ANALISIS DAN IMPLEMENTASI PENGGABUNGAN COOCCURENCE DAN PROBABILISTIC METHOD DALAM QUERY EXPANSION PADA INFORMATION RETRIEVAL

Putu Eka Doddy Suzanto[1], Yanuar Firdaus A.w.[2], Kusuma Ayu Laksitowening[3]

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

Abstrak

Information Retrieval (IR) merupakan bagian dari computer science yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Proses dalam Information Retrieval dapat digambarkan sebagai sebuah proses untuk mendapatkan relevant documents dari collection documents melalui pencarian query yang diinputkan user. Salah satu cara untuk meningkatkan performansi Information Retrieval ialah dengan mengoptimasi query inputan user, yaitu dengan cara menambahkan keyword-keyword baru pada query awal inputan user ataupun dengan merubah bobot dari setiap keyword yang ada pada query, atau yang sering disebut dengan Relevance FeedBack. Salah satu bentuk dari relevance Feedback adalah Query Expansion. Dalam Tugas Akhir ini, Penulis menggabungkan metode Co-Occurence dan Probabilistic untuk mendapatkan kata-kata yang akan digunakan untuk query expansion. Dari hasil pengujian didapatkan bahwa penggunaan metode gabungan ini dapat meningkatkan performansi dari sebuah Information Retrieval System. Selain itu, penggabungan kedua metode ini juga memberi nilai performansi yang lebih baik jika kedua metode ini digunakan secara terpisah.

Kata Kunci : Information Retrieval, Information Retrieval System, Query Expansion, Co-Occurence, Probabilistic, dan query.

Abstract

Information Retrieval (IR) is a computer science that related to retrieved information from collection documents based on the content of the documents. Process in Information Retrieval can be described as a process to collect relevant documents from collection documents based on user query. A way to increase the performance of Information Retrieval is with optimize query as known as Relevance Feedback. One of Relevance feedback is known as query expansion, which create a new query from original query by adding some keyword that relevan to original query. In this Final Project there are combine of Co-Occurence and Probabilistic Methods for getting new keyword. The result shows that the combine methods can increase perfoemance of the Information Retrieval System. Beside that, the combine of both methods can improve performance of each methods.

Keywords : Information Retrieval, Information Retrieval System, Query Expansion, Co-Occurence, Probabilistic, and query.

1. PENDAHULUAN

1.1 Latar Belakang

Menyusun ulang query(*query reformulation*) yang dimasukkan oleh user adalah hal yang sering dilakukan dalam *information retrieval*. Hal ini dilakukan untuk mengatasi ketidaksesuaian antara query yang dimasukkan oleh user dengan informasi yang ingin didapatkannya. *Query reformulation* yang sering dipakai adalah dengan *Query Expansion*, yaitu dengan memanjangkan query yang dimasukkan oleh user dengan menambahkan beberapa term kedalamnya. Query yang dimasukkan oleh user pada umumnya pendek dan *query expansion* dapat melengkapkan informasi yang ingin dicari user.

Dalam temu kembali dokumen beberapa pendekatan yang dapat digunakan dalam meranking term yang akan digunakan dalam *query expansion*. Dua diantaranya adalah melalui pendekatan *Cooccurence* dan *Probabilistic.* Pendekatan pertama merupakan pendekatan yang berbasis kepada pengukuran kemunculan kandidat term dan term dari query pada document yang didapatkan atau yang sering disebut sebagai *cooccurence approach*. Yang kedua adalah dengan pendekatan probalistik(*probabilistic approach*), pendekatan ini berbasis kepada distribusi sebuah term dalam *document collection* dan hasil pencarian teratas.

Pendekatan pertama menghasilkan term-term yang sering muncul bersamaan dengan term yang terdapat pada query user. Sedangkan untuk pendekatan kedua, *query expansion* menghasilkan term-term yang memiliki tingkat kepentingan yang tinggi. Karena kedua pendekatan ini menghasilkan term-term yang berbeda, maka sangat mungkin dilakukan penggabungan untuk mendapatkan hasil dari *query expansion* yang lebih baik.

1.2 Perumusan Masalah

Berdasarkan uraian diatas maka permasalahan yang muncul dan yang menjadi objek penelitian pada Tugas Akhir ini :

1. Bagaimana proses pencarian calon *additional term* yang akan dipakai dalam memanjangkan *query* dengan menggunakan metode *Co-Occurence* dan *Probabilistic*.
2. Bagaimana proses penggabungan metode *Co-Occurence* dan *Probabilistic* dalam pencarian calon *additional term*.
3. Bagaimana perbandingan performansi pencarian antara metode pada Tugas Akhir ini dengan *original matching function*-nya.
4. Bagaimana pengaruh performansi hasil penggabungan metode *Co-Occurence* dan *Probabilistic* terhadap masing-masing metode.

