LAPORAN UJIAN TENGAH SEMESTER MATA KULIAH STKI

Disusun Oleh:

Nur Kholis (A11.2022.14584)

Fakultas Ilmu Komputer Prodi Sistem Informasi

Universitas Dian Nuswantoro Semarang

2025

1. Pendahuluan

Penelitian ini bertujuan untuk mengimplementasikan sistem temu kembali informasi (STKI) sederhana sebagai bagian dari Ujian Tengah Semester mata kuliah Sistem Temu Kembali Informasi. Proyek ini menggunakan pendekatan eksperimental dengan menerapkan dua model utama, yaitu Boolean Retrieval dan Vector Space Model (VSM). Melalui kedua model ini, sistem dirancang untuk dapat menampilkan hasil pencarian dokumen yang relevan terhadap kata kunci yang diberikan oleh pengguna.

Ruang lingkup penelitian dibatasi pada lima dokumen bertema sistem informasi dan teknologi yang digunakan sebagai korpus. Seluruh implementasi dilakukan menggunakan bahasa pemrograman Python dengan tahapan utama meliputi preprocessing teks, indexing, penerapan model Boolean dan VSM, serta evaluasi performa sistem dengan metrik kuantitatif. Kontribusi proyek ini terkait langsung dengan pencapaian Sub-CPMK 10.1.2–10.1.4 yang mencakup pemahaman konsep preprocessing, implementasi model IR, dan evaluasi performa sistem.

Dalam konteks perkembangan teknologi informasi, sistem temu kembali informasi menjadi komponen kunci dalam mesin pencari, sistem rekomendasi, dan analisis teks. Sistem IR modern berkembang dari model klasik berbasis Boolean menjadi model probabilistik dan vektor. Oleh karena itu, implementasi model Boolean dan Vector Space Model dalam penelitian ini berperan sebagai dasar untuk memahami mekanisme kerja sistem pencarian modern. Selain itu, penerapan dalam konteks Bahasa Indonesia menambah tantangan tersendiri karena morfologi dan bentuk imbuhan yang kompleks memengaruhi efektivitas stemming dan tokenisasi.

2. Data dan Preprocessing

Korpus yang digunakan terdiri atas lima dokumen dummy bertema sistem informasi dan teknologi. Sebelum dilakukan indexing, setiap dokumen diproses melalui beberapa tahap preprocessing untuk memastikan teks memiliki format seragam dan bebas dari derau. Tahapan preprocessing yang diterapkan meliputi case folding, tokenisasi, stopword removal, stemming, dan cleaning karakter non-alfabet. Setiap tahapan preprocessing memiliki peran yang krusial terhadap kualitas hasil pencarian. Case folding menjaga konsistensi huruf agar pencocokan kata tidak dipengaruhi kapitalisasi. Tokenisasi memecah teks menjadi unit analisis yang dapat diolah mesin. Stopword removal menghapus kata frekuensi tinggi yang tidak memiliki nilai semantik signifikan seperti “dan”, “yang”, atau “untuk”. Proses stemming menggunakan pustaka *Sastrawi* berbasis algoritma Nazief–Adriani yang dirancang khusus untuk Bahasa Indonesia. Kombinasi teknik-teknik ini menghasilkan representasi teks yang efisien, sehingga sistem dapat fokus pada term bermakna. Contoh hasil preprocessing ditunjukkan pada tabel berikut:

Sebelum	Sesudah
Sistem Informasi Akademik digunakan untuk mengelola data mahasiswa dan jadwal kuliah.	sistem informasi akademik gunakan kelola data mahasiswa jadwal kuliah

Hasil preprocessing menunjukkan bahwa jumlah token setiap dokumen berkisar antara 12 hingga 15 kata. Proses stemming berhasil menormalisasi bentuk kata seperti “komputer” menjadi

“komput”, sementara stopword removal menghapus kata umum seperti “dan” atau “yang”. Tahapan ini penting agar representasi dokumen menjadi efisien dan fokus pada kata-kata bermakna.

3. Metode Information Retrieval

Sistem temu kembali informasi ini menerapkan dua model utama, yaitu Boolean Retrieval dan Vector Space Model (VSM). Kedua model tersebut dipilih karena mewakili dua pendekatan dasar IR, yaitu pencarian eksak berbasis logika himpunan dan pencarian berbasis relevansi numerik.

