ANALISIS DAN IMPELEMENTASI RELEVANCE FEEDBACK DALAM INFORMATION RETRIEVAL DENGAN MENGGUNAKAN BLIND RELEVANCE FEEDBACK TIPE 1 DAN TIPE 2

Yogi Saputra[1], Yanuar Firdaus A.w.[2], Warih Maharani[3]

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

Abstrak

Information Retrieval (IR) mempunyai sebuah kemampuan untuk memberikan informasi yang dibutuhkan, selama query yang dimasukkan dapat diubah menjadi sebuah set dokumen yang relevan terhadap user. Agar information retrieval dapat mengembalikan set dokumen yang relevan terhadap user, query yang dimasukkan haruslah mengandung kata yang mempunyai relevansi yang tinggi pula terhadap Index dokumen yang dimiliki oleh system information retrieval tersebut. Relevance Feedback (RF) dapat digunakan untuk membangun query yang lebih relevan.

Relevance Feedback(BRF) merupakan salah satu jenis RF yang sering digunakan untuk mendapatkan query yang lebih efektif. BRF 1 adalah metode RF yang umum digunakan, sedangkan BRF Tipe 2 adalah pengembangan dari BRF 1. Perbedaan yang signifikan antara BRF Tipe 1 dan Tipe 2 adalah query expansion collection yang digunakan. Pada BRF Tipe 1, query expansion collection yang digunakan adalah dokumen yang berada di index itu sendiri. Sedangkan pada BRF tipe 2, query expansion collection yang digunakan adalah document collection yang berada di luar document collection yang terdapat dalam index. Dalam Tugas Akhir ini, akan dilakukan analisis perbandingan performansi antara IR tanpa RF, IR menggunakan BRF Tipe 1, IR menggunakan BRF Tipe 2, dan IR menggunakan kombinasi BRF Tipe 1 dan Tipe 2, query expansion collection yang digunakan untuk BRF Tipe 2 adalah Wordnet.

Berdasarkan observasi yang telah dilakukan dapat diketahui bahwa akurasi BRF tipe 1 sedikit lebih baik daripada BRF Tipe 2, waktu yang diperlukan untuk melakukan proses RF untuk BRF Tipe 1 juga lebih cepat daripada BRF Tipe 2. Namun masih perlu penelitian lebih lanjut, apakah dengan menggunakan query expansion collection yang digunakan, BRF Tipe 2 dapat melebihi kemampuan BRF Tipe 1 dalam mengembalikan dokumen yang relevan. Semua BRF mempunyai akurasi retrieval yang lebih tinggi daripada IR tanpa menggunakan RF.

Kata Kunci : Information Retrieval, Relevance Feedback, Blind Relevance Feedback Tipe 1,

Abstract

Information Retrieval has an ability to provide a relevant document, as long as the query inputted into the system can be converted into a set of document that is relevant to user. For IR to be able to return a relevant set of document to user, the query inputted has to include a word that is of a high relevance to the document Index that are being used by the particular IR System. Relevance Feedback (RF) can be used to build a more relevant query.

Blind Relevance Feedback (BRF) is the common RF that is being used to get a more effective query. BRF Type 1 is the most common usage of RF, whereas BRF Type 2 are the improvement BRF, where the difference between them is the query expansion collection used. In BRF Type 1, the query expansion collection used, are the documents set itself, where in BRF Type 2, query expansion collection that are being used, is the document collection outside of the indexed document collections. In this Final Project analysis will be conducted to measure and compare the performance between IR without RF, IR with BRF Type 1, IR With BRF Type 2 and IR With BRF Type 1 and Type 2 combined. Query expansion collection used for BRF Type 2 is WordNet.

Based on observation that are already done, it has been known that BRF Type 1 are slightly better than BRF Type 2, time that are needed to proses BRF Type 1 are faster than the time for BRF Type 2. But this conclusion need a futher analysis, is a better query expansion collection can provide a better result for BRF Type 2 to surpass BRF Type 1 ability on retrieving relevant document. All of the BRF return a better retrival than IR alone without RF.

Keywords : Information Retrieval, Relevance Feedback, Blind Relevance Feedback Tipe 1,

1. Pendahuluan

1.1. Latar Belakang

Dalam era perkembangan teknologi informasi saat ini, keberadaan sebuah *search engine* telah menjadi sebuah bagian dari kehidupan sehari-hari. *Search engine* dapat membantu pengguna internet untuk menemukan informasi di Internet, dengan terlebih dahulu memasukkan sebuah kata kunci kedalam *search engine* tersebut [5].

