



PERBANDINGAN METODE COSINE SIMILARITY DAN JACCARD SIMILARITY PADA SISTEM TEMU KEMBALI INFORMASI Pencarian Arsip Dokumen

Melda, Sucipto, Menur Wahyu Pangestika

Teknik Informatika, Universitas Muhammadiyah Pontianak

Abstrak

Pengelolaan arsip dokumen pada instansi pemerintahan memegang peranan penting dalam mendukung pelayanan administrasi yang efektif. Namun, proses pencarian arsip yang masih dilakukan secara manual di Kantor Kecamatan Pontianak Selatan menyebabkan pencarian dokumen membutuhkan waktu yang lama dan berisiko menimbulkan kesalahan. Penelitian ini bertujuan merancang sistem temu kembali informasi berbasis web serta membandingkan kinerja metode Cosine Similarity dan Jaccard Similarity dalam pencarian arsip dokumen. Metode penelitian meliputi pengumpulan data melalui observasi, wawancara, dan studi literatur, dilanjutkan dengan tahapan preprocessing teks, pembobotan TF-IDF, implementasi algoritma, serta pengujian menggunakan confusion matrix dan User Acceptance Testing (UAT). Hasil penelitian menunjukkan bahwa metode Cosine Similarity memiliki performa yang lebih baik dibandingkan Jaccard Similarity dengan nilai accuracy 72,6%, precision 100,0%, recall 72,6%, F1-score 84,1%, dan waktu pemrosesan 0,091 detik. Sementara itu, Jaccard Similarity memperoleh accuracy 26,4%, precision 100,0%, recall 26,4%, F1-score 41,8%, dan waktu pemrosesan 0,709 detik. Dengan demikian, Cosine Similarity lebih efektif digunakan pada sistem pencarian arsip dokumen karena mampu menghasilkan tingkat relevansi dan efisiensi pencarian yang lebih baik.

Kata Kunci: Temu Kembali Informasi, Arsip Dokumen, TF-IDF, Cosine Similarity, Jaccard Similarity.

PENDAHULUAN

Arsip dokumen merupakan sumber informasi penting yang digunakan sebagai bukti administrasi

dan referensi dalam suatu instansi. Pengelolaan arsip yang baik diperlukan untuk mendukung pelayanan, menjaga keakuratan data, dan memudahkan pencarian dokumen saat dibutuhkan[1].

Di Kantor Kecamatan Pontianak Selatan mengelola rata-rata 57 surat per bulan (32 surat masuk dan 25 surat keluar). Namun, pengelolaan dan pencarian arsip yang masih manual memicu rendahnya efisiensi pelayanan, risiko kerusakan dokumen, serta lambatnya penemuan data relevan yang menghambat kinerja pelayanan publik.

Penerapan sistem temu kembali informasi berbasis web dapat mempermudah proses digitalisasi pencatatan, penyimpanan, dan pencarian data secara terintegrasi. Penelitian ini menerapkan algoritma Cosine Similarity berbasis pembobotan Term Frequency-Inverse Document Frequency (TF-IDF) dan Jaccard Similarity sebagai metode perbandingan untuk menelusuri data teks berdasarkan kata kunci.

Penelitian ini membandingkan metode Cosine Similarity dan Jaccard Similarity untuk mengukur tingkat kemiripan antara kata kunci pencarian dan dokumen arsip. Cosine Similarity menghitung kemiripan berdasarkan representasi vektor dokumen, sedangkan Jaccard Similarity menghitung kemiripan berdasarkan kesamaan term yang terdapat pada dokumen.

Perbandingan penelitian ini bertujuan untuk merancang sistem temu kembali informasi berbasis web pencarian arsip dokumen pada Kantor Kecamatan Pontianak Selatan serta membandingkan kinerja metode Cosine Similarity dan Jaccard Similarity menggunakan metrik accuracy, precision, recall, F1-score, dan waktu pemrosesan.

