Jurnal Hasil Penelitian LPPM *Untag* Surabaya
Januari 2018, Vol. 03, No. 01, hal 60- 62
jurnal.untag-sby.ac.id/index.php/jhp17

E-ISSN : 2502-8308
P-ISSN : 2579-7980

Kompresi Data Text Menggunakan Metode "Duplicate Word Indexing"

Puteri Noraisya Primandari
FakultasTeknik, Universitas 17 Agustus 1945 Surabaya
Email: puterinoraisya@untag-sby.ac.id

Abstract

A data is often important so it should not be removed from data storage as it is used as an archive. The large number of new data archive data can slowly reduce free space and slow down the data transfer process. In this research, an approach is done to manage archive data by doing compression to minimize size. The results of this study were tested on some randomly books shows the number of data text size decrease of 19.46%.
Keywords: data, archive, compression, text

Pendahuluan

Dalam dunia komputer, data membawa informasi yang disimpan dalam media penyimpanan (data storage). Pada kasus tertentu, data bersifat penting dimana data-data lama pada data storage tidak boleh dihapus karena digunakan sebagai arsip. Disatu sisi, pertumbuhan data baru yang eksplosif perlu disimpan secara terus-menerus, selain cenderung mengurangi ruang kosong (free space) dalam data storage juga akan memperlambat proses pemindahan data jika diperlukan.

Salah satu jenis data yang digunakan untuk menyimpan informasi penting adalah data text. Pada penelitian ini, penulis melakukan kompresi data pada jenis data teks yang difokuskan pada materi pembelajaran. Hal ini dikarenakan landasan penulisan materi berdasarkan pada tulisan manusia yang didalamnya terdapat beberapa kata yang duplikat, seperti: dan, yang, di, ke, dalam, dan lain sebagainya. Beberapa kata yang duplikat tersebut akan diindeks dan dikompresi menggunakan metode "Duplicate Word Indexing"

Kajian Literatur

Teknik kompresi pada data berjenis text yang paling lama dikenalkan oleh seorang mahasiswa MIT bernama David A. Huffman dalam papernya "A Method for the Construction of Minimum-Redundancy Codes" (Huffman, 1952). Metode kompresi Huffman dilakukan dengan pendekatan statistic dimana setiap data simbol (karakter) dikodekan dengan rangkaian bit, simbol dengan probabilitas yang tinggi akan memperoleh kode-kode pendek sedangkan simbol dengan probabilitas rendah akan memperoleh kode-kode panjang.

Metode kompresi lainnya juga dilakukan oleh Scott Hauck dan William D. Wilson dalam papernya "Runlength Compression Techniques for FPGA Configurations" (Wilson, 1999). Teknik kompresi ini dilakukan dengan cara menyederhanakan simbol berjalan (urutan kemunculan simbol) lalu menuliskan simbol tersebut sebanyak satu kali diikuti dengan jumlah kemunculannya.

Metode

Duplicate Word Indexing

Duplicate Word Indexing

Sebuah teks pada umumnya memiliki banyak pengulangan kata terutama pada penggunaan kata penghubung (dengan, dan, jika, dsb). Selain itu pengulangan kata juga terjadi karena sebuah teks cenderung membahas suatu topik tertentu sehingga kata-kata yang berhubungan dengan topik tersebut seringkali digunakan berulang-ulang.

Dalam sistem komputer, jumlah karakter dalam file teks berbanding lurus dengan ukuran file. Setiap satu huruf berformat ASCII disimpan dalam 8 bit memori (1 byte). Tujuan dari penelitian ini adalah untuk mengurangi karakter sebanyak-banyaknya sehingga mengurangi ukuran file.

Pada penelitian ini dilakukan kompresi text menggunakan teknik "Duplicate Word

Indexing” yaitu dengan cara mendaftar kata-kata duplikat dan memberi indeks pada masing-masing kata tersebut sehingga kata duplikat tidak perlu ditulis berulang kali cukup dituliskan indeksnya saja.

Memasukkan data teks

Dalam penelitian ini,format teks yang digunakan difokuskan pada format *txt*. Berikut adalah contoh teks yang akan dikompresi oleh sistem.

Keluarga saya memiliki kegemaran membaca buku. Saya senang membaca novel di sofa. Adik senang membaca komik di kamar. Selesai membaca, saya dan adik meletakkan buku di dalam lemari

Mendeteksi kataduplikat

Pada proses ini, setiap kata dispisah berdasarkan spasi atau tanda baca lainnya. Kemudian sistem melakukan pencarian kata-kata duplikat dalam teks.

Menyimpan dan memberi indeks pada kata duplikat

Kata-kata duplikat diindeks kedalam memori berdasarkan nama, posisi dalam teks dan jenis penulisan kata. Penelitian ini membagi tigajenis penulisankata dalam teksseperti padatabel 3.1.

