



PENERAPAN ALGORITMA TF-IDF DAN COSINE SIMILARITY UNTUK PENCARIAN ARSIP DOKUMEN DI KANTOR DESA KELAKAR

Ria Maisya Syarifah, Menur Wahyu Pangestika, Sucipto

Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer,

Universitas Muhammadiyah Pontianak

Abstrak

Pengelolaan arsip dokumen di Kantor Desa Kelakar hingga saat ini masih dilakukan tanpa sistem digital, sehingga menyebabkan kesulitan dalam pencarian dokumen, risiko kehilangan arsip, serta pemborosan waktu dan tenaga. Seiring meningkatnya kebutuhan akan pengelolaan arsip yang cepat dan akurat di Kantor Desa Kelakar, diperlukan sebuah sistem pencarian arsip digital yang mampu menampilkan dokumen secara relevan berdasarkan kata kunci pencarian. Penelitian ini menggunakan data berupa arsip dokumen administrasi desa dalam bentuk file PDF dan DOCX, termasuk dokumen hasil pemindaian yang diekstraksi menggunakan teknologi Optical Character Recognition (OCR). Tujuan penelitian ini adalah menerapkan algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan Cosine Similarity dalam sistem pencarian arsip dokumen berbasis web di Kantor Desa Kelakar. Hasil penelitian menunjukkan bahwa sistem mampu menampilkan dokumen yang relevan sesuai dengan kata kunci pencarian serta mempercepat proses temu kembali arsip. Hasil pengujian akurasi sistem menunjukkan nilai precision sebesar 0,001, recall sebesar 1,0, dan F1-score sebesar 0,0020, yang menunjukkan bahwa sistem mampu menampilkan dokumen yang relevan sesuai dengan kata kunci yang dimasukkan oleh pengguna. Selain itu, hasil white box testing menunjukkan bahwa seluruh fungsi dan alur logika program berjalan sesuai dengan rancangan yang telah ditetapkan tanpa ditemukan kesalahan fungsional yang signifikan. Dengan demikian, penerapan algoritma TF-IDF dan Cosine Similarity pada sistem pencarian arsip dokumen ini terbukti efektif dalam meningkatkan efisiensi dan kualitas pengelolaan arsip di Kantor Desa Kelakar.

Kata Kunci: Arsip Dokumen, TF-IDF, Cosine Similarity, OCR, Sistem Pencarian Arsip.

PENDAHULUAN

Dalam instansi atau organisasi, arsip berfungsi sebagai pendukung proses pengambilan keputusan, menunjang proses perencanaan, mendukung pengawasan, sebagai alat pembuktian, sebagai memori organisasi, dan dapat digunakan untuk kepentingan publik dan ekonomi. Arsip juga merupakan salah satu aset organisasi yang paling berharga. Oleh karena itu, dokumen harus ditata dengan baik untuk membangun manajemen organisasi yang efektif, efisien dan profesional untuk perbaikan organisasi. Tentunya hal ini harus dilakukan sesuai dengan prosedur pengarsipan yang benar untuk menjamin agar arsip tetap utuh, baik secara fisik maupun informasi.

Pengelolaan arsip di Kantor Desa Kelakar hingga saat ini masih dilakukan secara manual. Cara ini dinilai kurang efisien karena berpotensi menimbulkan berbagai permasalahan, seperti hilangnya dokumen penting, kerusakan fisik pada arsip akibat penyimpanan yang tidak optimal, serta terbengkalainya dokumen karena tidak adanya sistem pengelolaan yang terstruktur. Selain itu, metode manual juga dapat menyebabkan pemborosan sumber daya, baik dari segi waktu, tenaga, maupun biaya operasional. Seiring berjalannya waktu, penggunaan sistem digital dalam pengarsipan terbukti dapat mempermudah proses penyimpanan dan pencarian dokumen. Namun, dengan semakin banyaknya data yang tersimpan secara elektronik, penting untuk memastikan hanya dokumen yang relevan yang disimpan agar tidak terjadi penumpukan data yang tidak berguna dan untuk menjaga kualitas informasi yang diakses. Oleh karena itu, dibutuhkan sistem pengelolaan arsip yang lebih modern dan terkomputerisasi agar pengelolaan arsip di lingkungan kantor desa menjadi lebih efisien, aman, dan terstruktur.

