IMPLEMENTASI METODE DYANMIC WINDOW BASED PADA INFORMATION RETRIEVAL SSYTEMS

Panji Omara[1], Kusuma Ayu Laksitowening[2], Sri Widowati[3]

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

Abstrak

Pada saat ini dimana jumlah dokumen yang tersedia sangat besar, pencarian secara manual dapat dilakukan dengan membaca setiap dokumen pada koleksi dokumen untuk mendapatkan dokumen yang tepat dan sesuai kebutuhan. Namun, pencarian seperti itu membutuhkan waktu yang lama jika jumlah dokumen sangat banyak. Dan Information Retrieval Systems dapat memecahkan masalah tersebut.

Proses dalam Information Retrieval dapat digambarkan sebagai sebuah proses untuk mendapatkan relevant documents dari collection documents yang ada melalui pencarian query yang diinputkan user. Sistem information retrieval yang baik adalah sistem information retrieval yang mampu mengambil semua dokumen relevan dan kemudian mengurutkan dokumen relevan tersebut pada urutan teratas. Agar diperoleh sistem dengan dokumen relevan berada pada urutan teratas dibutuhkan suatu metode perangkingan dengan menggunakan metode penghitungan similiarity score yang efektif dalam menentukan nilai similiarity score suatu dokumen. Dalam tugas akhir ini akan digunakan metode Dynamic Window Based pada penghitungan nilai similiarity score nya .

Analisa yang dilakukan pada tugas akhir ini adalah membandingkan performansi dari metode Dynamic Window Based dan metode Probabilistik. Adapun untuk keakuratan sistem dalam mengambil dokumen yang relevan dapat dilihat dari nilai Precision,Recall sedangkan untuk kemampuan sistem dalam mengambil dokumen yang relevan dapat dilihat dari nilai IAP yang didapatkan. Menurut pengujian yang dilakukan performansi sistem dengan menggunakan metode Dynamic Window Based lebih baik dibandingkan dengan Probabilistik asalkan di dalam dokumen tersebut terdapat term conjoint yang akan membuat nilai similiarity score maksimal. Perubahan lebar window pada metode Dynamic Window Based juga akan berpengaruh terhadap hasil performansi sistemnya.

Kata Kunci : Information Retrieval System, Dynamic Window Based, similiarity score, Probabilistik,Precision, Recall, IAP.

Abstract

Nowadays where amount of document is huge, manual searching could be possible to do by read ing on by one document in document collection due to get document which is we searching for. But, if we do that way it needs a lot of time if document itself is huge. Information retrieval system can assist to solve this problem.

Process in information retrieval can be figure as a process to get relevant document from document collection by seraching the query which is input to systems by user. Good information retrieval systems is the systems that can get all relevant document and the result was ranked in top chart. To get good systems all we need is effective rank method when calculating similiarity score a document. In this final project will be using Dynamic Window Based method in calculating similiarity score.

The analysis conducted in this thesis is to compare the performance of the method of Dynamic Window-Based and Probabilistic methods. As for the accuracy of the system in taking the relevant documents can be seen from the value of Precision, Recall while for the system's ability to take the relevant documents can be seen from the IAP value obtained. According to tests performed using the method of system performance with the Dynamic Window-Based Probabilistic better than those provided in the document there are terms that will create value conjoint similiarity maximum score. Change window width Dynamic Window Based on the method will also influence the performance of the system.

Keywords : Information Retrieval System, Dynamic Window Based, similiarity score, Probabilistic,Precision, Recall, IAP.

8. 1. Pendahuluan

1.2 1.1 Latar Belakang Masalah.

Pada saat ini dimana jumlah dokumen yang tersedia sangat besar, pencarian secara manual dapat dilakukan dengan membaca setiap dokumen pada koleksi dokumen untuk mendapatkan dokumen yang tepat dan sesuai kebutuhan. Namun, pencarian seperti itu membutuhkan waktu yang lama jika jumlah dokumen sangat banyak. Salah satu solusi untuk mengatasi masalah tersebut adalah dengan menggunakan *Information Retrieval System*. Proses dalam *Information Retrieval* dapat digambarkan sebagai sebuah proses untuk mendapatkan *relevant documents* dari *collection documents* yang ada melalui pencarian *query* yang diinputkan *user*[2]. Sedangkan *query* dalam *Information Retrieval* merupakan kata kunci berupa satu atau beberapa term yang akan dicari. *Information retrieval systems* yang baik adalah sistem yang mampu mengambil dokumen relevan dan kemudian mengurutkan dokumen relevan tersebut pada urutan teratas[2]. Agar diperoleh sistem yang dapat mengambil dokumen yang relevan dan mengurutkan pada urutan teratas dibutuhkan suatu metode *matching* suatu *query* dengan koleksi dokumen menggunakan metode yang optimal dalam menentukan nilai *similiarity score* suatu dokumen.