METODE PENELITIAN

3.1 Tahapan Penelitian

Tahapan penelitian dimulai dari pengumpulan data arsip surat masuk dan surat keluar Kantor Kecamatan Pontianak Selatan melalui observasi, wawancara, dan studi literatur. Data

yang diperoleh kemudian diproses melalui tahapan preprocessing untuk mempersiapkan dokumen sebelum dilakukan pembobotan kata menggunakan TF-IDF. Selanjutnya, metode Cosine Similarity dan Weighted Jaccard Similarity diimplementasikan untuk menghitung tingkat kemiripan antara query dan dokumen arsip. Hasil pencarian kemudian dievaluasi menggunakan confusion matrix dan User Acceptance Testing (UAT).

Gambar 1 menunjukkan alur penelitian yang diawali dengan identifikasi masalah dan pengumpulan data. Selanjutnya dilakukan *preprocessing*, pembobotan TF-IDF, serta perhitungan kemiripan menggunakan metode *Cosine Similarity* dan *Weighted Jaccard Similarity*. Setelah itu, sistem diimplementasikan dan dilakukan pengujian menggunakan *confusion matrix* dan *User Acceptance Testing* (UAT). Tahap akhir penelitian adalah analisis hasil pengujian dan penarikan kesimpulan.



Gambar 1. Alir Penelitian**3.2 Preprocessing Dokumen**

Tahap *preprocessing* bertujuan untuk mengolah teks dokumen agar lebih terstruktur dan siap digunakan dalam proses perhitungan kemiripan. Tahapan preprocessing yang digunakan meliputi:

1. *Case Folding*, yaitu mengubah seluruh huruf menjadi huruf kecil.
2. *Tokenizing*, yaitu memecah kalimat menjadi kumpulan kata.
3. *Filtering*, yaitu menghapus kata yang tidak memiliki makna penting.
4. *Stemming*, yaitu mengubah kata berimbuhan menjadi kata dasar.

Hasil *preprocessing* digunakan sebagai dasar dalam proses pembobotan TF-IDF.

3.3 Pembobotan TF-IDF

Setelah preprocessing, setiap term diberikan bobot menggunakan metode TF-IDF. Metode ini mengukur tingkat kepentingan suatu kata berdasarkan frekuensi kemunculannya dalam dokumen dan keseluruhan koleksi dokumen. Perhitungan TF-IDF menggunakan persamaan (1) dengan contoh berikut:

Contoh:

Query "bantuan dana pemerintah"

Dokumen Arsip

- D1 "surat permohonan bantuan dana Pendidikan"
- D2 "laporan penggunaan dana bantuan pemerintah"
- D3 "surat pemberitahuan kegiatan desa"

Tabel 2. Hasil Perhitungan TF-IDF

Dokumen	TF-IDF
D1	0,070
D2	0,165
D3	0,000

Tabel 2 menunjukkan hasil perhitungan TF-IDF dengan menggunakan query "bantuan dana pemerintah" pada tiga dokumen.

3.4 Perhitungan Cosine Similarity

Setelah dilakukan pembobotan TF-IDF, tingkat kemiripan antara query dan dokumen dihitung menggunakan metode *Cosine Similarity*. Metode ini mengukur kesamaan berdasarkan sudut kosinus antara vektor query dan vektor dokumen. Semakin besar nilai yang dihasilkan, maka semakin tinggi tingkat relevansi dokumen terhadap query.

Perhitungan *Cosine Similarity* menggunakan vektor hasil pembobotan TF-IDF dengan persamaan (2) sebagai berikut:

Query "bantuan dana pemerintah"

- D1 "surat permohonan bantuan dana Pendidikan"
- D2 "laporan penggunaan dana bantuan pemerintah"
- D3 "surat pemberitahuan kegiatan desa"

Tabel 3. Hasil Perhitungan Cosine Similarity

Dokumen	Nilai Cosine Similarity
D1	0,47
D2	0,99
D3	0,00

Tabel 3 menunjukkan hasil perhitungan *Cosine Similarity* pada tiga dokumen dengan query "bantuan dana pemerintah". Hasil perhitungan berupa nilai kemiripan pada rentang 0 sampai 1.

