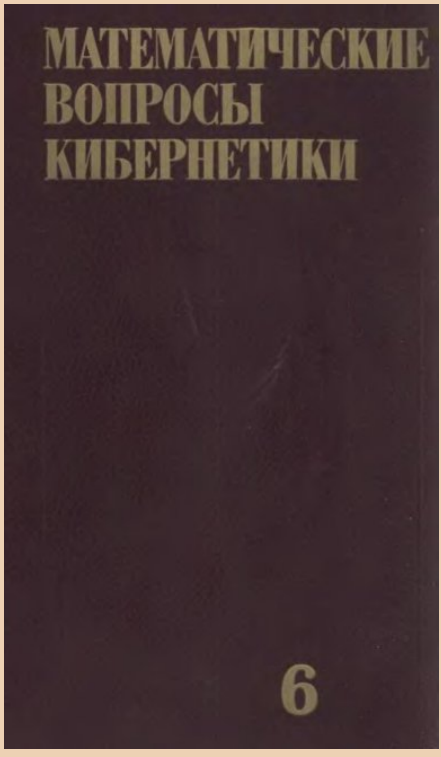
ИПМ им. М.В. Келдыша РАН • Электронная библиотека
Математические вопросы кибернетики • Выпуск 6



А. А. Марков

О некоторых мерах сложности и эффективности в алфавитном кодировании (доклад на VI Всесоюзной конференции по проблемам теоретической кибернетики - Саратов, 1983 г.)

Рекомендуемая форма библиографической ссылки: Марков А. А. О некоторых мерах сложности и эффективности в алфавитном кодировании (доклад на VI Всесоюзной конференции по проблемам теоретической кибернетики - Саратов, 1983 г.) // Математические вопросы кибернетики. Вып. 6. — М.: Наука, 1996. — С. 348–352. URL: http://library.keldysh.ru/mvk.asp?id=1996-348

и другим областям математики. Широко известна его монография «Введение в теорию кодирования», ставшая основным учебником по этой тематике в разных вузах страны. Александр Александрович подготовил восемь кандидатов наук. Последние годы он с большим увлечением работал над постановкой преподавания дискретной математики для математиков-теоретиков. Разработал и прочитал курс дискретной математики для студентов мехмата, руководил научно-методическим семинаром, разработал план специализации по дискретной математике на мехмате.

Александра Александровича отличала глубина и нестандартность мышления, он был интересным и остроумным собеседником, имел свою, совершенно не банальную точку зрения по многим вопросам не только науки, но и повседневной жизни. Со студенческих лет он соединял в себе стремление к истине и принципиальность с внимательным отношением к товарищам по работе.

Последние годы Александр Александрович был тяжело болен, но очень достойно и мужественно боролся с болезнью, продолжал работать с полной отдачей.

В архиве А. А. Маркова сохранился его доклад на VI Всесоюзной конференции по проблемам теоретической кибернетики, подводящий итоги важного этапа исследований, который завершился изданием монографии «Введение в теорию кодирования» и защитой докторской диссертации. Ниже приводится текст этого ранее не публиковавшегося доклада, хорошо передающий живую речь Александра Александровича.

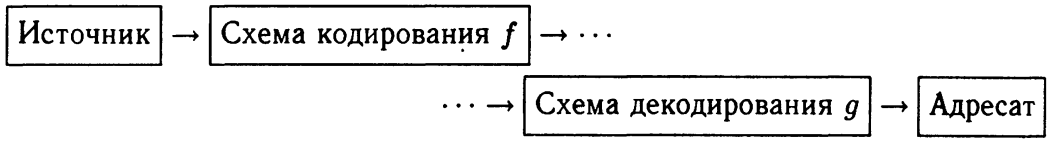
О НЕКОТОРЫХ МЕРАХ СЛОЖНОСТИ И ЭФФЕКТИВНОСТИ В АЛФАВИТНОМ КОДИРОВАНИИ

(Доклад на VI Всесоюзной конференции по проблемам теоретической кибернетики — Саратов, 1983 г.)

