ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ

Дубна

Р5 - 3985

Денеш Йожеф

НЕКОТОРЫЕ ТЕОРЕТИЧЕСКИЕ РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ КОМПРЕССИИ ИНФОРМАЦИИ

ЛАБОРАТОРИЯ ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ
И АВТОМАТИЗАЦИИ

1968

Денеш Йожеф P5-3985

Некоторые теоретические результаты экспериментов компрессии информации

В работе излагаются основные методы эффективного кодирования, применявшиеся для сжатия экспериментальной информации в ЦИФИ (Будапешт). Приведены блок-схема программы, оптимизирующий метод эффективного кодирования с учётом статистической неоднородности информации и некоторые результаты экспериментов. Изложена схема нового принципа сжатия информации, допускающей коммутативность (перестановку символов.).

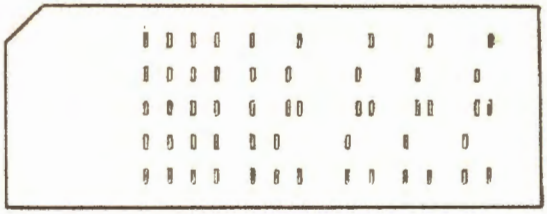
Препринт Объединенного института ядерных исследований.
Дубна, 1968.

Denes J. P5-3985

Some Theoretical Results of the Experiments on the Information Compression

The general methods of the effective coding, used for compression of the experimental information in ЦИФИ (Budapest), are presented. The block-diagram of the programme, which optimizes the method of effective coding with the account of statistic inhomogeneity of the information, and some experimental results are given. The new diagram for the information compression, permitting the symbols to be commutated, is considered.

**Preprint. Joint Institute for Nuclear Research.
Dubna, 1968**



P5 - 3985

Объединенный институт
ядерных исследований

ЛВТА

Денеш Йожеф*

НЕКОТОРЫЕ ТЕОРЕТИЧЕСКИЕ РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ КОМПРЕССИИ ИНФОРМАЦИИ

*ЦИФИ, Будапешт

Основой настоящей работы послужила лекция автора, прочитанная им в Объединенном институте ядерных исследований 30 мая 1968 г.

I. Основные понятия.

Обозначим элементы какого-нибудь множества H через I_1, $I_2 \ldots I_r$, где элементы H представляют какие-то данные, а K — множество, содержащее векторы $V_i = (a_1, a_2, \ldots a_n), i = \overline{1,r}$. Тогда преобразование H $\longrightarrow$ K будем называть кодированием, обратное ему — декодированием, где K — код и, наконец, каждое V_i называется кодовым словом длины n. Если n постоянно, то преобразование H $\longrightarrow$ K называется кодированием с постоянной длиной. Мы ограничим себя случаем, когда $a_j \ (j = \overline{1,n})$ может принимать значения 0 и I. В этом случае преобразование H $\longrightarrow$ K называется двоичным кодированием.

Любому элементу V_i соответствует вероятность p_i, с которой он встречается. Вероятности p_i образуют распределение P вероятностей, т.е. $\sum_{i=1}^{r} p_i = 1$. Если p_i имеют значительное отклонение от равномерного распределения (термин "значительное отклонение" определен точно в совместной работе К.Шерман и автора, см. /4/), то кодирование с постоянной длиной слова не приносит выигрыша в скорости передачи.

В последующей части нашей работы мы допустим, что P имеет значительное отклонение от равномерного распределения, и изучим эффективность преобразования H $\longrightarrow$ K' в зависимости от P, где K' —

код с переменной длиной слова. Эта проблема изучалась несколькими авторами, а результаты были суммированы Д.А.Новиком в его книге /12/ .

С помощью преобразования H $\longrightarrow$ K′ длина сообщения становится короче. Этот процесс для удобства будем называть компрессией данных. K' называется кодом, если он допускает однозначную дешифрацию, т.е., если V_j' и V_k' — два произвольных кодовых слова кода K', то послание $V_j' V_k'$ может быть однозначно разделено на отдельные кодовые слова K' . Это условие для кодов с постоянной длиной слова всегда справедливо, но для кодирования с переменной длиной — не всегда. Это можно проверить на следующем примере. Допустим, что

$$K' = \begin{cases} 0 \\ I \\ OI \end{cases} .$$

Тогда послание OIOI не является однозначно дешифруемым, поскольку это может быть OI или O/I/OI, имеются также другие возможности.