Dokumen kemudian diurutkan berdasarkan nilai kemiripan tertinggi sehingga dokumen yang paling relevan ditampilkan pada urutan pertama hasil pencarian.

3.5 Perhitungan Jaccard Similarity

Jaccard Similarity digunakan untuk mengukur tingkat kemiripan antara query dan dokumen berdasarkan kesamaan *term* yang dimiliki keduanya. Dalam penelitian ini digunakan *Weighted Jaccard Similarity* yang memanfaatkan bobot TF-IDF untuk meningkatkan representasi tingkat kepentingan setiap *term*. Perhitungan *Weighted Jaccard Similarity* menggunakan persamaan (4) sebagai berikut:

Setelah dilakukan *preprocessing*, diperoleh *term* sebagai berikut:

- Query (Q) = {bantuan, dana, pemerintah}

Dokumen yang digunakan adalah:

- Dokumen 1 (D₁) = {surat, permohonan, bantuan, dana, pendidikan}
- Dokumen 2 (D₂) = {laporan, penggunaan, dana, bantuan, pemerintah}
- Dokumen 3 (D₃) = {surat, pemberitahuan, kegiatan, desa}

Tabel 4 Hasil Perhitungan Weighted Jaccard Similarity

Dokumen	Nilai <i>Weighted Jaccard Similarity</i>
D1	0,26
D2	0,61
D3	0,00

Tabel 4 menunjukkan hasil perhitungan pada tiga dokumen dengan query {bantuan, dana, pemerintah} Nilai kemiripan berada pada rentang 0 sampai 1, di mana nilai yang mendekati 1 menunjukkan tingkat kemiripan yang semakin tinggi. Hasil perhitungan

kemudian digunakan untuk menentukan peringkat dokumen berdasarkan tingkat relevansinya terhadap query pencarian.

3.6 Pengujian Sistem

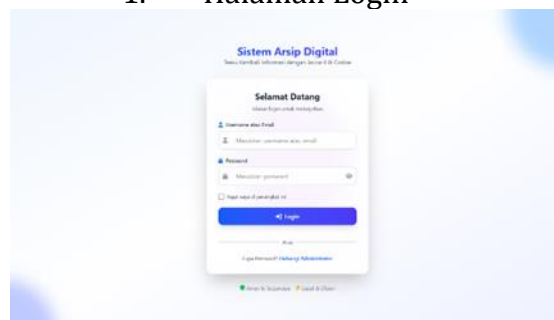
Pengujian sistem dilakukan untuk mengetahui tingkat kinerja sistem pencarian arsip dokumen. Evaluasi metode dilakukan menggunakan confusion matrix dengan parameter *accuracy*, *precision*, *recall*, dan *F1-score*. Selain itu, dilakukan pengukuran waktu pemrosesan untuk membandingkan efisiensi kedua metode.

Pengujian penerimaan pengguna dilakukan menggunakan *User Acceptance Testing* (UAT) kepada pegawai Kantor Kecamatan Pontianak Selatan untuk mengetahui tingkat kemudahan penggunaan dan kesesuaian sistem dengan kebutuhan pengguna.

7. Hasil Implementasi Sistem

Sistem temu kembali informasi arsip dokumen berhasil diimplementasikan berbasis web menggunakan *Laravel* dan *MySQL*. Sistem menyediakan fitur login, pengelolaan arsip surat masuk dan surat keluar, serta fitur pencarian dokumen menggunakan metode *Cosine Similarity* dan *Jaccard Similarity*. Proses pencarian dilakukan melalui tahapan *preprocessing*, pembobotan TF-IDF, dan perhitungan tingkat kemiripan dokumen terhadap query yang dimasukkan pengguna.

