BSc thesis

Batchelor's Programme in Computer Science

Tutkielman otsikko

Jussi Timonen

January 30, 2020

FACULTY OF SCIENCE

UNIVERSITY OF HELSINKI

Supervisor(s)

Prof. D.U. Mind, Dr. O. Why

Examiner(s)

Prof. D.U. Mind, Dr. O. Why

Contact information

P. O. Box 68 (Pietari Kalmin katu 5)

00014 University of Helsinki,Finland

Email address: info@cs.helsinki.fi

URL: http://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Batchelor's Programme in Computer Science

Tekijä — Författare — Author
Jussi Timonen

Työn nimi — Arbetets titel — Title
Tutkielman otsikko

Ohjaajat — Handledare — Supervisors
Prof. D.U. Mind, Dr. O. Why

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
BSc thesis	January 30, 2020	11 pages, 3 appendice pages

Tiivistelmä — Referat — Abstract

Tämä dokumentti on tarkoitettu Helsingin yliopiston tietojenkäsittelytieteen osaston opinnäytteiden ja harjoitustöiden ulkoasun ohjeeksi ja mallipohjaksi. Ohje soveltuu kanditutkielmiin, ohjelmistotuotantoprojekteihin, seminaareihin ja maisterintutkielmiin. Tämän ohjeen lisäksi on seurattava niitä ohjeita, jotka opastavat valitsemaan kuhunkin osioon tieteellisesti kiinnostavaa, syvällisesti pohdittua sisältöä.

Työn aihe luokitellaan ACM Computing Classification System (CCS) mukaisesti, ks. https://www.acm.org/about-acm/class, käyttäen komentoa \classification{}. Käytä muutamaa termipolkua (1–3), jotka alkavat juuritermistä ja joissa polun tarkentuvat luokat erotetaan toisistaan oikealle osoittavalla nuolella.

ACM Computing Classification System (CCS)
General and reference → Document types → Surveys and overviews
Applied computing → Document management and text processing → Document management → Text editing

Avainsanat — Nyckelord — Keywords
ulkoasu, tiivistelmä, lähdeluettelo

Säilytyspaikka — Förvaringsställe — Where deposited
Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information
Algorithms specialisation line

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Batchelor's Programme in Computer Science

Tekijä — Författare — Author
Jussi Timonen

Työn nimi — Arbetets titel — Title
Tutkielman otsikko

Ohjaajat — Handledare — Supervisors
Prof. D.U. Mind, Dr. O. Why

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
BSc thesis	January 30, 2020	11 pages, 3 appendice pages

Tiivistelmä — Referat — Abstract

Use this otherlanguage environment to write your abstract in another language if needed.

Topics are classified according to the ACM Computing Classification System (CCS), see https://www.acm.org/about-acm/class: check command \classification{}. A small set of paths (1–3) should be used, starting from any top nodes referred to bu the root term CCS leading to the leaf nodes. The elements in the path are separated by right arrow, and emphasis of each element individually can be indicated by the use of bold face for high importance or italics for intermediate level. The combination of individual boldface terms may give the reader additional insight.

ACM Computing Classification System (CCS)
General and reference → Document types → Surveys and overviews
Applied computing → Document management and text processing → Document management → Text editing

Avainsanat — Nyckelord — Keywords

ulkoasu, tiivistelmä, lähdeluettelo

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Algorithms specialisation line

Contents

1 Introduction

Enormous datasets are a common case in today's applications. Compressing the datasets is beneficial, because they naturally decrease memory requirements but also are faster when compressed data is read from disk (**Zob95**).

One of the leading methods of data compression is variable-length coding (**Sal99**), where frequent sequences of data are represented with shorter codewords. Because the sequences of data have different lenghts when compressed, it is not trivial to determine the exact location of a certain element. If this is required, the usual data compression algorithms are inefficient. Fortunately this is not a requirement compression algorithms usually need to fulfill.

However, random access of compressed data is needed in compressed data structures. In most compression methods, the only way to this is to decompress data from the beginning. An integer compressing method with fast random access is explained and compared existing state-of-the-art methods.

2 Basic Concepts

Variable-byte encoding (**Wil99**) is a method of compressing integers. It splits an integer into blocks of b bits and then encodes the block as $b + c$. c is a bit denoting whether b has the most significant bits of the number or not. A good dataset for variable-byte encoding is a set with both very small and very large numbers, since he compression comes from omitting the leading zeros in the bit array. This method loses in compression performance to other methods (**Bri09**), but they are very fast to decode and allow fast random access.

- Explain rank / select

3 Previous Work

bl

4 Algorithm

here

5 Results

hmll

5.1 Figures

Figure 5.1 gives an example how to add figures to the document. Remember always to cite the figure in the main text.