В большинстве практических случаев справедливо, что P имеет значительное отклонение от равномерного распределения, следовательно, наше предположение делает возможным изучение большинства практических случаев. С помощью ЭВМ ICT 1905 в Центральном институте физических исследований /Будапешт/ мы изучали вопросы компрессии экспериментальных данных как для физики низких энергий, так и для физики высоких энергий. Мы изучали также компрессию данных, не связанных с физическими исследованиями. Очевидно, что компрессия данных означает экономию времени, если они передаются, и экономию пространства памяти, если данные где-то хранятся.

2. Результаты экспериментов.

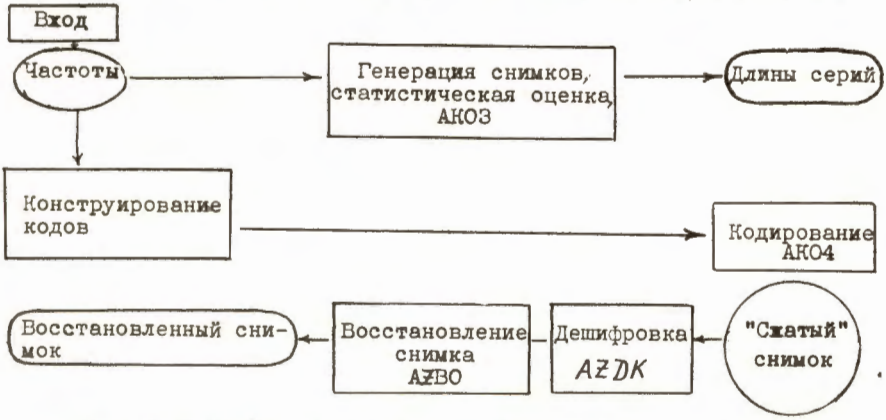
Мы приводим данные об экспериментах, сделанных в Центральном институте физических исследований /читатель, который интересуется подробностями, может просмотреть статьи, находящиеся в печати /4, II/.

Для компрессии снимков существует известный метод, называемый "Кодирование по длинам серий" /см. /I/ /. Его принцип можно изложить следующим образом. Рассмотрим снимок, представимый после сканирования в виде матрицы размером $n \times m$ с элементами a_{ik}. Каждая линия снимка разделена на n, а каждая колонка разделена на m полей одинакового размера. a_{ik} равно нулю или I, согласно соответствующему полю снимка, белому или черному (естественно, эта калибровка должна иметь такую плотность, чтобы большая часть полей получилась белой или черной. Очень редко может случиться, что одно поле одновременно частично белое и частично черное, т.е. цвет большинства полей считается достоверным). Матрица (a_{ik}), полученная таким образом, может быть распределена на следующие серии двоичных разрядов $a_{11}, a_{12}, \ldots a_{1n}, a_{21}, \ldots, a_{mn}$.

Эта серия двоичных разрядов называется цифровым преобразованием снимка /фототелеграф/. В таком снимке последовательные символы называются серией длиной m, если I-й и $(m+2)$-ой символы равны, а все другие отличаются от первого. Первый этап кодирования методом длин серий состоит в замене серий их длинами, так как распределение длин серий в каждом практическом случае имеет значимое отклонение от равномерного распределения. Возможно использовать кодирование с переменной длиной слова для кодирования длин серий.

Таким образом, кодирование может стать более эффективным. Очень

эффективен префикс-код **Хафмена**, его описание можно найти в работе /9/ . Для осуществления кодирования методом длин серий, главным образом, для снимков с пузырьковых камер, в ЦИФИ разработана программа, с помощью которой весь процесс компрессии может быть выполнен автоматически. Блок-схема программы выглядит следующим образом:



Выдача каждой программы помечена кружком, но, естественно, можно использовать программу для экономии времени без всякой внутренней выдачи. Эта программа определяет среднюю величину длины слова кода, число полей в снимке после компрессии и число черных полей в первоначальном снимке.

Существует такое расширение программы, которое способно производить компрессию информации с адаптируемой избыточностью, и с помощью этого компенсировать уменьшение способности к защите от ошибок для сжатой информации.

Существуют также генераторы случайных чисел для моделирования шумового эффекта, имеющегося в канале. Влияние этих генераторов на информацию таково, что она становится ошибочной, но избыточность

дает возможность обнаружить ошибку с помощью корректирующего кода /при определенном уровне шумов/. Расширенная программа дает пользующемуся машиной возможность знать, как много ошибок не может быть исправлено.

