1. PENDAHULUAN

1.1 Latar belakang

Perkembangan informasi digital saat ini berkembang dengan pesat. Setiap hari bahkan setiap menit jutaan informasi terus bertambah. Dengan semakin bertambahnya informasi yang tersedia muncul kebutuhan untuk menemukan informasi tertentu dari data yang banyak. Pencarian manual dengan membaca semua dokumen dalam tempat penyimpanan merupakan *perfect retrieval* tetapi bukan tindakan efektif. Oleh karena itu dibutuhkan *information retrieval* untuk pencarian informasi secara efektif.

Salah satu pembahasan dalam *information retrieval* adalah tentang perangkingan dokumen. Perangkingan dokumen melakukan pengurutan dokumen hasil retrieval yang disesuaikan dengan query yang dimasukkan oleh pengguna. Dokumen yang berada pada rangking teratas merupakan dokumen yang paling relevan dengan query sementara dokumen yang berada paling bawah merupakan dokumen yang kurang relevan dengan query. Beberapa penelitian yang membahas perangkingan dokumen seperti: perangkingan dengan menggunakan pencocokan N-gram terhadap kata dari query dan dokumen [5][11], menggunakan modul *crawler* [4], menggunakan *vector space model* [10].

Permasalahan pada penelitian ini adalah algoritma yang digunakan melakukan perangkingan dokumen dengan melakukan perbandingan pada kata per kata (*term by term*). Jadi *term-term* dianggap sebagai *term* yang *independen*t (tidak saling berhubungan). N-gram, melakukan pencocokan string dengan terlebih dahulu mentransformasikan *term-term* kedalam varian n-gram yaitu unigram, digram, trigram, quadgram. Pada *Vector Space Model* dokumen dianggap sebagai kumpulan *term-term*, *term-term* tersebut direpresentasikan dengan vektor matematika. Dimana, tingkat pentingnya sebuah *term* dalam dokumen dinyatakan dengan TF dan IDF. Jika diberikan query "fakultas informatika" maka saat menghitung jarak kedekatan query dengan dokumen menggunakan perhitungan *cosine similarity*, dokumen yang mengandung *term* dari query "fakultas" dalam jumlah tinggi bisa mendapat nilai similarity tinggi yang artinya akan menempatkan dokumen tersebut pada peringkat teratas, padahal dokumen tersebut belum tentu berhubungan dengan query "fakultas informatika".

Terlihat perbandingan dokumen dengan menggunakan kata per kata (*term by term* memiliki kelemahan yaitu hanya menggunakan kata saja, konteks yang dicari kurang lengkap dibandingkan dengan query yang diberikan. Salah satu solusi masalah ini adalah dengan melakukan *re-rangking* dokumen. Dengan melakukan *re-rangking* tersebut maka rangking dokumen yang dilakukan akan lebih sesuai dengan konteks *query* [6].

Pada tugas akhir ini, digunakan *re-rangking* dokumen sebagai kunci perangkingan dokumen bahasa Indonesia. Untuk perangkingan awal sebelum dilakukan *re-rangking* digunakan *vector space model*. *Vector space model* dipilih karena paling sederhana dan telah terbukti memiliki efektifitas dalam pencarian kata, dengan menampilkan hasil pencariannya berdasarkan kemiripan vector *query* dan vector dokumen. Setelah mendapatkan hasil perengkingan dari *Vector*

Space Model maka akan dilakuakan *re-rangking* dokumen dengan menerapkan suatu rumus yaitu pembobotan query.

1.2 Rumusan masalah

Berdasarkan pada latar belakang di atas, permasalahan yang akan diuraikan dan diteliti adalah:

1. Bagaimana cara menerapkan *Vector Space Model* pada Information retrieval.
2. Bagaimana tahap pengerjaan *Vector Space Model.*
3. Bagaimana cara melakukan *re-rangking* dokumen, setelah didapatkan hasil perangkingan awal dengan *Vector Space Model*.

1.3 Batasan Masalah

Adapun batasan masalah untuk tugas akhir ini adalah:

1. Query yang diinputkan berupa frase. Pada sistem menerima panjang query maksimal adalah tiga kata.
2. Dokumen yang akan dianalisis terbatas pada dokumen text bahasa Indonesia yang memiliki ejaan yang benar (EYD).
3. Data yang digunakan sebagai bahan analisis yaitu dokumen yang berasal http://www.tabloidpulsa.co.id/.
4. Penelitian ini berfokus pada *re-rangking* dokumen bahasa Indonesia, agar rangking yang dihasilkan lebih sesuai dengan konteks query pengguna.
5. Data yang akan dianalisis secara offline learning, yaitu sistem tidak terhubung ke internet.

1.4 Tujuan

Tujuan yang ingin dicapai dalam pengerjaan Tugas Akhir ini adalah sebagai berikut:

1. Menemukan kembali dokumen sesuai dengan query yang diinputkan user.
2. Melakukan *re-rangking* dokumen bahasa Indonesia, agar hasil perangkingan yang dihasilkan lebih sesuai dengan konteks query pengguna.
3. Mengevaluasi hasil *re-rangking* dokumen bahasa Indonesia.

1.5 Hipotesa

Hipotesis dalam penelitian tugas akhir ini adalah *re-rangking* dokumen dapat meningkatkan nilai *precision* dan *recall* dipertitik peringkat.

1.6 Metodologi Penyelesaian Masalah

Penyelesaian masalah dilakukan dalam beberapa tahap, secara garis besar sebagai berikut:

1. Tahap studi literatur

Pada tahap ini dilakukan studi literature yang relevan dengan permasalahan *information retrieval* khususnya *Vector Space Model*, pembobotan, proses indexing dan searching serta sumber-siumber lain yang relevan untuk menunjang penyelesaian tugas akhir ini. Sumber- sumber pustaka didapat dari buku, paper, maupun halaman web.

2. Tahap pengumpulan dataset
   Pada tahap ini dilakukan pengumpulan data berupa teks-teks dalam bahasa Indonesia. Teks tersebut berada dalam sebuah dokumen, yang nantinya dokumen tersebut akan digunakan untuk uji coba.
3. Tahap pre-processing
   Pada tahap ini pre-processing yang dilakukan pada data adalah cleaning dan parsing.
4. Tahap Analisis
   Pada tahap ini akan dianalisis hasil percobaan dan hasil simulasi menggunakan dua parameter, yaitu: *precision* dan *recall*. *S*erta penarikan kesimpulan.
5. Pembuatan Laporan
   Pada tahap ini dilakukan pembuatan laporan Tugas Akhir untuk mendokumentasikan tahap-tahap dari setiap kegiatan dan percobaan yang dilakukan.