УДК: 519.722

ЭФФЕКТИВНОЕ КОДИРОВАНИЕ ДЛИН СЕРИЙ ПРИ ФАКСИМИЛЬНОЙ ПЕРЕДАЧЕ ДАННЫХ ПО СЕТИ

М.П. Бакулина

В данной работе рассматривается задача эффективного кодирования длин серий для факсимильной передачи данных по сети. В качестве модели рассматривается марковский источник порядка n, порождающий символы с неизвестными условными вероятностями. В отличие от ранее известных мето-дов, предлагаемая конструкция кода позволяет достигать любой наперед заданной избыточности с сохранением небольшого объема памяти и небольшого среднего времени кодирования и декодирования.

Ключевые слова: сжатие данных, кодирование длин серий, объем памяти, скорость кодирования.

Введение

Задача эффективной передачи данных при цифровой факсимильной связи хорошо известна. Применительно к цифровым факсимильным сигналам для уменьшения объема передаваемых данных используются различные методы кодирования источника, а уменьшение объема данных достигается за счет применения кода, учитывающего статистические зависимости между элементами.

Полезная модель цифрового изображения предложена Дж. Капоном [1]. В ней каждая сканируемая строка рассматривается как марковская цепь первого порядка, в которой цвет каждого элемента изображения зависит лишь от цвета предыдущего элемента. На основе этой модели были созданы методы кодирования длин серий, в которых цифровое изображение рассматривается как последовательность чередующихся независимых серий из черных и белых элементов изображения. По сравнению с моделью Дж. Капона, модель кодирования длин серий учитывает зависимости более высокого порядка между соседними элементами одного цвета.

Для кодирования длин серий разработано много различных кодов (см., например, [2], [3]). В работе Р. Хантера и А. Х. Робинсона [4] были предложены международные стандарты кодирования для цифровой факсимильной связи. Ими используется одномерная схема кодирования, в которой длины серий кодируются с помощью модифицированного кода Хаффмана, что позволяет передавать типичные документы в форме черно-белых изображений, сканируемых с нормальным разрешением (3, 85 строк/мм, 1728 элементов на строку), со скоростью 4800 бит/с в среднем за время около 1 минуты.

В данной работе предлагается адаптивная схема кодирования данных для цифровой факсимильной связи, в которой используется двухэтапное кодирование длин серий. В отличие от ранее известных методов, предлагаемая схема позволяет достигать любой наперед заданной избыточности. Кроме избыточности, эффективность кода оценивается также объемом памяти V кодера и декодера (в битах), который требуется для реализации метода на компьютере со свободным доступом к памяти (это модель “обычного” компьютера), и средним временем T кодирования и декодирования одного символа источника информации, измеряемым числом бинарных операций над однобитовыми словами. В работе приводятся оценки эффективности кодирования, показывающие, что данный метод обладает небольшим объемом памяти и небольшим средним временем кодирования и декодирования.

Адаптивная схема кодирования длин серий

Пусть источник порождает сообщение $x_1 x_2 \ldots$, при этом вероятности появления символов неизвестны. Опишем адаптивную схему кодирования, осуществляемую в два этапа.

Метод основан на том, что каждую строку сканирования можно рассматривать как последовательность чередующихся серий 1 и 0, соответствующих сериям черных и белых элементов, причем вероятность появления 1 мала. Следовательно, на первом этапе можно добиться существенного сжатия поступающих входных данных для передачи их по сети. Достижение наперед заданной избыточности происходит на втором этапе за счет применения к полученной после первого этапа сокращенной последовательности какого-либо известного адаптивного кода. Мы будем использовать адаптивный арифметический код из [5], однако могут быть использованы и другие универсальные коды, применение которых дает тот же результат.

Рассмотрим первый этап кодирования. Введем в рассмотрение скользящее окно длины ω, представляющее собой слово $x_{t-\omega}x_{t-\omega+1}\ldots x_{t-1}$. Зафиксируем избыточность $r>0$ и определим $\omega=\lceil 2/r\rceil$. Пусть $N(1,i)$ – число вхождений 1 в окно $x_{t-\omega}x_{t-\omega+1}\ldots x_{t-1}$. Обозначим через $\widetilde{p}(1,i)$ оценку вероятности появления 1 в этом окне, а через l_i - длину блока, определяемую на основе статистики данного окна. Определим

$$\widetilde{p}(1,i)=\frac{N(1,i)+1}{\omega+2},$$
$$l_i=\left\lceil\frac{1}{\sqrt{\widetilde{p}(1,i)}}\right\rceil=\left\lceil\sqrt{\frac{\omega+2}{N(1,i)+1}}\right\rceil, \quad (1)$$
$$\widetilde{q}(i)=\widetilde{p}(0,i)=1-\widetilde{p}(1,i)$$

