ОСОБЕННОСТИ ПОСТРОЕНИЯ АЛГОРИТМОВ ЭНТРОПИЙНОГО КОДИРОВАНИЯ

Гарнышев И.Н.[1], Казанцев С.В.[2], Мальков Р.Ю.[3], Семенов И.Д.[4], Юдин С.В.[5]
Email: Garnyshev1160@scientifictext.ru

[1]Гарнышев Игорь Николаевич - сетевой инженер,
отдел администрирования сетей передачи данных,
Тинькофф Банк;
[2]Казанцев Сергей Владимирович - главный инженер,
департамент сетей передачи данных,
Сбербанк;
[3]Мальков Роман Юрьевич – эксперт,
Центр компетенций по облачным решениям,
Техносерв,
г. Москва;
[4]Семенов Иван Дмитриевич - старший инженер,
департамент сетей передачи данных,
Servers.com Лимассол, Кипр;
[5]Юдин Степан Вячеславович - администратор сети,
департамент технического обеспечения и развития инфраструктуры информационных систем,
Спортмастер, г. Москва

Аннотация: *в статье проведен анализ принципов энтропийного кодирования. Разработаны основы методологии кодирования дискретного информационного источника. Предложены алгоритмы определения условной энтропии и общего количества информации после обработки данных дискретного источника. Разработана схема определения диапазона значений ожидаемой длины для символьного набора дискретного источника. Показано, что разработанную методологию можно использовать для построения алгоритмов кодирования Хаффмана, арифметического кодирования и универсального кодирования.*
Ключевые слова: *дискретный информационный источник, функция вероятности, условная энтропия, общее количества информации, символьный блок, ожидаемая длина, неравенство Крафта-Макмиллана.*

PECULIARITIES OF THE ENTROPY CODING ALGORITHMS' DEVELOPMENT

Garnyshev I.N.[1], Kazantsev S.V.[2], Malkov R.Yu.[3], Semenov I.D.[4], Iudin S.V.[5]

[1]Garnyshev Igor Nikolaevich - Network Engineer,
DATA NETWORK ADMINISTRATION DEPARTMENT,
TINKOFF BANK;
[2]Kazantsev Sergei Vladimirovich - Senior Engineer,
NETWORK DEPARTMENT,
SBERBANK;
[3]Malkov Roman Yurevich – Expert,
CLOUD SOLUTIONS DEPARTMENT,
TECHNOSERV CLOUD,
MOSCOW;
[4]Semenov Ivan Dmitrievich - Senior Engineer,
NETWORK DEPARTMENT,
SERVERS.COM LIMASSOL, CYPRUS;
[5]Yudin Stepan Vyacheslavovich - Network Administrator,
DEPARTMENT OF TECHNICAL SUPPORT AND INFORMATION SYSTEMS INFRASTRUCTURE DEVELOPMENT,
SPORTMASTER, MOSCOW

Abstract: *the article includes analysis of the principles of entropy coding. The basics of the coding methodology for a discrete information source are developed. Algorithms for determining the conditional entropy and the mutual information of the discrete source processed data are proposed. The scheme of the expected length values range determining for a code string of a discrete source is developed. It is shown that the developed methodology can be used to develop Huffman coding, arithmetic coding and universal coding algorithms.*
Keywords: *discrete information source, probability function, conditional entropy, mutual information, code string, expected length, Kraft inequality.*

УДК 004.627

Введение

На сегодняшний день математические модели кодирования данных широко используются в системах передачи и хранения информации. Информация при этом может быть представлена в цифровых форматах с использованием символов и алфавитов, которые при построении обобщенной методологии следует рассматривать в самом широком смысле. Таким образом, выбор способа представления информации соответствует процессу кодирования и является наиболее важным аспектом разрабатываемой теории, а также соответствующего математического аппарата.

При ***анализе современных исследований***, проведенных в рамках данной темы, были рассмотрены основы кодирования дискретного информационного источника [1, 2] и, в частности, понятие энтропии и т.н. источники без памяти [1, 3-5]. Были рассмотрены методы построения математического аппарата на базе понятий условной вероятности и условной энтропии [6-10], а также общего количества информации источника [11].