Batasan masalah agar tidak meluasnya materi pembahasan dalam tugas akhir ini adalah sebagai berikut :

1. Koleksi dokumen dan kata kunci yang digunakan adalah dokumen dalam teks bahasa Inggris.
2. Koleksi dokumen menggunakan koleksi dokumen yang diambil dari ftp://ftp.cs.cornell.edu/SMART. Koleksi dokumen yang dipakai adalah

koleksi dokumen dengan kategori MED yang terdiri dari 1033 buah dokumen.

3. Pada koleksi dokumen tersebut sudah terdapat kumpulan kata kunci yang berjumlah sepuluh buah beserta *relevance judgement* untuk setiap kata kunci(*Query*) tersebut.
4. Stopword menggunakan daftar kata yang terdapat pada koleksi dokumen(yang diambil dari ftp://ftp.cs.cornell.edu/SMART).
5. Model sistem temu kembali informasi yang digunakan pada awal pencarian adalah model ruang vektor(Vector Space Model), dimana dokumen maupun kata kunci direpresentasikan sebagai vektor berdimensi *n,* dengan *n* adalah jumlah kata atau *term* pada kata kunci. *Relevance feedback* pada model ruang vektor dapat dijelaskan sebagai penggeseran vektor kata kunci mendekati vektor dokumen relevan dan menjauhi vektor dokumen tidak relevan[2].
6. Jumlah dokumen teratas yang akan diproses untuk mendapatkan calon *additional term* adalah lima(5) buah dokumen(peringkat 1 sampai dengan 5).
7. Jumlah *additional term* yang digunakan untuk setiap pencarian berjumlah lima(5) buah term.
8. Pengujian menggunakan parameter IAP(*Interpolated Average Precision*).

1.3 Tujuan

Secara umum tujuan penulisan yang ingin dicapai dalam tugas akhir ini adalah:

1. Merancang dan membangun suatu *Information Retrieval System* yang menggunakan *Query Expansion* dengan menggabungkan Metode *Co-Ocurrence* dan *Probabilistic*.

2. Menganalisis proses dari pencarian *Additional Term* dengan menggabungkan Metoda *Co-Ocurrence* dan *Probabilistic* yang kemudian digunakan untuk membuktikan apakah metode ini mampu memberikan nilai IAP yang lebih baik dari metode awalnya(*Vector Space Model*).

1.4 Metodologi Penyelesaian Masalah

Metodologi yang digunakan untuk menyelesaikan masalah dalam Tugas Akhir ini adalah :

1. Studi Literatur
   Studi literatur dari beberapa buku, jurnal, artikel yang membahas tentang *Information Retrieval, Query Expansion, Co-Occurence Method,* dan *Probabilistic Method.*

2. Analisis dan Desain
   Tahap ini meliputi analisis kebutuhan serta penyelesaian masalah untuk merancang perangkat lunak *search engine* dengan metoda *Query Expansion* dengan menggabungkan *Cooccurence* dan *Probalistic Method.* Desain perangkat lunak yang akan dibangun berdasarkan proses berikut :

3. Implementasi Sistem

Tahap ini meliputi pembangunan perangkat lunak yang telah dirancang pada tahap sebelumnya. Pembangunan perangkat lunak lebih ke arah *web-based* dengan menggunakan bahasa pemprograman PHP dan database MySQL. Implementasipenggabungan kedua metoda ini dilakukan secara seri seperti yang terlihat di gambar 2.

4. Analisis dan Pengujian

Pada tahapan ini yang dilakukan adalah melakukan pengujian terhadap perangkat lunak yang dibangun dan sekaligus melakukan analisis terhadap hasil pemrosesan perangkat lunak. Analisa performansi dari *search engine* ini setelah digunakan *Query Expansion* akan dinilai dari nilai IAP yang dihasilkan dari metode ini dan kemudian akan dibandingkan dengan nilai IAP dari *original matching function*-nya.

5. Penyusunan dan Laporan

Hasil penelitian akan disusun menjadi suatu laporan yang meliputi aspek-aspek dalam penelitian yaitu teori, perancangan dan implementasinya, serta membuat kesimpulan dari hasil penelitian tersebut.

