

BAB I PENDAHULUAN

Bab ini berisikan latar belakang dari penelitian yang akan dilakukan, latar belakang yang dijelaskan meliputi permasalahan dan usulan solusi yang akan dilakukan. Berikut penjelasan latar belakang dari penelitian ini.

1.1 Latar Belakang

Temu kembali informasi atau *Information Retrieval* adalah sebuah bidang ilmu yang mempelajari tentang cara untuk mendapatkan kembali informasi yang pernah tersedia. *Information Retrieval* digunakan untuk mencari dan menemukan informasi yang sesuai dan relevan dengan kebutuhan pengguna secara otomatis dari suatu kumpulan informasi [1]. Sistem temu kembali informasi berkaitan dengan pencarian suatu informasi yang isinya tidak terstruktur, yang selanjutnya ditemukan hasil dari informasi yang relevan dan sesuai dengan kata kunci (*keyword*) [2].

Pencarian informasi akan sangat sulit jika dilakukan secara manual dikarenakan informasi yang ada sangat banyak dan beragam [3]. Pencarian pada basis data (*database*) biasanya hanya mencari judul penelitian yang sesuai dengan kata kunci yang diberikan pengguna, misalnya jika kata kunci yang diberikan adalah “sistem informasi” maka akan ditampilkan semua dokumen yang mengandung kata “sistem informasi” namun sistem tidak dapat mengukur dokumen mana yang paling mirip berdasarkan kata kunci yang diberikan pengguna. Agar dapat melakukan pencarian berdasarkan kata kunci yang paling sesuai dan relevan, maka digunakan teknologi yang disebut *Information Retrieval* [4].

Rizki Tri Wahyuni [5] dalam penelitiannya memanfaatkan *Information Retrieval* untuk membuat sistem yang dapat mengklasifikasikan dokumen secara otomatis menggunakan algoritma *Cosine Similarity* dan metode pembobotan *Term Frequency-Inverse Document Frequency (TF-IDF)*, persentase rata-rata hasil ketepatan klasifikasi yang dilakukan oleh sistem adalah sebesar 98%. Shunzhi Zhu dkk pada tahun 2012 melakukan penelitian tentang perhitungan kemiripan antara kata kunci yang diberikan pengguna dengan informasi yang terdapat dalam dokumen dan pengelompokan dokumen menggunakan algoritma yang mereka usulkan yaitu *Sqrt-Cosine Similarity* [6]. Sahar Sohangir dkk pada tahun 2017 mengusulkan algoritma baru untuk mengukur tingkat kemiripan dokumen dengan

kata kunci yang diberikan pengguna yaitu *Improved Sqrt-Cosine Similarity*. Algoritma ini merupakan pengembangan dari *Sqrt-Cosine Similarity* [7].

Mengacu pada penelitian sebelumnya, penerapan algoritma *Cosine Similarity* untuk mengukur tingkat kemiripan suatu dokumen dengan kata kunci dapat dikatakan salah satu yang paling populer dan telah banyak digunakan. Perhitungan *Cosine Similarity* berdasarkan jarak *Euclidean*. Secara umum jarak *Euclidean* bukan matriks yang efektif dalam menangani probabilitas, dimana sering digunakan dalam analisis teks [7]. Shunzhi Zhu [6] dalam penelitiannya mengusulkan pengembangan dari algoritma *Cosine Similarity* yaitu *Sqrt-Cosine Similarity*. Mereka mengklaim bahwa sebagai pengukuran tingkat kemiripan, algoritma ini memberikan nilai antara nol dan satu, yang dinilai lebih baik dengan pendekatan berbasis probabilitas. Namun dalam beberapa kasus, perhitungan kemiripan *Sqrt-Cosine Similarity* tidak selaras dengan definisi dari pengukuran tingkat kemiripan itu sendiri [7].

Perhitungan algoritma *Improved Sqrt-Cosine Similarity* berdasarkan pada jarak *Hellinger*, dimana jarak *Hellinger* telah terbukti dapat bekerja lebih baik dibandingkan dengan algoritma yang berdasarkan jarak *Euclidean* seperti algoritma *Cosine Similarity*. Sahar Sohangir [7] dalam penelitiannya mengatakan *Hellinger distance (L1 norm)* dinilai lebih baik dari *Euclidean distance (L2 norm)* sebagai matrik untuk menangani data berdimensi tinggi. Berdasarkan hasil penelitian yang dilakukan oleh [7] algoritma *Improved Sqrt-Cosine Similarity* selalu mengungguli pengukuran tingkat kemiripan menggunakan algoritma *Cosine Similarity* dan juga pengukuran tingkat kemiripan menggunakan algoritma *Sqrt-Cosine Similarity*.

