RANCANG BANGUN MESIN PENCARI BAHASA JAWA MENGGUNAKAN METODE MATUSITA

Fatkhul Amin[1], Setyawan Wibisono [2], Wiwien Hadikurniawati[3]
[1,2,3] Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Stikubank
e-mail: [1]fatkhulamin@edu.unibank.ac.id, [2]setyawan@edu.unisbank.ac.id d , [3] wiwien@edu.unisbank.ac.id

Abstrak

Mesin pencari yang efektif akan menemukan dengan cepat informasi yang dibutuhkan dan hasil yang didapatkan memiliki relevansi dengan informasi yang dicari oleh pengguna. Mesin pencari dokumen teks bahasa jawametode Matusita dibuat untuk didapatkannya hasil yang efektif yaitu tingkat presisi yang tinggi atau akurat. Proses perhitungan menggunakan metode Matusita dilakukan dengan sebelumnya melalui proses tokenisasi, penyaringan dan pembuatan akar kata serta di hitung dengan algoritma Matusita. Mesin pencari metode Matusita setelah dilakukan percobaan dengan maksimal 5 suku kata menghasilkan rata-rata recall sebesar 0,02 dan rata-rata precision sebesar 0,41 yang berarti memiliki tingkat presisi tinggi. Implementasi dari algoritma matusita pada mesin pencari dokumen teks bahasa jawa dibuat dengan tampilan mesin pencari pada umumnya, yaitu akan dihasilkan sebuah pemeringkatan dari hasil pencarian dengan posisi teratas adalah dokumen dengan bobot tertinggi.

Kata Kunci*: Metode Matusita, Bahasa Jawa, mesin pencari*

1. PENDAHULUAN

Dokumen teks dalam sebuah mesin pencari ditempatkan dalam tempat tertentu yang dikenal dengan korpus. Penempatan dokumen ini dimaksudkan untuk mempermudah koneksi dalam sebuah sistem pencarian mesin pencari. Pencarian informasi berupad dokumen teks bergantung pada metode pencarian yang digunakan. Informasi-informasi yang diberikan oleh mesin pencari menghasilkan banyak sekali hasil pencarian yang berbeda-beda. Mesin pencari bekerja sesuai perannya yaitu memberikan informasi yang relevan bagi penggunanya. Informasi yang dihasilkan oleh mesin pencari memiliki hasil yang berbeda dalam keluarannya karena algoritma yang berbeda pula

Dokumentasi teks khususnya teks dalam bahasa jawa sulit ditemukan dewasa ini. Hal ini karena bahasa Jawa dari waktu ke waktu mulai ditinggalkan oleh masyarakat pada umumnya. Tidak hanya bahasa Jawa saja yang banyak ditinggalkan pada masa kini, tapi juga bahasa daerah yang lain di seluruh Indonesia. Perlunya bahasa jawa sebagai bahasa daerah yang paling banyak digunakan untuk di kumpulkan kata atau kalimat-kalimatnya dalam sebuah database yang bisa digunakan untuk dibuat mesin pencari sehingga bisa memberikan manfaat untuk orang banyak. Dokumen teks bahasa jawa akan sangat mudah untuk digunakan jika dibuat sebagai pusat dokumentasi yang dilengkapi dengan mesin pencari yang digunakan untuk mencari informasi berbahasa jawa. Model seperti ini akan membuat bahasa Jawa menjadi terpelihara dan bisa digunakan di era saat ini atau era internet. Mesin pencari berbahasa jawa dengan metode yang akurat akan bisa membantu proses pencarian dengan cepat dengan tingkat akurasi yang tinggi.

Mesin pencari dokumen teks dalam hal ini mesin pencari berbahasa jawa perlu dibuat untuk membantu dalam hal memelihara bahasa jawa dan memeudahkan implementasi mesin pencari berbahasa jawa kepada masyarakat. Mesin pencari dibuat dengan mempertimbangkan banyak hal seperti nilai kemanfaatan teknologi dalam kehidupan sehari hari. Adapun solusi untuk mengatasi masalah ini adalah dengan mesin pencari dokumen teks bahasa jawa menggunakan Metode Matusita agar hasil pencarian informasi memiliki tingkat akurasi yang tinggi dan proses pencarian yang cepat.

