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

Abstract

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.

Contents

List of Figures

List of Tables

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 [7] (abbreviated as RLE) is a 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 still usable in several areas like fax transmission, where RLE compression is combined with other techniques into Modified Huffman Coding [4]. 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 [1].

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 obviously the possible explosion in size, due to missing repetitions in the input string. Expanding the string to twice the original size is not really 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 originally 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 second objective might be to keep decoding speed close to the original run length encoding.

1.4 Structure of this work

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

- describe following structure
- use references
- try to keep idea of a recurrent theme

2. Principles of compression

To understand compression, one first has to understand some basic principles 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

TBD
...

2.1.1 Information Theory and Entropy

As Shannon described his analysis about the English language [8], he used the term Entropy closely aligned with its Definition in classical physics, where it is defined as the disorder of a system. Specifically speaking, it is assumed that a system has a set of states it can be in and it exists a probability distribution over those states. Shannon then defines the Entropy as:

$$H(S) = \sum_{s \in S} P(s)\, i(s)$$

where S describes all possible States, $P(s)$ is the likelihood of $s \in S$. So generally speaking it means that evenly distributed probabilities imply a higher Entropy and vice versa. It also implies that, given a source of information S, its average information content per message from S is also described by this formula.

In addition to that, Shannon defined the term self information $i(s)$ as:

$$i(S) = log_2 \frac{1}{P(s)}$$

indicating that the higher the probability of a state, less information can be contained. As an example, the statement "The criminal is smaller than 2 meters." is very likely but doesn't contain much information, whereas the statement "The criminal is larger than 2 meters." is not very likely but contains lots of information. With these definitions in mind, we can analyze some properties for the English language from an information theory perspective of view.

Combining those approaches, you will find that $\sum$ being a finite alphabet, and $P(s_i)$ describing the likelihood of s_i and $i(s_i)$ its self information, $i(s_i)$ also describes the amount of bits needed for this symbol, therefore its log_2.

$$H(\sum) = \sum_{i=1}^{n} P(s_i) \cdot log_2 \frac{1}{P(s_i)}$$

So the entropy also describes the theoretical amount of bits needed to persist a message from $\sum$, calculated in *bps* as Bits per Symbol.

TODO: fix caption

	$P(a)$	$P(b)$	$P(c)$	$P(d)$	$P(e)$	H
1.	0.2	0.2	0.2	0.2	0.2	2.322
2.	0.94	0.01	0.01	0.01	0.01	0.322

TODO: describe some relations between probability and entropy of information source

2.1.2 General Analysis

To evaluate the efficiency of a specific compression technique, we have to determine how much information the raw data contains. In this case for textual compression at first, we are talking about the English language. There have been broad analysis of ASCII Entropy, consisting of 96 different printable symbols for the English language, generating approaches for calculating the entropy [6].

If we assume a 96 Symbol alphabet and a uniform probability distribution, we have a quite high entropy of $log_2(96) = 6,6 bps$. In a empirically distribution generated by text analysis, the entropy comes down to $4.5\,bps$. Using an encoding with separated encoding per Symbol like the Huffman Coding, we can archive an entropy of $4.7\,bps$ which is only slightly worse that the theoretical entropy. By changing the assumption to blocks of symbols of length 8, we get 96^8 different blocks. With a probability distribution close to the English language we get an entropy as low as $1.3\,bps$ which implies a possible reduction of symbols to 3 printable characters without generating longer texts.

Consulting newer sources we will find that, up to 500 character long symbol based text analysis, with around 20.3 $Mio.$ different symbols, results in an entropy of around $1,58\,bps$ [3]. This gives a clear limit of how much compression we can theoretically expect from English text. Using a combination of different approaches like Huffman Coding and Run Length Coding, and encoding different bits of a single symbol with different approaches, we do no longer encode every symbol separately, which might result in a better compression.

2.1.3 Probabilistic Coding

The general idea behind Probability Coding is to analyze probabilities for messages and the encode them in bit strings according to their probability. This way, messages that are more likely and will repeat more often can be encoded in a smaller bit string representation. The generation of those probability distributions is considered part of the Analyzer module by the algorithm and will be discussed later on (chap: design, chap: impl). Probability coding compression can be discerned into fixed unique coding, which will represent each message with a bit string with the amount of bits being an integer, for example Huffman coding as type of prefix coding. In contrast to Huffman coding there are also arithmetic codes which can represent a message with a floating point number of bits in the corresponding bit string. By doing so they can “fuse” messages together and need less space to represent a set of messages. ...