Tabel 3.1.
Jenis PenulisanKata dalam Teks

Jenis Penulisan Kata	Contoh
Huruf pertama kata menggunakan huruf kapital	Saya
Semua kata menggunakan huruf kecil	Saya
Semua kata menggunakan huruf besar	SAYA

Kata duplikat diurutkan dari yang terpendek sampai terpanjang kemudian masing-masing diberi indeks.Hasil indeks seperti pada tabel 3.2 berikut.

Tabel 3.2.
Kata Duplikat dan Indeksnya

Indeks	Kata Duplikat
[0]	*di*
[1]	*adik*
[2]	*buku*
[3]	*saya*
[4]	*membaca*

***Filter*kata duplikat**

Tujuan dari *filter*kata adalah untuk mencari nilai optimal dari kompresi yaitu dengan memastikan ukuran setelah proses kompresi lebih kecil daripada ukuran normal. Setiap kata duplikat yang telah disimpan dalam memori dikenakan aturan sebagai berikut:

Jika jumlah huruf dalam kata *m* dan jumlah adalah *n* maka panjang total karakter tersebut dalam teks adalah:

$$p1 = \sum_{i=0}^{m} 1_i * \sum_{i=0}^{n} 1_i \qquad (1)$$

Ketika telah dikompresi dan kata tersebut memiliki indeks *i*, maka adalah *i* dalam tipe data String (bukan Integer) sehingga jumlah huruf dalam dapat disimbolkan sebagai *o*maka

$$p2 = \sum_{i=0}^{m} 1_i + \left(\sum_{i=0}^{o} 1_i * \sum_{i=0}^{n} 1_i \right) \qquad (2)$$

Dari *p1* dan *p2* maka dapat dihitung jumlah penghematan karakter pada kata tersebut setelah terjadi kompresi adalah

$$p3 = p2 - p1 \qquad (3)$$

Jika $p3 \geq 0$ kata duplikat tersebut tetap disimpan dalam memori, jika tidak maka kata akan dihapus.

Menuliskan kata duplikat di database ke dalam teks

Pada tahapan ini, sistem akan menulis ulang teks awal dengan menggantikan kata-kata duplikat sesuai indeknya dalam memori disertai simbolnya.

*Keluarga*1 memiliki kegemaran*5*2.`1\*4\*5 novel\*0 sofa.`3*4*5 komik*0 kamar.*

*Selesai*5,*1 dan*3 meletakkan*2*0 dalam lemari.*

Pada penelitian ini, pengujian sistem dilakukan pada data teks yang terdapat pada beberapa buku yang diambil secara acak. File tersebut akan disimpan dalam format *txt* kemudian dilakukan teknik kompresi data yang dijelaskan pada bab 3. Tabel 4 adalah judul dan halaman buku yang digunakan.

Tabel 3. Judul Buku

Nomor Buku	Judul Buku	Halaman
1	Sistem Informasi Geografis: Prinsip Dasar dan Pengembangan Aplikasi (Irwansyah, 2013)	1 - 13
2	Pengantar Manajemen (Priyono, 2007)	1 - 21
3	Teori dan Aplikasi: Pengolahan Citra (Abdul Kadir, 2013)	1 - 10

Tabel 4 menunjukkan jumlah kata duplikat dan ukuran file (dalam *byte*) sebelum dan sesudah dikompresi sehingga dapat dihitung prosentase penurunan ukuran file dalam persen.

Tabel 4. Jumlah dan Ukuran Data Teks Sebelum dan Sesudah Dikompresi

No. Buku	Jumlah Kata Duplikat	Ukuran Sebelum (*bytes*)	Ukuran Sesudah (*bytes*)	%
1	155	11,874	9,671	18,36
2	332	30,977	25,363	18,12
3	128	7,502	5,859	21,90

I. KESIMPULAN

Pada penelitian ini, penerapan kompresi text menggunakan teknik *Duplicate Word Indexing* terbukti dapat menurunkan ukuran file dari sesudah dan sebelum dikompresi dengan rata-rata prosentase penurunan yang diambil dari beberapa buku secara acak sebanyak 19.46%.

Referensi

Abdul Kadir, A. S. (2013). Teori dan Aplikasi : Pengolahan Citra. Yogyakarta: Andi.

Huffman, D. A. (1952). *A Method for the Construction of Minimum-Redundancy Codes.* Proceedings of the I.R.E.

Irwansyah, E. (2013). Sistem Informasi Geografis: Prinsip Dasar dan Pengembangan Aplikasi. Yogyakarta: digibooks.

Priyono. (2007). Pengantar Manajemen. Sidoarjo: Zifatama.

Wilson, S. H. (1999). *Runlength Compression Techniques for FPGA Configurations*. IEEE, (pp. 286 - 287). Napa Valley, CA, USA.