Improving Run Length Encoding (RLE) on bit level by preprocessing

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

Universität Trier
FB IV - Informatikwissenschaften
Lehrstuhl für Informatik I

Gutachter:	Prof. Dr. Ingo J. Timm
	xxxxxxxxxx
Betreuer:	xxxxxxxxxx

Vorgelegt am xx.xx.xxxx von:

Sven Fiergolla
Am Deimelberg 30
54295 Trier
sven.fiergolla@gmail.com
Matr.-Nr. 1252732

Zusammenfassung

Hier steht eine Kurzzusammenfassung (Abstract) der Arbeit. Stellen Sie kurz und präzise Ziel und Gegenstand der Arbeit, die angewendeten Methoden, sowie die Ergebnisse der Arbeit dar. Halten Sie dabei die ersten Punkten eher kurz und fokussieren Sie die Ergebnisse. Bewerten Sie auch die Ergebnissen und ordnen Sie diese in den Kontext ein.

Die Kurzzusammenfassung sollte maximal 1 Seite lang sein.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Introduction

Die Einleitung besteht aus der Motivation, der Problemstellung, der Zielsetzung und einem erster Überblick über den Aufbau der Arbeit.
TODO:
- explain compression ratio
- define unit of compression to quantify question and results

1.1 Motivation

In the last decades, digital data transfer became available everywhere and to everyone. This rise of digital data urges the need for data compression techniques or improvements on existing ones. Run-length encoding (abbreviated as rle) is simple coding schemes that performs lossless data compression. rle compression simply represents the consecutive, identical symbols of a string with a run, usually denoted by $\sigma\ i$, where σ is an alphabet symbol and i is its number of repetitions. To give an example, the string aaaabbaaabbbba can be compressed into rle format as $a^4b^2a^3b^4a^1$. Its simplicity and efficiency make run-length encoding is still used in several areas like fax transmission, where rle compression is combined with other techniques into Modified Huffman Coding [9]. Most fax documents are typically simple texts on a white background, rle compression is particularly suitable for fax and often achieves good compression ratios. Another appliance of rle is optical character recognition, in which the inputs are usually images of large scales of identically valued pixels. Other applications appear in bioinformatics, where rle compression is employed to speed up the comparison of two biological sequences.

1.2 Problem statement

Some strings like aaaabbbb archive a very good compression rate because the string only has two different characters and they repeat at least twice. Therefore it can be compressed to a^4b^4 so from 8 byte down to 4 bytes if you encode it properly. On the other hand, if the input is highly mixed characters with few or no repetitions at all like abababab, the run length encoding of the string is $a^1b^1a^1b^1a^1b^1a^1b^1a^1b^1$ which needs at least 16 bytes.

So the inherent problem with run length encoding is oviusly the possible explosion in size, due to missing repetitions in the input string. Expanding the string to twice the orifinal size is not realy a good compression so one has to make sure the input data is fitted for rle as compression scheme. One goal is to minimize the increase in size in the worst case scenario.

Also it should improve the compression ratio on data suited for run length encoding and perform better than the originaly proposed rle.

1.3 Main Objective

Was ist das Ziel der Arbeit. Wie soll das Problem gelöst werden?

- bessere kompressionsrate im vergleich zu konventionellem RLE
- gleiche oder ähnliche decoding zeit ?
- ansatz beschreiben

The main objectives that derives from the problem statement is to archive an improved compression ratio compared to regular run length encoding. To unify the measurements, the compression ratio is calculated by encoding all files listed in the Galgary Corpus and then normalize the results.

Since most improvements like permutations on the input, for example a revertable Burros-Wheeler transformation to increase the number of consecutive Symbols or a different way of reading the bytestream take quite some time, encoding speed will increase. A secod objective might be to keep decoding speed close to the original run legth encoding.

1.4 Gliederung/Aufbau der Arbeit

Was enthalten die weiteren Kapitel? Wie ist die Arbeit aufgebaut? Welche Methodik wird verfolgt?

2. Basic principles of compression

To understand compression one first hast to understand some basic pricipels of information theory like Entropy and different approaches to compress different types of data with different encoding and entropy. I will also show the key differences between probability coding and dictionary coding and a few comments on lossy compression.

2.1 Compression and Encoding fundamentals

...

2.1.1 Entropy and Unit of Compression

2.1.2 Probability Coding

2.1.3 Dictionary coding

2.1.4 Irreversible Compression

2.2 Run Length Encoding

...

2.3 Huffman Coding

...

2.4 State of the Art

Die Literaturrecherche soll so vollständig wie möglich sein und bereits existierende relevante Ansätze (Verwandte Arbeiten / State of the Art / Stand der Technik) beschreiben bzw. kurz vorstellen. Es soll aufgezeigt werden, wo diese Ansätze Defizite aufweisen oder nicht anwendbar sind, z.B. weil sie von anderen Umgebungen oder Voraussetzungen ausgehen.

