic Mean, Dice dan Kumar-Hassebrook menghasilkan bobot dokumen yang sama yaitu 0.111. pada metode Cosine menghasilkan bobot dokumen 0.104. Metode Harmonic mean, Dice dan Kumar-Hassebrook menghasilkan bobot lebih tinggi yaitu 0.111 dibandingkan dengan metode Cosine yang menghasilkan bobot 0.104.
e. Dokumen Teks Bahasa Jawa dengan No Dokumen LEO31 pada metode Harmonic Mean, Dice, Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.111.
f. Dokumen terambil paling sedikit oleh Metode Cosine yaitu 5 dokumen, Kumar-hassebrook 19 dokumen, Dice 20 dokumen dan Harmonic Mean 29 dokumen.

3.7. Diagram hasil komparasi Kinerja

Hasil kinerja komparasi algoritam similaritas Inner Product Family selengkapnya bisa dilihat pada gambar diagram 4.



Gambar 4. Bobot Dokumen semua metode pada keyword Wong seneng ngalah

5. KESIMPULAN

Dari apa yang sudah diuraikan serta penelitian yang telah penulis lakukan, maka penulis dapat menarik kesimpulan sebagai berikut:

a. Dokumen Teks Bahasa Jawa dengan No Dokumen ARI163, JL112014BD dan LEO31 pada metode Harmonic Mean, Dice, Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.822 (ARI163), 0.411 (JL112014BD) dan 0.111 (LEO31)
b. Dokumen Teks Bahasa Jawa dengan No Dokumen LEO63 pada metode Harmonic Mean dan Dice menghasilkan bobot dokumen yang sama yaitu 0.263. pada metode Kumar-Hassebrook dan Cosine menghasilkan bobot dokumen yang sama yaitu 0.065. Metode Harmonic mean dan Dice menghasilkan bobot lebih tinggi yaitu 0.263 dibandingkan dengan metode Kumar-Hassebrook dan Cosine yang menghasilkan bobot 0.065.
c. Dokumen Teks Bahasa Jawa dengan No Dokumen LEO14 pada metode Harmonic Mean, Dice dan Kumar-Hassebrook menghasilkan bobot dokumen yang sama yaitu 0.111. pada metode Cosine menghasilkan bobot dokumen 0.104. Metode Harmonic mean, Dice dan Kumar-Hassebrook menghasilkan bobot lebih tinggi yaitu 0.111 dibandingkan dengan metode Cosine yang menghasilkan bobot 0.104.
d. Dokumen terambil paling sedikit oleh Metode Cosine yaitu 5 dokumen, Kumar-hassebrook 19 dokumen, Dice 20 dokumen dan Harmonic Mean 29 dokumen.

DAFTAR PUSTAKA

[1] Amin, Fatkhul, dkk, 2017. A Hybrid Method Of Rule-Based And String Matching Stemmer For Javanese Language. Journal of Theoretical and Applied Information Technology. Vol.95. No 19. ISSN: 1992-8645
[2] Khuat Thanh Tung , 2015. 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, April 2015
[3] Mingyang, Dkk. 2005. Comparing Similarity Calculation Methods in Conversational CBR. IEEE International Conference on Information Reuse and Integration, Conf, ISBN: 0-7803-9093-8.
[4] Meadow, C.T., 1997. *Text Information Retrieval Systems*. Academic Press.New York.
[5] Tala, F.Z., 2003, *A Study of Stemming Effects on Information Retrieval in bahasa Indonesia*. Institut for logic, Language and Computation Universiteit van Amsterdam The Netherlands.
[6] Salton, G., 1989, *Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer*. Addison – Wesly Publishing Company, Inc. USA.
[7] Sung-Hyuk Cha, 2007, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. International Journal of mathematical Models and Methods in Applied Sciences. Issue 4 Volume Volume 1
[8] Vikas Thada , 2015. Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web, Department of Computer Science and Engineering Dr. K.N.M University,Newai, Rajasthan, India