УДК 621.391.037

В.Я. Певнев, М.В. Цуранов

Харьковский национальный университет внутренних дел

ПОСТРОЕНИЕ ОПТИМАЛЬНЫХ КОДОВЫХ ТАБЛИЦ

Рассматриваются системы сжатия информации, основанные на построении множества кодовых таблиц с шести битовыми символами. Предлагается использование таблицы переходов от одного алфавита к другому с целью повышения помехоустойчивости передаваемых сообщений

Ключевые слова: *кодовые таблицы, таблицы переходов, бит, помехоустойчивое кодирование.*

Введение

Современные технологии передачи информации позволяют осуществлять обмен данными на большие расстояния, создавать глобальные и корпоративные сети. С каждым годом количество устройств подключенных к глобальной информационной сети неуклонно растет, популярность социальных сетей и технологий мгновенного обмена сообщениями приводит к росту количества передаваемого трафика. Для оптимизации нагрузки на каналы связи большинство программ передачи текстовой информации используют алгоритмы сжатия данных.

Постановка проблемы. При сжатии информации искажения хотя бы одного бита передаваемого сообщения может значительно затруднить процесс восстановления сжатых данных. Поэтому в таком случае необходимо применять более совершенные алгоритмы помехоустойчивого кодирования. Однако любые коды вносят избыточность в передаваемое сообщение, в некоторых случаях избыточность вносимая кодом может быть равна или даже больше степени сжатия информации. Оценка эффективности алгоритмов сжатия применительно к реальным условиям канала связи и используемым методам помехоустойчивого кодирования является наиболее важным этапом в выборе обнаруживающих и корректирующих кодов для использования в канале связи.

Анализ литературы. Существующие алгоритмы сжатия (далее в статье будут рассматриваться только алгоритмы для сжатия текстовой информации) условно можно разделить на два класса [1]. Первый класс – это алгоритмы, основанные на статистических методах сжатия. Идея метода - часто повторяющиеся символы нужно кодировать более короткими цепочками битов, чем цепочки редких символов. Наиболее известным представителем данного класса является метод Хаффмена [1]. Ко второму классу относятся алгоритмы, в которых производится кодирование последовательности из двух или более ранее уже встречавшихся символов новым символом. К данным алгоритмам относятся алгоритмы Лемпеля-Зива [2].

Главной идеей всех преобразований в обоих классах методов сжатия является создание словарей. В этих словарях происходит запоминание повторяющихся символов или их цепочек, которые кодируются цепочками меньшей длины. Таким образом, при обмене информацией необходимо передавать не только сам текст, но и соответствующий словарь. В этом случае передаваемый файл становится очень чувствительным к различным искажениям. Искажение даже одного бита приводит к тому, что файл нельзя распаковать на приемной стороне. В этом случае необходимо прибегать к помехоустойчивому кодированию. Как известно [1], применение помехоустойчивого кодирования ведет к росту объема передаваемого сообщения. При этом коды могут быть с обнаружением ошибок и с исправлением ошибок. Первые в случае появления ошибки требуют повторной передачи сообщения, вторые – за счет избыточности восстанавливают сообщение.

Под методом сжатия данных подразумевается совокупность действий, обеспечивающих уменьшение объема данных. Процесс сжатия можно описать выражениями [3]:

$$y(t) = F_{сж}\left(x(t)\right); \qquad V_{y(t)} < V_{x(t)},$$

где x(t) – исходный набор данных; y(t) – набор сжатых данных; Fсж – совокупность действий, составляющих метод сжатия; Vx(t) – объем исходных данных; Vy(t) – объем сжатых данных.

Соответственно процесс восстановления будет иметь вид

$$x'(t) = F_{восст}\left(y(t)\right),$$

где $x'(t)$ – восстановленный набор данных; F – совокупность действий, составляющих метод восстановления.

Под помехоустойчивостью систем связи понимается способность системы связи противостоять вредному влиянию помех. Основными показателями помехоустойчивости систем связи являются: вероятность ошибочного (правильного) приема элемента сообщения (или бита), пакета данных и сообщения в целом. Влияние сжатия данных на помехоустойчи-

вость систем связи также может быть оценено по этим показателям.

В работе [4] предложен метод сжатия данных суть которого заключается в преобразовании кодовой таблицы, состоящей из 256 символов в несколько таблиц размером в 64 символа. В кодовых таблицах ASCII и KOI для отображения любого символа используется один байт. С его помощью можно передать латинский алфавит (строчные и прописные буквы), национальный алфавит (строчные и прописные буквы), цифры, знаки препинания, специальные знаки, элементы псевдографики. Если рассмотреть операционные системы Windows, Linux и MacOS то каждому символу, изображаемому на экране или передаваемому в линию связи, соответствует два байта (кодировка на освнове стандарта Unicode).

