

BAB I PENDAHULUAN

1.1 Latar Belakang

Saat ini pertumbuhan jumlah dokumen di Internet sangat pesat. Proses pencarian dokumen yang sangat banyak, tentunya memerlukan sebuah teknik khusus dalam mengambil / mencari dokumen yang relevan dengan permintaan pengguna. Salah satu teknik yang sangat umum digunakan dalam proses pencarian dokumen adalah *Information Retrieval* (IR). IR memberikan informasi tentang subjek yang diperlukan. Data termasuk teks, audio, video, dan dokumen lainnya. Tujuan dari penggunaan IR dalam pencarian dokumen salah satunya yaitu menghasilkan dokumen yang relevan dengan permintaan pengguna dalam waktu singkat dan tepat [1]. Pada proses pencarian dokumen dapat dilakukan dengan menggunakan sebuah *query*. *Query* adalah sebuah kata kunci yang digunakan untuk menampilkan data dari *database* kemudian akan diolah lebih lanjut. Data tersebut diambil dari tabel-tabel dalam *database* yang memudahkan pengguna dalam mengolah banyak data. Untuk mendapatkan dokumen yang relevan dengan permintaan pengguna, maka *query* harus relevan dengan dokumen yang diinginkan dan hal ini sangat dibutuhkan dalam proses pencarian dokumen. [2].

Permasalahan dari penggunaan *query* secara umum adalah dari sudut pandang pengguna atau *user*. Pengguna memiliki karakter yang berbeda-beda. Beberapa pengguna tidak bisa memasukkan kata kunci yang tepat untuk mengembalikan daftar hasil yang berkualitas tinggi. Mengambil informasi menggunakan sistem pencarian informasi sering membutuhkan kata kunci yang tepat dari berbagai bidang untuk mencapai hasil terbaik. Ratusan ribu dokumen yang tidak relevan akan dikembalikan oleh sistem dan tak jarang dianggap sebagai dokumen relevan jika kata kunci yang dipilih terlalu umum. Ketidakcocokan kosakata juga merupakan salah satu alasan ketidakefektifan sistem pencarian informasi. Ini adalah masalah mendasar untuk pengambilan informasi [3].

Query yang terlalu singkat dapat diselesaikan dengan cara *Query Expansion*, yaitu teknik menambahkan *query* tambahan pada *information retrieval* di dalam teknik *relevance feedback*. Teknik ini akan memodifikasi *query* untuk dapat memenuhi kebutuhan informasi yang dibutuhkan [4]. Dengan kata lain, proses merancang kembali *query* awal untuk meningkatkan kinerja sistem dalam proses

pengambilan informasi. Tujuan dari penggunaan *query expansion* secara khusus adalah untuk menambahkan istilah yang bermakna pada *query* awal yang memungkinkan untuk membantu pengguna dalam menghapus ambiguitas bahasa alami dan juga mengekspresikan konsep informasi dengan cara yang lebih rinci ke dalam *query* asli. Dengan menambahkan kata-kata terkait ke *query* asli juga dapat meningkatkan jumlah dokumen yang relevan teridentifikasi dan meningkatkan kemungkinan penemuan dokumen yang relevan [3]. Sehingga membuat *query* tersebut menjadi benar dan sesuai dengan yang diinginkan oleh pengguna. Tetapi tidak semua *query* perlu di perluas, terdapat beberapa faktor yang mempengaruhinya. Salah satunya adalah *query* yang singkat tetapi akurasinya tinggi, maka tidak perlu di perluas lagi.

A. Keikha dkk [5] dalam penelitiannya *query expansion* menggunakan metode *pseudo relevance feedback* dengan Wikipedia. Artikel Wikipedia yang paling terkait dengan *query* diidentifikasi dan dianggap sebagai dokumen umpan balik, Berdasarkan *query expansion* yang telah dilakukan, menghasilkan sebuah *query* yang telah di perluas. Untuk melakukan perhitungan persamaan dokumen, digunakanlah metode *cosine similarity* dan pembobotan menggunakan TF-IDF. Metode *cosine similarity* digunakan karena metode ini mempunyai nilai akurasi yang tinggi dimana kelebihan utamanya adalah tidak terpengaruh pada panjang pendeknya suatu dokumen dan pembobotan TF-IDF digunakan karena metode ini merupakan metode yang paling umum digunakan pada *information retrieval* untuk perhitungan bobot suatu kata. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat [6], [7].