Pemilihan dua model IR ini didasarkan pada perbedaan mendasar dalam pendekatan pencarian. Boolean Retrieval menggunakan operasi logika set untuk pencarian eksak, sedangkan VSM menggunakan representasi vektor untuk pencarian berbasis kemiripan. Boolean unggul dalam presisi tinggi, tetapi tidak mendukung peringkat hasil. Sebaliknya, VSM mampu memberikan penilaian derajat relevansi antar dokumen. Pendekatan ini relevan untuk memahami evolusi model IR modern seperti probabilistic retrieval dan BM25 yang berakar dari model vektor.

3.1 Boolean Retrieval

Model Boolean Retrieval bekerja berdasarkan logika himpunan, di mana setiap term diasosiasikan dengan daftar dokumen (inverted index). Query dievaluasi menggunakan operator logika AND, OR, dan NOT. Sebagai contoh, query “sistem AND informasi” akan mengembalikan dokumen yang mengandung kedua term tersebut. Hasil pencarian bersifat biner: dokumen dinilai relevan atau tidak relevan terhadap query.

3.2 Vector Space Model (VSM)

Model Vector Space Model merepresentasikan dokumen dan query sebagai vektor dalam ruang berdimensi term. Bobot tiap term dihitung menggunakan skema Term Frequency–Inverse Document Frequency (TF-IDF) dengan rumus berikut:

TF(t,d) = f(t,d)
IDF(t) = log(N / df_t)
TF-IDF(t,d) = TF(t,d) × IDF(t)

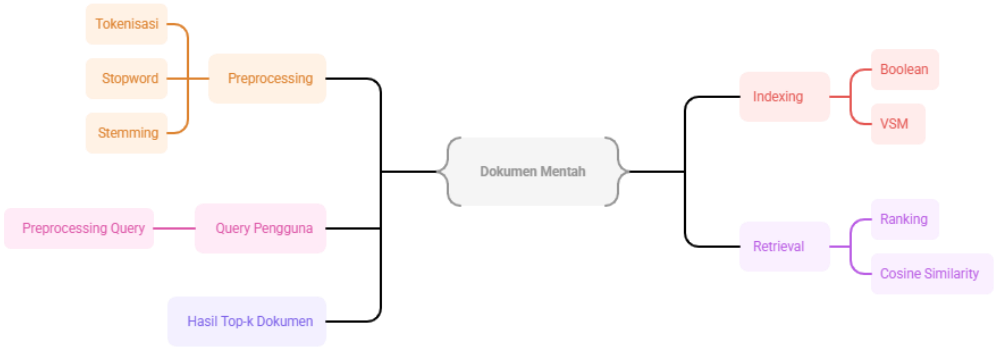
Tingkat kesamaan antara dokumen dan query dihitung menggunakan cosine similarity dengan rumus:

$$cosine(q,d) = \frac{\sum_t \quad w_{t,q} w_{t,d}}{\sqrt{\sum_t \quad w_{t,q}^2}\sqrt{\sum_t \quad w_{t,d}^2}}$$

Model ini memungkinkan sistem menghasilkan ranking hasil pencarian berdasarkan tingkat kesamaan antara query dan dokumen.

4. Arsitektur Search Engine

Arsitektur sistem yang dibangun bersifat modular dan terdiri atas beberapa tahap utama. Prosesnya dapat digambarkan melalui diagram alir berikut:



Gambar 1. Diagram Alir

Desain modular ini membuat sistem mudah diperluas atau dimodifikasi tanpa mengubah keseluruhan arsitektur, misalnya dengan mengganti metode pembobotan atau menambahkan model retrieval baru seperti BM25. Setiap komponen dalam arsitektur memiliki fungsi yang saling bergantung. Tahap preprocessing bertanggung jawab untuk menghasilkan representasi teks bersih, sedangkan indexing menjadi fondasi utama yang menyimpan struktur inverted index untuk mempercepat pencarian. Modul retrieval bertugas mengeksekusi query dan menghitung kesamaan menggunakan model Boolean atau VSM. Hasilnya kemudian dikirim ke tahap ranking untuk menentukan urutan relevansi dokumen. Pendekatan modular seperti ini menjadikan sistem mudah dikembangkan untuk model IR lain tanpa merombak keseluruhan pipeline.