2. TINJAUAN PUSTAKA

Penelitian terdahulu dilakukan oleh C Selvi dkk, (2018) melakukan riset dengan topik A novel similarity measure towards effective recommendation using Matusita coefficient for Collaborative Filtering in a sparse dataset. Langkah-langkah kesamaan konvensional tidak dapat memberikan hasil yang efektif rekomendasi untuk pengguna aktif dalam dataset yang jarang. Sebagai dataset jarang mengandung lebih sedikit item yang dinilai bersama, yang langkah-langkah ventilasi gagal untuk mempertimbangkan item no-nilai bersama nilai-nilai. Ukuran MCF yang diusulkan efisien memanfaatkan semua memberi peringkat informasi tanpa mempertimbangkan hanya yang disediakan pengguna nilai item yang dinilai bersama. Akibatnya, langkah MCF menawarkan rekomendasi yang efisien untuk pengguna aktif dengan menemukantetangga yang dapat diandalkan dan mengungguli yang serupa konvensional langkah-langkah ity. Analisis eksperimental pada dataset benchmark MovieLens dan Netflix membuktikan bahwa MCF yang diusulkan Mengukur menghapus masalah sparsity dan menyediakan efektifrekomendasi dengan peringkat pengguna dan item yang lebih sedikit.

Penelitian terdahulu tentang matusita juga dilakukan oleh S. Lhermitte , dkk (2011) melakukan penelitian dengan topik A comparison of time series similarity measures for classification and change detection of

ecosystem dynamics. Riset ini mendeskripsikan tentang beberapa pendekatan untuk mempelajari dinamika ekosistem berdasarkan jarak jauh merasakan seri waktu, ada kebutuhan yang kuat untuk yang lebih komprehensif pemahaman tentang langkah-langkah kesamaan yang ada. Ini langkah-langkahnya berkisar dari Minkowksi (DMan, dan DE) dan Mahalanobis (DMah) mengukur jarak, hingga korelasi (DCC), Komponen Utama Kesamaan analisis (PCA; DPCA) dan Fourier berdasarkan (DFFT, Dξ, DFk). Ini Pemahaman secara khusus penting karena banyak dari kesamaan ini tindakan berfungsi sebagai kriteria keputusan yang mendasari dalam beberapa seri waktu pengelompokan dan mengubah teknik deteksi dan pilihan kesamaan dapat mempengaruhi klasifikasi akhir dan mengubah hasil deteksi.

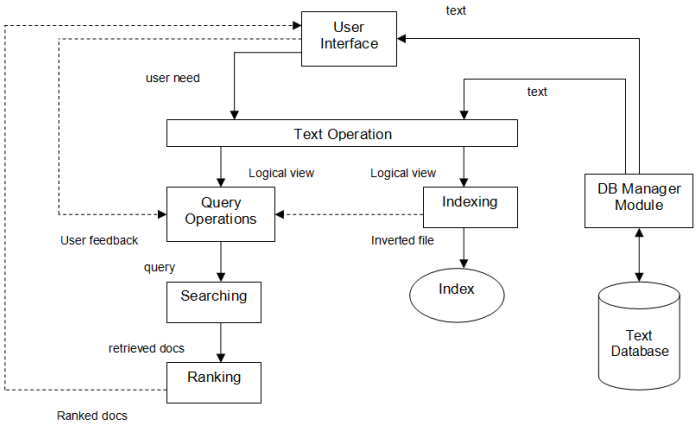
Oleh karena itu penelitian ini berfokus pada perbandingan kuantitatif Serangkaian waktu kesamaan sering digunakan mengukur D dalam fungsi yang bervariasi deret waktu dan karakteristik ekosistem, seperti amplitudo, waktu dan efek kebisingan. Perbandingan ini dengan simulasi Monte-Carlo berdasarkan himpunan bagian global MODIS Normalized Difference Vegetation indeks (NDVI) dan Indeks Vegetasi yang Ditingkatkan (EVI) dan Indeks Area Daun (LAI) data mengungkapkan empat kelompok utama pengukuran kesamaan seri waktu dengan sensitivitas yang berbeda: (i) DMan, DE, DPCA, dan DFk menghitung perbedaan antara deret waktu, (ii) DCC menilai temporal korelasi antara deret waktu, (iii) DMah juga mengkuantifikasi perbedaannya tetapi secara khusus menjelaskan korelasi temporal dan non-stasioneritas varian, iv) tindakan Fourier berdasarkan DFFT dan Dξ menghitung persamaan yang diturunkan berdasarkan komponen frekuensi tertentu.