Penggunaan sistem berbasis web dapat mempermudah dalam pengarsipan sebuah dokumen namun dengan banyaknya data yang ada perlu di perhatikan dokumen-dokumen yang relevan dengan yang dicari sehingga dapat menemukan dokumen yang tepat dan sesuai dengan query pencarian[2]. Penerapan algoritma *TF-IDF* mempermudah dalam pencarian data dengan tingkat relevansi yang tinggi sehingga tidak salah dalam mengambil data dokumen yang di perlukan. Selain itu, *Cosine similarity* digunakan sebagai metode dalam menghitung tingkat kesamaan (*similarity*) antar dokumen. Kelebihan dari algoritma cosine similarity adalah mampu mendeteksi tingkat kemiripan, dengan akurasi yang tinggi.

Beberapa penelitian sebelumnya, menggunakan algoritma TF-IDF dan *Cosine Similarity* pada Pemrosesan Query dan Pemeringkatan Judul Berita Terkait Gubernur Jawa Barat yang menghasilkan pengujian terhadap tiga query yang berbeda. Penelitian tersebut menunjukkan bahwa sistem mampu memberikan hasil yang sangat baik dengan rata-rata akurasi sebesar 99.75%, presisi 96.67%, recall 100%, dan f1 score 98.33%. Sementara itu, algoritma Sequential Search dan Fitur Autocomplete pada Aplikasi Arsip (E-Arsip) Perpustakaan Berbasis Web menghasilkan pencarian dengan nilai ke efektifan sebesar 100% dari 250 data, serta mampu menampilkan 4 sampai 10 ide atau saran dokumen sesuai kecocokan keyword pada data.

Kombinasi dari algoritma TF-IDF dan *cosine similarity* pada penelitian ini akan diterapkan menggunakan website yang mampu menemukan dokumen yang paling relevan tanpa memerlukan pencocokan kata yang persis, meningkatkan efisiensi pengelolaan arsip, mengurangi pemborosan sumber daya, serta keteraturan dokumen digital. Dengan demikian, penerapan algoritma

ini dapat membantu dalam pengelolaan arsip manual yang masih diterapkan di Kantor Desa Kelakar.

METODE PENELITIAN

Penelitian ini menggunakan metode pengembangan sistem pencarian arsip dokumen berbasis web dengan menerapkan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Cosine Similarity*. Penelitian dilakukan di Kantor Desa Kelakar dengan memanfaatkan arsip dokumen digital berupa file PDF, DOCX, dan dokumen hasil pemindaian (*scan*).

Tahapan penelitian dimulai dari identifikasi masalah, yaitu menganalisis kendala pengelolaan arsip yang masih dilakukan secara manual. Selanjutnya dilakukan studi literatur terkait sistem temu kembali informasi, algoritma *TF-IDF*, *Cosine Similarity*, dan teknologi *Optical Character Recognition* (OCR) sebagai dasar pengembangan sistem. Data penelitian diperoleh melalui pengumpulan arsip dokumen yang digunakan sebagai dataset pencarian.

Tahap berikutnya adalah pengolahan data. Dokumen hasil pemindaian terlebih dahulu diproses menggunakan OCR (*Optical Character Recognition*) untuk mengekstraksi teks agar dapat diolah oleh sistem. Setelah itu dilakukan tahap preprocessing yang meliputi *case folding*, *tokenizing*, *filtering* (*stopword removal*), *stemming*, dan *cleaning* untuk menghasilkan data teks yang lebih terstruktur dan siap diproses.

Data hasil preprocessing kemudian diberikan bobot menggunakan metode TF-IDF untuk menentukan tingkat kepentingan setiap kata dalam dokumen. Selanjutnya dilakukan perhitungan *Cosine Similarity* untuk mengukur tingkat kemiripan antara dokumen dan query pencarian pengguna. Nilai *similarity* yang dihasilkan digunakan sebagai dasar pemeringkatan dokumen sehingga

sistem dapat menampilkan hasil pencarian yang paling relevan.

