Natural Language Compression using Byte Codes

The thesis studies the properties of what has become known as Byte Codes. This is a basic problem of Coding Theory which is an important part of the larger field of Data Compression. In order to achieve some compression gain, the easy to handle fixed length codes have long ago been replaced by variable length codes, which are more difficult to handle and thus imply a processing slowdown. The Byte codes are a compromise: the codewords are still of variable length, but the lengths are restricted to multiples of bytes, so the handling is faster. The thesis reviews existing Byte codes and then goes on to suggest new variants. All these codes are compared according to various criteria, showing that the new suggested ones are competitive and often supersede the properties of the state of the art alternatives. My overall impression of the submitted work is a good one and I will recommend accepting it as a PhD thesis. Let me go now into details.

To summarize the content, the thesis starts by setting the background about codes in general and Byte codes and their properties in particular. This part, spanning about one third of the thesis, would probably not be necessary for a reader who is familiar with this field of research, say, in a journal publication, but since a PhD thesis should be a self-contained dissertation accessible also to readers of a broader audience, this background part is certainly justified for a thesis, and shows that it is up to date. The author then introduces the new notion of Open Dense Codes, generalizing previous work and adding his own contributions.

The author did a good job producing a coherent, well explained and well documented piece of work on a very specific topic which is treated with a praiseworthy precision for the smallest details. Indeed, the writing style is fluent and the English is fine in general, but could be improved. The thesis is well organized and well written (only a few typos or mistakes) and the author used with consistency a large variety of different notations. A number of well-chosen examples, tables and figures help to clarify the suggested algorithms, that are given formally and also explained.

Summarizing, this is the work of a mature computer scientist, having dealt with a very specific niche of the broad area of Data Compression. The fact that this work has already produced several publications at prestigious conferences and in a
journal speaks about itself. The thesis is fine as is, and there is no need to correct or amend anything. I can therefore recommend this thesis, even in its present form, to be recognized as a PhD thesis. Formally, the author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) I do recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

The few selected comments below are mainly technical and are intended to help the author, should he want to further publish his results. Following these suggestions is not a necessary requirement for my positive judgment, and the thesis is acceptable without amendments, as mentioned above. The comments are listed by order of appearance, not by importance. The list is far from being exhaustive. There are many similar and also additional instances that I have not listed here. I would suggest the author to ask a native English speaker to proofread his text.

- P.39 require a different compression approaches should be require different compression approaches
- P.43 Definition 4.1.1 Isn’t it rather 2^b-ary than b-ary?
- P.45 only natural portions of the text. Please explain.
- P.48 line 23: + k Is k at all defined here?
- P.56 in the constant time should be in constant time
- P.60 at the expense of an old word used previously but unused... Is there always such a word? Explain why, and if not, what to do in that case.
- P.61 can be at extended should be can be extended
- P.65 ...since the maximum number of codewords obtain the codewords with a size of one byte... Please rewrite, that does not make sense, probably wrong use of “obtain”.
- P.66 and $c$ is the average number... Use another letter instead of $c$. $c$ is reserved for $(s,c)$-codes.

- P.73 The paragraph starting "There are other rather efficient..." sounds like an introduction again. You say "so-called" when introducing a new notion, only for the first time. Here we are already on page 73, and the reader knows what Dense codes are. This looks as if it were copied from the introduction of some paper, without being adapted to the thesis.

- P.88 despite the efforts of the authors of [25]... Something wrong here. Maybe you wanted to say, despite my own efforts to get code from the authors of [25]? Because why should the authors of [25] do any effort to obtain any implementation? They did it themselves!

- P.89 The results in Figure 4.14 prove... I would rather say "show". To prove, in a mathematical thesis, means something else.

