Data Compression Algorithms

Introduction

Marcus Hutter (*1967)

DeepMind

Data compression

The process of converting an input data stream (the source stream, the original raw data) into output data stream (the compressed stream, the bitstream) that has a smaller size.

Compression algorithm = *encoding* (compression)
+ *decoding* (decompression)

Compression

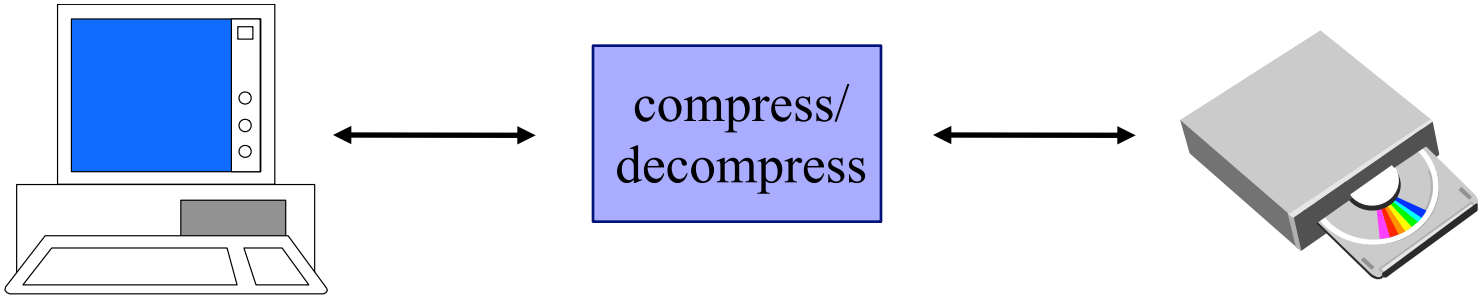
- *lossless*: the restored and original data are identical
- *lossy*: the restored data are a „reasonable“ approximation of the original

Methods

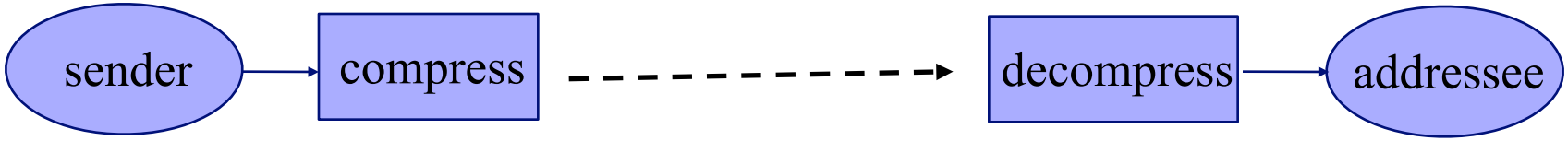
- *static* / *adaptive*
- *streaming* / *block*

Goals of data compression

✓ save storage



✓ reduce the transmission bandwidth



Measuring the performance

input data size u B

compressed data size k B (K bits)

measure	formula	example
compression ratio	$k / u * 100\%$	36 %
compression factor	$u : k$	3 : 1
compression gain	$(u-k)/u*100\%$	64 %
bpc (bits per char) bpp (bits per pixel) *average codeword length*	K / u	1.47 *b*/*c*
relative compression (*percent log ratio*)	100 ln (k / k')	10.5

size of data compressed by a standard algorithm

Performance: data corpora

Comparing lossless compression algorithms

Calgary Corpus (1987)

- 14 files: text, graphics, binary files

Canterbury Corpus (1997)
http://corpus.canterbury.ac.nz

- 11 files + artifical c. (4) + large c. (3) + miscellaneous c. (1)

Silesia Corpus (2003)
Sebastian Deorowicz, Politechnika Śląska, Gliwice
http://sun.aei.polsl.pl/~sdeor/index.php?page=silesia

- 18 files of sizes 6 – 51MB

Prague Corpus (2011)
Jan Holub et al., FIT ČVUT, Praha
http://www.stringology.org/projects/PragueCorpus/

- 30 files, 58 MB total

Canterbury corpus

soubor	kategorie	velikost (B)
alice29.txtd	English text	152089
asyoulik.txt	Shakespeare	125179
cp.html	HTML source	24603
fields.c	C source	11150
grammar.lsp	LISP source	3721
kennedy.xls	Excel Spreadsheet	1029744
lcet10.txt	Technical writing	426754
plrabn12.txt	Poetry	481861
ptt5	CCITT test set	513216
sum	SPARC Executable	38240
xargs.1	GNU manual page	4227