Тем не менее, в качестве ***нерешенной части общей задачи*** рассмотрено задание по разработке комплексной методологии кодирования дискретного информационного источника, что было предложено в качестве ***цели данного исследования***, для чего требуется построить алгоритмы определения условной энтропии и общего количества информации после обработки данных дискретного источника, а также схему определения диапазона значений ожидаемой длины для символьного набора дискретного источника.

1. Определение энтропии дискретного источника

Дискретным информационным источником (discrete information source) называется устройство или процесс, набор выходных значений которого представляет множество объектов, которые могут быть определены в рамках конечного алфавита A, состоящего из конечного множества $\{x_i\}$ элементов [1, 2]:

$$\begin{cases} A = \{x_i\} \\ i \in [1; I] \end{cases}. \qquad (1)$$

В том случае, когда множество элементов информационного источника достаточно велико, можно определить вероятность определения каждого из элементов X,принадлежащих множеству $\{x_i\}$ конечного алфавита A через соответствующую функцию распределения вероятности $P(x_i)$ или вектор вероятности $\overline{p_X}$ (рис. 1).

На основе множества $\{x_i\}$ и функции распределения вероятности $P(x_i)$ может быть вычислена функция энтропии $H(X)$ (рис. 2), которая определяет количество информации, соответствующее переменой:

$$H(X) = -\sum_{i=1}^{I} P(x_i) \log_2\big(P(x_i)\big). \qquad (2)$$

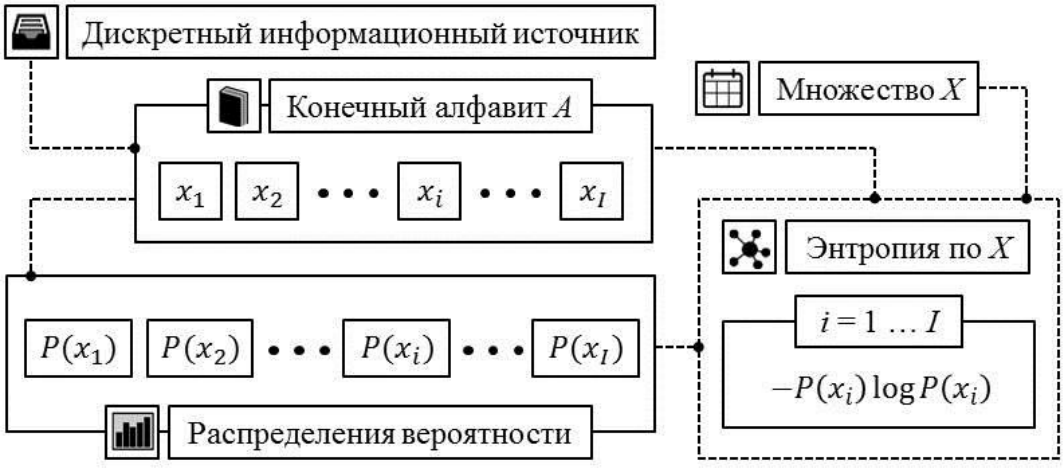

Рис. 1. Базовый алгоритм определения энтропии дискретного источника

Из чего можно вывести область значений функции энтропии как:

$$H(X) \in [0; \log_2(I)]. \qquad (3)$$

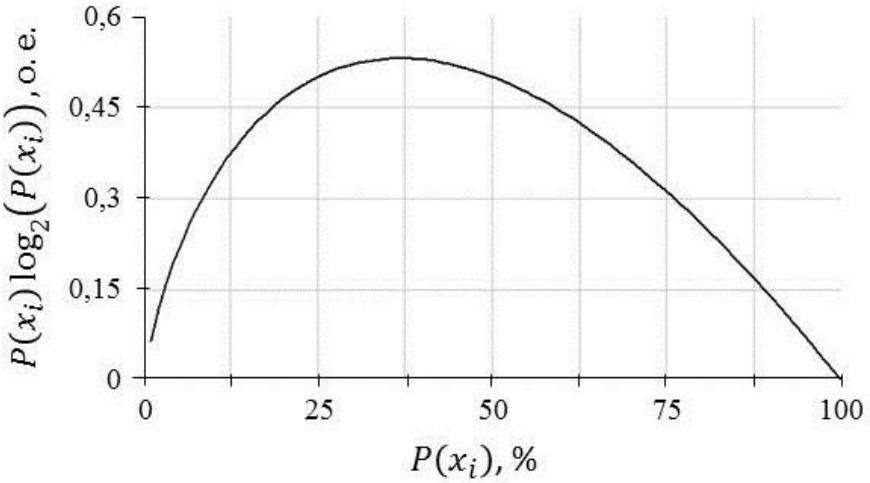

Рис. 2. График зависимости энтропии от функции вероятности по X

Таким образом, можно отметить, что значения, которые принимает функция энтропии как количества информации, что в случае использования двоичного логарифма представляется в битах, всегда неотрицательны, причем минимальное и максимальное значение $H(X)$ определяются в соответствии со следующими условиями:

$$\begin{cases} H(X) = H_{min} = 0 \text{ при } P(x_i) = 1 \\ H_{max} = \log_2(I) \text{ при } P(x_i) = \dfrac{1}{I} \text{ для } \forall i \end{cases} \quad (4)$$

Для дальнейшего развития математической модели энтропии дискретного источника необходимо ограничить анализ источниками выходные значения, которых являются независимыми [1, 3-5], т.е. источниками без памяти (memoryless source), но при этом характеризуются одним распределением вероятности (IID: Independent Identically Distributed).

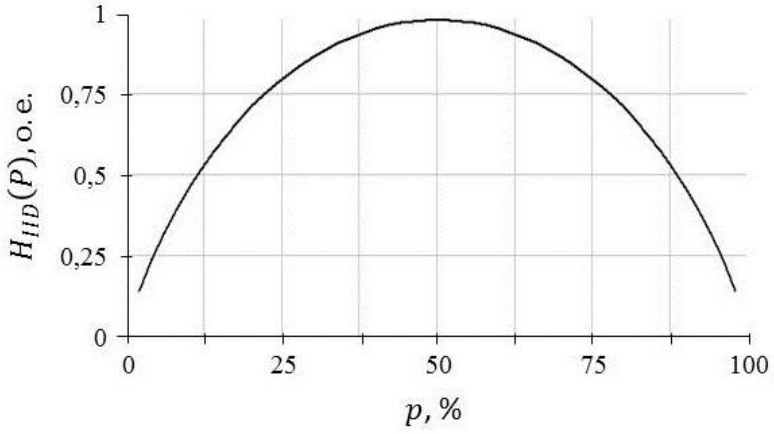

Рис. 3. График зависимости энтропии от функции вероятности для источников без памяти с одинаково распределенными выходами

Функцию энтропии для источников без памяти с одинаково распределенными выходами, где пределы функции распределения определяются величиной p можно записать как (рис. 3):

$$\begin{cases} H_{IID}(P) = -(p\log p + (1-p)\log(1-p)) \\ P \in [p; 1-p] \end{cases} \quad (5)$$

Как можно видеть, данный график является симметричным с максимумом при $p = 0{,}5$ (соответственно $P = 0{,}5$), где $H_{IID}(P) = 1$.

При кодировании информации зачастую бывает актуальным представлять код в виде серий (runs) идентичных элементов (например, при двоичном кодировании — в виде серий «0» или «1»). На основе

данных серий можно сформировать конечный алфавит, элементы которого определяют длину серии, как $A_0 = \{0, 1, 2 \dots n \dots N\}$.

Дальнейшее развитие математического аппарата подразумевает введение понятия условной энтропии (conditional entropy), которая определяется через условную вероятность [6-10]. Пусть у нас есть две случайные переменные X и Y, что принимают значения из наборов конечных алфавитов $A_x = \{x_1, x_2 \dots x_i \dots x_I\}$ и $A_y = \{y_1, y_2 \dots y_j \dots y_J\}$, соответственно. В таком случае условная энтропия определяется как:

$$H(Y|X) = -\sum_{i=1}^{I}\sum_{j=1}^{J} P(x_i, y_i) \cdot \log P(y_i|x_i) \,. \qquad (6)$$