Кодирование каждого очередного блока осуществляется на основе статистики, содержащейся в предыдущем окне. Если блок состоит из одних нулей, то его кодом является 0, иначе длина кодового слова равна l_i+1: началом его является 1, за которой следует тот же блок длины l_i. Затем сдвигаем окно на l_i символов вправо и на основе статистики полученного окна аналогичным образом кодируем следующий блок.

Рассмотрим второй этап кодирования. Представим последовательность $z_1z_2\ldots z_s, z_i\in\{0,1\}$, полученную после первого этапа кодирования, в виде

$$\underline{0}\ldots\underline{01}\underbrace{z_1\ldots z_{l_j}}_{l_j}\underline{0}\ldots\underline{01}\underbrace{z_1\ldots z_{l_j}}_{l_j}\ldots$$

В ней выделены блоки длины l_j, следующие за каждой выделенной 1, и “особые” символы $\underline{0},\underline{1}$, не входящие в блоки. Рассмотрим какой-нибудь блок $\underbrace{z_1\ldots z_{l_j}}_{l_j}$. Для определенности будем считать, что символам $z_1\ldots z_{l_j}$ этого блока соответствуют символы $x_i\ldots x_{i+l_i}$ исходной последовательности, оценки вероятностей которых определяются на основе статистики окна $x_{t-\omega}x_{t-\omega+1}\ldots x_{t-1}$, используемого на первом этапе. Определим

$$t=\lceil 4\log(30/r)\rceil,\quad \widehat{\omega}=\left\lfloor(40/r)^{8/3}\right\rfloor, \quad (2)$$
$$\widehat{q}(i)=(\lfloor\widetilde{q}(i)\cdot(2^{t/4}-2)+1\rfloor)\cdot 2^{-t/4},$$
$$\widehat{p}(1,i)=1-\widehat{q}(i),$$

где $\widetilde{q}(i)$ определяется формулой (1).

Для оценки вероятности символа z_n, находящегося в n-ой позиции после $(n-1)$ нулей в блоке $z_1\ldots z_{l_j}$, введем величины $\widehat{\pi}_n(1)$ и $\widehat{\pi}_n(0)$ $(n=1,\ldots,l_j)$, которые будем вычислять по следующим формулам:

$$\widehat{\pi}_n(1)=\frac{\widehat{p}(1,i)}{1-\widehat{q}(i)^{l_i-n+1}},\quad \widehat{\pi}_n(0)=1-\widehat{\pi}_n(1). \quad (3)$$

Обозначим через $|\delta_n|$ погрешность, обусловленную вычислением вероятностей $\widehat{\pi}_n(1)$ по формуле (3), а через $\widehat{\pi}'_n(1)$ величину, полученную при вычислении вероятностей $\widehat{\pi}_n(1)$ с погрешностью $|\delta_n|$:

$$\widehat{\pi}'_n(1)=\widehat{\pi}_n(1)\cdot(1+|\delta_n|).$$

Можно показать, что $|\delta_n|\le 2^{-3t/4}$.

Представим теперь каждую вероятность $\widehat{\pi}'_n(a)$ $(a\in\{0,1\})$ как двоичную дробь с t знаками. Кодирование символа z_n, находящегося в n-ой позиции после $(n-1)$ нулей в блоке, будем осуществлять адаптивным арифметическим кодом из [5] с помощью различных кодеров с вероятностями $\widehat{\pi}_n(1)$ и $\widehat{\pi}_n(0)$ для 1 и 0 соответственно. “Особые” символы 0 и 1 кодируются с вероятностями $\widetilde{q}(i)^{l_i}$ и $1-\widetilde{q}(i)^{l_i}$ для 0 и 1 соответственно, определяемыми из формулы (1). Наконец, символы в блоке, следующие после появления в этом блоке 1, кодируются с вероятностями $\widetilde{q}(j)$ и $\widetilde{p}(1,j)$ для 0 и 1 соответственно, определяемыми из (1) на основе статистики окна $x_{j-\omega}\ldots x_{j-1}$.