Penelitian lain dengan topik yang sama dilakukan oleh Piotr Porwik dkk, (2010) dengan topik A New Signature Similarity Measure Based on Windows Allocation Technique. Dalam tulisan ini metode perbandingan tanda tangan telah disajikan. Metode ini didasarkan pada yang baru ukuran kesamaan, di mana tambahan partisi tanda tangan ke jendela ukuran yang sesuai adalah diusulkan. Evaluasi penyelidikan dibawa ke kesimpulan bahwa metode yang diusulkan sangat efisien dibandingkan dengan metode lain. Dicapai hasil memberikan koefisien ERR terkecil. Harus juga menekankan, bahwa dalam pendekatan ini ditandatangani dengan titik diskrit yang berbeda (dengan panjang yang berbeda) dapat dibandingkan - apa yang tidak mungkin dalam metode lain. Diusulkan dalam pendekatan makalah ini memungkinkan untuk menghilangkan metode lain di mana panjang tanda tangan harus kompensasi, seperti algoritma DTW. Metode ini secara tidak perlu memperpanjang waktu komputasi. Dalam investigasi berikutnya tahapan seleksi dinamis parameter dan pemilihan fitur tanda tangan adalah direncanakan, di mana langkah-langkah tambahan dan kesamaan Koefisien juga akan diuji.

Pembuatan mesin pencari berbahasa jawa ini digunakan untuk memudahkan pencarian teks dokumen berbahasa jawa. Kumpulan kosa kata bahasa jawa akan ditempatkan dalam sebuah korpus untuk kemudian diolah menggunakan metode Matusita. Beberapa mesin pencari sebelumnya tentang obyek bahasa jawa pernah dilakukan, dan ini sebagai bahan penelitian unjuk tingkat akurasi metode matusita dibandingkan dengan implementasi metode pencarian sebelumnya.

3. METODE PENELITIAN

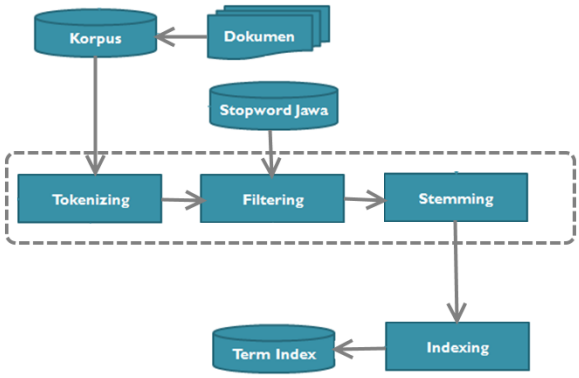
Mesin pencari dokumen teks bahasa jawa dibuat menggunakan metode Matusita. Arsitektur informasi dibuat sederahan dan mudah dipahami serta di implementasikan. Aplikasi semua proses dilakukan secara bertahap untuk didapatkannya hasil yang optimal. Proses Information Retrieval System (IRS) seperti pada gambar 1 menggunakan arsitektur yang sederhana. Sebelum dilakukannya proses temu kembali diperlukan pendefinisian database. Selanjutnya mengikuti tahapan proses; Dokumen-dokumen yang akan digunakan, Operasi yang akan digunakan dalam pencarian, dan model pengolahan teks (Baeza, 1999,h.9).