Canterbury corpus

The Artificial Corpus

a.txt	The letter 'a'	1
aaa.txt	The letter 'a', repeated 100,000 times.	100000
alphabet.txt	Enough repetitions of the alphabet to fill 100,000 characters	100000
random.txt	100,000 characters, randomly selected from [a-z|A-Z|0-9|!|] (alphabet size 64)	100000

Canterbury corpus

The Large Corpus

E.coli	Complete genome of the E. Coli bacterium	4638690
bible.txt	The King James version of the bible	4047392
world192.txt	The CIA world fact book	2473400

The Miscellaneous Corpus

pi.txt	The first million digits of pi	1000000

Data compression contests

Calgary Corpus Compression Challenge (1996)

- http://mailcom.com/challenge/
- (777,777.00 - X) / 333 for an archive of length X B that compresses the 14 file version of the Calgary corpus
- now (580,170.00 - X) / 111 $

Data compression contests

Calgary Corpus Compression Challenge (1996)

Size	Date	Name
759881	09/1997	Malcolm Taylor
692154	08/2001	Maxim Smirnov
680558	09/2001	Maxim Smirnov
653720	11/2002	Serge Voskoboynikov
645667	01/2004	Matt Mahoney
637116	04/2004	Alexander Rhatushnyak
608980	12/2004	Alexander Rhatushnyak
603416	04/2005	Przemysław Skibiński
596314	10/2005	Alexander Rhatushnyak
593620	12/2005	Alexander Rhatushnyak
589863	05/2006	Alexander Rhatushnyak
580170	07/2010	Alexander Rhatushnyak

Data compression contests II

Hutter Prize (2006) http://prize.hutter1.net

- create a self-extracting archive of the 100 MB prefix of English Wikipedia
- 500 € for each 1% improvement of the archive size

author	date	dec	size	comp. factor	RAM	time
Matt Mahoney	24.3.2006	paq8f	18'324'887	5.46	854MB	5h
Alexander Ratushnyak	25.7.2006	paq8hp5	17'073'018	5.86	900MB	5h
Alexander Ratushnyak	14.5.2007	paq8hp12	16'481'655	6.07	936MB	9h
Alexander Ratushnyak	23.5.2009	decmprs8	15'949'688	6.27	936MB	9h
Alexander Ratushnyak	4.11.2017	phda9	15284944	6.54	1048MB	5h

Data compression contests II

Hutter Prize : update

- create a self-extracting archive of the 1 GB prefix of English Wikipedia
- run in ≤50 hours, a single CPU core and <10GB RAM and <100GB HDD

author	date	dec	size	comp. factor	RAM	time
Alexander Ratushnyak	4.7.2019	phda9v1.8	116'673'681	8.58	6.3GB	23h
Artemiy Margaritov	31.5.2021	starlit	115'352'938	8.67	10GB	50h
Saurabh Kumar	16.7.2023	fast cmix	114'156'155	8.76	8.4GB	43h

Hutter prize

Goal: encourage research in AI

Marcus Hutter: being able to compress well is closely related to acting intelligently

- M. Hutter, *Towards a Universal Theory of Artificial Intelligence based on Algorithmic Probability and Sequential Decisions,* Proceedings of the 12th European Conference on Machine Learning, 226-238, 2000
- the optimal behaviour of a goal-seeking agent in an unknown but computable environment
- guess at each step that the environment is controlled by a shortest program consistent with all interaction so far

Kolmogorov complexity (algorithmic information theory)

Compressing natural language text – Turing imitation game alternative

Limits of lossless compression

Encoding f: {n-bit strings} $\rightarrow$ {strings of length $< n$}

$|\,\mathrm{Dom}\, f\,| = 2^n$

$|\,\mathrm{Im}\, f\,| \leq 2^n - 1$

$\Rightarrow$ f cannot be injective!

Let $\mathrm{M} \subseteq \mathrm{Dom}\, f$ such that $\forall s \in \mathrm{M}, |\, f(s)\,| \leq 0.9n$

f injective on $\mathrm{M} \Rightarrow |\mathrm{M}| \leq 2^{1+0.9n} - 1$

$n = 100, |\mathrm{M}|/\, 2^n < 2^{-9}$

$n = 1000, |\mathrm{M}|/\, 2^n < 2^{-99} \approx 1.578 \cdot 10^{-30}$