RELEVANCE FEEDBACK DENGAN ALGORITMA ROBERTSON AND SPARCK JONES PADA INFORMATION RETRIEVAL

Elvandi Gunata Siallagan[1], Yanuar Firdaus A.w.[2], Kusuma Ayu Laksitowening[3]

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

Abstrak

Semakin meningkatnya jumlah informasi pada masa kini menimbulkan permasalahan berupa menemukan dokumen yang relevan dengan kebutuhan secara tepat dan cepat. Sistem temu kembali dapat dibangun untuk menyelesaikan masalah tersebut.

Relevanve feedback merupakan teknik dimana pengguna melakukan feedback atas hasil pencarian dan menggunakan data feedback tersebut untuk memformulasikan query baru guna melakukan pencarian kembali. Data feedback yang digunakan pada algoritma Robertson dan Sparck Jones ialah dokumen relevan yang ditemukan pada pencarian sebelumnya. Algoritma ini melakukan query reweighting dan query expansion untuk membentuk query baru.

Algoritma ini dapat meningkatkan ataupun bahkan menurunkan performansi sistem seperti nilai precision, recall dan IAP yang diperoleh pada pencarian awal. Hasil terbaik penerapan algoritma dalam nilai precision ini ialah meningkatkan 232.35% nilai precision pencarian awal sedangkan hasil terburuknya ialah menurunkan -39.8% nilai precision pencarian awal. Hasil terbaik penerapan algoritma ini dalam nilai recall ialah meningkatkan 11.38 % nilai recall pencarian awal dan hasil terburuk ialah menurunkan 23.2%. Hasil terbaik penerapan algoritma ini dalam nilai IAP ialah meningkatkan 174.8% nilai IAP pencarian awal sedangkan hasil terburuknya ialah menurunkan 16.67% nilai IAP pencarian awal. Dengan algoritma ini, waktu pencarian akan semakin meningkat sesuai jumlah dokumen feedback dan penambahan term query.

Kata Kunci : information retrieval, relevance feedback, Robertson dan Sparck Jones,

Abstract

Growing of amount of informations at present generates problem in the form of finding relevant document to the need accurately and quickly. Information retrieval system can assist to solve this problem.

Relevance feedback is a technique where user do feedback and using these feedback for reformulate new query which will used for new searching. The document feedback used in Robertson and Sparck Jones algorithm is the relevant document which found. This algorithm do query reweighting and query expansion to form new query.

This algorithm can increase and reduce the searching performance like precision, recall, and IAP value which obtained at initial searching. The best result of precision value caused this algorithm is increasing 232.35% precision value of inital searching while the worst result is reduce -39.8% precision of initial searching. The best result of recall value caused this algorithm is increasing 11.38% recall value of inital searching while the worst result is reduce 23.2% recall. The best result of IAP value caused this algorithm is increasing 174.8% recall IAP of inital searching while the worst result is reduce 16.67% IAP. With this algorithm, searching time will increase fit to amount of document feedback and the expansion of term query.

Keywords : information retrieval, relevance feedback, Robertson dan Sparck

1. PENDAHULUAN

1.1. Latar Belakang Masalah

Jumlah informasi yang sangat banyak menimbulkan permasalahan yakni sulitnya menemukan informasi yang dibutuhkan secara tepat dan cepat. Dengan jumlah informasi yang sangat banyak tidak memungkinkan bagi seseorang mencari dokumen yang relevan dengan kebutuhannya secara manual yakni dengan membaca satu persatu dokumen untuk menemukan dokumen yang sesuai atau relevan terhadap kebutuhannya. Pencarian dokumen dengan cara seperti ini membutuhkan waktu yang lama sebanding dengan jumlah dokumen yang menjadi sumber pencarian.

Sistem temu kembali informasi atau *information retrieval system* merupakan sistem untuk menemukan kembali dokumen-dokumen yang sesuai dengan kebutuhan pengguna. Dengan menginputkan topik mengenai informasi yang dibutuhkannya, pengguna dapat menemukan dokumen yang relevan terhadap kebutuhannya tersebut. Topik yang menggambarkan kebutuhan pengguna sering disebut sebagai *keywords* atau *query*. Berdasarkan *query* tersebut, maka sistem temu kembali akan berusaha menemukan dan menampilkan dokumen-dokumen yang relevan sesuai dengan *query* tersebut. Adapun hasil pencarian tersebut akan ditampilkan secara berurut berdasarkan tingkat kerelevanannya terhadap *query*.