1. Halaman Login



Gambar 2. Login

Gambar 2 berfungsi sebagai autentikasi awal untuk memastikan hanya pengguna berwenang yang dapat mengakses sistem pengelolaan arsip. Pengguna diwajibkan memasukkan kredensial berupa *username* atau email dan kata sandi yang valid. Setelah verifikasi berhasil, sistem akan mengarahkan pengguna ke halaman *dashboard* sesuai dengan hak akses masing-masing guna menjaga keamanan dan kendali data.

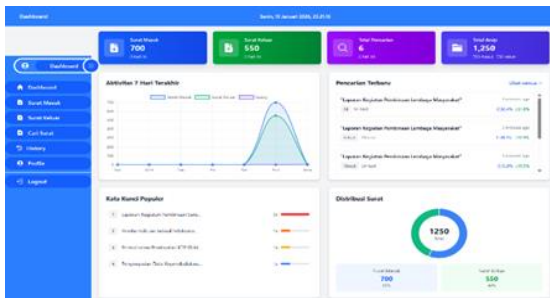
2. Halaman Beranda



Gambar 3. Beranda

Gambar 3 merupakan tampilan utama setelah login yang menyajikan ringkasan informasi data arsip, seperti jumlah surat masuk dan surat keluar. Antarmukanya dirancang sederhana dan informatif guna memudahkan pengguna menavigasi menu lain secara cepat.

3. Halaman Dashboard



Gambar 4. Dashboard

Gambar 4 merupakan tampilan *dashboard* yang menyajikan ringkasan data penting, seperti jumlah surat masuk, surat keluar, dan aktivitas terbaru dalam sistem pengelolaan arsip. Dirancang secara sederhana dan informatif, halaman ini berfungsi memudahkan

pemantauan aktivitas sekaligus mempercepat akses ke berbagai fitur utama guna meningkatkan efisiensi pengelolaan data.

8. Hasil Pengujian Sistem

Pengujian sistem dilakukan menggunakan *confusion matrix* untuk mengukur nilai *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil pengujian menunjukkan bahwa metode *Cosine Similarity* memperoleh *accuracy* 72.6%, *precision* 100.0%, *recall* 72.6%, dan *F1-score* 84.1%. Sementara itu, metode *Jaccard Similarity* memperoleh *accuracy* 26.4%, *precision* 100.0%, *recall* 26.4%, dan *F1-score* 41.8%.

Selain pengujian kinerja sistem, dilakukan pula *User Acceptance Testing* (UAT) terhadap 25 responden.

Tabel 5. Hasil Pengujian UAT

Kategori	Jumlah	Bobot	Skor
TB	0	1	0
KB	8	2	16
CB	43	3	129
B	91	4	364
SB	111	5	555
Jumlah	253		1064

Tabel 5 menunjukkan hasil pengujian sistem memperoleh skor 1064 dari nilai maksimum 1250 dengan tingkat kelayakan sebesar 85.12%. Berdasarkan klasifikasi *Skala Likert*, nilai tersebut termasuk dalam kategori "Sangat Layak", sehingga sistem dapat diterima dengan baik oleh pengguna.

9. Hasil Perbandingan Metode *Cosine Similarity* dan *Jaccard Similarity*

Tabel 6. Hasil Perbandingan

Matrix Evaluasi	<i>Cosine Similarity</i>	<i>Weighted Jaccard Similarity</i>
-----------------	--------------------------	------------------------------------

Precision	100%	100%
Accuracy	72.6%	26.4%
Recall	72.6%	26.4%
F1-Score	84.1%	41.8%
Waktu	0.091 detik	0.709 detik

Tabel 6 menunjukkan pengujian sistem temu kembali arsip menggunakan query "surat bantuan dana pemerintah" menunjukkan bahwa *Cosine Similarity* dan *Jaccard Similarity* sama-sama menghasilkan *precision* 100%. Namun, *Cosine Similarity* lebih unggul dengan nilai *recall* (72.6%), *F1-score* (84.1%), akurasi (72.6%), serta waktu pemrosesan yang lebih cepat (0.091 detik).

