звеном является разработка новой учебной программы в условиях университетского образования, например, в Сумском государственном университете, где имеются такие факультеты, как медицинский,физико-технологический, инженерный, экономический с кафедрами электроники, теоретической и прикладной физики, математики и других. Практическому здравоохранению жизненно необходимы врачи с инженерным образованием и инженеры с медицинским. Поэтому существенным представляется и создание соответствующих курсов освоения смежных профессий в упомянутых отраслях.

Уже сегодня комплексное решение проблемы осуществляется путем проведения совместных научных, исследовательских, проектных и практических работ кафедр хирургического и терапевтического профиля медицинского факультета и промышленной электроники Сумского государственного университета, а также Сумского производственного объединения "Электрон" - АО "Selmi". Кафедра промышленной электроники может подготовить достаточное количество специалистов по приборам медицинской техники и способна создать комплексную лабораторию для проведения как учебного процесса, так и научно-исследовательской работы. При этом приборостроители ПО "Электрон", имеющие большой опыт создания сложных научно-технических комплексов, смогут освоить серийный выпуск отдельных приборов и комплексных модулей для крупных клиник, лечебных и диагностических центров.

Таким образом, в настоящее время на Украине имеются реальные возможности создать учебно-научно-производственное объединение с целью решения неотложных задач медицины с помощью электронной техники.

SUMMARY

There is a real opportunity in Ukraine to build up a science technology centre for developments of modern original medicinal diagnostic equipment and for creation of diagnostic laboratories, using electron microscopes,mass-spectrometers and analytical equipment. Besides,the development of information, control and artifical intelligense systems with the help of PC is also available.

Поступила в редколлегию 28 декабря 1995 г.

УДК 621.391.1

КОМБИНАТОРНОЕ СЖАТИЕ ИНФОРМАЦИИ

Борисенко А.А., проф.

Сжатие информации, генерируемой вероятностными источникамиинформации, является предметом анализа во многих работах по теории информации, например [1]. Однако использование на практике результатов этих исследований сталкивается с трудной задачей определения конкретных значений вероятностей $p(a_i)$ букв a_i алфавита $A=\{a_1,a_2,\ldots,a_i,\ldots a_k\}$ для генерируемых слов. Обычно для ее решения используется метод статистических испытаний. Однако этот метод, кроме своей трудоемкости, обладает и значительными погрешностями в определении вероятностей букв. Это вызвано как ограниченностью величин выборок, на которых производятся испытания, так и структурными особенностями передаваемой информации, которая обычно представлена последовательностью кадров информации - структурированных информационных единиц - текстов,

телевизионных изображений, графиков, таблиц и т.п. Поэтому среднее распределение частот букв на всей их бесконечной последовательности может резко отличаться от распределения частот букв a_i в кадрах информации.

Например, по факсимильной связи передается изображение чертежа, а за ним поясняющая его текстовая информация. Очевидно, что частоты появления букв в этих случаях и даже алфавиты, в которые они входят, будут значительно отличаться. Следовательно, для большей эффективности кодирования важнее использовать распределение частот букв внутри кадра, чем среднее значение частот в бесконечной последовательности этих букв.

Для решения задач оптимального кодирования таких структурированных сообщений, как в приведенном примере, предлагается избыточное универсальное кодирование [1]. Избыточность кода можно значительно уменьшить, если оптимальное кодирование производить применительно к каждому кадру информации в отдельности. Поток информации для этого бесконечного вероятностного источника представляется в виде неограниченной последовательности сообщений кадров, а сам вероятностный источник преобразуется в два взаимосвязанных - конечный комбинаторный источник и источник чисел букв в сообщениях длины n.

Конечным комбинаторным источником называется источник, порождающий с равной вероятностью $p(L)=1/n$ слова

$$L=l_1,l_2,\ldots,l_j,\ldots,l_n;\ l_j \in A=\{a_1,a_2,\ldots,a_k\}.$$

Для такого источника характерно, кроме наличия множества R разрешенных слов L, наличие множества $A^n \backslash R$ запрещенных слов $\hat{L}=\hat{l}_1,\hat{l}_2,\ldots,\hat{l}_j,\ldots,\hat{l}_n$, $\hat{L} \in A^n \backslash R$, $\hat{l}_j \in A$.

Разрешенные слова задаются предикатом p на словах множества A^n. Источник чисел букв генерирует информацию об этом предикате. Его наличие позволяет выделить конкретный кадр из генерируемой вероятностным источником последовательности кадров и вычислить вероятности $p(a_i)$ букв в этом кадре. В простейшем случае источник чисел букв генерирует числа n_i букв a_i, содержащихся в генерируемом комбинаторном источнике – кадре информации. Очевидно, что в этом случае

$$\sum_{i=1}^{k} n_i = n \qquad (1)$$

Зная величины n_i, можно вычислить количество последовательностей в множестве R:

$$|R| = \frac{n!}{n_1!\ n_2!\ \ldots\ n_i!\ \ldots n_k!}. \qquad (2)$$

Любые дополнительные ограничения, входящие в предикат p, могут только уменьшить значение |R| в пределе до 1. Например, одним из таких ограничений может быть условие, что сумма кодов букв a_i в генерируемом кадре информации должна равняться некоторому целому числу Z.

