MODEL MARKOV CHAIN PADA INFORMATION RETRIEVAL

Minda Septiani[1], Yanuar Firdaus A.w.[2], Retno Novi Dayawati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Information Retrieval merupakan metode pencarian dokumen yang fleksibel dan dalam pengembangannya, dapat disesuaikan dengan kebutuhan user pada informasi dan terhadap jenis dokumen yang ada. Dengan mengembangkan konsep matematis dari model Information Retrieval, maka dapat dihasilkan model dan konsep Information Retrieval baru yang dapat digunakan dalam melakukan pencarian dokumen terhadap koleksi dokumen.

Penggabungan antara konsep Information Retrieval dan Model Markov Chain dapat menjadikan satu model temu kembali dokumen yang dapat diandalkan, perlakuan model ini terhadap keywords yang dimasukkan user pun dapat menjadi bahan pertimbangan penelitian lebih lanjut terhadap model tersebut.

Dalam tugas akhir ini dilakukan pengujian terhadap aplikasi yang menerapkan konsep Information Retrieval dan Model Markov Chain sebagai model matematis yang diterapkan pada tahap Matrix Formulation. Penggunaan step awal pada aplikasi yang menerapkan Model Markov Chain pada Information Retrieval menghasilkan aplikasi dengan performansi yang paling baik, karena dengan meningkatnya ukuran step yang digunakan, maka performansi aplikasi semakin menurun.

Didapatkan pula bahwa nilai IAP dan Recall aplikasi yang menerapkan Model Markov Chain pada Information Retrieval lebih baik dibandingkan dengan nilai IAP dan Recall aplikasi Infromation Retrieval yang menerapkan TFIDF.

Nilai tertinggi Precision aplikasi yang menerapkan Model Markov Chain pada Information Retrieval pada koleksi dokumen ADI adalah 0,097, sedangkan pada koleksi dokumen CRAN adalah 0,0668. Nilai tertinggi Recall aplikasi pada koleksi dokumen ADI adalah 1, sedangkan pada koleksi dokumen CRAN adalah 0,81. Nilai tertinggi IAP aplikasi yang menerapkan Model Markov Chain pada Information Retrieval pada koleksi dokumen ADI adalah 0,39, sedangkan pada koleksi dokumen CRAN adalah 0,5969.

Kata Kunci : Information Retrieval, Query expansion, Markov Chain, Precision, Recall, Query.

Abstract

Information Retrieval is a method which is used for search document, and can be fited for user need of information and documents. By developing the mathematic concept of Information Retrieval method, we can get the improvement of document search application from document collection.

Combination of Information Retrieval and Markov Chain Model can make one new Information Retrieval method, the way of this combination model could be researched and implemented.

At this final task, we implemented the combination of Information Retrieval and Markov Chain Model as mathematic model which is used in the Matrix Formulation calculation. Using the first step for Markov Chain Model in Information Retrieval, application get best performance, because if we set higher step for application, then performance of application will get worse.

IAP and Recall of apllication which implement the Markov Chain Model in Information Retrieval are better than application which implement TFIDF.

The highest Precision Markov Chain Information Retrieval application for ADI document collection is 0,097, and for CRAN document collection is 0,0668. The highest Recall Markov Chain Information Retrieval application for ADI document collection is 1, for CRAN document collection is 0,81. The highest IAP Markov Chain Information Retrieval application for ADI document collection is 0,39, and for CRAN document collection is 0,5969.

Keywords : Information Retrieval, Query expansion, Markov Chain, Precision, Recall, Query.

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*).

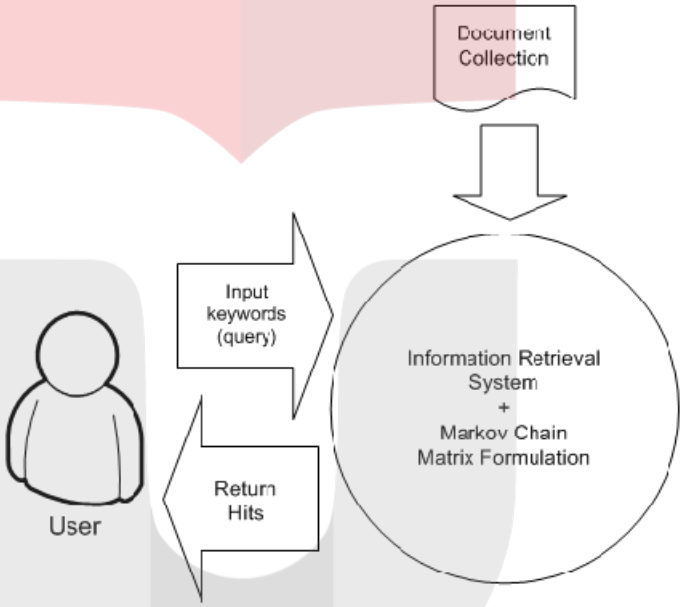
1.4. Metodologi Penyelesaian Masalah

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 kerelevanan dokumen-dokumen yang dihasilkan, adalah dengan melakukan perhitungan *Precision* dan *Recall*.
c) Untuk menganalisis perangkingan 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

5. Penutup

Pada bab ini akan diuraikan hal yang dapat disimpulkan dari pelaksanaan Tugas Akhir ini. Selain itu diuraikan pula beberapa saran yang dapat digunakan dalam pengembangan Tugas Akhir di masa mendatang.