Gambar 1. The Process of Retrieving Information (Baeza, 1999,h.10)

3.1. Arsitektur Informasi

Arsitektur mesin pencari metode matusita dilakukan dengan pengumpulan dokumen teks dokumen berbahasa Jawa dalam sebuah database atau yang dikenal dengan korpus. Selanjutnya korpus akan diproses melalui beberapa tahapan proses untuk mendapatkan kata dasar dari kata yang dimasukkan. Kode program yang telah dibuat akan membuat kata bentukan berubah menjadi kata dasar. Gambar 2 menunjukkan arsitektur mesin pencari.



Gambar 2. Arsitektur Informasi mesin pencari

3.2. Metode Matusita

Formula Metode Matusita diperlihatkan dalam bentuk notasi himpunan yang dapat digunakan rumus (1):

$$= \sqrt{2 - 2\sum_{i=1}^{d} \sqrt{P_i Q_i}} \qquad (1)$$

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.

4. HASIL DAN PEMBAHASAN

4.1. Tampilan Mesin Pencari Metode Matusita

Tampilan mesin pencari dibuat seperti tampilan mesin pencari pada umumnya yaitu sederhanan dan pengguna mudah menggunakan. Kotak button dengan label cari digunakan untuk memproses setelah query di input. Tombol button cari jika sudah diklik akan menampilkan dokumen hasil pencarian. Gambar rancangan tampilan dapat dilihat pada gambar 3.

Gambar 3. Arsitektur Informasi mesin pencari

4.2. Hasil Pencarian

Hasil pencarian informasi dokumen teks bahasa jawa menggunakan mesin pencari bahasa jawa metode matusita didapatkan bahwa data yang terambil sedikit dan data yang relevan lebih banyak. Artinya data relevan yang didapatkan lebih banyak dengan didasarkan pada kuantitas dokumen yang terambil untuk kata kunci tertentu dihubungkan dengan relevansi pencarian dengan hasil pencarian. Metode matusita memungkinkan hasil yang didapatkan atau hasil yang terambil efektif sesuai dengan formulasi matusita bahwa akan didapatkan hasil dokumen sedikit dan akurat.

Pada percobaan mesin pencari digunakan atau dimasukkan beberapa term bahasa jawa dengan kategori term 1 suku kata, 2 suku kata, 3 suku kata, 4 suku kata, dan 5 suku kata dengan begitu informasi yang didapatkan bisa dibandingkan dan dianalisis tingkat efektifitas mesin pencari yang di buat. Berikut ini tabel hasil pencarian (tabel 1) menggunakan mesin pencari matusita dengan kata kunci; 1 *term* yaitu *"sabar", "becik",* 2 *term* yaitu "*seneng ngalah*" , "seneng sabar" 3 *term* "*wong angel sukses*". " wong seneng ngalah" 4 *term* "*wong sukses seneng ngalah*". 5 *term* "*wong sing sukses seneng ngalah*".

Tabel 1. Hasil Pencarian Term

No	Term (kata kunci)	Terabil	Relevan	Tidak Relevan
1	sabar	9	3	6
2	becik	11	2	9
3	Seneng nesu	7	1	6
4	Wong Sabar	17	12	5
5	wong angel sukses	5	1	4
6	Wong Seneng Ngalah	7	3	4
7	Wong sukses seneng ngalah	8	4	4
8	Wong sukses sing seneng ngalah	12	9	3

Hasil pencarian dikatakan relevan secara subyektif dengan dasar pada deklarasi arti kata untuk menentukan relevan atau tidaknya suatu term yang dicari dengan hasil yang didapatkan. Hasil yang relevan setidaknya terma hasil pencarian memiliki arti atau makna yang sama dengan deklarasi arti kata. Sedangkan term tidak relevan karena kata yang muncul tidak memiliki arti yang sama atau kata yang dicari tidak tampil dalam hasil pencarian. Tabel 2 menunjukkan tabel deklarasi persepsi term beserta artinya.