Namun ada kalanya pengguna tidak mampu mengekspresikan informasi yang dibutuhkannya dalam bentuk *query* bahkan ada kalanya pengguna tidak mengetahui informasi apa saja yang terdapat dalam sumber pencarian [12]. Untuk mengatasi hal tersebut pengguna dapat menandai atau memilih contoh jenis dokumen yang dibutuhkannya. Pemilihan dokumen ini dapat dianggap sebagai ekspresi kebutuhan user. Pemilihan dokumen ini dapat dilakukan saat pengguna telah disediakan hasil pencarian atas *query* sebelumnya. Dengan contoh jenis dokumen tersebut, sistem dapat menggunakannya untuk mencari dokumen-dokumen lain yang sejenis untuk memenuhi kebutuhannya. Hal inilah yang memulai timbulnya *relevance feedback.*

Secara sederhana *relevance feedback* ialah proses pengguna menandai atau memilih dokumen yang relevan terhadap kebutuhannya dan menginputkannya kembali ke *information retrieval system* [12]. Data yang terdapat di dokumen yang difeedbackkan oleh pengguna tersebut dapat digunakan untuk mencari dokumen lain yang relevan bahkan dapat memperbaiki urutan hasil pencarian sebelumnya. Data yang terdapat di dokumen yang difeedbackkan tersebut digunakan untuk mengubah *query* sebelumnya yang diinputkan pengguna. Dengan terbentuknya query baru, maka pencarian ulang dapat dilakukan kembali.

Modifikasi terhadap *query* lama ini dapat dilakukan dengan menambah kata baru pada *query* lama, melakukan pembobotan ulang terhadap *query* lama. Perbaikan kata kunci dengan penambahan kata dan pembobotan ulang kata kunci dapat dilakukan dengan menggunakan algoritma Robertson and Sparck Jones. Algoritma ini dapat menghasilkan *query* baru dengan menggunakan data *feedback* dari pengguna dan *query* lama. Query baru tersebut akan digunakan untuk pencarian ulang.

1.2. Perumusan Masalah

Permasalahan yang dijadikan sebagai objek penelitian pada tugas akhir ini adalah:

1. Bagaimana menemukan dokumen yang relevan dengan menggunakan *relevance feedback.*
2. Bagaimana pengaruh *relevance feedback* terhadap performansi hasil pencarian dokumen yang relevan.
3. Bagaimana menganalisis performansi dari penerapan *relevance feedback* dengan algoritma Robertson dan Sparck Jones pada *information retrieval*.

Batasan masalah dari tugas akhir ini adalah:

1. Koleksi dokumen dan *query* yang digunakan adalah dokumen berbentuk teks dalam bahasa Inggris.
2. Koleksi dokumen yang menjadi objek pencarian ialah koleksi dokumen yang diambil dari ftp://ftp.cs.cornell.edu/SMART. Koleksi dokumen yang digunakan yakni: ADI, CISI, CRAN, dan MED. Pada dokumen tersebut sudah terdapat kumpulan *query* dan *relevance judgement* untuk setiap *query*.

1.3. Tujuan

Tujuan pembuatan tugas akhir ini adalah:

1. Merancang dan membangun perangkat lunak untuk menerapkan *relevance feedback* dengan algoritma Robertson dan Sparck Jones pada *information retrieval*.
2. Menganalisis performansi relevance feedback dengan algoritma Robertson and Sparck Jones pada information retrieval yang diterapkan pada perangkat lunak.

1.4. Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang digunakan dalam menyelesaikan penelitian ini adalah:

1. Studi literatur

   Langkah ini bertujuan untuk memahami dasar teori mengenai *information retrieval*, *relevance feedbback*, dan algoritma Robertson dan Sparck Jones serta hal lain yang mendukung penyelesaian tugas akhir ini. Sumber dasar teori dapat berupa buku, *paper*, maupun halaman web.
2. Analisis kebutuhan perangkat lunak

   Yaitu melakukan analisis kebutuhan perangkat lunak yang akan dibangun, agar didapatkan gambaran umum seperti apa perangkat lunak yang ingin dibangun.
3. Perancangan perangkat lunak

   Berdasarkan kebutuhan yang telah diidentifikasi, maka dapat dirancang perangkat lunak yang sesuai untuk memenuhi kebutuhan. Rancangan perangkat lunak dapat menjadi panduan saat implementasi perangkat lunak.
4. Implementasi

   Pada tahapan ini dilakukan pembangunan perangkat lunak yang telah dirancang dengan menggunakan teknik pemrograman tertentu. Pada

tahapan ini dibangun sistem yang dapat menangani proses *indexing* terhadap data dokumen yang menjadi objek pencarian, proses formulasi *query*, pencarian dan pengembalian dokumen yang relevan. Kemampuan untuk melakukan proses *relevance feedback* juga diimplementasikan pada sistem.