Metode yang digunakan dalam proses penghitungan nilai *similiarity score* saat ini sebagian besar hanya memperhitungkan kemunculan suatu *term* dalam dokumen namun tidak memperhitungkan jarak antara term yang muncul dalam dokumen. Padahal jarak antar *term query* didalam dokumen sangat menentukan tingkat *similiarity* antara dokumen dengan *query* terutama bila term tersebut berbentuk frasa. Jarak *term* satu dengan *term query* yang lain yang jauh dapat dikatakan tidak bisa menunjukan arti yang jelas. Biasanya di dalam *query* terdiri dari beberapa *term query*, dimana semua *term query* tersebut menunjukan sebuah arti[11]. Misal terdapat *query "brain cancer dangerous"* dan dalam dokumen 1 terdapat *term "brain cancer"* dalam dokumen 2 terdapat *term "cancer doesn't have negative effect to brain"* maka dari contoh diatas didapatkan dokumen 1 dengan jarak antar *term* yang kecil lebih relevan dibanding dengan dokumen 2 yang mempunyai jarak antar *term* lebih besar. *Term* yang muncul dengan jarak berdekatan dalam dokumen berkontribusi lebih besar terhadap nilai *similiarity* dibandingkan dengan *term* yang muncul dengan jarak yang berjauhan. Semakin dekat jarak suatu *term* dalam dokumen maka semakin besar nilai *similiarity* antara *query* dan dokumen[11].

Oleh sebab itu dalam Tugas Akhir ini akan diimplementasikan dan dianalisis performansi dari penerapan metode *Dynamic Window Based Method* dalam *IR System*.

Dalam menganalisa hasil penerapan dari metode *Dynamic Window Based* akan digunakan metode yang lain yaitu Probabilistik sebagai pembanding.

1.3 1.2 Perumusan Masalah.

Rumusan masalah dari tugas akhir ini adalah :

a) Bagaimana menerapkan metode *Dynamic Window Based* dalam *information retrieval systems*? Dalam menganalisis performansi nya akan digunakan metode Probabilistik sebagai pembanding.

b) Bagaimanakah perbandingan performansi sistem dengan metode *Dynamic Window Based* dibandingkan dengan metode Probabilistik?
c) Bagaimana pengaruh ukuran lebar *window* terhadap performansi *information retrieval system*.

1.4 1.3 Batasan Masalah.

Adapun batasan masalah dari tugas akhir ini adalah :

a) Penelitian pada Tugas Akhir ini berfokus pada pengimplementasian metode *Dynamic Window Based* pada IRS serta analisis performansinya dengan menggunakan parameter *Precision, Recall*, serta *IAP.*
b) Batasan lebar *window* yang digunakan adalah lebih besar dari 1 dan kurang dari sama dengan 20, karena rata-rata panjang *query* yang ada tidak melebihi 20 *term*.
c) Mengasumsikan bahwa hasil dari proses *preprocessing* yang digunakan sudah benar.
d) Perhitungan nilai performansi sistem dibatasi/dihitung hanya pada 20 dokumen yang ter-*retrieved* pertama, karena diharapkan performansi sistem akan lebih terlihat untuk masing-masing metode.
e) Koleksi dokumen yang digunakan dalam tugas akhir ini yang berupa kumpulan dokumen dan *query* menggunakan bahasa Inggris. Koleksi dokumen yang digunakan berbentuk *.txt dan dokumen yang digunakan merupakan dokumen yang termasuk *unstructured document*. Dalam Tugas Akhir ini digunakan koleksi dokumen yang berasal dari ftp.cs.cornell.edu/pub/smart . Di dalam koleksi dokumen yang digunakan terdapat kumpulan dokumen dan *query* nya beserta *relevance judgment* untuk tiap-tiap dokumen. *Query* yang digunakan merupakan *simple query*.

1.5 1.4 Tujuan.