Расписав функцию $P(x_i, y_i)$ через $P(y_i|x_i)$ и используя неравенство Дженсена для произведения $P(y_i|x_i) \cdot \log P(y_i|x_i)$, которое представляет собой вогнутую функцию, получим:

$$\begin{cases} H(Y|X) = -\sum_{i=1}^{I} P(x_i) \sum_{j=1}^{J} P(y_i|x_i) \cdot \log P(y_i|x_i) \\ H(Y|X) \geq H(Y|X) \end{cases} \qquad (7)$$

Данный подход, представленный системой уравнений и неравенств (7), может применяться для любого количества переменных (рис. 4):

$$\begin{cases} P(X_1,\ X_2,\ \dots,\ X_N) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \dots P(X_N|X_1, X_2, \dots, X_{N-1}) \\ H(X_1,\ X_2,\ \dots,\ X_N) = H(X_1) + H(X_2|X_1) + \dots + H(X_N|X_1, X_2, \dots, X_{N-1}) \end{cases} \qquad (8)$$

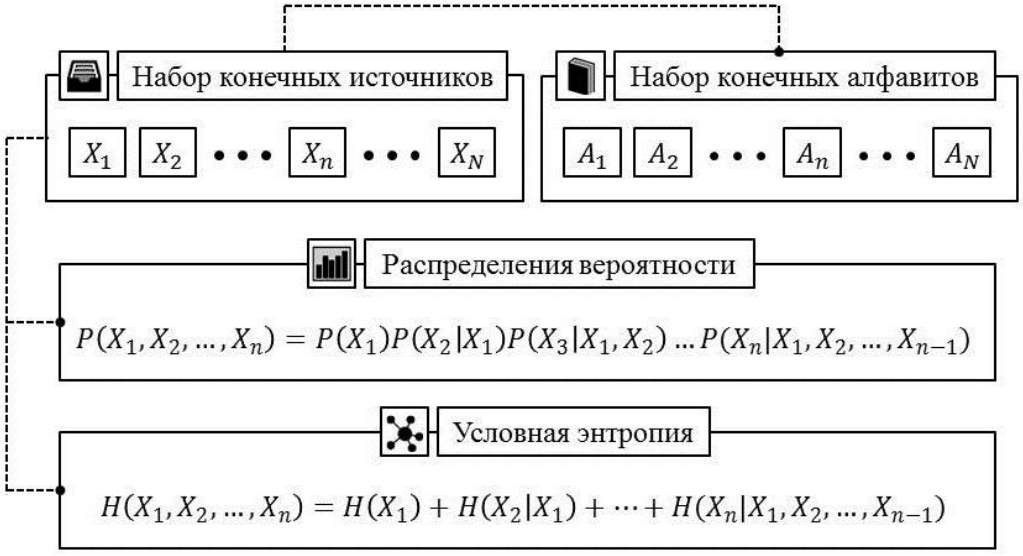

Рис. 4. Базовый алгоритм определения условной энтропии

Алгоритм, представленный на рис. 4, можно рассматривать как частный случай применения цепного правила (chain rule) для определения условной энтропии. Рассматриваемый подход позволяет выразить общий объем информации как сумму информационных вкладов от отдельной переменной X_n в том случае, когда известны значения предыдущих переменных $\{X_1, X_2 \dots X_{n-1}\}$. Данный алгоритм является эффективным инструментом математического анализа в том случае, когда переменные X_n не являются независимыми.

2. Схема определения общего количества информации источника

Выражения, полученные для расчета условной вероятности и условной энтропии, позволяют ввести понятие общего количества информации (mutual information). Так, для пары случайных переменных X и Y, которые принимают значения из наборов конечных алфавитов $A_x = \{x_1, x_2 \dots x_i \dots x_I\}$ и $A_y = \{y_1, y_2 \dots y_j \dots y_J\}$ общее количество информации определяется как [1, 11]:

$$\left[\begin{array}{l} I(X;Y) = \sum_{x=x_i}^{x_I} \sum_{y=y_j}^{y_J} P(x,y) \log\left({P(y|x)}/{P(y)} \right) \\ \left[\begin{array}{l} I(X;Y) = H(Y) - H(Y|X) \\ I(X;Y) = H(X) - H(X|Y) \end{array}\right. \end{array}\right. \quad (9)$$

Исходя из представлений об условной энтропии систему уравнений (8) можно также дополнить следующим набором условий:

$$\left[\begin{array}{l} I(X;Y) \geq 0 \\ I(X;Y) = I(Y;X) \end{array}\right., \quad (10)$$

причем $I(X;Y) = 0$ в том случае, когда X и Y независимы, и, $H(Y|X)$ и, значит, $H(X|Y)$ достигают максимально возможного значения ($H(X)$ и $H(Y)$, соответственно).

