

Implementasi Metode Stemming Tala Dan Fungsi Jaccard Pada Aplikasi Katalog Perpustakaan

Pipin Farida Ariyani¹⁾, Annisa Rahmala²⁾, Noni Juliasari³⁾

^{1),2),3)} Teknik Informatika, Fakultas Teknologi Informasi, Universitas Budi Luhur
Jl. Raya Ciledug, Petukangan Utara, Kebayoran Lama, Jakarta Selatan 12260
Email : pipin.faridaariyani@budiluhur.ac.id

Abstrak. Koleksi buku dalam perpustakaan semakin meningkat sehingga dibutuhkan kategorisasi untuk mempermudah penyusunan dan pencarian buku. Keberadaan sistem penyediaan informasi untuk kemudahan pencarian data buku akan dibutuhkan, namun informasi yang berkualitas dipengaruhi oleh relevansi, keakuratan, dan ketepatan waktu. Penelitian ini dirancang aplikasi katalog perpustakaan untuk mencari informasi mengenai koleksi buku yang ada di perpustakaan menggunakan tools Bahasa pemrograman PHP dan basis data MySQL. Dokumen bahasa Indonesia digunakan sebagai sumber acuan. Untuk kebutuhan pencarian informasi yang relevan dari suatu kumpulan informasi diperlukan proses stemming, pembobotan, dan pengelompokan dalam pencarian dokumen. Proses stemming menggunakan metode Tala, sementara metode Logaritmik Term Frequency berbasis Vector Space Model digunakan untuk menentukan bobot kata suatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut. Semakin besar jumlah kemunculan suatu kata dalam dokumen akan memberikan nilai kesesuaian yang semakin besar. Fungsi Jaccard digunakan untuk menghitung tingkat kesamaan antar query dan term di dokumen yang berpengaruh pada pemeringkatan hasil kategorisasi dokumen yang relevan. Proses penelusuran katalog ini dapat mempercepat perolehan temuan atau informasi mengenai suatu bahan pustaka dari mulai judul hingga ringkasan ebook tersebut. Dari hasil uji coba proses pencarian diperoleh rata-rata perhitungan waktu tanggap aplikasi adalah 5.66 detik dan hasil recall terhadap pencarian dokumen yang didapat yaitu rata-rata 83%.

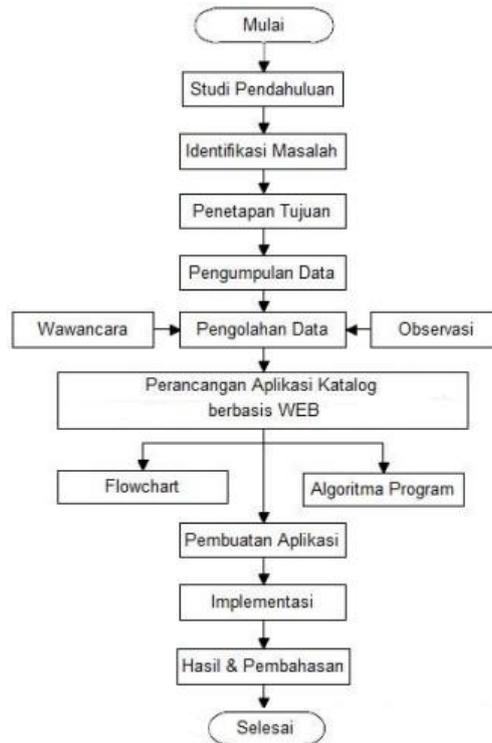
Kata kunci: Pencarian Informasi, Stemming Tala, Jaccard, Bobot Kata, Logaritmik Term Frequency

1. Pendahuluan

Pengelolaan data buku merupakan proses bisnis utama dalam sebuah perpustakaan. Semakin lama sebuah perpustakaan beroperasi maka akan semakin banyak dan beragamnya buku yang tersedia. Saat koleksi meningkat, maka muncul masalah dalam hal penyimpanan dan pencarian data buku yang dibutuhkan dalam waktu singkat. Hal ini diperparah jika pada perpustakaan tidak ada alat bantu pengelolaan data dalam bentuk sistem terkomputerisasi atau aplikasi. Namun untuk perpustakaan yang sudah didukung aplikasi pengelolaan data terkomputerisasi pun masih saja mengalami kendala untuk pencarian data buku yang relevan dan memiliki tingkat akurasi yang baik sesuai yang diinginkan oleh pencari data. Oleh sebab itulah, dalam penelitian ini kami berupaya merancang sebuah aplikasi katalog perpustakaan untuk menghasilkan sebuah temuan atau informasi mengenai suatu bahan pustaka yang relevan dan akurat.

