Byte pair encoding

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]

1 Byte pair encoding example

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=aa

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.

2 References

[1] Philip Gage, *A New Algorithm for Data Compression*. “Dr Dobbs Journal”.

[2] 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”.

3 Text and image sources, contributors, and licenses

3.1 Text

- **Byte pair encoding** *Source:* https://en.wikipedia.org/wiki/Byte_pair_encoding?oldid=643343918 *Contributors:* Damian Yerrick, Mboverload, Grotte, Matt Mahoney, MarkHudson, XLerate, SmackBot, MindlessXD, MisterHand, Tan90deg, INVERTED, Sterrys, R'n'B, Addbot, Erik9bot, Helpful Pixie Bot, TerryAlex and Anonymous: 7

3.2 Images

- **File:Symbol_template_class.svg** *Source:* https://upload.wikimedia.org/wikipedia/en/5/5c/Symbol_template_class.svg *License:* Public domain *Contributors:* ? *Original artist:* ?

3.3 Content license

-