Соответственно, используя цепное правило для определения условной энтропии, можно получить выражение для общего количества информации, которое позволяет оперировать произвольным количеством переменных:

$$I(Y;X_1,\dots,X_n) = I(Y;X_1) + I(Y;X_2|X_1) + \cdots + I(Y;X_n|X_1,X_2,\dots,X_{n-1}) \quad (11)$$

На основе (8) и (10) можно вывести формулу для энтропии двух переменных, как функции от энтропии каждой из переменной и общего количества информации:

$$H(X,Y) = H(X) + H(Y) - I(X;Y) \quad (12)$$

Полученные уравнения позволяют доказать, что при обработке дискретного источника общее количество информации не увеличивается (рис. 5).

Рис. 5. Алгоритм определения общего количества информации после обработки данных дискретного информационного источника

Пусть Z — процесс обработки Y и, таким образом, связана с X только через Y. Тогда $I(X;Z|Y) = 0$, а $I(X;Y|Z) \geq 0$. Соответственно, при обработке Y общее количество информации $I(X,Y)$ не увеличивается $I(X;Z) \leq I(X;Y)$.

3. Обобщенный алгоритм кодирования дискретного источника

Для практического применения разработанной математической модели на базе понятия энтропии дискретного источника необходимо рассматривать анализ символьных блоков. Если отдельный информационный элемент состоит из N бинарных символов, то представление о нем можно составить через $N \cdot H$. Процесс анализа символьных блоков дискретного информационного источника будет называться кодирование источника (source coding).

Рассмотрим символьный набор бинарных элементов X^{01}, представленный множеством символов $\{x_i^{01}\}$ для $i \in [1;I]$, каждый из которых определяется длиной $l(x_i^{01})$. Для символьного набора можно определить значение ожидаемой длины [1, 2, 12]:

$$L(X^{01}) = \sum_{i=1}^{N} P(x_i^{01}) l(x_i^{01}) \qquad (13)$$

Далее в рамках развития модели предлагается использовать неравенство Крафта-Макмиллана, которое даёт необходимое условие существования разделимых и префиксных кодов, обладающих заданным набором длин кодовых слов:

$$\sum_{i=1}^{I} 2^{-l(x_i^{01})} \leq 1 \rightarrow \sum_{i=1}^{I} 2^{l_m - l(x_i^{01})} \leq 2^{l_m}, \qquad (14)$$

где l_m — максимальная длина символьного набора. Полученное неравенство выводится на основе того если строка имеет длину $l(x_i^{01})$, из условия префикса следует, что ни одно из $2^{l_m - l(x_i^{01})}$ не содержится в коде, а два расширения разных строк кода никогда не равны, так как это также нарушило бы условие префикса.

Рассмотрим тот случай, когда значение $\log P(x_i^{01})$ представлено целыми числами и соответственно может быть принято как эквивалент длины $l(x_i^{01})$. В таком случае неравенство Крафта-Макмиллана переходит в равенство и, таким образом, значение ожидаемой длины соответствует значению энтропии (рис. 6).

