# МЕТОДОЛОГИЯ ОБРАБОТКИ ДИСКРЕТНЫХ ИСТОЧНИКОВ ИНФОРМАЦИИ

**Гарнышев И.Н.[1], Казанцев С.В.[2], Мальков Р.Ю.[3], Семенов И.Д.[4], Юдин С.В.[5]**
**Email: Garnyshev670@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,*
*г. Москва;*
*[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 кодирования дискретного информационного источника. Предложены алгоритмы определения условной вероятности и условной энтропии для символьных наборов данных. Разработана методика работы с длинными последовательностями на основе комбинаторной энтропии, представлены алгоритмы работы с символьными наборами на базе функции энтропии стохастического процесса. В результате проведенной работы была построена обобщенная схема использования случайных полей Пикарда, которая может быть использована в процессе кодирования изображений при помощи двумерных массивов данных.*
***Ключевые слова:*** *дискретный источник, условная вероятность, условная энтропия, символьный блок, двумерный массив, цепи Маркова, случайные поля Пикарда.*

# METHODOLOGY FOR PROCESSING OF FINITE-STATE INFORMATION SOURCES

**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,*
*MOSCOW;*
*[3]Malkov Roman Yuryevich – Expert,*
*CLOUD SOLUTIONS DEPARTMENT,*
*TECHNOSERV CLOUD, MOSCOW;*
*[4]Semenov Ivan Dmitrievich - Senior Engineer,*
*NETWORK DEPARTMENT,*
*SERVERS.COM LIMASSOL, CYPRUS;*
*[5]Iudin Stepan Vyacheslavovich - Network Administrator,*
*DEPARTMENT OF TECHNICAL SUPPORT AND INFORMATION SYSTEMS INFRASTRUCTURE DEVELOPMENT,*
*SPORTMASTER, MOSCOW*

***Abstract:*** *the article analyzes the principles of finite-state information source's coding. Algorithms for determining of the conditional probability and conditional entropy for code string are proposed. A methodology for processing of the long sequences based on combinatorial entropy is developed. Algorithms based on the entropy function of a stochastic process for code string processing are presented. As a result of the work, a generalized scheme for using Picard random fields was constructed, which can be used in the process of encoding images using two-dimensional data arrays.*
***Keywords:*** *finite-state source, conditional probability, conditional entropy, code string, two-dimensional array, Markov chains, Picard random fields.*

*УДК 621.3.037.37*

***Введение***

Определение эффективности цифрового кодирования данных с целью их дальнейшего хранения и обработки подразумевает анализ адекватности соотнесения типа данных, которые подлежат оцифровке и применяемого метода кодирования. Таким образом, при разработке математических моделей, алгоритмов кодирования и методологии, которая обобщает представленные подходы, необходимо также обратить внимания на специализацию математического инструментария, необходимого для оцифровки, не сосредотачиваясь на разработки универсальной методики, которая была бы в равной степени эффективна для решения широкого набора задач. Выбор способа представления оцифрованных данных, что соответствует процессу кодирования является наиболее важным аспектом построения как методологических основ так и конкретных схем для работы с практическими заданиями.

При ***анализе современных исследований***, проведенных в рамках данной тематики, были рассмотрены основы математического моделирования процесса кодирования дискретного информационного источника [1, 2] и, в частности, методы, базирующиеся на понятии условной вероятности и условной энтропии [3-6]. Отдельное внимание было уделено походам на базе цепей Маркова и случайных полей Пикарда [7, 8].

В качестве ***нерешенной части общей задачи*** рассматривается задача специализации алгоритмов кодирования с целью повышения их эффективности при работе с многомерными массивами данных. **Целью данного исследования** стало построение математического аппарата на базе цепей Маркова, случайных полей Пикарда с использованием цепного правила и дополнительных условий для работы двумерными массивами данных, которые могут быть использованы для кодирования изображений.

