

1. Pendahuluan

1.1. Latar Belakang Masalah

Information Retrieval merupakan istilah untuk mempelajari sistem pencarian sehingga mendapat informasi yang dicari, mulai dari *indexing*, *searching* (pencarian), dan *recalling* (pemanggilan data kembali). Pada proses pencarian, dibutuhkan *query*, sebagai *keyword* yang kemudian dicocokkan dengan *document collection*. Dengan memasukkan *query* pada aplikasi yang menerapkan *Information Retrieval*, *user* akan diberikan dokumen-dokumen yang relevan dengan *query* tersebut.

Pada penerapan *Information Retrieval* sederhana, mesin pencari hanya mengandalkan *query* yang dimasukkan oleh *user* dalam proses pencarian sebagai informasi yang digunakan dalam proses pencarian. Padahal terdapat pula kemungkinan dokumen-dokumen yang relevan yang tidak mengandung *term-term* dari *query* yang dimasukkan oleh *user*. Sehingga, bisa jadi mesin pencari tidak mendapatkan dokumen-dokumen relevan tersebut. Dokumen-dokumen relevan terhadap *query* tidak selalu mengandung *term-term* pada *query*, sehingga dibutuhkan *term-term* baru selain dari pada *term-term query* yang dapat merepresentasikan informasi yang dibutuhkan oleh *user*. Di mana *term-term* baru tersebut memiliki keterkaitan dengan *term-term* yang ada pada *query*. Hal tersebut yang menjadi alasan pengembangan *Information Retrieval System*.

Markov Chain merupakan salah satu model probabilistik yang saat ini penerapan dan pengembangannya telah dilakukan dalam berbagai bidang dan ilmu pengetahuan. Model ini menggambarkan suatu rangkaian proses, di mana kejadian di masa yang akan datang tidak bergantung pada proses yang di masa lalu, tapi bergantung pada proses di masa sekarang. Hal tersebut dapat pula diterapkan pada suatu aplikasi *Information Retrieval*, karena penerapan perhitungan pada Markov Chain dapat diaplikasikan pada pengembangan *query* berdasarkan distribusi *term* pada suatu *document collection*.

Markov Chain pada *Information Retrieval* merupakan salah satu contoh penerapan *query expansion*. Pada dasarnya, Markov Chain pada *Information Retrieval* menggambarkan hubungan antara dokumen dengan *term*, dan hubungan antara *term* dengan dokumen. Hubungan tersebut dinyatakan dalam perhitungan probabilitas antara dokumen dengan *term*, dan *term* dengan dokumen. Nilai probabilitas dari *term* baru yang akan dihasilkan tersebut bergantung pada distribusi *term-term* pada *query* dan jumlah kemunculan *term* pada dokumen.

Dengan menerapkan model Markov Chain pada *Information Retrieval*, maka dapat ditentukan *term-term* baru. *Term-term* baru tersebut dapat membantu *user* dalam menemukan dokumen-dokumen hasil temuan yang relevan yang tidak mengandung *query* yang dimasukkan oleh *user*.

1.2. Rumusan Masalah

Permasalahan yang dijadikan objek penelitian kali ini adalah :

1. Bagaimana penerapan Markov Chain dalam menghasilkan *term-term* baru berdasarkan distribusi *term* pada *query* yang dimasukkan oleh *user* dan berdasarkan distribusi *term* pada *document collection*?

2. Bagaimana hasil analisa performansi aplikasi yang menerapkan Markov Chain pada *Information Retrieval*?

Batasan masalah dalam penelitian ini adalah :

1. Jenis dokumen yang digunakan adalah jenis *free-text document* berbahasa Inggris. Dan *document collection* yang digunakan, diambil dari <ftp://ftp.cs.cornel.co.uk/smart>.
2. Aplikasi tidak melakukan penanganan terhadap *Boolean Operations*.
3. Performansi yang diukur pada aplikasi di setiap *step* Markov Chain diukur berdasarkan tingkat relevansi dokumen yang dihasilkan, hal tersebut berdasarkan nilai *Precision*, *Recall*, dan *IAP (Interpolated Average Precision)*.
4. Aplikasi dijalankan secara *offline*.

