УДК 621.391.1

О РАЗЛОЖЕНИИ БЕРНУЛЛИЕВСКИХ ИСТОЧНИКОВ ИНФОРМАЦИИ

Борисенко А.А., проф.

Проблема кодирования источников сообщений в теории информажии занимает центральное место. Они подразделяются на два класса: комбинаторные и вероятностные [1]. В комбинаторных сообщения делятся на запрещенные и разрешенные. При этом вероятности разрешенных равны между собой. Вероятностные источники в отличие от комбинаторных не имеют запрещенных комбинаций, а разрешенные могут отличаться своими вероятностями. Если при этом генерируемые сообщения независимы друг от друга, то такие источники называются источниками Бернулли [1]. Они сравнительно грубо моделируют реальные процессы генерирования информации, но в силу своей простоты имеют довольно широкое применение на практике. Собственно для этих источников Шеннон предложил свою формулу вычисления энтропии [2]:

$$H = -\sum_{j=1}^{N} P_j \log_2 P_j \,, \qquad (1)$$

где P_j - вероятность передаваемых сообщений;

N - их число.

Однако в случае большого количества сообщений $N = k^n$, где n - длина сообщений; k - число букв в алфавите $A = \{a_1, a_2, \ldots, a_k\}$ источника информации, возникают значительные вычислительные сложности, как при вычислении (1), так и при решении задач оптимального кодирования. Кроме того, на практике вероятности P_j могут значительно изменяться в зависимости от формы передаваемой информации - графической, текстовой, документальной, художественной и т.д. Поэтому известные методы оптимального кодирования [2] в данном случае не могут быть применены, так как они рассчитаны на постоянное значение P_j. В таких случаях применяется универсальное кодирование [1]. Это кодирование рассчитано на некоторый разброс вероятностей каждого сообщения и является оптимальным при некоторой минимальной не равной нулю остаточной избыточности. Оно является обобщением обычных методов оптимального кодирования, рассчитанных на постоянные P_j, а значит, в своей основе использует префиксное кодирование. В то же время существуют методы сжатия информации, например нумерационные [3], рассчитанные на случай равновероятных сообщений. Эти методы не охватываются универсальным кодированием, что сужает область их применения.

В данной работе предлагается метод оптимального кодирования, использующий в своей основе разложение бернуллиевского источника на два взаимосвязанных, решающих задачи снижения сложности вычисления энтропии и равновероятного универсального кодирования.

В основу метода положено преобразование вероятностного источника A^*, генерирующего двоичные сообщения длиной n из их общего числа 2^n, в два других - источники A и B. Смысл такого преобразования состоит в том, что исходное множество из 2^n двоичных сообщений разбивается на $(n+1)$ классов эквивалентности в соответствии с соотношением

$$2^n = C_n^0 + C_n^1 + \ldots + C_n^k + \ldots + C_n^n \,. \qquad (2)$$

Представителем класса эквивалентности в этом случае выступает число К единиц в двоичных кодовых комбинациях.

В результате каждое из 2^n двоичных сообщений источника представляется в виде числа K, содержащихся в этом сообщении единиц и относящейся к нему равновесной кодовой комбинации. Соответственно источник A генерирует эти комбинации, а источник B числа K. При этом число генерируемых сообщений в источниках A и B сокращается от 2 до n и, кроме того, как будет показано ниже, равновесные кодовые комбинации, принадлежащие к одному и тому же классу эквивалентности, будут равновероятными и, следовательно, к ним можно применять структурные методы сжатия, использующие коды с равной длиной слов.

Утверждение 1. Вероятностный стационарный бернуллиевский источник A^* с длиной генерируемых слов n и энтропией

$$H(A^*) = -\sum_{j=1}^{2^n} p_j \log_2 p_j , \qquad (3)$$

где p_j - вероятность генерирования источником A^* j-й двоичной кодовой комбинации, представим вероятностным бернуллиевским источником B с энтропией

$$H(B) = -\sum_{k=1}^{n} \hat{P}_k \log_2 \hat{P}_k , \qquad (4)$$

где $\hat{P}_k = \sum_{j=1}^{C_n^k} P_{kj}$ - вероятность появления $K = 0, 1, \ldots, n$ единиц в генерируемой источником A^* j-й, $j = 1, 2, \ldots, 2^n$, двоичной кодовой комбинации;