2.1.4 Dictionary Coding

Dictionary Coding is best suited for Data with a small amount of repeating patterns like a text representing specific words. In this case it is very effective to save the patterns just once and refer to them if the are used later on in the text. If we know a lot about the text itself in advance, a static dictionary is sufficient but in general we want to compress an arbitrary text. This requires a more general approach as used by the well known LZ77 and LZ88 algorithms described by Jacob Ziv and Abraham Lempel [9]. They use a so called sliding window principle, a buffer moving across the data to encode in the case of LZ77 or a dynamic dictionary in the implementation of LZ78. Both of them are still used in modified versions in real world applications as described in Section 2.4.

The main difference between probabilistic coding and dictionary coding is, that the so far presented probability based methods like RLE or Huffman Coding are working on single character or byte whereas the dictionary based methods encode groups of characters of varying length. This is a clear advantage in terms of compression performance on every day data because of its repeating patterns most textual data has, as shown in Section 3.1.

2.1.5 Irreversible Compression

Irreversible Compression or also called "Lossy Compression" is a type of compression which looses information in the compression process and is therefore not completely reversible, hence the name. There are a lot of use cases for these type of compression algorithms, mostly used for images, video and audio files. These files typically contain information almost not perceptible for an average human like really small color differences between two pixels of an image or very high frequencies in an audio file. By using approximations to store the information and accept the loss of some information while retaining most of it, lossy image compression algorithms can get twice the compression performance compared to lossless algorithms. Due to the fact that most lossy compression techniques are not suited for text compression which is the main use case for this work, we will not elaborate any further on this topic.

2.2 Run Length Coding

Run length coding might count as the simplest coding scheme that makes use of context and repeating occurrences as described in the Section 1.1. While the example made earlier was in textual representation, it is best suited and mostly used for pallet based bitmap images [2] such as fax transmissions or computer icons.

2.2.1 The history

The ITU-T T4 (Group 3) standard for Facsimile (fax) machines [5] is still in force for all devices used over regular phone lines. Each transmission sends a black and white image, where each pixel is called a *pel* and with a horizontal resolution of 8.05 $\frac{pels}{mm}$ and the vertical resolution depending on the mode. To encode each sequence of black and white pixels, the T4 standard uses RLE to encode each sequence of black and white pixels and since there are only two values, only the run length itself has to be encoded. It is assumed that the first run is always a white run, so there is

a dummy white pel at the beginning of each sequence. For example, the sequence *bbbbwwbbbbb* can be encoded as 1,4,2,5 with the leading white dummy pixel.

To further reduce the RLE encoded sequence, a probability coding procedure like Huffman coding can be used do code the sequence, since shorter runs like lengths are generally more common than very long runs, which is also done by the T4 standard. Based on a broad analysis of these run lengths, the T4 defined a set of static Huffman codes, using a different codes for the black and white pixels.

run length	white codeword	black codeword
0	00110101	0000110111
1	000111	010
2	0111	11
3	1000	10
4	1011	011
...		
20	0001000	00001101000
...		
64+	11011	0000001111
128+	10010	000011001000

A small subset of the static T4 Huffman codes are given in Table 2.2.1 which also shows the most likely runs to occur. Runs of more than 64 have to be encoded in multiple Huffman codes, for example a run of 150 has to be encoded with the Huffman code of 128 followed by the code for 22.

By combining RLE with a probabilistic approach, the ratio of compression rises because we no longer have to encode a run of length 1 or 2 with a fixed size of bits each instead we can write a recognizable Huffman code of varying length, which will be explored in grater detail later on as well as the decoding of the shown Huffman codes.

2.2.2 Limitations

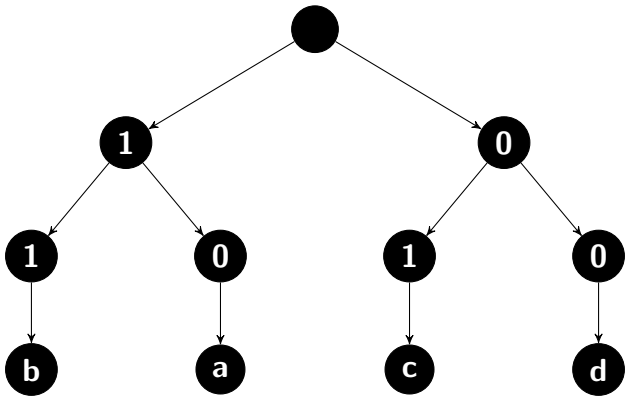
As mentioned in the Section 1.2, run length encoding is rarely used for regular text or continuous tone images because its potentially increase in size, due to non repetitive characters or bytes. To reduce this problem there are several approaches known, partly implemented and used some of which will be addressed like a Burrows-Wheeler-Transformation.