5. Pengujian dan analisis hasil.
   Pengujian dilakukan untuk memperhatikan pengaruh *relevance feedback* terhadap infromation *retrieval*. Pada proses ini akan dibandingkan nilai *recall*, *precision* dan *Interpolated Average Precision* (IAP) dan waktu yang dibutuhkan sebelum dan setelah proses *relevance feedback*. Pengujian ini akan dilakukan berulang kali dengan jumlah *feedback* yang berbeda untuk menentukan parameter jumlah *feedback* yang baik dalam meningkatkan performansi sistem.
6. Penyusunan laporan tugas akhir.

5. Penutup

Pada bab ini akan disimpulkan hasil pengerjaan Tugas Akhir ini dan beberapa saran yang dapat mengembangkan Tugas Akhir ini.

5.1. Kesimpulan

Berdasarkan analisis terhadap hasil pengujian perangkat lunak dalam Tugas Akhir ini dapat dihasilkan beberapa kesimpulan, yakni:

1. Penggunaan algoritma Robertson dan Sparck Jones dapat diterapkan pada *relevance feedback* di *information retrieval*.
2. *Relevance feedback* dengan menggunakan algoritma Robertson dan Sparck Jones tidak selalu meningkatkan nilai *precision* yang diperoleh saat pencarian awal. Hal ini dibuktikan bahwa pada koleksi data Cran dan Med, pengaruh *relevance feedback* tidak selalu menghasilkan nilai *precision* yang lebih besar daripada nilai *precision* pada pencarian awal. Namun dengan memperhatikan jumlah penurunan dan peningkatan nilai *precision* akibat relevance *feedback*, maka disimpulkan bahwa relevance *feedback* lebih banyak atau hampir secara keseluruhan mampu meningkatkan nilai *precision*.
3. Penambahan jumlah *term* yang berbeda pada *query expansion* menghasilkan nilai *precision* yang berbeda pada saat menggunakan jumlah dokumen *feedback* yang sama. Walaupun menggunakan jumlah dokumen *feedback* yang sama, namun penggunaan jumlah *term* yang lebih banyak mengakibatkan nilai *precision* lebih kecil dibandingkan penggunaan *term* yang lebih sedikit.
4. Penggunaan algoritma Robertson dan Sparck Jones pada *relevance feedback* tidak selalu menghasilkan nilai *recall* yang lebih baik dibandingkan nilai *recall* pencarian awal. Hal ini terbukti pada saat menggunakan dokumen *feedback* sebanyak 1 dokumen, semua nilai *recall*nya lebih kecil daripada *recall* pencarian awal. Nilai *recall* yang lebih baik daripada pencarian awal akan dihasilkan pada saat penggunaan dokumen *feedback* sebanyak diatas 5 buah.
5. Penambahan jumlah *term* yang berbeda pada *query expansion* menghasilkan nilai *recall* yang berbeda pada saat menggunakan jumlah dokumen *feedback* yang sama. Walaupun menggunakan jumlah dokumen *feedback* yang sama, tetapi penggunaan jumlah *term* yang lebih banyak ternyata menghasilkan nilai *recall* yang lebih baik daripada penambahan jumlah *term* yang sedikit.
6. Algoritma Robertson dan Sparck Jones yang digunakan pada *relevance feedback* tidak selalu menghasilkan nilai IAP yang lebih baik dibandingkan hasil *initial searching*. Nilai IAP hasil *relevance feedback* akan terjamin lebih besar daripada nilai IAP hasil *initial searching* saat digunakan dokumen *feedback* lebih besar dari 5 buah dokumen.
7. Penggunaan jumlah *term* yang berbeda pada *query expansion* akan menghasilkan nilai IAP yang berbeda walaupun menggunakan jumlah dokumen *feedback* yang sama. Walaupun menggunakan jumlah dokumen *feedback* yang sama, namun *relevance feedback* dengan menggunakan

jumlah *term* yang lebih banyak menghasilkan nilai IAP yang lebih tinggi daripada menggunakan jumlah *term* yang lebih sedikit.