Tujuan dari tugas akhir ini adalah :

a) Mengimplementasikan metode *Dynamic Window Based* pada *information retrieval systems*.
b) Menganalisis performansi penggunaan metode *Dynamic Window Based* pada *information retrieval system.* Dalam menganalisa performansi hasil penerapan dari metode *Dynamic Window Based* akan digunakan metode yang lain yaitu metode Probabilistik sebagai pembanding.
c) Menganalisis pengaruh ukuran lebar *window* terhadap performansi *information retrieval systems* yang diukur dengan menggunakan parameter F-Measure dan IAP.

1.6 1.5 Metodologi Penyelesaian Masalah.

1.7 Metodologi penyelesaian masalah yang akan digunakan adalah :

a. Studi literatur

Merupakan tahapan dalam mempelajari konsep dan teori pendukung untuk memecahkan permasalahan. Pencarian sumber dan referensi berupa buku, makalah, jurnal dan media yang lain seperti internet yang berhubungan dengan konsep *infromation retrieval*, dan metode *Dynamic Window Based* , serta informasi lainnya yang menunjang pembuatan tugas akhir ini.

b. Pengumpulan data

Pengumpulan dokumen yang akan dijadikan sebagai data set untuk analisis dan pengimplementasian metode *Dynamic Wndow Based* dalam *information retrieval system.*

c. Perancangan Sistem.

Mempersiapkan data set yang akan digunakan dalam proses *matching* menggunakan metode *Dynamic Window Based*. Kemudian dilakukan perancangan sistem yang nantinya akan dilakukan proses implementasi metode *Dynamic Window Based* pada IRS.

d. Implementasi.

Tahapan implementasi ini akan dilakukan proses pengimplementasian metode *Dynamic Window Based* pada IRS. Di tahap ini terdiri dari :

- Dokumen uji yang telah disiapkan sebelum dilakukan proses perhitungan nilai *similiarity score* yaitu dilakukan proses *tokenisasi, stopword removal, stemming, term weighting* terlebih dulu.
- Menerapkan metode *Dynamic Window Based* dalam IRS untuk perhitungan nilai *similiarity score* dengan mengaplikasikannya kedalam kode program.

e. Testing dan analisis Hasil.

Dilakukan proses pengujian terhadap sistem yang sudah dibangun apakah sistem sudah sesuai dengan yang diharapkan.

- Untuk menganalisis performansi metode *Dynamic Window Based* diinputkan suatu *query* ke dalam sistem kemudian nilai *similiarity* terhadap dokumen yang didapatkan dihitung nilai *precision, recall, dan IAP* nya. Kemudian dilakukan prosedur yang sama dengan lebar *window* yang berbeda, untuk mengetahui pengaruh lebar *window* terhadap performansi sistem. Hasil pengujian dianalisis berdasarkan parameter *precision, recall, dan IAP*. Dari hasil analisis tersebut diambil kesimpulan mengenai penerapan metode *Dynamic Window Based* pada *IR System*.

1.8 1.6 Sistematika Penulisan.

Sistematika Penulisan Tugas Akhir ini terdiri dari lima bab dengan disertai lampiran terkait pelaksanaan tugas akhir 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 sistem.

BAB II Dasar 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.

BAB IV Implementasi, Pengujian dan Analisis

Bab ini membahas implementasi detail sistem dan pengujian terhadap sistem.

BAB V Kesimpulan dan Saran

Berisi tentang kesimpulan dan saran sebagai hasil dari analisis dan implementasi Tugas Akhir.

9.

13.5. Kesimpulan dan Saran

1.16 5.1 Kesimpulan.

Kesimpulan yang didapatkan dari pengujian dan analisa terhadap *information retrieval system* yang dibangun adalah :

1. Nilai performansi sistem pada semua *dataset* tertinggi didapatkan dengan menggunakan metode *dynamic window based* dibandingkan metode probabilistik. Hal ini disebabkan karena metode *dynamic window based* selain memperhitungkan kemunculan *term query* dalam dokumen namun juga memperhatikan jarak antar *term query* dalam dokumen. Sehingga dengan menggunakan metode *dynamic window based* sistem mampu me-

retrieved dokumen yang relevan pada urutan teratas lebih baik yang menyebabkan nilai performansi yang didapatkan tinggi.