При использовании указанного метода новые кодировочные таблицы или, как их принято называть, используя терминологию методов сжатия данных, словари содержат символы, которые можно кодировать с помощью 6 бит [5]. Данные словари должны размещаться в программе, обеспечивающей сжатие-восстановление данных. Очевидно, что в этом случае отпадает необходимость их передачи со сжатым файлом. Искажение одного или нескольких бит приведет только к потере одного или нескольких символов.

Данный метод позволяет избежать накапливания ошибок при восстановлении информации, однако он очень чувствителен к искажению последовательности смены кодовой таблицы. В методе [4] все последовательности смена кодовых таблиц имеют минимальное кодовое расстояние равное 1, что не позволяет обнаружить одиночные ошибки без применения помехоустойчивого кодирования.

В работе [4] проведен анализ длины передаваемого сжатого сообщения, которая по мнению авторов в большинстве современных систем не превышает 1024 бита.

В работе [3] был произведен анализа различных методов сжатия, в результате которого были получены следующие утверждения:

- с увеличением степени сжатия вероятность правильного приёма сообщения в системе связи с коммутацией пакетов увеличивается;
- с увеличением объема сообщения вероятность правильного приема сжатого сообщения увеличивается;
- с уменьшением длины пакета вероятность правильного приема сжатого сообщения увеличивается;
- с уменьшением вероятности искажения символа в канале связи прирост вероятности правильного приёма сжатого сообщения становиться меньше.

Таким образом [4], можно сделать вывод, что при сжатии данных вероятность ошибочного приема сообщения уменьшается, но цена ошибки увеличивается.

Для передачи небольших сообщений в надежных каналах связи наиболее целесообразно будет использовать метод описанный в [4]. Тем не менее при наличии даже одиночных ошибок в канале метод сжатия указанный выше имеет значительные ограничения.

Цель статьи – обосновать выбор оптимальных кодовых таблиц для алгоритма сжатия текстовой информации при ее передачами в канале связи с ошибками.

Изложение основного материала

Распределение ошибок. Для моделированиях групповых ошибок в реальных каналах связи было разработано большое количество моделей, которые математическими методами основываясь на законах распределения случайных величин, описывают потоки ошибок в реальных каналах. Более близкие к реальной работе канала связи результаты появления ошибок можно получить, если строить математическую модель потока ошибок, оперируя математическими понятиями, близкими к физическим явлениям, происходящим в канале связи. Существует ряд моделей, которые в определенной степени учитывают физику явлений, приводящих к искажению передаваемой информации. Наиболее известные среди них модели Гильберта [6], Эллиота-Гильберта[7], Фричмана-Свободы,[7] Флойлиха-Беннета[8], Попова-Турина[6].

Распределение групповых ошибок. В работе [9] были представлены результаты эксперимента в результате которого удалось доказать, что ошибки имеют тенденцию к группированию, однако в реальных системах количество рядом идущих ошибок в одном блоке достаточно мало. Что позволяет использовать коды обнаруживающие одиночные ошибки в реальных системах на достаточно малом размере кодируемого блока данных. В ходе эксперимента выяснилось, что при вероятности возникновения одиночной ошибки в канале равной $P_{ош} \approx 3*10^{-2}$, практически каждое из 10 полученных сообщений размером 8 бит будет невосстанавливаемым [9], поскольку возникает двойная ошибка, место возникновения которой определить очень сложно. При этом уже все закодированное сообщение восстановить практически невозможно.

При улучшении качества канала количество таких ошибок резко падает. Так при вероятности возникновения одиночной ошибки в канале равной $P_{ош} \approx 10^{-3}$ невосстанавливаемым является один из 50 блоков при использовании миниблоков размером 4 бита или один из 20 блоков при использовании миниблока размером 8 бит при передачи блока, размером 8192 бит [9].

Выбор длинны кодируемого блока. Наиболее важным параметром, при использовании в реальных

системах, является размер кодируемого блока. При использовании кодовых таблиц рассмотренных в [4] наиболее подходящим размером контролируемого блока становится 4 бита [10]. Основным аргументом для выбора такого размера служит принцип восстановления ошибочно принятых битов – чем больше размер контролируемого блока, тем больше вариантов необходимо просмотреть. Количество просматриваемых вариантов можно определить как

$$K = D * q$$

где K – количество просматриваемых вариантов; q – количество искаженных символов в блоке; D – размер контролируемого блока.