Tahap akhir adalah implementasi dan pengujian sistem. Sistem dikembangkan menggunakan *framework Flask*, database *MongoDB*, serta diimplementasikan dalam aplikasi berbasis web. Evaluasi dilakukan menggunakan metrik *precision*, *recall*, dan *F1-score* untuk mengukur akurasi hasil pencarian, serta *white box testing* untuk memastikan fungsi sistem berjalan sesuai rancangan.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil implementasi sistem pencarian arsip dokumen menggunakan algoritma *TF-IDF* dan *Cosine Similarity* pada Kantor Desa Kelakar. Evaluasi dilakukan untuk mengetahui kemampuan sistem dalam menemukan dokumen yang relevan berdasarkan kata kunci pencarian serta mengukur performa sistem melalui metrik pengujian.

Implementasi Sistem Pencarian Arsip

Tahap implementasi dimulai dengan proses unggah dokumen ke sistem. Sistem mendukung dokumen dalam format PDF, DOCX, dan hasil scan. Untuk dokumen hasil pemindaian dilakukan ekstraksi teks menggunakan *Tesseract OCR*, sedangkan dokumen digital diproses melalui ekstraksi teks secara langsung. Selanjutnya dilakukan preprocessing yang terdiri atas *case folding*, *tokenizing*, *filtering*, *stemming*, dan *cleaning* sebelum dilakukan pembobotan menggunakan TF-IDF.

Setelah seluruh dokumen diproses, sistem menghitung tingkat kemiripan menggunakan *Cosine Similarity* dan menampilkan dokumen berdasarkan skor *similarity* tertinggi.

Tabel 1. Tahapan Pengolahan Data Sistem

Tahapan	Proses	Output
Input data	Upload dokumen PDF, DOCX, Scan	Dokumen tersimpan
OCR	Ekstraksi teks dari dokumen scan	Teks digital
Preprocessing	Case folding, tokenizing, filtering, stemming, celaning	Data bersih
TF-IDF	Pembobotan kata	Vektor dokumen
Cosine Similarity	Perhitungan kemiripan	Ranking dokumen

Berdasarkan Tabel 1, sistem melakukan transformasi dokumen secara bertahap hingga menghasilkan daftar dokumen yang telah diranking berdasarkan relevansi terhadap query pencarian.

Hasil Pengujian Sistem

Pengujian akurasi sistem pencarian arsip dokumen dilakukan untuk mengetahui tingkat keberhasilan sistem dalam menemukan dokumen yang relevan berdasarkan kata kunci (*query*) yang diberikan oleh pengguna. Pengujian ini menggunakan metrik *Precision*, *Recall*, dan *F1-Score*, dengan metode pembobotan TF-IDF dan perhitungan tingkat kemiripan menggunakan *Cosine Similarity*. Dataset yang digunakan dalam pengujian terdiri dari 1.000 dokumen arsip digital yang seluruhnya berhasil diindeks oleh sistem.

Tabel 2. Hasil Pengujian Akurasi Sistem

Metrik	Nilai
Precision	0,001
Recall	1,0
F1-Score	0,0020

Berdasarkan Tabel 2, sistem memperoleh nilai recall sebesar 1,0 yang menunjukkan bahwa sistem mampu menemukan seluruh dokumen yang sesuai dengan kata kunci pencarian.

Namun nilai precision sebesar 0,001 menunjukkan bahwa masih terdapat dokumen yang kurang relevan ikut ditampilkan dalam hasil pencarian. Nilai F1-score sebesar 0,0020 menggambarkan keseimbangan antara precision dan recall yang masih dapat ditingkatkan melalui optimasi preprocessing dan pembobotan.

Analisis Penerapan TF-IDF

Pada sistem ini, pembobotan TF-IDF dilakukan terhadap seluruh dokumen arsip untuk membentuk representasi vektor dokumen. Query pengguna tidak dihitung ulang nilai IDF-nya, melainkan ditransformasikan ke dalam ruang vektor yang sama menggunakan model TF-IDF yang telah dibangun dari dokumen. Pendekatan ini bertujuan agar proses perhitungan kemiripan antara query dan dokumen dapat dilakukan secara konsisten.