7. Pembahasan

Berdasarkan hasil pengujian, metode *Cosine Similarity* menunjukkan performa yang lebih baik dibandingkan *Weighted Jaccard Similarity*. Kedua metode menghasilkan nilai *precision* yang sama, yaitu 100%, yang menunjukkan bahwa dokumen yang ditampilkan relevan dengan kebutuhan pengguna. Namun, *Cosine Similarity* menghasilkan nilai *recall* sebesar 72.6%, *accuracy* sebesar 72.6%, dan *F1-score* sebesar 84.1%, sehingga mampu menemukan lebih banyak dokumen relevan, sedangkan *Weighted Jaccard Similarity* menghasilkan nilai *recall* sebesar 26.4%, *accuracy* sebesar 26.4%, dan *F1-score* sebesar 41.8%. Selain itu, waktu pemrosesan *Cosine Similarity* lebih cepat, yaitu 0.091 detik, sedangkan *Weighted Jaccard Similarity* membutuhkan waktu 0.709 detik. Hasil tersebut dipengaruhi oleh penggunaan pembobotan TF-IDF pada *Cosine Similarity*, sehingga kata-kata yang lebih penting dalam dokumen memiliki bobot yang lebih besar dan menghasilkan pencarian yang lebih relevan dengan

kata kunci yang dimasukkan pengguna. Sebaliknya, *Weighted Jaccard Similarity* membandingkan bobot kata antara query dan dokumen sehingga lebih sensitif terhadap perbedaan kosakata. Hal ini menyebabkan nilai kemiripan yang dihasilkan oleh *Weighted Jaccard Similarity* cenderung lebih rendah ketika query dan dokumen menggunakan pilihan kata yang berbeda meskipun memiliki makna yang sama. Oleh karena itu, *Cosine Similarity* lebih efektif dan efisien untuk diterapkan pada sistem pencarian arsip dokumen karena mampu menghasilkan tingkat relevansi pencarian yang lebih baik serta waktu respons yang lebih cepat.

SIMPULAN

Penelitian ini berhasil mengembangkan sistem temu kembali informasi arsip dokumen berbasis web menggunakan metode *Cosine Similarity* dan *Jaccard Similarity* pada Kantor Kecamatan Pontianak Selatan. Hasil pengujian menunjukkan bahwa metode *Cosine Similarity* memiliki kinerja yang lebih baik dibandingkan *Jaccard Similarity* dengan nilai *accuracy* sebesar 72.6%, *precision* sebesar 100%, *recall* sebesar 72.6%, dan *F1-score* sebesar 84.1%, serta waktu pemrosesan yang lebih cepat, yaitu 0.091 detik. Sedangkan, *Jaccard Similarity* memperoleh nilai *accuracy* sebesar 26.4%, *precision* sebesar 100%, *recall* sebesar 26.4%, *F1-score* sebesar 41.8%, dan waktu pemrosesan sebesar 0.709 detik. Selain itu, hasil *User Acceptance Testing* (UAT) memperoleh nilai kelayakan sebesar 85.12% dengan kategori "Sangat Layak", sehingga sistem dinilai mampu membantu proses pencarian arsip dokumen secara efektif sesuai dengan kebutuhan pengguna.

Untuk pengembangan selanjutnya, penelitian dapat menggunakan jumlah dataset yang lebih besar dan lebih beragam agar hasil

evaluasi metode menjadi lebih komprehensif. Selain itu, penelitian dapat membandingkan metode temu kembali informasi lainnya serta menambahkan fitur pencarian semantik, rekomendasi dokumen, atau pencarian berbasis sinonim guna meningkatkan relevansi hasil pencarian dan memudahkan pengguna dalam menemukan arsip yang dibutuhkan.