Мы преобразовали в цифровую форму несколько типов снимков с помощью фототелеграфа "Гамма", который был оснащен специальным электронным устройством для того, чтобы приспособить фототелеграф для связи с перфорирующим оборудованием "Фацит". Мы использовали 8- канальную бумажную ленту как входное устройство на ЭВМ. В ближайшем будущем оборудование будет улучшено за счет замены бумажной ленты и обеспечения связи он-лайн фототелеграфа с ЭВМ. С другой стороны, будут сделаны улучшения за счет применения приборов с более высокой разрешающей способностью.

Как можно видеть из блок-схемы, программа дает возможность производить модели простых снимков /содержащих только линии/ автоматически с помощью подпрограммы АКОЗ. В некоторой части экспериментов объектом компрессии были снимки, полученные автоматически. Подпрограмма АКОЗ делает модель снимка размером 400 х 400, что означает снимок, содержащий 160 000 полей. Естественно, эффективность компрессии зависит от соотношения белых и черных полей, то есть от соотношения нулей и единиц в преобразованном цифровым кодом снимке. В исключительном случае /который, правда, не типичен/, очень простой снимок был сокращен до 3200 полей, что означает фактор компрессии, близкий к 50, так как первоначальный снимок содержал 160 000 полей. В большинстве случаев фактор компрессии был порядка 5-6. Все результаты экспериментов по компрессии данных физики высоких энергий опубликованы в статье /11/ . Мы также применяли компрессию данных к экспериментальным результатам, полученным с по-

мощью 512- канального анализатора, и текстам на венгерском языкѐ.

Стандартная программа для компрессии любой программы для ЭВМ ICT 1905 / с целью хранения или передачи/ находится в стадии подготовки. В этой области будет также осуществляться сотрудничество с ОИЯИ.

Готовятся другие программы, которые, кроме кодирования методом длин серий, используют другие различные методы кодирования. Главным достоинством этих методов является то, что они не зависят от статистики. Это свойство не выполняется для кодирования методом длин серий. Можно коротко пояснить алгоритм новой программы следующим образом.

Пусть a, a_2 .... a_r - произвольно преобразованная в цифровой код картинка или любая другая информация, состоящая из 0,I. Если справедливо условие существенного отличия от равномерного распределения $p_0 \gg p_1$, то эффективная компрессия данных может быть выполнена с помощью следующего преобразования:

$$T = \begin{cases} 00 \rightarrow 0 \\ 01 \rightarrow 11 \\ 1 \rightarrow 10 \end{cases}.$$

Это означает, что если $a_1 = 1$, то $T(a_1) = 10$, если $a_1 = 0$ и $a_2 = 0$, то $T(a_1 a_2) = 0$. Наконец, если $a_1 = 0$ и $a_2 = 1$, то $T(a_1 a_2) = 11$.

Можно пользоваться методом итеративного применения преобразования T и достигать все более эффективной компрессии. Однако по достижению определенного минимума итерация даст более длинное сообщение, и длина его будет стремиться к бесконечности, если число итераций стремится к бесконечности.

Для иллюстрации этого положения мы приводим следующий пример: если u обозначает информацию, которая должна быть сжата, то предположим, что

u = 00/00/00/I/00/00/I/00/00/0I/00/00/

Тогда $T(u)$ = 000I000I000II00

$T^2(u)$= 0II0II0III00

$T^3(u)$= IIIOIIIOIIIOI00 .

Если длина u обозначена через l_0 , а длина $T^m(u)$ – через l_m , тогда в нашем примере l_0 = 22, l_1 = I5, l_2 = I2, l_3 = I5.

Интересно заметить,что Леммелу удалось найти некоторую решетчато-упорядоченную алгебраическую структуру, операция которой обладает именно этим свойством /I0/. Е.Л.Блох изучал этот метод с точки зрения теории вероятности /2/.

3. Теоретические исследования.

Сжатая информация содержит, естественно, меньше избыточности и, следовательно, меньше защищена от ошибок, чем первоначальные данные. Некоторые авторы изучали возможность обнаружения ошибок и исправления сжатой информации /см., например, /3/ /8/ /I4/.

В работе /4/ можно найти более эффективный метод защиты от ошибок для сжатой информации, чем тот, о котором упоминалось выше. Принцип метода, описанного в /4/, состоит в том, что последовательные знаки рассматриваются как символы информации кодовых слов кода (n,K) . Сжатая информация кодируется кодом (n,K) /Из-за отсутствия места мы не можем привести все используемые понятия обнаружения ошибки и корректирующих кодов. Если читатель не знаком с этими

понятиями, он может посмотреть работы /7,13/ .

Чтобы продемонстрировать этот метод защиты от ошибки, мы можем привести следующий пример. Пусть H = $\{ I_1, I_2, I_3, I_4 \}$ будет набором информации, а $P_I = 0{,}5$ $P_2 = 0{,}2$ $P_3 = 0{,}15$ $P_4 = 0{,}15$ – соответствующие вероятности.

