Codificación de pares de bytes

La codificación de pares de bytes[1] o la **codificación de digram**[2] es una forma simple de compresión de datos en la que el par más común de bytes consecutivos de datos se reemplaza con un byte que no ocurre dentro de esos datos. Se requiere una tabla de reemplazos para reconstruir los datos originales. El algoritmo fue descrito públicamente por primera vez por Philip Gage en un artículo de febrero de 1994 "Un nuevo algoritmo para la compresión de datos" en el *C Users Journal*.[3]

Se ha demostrado que una variante de la técnica es útil en varias aplicaciones de procesamiento de lenguaje natural, como GPT, GPT-2 y GPT-3 de OpenAI.[4]

Ejemplo de codificación de par de bytes

Supongamos que queremos codificar los datos

```
aaabdaaabac
```

El par de bytes "aa" ocurre con mayor frecuencia, por lo que será reemplazado por un byte que no se usa en los datos, "Z". Ahora tenemos los siguientes datos y tabla de reemplazo:

```
ZabdZabac
Z = aa
```

Luego repetimos el proceso con el par de bytes "ab", reemplazándolo con Y:

```
ZYdZYac
Y = ab
Z = aa
```

Podríamos detenernos aquí, ya que el único par de bytes literal que queda solo ocurre una vez. O podríamos continuar el proceso y usar codificación recursiva de pares de bytes, reemplazando "ZY" con "X":

```
XdXac
X = ZY
Y = ab
Z = aa
```

Estos datos no se pueden comprimir aún más mediante la codificación de pares de bytes porque no hay pares de bytes que se producen más de una vez.

Para descomprimir los datos, simplemente realice los reemplazos en el orden inverso.

Véase también

- Emparejamiento recursivo
- Algoritmo de Sequitur

Referencias

1. Philip Gage, *A New Algorithm for Data Compression*. «Dr Dobbs Journal» (http://www.drdobbs.com/article/print?articleId=184402829&dept_url=/).
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» (https://web.archive.org/web/20160326130908/http://www.csse.monash.edu.au/cluster/RJK/Compress/problem.html). Archivado desde el original (http://www.csse.monash.edu.au/cluster/RJK/Compress/problem.html) el 26 de marzo de 2016.
4. Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav *et al.* (4 de junio de 2020). «Language Models are Few-Shot Learners» (http://arxiv.org/abs/2005.14165). *arXiv:2005.14165 [cs]*.

Enlaces externos

- Esta obra contiene una traducción derivada de «Byte pair encoding» de Wikipedia en inglés, publicada por sus editores (https://en.wikipedia.org/wiki/Byte_pair_encoding?action=history) bajo la Licencia de documentación libre de GNU y la Licencia Creative Commons Atribución-CompartirIgual 4.0 Internacional (https://creativecommons.org/licenses/by-sa/4.0/deed.es).
- Un nuevo algoritmo para la compresión de datos; Gage 1994 (https://www.derczynski.com/papers/archive/BPE_Gage.pdf)