Penggunaan metode *pseudo relevance feedback* di sebagian besar kasus dokumen umpan balik pada kenyataannya banyak mengandung dokumen yang tidak relevan dengan *query* [5]. Untuk menyelesaikan masalah ini, maka digunakan artikel Wikipedia bahasa Indonesia sebagai dokumen umpan balik untuk menghindari masuknya dokumen yang tidak relevan dalam koleksi dokumen. Bahasa Indonesia memiliki banyak sekali kata-kata yang ambigu, sehingga wikipedia bahasa Indonesia perlu digunakan sebagai acuan agar kata-kata yang ambigu atau kata yang mengandung istilah dapat memiliki suatu makna tertentu.

Dengan menggunakan metode *pseudo relevance feedback* dengan dataset wikipedia bahasa Indonesia, diharapkan untuk kasus *query expansion* dalam bahasa Indonesia ini dapat mencapai nilai rata-rata presisi yang tinggi. Dengan bantuan dari Wikipedia bahasa Indonesia, diharapkan kata - kata yang tidak baku tersebut menjadi sebuah kata yang dapat digunakan oleh pengguna sebagai *query* yang benar. Hasil dari pencarian berdasarkan *query* tersebut sesuai dengan yang diinginkan pengguna.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Permasalahan yang umum terjadi adalah penggunaan *query* yang terlalu singkat atau ambigu. Sehingga hasil *query* yang diinginkan tidak sesuai dengan dokumen yang didapatkan.

1.2.2 Usulan Solusi

Solusi dari permasalahan tersebut adalah *Query Expansion* menggunakan metode *pseudo relevance feedback* dengan dataset wikipedia bahasa Indonesia.

1.2.3 Pertanyaan Penelitian

Seberapa besar rata-rata presisi metode *pseudo relevance feedback* dengan dataset Wikipedia bahasa Indonesia dalam melakukan *Query Expansion* bahasa Indonesia?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

Bertujuan untuk mengukur nilai rata-rata presisi metode *pseudo-relevance feedback* dengan dataset Wikipedia bahasa Indonesia jika diterapkan dalam *Query Expansion* bahasa Indonesia

1.3.2 Manfaat

Diharapkan kita dapat menentukan metode yang sesuai dengan *Query Expansion* untuk bahasa Indonesia. Jika dengan penggunaan metode *pseudo-relevance feedback* dengan dataset Wikipedia memiliki nilai rata-rata presisi yang sangat besar, maka dapat di terapkan dalam *Query Expansion* bahasa Indonesia.

1.4 Batasan-batasan

Batasan masalah pada penelitian ini adalah sebagai berikut :

1. Bahasa yang akan digunakan adalah bahasa Indonesia
2. Hanya menghitung nilai presisi metode *pseudo relevance feedback* dengan dataset Wikipedia bahasa Indonesia untuk *query expansion* bahasa Indonesia
3. Dokumen yang digunakan sebanyak 100 artikel bahasa Indonesia tentang pariwisata Madura
4. Menggunakan bahasa pemrograman *php* dan *database MySQL*
5. Data wikipedia yang digunakan adalah abstrak dari artikel di wikipedia bahasa Indonesia sebanyak 50.000 artikel.
6. Jumlah maksimal kata untuk kata kunci adalah 25 kata

1.5 Sistematika Laporan

Sistematika laporan penelitian ini adalah sebagai berikut :

1. BAB I PENDAHULUAN

Bab ini membahas mengenai Latar belakang, Perumusan masalah, Tujuan dan Manfaat, Batasan masalah, dan Sistematika laporan.

2. BAB II KAJIAN PUSTAKA

Bab ini membahas tentang Landasan teori, Penelitian sebelumnya dan Contoh penggunaan metode

3. BAB III METODE USULAN

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

4. BAB IV HASIL DAN PEMBAHASAN

Pada bab ini membahas tentang proses pembuatan sistem, penjelasan beberapa kode program, dan hasil pengujian

5. BAB V PENUTUP

Pada bab ini membahas tentang kesimpulan dan saran dari penelitian ini.