DAFTAR PUSTAKA

- [1] Y. A. Hapsari, P. Suharso, and) Korespondensi, "Analisis Pengelolaan Arsip Dinamis di Kantor Kelurahan Pojoksari Kecamatan Ambarawa," *ANUVA*, vol. 5, no. 4, pp. 555-568, 2021.
- [2] S. Purwaningrum, A. Susanto, and A. Kristiningsih, "Pengaruh Synonym Recognition dalam Deteksi Kemiripan Teks Menggunakan Winnowing dan Cosine Similarity," 2023.
- [3] M. Farras Majid and A. Solichin, "Aplikasi Pendeteksi Kalimat Kasar Bahasa Indonesia Pada File Audio Menggunakan Jaccard Similarity Dan N-Gram," *Jurnal TICOM: Technology of Information and Communication*, vol. 12, no. 1, 2023.
- [4] Ibrahim Farhan, "Perbandingan Metode Cosine Dan Jaccard Similarity Pada Sistem Rekomendasi Ayat Al-Qur'an Terhadap Curahan Hati Pengguna."
- [5] M. A. Ma'ruf and A. Qoiriah, "Perbandingan Algoritma Cosine Similarity dan Euclidean Distance pada Sistem Rekomendasi Film dengan Metode Item-Based Collaborative Filtering," *Journal of Informatics and Computer Science*, vol. 04, 2022.
- [6] S. Purwaningrum, A. Susanto, N. Wachid, and A. Prasetya, "Perbandingan Dice Similarity dan Jaccard Coefficiencie Terhadap Algoritma Winnowing untuk Deteksi Kesamaan Dokumen Teks Bahasa Indonesia," 2021.
- [7] D. Iskandar and A. Kurniawati, "Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 1, pp. 133-144, Feb. 2025, doi: 10.25126/jtiik.2025129143.
- [8] R. Putra and Z. Mustika, "Similarity Judul Tugas Akhir Menggunakan Metoda Cosine, Jaccard, Rabin-Karp Pada Jurusan TI," *JITSI: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 5, 2024, doi: 10.62527/jitsi.5.4.291.
- [9] N. Setiawan and F. Amin, "Sistem Temu Kembali Informasi Jurnal Ilmiah Unisbank Dengan Metode Cosine Similarity," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 6, no. 3, pp. 450-460, Jul. 2024, doi: 10.47233/jteksis.v6i3.1391.
- [10] R. Rismayani, H. SY, T. Darwansyah, and I. Mansyur, "Implementasi Algoritma Text Mining dan Cosine Similarity untuk Desain Sistem Aspirasi Publik Berbasis Mobile," *Komputika : Jurnal Sistem Komputer*, vol. 11, no. 2, pp. 169-176, Aug. 2022, doi: 10.34010/komputika.v11i2.6501.
- [11] Selfi Fitria Sari and Riyan Sisiawan Putra, "LITERATURE REVIEW SISTEM PENGELOLAAN ARSIP DIKANTOR KELURAHAN KEBOLE DAN KECAMATAN WANASARI KABUPATEN BREBES," 2022.
- [12] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks," *Sintesia*, vol. 1, pp. 81-88, 2022.
- [13] F. Teknik, "PENERAPAN TEKS MINING DAN COSINE SIMILARITY UNTUK MENENTUKAN KESAMAAN DOKUMEN SKRIPSI APPLICATION OF TEXT MINING AND COSINE SIMILARITY TO DETERMINE THE SIMILARITY OF THESIS DOCUMENTS," 2024.
- [14] I. M. Siregar, D. Pratama, and C. Himawan, "SINTECH Journal | 101 Penggunaan Jaccard Similarity Coefficient dalam Optimasi Proses Rekrutmen Karyawan Berbasis Profil dan Kompetensi," 2024, [Online]. Available: <https://doi.org/10.31598>
- [15] Joko Handoyo, Aris Rakhmadi, and Tri Rochmadi, "Kerangka Kerja Penalaran Berbasis Kasus yang Ditingkatkan dengan Weighted Jaccard Similarity untuk Diagnosis Malnutrisi pada Balita," Jan. 2026.