Dalam penelusuran informasi, akan dilihat dari kesamaan dokumen yang ada sehingga jumlah dokumen relevan yang ditelusuri akan dipengaruhi oleh jumlah kata kunci yang digunakan. Pada beberapa penelitian sebelumnya terkait penelusuran kesamaan dokumen, Sugiyanto dkk [1] telah mencoba melakukan penelusuran kesamaan dokumen pada data skripsi untuk mendeteksi plagiarisme yang dipublikasi dengan judul *paper* "Analisa Performa Metode Cosine Dan Jaccard Pada Pengujian Kesamaan Dokumen". Penelitian serupa [2] juga dilakukan oleh Putri dkk dengan pendekatan berbeda, pada artikel publikasi berjudul "Information Retrieval Tugas Akhir dan Perhitungan Kemiripan dokumen Mengacu Pada Abstrak Menggunakan Vector Space Model". Kemudian penelitian [3] yang membangun suatu aplikasi perangkat lunak yang dapat melakukan klasifikasi data teks dengan menggunakan pembobotan TF-IDF terhadap sumber informasi teks elektronik Berbahasa Indonesia yang diunggah secara terpadu dan selektif. Terkait perancangan katalog perpustakaan, telah ada penelitian [4] yang mengimplementasikan metode Rocchio Relevance Feedback untuk memberikan rekomendasi bahan pustaka yang relevan dengan keinginan pengguna berdasarkan *input keyword* dari *user*.

Adapun metode yang akan digunakan dalam aplikasi merupakan implementasi dari konsep *information retrieval* dengan tahapan dapat dilihat pada Gambar 1 (satu) sebagai berikut :



Gambar 1. Alur pikir penelitian

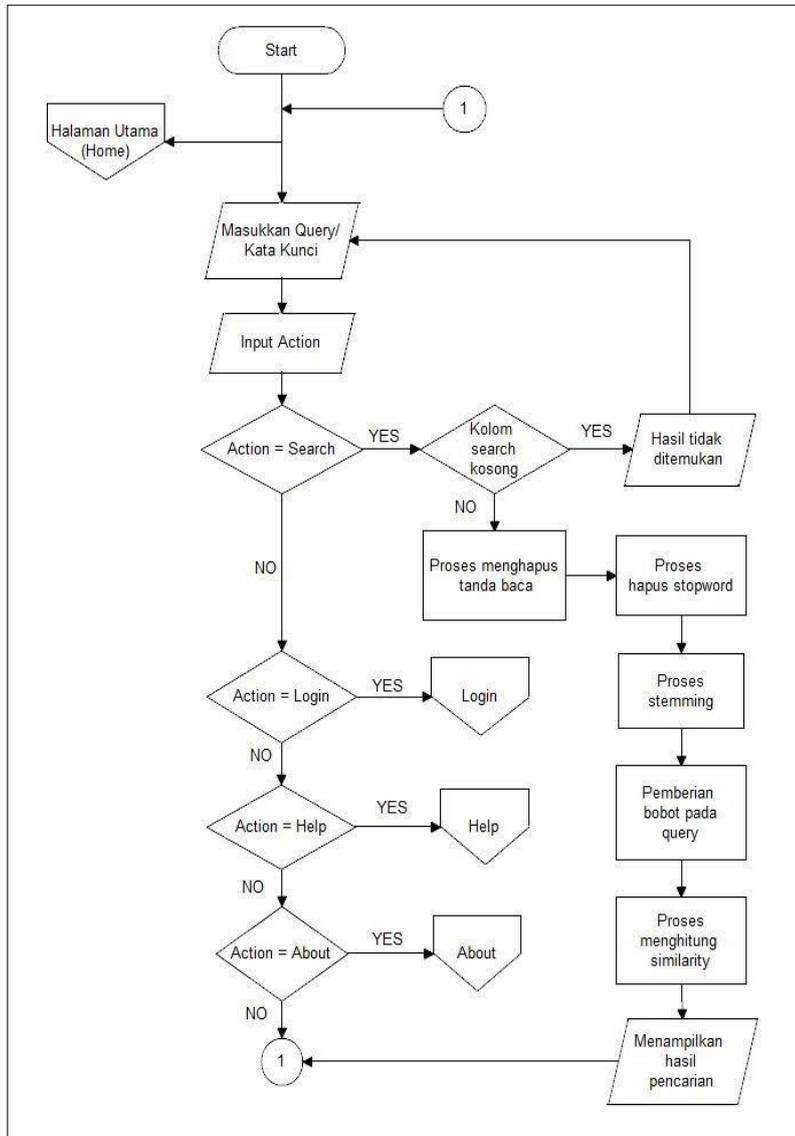
Metode yang akan diterapkan pada sistem penelusuran katalog ini yaitu metode Tala karena metode ini lebih sederhana dan teknik aturan imbuhan sangat mudah untuk diimplementasikan dan tidak membutuhkan waktu komputasi yang tinggi. Selain itu penelitian juga menggunakan *corpus* yang relatif kecil seperti ringkasan *ebook* atau artikel maka metode Tala ini tepat sekali digunakan sebagai alur program yang dapat meningkatkan kualitas indeks dan performa *stemming* berbasis aturan yang relatif stabil dengan jumlah dokumen yang berkembang. Metode Tala ini digunakan sebagai proses *stemming* atau menghilangkan imbuhan dari sebuah kata dan metode ini adalah adopsi dari algoritma *stemmer* bahasa Inggris terkenal Porter. Kemudian juga terdapat pembobotan sebuah kata menggunakan metode Logarithmic Term Frequency. Tahap terakhir yaitu pencarian *similarity* antar dokumen dan *query* menggunakan metode Jaccard. Adapun pembobotan yang dilakukan berbasis Vector Space Model merupakan model *information retrieval system* yang merepresentasikan dokumen dan *query* dalam bentuk vektor dimensional [5].