А. А. МАРКОВ

(Н. НОВГОРОД)

В теории алфавитного кодирования изучается следующая математическая модель канала связи:



Имеется q-ичный алфавит канала $A = \langle a_1, \ldots, a_q \rangle$ $(q \geqslant 2)$, источник производит сообщения в языке $L \subseteq B^*$, где $B = \langle b_1, \ldots, b_m \rangle$ $(m > q)$ — алфавит языка L, а кодирование сообщений осуществляется побуквенно в соответствии со следующей схемой:

$$f = f_V\colon \begin{cases} b_1 & \to & v_1, \\ b_2 & \to & v_2, \\ \ldots\ldots\ldots & & \\ b_m & \to & v_m \end{cases}$$

(схема f_V определяется кодом $V = \langle v_1, v_2, \ldots, v_m \rangle$). Можно сказать, что алфавитное кодирование есть гомоморфизм свободной полугруппы B^* в A^*, порожденный схемой f_V.

Цель доклада — рассказать о современном состоянии теории алфавитного кодирования и ее связи с вероятностной теорией информации. Подробнее остановлюсь на некоторых принципиальных вопросах. Ограничусь теми только вопросами, которые связаны со сжатием информации, но замечу, что малоисследованная тематика помехоустойчивости алфавитного кодирования тоже заслуживает внимания.

Сжатие информации основывается на том, что в языках обычно имеется значительная избыточность. Характер избыточности может быть двоякий. Во-первых, некоторые буквы и их сочетания могут встречаться часто, другие — редко, это обстоятельство отражается в вероятностных моделях языков, которые вслед за К. Шенноном обычно и исследуются в теории информации. Во-вторых, избыточность заложена обычно и в семантике языка, включая его структурные свойства, грамматику, синонимию. Например, в сообщениях на русском языке не могут встречаться фрагменты «ЯА», «ГГ», «ППП» и т. п. Использование второго вида избыточности в последнее время изучалось в теории алфавитного кодирования, о работах последнего времени в этом направлении я и расскажу. Вопросов использования синонимии касаться не буду, эта тема разработана мало.

Основное требование, которое предъявляется к кодированию, — взаимная однозначность f_V на L. Буду рассматривать кодирование только регулярных языков, т. е. представимых в конечных автоматах. Замечу: область развития общих методов алфавитного кодирования на другие модели языков не может быть расширена очень сильно, так как уже для класса детерминированных контекстно-свободных языков проблема взаимной однозначности алфавитного кодирования алгоритмически неразрешима (Л. П. Жильцова [1]). Дж. Хартманис и Дж. Е. Хопкрофт написали работу «О причинах неразрешимости некоторых проблем теории языка» [2]; здесь имеются аналогичные причины.

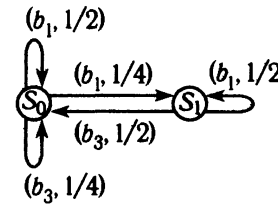


Рис. 1

Основную постановку задачи проиллюстрирую на примере кодирования вероятностного эргодического источника (рис. 1), где

$$m = 3, \quad q = 2, \quad L = B^* \setminus B^*(b_1 b_2)B^*, \quad H_{\text{ИСТ}} = \frac{4}{3}, \quad \pi_1^* = \pi_2^* = \pi_3^* = \frac{1}{3}.$$

Стоимость кодирования F сообщений, порождаемых источником, определяется как

$$C = \lim_{N\to\infty} \frac{1}{N} \sum_{\alpha \in B^N \cap L} p(\alpha) \cdot |F(\alpha)|,$$

а разность $R = C - H_{\text{ИСТ}}$ называется избыточностью кодирования. В случае алфавитного кодирования эргодического источника имеем

$$C = \sum_{i=1}^{m} \pi_i^* \cdot |v_i|.$$

Согласно теореме Шеннона, при любом конечно-автоматном кодировании избыточность неотрицательна, а при N-блочном кодировании может быть сделана сколь угодно близкой к нулю за счет выбора достаточно большого N. При этом, если источник отличен от бернуллиевского, как в рассматриваемом примере, то согласно Р. Е. Кричевскому [3] существует константа c_0 такая, что при достаточно больших N избыточность N-блочного кодирования $R_N \geqslant c_0/N$.