Tabel 2. Tabel Deklarasi term

Jawa	Indonesia	DEKLARASI (sumber KKBI, https://kbbi.web.id)
sabar	Sabar	sabar/sa·bar/ a 1 **tahan menghadapi cobaan** (tidak lekas marah, tidak lekas putus asa, tidak lekas patah hati); tabah: ia menerima nasibnya dengan --; hidup ini dihadapinya dengan --; 2 **tenang; tidak tergesa-gesa; tidak terburu nafsu**: segala usahanya dijalankannya dengan --;
becik	Baik	baik /ba·ik / 1 a **elok; patut; teratur (apik, rapi, tidak ada celanya, dan sebagainya)**
Seneng	Bahagia	bahagia/ba·ha·gia/ 1 n **keadaan atau perasaan senang dan tenteram (bebas dari segala yang menyusahkan)**: -- dunia akhirat; hidup penuh --; 2 a **beruntung; berbahagia**
Ngalah	Mengalah	mengalah/me·nga·lah/ v **mengaku kalah; dengan sengaja kalah (menyerah); tidak mempertahankan pendapat (tuntutan dan sebagainya);**
Wong	Orang	orang n 1 manusia (dalam arti khusus); 2 manusia (ganti diri ketiga yang tidak tentu): jangan lekas percaya pada mulut --; 3 dirinya sendiri; manusianya sendiri: saya tidak bertemu dengan -- nya; 4 kata penggolong untuk manusia: lima -- nelayan; 5 anak buah (bawahan): mereka itu -- nya Pak Camat; 6 rakyat (dari suatu negara); warga negara: -- Pakistan; 7 manusia yang berasal dari atau tinggal di suatu daerah (desa, kota, negara, dan sebagainya): dia -- Bogor; suaminya -- Eropa; 8 suku bangsa; 9 manusia lain; bukan diri sendiri; bukan kaum (golongan, kerabat) sendiri: jangankan anak sendiri, anak -- pun saya tolong; negeri -- , negeri lain (bukan negeri kita); 10 cak karena (sebenarnya):
Angel	Sulit	sulit/su·lit/ a 1 sukar sekali; susah (diselesaikan, dikerjakan, dan sebagainya): pekerjaan yang -- diselesaikan; rasanya -- baginya untuk memberitahukan hal itu kepadamu; 2 susah dicari; jarang terdapat: obat semacam itu -- didapat; 3 dirahasiakan (sukar diketahui dan sebagainya); tersembunyi: tempat -- pun ia tahu; ia dapat mengetahui hal yang --; 4 gelap (rahasia, tidak terang-terangan): apa yang mereka lakukan itu merupakan perbuatan yang --; 5 dalam keadaan yang sukar (genting, gawat, dan sebagainya):
Sukses	Sukses	sukses/suk·ses/ /suksés/ a berhasil; beruntung:

4.3. Hasil Uji Recall dan Precission

Studi hasil pencarian dengan perhitungan uji recall dan precission dilakukan pada satu keyword dengan 5 suku kata menghasilkan beberapa informasi untuk analisis. Hasil pencarian dokumen dengan keyword "wong sukses sing seneng ngalah", menunjukkan dokumen dengan bobot tertinggi adalah dokumen letak dokumen CAN54 (bobot -13.5). Dokumen CAN54 (dokumen teks yang berasal dari kategori palintangan CANCER) memiliki bobot tertinggi atau memiliki tingkat kemiripan tertinggi dibandingkan dengan dokumen lain yang ada pada korpus.