Kelemahan dari *search engine* adalah, *search engine* hanya akan mencari informasi yang dicari dengan tepat, dengan kata kunci atau disebut juga dengan *query* yang tepat pula. Disinilah terjadi masalah pada proses pencarian informasi melalui *search engine*. User biasanya menginputkan *query* yang terlalu panjang atau pendek dan tidak spesifik. *Query* yang dimasukkan biasanya bermakna sangat luas dan tidak menggambarkan secara spesifik informasi yang dicari. Untuk itulah dalam Information Retrieval, diperlukan sebuah *Relevance Feedback*. *Relevance Feedback* adalah sebuah teknik, dimana kata kunci (*Query* Awal) akan diformulasi ulang untuk menghasilkan *query* yang lebih baik, berdasarkan dari dokumen yang telah berhasil diretrieve dari *query* awal [5].

Dalam Tugas Akhir ini, akan dibahas tentang Information Retrieval dengan menggunakan *Blind Relevance Feedback*. *Blind Relevance Feedback* merupakan salah satu metode *Relevance Feedback* dalam Information Retrieval, dimana user tidak perlu memberikan *feedback* secara aktif kepada sistem, tetapi sistem akan melakukan *Query Expansion* dengan menganggap bahwa dokumen yang mempunyai ranking tinggi adalah dokumen yang relevan terhadap pencarian user. Alasan dipilihnya metode *Blind Relevance Feedback* adalah karena,dibandingkan dengan metode *Relevance Feedback* yang lain, *Blind Relevance Feedback* mempunyai keunggulan mengotomasi bagian manual dari *Relevance Feedback*, dan tidak memerlukan assessor, sehingga hasil *Relevance Feedback* tidak akan terpengaruh faktor Assesor. Teknik *Blind Relevance Feedback* yang digunakan pada tugas akhir ini adalah *Blind Relevance Feedback* Type 1 dan 2, dimana perbedaan dari BRF*(Blind Relevance Feedback*) tipe 1 dan 2 adalah *Query Expansion Collection* yang digunakan [2].

1.2. Perumusan Masalah

Perumusan masalah dalam tugas akhir ini adalah sebagai berikut :

1. Bagaimana perbandingan performa *Information Retrieval* menggunakan masing-masing tipe *Blind Relevance Feedback*, dan Kombinasi dari kedua tipe *Blind Relevance Feedback* dalam meretrieve dokumen yang dicari?
2. Bagaimana mengetahui tipe BRF mana yang tepat untuk digunakan, apakah tipe BRF yang berbeda akan mengembalikan hasil yang berbeda?
3. Bagaimana menentukan saat yang tepat untuk menggunakan BRF tipe 1 atau tipe 2, dan menentukan saat yang tepat untuk tidak menggunakan BRF ?

Adapun batasan masalah dalam Tugas Akhir ini adalah :

1. *Relevance Feedback* yang digunakan hanya *Blind Relevance Feedback*.
2. Dataset yang digunakan adalah TREC WT10g [WTX010~WTX019] Berjumlah 500 Files, dimana *data collection* yang didapatkan adalah *data*

collection dari hasil crawl situs di Internet. Sedangkan *Query* uji yang digunakan adalah *query* dari TREC web track.

3. *Query Expansion Collection* yang digunakan dalam *Blind Relevance Feedback* Type 2 adalah WordNet.
4. Tugas Akhir ini berfokus pada *Blind Relevance Feedback*, sehingga dalam Tugas Akhir ini tidak akan membahas secara mendalam proses indexing,dan proses *preprocessing* lainnya sebelum proses *Relevance Feedback.*

1.3. Tujuan

Secara umum, Tujuan dari pembuatan Tugas Akhir ini adalah :