Figure 5.1: University of Helsinki flame-logo for Faculty of Science.

5.2 Tables

Table 5.2 gives an example how to report experimental results. Remember always to cite the table in the main text.

Table 5.1: Results with 100k entries (in milliseconds).

Experiment	128	256	32768	65536	2^{24}	$2^{32} - 1$
$7 bit V Byte encoding$	34.97	49	53.04	52.18	53.08	76.21
$8 bit V Byte encoding$	33.57	32.47	42.96	43.11	46.15	65.14
$7 bit V Byte encoding with array$	33.39	46.85	51.24	49.03	48.93	66.84
$8 bit V Byte encoding with array$	32.52	31.88	41.54	39.94	41.15	52.86

Table 5.2: Results with 1M entries (in milliseconds).

Experiment	128	256	32768	65536	2^{24}	$2^{32} - 1$
7bitV Byteencoding	38.17	55.09	64.38	65.36	68.08	159
8bitV Byteencoding	37.09	37.75	53.44	54.6	59.32	148.7
7bitV Byteencodingwitharray	38.09	55.42	62.22	61.25	71.72	135.01
8bitV Byteencodingwitharray	36.13	36.83	50.58	50.73	56.93	103.18

6 Citations

6.1 Citations to literature

References are listed in a separate .bib-file. In this case it is named `bibliography.bib` including the following content:

```
@article{einstein,
    author =        "Albert Einstein",
    title =         "{Zur Elektrodynamik bewegter K{\"o}rper}. ({German})
        [{On} the electrodynamics of moving bodies]",
    journal =       "Annalen der Physik",
    volume =        "322",
    number =        "10",
    pages =         "891--921",
    year =          "1905",
    DOI =           "http://dx.doi.org/10.1002/andp.19053221004"
}

@book{latexcompanion,
    author      = "Michel Goossens and Frank Mittelbach and Alexander Samarin",
    title       = "The \LaTeX\ Companion",
    year        = "1993",
    publisher   = "Addison-Wesley",
    address     = "Reading, Massachusetts"
}

@misc{knuthwebsite,
    author      = "Donald Knuth",
    title       = "Knuth: Computers and Typesetting",
    url         = "http://www-cs-faculty.stanford.edu/%7Eknuth/abcde.html"
}
```

In the last reference url field the code `%7E` will translate into `~` once clicked in the final pdf.

References are created using command `\cite{einstein}`, showing as (**einstein**). Other examples: (**latexcompanion**; **knuthwebsite**).

Citation style can be negotiated with the supervisor. See some options in `https://www.sharelatex.com/learn/Bibtex_bibliography_styles`.

6.2   Crossreferences

Appendix B on page i contains some additional material.

7 From tex to pdf

In Linux, run `pdflatex filename.tex` and `bibtex filename.tex` repeatedly until no more warnings are shown. This process can be automised using make-command.

8 Conclusions

It is good to conclude with a summary of findings. You can also use separate chapter for discussion and future work. These details you can negotiate with your supervisor.

Appendix A   Template instructions

In the HY-CS-main.tex file you will find the following STEPS 0–5. Bellow instructions what each STEP means and how to set up your thesis by following these STEPS.

STEP 0 – Clone the thesis template

- One template for all thesis types: https://www.overleaf.com/read/hzgngkgshqwh

STEP 1 – BSc or MSc thesis?

1. Select whether your are writing BSc (tkt for new, tktl for old) or MSc (csm for new, cs for old, dsm for data science) thesis.

2. Select your language: finnish, english, or swedish.

3. If you are writing MSc select your line / track.

STEP 2 – Set up your personal information

1. Write the working title of your thesis.

2. Write your name to the author field.

3. Write the names of your supervisors and examiners of the thesis.

STEP 3 – Write your abstract here

- You can also have the abstract in multiple languages with otherlanguages-environment. Bellow example how to add an english abstract to a thesis written in some other language than english:

```
\begin{otherlanguage}{english}
\begin{abstract}
Your abstract text goes here.
\end{abstract}
\end{otherlanguage}
```

STEP 4 – Writing your thesis

1. There are some writing instructions in [bsc/msc]_[finnish/english]_contents.tex files.

2. You can delete the contents of [bsc/msc]_[finnish/english]_contents.tex file and write your thesis inside that file.

STEP 5 – Set your bibliography style

- The default is Numbering alphabetic order, which should be used in most cases.

Appendix B Sample Appendix

usually starts on its own page, with the name and number of the appendix at the top. The appendices here are just models of the table of contents and the presentation. Each appendix Each appendix is paginated separately.

In addition to complementing the main document, each appendix is also its own, independent entity. This means that an appendix cannot be just an image or a piece of programming, but the appendix must explain its contents and meaning.