Произведем разбиение множества последовательностей R на K классов эквивалентности Ra_{i_1}, представителем каждого из которых будет одна из букв a_{i_1} алфавита A, так что

$$|R|=\sum_{i=1}^{k}|Ra_{i_1}|,\ |Ra_{i_1}|\geq 1. \qquad (3)$$

Отношение $|Ra_{i_1}|/[R]$ в этом случае будет вероятностью $p(a_{i_1})$ появления первой буквы a_{i_1} в генерируемом конечным комбинаторным источником кадре информации:

$$p(a_{i_1}) = |Ra_{i_1}|/|R|. \qquad (4)$$

Для рассматриваемых ограничений в виде чисел букв в сообщениях

$$|Ra_{i_1}| = \frac{(n-1)!}{n_1!\ n_2!\ \dots\ (n_i-1)!\ \dots n_k!} \qquad (5)$$

и соответственно из (2) и (4) следует, что

$$p(a_{i_1}) = n_i/n. \qquad (6)$$

Продолжим разбиение полученных классов на новые подклассы до тех пор, пока последний подкласс будет содержать всего лишь одну последовательность. Каждый из подклассов в этом случае кодируется одной из букв алфавита A: $a_{i_1 i_2},\dots, a_{i_1 i_2\dots i_j},\dots, a_{i_1 i_2\dots i_j\dots i_n}$.

Очевидно, что

$$|Ra_{i_1 i_2\dots i_{j-1}}| \geq |Ra_{i_1 i_2\dots i_j}|. \qquad (7)$$

При этом вероятность появления буквы на j-ом такте генерирования

$$p(a_{i_1 i_2\dots i_j}) = \frac{|Ra_{i_1 i_2\dots i_j}|}{|Ra_{i_1 i_2\dots i_{j-1}}|}. \qquad (8)$$

Так как в общем случае $|R|\neq|A^n|$, то появление той или иной буквы $a_{i_1 i_2\dots i_j}\in A$ зависит от уже сформированных предшествующих ей букв в генерируемой последовательности.

Поэтому конечный комбинаторный источник следует рассматривать как марковский n-го порядка, а вероятности $p(a_{i_1 i_2\dots i_j})$ – как условные вероятности $p(a_{ij}/i_1 i_2\dots i_{j-1})$, которые могут принимать на разных тактах генерирования разные значения.

Вероятность p(L) каждой последовательности L из множества |R| равна произведению вероятностей $p(a_{i_1 i_2\dots i_j})$ всех букв,входящих в нее:

$$p(a_{i_1})p(a_{i_1 i_2})\dots p(a_{i_1 i_2\dots i_j})\dots p(a_{i_1 i_2\dots i_n}) = 1/n. \qquad (9)$$

Знание вероятностей букв $a_{i_1 i_2\dots i_n}$ на каждом такте генерирования позволяет решать задачу оптимального кодирования кадра информации с целью его максимального сжатия.

Пределы сжатия определяются энтропией сообщений слов L, которая равна сумме энтропий букв на всех тактах их генерирования.

На первом такте это будет безусловная энтропия

$$h_1 = -\sum_{i_1=1}^{k} p(a_{i_1}) log p(a_{i_1}) = -\sum_{i_1=1}^{k} \frac{|Ra_{i_1}|}{|R|} log_2 \frac{|Ra_{i_1}|}{|R|}. \qquad (10)$$

На втором такте энтропия буквы будет уже условной:

$$h_1=-\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\frac{|Ra_{i_1i_2}|}{|R|}\log_2\frac{|Ra_{i_1i_2}|}{|Ra_{i_1}|}. \qquad (11)$$

Соответственно на j -ом такте энтропия

$$h_j=-\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\dots\sum_{i_j=1}^{k}\frac{|Ra_{i_1i_2\dots i_j}|}{|R|}\log_2\frac{|Ra_{i_1i_2\dots i_j}|}{|Ra_{i_1i_2\dots i_{j-1}}|}. \qquad (12)$$

а на n-ом

$$h_n=\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\dots\sum_{i_{n-1}=1}^{k}\frac{|R_{i_1i_2\dots i_{n-1}}|}{|R|}\log_2|Ra_{i_1i_2\dots i_{n-1}}|. \qquad (13)$$