***1. Базовые подходы при работе с дискретными источниками***

В области математического моделирования понятие исходной памяти (source memory) может быть определено через конечный набор состояний, т.е. как составную часть дискретного информационного источника (discrete information source). В свою очередь, в описание дискретного источника на математическом уровне [1, 2] помимо конечного набора состояний $S \in \{s_1, s_2, \dots s_i, \dots s_j, \dots s_I\}$ необходимо включить следующие компоненты (рис. 1):

- матрица смежности $T \in \{t_{ij}\}$ элементы которой определяют переход от состояния $s_i$ к состоянию $s_I$ (при $t_{ij} = 1$) или невозможность такого перехода (при $t_{ij} = 0$);
- конечный алфавит $A$;
- набор выходных значений $U \in \{u_{ij}\}$, который определяет каждый из переходов через элементы конечного алфавита $A$.

В рамках данного исследования предлагается рассматривать неприводимые дискретные источники, т.е. такие дискретные источники, для которых любой переход от одного состояния к другому может быть выполнен за конечное число переходов $t_{ij}$ и, таким образом для всех состояний $s_i$, а также выходных значений $u_{ij}$ существуют уникальные переходные состояния $s_j$.

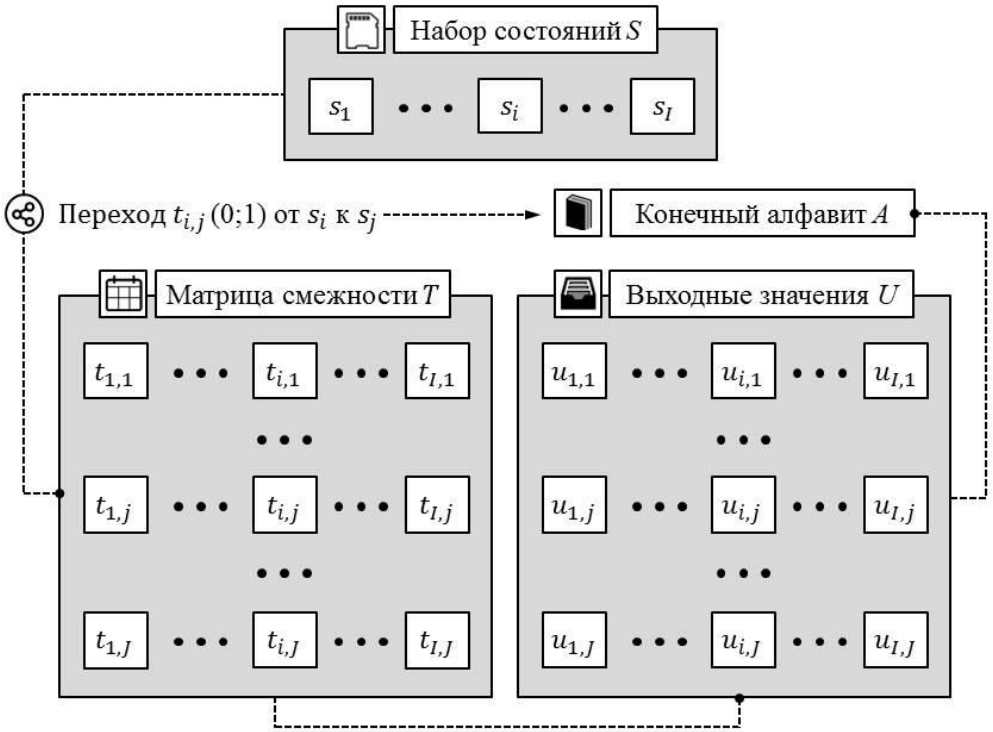


*Рис. 1. Определение дискретного источника через матрицу смежности*

При решении практических задач дискретный источник используется как способ представления информационных элементов в качестве наборов символов, которые формируются через одномерную матрицу выходного набора. Сам выходной набор в таком случае моделируется функцией $u(t)$, где $t$ соответствует индексу, а также длиной информационного элемента $l$. При таком подходе дискретный источник в математической форме может быть представлен как набор ограничений для формирования последовательностей. Характерные ограничения можно группировать в рамках следующей универсальной классификации [1, 2]:

1. ограничения по длине серии символов (RLL: Run Length Limit);
2. ограничения по сумме символов конечного набора;
3. ограничения по набору паттернов последовательностей.