Hasil perhitungan Recall untuk keyword "*wong sukses sing seneng ngalah*" adalah sebagai berikut;

$$Recall\ (R) = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ relevant\ items\ in\ collection}$$

$$Recall = \frac{9}{253} = 0.04$$

Hasil perhitungan Precision untuk keyword "*wong sukses sing seneng ngalah*" adalah sebagai berikut;

$$Precision\ (P) = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ \ items\ retrieved}$$

$$Precision = \frac{9}{12} = 0.75$$

Hasil perhitungan rata-rata untuk Recall dan precision adalah sebagai berikut;

$$Rata - rata\ Recall = \frac{0.14}{8} = 0.02$$

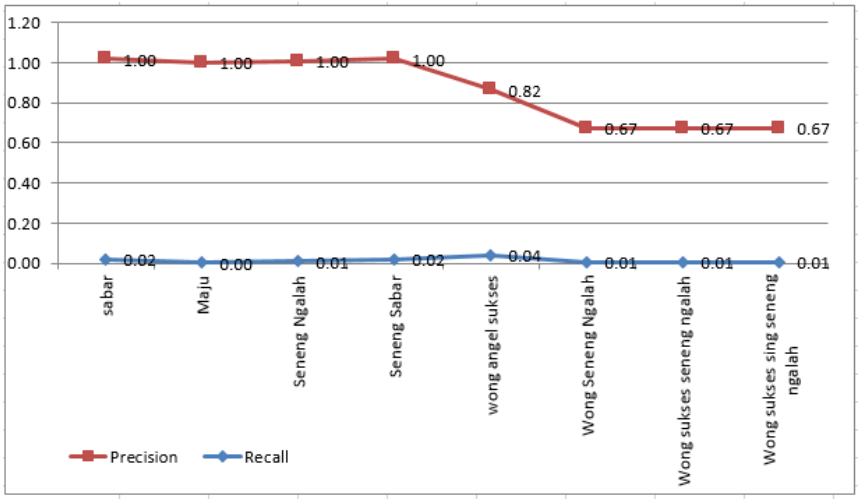
$$Rata - rata\ Precision = \frac{3.24}{8} = 0.41$$

Hasil pencarian informasi dokumen teks bahasa jawa menggunakan metode matusita setelah dilakukan uji recall dan precission didapatkan hasil seperti terlihat pada tabel 3.

Tabel 3. Hasil Perhitungan uji Recall dan Precission

No	Query	Recall	Precision
1	sabar	0.01	0.33
2	Becik	0.01	0.18
3	Seneng Ngalah	0.00	0.14
4	Seneng Sabar	0.05	0.71
5	wong angel sukses	0.00	0.20
6	Wong Seneng Ngalah	0.01	0.43
7	Wong sukses seneng ngalah	0.02	0.50
8	Wong sukses sing seneng ngalah	0.04	0.75

Gambar 4 menunjukkan hasil uji perhitungan recall dan precission dalam bentukdiagram garis.



Gambar 4. Diagram Hasil Uji recall dan Precission

5. KESIMPULAN

Mesin pencari dengan metode Matusita menghasilkan pencarian dokumen teks dengan tingkat akurasi atau *precision* = 0,41. Hasil Uji mesin pencari dengan uji *recall* dan uji *precision* menunjukkan hasil pencarian dokumen teks memiliki rata-rata *recall* = 0,02 dan rata-rata *precision* = 0,41.

DAFTAR PUSTAKA

[1] C Selvi, dkk. 2018. A novel similarity measure towards effective recommendation using Matusita coefficient for Collaborative Filtering in a sparse dataset. Sadhana. Vol V. 43:202

[2] Kadir, A., 2001. Dasar Pemrograman Web Dinamis menggunakan PHP. Penerbit Andi. Yogyakarta.

[3] Manning, C., Raghavan, P., 2007. An Introduction to Information Retrieval, Stanford. USA.

[4] Meadow, C.T., 1997. Text Information Retrieval Systems. Academic Press.New York.

[5] Piotr Porwik, dkk. 2010. A New Signature Similarity Measure Based on Windows Allocation Technique. Applications (IJCISIM). ISSN: 2150-7988 Vol.2 (2010), pp.297-305

[6] 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.

[7] Salton, G., 1989, Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer. Addison – Wesly Publishing Company, Inc. USA.

[8] S. Lhermitte, dkk. 2011. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. Elsevier, 2011, Remote Sensing of Environment 115 (2011) 3129–3152