Рассмотрим для этого источника алфавитное кодирование по схеме

$$f_V\colon \begin{cases} b_1 & \to & 1, \\ b_2 & \to & 0, \\ b_3 & \to & 10. \end{cases}$$

Легко видеть, что f_V взаимно однозначно на L, причем декодирование может быть реализовано конечным автоматом (рис. 2).

Стоимость этого кодирования

$$C = \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 2 = \frac{4}{3} = H_{\text{ист}},$$

т. е. оно имеет нулевую избыточность. Если же не использовать грамматику языка L и кодировать алфавит B префиксным кодом по алгоритму Хаффмана, то получим

$$b_1 \to 1, \qquad b_2 \to 01, \qquad b_3 \to 00,$$

$$C^X = \frac{5}{3}, \quad R = \frac{1}{3}, \quad \zeta(L, B^*, \pi^*) = \frac{C(L)}{C^X} = \frac{4/3}{5/3} = 0{,}8.$$

С решением задачи оптимального алфавитного кодирования связано несколько вопросов; рассмотрим их.

Пусть $f_L(x_1, \ldots, x_m)$ — характеристическая функция спектров длин элементарных кодов; переменные принимают в качестве значений неотрицательные целые числа, а значение $f_L(x_1, \ldots, x_m)$ равно 1, если существует дешифруемый для L код $V = \langle v_1, \ldots, v_m\rangle$ с $|v_1| = x_1, \ldots, |v_m| = x_m$, и равно нулю в противном случае. Поскольку из $\widetilde{x} \leqslant \widetilde{x}'$ следует $C(\widetilde{x}) \leqslant C(\widetilde{x}')$, минимум стоимости алфавитного кодирования всегда достигается на кодах, спектры которых являются нижними единицами функции f_L. Множество $M(L)$ нижних единиц f_L конечно в силу теоремы Диксона и называется матрицей оптимального кодирования языка L.

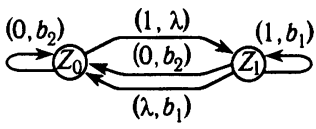


Рис. 2

Вопрос 1. Оценить сложность расшифровки $M(L)$.

Вопрос 2. Оценить сложность минимизации линейной формы на конечном множестве $M(L)$.

Вопрос 3. Какова эффективность учета элементов модели языка сообщений, т. е. как влияют на стоимость оптимального кодирования различные факторы, учтенные в модели L, и какие из них содержат полезную информацию?

Результативность исследований по вопросу 3 зависит от того, насколько изучены вопросы 1 и 2. Среди факторов, несущих полезную информацию и необходимых среди любой полезной информации для широкого класса языков, отмечу наличие в языке запрещенных фрагментов.

Успехи в решении вопроса 2 зависят от успехов в решении вопроса 1 и теоретических исследованиях в области дискретной оптимизации. Отмечу подходы, развитые М. Ю. Мошковым [4] и Е. Г. Воробьевой [5]. Подчеркну, что учет структурной модели языка вовсе не обязательно усложняет оптимизационную задачу по сравнению с алгоритмом Хаффмана, но может ее даже ощутимо упростить.

Исследования по вопросу 1 составляют основное математическое содержание собственно теории алфавитного кодирования; на них остановлюсь подробнее.

Здесь представляют интерес в первую очередь условия взаимной однозначности алфавитного кодирования регулярных языков. Имеются комбинаторно-логические условия взаимной однозначности конечно-автоматного

типа. В качестве меры сложности распознавания взаимной однозначности обычно рассматривают длину кратчайшего слова в алфавите канала, которое может быть неоднозначно декодировано (если такое существует). Для этой характеристики получаются квадратичные верхние оценки как в зависимости от суммы длин элементарных кодов, так и от числа состояний регулярного источника, порождающего язык сообщений, причем по порядку они не могут быть улучшены (Л. Г. Киселева [6], В. И. Шевченко [7]).