1.3. Tujuan

Hal-hal yang ingin dicapai dalam penelitian kali ini adalah :

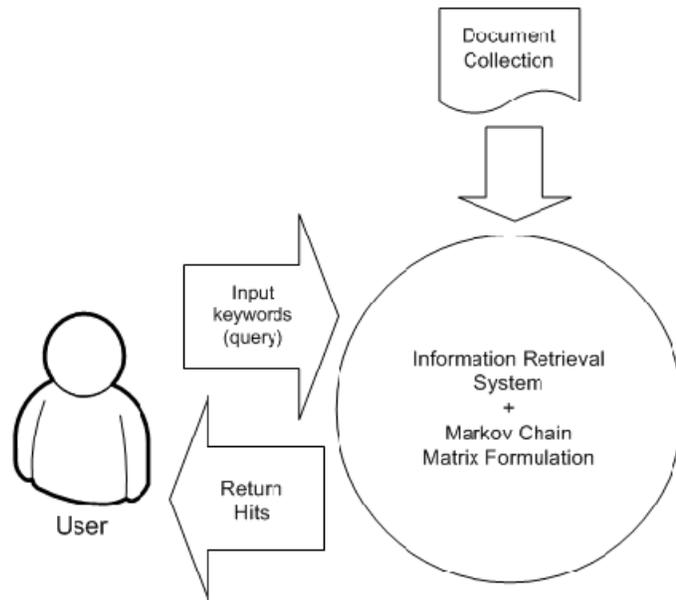
1. Menganalisis dan mengimplementasikan model perhitungan Markov Chain pada *Information Retrieval* dalam menghasilkan *term-term* baru berdasarkan *query* yang dimasukkan oleh *user*. Kemudian melakukan perangkingan dokumen hasil pencarian.
2. Menganalisis performansi aplikasi pada setiap *step* dan pada setiap perubahan jumlah *query expansion* dan *document expansion* yang dijalankan oleh aplikasi dengan menggunakan perhitungan *Precision*, *Recall* dan *IAP (Interpolated Average Precision)*.

1.4. Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam proses penelitian kali ini adalah :

1. Studi Literatur
Dilakukan pembelajaran, dan pengumpulan referensi, guna mempelajari konsep mengenai *Information Retrieval* dan mengenai teori Markov Chain. Juga dilakukan pemahaman mengenai karakteristik dari model Markov Chain pada *Information Retrieval*.
2. Analisis Kebutuhan Aplikasi dan Desain
Pada tahap ini dilakukan analisis kebutuhan terhadap aplikasi yang menerapkan Model Markov Chain pada *Information Retrieval*. Juga dilakukan desain aplikasi Model Markov Chain pada *Information Retrieval*.
3. Implementasi
Pada tahap ini dilakukan implementasi aplikasi berdasarkan hasil analisis kebutuhan aplikasi.
4. Pengujian dan analisis hasil
Melakukan pengujian aplikasi yang telah dibangun dan kemudian menganalisis hasil performansi aplikasi.
 - a) Dilakukan pengujian aplikasi dengan menggunakan data berupa koleksi dokumen yang telah tersedia, kemudian dilakukan analisis performansi aplikasi.

- b) Untuk menganalisis tingkat korelevanan dokumen-dokumen yang dihasilkan, adalah dengan melakukan perhitungan *Precision* dan *Recall*.
 - c) Untuk menganalisis perbandingan dokumen hasil pencarian dilakukan perhitungan IAP (*Interpolated Average Precision*).
5. Penyusunan laporan tugas akhir.
Melakukan penyusunan laporan berdasarkan hasil analisis kebutuhan aplikasi, beserta implementasi aplikasi dan hasil pengujian aplikasi yang menerapkan Model Markov Chain pada Information Retrieval. Gambar 1.1 menggambarkan mekanisme kerja aplikasi secara garis besar :



Gambar 1.1 : Gambaran Umum Kerja Aplikasi