IMPLEMENTASI INDEX COMPRESSION MENGGUNAKAN VARIABLE BYTE CODE

Adelino Thesaria[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. Berbagai pendekatan untuk meningkatkan performansi Information Retrieval (IR) telah dilakukan. Salah satu cara untuk meningkatkan performansi tersebut adalah dengan kompresi indeks (index compression). Dua jenis teknik kompresi pada Information Retrieval System yaitu lossy compression dan lossless compression. Pada lossless compression semua informasi tetap terjaga, contohnya posting file compression.

Indeks merupakan bagian paling utama dalam Information Retrieval System. Tujuan menyimpan indeks adalah untuk mengoptimalkan kecepatan dan kinerja dalam menemukan dokumen yang relevan untuk permintaan pencarian. Tanpa indeks, mesin pencari akan memindai setiap dokumen, yang akan memerlukan banyak waktu dan daya komputasi. Namun, semakin banyaknya dokumen yang masuk semakin bertambah juga kapasitas indeks. Index Compression adalah teknik yang digunakan untuk lebih mengefisienkan indeks, baik dari kapasitas maupun performansi dari Information Retrieval System. Dengan mengkompresi indeks, dapat mengurangi kapasitas yang digunakan hingga 75%. Index Compression juga dapat meningkatkan kecepatan transfer dari disk ke memori.

Dalam tugas akhir ini, akan dilakukan implementasi Index Compression menggunakan Variable Byte Code. Variable byte code merupakan salah satu teknik dalam kompresi indeks yang diterapkan pada Information Retrieval guna mengurangi kapasitas disk yang terpakai dan pemakaian waktu pencarian yang lebih cepat. Oleh karena itu Diharapkan setelah menggunakan Index Compression menggunakan Variable Byte Code, kapasitas indeks akan berkurang dan performansi dari Information Retrieval System meningkat.

Kata Kunci : Information Retrieval, Information Retrieval System, Index Compression, lossless compression, Variable Byte Code.

Abstract

Information Retrieval (IR) is part of computer science related to the retrieval of information from documents that are based on the content and context of the documents themselves. Processes in Information Retrieval can be described as a process to obtain relevant documents from the collection of documents through search queries entered by users. Various approaches to improve the performance of Information Retrieval (IR) has been performed. One way to improve performance is to compress the index (index compression). Two types of compression techniques on Information Retrieval System is lossy compression and lossless compression. In lossless compression of all information will be maintained, for example, posting the file compression.

Index is the most important part in the Information Retrieval System. The aim is to save the index to optimize the speed and performance in finding relevant documents for the search query. Without indexes, search engines will scan every document, which will require much time and computing power. However, a growing number of incoming documents also increases the capacity of the index. Index Compression is a technique used to minimize the index, both of capacity and performance of Information Retrieval System. By compressing the index, can reduce the capacity used up to 75%. Index Compression can also increase transfer speed from disk to memory.

In this thesis, will be implemented using Variable Compression Index Byte Code. Variable byte code is one of the techniques applied in the compression index on Information Retrieval in order to reduce the unused disk capacity and usage of a faster search time. Therefore it is expected that after using the Index Compression using Variable Byte Code, the capacity will be reduced and the performance index of the Information Retrieval System to increase.

Keywords : Information Retrieval, Information Retrieval System, Index Compression, lossless compression, Variable Byte Code.

1. PENDAHULUAN

1.1 Latar Belakang

Pada saat ini, kebutuhan Internet pada setiap individu terus meningkat. Hal ini disebabkan banyaknya fasilitas dan kemudahan yang ditawarkan dari dunia Internet itu sendiri, salah satunya adalah informasi. Setiap orang membutuhkan informasi dengan cepat dan tepat. Dulu kita dapat mengelompokkan informasi dengan manual, menyusunnya berdasar judul, pengarang, tanggal, dan sebagainya, kemudian dicari satu-persatu. Namun apa yang terjadi jika informasi yang ada jumlahnya ribuan, ratusan ribu, bahkan jutaan. Untuk itu dikembangkanlah suatu sistem yang memudahkan kita, suatu sistem yang dapat bekerja secara otomatis, yang disebut dengan sistem *Information Retrieval*.