В результате моделирования произведенного в [9,10] вероятность появления контролируемых блоков с более чем одной ошибкой, при передаче сообщения размером 1 килобайт при вероятности искажения одного символа 10^{-3}, не превышает 0.02. При этом среднее количество искаженных контролируемых блоков составило чуть более 8, а максимальное их число равняется 18.

Результаты, полученные в [9,10] получены при имитационном моделировании канала связи. Предполагалось, что канал симметричный, закон распределения ошибок равномерный. Высокая достоверность результата достигнута за счет проведения 5000 опытов в одной точке.

Исходя из результатов эксперимента [9,10], можно сделать вывод, что групповые ошибки становятся одиночными, при условии использования достаточно малых контролируемых блоков. Исходя из этого использование 6 битных кодовых таблиц, является оптимальным в каналах с вероятностью ошибки равной $P_{ош} \approx 10^{-3}$.

С точки зрения организации данных в ЭВМ следует разбить предлагаемую последовательность на две части по 3 бита в каждой и к каждой добавить бит проверки на четность. Таким образом, можно построить кодограмму, которая будет состоять из контролируемых блоков размером в 4 бита. Также последовательность из 8 бит будет удобна для последующей обработки и передачи каналами связи.

Порядок формирования кодовых таблиц. В работах [4,10] указан порядок формирования кодовых таблиц, при условии, что для кодирования символа используется 6 бит и 2 бита для помехоустойчивого кодирования. Проанализируем типовую кодовую таблицу представленного метода. В табл. 1 представлены коды перехода на другой алфавит, используемые в методе описанном в [4]. Наибольший интерес представляют коды перехода на другой алфавит, поскольку ошибочное декодирование данного кода приведет к необходимости к необратимым потерям и невозможности расшифровать весь текст сообщения. Например, если при отправке сообщения вместо кода 111100 (переход к алфавиту 4) будет искажен четвертый бит, то мы получим код 111000 (переход к алфавиту 3). В результате все последующие переданные сообщения будут декодироваться неправильно, и в результате проверки потребуется повторная передача символов либо повторное декодирование принятых ранее сообщений. Данную ошибку призваны исправить два проверочные бита, однако если ошибка возникнет одновременно в двух мини блоках по 4 бита, то исправить данное сообщение будет невозможно.

Таблица 1

Кодовые символы перехода на новый алфавит

№ в кодовой таблице	Двоичное представление	Символы алфавита
51	110011	Пер-д к алфав. 1
52	110100	Пер-д к алфав. 2
56	111000	Пер-д к алфав. 3
60	111100	Пер-д к алфав. 4
63	111111	Пер-д к алфав. 5

Кроме описанных выше проблем представленный в [4,10] метод кодирования имеет и другие существенные недостатки:

- символы смены алфавита дублируются во всех кодовых таблицах, что значительно сокращает количество символов, которые можно выделить для других символов;
- в кодовой таблице можно выделить ограниченное число записей для символов перехода на другой алфавит, что сокращает число одновременно используемых алфавитов;
- минимальное кодовое расстояние между символами смена алфавита равняется 1, что не позволит обнаружить даже одинарную ошибку [1], в случае использования контрольной суммы минимальное кодовое расстояние будет увеличено до 2, однако это не позволит обнаруживать двойные ошибки.

Порядок формирования таблицы переходов. Для устранения указанных выше недостатков авторами был предложен метод, заключающийся в формировании отдельной кодовой таблицы с символами перехода на другой алфавит. Предложенный метод заключается в том, что вместо того чтоб вставлять код перехода на другой алфавит в каждую кодовую таблицу, создается специальная таблица, содержащая коды всех алфавитов используемых в системе. Для перехода к кодовой таблице в каждом алфавите существует единая комбинация. Для повышения защищенности кода перехода к таблице алфавитов, авторы предлагают использовать комбинацию, состоящую из одних нулей.

Получив подобную комбинацию, декодер будет знать, что последующий полученный байт необходимо декодировать из отдельной кодовой таблицы. Пример кодовой таблицы для 9 алфавитов рассмотрен в табл. 2.