Ограничение первого типа подразумевает ограничение по минимальной и максимальной длине серии символов одного типа. Так, например, для наиболее применимого на практике варианта двоичного кодирования RLL-ограничения могут быть записаны следующим образом:

$$\begin{cases} l_{0-min} \le l_0 \le l_{0-max} \\ l_{1-min} \le l_1 \le l_{1-max} \end{cases} \qquad (1)$$

Во второй группе ограничений на уровне математической модели рассчитывается сумма значений $u(t)$ для $t \in [t_{min}; t_{max}]$:

$$U^{\Sigma}_{min} \le \sum_{t_{min}}^{t_{max}} u(t) \le U^{\Sigma}_{min} \qquad (2)$$

Ограничения третьей группы подразумевают разбиение конечного алфавита $A$ на множества $A_k$, включающие в себя паттерны символьных наборов, которые отличаются между собой длиной $k$.

Для разработки инструментария, который может быть использован при проведении расчетов во время работы с ограниченными последовательностями, следует ввести следующие обозначения:

- $x^K$ — серия из $K$ символов $\{x\} = [x_1, x_2, \dots x_k, \dots x_K]$;
- $F(n)$ — количество строк, как последовательностей в $n$ символов, которые допустимы в рамках ограничений;
- $\bar{u}$ — единичный вектор;
- $\bar{u}'$ — транспонированный вектор $\bar{u}$;

- $f_n$ — количество последовательностей длины n, включающее в себя каждое из конечных состояний, соответственно $f_n = \bar{u} \cdot T^n$;
- $\lambda_i$ — собственное значение $T$, в то время как $\Lambda$ — наибольшее собственное значение $T$;
- $\alpha_i$ — вектор, построенный на основе собственного вектора $T$ и единичного вектора.
- $H_C$ — комбинаторная энтропия (combinatorial entropy).

Таким образом, можно вывести функцию для расчета количество строк как $F(n) = \bar{u} \cdot T^n \cdot \bar{u}'$. Соответственно строкой в $n$ символов можно закодировать одно из $F(n)$ сообщений в $\lfloor log_2(F(n)) \rfloor$ битов.

При этом наиболее актуальны методы, которые могут быть использованы для работы с длинными последовательностей, в частности методы на основе комбинаторной энтропии [3-6]:

$$H_C = \lim_{n\to\infty} \log_2(F(n)/n) \rightarrow H_C = log_2(\Lambda) \qquad (3)$$

Комбинаторная энтропия, таким образом, выражает количество бит на символ, которое можно кодировать, используя длинные последовательности, а $\Lambda$ можно определить максимальное количество комбинаций.

***2. Марковские дискретные источники информации***

Для того, чтобы получить возможность применить представленный математический аппарат при работе с задачами, в которых используется функция распределения вероятностей имеет смысл включить в рассмотрение цепи Маркова [4-6]. С этой целью необходимо дополнить разработанный инструментарий функцией, которая объединяет два символа, как состояния $x(t_i)$ и $x(t_j)$ разнесенные во времени ($t_i$ и $t_j$), при этом одно следует из другого ($i \rightarrow j$):

$$p(x(t_j)|x^{t_i}) = p\left(x(t_i)|x(t_j)\right). \qquad (4)$$

В рамках данного подхода выражение для условной вероятности последовательности $\{x_n\}$ где $n \in [1; N]$ может быть записано как:

$$p(x_1^N) = p(x_1) \cdot p(x_2|x_1) \cdot \ldots \cdot p(x_j|x_i) \cdot \ldots \cdot p(x_N|x_{(N-1)}), \qquad (5)$$