Information Retrieval (IR) yang memiliki arti temu kembali informasi, merupakan bagian dari *computer science* yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Setiap proses dalam *Information Retrieval* dapat digambarkan sebagai sebuah proses untuk mendapatkan *relevant documents* dari *document collections* melalui pencarian *query* yang diinputkan user. Dari setiap informasi yang dimasukkan akan dikumpulkan di dalam indeks untuk memudahkan pencarian.

Indeks merupakan bagian paling utama dalam *Information Retrieval System.* Tujuan menyimpan indeks adalah untuk mengoptimalkan kecepatan dan kinerja dalam menemukan dokumen yang relevan untuk permintaan pencarian. Tanpa indeks, mesin pencari akan memindai setiap dokumen, yang akan memerlukan banyak waktu dan daya komputasi. Namun, semakin banyaknya dokumen yang masuk semakin bertambah juga kapasitas indeks. *Index Compression* adalah teknik yang digunakan untuk lebih mengefisienkan indeks, baik dari kapasitas maupun performansi dari *Information Retrieval System*. Dengan mengkompresi indeks, dapat mengurangi kapasitas yang digunakan hingga 75%. *Index Compression* juga dapat meningkatkan kecepatan transfer dari disk ke memori.

Dalam tugas akhir ini, akan dilakukan analisis dan implementasi *Index Compression* menggunakan *Variable Byte Code*. *Variable Byte Code* merupakan salah satu teknik *Index Compressing* dengan cara mengkompresi posting file dan digunakan oleh banyak sistem komersial / riset. Dalam mengkrompesi posting file, *Variable Byte Code* lebih efisien dan mudah diimplementasikan dibandingkan dengan teknik posting file lainnya, seperti *gamma code*. Diharapkan setelah menggunakan *Index Compression*, kapasitas indeks akan berkurang dan performansi dari *Information Retrieval System* meningkat.

1.2 Perumusan Masalah

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

1. Bagaimana proses pengindeksan dan kompresi dokumen dengan menggunakan *Variable Byte Code* dalam *Information Retrieval System*.
2. Bagaimana proses pencarian indeks yang sudah dikompresi menggunakan *Variable Byte Code.*
3. Bagaimana performansi dari *Information Retrieval System* setelah dilakukan pengkompresian indeks dengan menggunakan *Variable Byte Code* dilihat dari waktu pencarian, kapasitas indeks, dan nilai perbandingan *precision, recall*, dan IAP.

Batasan masalah agar tidak meluasnya materi pembahasan dalam tugas akhir ini ialah:

1. Koleksi dokumen dan query yang digunakan menggunakan bahasa Inggris.
2. Simulasi yang dibuat berbasis web menggunakan PHP dan basisdata MySQL
3. Analisis data dokumen dilakukan terhadap standar koleksi dokumen untuk *information retrieval* yang didapat dari: ftp://ftp.cs.cornel.edu/pub/smart/med/ yang bertipe file ".txt" dan meliputi Med *collection* dimana sudah terdapat kumpulan *query* beserta *relevance judgment* untuk setiap *query* tersebut.
4. *Query* pengujian (*query* uji) untuk pengukuran performansi sistem sudah ditentukan sebelumnya.

1.3 Tujuan

Tujuan pembuatan tugas akhir ini adalah :

1. Merancang dan membangun suatu perangkat lunak berupa search engine untuk pengimplementasian *Index Compression* dengan menggunakan *Variable Byte Code* dalam *Information Retrieval System.*
2. Melakukan analisis performansi dari *Information Retrieval System* setelah dilakukan kompresi indeks dengan menggunakan *Variable Byte Code* dilihat dari waktu pencarian, kapasitas indeks, dan nilai perbandingan *precision, recall*, dan IAP.

1.4 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang digunakan dalam menyelesaikan penelitian ini adalah:

1. Tahap Pengumpulan Data dan Studi Literatur
   Pada tahap ini dilakukan pengumpulan data dan mempelajari sumber-sumber pustaka yang ada, yang dapat dijadikan referensi mengenai *Information Retrieval* khususnya *Index Compression, Variable Byte*

Code, serta sumber-sumber lain yang menunjang penyelesaian tugas akhir ini. Sumber-sumber pustaka dapat berupa buku, paper, maupun halaman web.