Penelitian ini mencoba untuk membangun suatu sistem temu kembali informasi yang diimplementasikan dalam sebuah sistem penyimpanan dokumen penelitian berbasis web. Dalam penelitian ini data dokumen penelitian yang akan digunakan hanya mengambil judul serta isi bagian abstrak dari dokumen penelitian. Algoritma yang digunakan adalah algoritma *Improved Sqrt-Cosine Similarity* karena beberapa hasil penelitian sebelumnya memperlihatkan bahwa algoritma *Improved Sqrt-Cosine Similarity* mampu memberikan tingkat presisi terbaik dalam pencarian dokumen dengan metode pembobotan kata *TF-IDF*.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan di atas, maka dapat disimpulkan perumusan masalahnya adalah:

1.2.1 Permasalahan

Suatu sistem temu kembali dokumen umumnya menggunakan algoritma *Cosine Similarity* yang berdasarkan jarak *Euclidean*. Jarak *Euclidean* dinilai kurang efektif dalam menangani analisis teks. Terdapat pengembangan dari algoritma *Cosine Similarity* yaitu *Sqrt-Cosine Similarity*, namun pada beberapa kasus, perhitungan tingkat kemiripan menggunakan *Sqrt-Cosine Similarity* tidak selaras dengan definisi dari pengukuran tingkat kemiripan itu sendiri.

1.2.2 Metode Usulan

Metode yang akan digunakan untuk mengatasi permasalahan di atas ialah menggunakan algoritma *Improved Sqrt-Cosine Similarity* dengan metode pembobotan *TF-IDF*.

1.2.3 Pertanyaan Penelitian

Berapa rata-rata presisi sistem temu kembali dokumen yang menerapkan algoritma *Improved Sqrt-Cosine Similarity* dengan metode pembobotan *TF-IDF* untuk mencari dokumen penelitian.

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

Penerapan algoritma *Improved Sqrt-Cosine Similarity* dengan metode pembobotan *TF-IDF* ini bertujuan untuk memperoleh sistem yang mampu lebih presisi dalam mencari dan menemukan dokumen penelitian.

1.3.2 Manfaat

Manfaat dari penelitian ini ialah sebagai berikut:

1. Dengan selesainya perancangan dan pembangunan Sistem Temu Kembali Informasi ini dapat membantu pengguna untuk mencari dokumen penelitian dengan hasil yang lebih presisi dan lebih efisien.

2. Mempermudah dalam pencarian informasi dokumen penelitian karena dapat mempresentasikan hasil informasi secara terurut berdasarkan kemiripan antara kata kunci dengan informasi yang terdapat pada dokumen penelitian.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini sebagai berikut:

1. Koleksi dokumen (*corpus*) yang akan digunakan dalam penelitian ini dalam bentuk *database* dokumen penelitian/skripsi yang berasal dari Portal Tugas Akhir Trunojoyo.
2. Proses *stopword removal* hanya mendukung bahasa Indonesia.
3. Data dokumen penelitian yang akan digunakan hanya mengambil isi dari bagian judul dan abstrak.
4. Dokumen yang akan digunakan dalam penelitian ini sebanyak 683 dokumen penelitian serta berhubungan dengan bidang informatika.
5. *Stemming* (pemotongan imbuhan) tidak dilakukan pada dokumen penelitian.

1.5 Sistematika Proposal

Sistematika penulisan proposal ini terdiri dari beberapa bab, yaitu:

BAB I PENDAHULUAN

Pada bab ini membahas tentang latar belakang, rumusan masalah, tujuan dan manfaat, batasan masalah, serta sistematika dalam penulisan laporan.

BAB II KAJIAN PUSTAKA

Pada Bab ini berisi kajian pustaka dan landasan teori. Tinjauan pustaka berisi pustaka singkat dan jelas atas pustaka yang mendasari bidang kajian sebelumnya. Sedangkan landasan teori berisi teori yang mendukung penelitian.

BAB III METODE USULAN

Pada bab ini membahas tentang metode yang diusulkan (algoritma), arsitektur sistem (*Input, Proses, Ouput*), data penelitian, tahapan penelitian, skenario pengujian serta rencana pengujian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil dari penelitian yang dilakukan. Hasil yang diperoleh berupa kekurangan, kelebihan, dan uji coba yang telah dilakukan.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil uji coba serta saran yang dapat digunakan untuk pengembangan penelitian berikutnya.