2. Pembahasan

Aplikasi katalog perpustakaan yang akan dibangun dimaksudkan untuk membantu para siswa maupun petugas perpustakaan dalam menelusuri sejumlah koleksi buku di perpustakaan. Dengan melakukan pembobotan dari masing-masing judul dan ringkasan *ebook* yang sesuai dengan kata kunci yang dicari, sehingga akan menampilkan buku yang relevan dengan kata kunci yang dimasukkan oleh pengguna.

2.1. Alur Proses

Flowchart ini menjelaskan alur proses dari halaman utama yaitu *searching file* selain itu terdapat pilihan menu yang tersedia diantaranya menu “Login”, “Help”, dan “About” seperti yang terlihat pada gambar 2 (dua) berikut :



Gambar 2. Flowchart Menu Utama

Untuk mengimplementasikan aplikasi katalog perpustakaan berbasis web ini dibutuhkan 3 (tiga) metode yaitu Tala, Jaccard, dan Logarithmic Term Frequency dengan 6 (enam) tahap, berikut ini :

1) *Tokenizing*

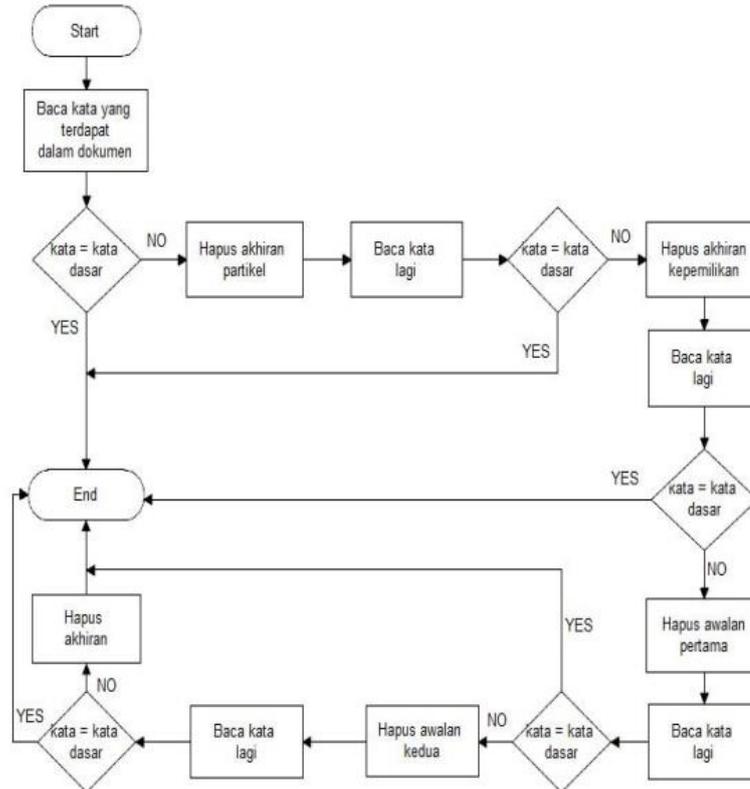
Proses dimana isi dari dokumen yang berisikan kumpulan kalimat dipecah menjadi kata per kata. Disini tanda baca dan spesial karakter lainnya dihilangkan dari kumpulan kata yang ada.