8. Proses *relevance feedback* tidak selalu membutuhkan waktu yang lebih banyak dibandingkan waktu *initial searching*. Waktu yang dibutuhkan untuk pencarian akan meningkat saat digunakan jumlah *term* yang lebih banyak dibandingkan jumlah *term* yang sedikit walaupun menggunakan jumlah dokumen *feedback* yang sama.

5.2. Saran

Untuk mengembangkan Tugas Akhir ini, maka berikut beberapa saran yang dapat diperhatikan.

1. Sumber pencarian atau koleksi dokumen yang digunakan agar tidak hanya dokumen file teks berformat .txt.
2. Perangkat lunak mampu menangani jika pengguna melakukan pencarian dokumen relevan dengan beberapa kali *feedback*.
3. Jika pada perangkat lunak ini, *term* baru yang akan ditambahkan pada *query expansion* dipilih pada peringkat teratas secara otomatis. Namun hal ini dapat dikembangkan jika perangkat lunak dapat menangani pengguna yang ingin memilih *term* baru pada *query expansion* secara manual.

Daftar Pustaka

[1] Berry, Michael W. and B. Murray, 1999, *Understanding Search Engine:Mathematical Modeling and Text Retrieval,* University of Tennessee.

[2] Buckley, C. (et.al.), The Effect of Adding Relevance Information in a Relevance Feedback Environment, In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 292-300, 1994. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.7537.pdf . Didownload pada tanggal 15 Januari 2009.

[3] Fuhr, Norbert, *Probabilistic Models in Information retrieval*.

[4] Greengrass, E., 2000, Information Retrieval: A Survey. http://clgiles.ist.psu.edu/IST441/materials/texts/IR.report.120600.book.pdf. Didownload pada tanggal 15 Januari 2009.

[5] Hiemstra, D. and Stephen Robertson, Relevance Feedback for Best Match *Term* Weighting Algorithms in Information Retrieval. http://www.ercim.org/publication/ws-proceedings/DelNoe02/hiemstra.pdf. Didownload pada tanggal 15 Januari 2009.

[6] Jones, K. Sparck (et.al), 2000, A probabilistic model of information retrieval:development and comparative experiments. London

[7] Manning, Christopher D. (et.al.), 2008, *Introduction to Information Retrieval,* USA: Cambridge University Press.

[8] Robertson, S.E. (et.al), Okapi at Trec 3, London.

[9] Robertson, S.E. and K. Sparck Jones, 1976, Relevance Weighting of Search *Terms*, In *Journal of the American Society for Information Science*, 27, 129-146.

[10] Robertson, S.E., and K. Sparck Jones, Simple, proven approaches to text retrieval. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-356.pdf. Didownload pada tanggal 15 Januari 2009.

[11] Rocchio J. J., 1971, Relevance Feedback in Information Retrieval, In Salton, G. (Ed.): *The SMART Retrieval System. Experiments in Automatic Document Processing*,313-323, New Jersey: Prentice Hall.

[12] Ruthven, Ian and Mounia Lalmas M., A survey on the use of relevance *feedback* for information access system. http://inex.is.informatik.uni-duisburg.de:2004/pdf/ker_ruthven_lalmas.pdf. Didownload pada tanggal 15 Januari 2009.

[13] Salton, Gerald and Chris Buckley, Improving Retrieval Performance by Relevance Feedback, In *Journal of The American Society for Information Science*, 41 (4), 288-297, 1990.

http://users.cs.fiu.edu/~vagelis/classes/COP6776/publications/jasistSalton1990.pdf. Didownload pada tanggal 15 Januari 2009.

[14] Singhal, Amit, Modern Information Retrieval: A Brief Overview. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.7537.

Didownload pada tanggal 15 Januari 2009.

[15] van Rijsbergen, C.J., 1979, *Information Retrieval*, University of Glasgow.

[16] Vinay, Vishwa (et.al.), Evaluating Relevance Feedback Algorithms for Searching on Small Displays, In D.E. Losada and J.M. Fernandez-Luna (Eds.): *ECIR 2005, LNCS 3408*, 185-199, 2005. http://www.adastral.ucl.ac.uk/~icox/papers/2005/ECIR2005.pdf. Didownload pada tanggal 15 Januari 2009.

[17] Vinay, V. (et.al.), Comparing Relevance Feedback for Web Search.

[18] Zhai, ChengXiang, 2007, A Brief Review of Information Retrieval Models. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.89.5804. Didownload pada tanggal 15 Januari 2009