Оценивая избыточность кодирования, можно показать, что для построенного метода α со скользящим окном длины $\widehat{\omega}$ и точностью представления вероятностей t, определяемыми формулами (2), избыточность построенной адаптивной схемы не превосходит величины r, где $0\le r\le 1$. При этом среднее время T кодирования и декодирова-

ния данного метода удовлетворяет неравенству

$$T \le C_1 \cdot \sqrt{p} \cdot \log(1/rp) \cdot \log\log(1/rp) \cdot \\ \cdot \log\log\log(1/rp) + C_2,$$

где C_1, C_2-константы. Полученная оценка показывает, что средняя скорость кодирования и декодирования в $\sqrt{p}$ раз лучше времени ранее известных методов кодирования длин серий. Оценим общий объем памяти кодера и декодера. Так как на втором этапе алгоритма в памяти не нужно хранить кодеры, соответствующие каждой длине блока, то память второго этапа не превосходит C/r, где C - константа. Однако на первом и втором этапах кодирования в памяти необходимо хранить окно длины ω и $\widehat{\omega}$, что приводит к существенному увеличению общего объема памяти.

Построим теперь адаптивный метод кодирования β, для которого общий объем памяти кодера и декодера существенно меньше, чем в методе α. Пусть $\nu_i(x_1 \dots x_\omega)$ - частота встречаемости буквы $a_i \in A$ в слове $x_1 \dots x_\omega$, где $A = \{a_1, \dots a_k\}$. Определим оценки вероятностей $\widehat{p}(a_i)$ как

$$\widehat{p}(a_i) = \frac{\nu_i(x_1 \dots x_\omega) + 1}{\omega + k}.$$

Метод β основан на описанном выше алгоритме α, но на первом и втором этапах этого алгоритма вместо скользящего окна используются только счетчики частот встречаемости букв в окне $x_1 \dots x_\omega$, то есть в памяти хранятся только частоты. На втором этапе кодирования метода β вновь используется адаптивный арифметический код из [5]. Так как для записи частоты встречаемости одной буквы достаточно $\lceil \log \omega \rceil$ бит, то для метода β общий объем памяти кодера и декодера $V \le C_3 \cdot \log(1/r)$, где C_3 - константа.

Скорость кодирования и декодирования для метода β такая же, как и для метода α.

Работа частично поддержана Российским фондом фундаментальных исследований (грант № 11-07-00183а)

СПИСОК ЛИТЕРАТУРЫ

1. Capon, J. A probabilistic model for run-length coding of pictutes // J. Capon. – IRE Trans. Inform. Theory. - 1959, vol. IT-5. - P. 157-163.
2. Rothgordt, U., Intermediate ternary code: A redundancy reducing run-length code for digital facsimile/ U. Rothgordt, G. Renelt// Electron. Lett. – 1997, vol. 13. - P. 747-750.
3. Takagi, M. A highly efficient run-length coding scheme for facsimile transmission // M.Takagi, T.Tsuda – Electron. Commun. Jap. – 1975, vol.58 A, N 2. - P. 30-38.
4. Хантер, Р. Международные стандарты кодирования для цифровой факсимильной связи // Р. Хантер, А. Х. Робинсон – ТИИЭР – 1980, Т. 68, № 4 – С. 112-129.
5. Witten, I. H. Arithmetic coding for data compression // I. H. Witten, R. Neal, J. G. Cleary – Comm. ACM.- 1987, vol.30, N 6. – P. 520-540.

к.ф.-м.н. ***Бакулина М.П.*** *, научный сотрудник, тел. 8-961-215-79-36, marina@rav.sscc.ru - Институт Вычислительной Математики и Математической Геофизики СО РАН*

УДК: 004.052

МЕТОДИКА КЛАССИФИКАЦИИ ИС, ОБРАБАТЫВАЮЩИХ КОНФИДЕНЦИАЛЬНУЮ ИНФОРМАЦИЮ

Миронова В.Г.

Одним из этапов проведения предпроектного обследования информационных систем обработки конфиденциальной информации является их классификация. В статье предложен оригинальный способ определения классификационных признаков и формирования класса информационной системы обработки конфиденциальной информации.

Ключевые слова: конфиденциальная информация, информационная система, критерии классификации, класс.

В настоящее время проблема обеспечения информационной безопасности (ИБ) в информационных системах (ИС) обработки конфиденциальной информации стоит очень остро. Это обусловлено, прежде всего, большим количеством потенциальных угроз ИБ,