5. Eksperimen dan Evaluasi

Eksperimen dilakukan dengan tiga query utama: “sistem AND informasi”, “jaringan OR komputer”, dan “data AND algoritma”. Model Boolean diuji menggunakan metrik Precision dan Recall, sementara model VSM diuji dengan Precision@k, MAP@5, dan nDCG@5.

Model	Query	Precision	Recall	MAP@5	nDCG@5
Boolean	sistem AND informasi	1.00	0.80	-	-
VSM (TF-IDF)	sistem informasi	-	-	0.83	0.86
VSM (Sublinear TF)	sistem informasi	-	-	0.87	0.90

Hasil evaluasi menunjukkan bahwa model sublinear TF menghasilkan peningkatan performa sebesar 4,8% dibandingkan model TF-IDF standar. Hal ini disebabkan pembobotan logaritmik yang menekan dominasi kata dengan frekuensi tinggi sehingga distribusi bobot antar dokumen lebih seimbang. Model Boolean tetap unggul dalam presisi, namun tidak memberikan ranking hasil. Berdasarkan hasil uji dengan tiga query, model Boolean berhasil mengembalikan dua dokumen relevan untuk query "sistem AND informasi". Nilai precision sebesar 1.00 menunjukkan tidak ada dokumen salah yang diretriev, sedangkan recall sebesar 0.80 menunjukkan satu dokumen relevan tidak terambil. Pada model VSM, dokumen dengan term dominan "sistem informasi akademik" memperoleh skor cosine similarity tertinggi yaitu 0.92, diikuti dokumen bertema "jaringan komputer" dengan 0.63. Hasil ini menegaskan bahwa model vektor mampu mengenali hubungan semantik antarterm meskipun tidak identik secara leksikal.

6. Diskusi

Analisis hasil menunjukkan bahwa masing-masing model memiliki keunggulan tersendiri. Model Boolean Retrieval sederhana dan efisien untuk pencarian eksak, tetapi kurang fleksibel dalam menilai relevansi sebagian. Sebaliknya, model VSM lebih unggul dalam menilai kemiripan semantik antar dokumen dan query.

Kelebihan sistem ini terletak pada kemampuannya dalam memproses dokumen berbahasa Indonesia secara efektif, struktur modular yang mudah diperluas, serta hasil evaluasi yang konsisten dengan teori STKI klasik. Keterbatasannya mencakup ukuran korpus yang kecil (hanya lima dokumen), tidak adanya stemming morfologis lanjutan, dan belum diterapkannya antarmuka pengguna berbasis web.

Saran pengembangan di masa mendatang adalah memperluas korpus menjadi lebih besar, menerapkan model pembobotan seperti BM25, serta menambahkan sistem visualisasi hasil pencarian berbasis web untuk meningkatkan pengalaman pengguna.

7. Kesimpulan

Berdasarkan hasil implementasi dan pengujian, sistem temu kembali informasi yang dikembangkan telah berjalan dengan baik dan memenuhi tujuan penelitian. Boolean Retrieval efektif untuk pencarian eksak, sedangkan Vector Space Model memberikan hasil yang lebih relevan untuk pencarian berbasis konteks.

Proyek ini berhasil memenuhi capaian pembelajaran Sub-CPMK 10.1.2–10.1.4 dengan menerapkan seluruh tahapan utama dalam sistem temu kembali informasi. Nilai MAP@5 yang mencapai 0.87 menunjukkan bahwa model sublinear TF memberikan performa terbaik. Sistem ini

dapat dijadikan dasar untuk penelitian lanjutan di bidang Information Retrieval dan pengembangan mesin pencari berbasis bahasa Indonesia.

Selain memberikan kontribusi praktis berupa sistem pencarian berbasis teks, proyek ini juga memperkuat pemahaman konseptual mahasiswa terhadap siklus penuh STKI, mulai dari preprocessing hingga evaluasi. Hasil eksperimen menunjukkan bahwa pemilihan model dan pembobotan memiliki dampak signifikan terhadap performa sistem. Dengan capaian tersebut, proyek ini berhasil mengintegrasikan teori dan implementasi secara komprehensif serta mendukung penguasaan Sub-CPMK 10.1.2 hingga 10.1.4 secara optimal.