Rank	Judul Dokumen	Bobot TF-IDF Dokumen (berdasarkan term query):
Rank #1	Surat permintaan pembayaran bereskan kegiatan 17 agustus 2024.pdf	agustus + Dok: 0.2578 Query: 0.8366
		bayar + Dok: 0.0964 Query: 0.3130
		minta + Dok: 0.0980 Query: 0.3180
		minta bayar + Dok: 0.0980 Query: 0.3180
Rank #2	laporan permintaan pembayaran keperluan kegiatan lomba 17 Agustus .pdf	agustus + Dok: 0.1811 Query: 0.8366
		bayar + Dok: 0.0678 Query: 0.3130
		minta + Dok: 0.0610 Query: 0.3180
		minta bayar + Dok: 0.0558 Query: 0.3180

Gambar 1. Hasil Bobot TF-IDF untuk Dokumen

Pada gambar 1 merupakan Hasil pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) pada dua dokumen yang diperoleh dari proses pencarian berdasarkan kata kunci (*query*) yang dimasukkan oleh pengguna.

Pada Rank #1, dokumen berjudul "*Surat permintaan pembayaran bereskan kegiatan 17 agustus 2024.pdf*" memiliki nilai bobot TF-IDF yang lebih tinggi dibandingkan dokumen lainnya. Hal ini terlihat dari nilai pada beberapa kata kunci seperti "*agustus*" dengan bobot 0,2578, serta kata "*bayar*", "*minta*",

dan “*minta bayar*” yang juga memiliki nilai cukup signifikan. Nilai bobot pada query ditampilkan lebih besar karena kata-kata tersebut merupakan inti pencarian yang diberikan oleh pengguna. Tingginya bobot pada dokumen ini menunjukkan bahwa kata-kata kunci tersebut sering muncul dan memiliki tingkat kepentingan yang tinggi dalam dokumen, sehingga dokumen ini dianggap paling relevan.

Sementara itu, pada Rank #2, dokumen berjudul “*laporan permintaan pembayaran keperluan kegiatan lomba 17 Agustus.pdf*” memiliki nilai bobot TF-IDF yang lebih rendah untuk setiap kata kunci dibandingkan dengan dokumen pada peringkat pertama. Sebagai contoh, kata “*agustus*” memiliki bobot 0,1811, dan kata lainnya seperti “*bayar*” dan “*minta*” juga menunjukkan nilai yang lebih kecil. Hal ini menandakan bahwa frekuensi kemunculan dan tingkat kepentingan kata kunci dalam dokumen kedua tidak sekuat pada dokumen pertama.

Dengan demikian, dapat disimpulkan bahwa dokumen dengan bobot TF-IDF yang lebih tinggi pada kata kunci yang relevan akan memiliki peluang lebih besar untuk berada pada peringkat teratas dalam hasil pencarian. Sistem ini menunjukkan bahwa metode TF-IDF mampu mengidentifikasi tingkat kepentingan kata dalam dokumen secara efektif sehingga mendukung proses temu kembali informasi yang lebih akurat.

Analisis Penerapan Cosine Similarity

Perhitungan tingkat kemiripan antara query dan dokumen dilakukan menggunakan metode Cosine Similarity. Nilai cosine similarity dihitung dengan membandingkan vektor TF-IDF query terhadap vektor TF-IDF seluruh dokumen yang telah diindeks. Implementasi perhitungan cosine similarity pada sistem ini menggunakan

fungsi `cosine_similarity()` dari library `scikit-learn`.

Tabel 3. Ranking Hasil Pencarian

Ranking	Dokumen	Skor Similarity
1	Surat permintaan pembayaran beres kegiatan 17 agustus	1,00
2	Surat permintaan pembayaran keperluan kegiatan lomba 17 agustus	0,71

Berdasarkan data pada tabel 2, untuk dokumen peringkat pertama diperoleh nilai dot product sebesar 0,3082 yang berasal dari hasil perkalian setiap bobot TF-IDF antara query dan dokumen, yaitu pada term *agustus*, *bayar*, *minta*, dan *minta bayar*. Selanjutnya dihitung panjang vektor query dengan menjumlahkan kuadrat setiap komponennya dan diakar, yang menghasilkan nilai mendekati 1, sedangkan panjang vektor dokumen pertama juga dihitung dengan cara yang sama. Hasil akhir Cosine Similarity diperoleh dengan membagi nilai dot product dengan hasil perkalian kedua panjang vektor tersebut, sehingga menghasilkan nilai sebesar 0,3082. Proses yang sama dilakukan pada dokumen peringkat kedua, di mana diperoleh nilai dot product sebesar 0,2098 dan menghasilkan nilai Cosine Similarity sebesar 0,2098. Dengan demikian, dapat disimpulkan bahwa dokumen pertama memiliki tingkat kemiripan yang lebih tinggi terhadap query dibandingkan dokumen kedua, sehingga sistem menempatkannya pada peringkat teratas sebagai hasil pencarian yang paling relevan.

