RAREFIED DATA COMPRESSION METHOD

Eugeny L. Yakimovitch

In this paper a new data compression method called modular encoding is introduced. A modification of the method with the use of enumerative encoding and its applications to redundant data compression are discussed. Theoretical constraints for calculation the compression ratio are provided.

Introduction

A study of data compression methods with possible application to optimization of the data transmission process is very important. Efficiency of modern computer systems considerably depends on the existing infrastructure between separate interacting objects. In most cases, an abstract computer network can be considered as a model of such infrastructure, in which it is possible to observe the data transmission processes in details. In this connection any optimization of such process may turn out to be a significant improvement, including the application of data compression methods.

It is known that data compression costs time to be evaluated, thus any application of data compression methods is limited with time constraints. This fact makes development of robust data compression methods especially attractive.

In this paper a new data compression method is proposed which can be effectively applied to compress rarified data, transmitted through the network. It is assumed that redundant characteristics of the data can be analyzed and estimated in advance.

The method itself is a kind of a new entropy encoding method, based on modular arithmetic. Further in this paper, the discussion of such modular encoding method is stated and some modification of this method, including the use of enumerative encoding, is introduced.

Modular encoding

Let σ stands for an information source to be coded by method. For instance, σ is a string of symbols $\sigma = (s_1, s_2, .., s_n), |\sigma| = n$, where each symbol $s \in A$, A is an alphabet of the source σ and $|\sigma|$ denote length of σ. Let λ be a length in bits of each symbol $s \in A$ and $\overline{A}$ be a set of symbols that are missing in alphabet $\forall s \in \overline{A} : s \notin A$ of σ (had never occurred in a block). If for coded σ there exists a redundancy constraint such that $l = |\overline{A}| > 0$, then information in σ can be efficiently compressed into string $\mu(\sigma)$ with length in bits equal to

$$|\mu(\sigma)| \leq \log_2((2^\lambda - l)^n) + 2^\lambda, \qquad (1)$$

where 2^λ is a length of corresponding characteristic vector of alphabet A.

To put the statement above in a more simple way, the introduced modular encoding can be seen as simple transition of a numeric value form higher base into lower. This is an especially valuable feature of the possible implementation

of the method because the major part of the coding and decoding process can be accomplished by simple fast modular arithmetic operations.

Actually, to achieve a good ratio by the introduced data compression method, a more accurate constraint should be set. Thus the application of compression method would be rational only if the next constraint is satisfied:

$$\begin{cases} \Upsilon(n,\lambda,l) = n - \log_2((2^\lambda - l)^n) - 2^\lambda \\ \Upsilon(n,\lambda,l) > 0 \end{cases} \qquad (2)$$

Obviously a study of function $\Upsilon(n,\lambda,l)$ is an optimization problem.

Enumerative encoding

Further analysis of the introduced method results in a simple modification. The length of characteristic vector in (2) calculated as 2^λ can be significantly reduced due to the application of enumerative encoding.

A method of enumerative encoding was numerously reinvented by different researchers because of its fundamental properties. For the needs of the current application the binary case of enumerative coding is used. Basically it states, that every binary string $\varsigma \in \{0,1\}$ of length v and number of ones o can be coded as a number of ς in set of all possible combinations $B_o^v, |B_o^v| = \binom{v}{o}$, where $\binom{v}{o}$ is a binomial coefficient.

A new length of ς encoded by binary enumerative coding is computed by formula:

$$|\beta(\varsigma)| \leq \log_2\binom{v}{o} + \log_2(v), \qquad (3)$$

where v is a length of ς and o is a number of ones in it.

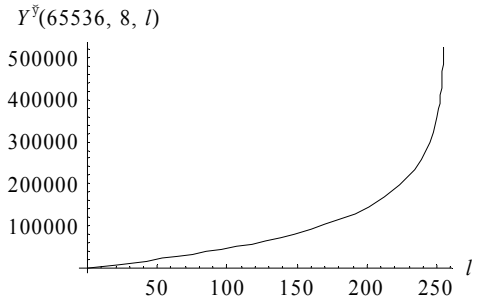
The method is very efficient for the asymmetric case $(o >> \frac{u}{2}) \| (o << \frac{u}{2})$, when number of zeros significantly differs form number of ones.

Modification and Conclusions

Modification of the modular encoding by extra encoding a characteristic vector of alphabet A with enumerative coding leads to another constraint, that can be noted as:

$$\begin{cases} \Upsilon'(n,\lambda,l) = n - \log_2((2^\lambda - l)^n) - \log_2\binom{2^\lambda}{l} - \log_2(2^\lambda) \\ \Upsilon'(n,\lambda,l) > 0 \end{cases} \qquad (4)$$

To prove that the discussed theory has a practical value, a study of redundancy case can be done. For instance, let the length of σ be *65536* and number of bits for each symbol will be $\lambda = 8$. The result of application of the modified modular encoding to σ with $0 \leq l \leq 255$ is shown in the plot on picture 1.



Picture 1. Theoretical estimation of modified modular encoding ratio.

All the discussed methods are reversible and can be successfully applied for robust data compression of the redundant data. From this point of view the method can be compared to existing method of entropy coding such as [1, 2]. The main difference is that the redundancy constraints such as (2) or (4) may require a highly rarified information source. Another disadvantage of the method is that it is off-line, thus not directly applicable for the compression of network data.

Further research will be conducted towards practical implementation of the method. An adaptive technique may be applied to solve the pointed difficulties of the method.

Literature:

1. Golomb, S.W. (1966). , Run-length encodings. IEEE Transactions on Information Theory, IT--12(3):399—401
2. R. F. Rice (1971) and R. Plaunt, "Adaptive Variable-Length Coding for Efficient Compression of Spacecraft Television Data, " IEEE Transactions on Communications, vol. 16(9), pp. 889-897, Dec. 1971.

Eugeny Ludvigovich Yakimovitch, Teaching assistant of Belarusian State University of Informatics and Radioelectronics, master in computer sciences, eugeny.yakimovitch@gmail.com