Incremental encoding

From Wikipedia, the free encyclopedia

Incremental encoding, also known as **front compression**, **back compression**, or **front coding**, is a type of delta encoding compression algorithm whereby common prefixes or suffixes and their lengths are recorded so that they need not be duplicated. This algorithm is particularly well-suited for compressing sorted data, e.g., a list of words from a dictionary.

For example:

Input	Common prefix	Compressed output
myxa myxophyta myxopod nab nabbed nabbing nabit nabk nabob nacarat nacelle	no preceding word 'myx' 'myxop' no common prefix 'nab' 'nabb' 'nab' 'nab' 'nab' 'na' 'nac'	0 myxa 3 ophyta 5 od 0 nab 3 bed 4 ing 3 it 3 k 3 ob 2 carat 3 elle
64 bytes		46 bytes

The encoding used to store the common prefix length itself varies from application to application. Typical techniques are storing the value as a single byte; delta encoding, which stores only the change in the common prefix length; and various universal codes. It may be combined with other general lossless data compression techniques such as entropy encoding and dictionary coders to compress the remaining suffixes.

Applications [edit]

Incremental encoding is widely used in information retrieval to compress the lexicons used in search indexes; these list all the words found in all the documents and a pointer for each one to a list of locations. Typically, it compresses these indexes by about 40%.[1]

As one example, incremental encoding is used as a starting point by the GNU locate utility, in an index of filenames and directories. The GNU locate utility further uses bigram encoding to further shorten popular filepath prefixes.

References [edit]

1. ^ Ian H. Witten, Alistair Moffat, Timothy C. Bell. Managing Gigabytes. Second edition. Academic Press. ISBN 1-55860-570-3. Section 4.1: Accessing the lexicon, subsection Front coding, pp.159–161.

This computer storage–related software article is a stub. You can help Wikipedia by expanding it.

Categories: Lossless compression algorithms | Database index techniques | Storage software stubs