P_{kj} - вероятность генерирования источником j-й, $j = 1, 2, \ldots, C_n^k$, двоичной кодовой комбинации с К единицами;

$$P_{kj} = P_k = q^k (1-q)^{n-k} ,$$

q - вероятность появления единицы в двоичной кодовой комбинации j, и конечном комбинаторном источнике A, генерирующим двоичные равновесные слова длины n, с условной энтропией

$$H(A/B) = \sum_{k=1}^{n} \hat{P}_k \log_2 C_n^k . \qquad (5)$$

Доказательство. Так как $P_{kj} = P_k$ для всех двоичных кодовых комбинаций с K единицами, то

$$\hat{P}_k = \sum_{j=1}^{C_n^k} P_{kj} = C_n^k P_k \qquad (6)$$

и соответственно

$$H(B) = -\sum_{k=0}^{n} C_n^k P_k \log_2 C_n^k P_k .$$

Условная энтропия источника A

$$H(A/B) = -\sum_{k=0}^{n} \sum_{k=1}^{C_n^k} \hat{P}_k P_{j/k} \log_2 P_{j/k} , \qquad (8)$$

где $P_{j/k}$ - условная вероятность генерирования источником A равновесной кодовой комбинации при условии, что для нее источником B определено число единиц K.

Так как

$$\hat{P}_k P_{j/k} = P_k , \qquad (9)$$

то

$$P_{j/k} = \frac{P_k}{\hat{P}_k} = \frac{P_k}{P_k C_n^k} = \frac{1}{C_n^k} . \qquad (10)$$

Соответственно

$$H(A/B) = \sum_{k=0}^{n} \hat{P}_k \log_2 C_n^k . \qquad (11)$$

Взаимная энтропия

$$H(A,B) = H(A/B) + H(B) = -\sum_{k=0}^{n} C_n^k P_k \log_2 C_n^k P_k +$$

$$+ \sum_{k=0}^{n} C_n^k P_k \log_2 C_n^k = -\sum_{k=0}^{n} C_n^k P_k \log_2 P_k \quad . \qquad (12)$$

Так как каждому значению $\hat{P}_k$ соответствует C_n^k вероятностей $P_j = P_k$ исходного источника A, то

$$H(A/B) + H(B) = -\sum_{k=1}^{n} C_n^k P_k \log_2 P_k = -\sum_{j=1}^{2^n} P_j \log_2 P_j = H(A^*) \quad , \qquad (13)$$

что и требовалось доказать.

Примем без доказательства следующее утверждение.

Утверждение 2. Максимального значения энтропии $H(A/B)$ и $H(B)$ достигают при $q = 0,5$ и соответственно равны:

$$H_{\max}(A/B) = \sum_{k=0}^{n} \frac{C_n^k}{2^n} \log_2 C_n^k ; \qquad (14)$$

$$H_{\max}(B) = \sum_{k=0}^{n} \frac{C_n^k}{2^n} \log_2 \frac{C_n^k}{2} .$$

Объединенная величина избыточной информации, содержащейся в источниках A и B

$$I(A,B) = I(A^*) = n - H(A,B) = H_{\max}(A,B) - H(A,B) =$$
$$= H_{\max}(B) - H(B) + H_{\max}(A/B) - H(A/B) =$$
$$= I(B) + I(A) = \sum_{k=1}^{n} \hat{P}_k \log_2 2^n P_k .$$

Значения избыточностей источников A и B определяются из выражений:

$$I(A) = H_{\max}(A/B) - H(A/B) = \sum_{k=0}^{n} (2^{-n} - P_k) C_n^k \log_2 C_n^k ; \qquad (17)$$

$$I(B) = I(A,B) - I(A) = n - H(A,B) - I(A) = H_{\max}(B) - H(B) =$$
$$= -\sum_{k=0}^{n} \frac{C_n^k}{2^n} \log_2 \frac{C_n^k}{2^n} + \sum_{k=0}^{n} \hat{P}_k \log_2 \hat{P}_k . \qquad (18)$$

Таким образом, разбиение двоичных сообщений бернуллиевского вероятностного источника на классы эквивалентности с признаком числа K единиц в двоичных кодовых комбинациях преобразует его в два новых бернуллиевских источника, один из которых вероятностный, а второй - комбинаторный, с уменьшением числа слагаемых в выражениях для их энтропий от 2^n до n. В результате число операций при вычислении энтропий этих источников уменьшается от 2^n до $2n$, что значительно уменьшает вычислительные трудности. Кроме того, наличие комбинаторного источника A позволяет решить задачу его оптимального кодирования равномерными кодами.

SUMMARY

The method increasing the optimal coding speed for a probable source is proposed. The probable source having large length messages is transformed into two interdependent universal sources, one of which is a combinatorial one.

СПИСОК ЛИТЕРАТУРЫ

1. Кричевский Р.Е. Сжатие и поиск информации.-М.: Радио и связь, 1989.- 168 с.
2. Галлагер Р. Теория информации и надежная связь /Пер. с англ. - М.: Сов. радио, 1974.- 720 с.
3. Амелькин В.А. Методы нумерационного кодирования.- Новосибирск: Наука, 1986.- 156 с.

Поступила в редколлегию 9 октября 1995 г.