Префикс-код, соответствующий этому распределению вероятности, выглядит следующим образом /4/ :

$I_I \rightarrow 0$

$I_2 \rightarrow 11$

$I_3 \rightarrow 100$

$I_4 \rightarrow 101$.

Для защиты от ошибки был применен код Хемминга /4,7/. Кодовые слова его следующие:

0000000

0001111

0010110

0011001

0011001

0100101

0101010

0110011

0111100

1000011

1001100

1010101

1011010

1100110

1101001

1110000

1111111.

Используя префикс-код и код Хемминга, описанный выше, мы запретим кодирование следующих серий информации

$I_3\ I_1\ I_1\ I_3\ I_4\ I_1\ I_2\ I_3\ I_1\ I_2$.

После применения префикс-кода можно получить следующую серию из 0 и 1:

10000100101011100011 .

После применения кода Хемминга /4,7/ можно получить:

1000011/0100101/1010101/1110000/0011001.

В работе /4/ имеется таблица для сравнения среднего значения длин кода и вероятностей ошибок, полученных несколькими методами защиты от ошибок. Мы приводим здесь таблицу для того, чтобы пояснить обозначения:

Q обозначает вероятность того, что обнаруженная ошибка встречается,

ℓ обозначает среднюю длину кодового слова.

Как специальный фактор эффективности авторы ввели обозначение E. E определяется следующим образом:

$$E = -\frac{\ell_q Q}{\ell}$$ /см. [4] /.

Следующая таблица содержит значения параметров ℓ, Q, E для нескольких методов, которые здесь не названы. Поскольку было бы сложным объяснять каждый из них, мы упомянем только, что в последнем ряду таблицы имеются параметры, относящиеся к методам, описанным в работе /4/ :

ℓ	Q	E
4	$4\ 10^{-3}$	0,599
5,02	$9,97\ 10^{-6}$	0,995
7	$2,1\ 10^{-5}$	0,668
3,06	$1,7\ 10^{-2}$	0,579
7,06	$1,11\ 10^{-2}$	0,277
12	$1,2\ 10^{-2}$	0,160
9,18	$3,82\ 10^{-5}$	0,482
5,86	$2,14\ 10^{-8}$	1,309

Хорошо известно, что, используя кодирование с переменной длиной слова, можно сжать информацию. Средняя длина слова после кодирования зависит от распределения вероятности исходного события.

Автор имеет несколько неопубликованных результатов относительно усовершенствования кодирования с переменной длиной слова, выбирая набор информации не совсем обычным путем. Чтобы дать некоторое представление относительно этого принципа, мы покажем, как этот принцип может быть применен для улучшения кодирования методом длин серий.

Пусть m – выбранное соответствующим образом положительное целое число. Мы берем снимок, преобразованный в цифровой код, причем максимум длины серий m пусть будет неизменным. Все серии с длиной, больше, чем m, будут разделены на серии длины m таким образом, что длина последней серии должна быть, по крайней мере, меньше, чем m. Таким образом, можно получить новое распределение вероятностей относительно нового набора информации,

а именно, элементы набора информации есть только серии длины, в большинстве равной m . Если значимое отклонение от равномерного распределения будет больше для распределения преобразования, чем первоначальное, может быть осуществлена более эффективная компрессия данных. Наш опыт и теоретические работы показывают, что для соответствующего m это преобразование распределения вероятности приводит к такому распределению, которое делает возможным более эффективную компрессию данных, чем первоначальная.

Имеется другая сторона неопубликованных теоретических исследований, которая связана с алгебраическими результатами автора в работах, опубликованных по вопросу преобразования полугрупп (см. /5, 6/). Эти результаты применяются для получения новых результатов в алгебраической теории информации.

Приведем несколько главных идей теории /описание этой теории еще не выполнено , и настоящее изложение является первым сообщением основных мыслей/.

Язык человеческого общения таков, что два вида информации, например, слова, находятся в сопоставлении таким образом, что длина двух закодированных элементов информации больше, чем длина одного из них. Но в случае матриц произведение двух матриц размером $n \times n$ имеет также размер $n \times n$.

В чем проблема? Нужно найти однозначно сомножители, если произведение дано. Но подобная же проблема встает, когда мы используем кодирование переменной длины и вынуждены ограничивать себя такими кодами, которые однозначно дешифруемы. Однозначная дешифруемость есть специальная декомпозиция.