2. Analisis dan Desain
   Pada tahap ini dilakukan analisis kebutuhan dan perancangan perangkat lunak yang akan dibangun. Perancangan yang dilakukan adalah perancangan arsitektur perangkat lunak..
3. Tahap Implementasi Sistem
   Pada tahap ini meliputi pembangunan perangkat lunak yang telah dirancang berdasarkan analisis dan perancangan yang telah dilakukan.
4. Tahap Analisis dan Pengujian
   Pada tahap ini akan dilakukan pengujian terhadap perangkat lunak yang telah dibangun, dan kemudian menganalisis hasil performansi yang didapatkan. Pengukuran performansi adalah precision-recall dan time space. Tujuan pengujian adalah untuk mengetahui bagaimana performansi perangkat lunak yang dibangun setelah mengkompresi indeks dengan *Variable Byte Code*.
5. Tahap Penyusunan 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 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 system.

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 dan Pengujian Sistem
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.

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:

1. Penerapan Index Compression dengan menggunakan metode *Variable Byte Code* dapat mengurangi kapasitas disk yang terpakai untuk indeks dan dapat meningkatkan kecepatan mendapatkan dokumen yang sesuai dengan *query*.
2. Untuk mendapatkan hasil kompresi yang lebih maksimal, maka perbandingan antara jumlah term dan jumlah dokumen dalam suatu koleksi dokumen harus semakin besar.
3. Meskipun performansi yang didapat melalui metode *Variable Byte Code* tidak mengalami peningkatan dalam hal precision, recall, dan nilai IAP jika dibandingkan dengan *Information Retrieval System* yang biasa (tidak menggunakan metode *Variable Byte Code*, kecepatan IRS menggunakan metode Variable Byte Code mengalami peningkatan (lebih cepat), sehingga disarankan menggunakan metode ini karena mendapatkan kapasitas disk yang lebih kecil untuk indeks dan kecepatan akses yang lebih cepat dibandingkan dengan IRS biasa.

5.2 Saran

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

1. Digunakan file dokumen yang lebih banyak, sehingga memaksimalkan kompresi yang dilakukan dan lebih terlihat jelas kapasitas disk yang berkurang dan kecepatan yang didapat.
2. Mengkombinasi metode *Variable Byte Code* dengan metode lain sehingga didapatkan performansi dari segi pendapatan dokumen (*precision*, *recall*, dan nilai IAP) lebih baik.
3. Jenis dokumen yang dicari tidak hanya berupa teks saja.
4. Proses *stemming* dan *word token* yang digunakan dikembangkan lagi.

DAFTAR PUSTAKA

[1] Buettcher, Stefan, *Inverted Files*, 2006, University of Waterloo, Canada.

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

[3] *Index Compression*, 2009, Dept of Computer Science and Engineering Hanyang University.

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

[5] Manning, Christopher D, Prabhakar Raghavan, Hinrich Schutze, *Introduction to Information Retrieval*, 2008, New York: Cambridge University Press.

[6] Moffat, Alistair, Justin Zobel, *Self-Indexing Inverted Files for Fast Text Retrieval*, 1994.

[7] Scholer, Falk, Hugh E. Williams, John Yiannis, and Justin Zobel, *Compression of Inverted Indexes for Fast Query Evaluation*, 2002, School of Computer Science and Information Technology, RMIT University, Australia

[8] Trotman, Andrew, *Compressing inverted files*, 2003, Department of Computer Science, University of Otago, New Zealand

[9] Van Rijsbergen, C.J., 1979, Information Retrieval. Department of Computing Science, University of Glasgow.

[10] Witten, Ian H., Moffat, Alistair, Bell,Timothy C., "Managing Gigabytes: Compressing and Indexing Documents and Images", second edition. Morgan Kaufmann Publishers, Academic Press,1999

[11] Zhang, Jiangong, Xiaohui Long, Torsten Suel, *Performance of Compressed Inverted List Caching in Search Engines*, 2008, CIS Department, Polytechnic University, USA