Другой класс составляют так называемые спектральные условия (т. е. условия вида $Q(|v_1|, \ldots, |v_m|)$, где Q — некоторый предикат) — необходимые условия взаимной однозначности, вытекающие из мощностных соображений. Тот факт, что при взаимно однозначном кодировании число сообщений языка, которым сопоставлены кодовые комбинации длины N, не может быть больше q^N, записывается в виде бесконечной системы спектральных неравенств. Вся система для теоретических исследований малоэффективна, но выводимое из нее в качестве следствия асимптотическое спектральное неравенство для многих регулярных языков является необходимым и достаточным условием реализуемости спектра дешифруемым для L кодом. При $L = B^*$ это — неравенство Мак-Миллана.

Для многих регулярных языков, однако, аналитического описания $M(L)$ асимптотическим спектральным неравенством не получается, и для расшифровки ее требуется перебор. В качестве меры сложности перебора рассматривают максимум значения компонент векторов $M(L)$, называемый значностью матрицы оптимального кодирования: $k(L)$.

Для получения верхних оценок $k(L)$ предложен общий метод — принцип равномерного пополнения. Этот принцип состоит в том, что в любом дешифруемом для L коде любой набор элементарных кодов можно заменить равномерным кодом с сохранением дешифруемости для L. Класс языков, для которых справедлив принцип равномерного пополнения, шире, чем класс языков, для которых f_L монотонна. Однако этот принцип не может быть обоснован для всех регулярных языков; примеры его нарушения имеются уже среди языков, порождаемых регулярными источниками с двумя состояниями.

Принцип равномерного пополнения удалось обосновать для класса вполне регулярных языков — порождаемых регулярными источниками, у которых алфавит языка, порожденного любой компонентой циклической связности, если не пуст, то совпадает с алфавитом всего языка. Данный класс содержит все конечные языки и все языки, которые порождаются эргодическими источниками. Непосредственным использованием оценок размерности равномерного пополнения для вполне регулярных языков получаются и верхние оценки $k(L)$.

За пределами вполне регулярных языков остаются, например, языки $B^* \cdot \{c\}^*$, где буква c не принадлежит B. Для этих языков А. А. Кочетов получил полное описание матриц оптимального двоичного кодирования [8]. Это потребовало развития специальных методов и использования специальных арифметических функций.

СПИСОК ЛИТЕРАТУРЫ

1. Жильцова Л. П. Об алфавитном кодировании контекстно-свободных языков // Комбинаторно-алгебраические методы в прикладной математике / Под ред. Ал. А. Маркова. — Горький, 1983. — С. 106–122.
2. Hartmanis J., Hopcroft J. E. What makes some language theory problems undecidable // J. of Computer and System Science. — 1970. — V. 4, № 4. — P. 368–376. [Рус. пер.: Хартманис Дж., Хопкрофт Дж. Е. О причинах неразрешимости некоторых проблем теории языка // Кибернетич. сб. Новая серия. Вып. 8. — М.: Мир, 1971. — С. 232–243.]
3. Кричевский Р. Е. Лекции по теории информации. — Новосибирск, 1970.

4. Мошков М. Ю. О задаче минимизации линейной формы на конечном множестве // Комбинаторно-алгебраические методы в прикладной математике / Под ред. Ал. А. Маркова. — Горький, 1985. — С. 98–119.
5. Воробьева Е. Г. Монотонные преобразования в задачах дискретной оптимизации и минимизация стоимости кодирования в классе самосинхронизирующихся кодов. — Горький, 1985. — Деп. ВИНИТИ, № 3398–85.
6. Киселева Л. Г. Алгебраическое исследование простейших кодов и бескоэффициентных уравнений в словах // Матем. сборник. — 1979. — Т. 108 (150), № 4. — С. 529–550.
7. Шевченко В. И. О сложности распознавания алфавитной редукции регулярных языков // Комбинаторно-алгебраические методы в прикладной математике / Под ред. Ал. А. Маркова. — Горький, 1983. — С. 208–217.
8. Кочетов А. А. Алфавитное двоичное кодирование произведения свободной полугруппы и свободной моногенной полугруппы // Материалы Всесоюзного семинара по дискретной математике и ее приложениям. — М.: Изд-во МГУ, 1986. — С. 197.