2. Perubahan lebar *window* pada metode *dynamic window based* akan berpengaruh terhadap performansi sistem, yang diukur dengan menggunakan parameter *precision, recall,* serta *IAP*. Performansi sistem akan semakin tinggi jika dokumen yang relevan ter-*retrieved* pada urutan teratas. Semakin besar nilai *similiarity score* suatu dokumen maka dokumen tersebut akan ter-*retrieved* pada urutan teratas. Nilai *similiarity score* suatu dokumen akan semakin tinggi jika lebar *window* yang digunakan lebarnya sama atau hampir sama dengan ukuran persebaran *term* yang *conjoint* serta lebar *window* mampu mencakup *term query* yang tidak *conjoint* dalam *window* tersebut sehingga persebarannya tidak terlalu besar jaraknya, karena akan menyebabkan nilai *tightwin* besar, dengan semakin besar nilai *tightWin* maka nilai *similiairty score* akan semakin kecil.
3. Kekurangan pada system dengan menggunakan metode *dynamic window based* ini adalah ketika di dalam dokumen tersebut terdapat *term-term query* yang conjoint namun *term-term query* tersebut bukan berupa frasa sehingga mempengaruhi kerelevanan suatu dokumen terhadap *query*.

1.17 5.2 Saran.

1. Perlu dilakukan tambahan proses klasifikasi *term query* mana saja yang merupakan inti dari *query* pencarian yang ada didalam proses penghitungan *similiarity score* nya. Hal ini dilakukan agar dokumen yang ter-retrieved benar-benar merupakan dokumen yang relevan terhadap query, sehingga nilai performansi yang didapatkan akan semakin maksimal/lebih tinggi.
2. Dapat menggunakan kumpulan dokumen berformat lain, sehingga tidak terbatas pada dokumen dengan format .txt.
3. Dapat menggunakan karakteristik *query* yang berbeda tidak terbatas pada *simple query*.
4. Dapat mencoba membandingkannya dengan metode yang lain selain metode probabilistic.

14.DAFTAR PUSTAKA

[1] Chipaila. Kisesa. Amuel.2006. *Information Retrieval Using Relevance Feedback (Rocchio Technique).*

[2] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/information-retrieval-book.html Diakses 20 Maret 2010 pukul 15.00 WIB

[3] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[4] 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.

[5] Hyusein, Byurhan, Patel, Ahmad (2003) Web Document Indexing and Retrieval, LNCS 2588 pp. 573-579, Springer Verlag Berlin

[6] Manning, Christopher D, Ragnavan Prabhakar, Schutze, Hinrich (2008) *Introduction to Information Retrieval*, Cambridge University Press)

[7] Murad, Azmi MA., Martin, Trevor. (2007) Word Similarity for Document Gouping using Soft Computing. IJCSNS International Journal of Computer
Science and Network Security, Vol.7 No.8, August 2007, pp. 20- 27

[8] Kenneth W.Church, William A.Gale, (1995) *Inverse Document Frequency(IDF): A Measure of Deviations from Poisson*. AT&T Bell Laboratories.

[9] Liu,Yiqun; Wang Canhui; Zhang Min; Ma Sahoping.2004. *Finding Abstract Field of Web Pages and Query Specific Retrieval*.
http://trec.nist.gov/pubs/terc13/papers/tsinghua-ma.web.pdf
Diakses 9 Maret 2010 pukul 20.00 WIB.

[10] Peter Yeung**.** *Weighting Document Genre in Enterprise Search*
http://plg1.cs.uwaterloo.ca/~p2yeung/peter_yeung.pdf.
Diakses pada 8 Maret 2010 pukul 23.32 WIB

[11] Qianli Jin;Jun Zhao; Bo Xu.2005.*Window-based Method for Information Retrieval.*
www.nlpr.ia.ac.cn/2005papers/gjhy/gh74.pdf
Diakses 9 Maret 2010 pukul 20.05

[12] Ramos, Juan. (2000). *Using TF-IDF to Determine Word Relevance in Document Queries.* Department of Computer Science, Rutgers University.
www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424...
Diakses 20 Maret 2010 pukul 15.00 WIB

[13] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley,1999.

[14] Salton, G. And C. Buckley.(1988). *Term-weighting approaches in automatic text retrieval*. Information Processing & Management.

[15] Warren R. Greiff. (1998). *A theory of term weighting based on exploratory data analysis*. In Proceedings of SIGIR-98.

[16] Zuckley, C. (et.al.), Evaluating Evaluation Measure Stability. Di download pada tanggal 14 Desember 2010