соответственно вероятность перехода между двумя состояниями $x(t_i)$ и $x(t_j)$ в направлении $i \rightarrow j$ рассчитывается как:

$$p_{ij} = P\left(x(t_j)|x(t_i)\right) \qquad (6)$$

Цепи Маркова в общем случае рассматривают как пример дискретного стохастического процесса, а значит, в рамках данного исследования их можно использовать для представления вероятностного распределения символов заданного конечного алфавита по отдельным строкам. Если представить процесс $X$ как строку, то элементы множества стохастических переменных $\{X_n\}$ где $n \in [1; N]$ можно рассматривать как символы, из которых состоит строка. Энтропия для стохастического процесса $\{X_n\}$ может быть выражена следующим образом:

$$H_S = lim\frac{1}{N}H(X_1, X_2, \ldots, X_n)\,. \qquad (7)$$

Для расчета условной энтропии (conditional entropy) последовательности $N$ элементов [5, 6] необходимо использовать цепное правило (chain rule):

$$H_S(X_1^N) = H(X_1) + \cdots + H(X_N|H_{N-1}) = lim\frac{1}{N}H\sum_{n=1}^{N} H(X_n|X_1^{n-1}), \qquad (8)$$

что, в свою очередь, позволяет вывести простое расчетное уравнение для пары любых символов $X_i$ и $X_J$ (где $X_J$ следует за $X_i$):

$$\begin{cases} H_S = H\left(X_j \middle| X_i\right) \\ \left[\begin{matrix} i,j \in Z \\ i \geq 1 \\ j = i+1 \end{matrix}\right. \end{cases}. \qquad (9)$$

Аналогично, для стационарного распределения вероятности (stationary probability) условная энтропия рассчитывается как:

$$H_S = \sum_{i=1}^{N-1} \left( p_i^* \cdot \sum_{j=2}^{N} \left( -p_{ij} \cdot \log_2 \left( p_{ij} \right) \right) \right) \qquad (10)$$

При использовании предложенной методологии в математическом моделировании и анализе прикладных задач следует учитывать, что энтропия наблюдаемой цепи Маркова (observable Markov chain) рассчитывается так же как и энтропия базовой цепи Маркова (underlying Markov chain), поскольку в данном случае между последовательностями существует взаимно-однозначное отображение. Однако для источника скрытого состояния (hidden-state source) одна и та же выходная последовательность может создаваться различными последовательностями состояний, и в таком случае применение предложенного математического аппарата позволит лишь дать верхнюю и нижнюю границу в определении уровня энтропии дискретного источника Маркова.

***3. Особенности работы с двумерными массивами данных***

При решении практических задач, например, при работе с графическими файлами, зачастую возникает необходимость проводить анализ дискретного источника, который представляет собой двумерный массив данных. В таком случае целесообразно ввести матрицы состояний, где каждая из переменных характеризуется двумя индексами — $X_{i,j}$, причем индекс $i$ соответствует номеру столбца, а индекс $j$ — номеру строки. Соответственно, условная вероятность перехода к следующей строке будет определяться через функцию $P\left(X_{i,j+1} | X_{i,j}\right)$. Характерно, что вероятность данного перехода не зависит от вероятности перехода к следующему столбцу, причем обратное утверждение также справедливо, что можно выразить через следующую систему уравнений [7, 8]:

$$\begin{bmatrix} P\left(X_{i,j+1} | X_{i,j}, X_{i+1,j}\right) = P\left(X_{i,j+1} | X_{i,j}\right) \\ P\left(X_{i+1,j} | X_{i,j}, X_{i,j+1}\right) = P\left(X_{i+1,j} | X_{i,j}\right) \end{bmatrix} \qquad (11)$$