SIMPULAN

Berdasarkan hasil perancangan, implementasi, dan pengujian sistem yang telah dilakukan, dapat disimpulkan

bahwa permasalahan pengelolaan arsip dokumen di Kantor Desa Kelakar yang sebelumnya dilakukan secara manual dan menimbulkan kesulitan dalam pencarian dokumen, risiko kehilangan arsip, serta ketidakefisienan waktu dan tenaga, dapat diatasi melalui pengembangan sistem pencarian arsip dokumen berbasis *web*. Sistem yang dibangun mampu mengelola arsip secara terstruktur dan mendukung proses pencarian dokumen secara lebih cepat dan sistematis. Penerapan algoritma *Term Frequency-Inverse Document Frequency* (TF-IDF) dalam proses pembobotan kata serta *Cosine Similarity* dalam perhitungan tingkat kemiripan dokumen terbukti mampu meningkatkan relevansi hasil pencarian berdasarkan kata kunci yang dimasukkan oleh pengguna.

Hasil pengujian akurasi sistem menunjukkan nilai *precision* sebesar 0,001, *recall* sebesar 1,0, dan *F1-score* sebesar 0,0020. Nilai *recall* yang tinggi menunjukkan bahwa sistem mampu menemukan seluruh dokumen yang relevan, sedangkan nilai *precision* menunjukkan bahwa masih terdapat beberapa dokumen yang kurang relevan dalam hasil pencarian.

Dengan demikian, penerapan algoritma TF-IDF dan *Cosine Similarity* pada sistem pencarian arsip dokumen berbasis web ini terbukti efektif dalam meningkatkan efisiensi, kecepatan, dan kualitas pengelolaan arsip dokumen di Kantor Desa Kelakar, serta dapat menjadi solusi digital yang mendukung peningkatan pelayanan administrasi desa.

DAFTAR PUSTAKA

A. H. Nasrullah (2024). "Integrasi Tf-Idf & Algoritma Cosine Similarity Untuk Deteksi Tingkat Kemiripan Judul Penelitian (Studi Kasus Mahasiswa Fakultas Ilmu Komputer UNISAN Gorontalo)," *Information Technology Education Journal*, vol. 3, no. 3.

C. Saputra, H. Irsyad & A. Rahman (2025). "Pemrosesan Query & Pemeringkatan Judul Berita Terkait Gu-bernur Jawa Barat Menggunakan TF-IDF & Cosine Similarity," *JurnalApplied Information Technology and Computer Science*, vol. 4.

D., H. S. F. C., & N. A. Kurniadi (2020). "Implementasi Algoritma Cosine Similarity pada Sistem Arsip Dokumen di Universitas Islam Sultan Agung," *Jurnal Transformatika*, vol. 17, no. 2.

F. A. Nugroho, F. Septian, D. A. Pungkastyo, & J. Riyanto (2021). "Penerapan Algoritma Cosine Similarity untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian & Pengabdian Kepada Masyarakat," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, p. 529.

M. A. Rofiqi, Abd. C. Fauzan, A. P. Agustin, & A. A. Saputra (2019). "Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, vol. 1, no. 2, pp. 58–64.

M. F. Sjojfan, F. Fauziah, & R. T. Aldisa (2023). "Kombinasi Algoritma Sequential Search & Fitur Autocomplete Pada Aplikasi Arsip (E-Arsip) Perpustakaan Berbasis Web," *Journal of Information System Research (JOSH)*, vol. 4, no. 4, pp. 1262–1269.

V., Z. Z., & S. H. S. Vika Sari (2024). "Implementasi Manajemen Kearsipan dalam Meningkatkan Kualitas Layanan Administrasi Sekolah di SMAN 1 Karimun," *Bahasa Indonesia. MUMTAZ - Education Management and Islamic Studies*, vol. 4, no. 2.