2.3 Prefix Coding

Prefix codes make use of an untypical idea in computer science. Usually we deal with fixed length codes like 7 bit ASCII or 32 bit Integer representations which map each possible value into that fixed amount of bits. For compression it would be a benefit if we could write codes of variable length. The problem with these varying length codes is that as soon as they are part of a sequence, it becomes very hard to tell where one code starts and finishes, resulting in ambiguous encoding. For example the set of codes $\{(a, 1), (b, 01), (c, 101), (d, 011)\}$ and the encoded message is 1011, there is no way to tell which message was encoded.

One way to address this problem is by adding extra stop symbols or encoding a length before each code but those just add additional data. Another approach is to use so called prefix codes and make no code which is prefix of another code, which makes them uniquely decodable.

2.3.1 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.

State of the art compression
- techniques
- use cases

2.5 Limits of compression

- existence of not compressable strings
-https://www.quora.com/Is-there-a-theoretical-limit-to-data-compression-If-so-how-was-it-found
- unable to compress random data
- Kolmogorow-Komplexität

3. Analysis

The following chapter contains a detailed analysis of the problem and some fundamental requirements for the algorithm.

3.1 Initial Findings

Originally developed and used on black and white pallet images containing only two values, on a binary level so to speak, we want to use it for rather arbitrary data, mostly text. The initial algorithm is not suited for that because continuous text as binary representation does not contain runs of any kind, which could be compressed. The ASCII representation of the letter 'e' which is the most common in general English, has the value 130 or '01100101' as 7-bit ASCII or '001100101' as byte value of the UTF-8 representation. And this is the case for most printable characters as they al have a value between 32 and 127 (or 255 for the extended ASCII).

RLE could also work on a byte level, because there should be repetitions of any kind. Byte level RLE encodes runs of identical byte values, ignoring individual bits and word boundaries. The most common byte level RLE scheme encodes runs of bytes into 2-byte packets. The first byte contains the run count of 1 to 256, and the second byte contains the value of the byte run. If a run exceeds a count of 256, it has to be encoded twice, one with count 256 and one with any further runs.

However after some analysis of the corpus data, it was shown that most runs had a value of one and almost no runs larger that 4 occurred, which lead to the conclusion, two bit for the run count should be plenty, which is also shown in Table 3.1. To have a comparison to some extent, the initial performance analysis was performed on the Calgary Corpus but the results where very underwhelming as expected, as shown in the table below. Even with a run size of just two bits, there is still a increase in size of about 9% and uses 8.74 $\frac{bits}{symbol}$. This is still useful as a kind of a base line.

bits per rle number	compression ratio in %	bits per symbol in $\frac{bits}{symbol}$
8	165	13.2
7	154	12.38
6	144	11.57
5	134	10.77
4	125	10.00
3	116	9.29
2	109	8.74

...

3.2 Improvements by Preprocessing

- First improvements due to byte remapping
- burrows wheeler transformation

...

3.3 Further Improvements

- combining different compression techniques

...

3.4 Summary

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 objective a few ideas to improve run length encoding where found. The first real difference is the conception of reading the data in chunks and then encode it in parallel, 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 Parallel Byte Reading

...

4.1.1 First Ideas

...

4.1.2 New Perspective

...

4.1.3 Performance Improvements

...

4.2 Preprocessing

...

4.2.1 Burrows Wheeler Transformation

...

4.2.2 Byte Mapping to reduce Input space

...

4.2.3 Dynamic Encoding

...

4.3 Alternative Compression for partial data

...

4.3.1 Huffman Coding

...

4.4 Summary

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

5.3 Implementation Evaluation

- evaluation of implementation choices made ...

6. Evaluation

Hier erfolgt der Nachweis, dass das in Kapitel 4 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)

Bibliography

[1] Al-Okaily A, Almarri B, Al Yami S, and Huang CH. Toward a better compression for dna sequences using huffman encoding. *J Comput Biol.*, 24(4):280–288, 2017.

[2] MJ Dallwitz. An introduction to computer images. *TDWG Newsletter*, 7(10):2, 1992.

[3] Fabio G. Guerrero, editor. *A new look at the classical entropy of written English*, 2009.

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

[5] International Telecommunication Union (ITU). *T.4 : Standardization of Group 3 facsimile terminals for document transmission*, 2004.

[6] Maciej Liśkiewicz and Henning Fernau. *Datenkompression.*

[7] A. Harry Robinson and C. Cherry, editors. *Results of a prototype television bandwidth compression scheme, Proceedings of the IEEE 3*, volume 55, March 1967.

[8] C. E. Shannon. *Prediction and entropy of printed English*, volume 30. Jan 1951.

[9] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977.

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