

Main page
Contents
Featured content
Current events
Random article
Donate to Wkipedia
Wkipedia store

Interaction

Help

About Wikipedia

Community portal

Recent changes

Contact page

Tools

What links here Related changes

Upload file

Special pages

Permanent link

Page information

Wikidata item

Cite this page

Print/export

Create a book
Download as PDF
Printable version

Languages

العربية

日本語

中文

Article Talk Read Edit View history Search Q

Byte pair encoding

From Wikipedia, the free encyclopedia

Byte pair encoding^[1] or digram coding^[2] is a simple form of data compression in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur within that data. A table of the replacements is required to rebuild the original data. The algorithm was first described publicly by Philip Gage in a February 1994 article "A New Algorithm for Data Compression" in the *C Users Journal*.^[3]

Byte pair encoding example [edit]

Suppose we wanted to encode the data

aaabdaaabac

The byte pair "aa" occurs most often, so it will be replaced by a byte that is not used in the data, "Z". Now we have the following data and replacement table:

ZabdZabac

Z=aa

Then we repeat the process with byte pair "ab", replacing it with Y:

ZYdZYac

Y=ab

Z=aa

We could stop here, as the only literal byte pair left occurs only once. Or we could continue the process and use recursive byte pair encoding, replacing "ZY" with "X":

XdXac

X=ZY

Y=ab

Z=a

This data cannot be compressed further by byte pair encoding because there are no pairs of bytes that occur more than once.

To decompress the data, simply perform the replacements in the reverse order.

References [edit]

- 1. ^ Philip Gage, A New Algorithm for Data Compression. "Dr Dobbs Journal" ₺.
- ^ Ian H. Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes. New York: Van Nostrand Reinhold, 1994. ISBN 978-0-442-01863-4.
- 3. ^ "Byte Pair Encoding" ₺.

v· t· e	Data compression methods [hide]		
Lossless	Entropy type	Unary · Arithmetic · Golomb · Huffman (Adaptive · Canonical · Modified) · Range · Shannon · Shannon–Fano · Shannon–Fano–Elias · Tunstall · Universal (Exp-Golomb · Fibonacci · Gamma · Levenshtein)	
	Dictionary type	Byte pair encoding · DEFLATE · Lempel–Ziv (LZ77 / LZ78 (LZ1 / LZ2) · LZJB · LZMA · LZO · LZRW · LZS · LZSS · LZW · LZWL · LZX · LZ4 · Statistical)	
	Other types	BWT · CTW · Delta · DMC · MTF · PAQ · PPM · RLE	
Audio	Concepts	Bit rate (average (ABR) · constant (CBR) · variable (VBR)) · Companding · Convolution · Dynamic range · Latency · Nyquist–Shannon theorem · Sampling · Sound quality · Speech coding · Sub-band coding	

	Codec parts	A-law \cdot μ -law \cdot ACELP \cdot ADPCM \cdot CELP \cdot DPCM \cdot Fourier transform \cdot LPC (LAR \cdot LSP) \cdot MDCT \cdot Psychoacoustic model \cdot WLPC	
lmage	Concepts	Chroma subsampling · Coding tree unit · Color space · Compression artifact · Image resolution · Macroblock · Pixel · PSNR · Quantization · Standard test image	
	Methods	Chain code · DCT · EZW · Fractal · KLT · LP · RLE · SPIHT · Wavelet	
Video	Concepts	Bit rate (average (ABR) · constant (CBR) · variable (VBR)) · Display resolution · Frame · Frame rate · Frame types · Interlace · Video characteristics · Video quality	
	Codec parts	Lapped transform · DCT · Deblocking filter · Motion compensation	
Theory	Entropy · Kolmogorov complexity · Lossy · Quantization · Rate–distortion · Redundancy · Timeline of information theory		
⑥ Compression formats ⋅ ⑥ Compression software (codecs)			

Categories: Lossless compression algorithms

This page was last modified on 20 January 2015, at 09:03.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Privacy policy About Wikipedia Disclaimers Contact Wikipedia Developers Mobile view