- P.92 in the constant time should be in constant time

- P.92 in neighborhood of should be in the neighborhood of

Ramat Gan, June 27, 2014

Prof. Shmuel T. Klein
Department of Computer Science
Bar Ilan University
Ramat Gan 52009, Israel
Dissertation Opponent's Review

Dissertation title: Natural Language Compression Using Byte Codes

Author: Petr Procházka

Reviewer: Jiří Dvorský

Up-to-datedness of the dissertation

Up-to-datedness of the dissertation is evident. Data compression is widely used in all areas of electronic communication, data storage, data transmission, searching through data etc. While conventional lossless data compression is usually done using common tools such as gzip or bzip2, which defines de-facto industrial standard, there is still a large space for research in highly specialized data compression methods which meet other requirements e.g. fast decompression, fast direct searching.

Formal structure and organization of the dissertation

The dissertation is conceptually divided into five chapters. The first two chapters introducing reader to the dissertation topic, establish basic notations and provides comprehensive state of the art. The third chapter describes novel approaches to dense coding. The large chapter four, forming core of the work brings new, original compression methods developed by the author. The conclusion and possible future work is contained in the fifth chapter.

Completion of the dissertation objectives

The objectives of the dissertation were fully met.

Assessment of the methods used in the dissertation

Methods used in the dissertation are fully compliant with the current state of knowledge in the data compression and appropriately develop this part of computer science. Newly developed methods are clearly described subsequently tested on standard text corpora. Authors own ideas are clearly separated from the assumed ones whose sources are properly cited.

Evaluation of the results and contributions of the dissertation

Results and contributions of the dissertation represents:

- establishment of Open Dense Code,
- definition of two new adaptive dense codes – Two-Byte Dense Code (TBDC) and Self-tunning Dense Code (STDC),
• definition of semi-adaptive version of TBDC, and
• modification of STBDC for compression of a set of small text files.

All of the presented compression methods are optimized for the compression of the natural language content. The methods are also very fast in compression and decompression.

Remarks, objections, notes, and questions for the defense

I consider the dissertation very well written, the results were published in the prestigious Data Compression Conference (DCC), therefore, I have no doubt about the quality of the dissertation. So I have rather general questions:

1. The newly developed methods are proposed to be used for XML (HTML) compression, but there are no such file in the test corpora. Are there any results for XML compression?

2. The test files corpora consist of text files written in English. That ok. But what about other languages? For example Czech?

3. Is there any possibility implement proposed compression or decompression methods in hardware?

The overall evaluation of the dissertation

I am convinced that the presented Ph.D. thesis represents excellent study providing valuable contribution to the state of scientific knowledge in the area. The author has demonstrated his potential to individual scientific work.

Conclusion

The author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (a) of the Law Nr. 111/1998 (The Higher Education Act) I do recommended the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

Olomouc, August 4th 2014

Jiří Dvorský
Department of Computer Science
VSB – Technical University of Ostrava
Opponent's report

on PhD Thesis

"Natural Language Compression Using Byte Codes"

by

Petr Procházka

The thesis provides a very thorough coverage of one particular area of lossless data compression, namely text compression using dense codes.

Formal structure and organization.

The text of the thesis is organized into five chapters. The first two of them serve as a necessary introduction into the topic of the dissertation. This includes a formulation of the main problems studied as well as an introduction of fundamental concepts, notations and results from stringology, data compression, information theory and linguistics. This part is concluded with a comprehensive survey of previous results from the field of data compression that are relevant to author’s research.

The next two chapters constitute the core of the thesis. The author describes the concept of Open Dense Code, designed to serve as a unifying framework for various dense code schemes, covering both those designed previously as well as novel ones suggested in this thesis:

- two adaptive methods called Two-Byte Dense Code and Self-Tuning Dense Code, aimed at compression of small files with natural language content;
- a semi-adaptive version of the latter, focused on compression of large files, possibly created as a concatenation of texts of different nature;
- a further modification of the previous method, designed to compress a huge set of small files, which allows direct search on the compressed text and decompression of an arbitrary part of the file.

In concluding remarks, the author summarizes the original contribution of the thesis and suggest several topics for further research.

Up-to-dateness of the dissertation.