1. Untuk mengukur dan menganalisis sejauh mana peningkatan hasil *retrieve* dokumen yang dihasilkan oleh *Blind Relevance Feedback* Type 1 dan 2, jika dibandingkan dengan hasil retrieve dokumen tanpa menggunakan *Blind Relevance Feedback.*
2. Untuk menganalisis, tipe BRF mana yang tepat untuk digunakan dalam Topik Uji Studi Kasus Tugas Akhir ini. Jika ada perbedaan performa antara BRF tipe 1 dan BRF tipe 2, maka hasil analisa juga diharapkan dapat menunjukkan faktor-faktor yang mempengaruhi perbedaan performa BRF tersebut.
3. Untuk menganalisis, kapan saat yang tepat untuk menggunakan BRF tipe 1, atau Tipe 2. Dan kapan saat untuk tidak menggunakan BRF sama sekali.
4. Secara khusus, diharapkan setelah Tugas akhir ini selesai, dapat dibuktikan bahwa *Information Retrieval* yang menggunakan *Blind Relevance Feedback*, menghasilkan *Information Retrieval* yang lebih baik dalam kemampuannya meretrieve dokumen yang relevan daripada sebelum menggunakan *Blind Relevance Feedback.*

Hipotesa awal untuk Tugas Akhir ini adalah, *Information Retrieval* dengan menggunakan *Blind Relevance Feedback* , akan mengembalikan atau meretrieve dokumen yang lebih relevan bagi user dibandingkan dengan *Information Retrieval* yang tidak menggunakan *Blind Relevance Feedback*, terlepas dari tipe dari *Blind Relevance Feedback* yang digunakan.

1.4. Metodologi Penyelesaian Masalah

Metodologi yang digunakan dalam menyelesaikan Tugas Akhir ini adalah :

1. Studi Literatur
   Mempelajari landasan teori dari referensi-referensi yang ada tentang *Information Retrieval, Relevance Feedback, Blind Relevance Feedback*, dan literatur-literatur lainnya yang membantu baik itu dalam penelitian maupun dalam pembuatan aplikasi.
2. Pembangunan perangkat lunak
   a) Analisis dan Perancangan
      Melakukan analisis dan perancangan perangkat lunak dengan menggunakan metode *Unified Modelling Language* (UML)
   b) Pengkodean
      Implementasi dari hasil perancangan ke dalam pemrograman komputer dengan menggunakan teknik pemrograman berorientasi objek.
   c) Pengujian
      Adapun skenario pengujuan yang akan dilakukan adalah sebagai berikut:
      1) Melakukan penyimpanan *data collection*.

2) Melakukan indexing terhadap informasi-informasi yang terdapat dalam *data collection*.
3) Melakukan proses pencarian terhadap *data collection* dengan menggunakan *query* uji yang telah ditentukan. Pengujian pada tahap ini dilakukan terhadap *Information Retrieval* tanpa *Blind Relevance Feedback*, dan *Information Retrieval* dengan menggunakan *Blind Relevance Feedback.*
4) *Query* uji sebisa mungkin akan dibuat untuk tidak memihak, dalam artian *query* uji tidak akan membuat sebuah IR lebih baik dari IR yang lainnya.
5) Melakukan analisis dari hasil pengujian aplikasi.

3. Analisis Hasil
Pada tahap ini, akan dilakukan analisis hasil *query expansion* dengan menggunakan beberapa sampel kasus sebagai berikut :
a. *Query* uji dimasukkan pada sistem tanpa menggunakan *Blind Relevance Feedback.*
b. *Query* uji dimasukkan pada sistem dengan menggunakan *Blind Relevance Feedback.*

Parameter yang digunakan dalam pengujian sistem ini adalah :
1) *Precision*
Rasio dari seberapa banyak dokumen yang ter-*retrieve*, relevan terhadap user.
2) *Recall*
Rasio dari seberapa banyak dokumen relevan yang seharusnya ter-*retrieve*, berhasil di *retrieve*.
3) MAP (*Mean Average Precision*)
Mean dari *Average Precision* setelah dokumen yang relevan berhasil di *retrieve*. Nilai MAP didapat dari *Average precision* dibagi jumlah set *Query* uji yang digunakan.
4) nDCG
5) R-*Precision*
Sebuah pengukuran yang mengkomplemenkan MAP. Menggambarkan keseluruhan *Recall* dari *Top Ranked Document*.