5.1. Kesimpulan

Berdasarkan hasil analisis dan pengujian aplikasi yang dilakukan dalam Tugas Akhir ini dapat diambil beberapa kesimpulan yaitu :

1. Penerapan aplikasi Markov Chain Model pada Information Retrieval menghasilkan kesimpulan bahwa perhitungan matematis Model Markov Chain dalam menghasilkan term baru dan dalam mencari dokumen pada aplikasi Information Retrieval telah dapat diterapkan berdasarkan perhitungan pada tahap Matrix Formulation.
2. Penerapan aplikasi Markov Chain Model pada Information Retrieval menghasilkan kesimpulan, sebagai berikut:
   - Aplikasi yang menerapkan Model Markov Chain pada Information Retrieval terbukti memiliki nilai Recall dan IAP yang lebih baik dibanding dengan aplikasi Information Retrieval yang menerapkan TFIDF.
   - Semakin besar jumlah *term* baru, jumlah dokumen yang diambil, maka semakin rendah nilai *Precision* aplikasi. Sedangkan jika ukuran step yang digunakan bertambah maka nilai Precision aplikasi meningkat.
   - Semakin besar jumlah *term* baru dan jumlah dokumen yang diambil, maka semakin tinggi nilai *Recall* aplikasi, sedangkan semakin besar ukuran step yang digunakan, maka semakin rendah *Recall* aplikasi.
   - Semakin besar jumlah dokumen yang terambil, semakin besar jumlah term baru dan semakin besar ukuran step yang digunakan pada aplikasi, maka semakin rendah nilai IAP aplikasi.

5.2. Saran

Untuk pengembangan Tugas Akhir di masa mendatang, penulis menyarankan hal-hal sebagai berikut:

1. Aplikasi selanjutnya diharapkan dapat menggunakan metode perhitungan perkalian matriks yang lebih efisien.
2. Aplikasi selanjutnya diharapkan menggunakan metode penyimpanan data ke *database* yang lebih efisien.
3. Koleksi dokumen yang digunakan agar tidak hanya dokumen file teks berformat .txt.
4. Koleksi dokumen yang digunakan menggunakan bahasa lain, contohnya bahasa Indonesia.

6. Daftar Pustaka

[1] Abdurachman, Edi. Konsep Dasar Markov Chain serta Kemungkinan Penerapannya di Bidang Pertanian. Didownload pada 6 Desember 2009.

[2] Belkin, N.J. "Anomalous State of Knowledge as a Basis for *Information Retrieval*", *Canadian Journal of Information Sciences*, 5, 1980, 133-143.

[3] Berger, Adam and John Lafferty. Infromation Retrieval as Statistical Translation. Didownload pada 6 Desember 2009

[4] Blanco, Roi and Alvaro Barreiro. Probabilistic *Document* Length *Prior*s for Language Models. Didownload pada 10 Oktober 2010.

[5] Lafferty, John. Chengxiang Zhai. *Document* Models, *Query* Models, and Risk Minimization for *Information Retrieval*. Didownload pada 6 Desember 2009.

[6] Lehmann, Alain and John Shawe-Taylor. A Probabilistic Model for Text Kernels. Didownload pada 10 Oktober 2010.

[7] Manning, D. Christopher, Prabhakar Raghavan, Hinrich Scutze. An Introduction to *Information Retrieval*. Cambridge University Press. England. 2008.

[8] Miller, David R. H., Tim Leek, and Richard M. Schwartz. A Hidden Markov Model *Information Retrieval* . Didownload pada 6 Desember 2009.

[9] Rabiner, Lawrence. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Didownload pada 6 Desember 2009.

[10] Song, fei and W. Bruce Croft. A General Language Model for *Information Retrieval*. Didownload pada 6 Desember 2009.

[11] Thompson, Kevyn Collins and Jamie Callan. *Query expansion* using Random Walk Models. Didownload pada 6 Desember 2009.

[12] Wikipedia. http://en.wikipedia.org/wiki/Information_retrieval. didownload pada 6 Desember 2009