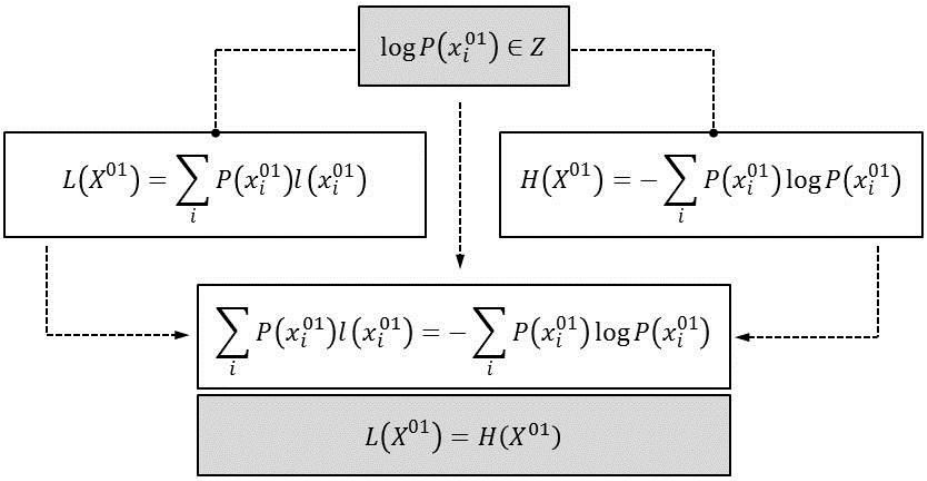

Рис. 6. Схема определения ожидаемой длины для целого $log\, P(x_i^{01})$

Для расширения диапазона допустимых значений $\log P(x_i^{01})$ необходимо ввести понятие $\lceil \log P(x_i^{01}) \rceil$ как округление $\log P(x_i^{01})$ до ближайшего целого числа (рис. 7), т.е. из системы уравнений и неравенств:

$$\begin{cases} l(x_i^{01}) = \lceil -log\, P(x_i^{01}) \rceil \leq -log\, P(x_i^{01}) + 1 \\ H(X^{01}) - L = \sum_i P(x_i^{01}) \left[log \frac{1}{P(x_i^{01})} - l_i \right] \\ \sum_i P(x_i^{01}) \left[log \frac{1}{P(x_i^{01})} - l_i \right] = \sum_i P(x_i^{01}) log \left(\frac{2^{-l_i}}{P(x_i^{01})} \right) \end{cases} \qquad (15)$$

и системы неравенств:

$$\begin{cases} \sum_i P(x_i^{01}) log \left(\frac{2^{-l_i}}{P(x_i^{01})} \right) \leq log \sum_i 2^{-l_i} \\ log \sum_i 2^{-l_i} \leq 0 \end{cases} \qquad (16)$$

можно получить:

$$H(X^{01}) \leq L \leq H(X^{01}) + 1 \qquad (17)$$

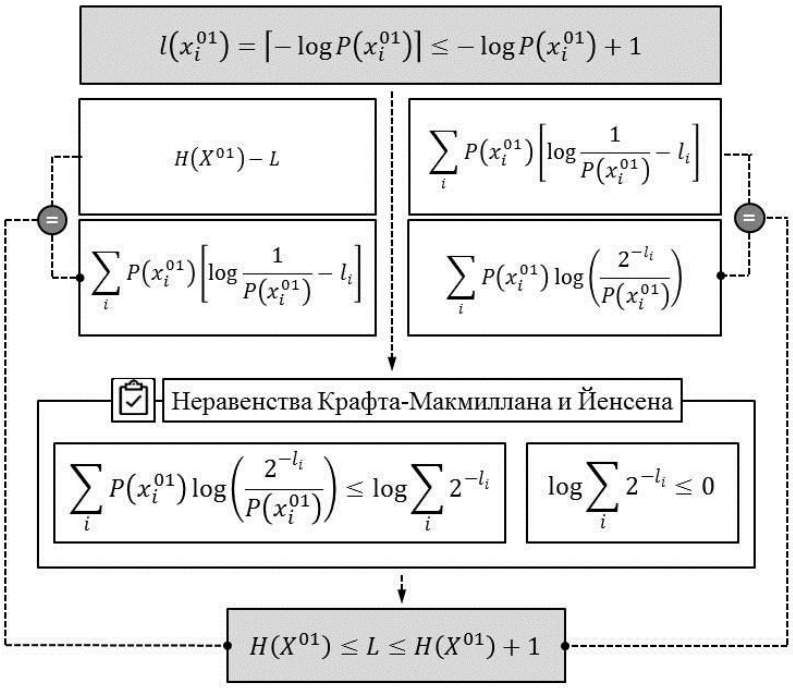

Рис. 7. Схема определения диапазона значений ожидаемой длины

Построенная обобщенная модель кодирования дискретного источника может быть использована при построении алгоритмов на основе методов кодирования Хаффмана и арифметического кодирования, в частности арифметического кодирования с конечной точностью и арифметического декодирования.