1.5 Sistematika Penulisan

Sistematika Penulisan Tugas Akhir ini terdiri dari 5 Bab, yaitu:

BAB I Pendahuluan

Bab ini membahas kerangka penelitian dalam tugas akhir, meliputi latar belakang, perumusan masalah, batasan masalah, tujuan perancangan dan metodologi yang digunakan dalam perancangan system.

BAB II Landasan Teori

Bab ini menjelaskan seluruh teori yang menjadi landasan konseptual dan mendukung penyelesaian tugas akhir ini.

BAB III Analisis dan Perancangan Sistem

Bab ini membahas mengenai pengumpulan data analisis dan perancangan perangkat lunak yang terdiri dari perancangan struktur data, perancangan modul dan *interface*.

BAB IV Implementasi dan Pengujian Sistem

Bab ini membahas implementasi detail sistem dan pengujian terhadap sistem.

BAB V Kesimpulan dan Saran

Berisi tentang kesimpulan dan saran yang dapat diambil dari keseluruhan sistem yang telah dibuat.

5. KESIMPULAN DAN SARAN

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 perangkat lunak yang dilakukan dalam Tugas Akhir ini dapat diambil beberapa kesimpulan yaitu :

a. Pendekatan yang yang digunakan pada metode *Co-Occurence* memberikan nilai performansi yang hampir sama.
b. Pendekatan yang paling baik pada metode *Probabilistic* adalah pendekatan Bose-Einstein.
c. Penggunaan metode penggabungan *Co-Occurence* dan *Probabilistic* pada VSM dapat meningkatkan performansi pada *Information Retrieval System* (IRS) jika dibandingkan dengan pencarian menggunakan model ruang vektor saja.
d. Penggabungan metode *Co-Occurence* dan *Probabilistic* meningkatkan performansi pencarian metode masing-masing.

5.2 Saran

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

a. Dokumen yang digunakan lebih banyak dan lebih luas sehingga memaksimalkan kinerja metode *Co-Occurence*.
b. Proses *stemming* yang digunakan dapat dikembangkan lagi.
c. Pencarian *synonym words* akan meningkatkan proses pencarian.
d. Perangkat lunak tidak hanya menangani *word indexing* tapi juga dapat menangani *phrase indexing*.
e. Jenis dokumen yang dicari tidak hanya berupa teks saja.

DAFTAR PUSTAKA

[1] Amin Aly,Abdelmgeid, *Using A Query Expansion Technique to Improve Document Retrieval*, International Journal "Information Technologies and Knowledge" Vol.2, 343-348, 2008, Eygpt: El-Minia University.

[2] Baeza-Yates, R., and Ribeiro-Neto, B., *Modern Information Retrieval*, 1999, ACM Press, NY, USA.

[3] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi, *An information-theoretic approach to automatic query expansion*, ACM Trans. Inf. Syst. 19(1):1-27, 2001.

[4] C. J. Van Rijsbergen, *A theoretical basis for cooccurence data in information retrieval.* Journal of Documentation. 106-119. 1997.

[5] Gianni Amati and Cornelis Joost Van Rijsbergen, *Probabilistic models of information retrieval based on measuring the divergence from randomness*, ACM Trans. Inf. Syst. 357-389, 2002.

[6] Helen J. Peat and Peter Willett, *The limitations of term co-occurrence data for query expansion in document retrieval systems*. JASIS, 42(5):378–383, 1991.

[7] Ingwersen, Peter. 2005. *The Turn : System-Oriented Information Retrieval*, Book Series The Information Retrieval Series Vol. 18. Springer Netherlands Publishers, Netherlands.

[8] J. Callan. Distributed Information Retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, 127-150, 2000. Kluwer Academic Publishers.

[9] Kim, B. M., Kim, J. Y. and Kim, J., *Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference*, Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, Canada, Vol. 2, pp.715-720, 2001.

[10] Lin, Hsi-Ching, Li-Hui Wang, Shyi-Ming Chen, *Query Expansion for Document Retrieval by Mining Additional Query Terms*, Information and Management Sciences Vol. 19, No. 1, pp. 17-30, 2008, Taiwan : National Taiwan University.

[11] R.Perez-Aguera,Jose, Lourdes Araujo, *Comparing and Combining Methods for Automatic Query Expansion*, Research in Computing Science 33, pp. 177-188,2008, Universidad Complutense de Madrid.

[12] Rocchio, J.Y., *Relevance Feedback in Information Retrieval*, The SMART Retrieval System. Engelwood Cliff, N.J.: Prentice Hall, PP. 313-323, 1971.

[13] Thomas M. Cover, Joy A. Thomas, *Elements of information theory.*Wiley-Interscience, NewYork.1991.