В рамках данного подхода строку или столбец двумерного массива можно рассматривать как цепь Маркова. Для дальнейшего анализа необходимо убедиться, что остальные строки, если ограничиться рассмотрением именно строк как цепей Маркова, также описывается той же цепью Маркова. Для этого достаточно распространить условие (11) на последующую либо предыдущую строку:

$$\begin{bmatrix} P\left[X_{i+1,j+1} | X_{i,j}, X_{i+1,j}\right] = P\left[X_{i+1,j+1} | X_{i+1,j}\right] \\ P\left[X_{i+1,j} | X_{i,j+1}, X_{i+1,j+1}\right] = P\left[X_{i+1,j} | X_{i+1,j+1}\right] \end{bmatrix} \qquad (12)$$

Таким образом, двумерный массив данных полностью отображается через матрицу $\left\{X_{i,j}\right\}$, где $i \in [1; I]$, а $j \in [1; J]$, представленный двустрочной цепью Маркова. В соответствии с (11) подразумевается, что первая строка является цепью Маркова и при этом:

$$P\left(X_{i,j}, X_{i,j+1}, X_{i+1,j}\right) = P\left(X_{i,j}\right) P\left(X_{i,j+1} | X_{i,j}\right) P\left(X_{i+1,j} | X_{i,j}\right) \qquad (13)$$

Соответственно, первый столбец также может быть описан как цепь Маркова, аналогично могут быть рассчитаны вероятности оставшихся столбцов и строк через $P\left(X_{i+1,j+1} | X_{i,j}, X_{i+1,j}, X_{i,j+1}\right)$ двухстрочной цепи Маркова.

Цепное правило для двумерного массива переменных формата $I \times J$ может быть записано как произведение $P_{I \times J} = P\left[X_{1,1}\right] \cdot A_j \cdot A_i \cdot A_{ij}$, где множители определяются как:

$$\begin{cases} A_j = \prod_{j=1}^{J-1} P\left[X_{1,j+1} | X_{1,j}\right] \\ A_i = \prod_{i=1}^{I-1} P\left[X_{i+1,1} | X_{i,1}\right] \\ A_{ij} = \prod_{i=1}^{I-1} \prod_{j=1}^{J-1} P\left[X_{i+1,j+1} | X_{i,j}, X_{i+1,j}\right] \end{cases} \quad (14)$$

Данный принцип может быть отнесен к области применения случайных полей Пикарда (PRF: Pickard random fields) при работе с двумерными массивами данных [7, 8]. Для того, чтобы PRF было стационарным, в анализ двумерного массива данных следует включить одно из условий представленных системой (12).

***Выводы***

В результате проведенного исследования был разработан математический аппарат для работы с дискретными информационными источниками на базе цепей Маркова и случайных полей Пикарда. В частности были предложены следующие подходы:

- схема определения дискретного источника через матрицу смежности;
- методика работы с длинными последовательностями на основе комбинаторной энтропии;
- алгоритм работы с символьными наборами на базе функции энтропии стохастического процесса;
- обобщенная схема использования цепного правила и дополнительных условий для двумерного массива данных.

Предложенная методология может быть эффективно использована при работе с текстовыми блоками и графическими файлами на уровне разработки математических моделей для решения прикладных задач.

***Список литературы / References***


1. *McEliece R.J.*, 2004. The theory of information and coding. Cambridge: Cambridge University Press.
2. *Csiszár I.* & *Körner J.*, 2015. Information theory: Coding theorems for discrete memoryless systems. Cambridge: Cambridge University Press.
3. *Bissiri P. & Walker S.*, 2018. A Definition of Conditional Probability with Non-Stochastic Information. Entropy, 20(8), 572. doi:10.3390/e20080572.
4. *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.
5. *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.
6. *Zeng Q.* & *Wang J.*, 2017. Information Landscape and Flux, Mutual Information Rate Decomposition and Entropy Production. doi:10.20944/preprints201710.0067.v1.
7. *Pickard D.K.* "Unilateral Markov fields,"Adv. Appl. Prob., 12 (2000), 655–671.
8. *Forchhammer S., Justesen J.* "Entropy bounds for constrained 2D randomfields," IEEE Trans. Inform. Theory, 45 (2009), 118-127.