This thesis may serve as a convincing counterexample to the — occasionally presented — statement that the problem of lossless data compression has been already resolved. The up-to-dateness of the problems studied in the dissertation is particularly apparent from the second and third sections of
chapter four. This part of the text deals with compression methods which facilitate direct searching on the compressed text, with particular attention paid to fashionable applications in the field of web search engines. The relationship to contemporary research may be documented by a number of current references, including quite recent contributions to *Data Compression Conference*, the most prestigious international forum in this field.

**Assessment of the methods used in the dissertation.**

The exposition of the thesis is presented in fairly comprehensible and concise way. Even though English is not author's native language, he succeeds in making his meaning clear. There are only rare ambiguities that can be usually clarified from the context.

The exposition is illustrated with a number of examples, tables and figures. All proposed algorithms are both explained in the text and described more formally using a pseudocode.

It should be noted that the thesis contains only one theorem with a formal — and rather straightforward — proof. This, however, is certainly not a deficiency. The author follows an approach that is quite common in this field: the newly designed methods are implemented and experimentally evaluated on a set of files from standard data corpora. The testbed includes selected files from Calgary and Canterbury corpora as well as four larger files, obtained by concatenation of selected texts from Project Gutenberg. Particular source codes of author's implementation or data used in the experiments are not included.

**Contribution of the dissertation.**

The original contribution of the dissertation is twofold. First, the author introduces the idea of Open Dense Code, which provides a unifying concept covering both existing and newly-designed dense codes. Second, he describes and evaluates four new methods, fitting into this general scheme.

I particularly appreciate the fact that the experiments are to used to make a recommendation on the superiority of particular compression methods for specific data (small files, large concatenation of different texts) or specific applications (direct search on the compressed text, random decompression of an arbitrary part of the text). This, in my opinion, is a meaningful approach to perform up-to-date research in the field of lossless data compression.

**Remarks, objections, notes.**

I have found several formal deficiencies in the text, mostly in Section 2.1. They are, however, only of a minor nature, and may be fixed by local modifications
or even deletion (e.g. the notion of *optimal code* from Definition 2.1.12 does not seem to be used in the rest of the thesis).

Section 2.2 introduces a classification of natural languages, based on their degree of inflection. In this respect, one cannot avoid mentioning a fairly recent research on syllable-based compression due to Jan Lánský, which discovered superiority of this approach for languages with rich morphology. It is natural to ask whether author's experiments revealed any sensitivity of his methods towards the language of the compressed text.

Section 4.1.3 describes test files used in subsequent experiments. The author selected several natural language files from two classical data corpora and formed four of his own by concatenating some Project Gutenberg texts. It may be of interest that a much larger instance of the latter may be found in *Pizza & Chili* corpus, created by Ferragina and Navarro as a testbed for compressed indexes.

The results of experiments are carefully described in Chapter 4. However, to play fair with prospective readers (and maybe followers), it may be worth considering to share not only ideas and experimental results, but also source codes and test files used in the experiments, e.g. in the form of an electronic supplement to the dissertation.

**The overall evaluation of the dissertation.**

The thesis provides a noticeable contribution to the field of natural language compression.

The quality of the dissertation is best verified by the fact that all its crucial parts were either presented at conferences, or published in a journal. The list includes four international conferences, one local workshop and a paper in a respectable impacted journal. This, in my opinion, is more than a sufficient publication activity for a PhD candidate.

The impact of author's results may be also documented by three citations, including one reference in an impacted journal. Another evidence of author's establishment in research community is provided by his co-authorship of a full contribution to this year's Data Compression Conference (which, however, does not seem to be covered by this thesis).

**Conclusion**

The author of the dissertation clearly demonstrates a talent for creative scientific research. Without hesitation, I recommend the thesis for presentation and defense with the aim of receiving the PhD degree.
Prague, 30 June 2014

doc. RNDr. Tomáš Dvořák, CSc.
Faculty of Mathematics and Physics
Charles University in Prague