Je nach Art der Abschlussarbeit kann es auch sinnvoll sein, diesen Abschnitt in die Einleitung zu integrieren oder als eigenes Kapitel aufzuführen.

Beispiel, wie mit LaTeX zitiert werden kann: [4, 5, 12]

State of the art compression - techniques
- use cases

3. Analysis

In diesem Kapitel sollen zunächst das zu lösende Problem sowie die Anforderungen und die Randbedingungen einer Lösung beschrieben werden (eine präzisierte Aufgabenstellung). ...

3.1 Prerequisites

Anforderungen und Randbedingungen ...

3.2 Initial Findings

- no matrix representation but chunks of type byteArray
- static encoding difficulties
...

3.3 Improvements by Preprocessing

- First improvements due to byte remapping
- burrows wheeler transformation
...

3.4 Further Improvements

- combining different compression techniques
...

3.5 Summarization

Am Ende sollten ggf. die wichtigsten Ergebnisse nochmal in *einem* kurzen Absatz zusammengefasst werden.

4. Conceptual Design

In diesem Kapitel erfolgt die ausführliche Beschreibung des eigenen Lösungsansatzes. Dabei sollten Lösungsalternativen diskutiert und Entwurfsentscheidungen dargelegt werden.

To archive the main obejctive a few ideas to improve run lenthg encoding where found. The first real difference is the conception of reading the data in chunks and thee encode it vertically, meaning all the most significant bits in one chunk, then all second most significant bits and so on. The second change was to switch between the encoding of runs of characters bytes to count runs of ones and zeros so basically the same mechanism but on bit level.

4.1 Vertical Byte Reading

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

4.2 Preprocessing

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto

odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

4.2.1 Burrows Wheeler Transformation

4.2.2 Byte Mapping to reduce Input space

4.2.3 Dynamic Encoding

4.2.4 Alternative Encoding / Compression

4.3 Summarization

Am Ende sollten ggf. die wichtigsten Ergebnisse nochmal in *einem* kurzen Absatz zusammengefasst werden.

5. Implementation

- some detailed implementation details
...

5.1 Implementation Decisions

- why kotlin - performance improvements with the graalvm
- other decisions ...

5.2 Implementation Detail

- detailed information about specific modules and classes
...

5.2.1 Parsing

- explain universal parsing

5.2.2 Burrows Wheeler Transformation

- TBD

5.2.3 Byte Remapping

- show use case

5.2.4 Dynamic Encoding

- different sizes and optima for different files and chunk sizes

6. Evaluation

Hier erfolgt der Nachweis, dass das in Kapitel **??** entworfene Konzept funktioniert. Leistungsmessungen einer Implementierung werden immer gerne gesehen.

6.1 Functional Evaluation

- Mathematical Comparison
- Comparison of encoded file sizes
- comparing to regular REL and Huffman coding
...

6.2 Benchmarks

- Benchmark with the Galgary corpus
http://www.data-compression.info/Corpora/
...

6.3 Conclusion

Am Ende sollten ggf. die wichtigsten Ergebnisse nochmal in *einem* kurzen Absatz zusammengefasst werden.

7. Discussion

(Keine Untergliederung mehr)

Literaturverzeichnis

[1] ATM service categories: The benefits to the user. White Paper, The European Market Awareness Committee, May 1996.

[2] Autor. Titel. *Journaltitel*, Nummer des Jahrgangs(Nummer der Ausgabe):Seitenzahlen, December 1993.

[3] Autor. Titel. In *Buchtitel*. Verlag, 1994.

[4] Gerold Blakowski and Ralf Steinmetz. A media synchronization survey: Reference model, specification, and case studies. *IEEE Journal on Selected Areas in Communication*, 14(1):5–35, January 1996.

[5] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick. A framework for qos-based routing in the internet. RFC 2386 (Informational), August 1998.

[6] Kurt Gödel. *Titel*. Verlag, 1957.

[7] David Hutchison, Geoff Coulson, Andrew Campbell, and Gordon S. Blair. *Quality of Service Management in Distributed Systems*, chapter 11, pages 273–302. Addison Wesley, 1994. Editor: Morris Sloman.

[8] David E. McDysan and Darren L. Spohn. *ATM: Theory and Application*. McGraw-Hill, New York, 1995.

[9] A. Harry Robinson Roy Hunter, editor. *International digital facsimile coding standards, Proceedings of the IEEE 68*, 1980.

[10] Fred Stenz et al. *Technische Beschreibung für System 0815*, 1998.

[11] Fred Stenz, Willi Weich, and Dieter Drollig, editors. *About Time*, 1985.

[12] Fred Stenz, Willi Weich, Dieter Drollig, Kurt Klein, and Guenter Ganz. *Technische Beschreibung für System 4711*, 1998.

[13] Ludwig van Beethoven. *Titel*. Verlag, 1812.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Bachelor-/Masterarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe. Die Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vor-gelegt. Sie wurde bisher auch nicht veröffentlicht.

Trier, den xx. Monat 20xx