Можно заметить, что все значимые тексты любого человеческого языка образуют

свободную полугруппу, образующими элементами которой являются все слова данного языка, включая промежутки и знаки препинания. В структуре языка, т.е. в специальном случае свободных полугрупп существует однозначная декомпозиция факторов, так называемая однозначная первичная факторизация, поскольку произвольные тексты могут быть разложены исключительно на слова, промежутки и знаки препинания.

Естественно, встает следующий вопрос, который, насколько автору известно, не изучен, а именно : что является необходимым и достаточным условием в алгебраической структуре для существования однозначной первичной факторизации, а если она существует, то какова плотность элементов?

Поскольку в произвольной алгебраической структуре с однозначной первичной факторизацией можно осуществить компрессию информации, кодированной с помощью элементов алгебраической структуры , то эффективность информации зависит от плотности элементов.

Наши исследования связаны в настоящее время с проблемами, упомянутыми выше, и мы исследуем несколько классов полугрупп с целью найти такие полугруппы /по крайней мере, одну/, в которой может быть осуществлена первичная факторизация, а плотность элементов выше, чем в свободных полугруппах, используемых как коды переменной длины.

Если допустить коммутативность, например, для данных в области низких энергий, то можно получить полезный,с точки зрения компрессии данных,класс алгебраических структур посредством использования мультипликативных полугрупп натуральных чисел. Основной идеей настоящего исследования является использование этих структур для компрессии данных.

4. Некоторые замечания о технической реализации

Использовалось оборудование для преобразования снимков в цифровой код. Испытывается устройство для кодирования длины измерения и для описанного улучшенного метода

Разработка устройства для защиты от ошибок в сжатой информации будет завершена в ближайшем будущем.

Наши будущие проекты - разработать устройства для компрессии данных - могут быть сформулированы следующим образом:

I. Осуществление связи он-лайн между преобразователем в цифровой код и ЭВМ.

2. Разработка устройства, которое сжимает информацию методом, не зависящим от статистики.

3. Разработка быстрого умножителя и факторизатора.

4. Разработка эффективного компрессора информации, использующего некоммутативную алгебраическую конструкцию с однозначной первичной факторизацией.

Автор благодарен своим ассистентам Золтану Мочи и Эмёке Коренчи за написание программ. Автор выражает при этом свою благодарность Лаборатории вычислительной техники и автоматизации ОИЯИ, особенно заместителю директора Н.Н.Говоруну за предоставленную возможность выступить на семинаре по данной теме и написания препринта.

Автор благодарит также Г.А.Ососкова и Г.Н.Савину за перевод, редакцию и оформление препринта.

ЛИТЕРАТУРА

I. H.Wyle, T.Erb., Banow. Reduced Time Facsimile Transmission by Digital Coding (IRE Trans. CS-9, No 3, 215-222, (1961).

2. Э.Л.Блох. Последовательное кодирование. Проблемы пер. информации. Вып. 5 (1960) 55-69

3. L.Calabi. Combinatorial properties of variable length error correcting codes. Scientific Report No2, Contract AF 19(628) 3826, Dec.1964 Parke Math.Lav.AFCRL 65-28

4. J.Denes, K.Schermain: Variable length encodings to appear.

5. J.Denes. Transformations and transformation semigroups. Theory of graphs. Proc.Conf.Graph Theory held at Tihany, Sept. 1966.

6. J.Denes. Combinotorial properties of transformation semigroups (to appear in J.Combinational Theory)

7. J.Denes, MSzokolay. Praktischen und theoretischen Problemen des Datenubertragung. VEB Verlag Technik, Berlin (to appear)

8. E.N.Gilbert, E.F.Moore. Variable length binary encodings. Bell System Techn., J., 1959 , vol.38, No 4, 933-967.

9. D.A.Huffmann. A method for the construction of minimum redundancy codes. Proc.IRE, 1952, 1098-1101

10. A.E.Laemmel. Application of lattice-ordered semigroups to codes and finite state transducers. Proc.Sympos.Meth. Theory of Automation, New York, 1962. Polytechnic Press of Polytechnic Inst. of Brooklyn, Brooklyn, 1963.

11. Z.Mocsi. Valtozo szohosszusagu kodolas alkalmazasa az alacsony es nagy energias fizikaban. (Hungarian) to appear.

12. Д.А.Новик. Эффективное кодирование. Изд.Энергия, 1965.

13. У.Питерсон. Коды, исправляющие ошибки. Пер. с англ. Изд. Мир, 1964.

14. J.A.Riley. The theory of index correcting codes. Appendix of /3/.

Рукопись поступила в издательский отдел

16 июля 1968 года