2) *Stopword List Removal*

Proses untuk menghilangkan kata yang sering digunakan dalam dokumen tetapi tidak begitu bermakna.

3) Stemming

Sebuah proses untuk mereduksi bentuk dari sebuah kata menjadi bentuk gramatikal atau menjadi kata dasarnya. Alur proses metode *stemming* yang digunakan yaitu *stemmer* Tala dapat dilihat pada gambar 3 (tiga).



Gambar 3. Alur Proses Stemming

4) Pembobotan Kata

Proses untuk memberikan pembobotan setiap kata didalam dokumen setelah melewati tiga proses diatas.

$$l_{tf} = 1 + \log (tf) \dots\dots\dots (1)$$

Keterangan :

tf = Nilai frekuensi *term* didalam dokumen

l_{tf} = Nilai bobot suatu *term* terhadap dokumen

5) Perhitungan Similaritas

Perhitungan similaritas (kemiripan) antara *query* permintaan *user* dengan dokumen yang tersimpan dalam *database*. Perhitungan dilakukan menggunakan rumus berbasis *vector space model* dengan fungsi Jaccard.

$$\frac{(\sum(X_i.Y_i))}{((\sum((X_i)^2)) + (\sum((Y_i)^2)) - (\sum(X_i.Y_i)))} \dots\dots\dots (2)$$

6) Clustering

Proses pengelompokkan pemisahan dokumen yang relevan dan tidak relevan berdasarkan kata kunci yang dimasukkan oleh pengguna.

2.2. Tampilan Sistem

Aplikasi katalog perpustakaan ini dirancang menggunakan pemrograman PHP yang berjalan di *web browser localhost* dengan berbagai fitur yang tersedia yaitu “Upload File”, “Download File”, kolom pencarian koleksi pustaka, dan menu yang tersedia yaitu “Add/Change Staff”, “Change Password Admin”, “Melihat Daftar Ebook”. Selain itu terdapat menu “Help” untuk memberikan informasi mengenai penggunaan katalog perpustakaan. Berikut adalah tampilan dari *prototype* sistem untuk hasil pencarian dokumen bahan pustaka :



Gambar 4. Hasil Proses Pencarian Dokumen

2.3. Pengujian

Langkah pengujian adalah dengan mengujicobakan data *sample* yang terdiri dari 100 file pdf yang terbagi menjadi 7 (tujuh) kategori *file* yaitu ringkasan 14 *file* kebudayaan, 16 *file* administrasi perkantoran, 13 *file* TKJ, 10 *file* desain, 13 *file* sains, 18 *file* kesehatan, dan 16 *file* teknologi. Kemudian dilakukan ujicoba terhadap *query* pencarian dokumen, lalu dilakukan pengukuran dengan menghitung tingkat performansi pencarian dengan metode Recall dan Precision serta performansi waktu tanggap pengeksekusian sistem. Berikut adalah hasil dari ujicoba terhadap 2 (dua) jenis *query*:

1) Efektivitas Program

Adapun hasil uji coba program dengan menggunakan 2 (dua) *query* untuk mewakili setiap kategori buku yang ada didapatkan hasil nilai Recall dan Precision sebagai berikut :