Выводы

В результате проведенного анализа были разработаны основы методологии кодирования дискретного информационного источника. В частности были предложены:

- базовый алгоритм определения условной энтропии;
- алгоритм определения общего количества информации после обработки данных дискретного источника;
- обобщенная схема определения диапазона значений ожидаемой длины для символьного набора дискретного источника.

Данную методологию можно использовать для разработки алгоритмов кодирования Хаффмана, арифметического кодирования и универсального кодирования.

Список литературы / References

1. *Csiszár I., & Körner J.*, 2015. Information theory: Coding theorems for discrete memoryless systems. Cambridge: Cambridge University Press.
2. *McEliece R.J.*, 2004. The theory of information and coding. Cambridge: Cambridge University Press.
3. *Zhong Y., Alajaji F. & Campbell L.L.*, 2007. Error Exponents for Asymmetric Two-User Discrete Memoryless Source-Channel Systems. 2007 IEEE International Symposium on Information Theory. doi:10.1109/ isit.2007.4557472.
4. *Haghighat J., Hamouda W. & Soleymani M.*, 2006. Random Binning and Turbo Source Coding for Lossless Compression of Memoryless Sources. *IEEE* Vehicular Technology Conference. doi:10.1109/vtcf.2006.366.
5. *Sungkar M. & Berger T.*, 2018. Discrete Reconstruction Alphabets in Discrete Memoryless Source Rate-Distortion Pblems., 2018 IEEE International Symposium on Information Theory (ISIT). doi:10.1109/isit.2018.8437835.
6. *Bissiri P. & Walker S.*, 2018. A Definition of Conditional Probability with Non-Stochastic Information. Entropy, 20 (8), 572. doi:10.3390/e20080572.
7. *Gu Y.H. & Wu W.M.*, 2011. DDoS Detection and Prevention Based on Joint Entropy and Conditional Entropy. Key Engineering Materials, 474-476, 2129-2133. doi:10.4028/www.scientific.net/kem.474-476.2129.
8. *Yan K.*, 2015. Conditional entropy and fiber entropy for amenable group actions. Journal of Differential Equations, 259(7), 3004-3031. doi:10.1016/j.jde.2015.04.013.

9. *Patil G.,* 2013. Conditional Entropy ProfilesBased in part on the article “Conditional entropy profiles” by G. P. Patil, which appeared in the Encyclopedia of Environmetrics. Encyclopedia of Environmetrics. doi:10.1002/9780470057339.val005m.pub2.
10. *Zhou X.*, 2016. A formula of conditional entropy and some applications. Discrete and Continuous Dynamical Systems, 36(7), 4063-4075. doi:10.3934/dcds.2016.36.4063.
11. *Zen, Q. & Wang J.,* 2017. Information Landscape and Flux, Mutual Information Rate Decomposition and Entropy Production. doi:10.20944/preprints201710.0067.v1.
12. *Wiegand T. & Schwarz H.,* 2011. Source coding: Part I of fundamentals of source and video coding. Boston: Now.
13. *Moffat A.,* 2016. Huffman Coding. Encyclopedia of Algorithms, 938-942. doi:10.1007/978-1-4939-2864-4_633.
14. *Tamir D.,* 2018. Delta-Huffman Coding of Unbounded Integers. 2018 Data Compression Conference. doi:10.1109/dcc.2018.00081.
15. *Salman N.H.,* 2016. New Image Compression/Decompression Technique Using Arithmetic Coding Algorithm. Journal of Zankoy Sulaimani - Part A, 19(1), 263-272. doi:10.17656/jzs.10604.
16. *Al-Rababaa A., Laval C.U. & Dube D.,* 2015. A finite-precision adaptation of bit recycling to arithmetic coding. 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). doi:10.1109/isspit.2015.7394382.