Таблица 2

Кодовая таблица символов перехода на новый алфавит

№ в кодовой таблице	Двоичное представление	Символы алфавита
0	100001	Пер-д к алфав. 1
1	001100	Пер-д к алфав. 2
2	110000	Пер-д к алфав. 3
3	110101	Пер-д к алфав. 4
4	111111	Пер-д к алфав. 5
5	011110	Пер-д к алфав. 6
6	101101	Пер-д к алфав. 7
7	010010	Пер-д к алфав. 8
8	000011	Пер-д к алфав. 9

Как видно из таблицы минимальное кодовое расстояние для каждой комбинации, без учета проверочных бит, равно 2. Что позволяет обнаруживать одиночные ошибки, в случае использования двух проверочных бит кодовое расстояние увеличивается до 4. Данный фактор позволяет говорить, что использование предлагаемой кодовой таблицы, для символов перехода на другой алфавит, позволит обнаружить двойную ошибку. Это позволит увеличить достоверность декодирования символов перехода на другой алфавит и снизит вероятность повторной передачи данных в результате их неправильного декодирования.

Также следует заметить, что в табл. 2 рассмотрено 9 алфавитов, что практически в два раза превышает количество алфавитов используемых в таблицах метода [4]. Использование предлагаемого метода кодирования позволяет существенно расширить количество одновременно используемых алфавитов либо количество одновременно используемых символов. Это позволит более эффективно использовать предложенную в [4] методику кодирования в международных системах обмена информацией и с некоторых национальными алфавитами.

Еще одним преимуществом предложенной методики по сравнению с изложенной в [4] является меньшая избыточность данных в каждой кодовой таблице, поскольку во всех таблицах повторяется только одна кодовая комбинация, переход к таблице алфавитов.

Выводы

Построение оптимальных кодовых таблиц позволяет уменьшить избыточность передаваемых сообщений. Введение новой таблицы переходов позволяет увеличить устойчивость к ошибочному переходу на другой алфавит. Это достигается за счет увеличения кодового расстояния между элементами таблицы переходов.

Список литературы:

1. Хэмминг Р.В. Коды с обнаружением и исправлением ошибок / Коды с обнаружением и исправлением ошибок. – М.:ИЛ,1956. – с.7 – 22.

2. Кричевский Р.Е. Сжатие и поиск информации. – М.:Радио и связь, 1989. – 168 с.

3. Стрюк А.Ю., Кабакчей Р.М., Клименко К.С. Влияние сжатия данных на помехоустойчивость приема сообщений//. Системи обробки інформації. Х. – 2002, випуск 5(21) – С. 277 – 280.

4. Пєвнєв В.Я., Яценко І.Л. Про один засіб стиску текстової інформації//Вісник Житомирського інженерно – технологічного інституту. Житомир. – 2002, №IV(23) – С.206 – 209.

5. Пєвнєв В.Я., Яценко І.Л. Побудова таблиць, що кодують// Тез. Докл. Междун. . Конф. «Проблемы информатики и моделирования – 2003», Харьков, НТУ «ХПИ», 2003

6. Гильберт Э.Н. Пропускная способность канала связи с пакетными ошибками./Кибернетический сборник, № 9. – М.: Мир,1964.

7. Статистика ошибок при передачи цифровой информации: Пер. С англ. Под ред. С.И.Самойленко. – М.:Мир,1966.

8. Советов Б.Я., Стах В.М. Построение адаптивных систем передачи информации для автоматизированного управления. – Л.: Энергоиздат, 1982.

9. Пєвнєв В.Я., Цуранов М.В. Экспериментальные исследования моделей групповых ошибок в каналах связи // Вісник НТУ „ХПІ”. Збірник наукових праць. Харків:НТУ”ХПІ”. – №49. – 2011. – С. 115 – 121

10. Певнев В.Я., Яценко И.Л. Метод восстановления информации при обмене данными в распределенных вычислительных системах//Вісник. Кременчуцького державного політехнічного університету. Кременчук. – 2003, Вип.3/2003 (20) – С. 19 – 21.

Поступила в редколлегию 14.03.2012

Рецензент: д-р техн. наук, проф. А.А. Серков, НТУ «ХПИ», Харьков.

ПОБУДОВА ОПТИМАЛЬНИХ КОДОВИХ ТАБЛИЦЬ

В.Я. Певнєв, М.В. Цуранов

Розглядаються системи стиснення інформації, засновані на побудові множини кодових таблиць з шести бітовими символами. Пропонується використання таблиці переходів від одного алфавіту до іншого з метою підвищення завадостійкості переданих повідомлень

Ключові слова: *кодові таблиці, таблиці переходів, біт, завадостійке кодування.*

CONSTRUCTION OF OPTIMAL CODE TABLES

V.Ya. Pevnev, M.V. Tsuranov

We consider the data compression system based on building a set of code tables with six bit characters. It is proposed to use a table of transitions from one alphabet to another in order to improve noise immunity of transmitted messages

Keywords*: code table, transition table, the bit noise-stable coding.*