Tabel 1. Nilai Efektivitas Program

<i>Query</i>	<i>Query 1</i>	<i>Query 2</i>	<i>Query 3</i>	<i>Average</i>
<i>Recall</i>	0.81	0.9	0.78	0.83
<i>Precision</i>	0.62	0.36	0.64	0.54

Pengujian ketepatan (*precision*) ialah perbandingan jumlah relevan yang didapatkan sistem dengan jumlah dokumen seluruh dokumen yang terambil oleh sistem baik relevan maupun tidak relevan. Pengujian kelengkapan (*Recall*) ialah menemukan seluruh dokumen yang relevan dalam koleksi. Dari data yang diujikan, dapat disimpulkan bahwa rata-rata *recall* adalah 83% dan rata-rata *precision* adalah 54%.

2) Efisiensi Program

Adapun hasil uji coba program dengan menggunakan 3 (tiga) *query* untuk mewakili setiap informasi dari 4 (empat) kategori dokumen yang ada didapatkan hasil perhitungan waktu tanggap sebagai berikut :

Tabel 2. Nilai Efisiensi Waktu Tanggap Program

<i>Query</i>	<i>Query 1</i>	<i>Query 2</i>	<i>Query 3</i>	<i>Average</i>
Waktu tanggap	6.6313	4.1492	6.2084	5.66

Performansi waktu tanggap digunakan untuk mengukur waktu saat pengeksekusian program. Untuk dapat menghitung waktu eksekusi proses pencarian dokumen menggunakan fungsi *microtime* yang telah disediakan oleh PHP. Cara paling mudah adalah dengan membuat sebuah variabel dari nilai *return* fungsi *microtime* di awal perintah yang ingin dijalankan, kemudian ambil lagi nilai yang sama dibawah perintah yang ingin dijalankan dengan variabel lain, dan selisih nilai variabel terakhir dengan yang pertama dikurang maka hasilnya yaitu lama eksekusi skrip tersebut. Perhitungan waktu dimulai ketika proses pencarian dokumen dimulai dan berakhir pada saat proses

pencarian dokumen selesai. Dari data yang diujikan, disimpulkan bahwa rata-rata perhitungan waktu tanggap adalah 5.66 detik.

3. Kesimpulan

Berdasarkan pembahasan yang dilakukan maka dapat disimpulkan beberapa hal sebagai berikut :

- 1) Pencarian dokumen yang dilakukan bisa berdasarkan isi konten dokumen sehingga *user* tidak hanya mengandalkan judul *ebook* atau artikel saja sebagai *keyword* dalam proses pencarian.
- 2) Kombinasi penerapan metode *stemming*, pembobotan kata dan juga teknik similaritas yang digunakan sudah mampu memberikan hasil pencarian dokumen yang baik.
- 3) Data uji perlu diperbanyak untuk mendapatkan hasil evaluasi yang lebih akurat, terutama untuk menguji relevansi dokumen dari teknik similaritasnya.

Daftar Pustaka

- [1]. Sugiyamto, B. S. dan A. S., 2014. *Analisa Performa Metode Cosine dan Jaccard pada Pengujian Kesamaan Dokumen*, Jurnal Masyarakat Informatika, 5(10), hal. 1-8.
- [2]. Putri Elfa Mas' udia, Martono Dwi Atmadja, Lis Diana Mustafa, 2017. *Information Retrieval Tugas Akhir Dan Perhitungan Kemiripan Dokumen Mengacu Pada Abstrak Menggunakan Vector Space Model*, Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer, 8(1), hal. 355-362.
- [3]. Susandi, D. dan Sholahudin, U., 2016. *Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani, KNN dan Fungsi Similarity Cosine untuk Pembobotan IDF dan WIDF pada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia*. Jurnal ProTekInfo, 3(1), hal. 22-29.
- [4]. Pausta Yugianus, Harry Soekotjo Dachlan, Rini Nur Hasanah., 2013. Pengembangan Sistem Penelusuran Katalog Perpustakaan Dengan Metode Rocchio Relevance Feedback, Jurnal EECCIS, 7(1).
- [5]. Utomo, M. S., 2013. Implementasi Stemmer Tala pada Aplikasi Berbasis Web. Jurnal Teknologi Informasi DINAMIK, 18(1), hal. 41-45