Очевидно, что $h_{n+1}=0$. Это условие является признаком конца кадра. Совместная энтропия двух букв a_{i_1} и a_{i_2}:

$$H(L_2)=h_1+h_2=-\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\frac{|Ra_{i_1}||Ra_{i_2}|}{|R||Ra_{i_1}|}\log_2\frac{|Ra_{i_1i_2}|}{|R|}- \qquad (14)$$

$$-\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\frac{|Ra_{i_1}||Ra_{i_1i_2}|}{|Ra_{i_1}||R|}\log_2\frac{|Ra_{i_1i_2}|}{|Ra_{i_1}|}=-\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\frac{|Ra_{i_1i_2}|}{|R|}\log_2\frac{|Ra_{i_1i_2}|}{|R|}.$$

Соответственно

$$H(L_j)=h_1+h_2+\dots+h_j=-\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\dots\sum_{i_j=1}^{k}\frac{|Ra_{i_1i_2\dots i_j}|}{|R|}\log_2\frac{|Ra_{i_1i_2\dots i_j}|}{|R|} \qquad (15)$$

и

$$H(L_n)=-\sum_{i_1=1}^{k}\sum_{i_2=1}^{k}\dots\sum_{i_n=1}^{k}\frac{1}{|R|}\log_2\frac{1}{|R|}=\log_2|R|. \qquad (16)$$

Величина избыточной информации в генерируемых последовательностях L:

$$I=n\log_2 k-\log_2|R| \qquad (17)$$

и, соответственно, коэффициент сжатия

$$\mu=\frac{I}{H}=\frac{\log_2 k^n/|R|}{\log_2|R|} \qquad (18)$$

Таким образом, сущность предлагаемого метода сжатия информации состоит в представлении сжимаемых последвательностей в виде кадров информации, определения условных вероятностей появления тех или иных букв в них и после этого оптимальном кодировании указанных букв на каждом такте на основе известных методов оптимального кодирования Шеннона - Фано или Хаффмэна [4].

Этот метод представим в виде конкретного **алгоритма:**

1. Определяется длина кодируемого кадра информации.
2. Находятся значения $|Ra_{i_1 i_2 \ldots i_j}|$ и $|Ra_{i_1 i_2 \ldots i_{j-1}}|$, $j=1$.
3. Вычисляются вероятности $p(a_{i_1 i_2 \ldots i_j})$,
   если $|Ra_{i_1 i_2 \ldots i_{j-1}}| > 1$.
   В случае, если $|Ra_{i_1 i_2 \ldots i_{j-1}}| = 1$,то конец.
4. Генерируется с вероятностью $p(a_{i_1 i_2 \ldots i_j})$ буква $a_{i_1 i_2 \ldots i_{j-1}}$.
5. Производится кодирование буквы $a_{i_1 i_2 \ldots i_{j-1}}$ одним из известных методов оптимального кодирования.
6. $j=j+1$. Переход к пункту 2.

Как следует из алгоритма, рассматриваемый метод сжатия обладает повторяемостью выполняемых операций, относительной простотой, не требует определения частот появления букв, так как вероятности вычисляются, учитывает зависимость между буквами, удобен для аппаратной реализации. Особым свойством метода является то, что его сложность не растет с увеличением длины слов. Все это приводит к повышению эффективности средств сжатия информации, использующих этот метод, по сравнению с аналогичными средствами, использующими известные методы оптимального кодирования.

Эффективность метода показана в работах [2,3] при решении задач сжатия двоичных изображений и текстовой информации.

SUMMARY

The method of data compression based on the being methods of optimal coding is supposed. Its features are universality, adaptability to the symbols' frequency, accounting dependencies between the symbols, the simple technical realization, the inconsiderable increase of computing when compessing the long dinary sequences. The used method's idea is the transformation of the stochastic source into the combinatorial one and the secondary stochastic one that has considedably the lesser quantity of the codewords. Then the messages of the combinatorial source are compressed by separating into the equivalent classes and the step-type compression.

СПИСОК ЛИТЕРАТУРЫ

1. Кричевский Р.Е. Сжатие и поиск информации. - М: Радио и связь, 1989.- 168 с.
2. Борисенко А.А., Губарев С.И., Стеценко Л.А. Об одном способе сжатия текстовой информации / Проблемы бионики / Респ. межвед. научн.- техн. сб.- Харьков: Выща школа, ХГУ. - 1984, вып. 33, с. 52-54.
3. Борисенко А.А., Губарев С.И. О некоторых возможностях позиционных систем счисления / АСУ и приборы автоматики / Респ. межвед. научн. - техн. сб.- Харьков: Выща школа, ХГУ. - 1987, вып. 84, с. 95-99.
4. Цымбал В.П. Теория информатики и кодирование. - Киев: Выща школа.- 1977, 288 с.

Поступила в редколлегию 6 февраля 1996 г.