4. Pengambilan Kesimpulan dan Pembuatan Laporan
Menarik kesimpulan dari hasil analisis yang telah dilakukan terhadap sistem, serta mendokumentasikan hasil perancangan, implementasi, pengujian, dan analisis ke dalam suatu laporan.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Dari hasil pengujian dan analisis yang telah dilakukan dapat diambil beberapa kesimpulan sebagai berikut:

1. Kemampuan *Information Retrieval* dengan menggunakan *Blind Relevance Feedback* dalam menghasilkan set dokumen yang relevan untuk user lebih baik dari *Information Retrieval* tanpa menggunakan *Relevance Feedback*.
2. Setelah dilakukan perbandingan performa antara *BRF* tipe 1,tipe 2 dan tipe 1+2, dapat disimpulkan bahwa BRF tipe 1+2 adalah BRF yang menghasilkan nilai *retrieval* terbaik.
3. BRF Tipe 1, Tipe 2 dan Tipe 1+2 dapat digunakan pada semua sistem information retrieval, dan dapat digunakan setiap saat untuk mendapatkan *query* yang lebih relevan.
4. Dalam setiap skenario pengujian, BRF selalu menunjukkan peningkatan hasil *retrieval* dokumen dibandingkan dengan *Information Retrieval* tanpa *Relevance Feedback*.

5.2. Saran

Saran yang dapat diberikan untuk melakukan pengembangan berikutnya antara lain sebagai berikut:

1. Untuk mendapatkan hasil yang lebih baik dalam *Blind Relevance Feedback*, dapat dilakukan pengujian dengan mengubah kombinasi parameter *Blind Relevance Feedback* yang digunakan.
2. Untuk mendapatkan hasil yang lebih baik dalam *Blind Relevance Feedback* tipe 2, perlu dilakukan pengujian lebih lanjut dengan menggunakan *Query Expansion Collection* yang berbeda-beda.

Daftar Pustaka

[1] Allan, James., *Incremental relevance feedback for information filtering*, Diunduh pada: http://eprints.kfupm.edu.sa/45645/1/45645.pdf, 15 oktober 2009

[2] He,Daqing and Peng,Yefei. *Comparing Two Blind Relevance Feedback Techniques, Diunduh pada :* http://www.sis.pitt.edu/~daqing/docs/pp109-he.pdf 15 oktober 2009

[3] Karls,Eberhard. *Relevance Feedback and Query Expansion*. Diunduh pada : http://www.sfs.uni-tuebingen.de/~parmenti/slides/slides9-1x4.pdf 15 oktober 2009

[4] Lin, Jimmy. Murray, G. Craig., *Assesing the term independence Assumption using Blind Relevance Feedback.* Diunduh pada : http://www.umiacs.umd.edu/~jimmylin/publications/Lin_Murray_SIGIR2005_poster.pdf 15 oktober 2009

[5] Manning,Christopher D., Raghavan Prabhakar, and Hinrich, Schutze, 2009, *an introduction to information retrieval*. United Kingdom, Cambridge University.

[6] Peng,Yefei and Mao,Ming. *Blind Relevance Feedback With Wikipedia : Enterprise Track. Diunduh pada:* http://trec.nist.gov/pubs/trec17/papers/yahoo.ent.NEW.pdf 21 oktober 2009

[7] R. Baeza-Yates and G.Navarro, 1999, *Modern Information Retrieval*. New York, Addison-Wesley.

[8] Salton, Gerard and Buckley, Christopher. *Improving retrieval performance by relevance feedback,* Diunduh pada : http://www.umiacs.umd.edu/~jimmylin/LBSC796-INFM718R-2006-Spring/papers/Salton90.pdf 15 oktober 2009

[9] Thomas M Cover and Joy A. Thomas., 1991, *Elements of Information Theory*. New York, John Wiley & sons.

[10] Wikipedia. *Discounted Cummulative Gain*. Diunduh pada: http://en.wikipedia.org/wiki/Discounted_cumulative_gain, 15 oktober 2009

[11] Wikipedia. *Information Retrieval*. Diunduh pada: http://en.wikipedia.org/wiki/Information_retrieval, 15 oktober 2009

[12] Wikipedia. *Relevance Feedback*. Diunduh pada: http://en.wikipedia.org/wiki/Relevance_feedback. 15 Oktober 2009

[13] Wikiperdia. *Wordnet.* Diunduh pada : http://en.wikipedia.org/wiki/WordNet, 8 Mei 2011.

[14] William B. Frakes and R. Baeza-Yates, 2007, *Information Retrieval Data Structures and Algorithm*. New York, Addison-Wesley.

[15] Z.Gu and M. Luo. *Comparison of using passages and documents for blind relevance feedback in information retrieval*. Diunduh pada: http://portal.acm.org/citation.cfm?id=1008992.1009081 15 oktober 2009