Лекция 5. СЖАТИЕ ДАННЫХ. АРХИВАТОРЫ

Программы архивации

Назначение программ архивации (программ-архиваторов) — экономить место на диске за счет сжатия (упаковки) одного или нескольких исходных файлов в архивный файл. Программы-архиваторы используются для хранения в упакованном виде больших объемов информации, которая понадобится только в будущем; переноса информации между компьютерами с помощью дискет или электронной почты; создания в сжатом виде резервных копий файлов. В результате работы программ-архиваторов создаются архивные файлы (архивы).

Кластер — здесь: минимальная единица дискового пространства, выделяемого ОС для записи файлов. Как правило, файл при хранении занимает несколько кластеров. Объем кластера жесткого диска для ОС Windows может быть 4К, 8К, 16К, 32К.

В основе работы программ-архиваторов лежит процедура поиска и перекодирования одинаковых фрагментов содержимого файлов. Существует множество разнообразных подходов к сжатию данных. В качестве иллюстрации простейшего метода сжатия данных опишем механизм энтропийного кодирования.

Суть этого кодирования заключается в представлении часто встречающихся символов (сочетаний символов) короткими кодами, а редко встречаемых — более длинными. Предположим, что в исходной кодируемой последовательности встречаются только n символов: S_1, S_2,..., S_{n-2}, S_{n-l}, S_n с вероятностью появления p_1, p_2, ..., p_{n-2}, p_{n-1}, p_n. Для простоты будем считать, что символы отсортированы в порядке убывания этой вероятности. Объединим два символа S_{n-1} и S_n с наименьшими вероятностями появления в один комбинированный символ S и рассчитаем вероятность его появления $p_{n-1}+p_n$. В результате получим последовательность из $n-1$ символов: S_1, S_2,..., S_{n-2}, S. В дальнейшем этот символ S участвует в обработке наравне с

исходными. Символы опять сортируют в порядке убывания вероятности их появления, и повторяют процедуру объединения до тех пор, пока не останется только два символа. Затем в обратном порядке происходит кодирование исходных символов. Поясним это следующим примером.

Пример 1.1. Имеется файл, содержащий только символы S_1, ..., S_4 с вероятностями их появления {0,6 0,2 0,15 и 0,05} соответственно. Требуется представить эти символы сокращенным кодом.

Решение. Упорядочим символы по возрастанию вероятностей их появления и объединим по приведенному алгоритму (рис. 1.3).

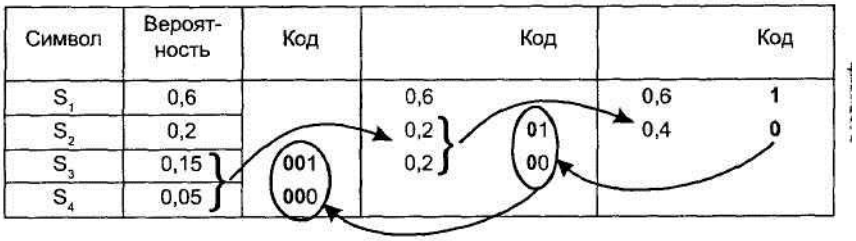

РИС 1.3 Иллюстрация процесса *кодирования информации*

На последнем шаге объединения установим, что символ S_1 кодируется 1 (единицей), а код остальных символов имеет первый 0 (нуль). Затем установим, что символ S_2 кодируется кодом 01, а коды символов S_3 и S_4 имеют первые 00. Аналогично, на последнем шаге определяем, что код символа S_3 равен 001, а символа S_4 — 000.

В результате символ S_1 кодируется 1 (1 битом), символ S_2 — 01 (2 битами), а символы S_3 и S_4 кодируются 001 и 000 (3 битами каждый).

Оценим эффект уменьшения размера исходного текста при таком кодировании. Предположим, имеется последовательность из 1000 таких символов S_1, ..., S_4. Тогда символ, обозначенный через S_1, присутствует в последовательности 600 раз, символ S_2 — 200 раз и символы S_3 и S_4 по 150 и 50 раз соответственно. Расчет не сложный, достаточно вероятность появления символа умножить на объем выборки. Общая длина закодированной последовательности равна: (1x600) + (2x200)+ (3x150) + (3x50) = 1600 бит.

Если кодировать символы без учета вероятности их встречаемости,

например символ S_1 кодируется через 00, символ S_2 через 01 и символы S_3 и S_4 через 10 и 11 соответственно (каждый символ кодируется ровно двумя битами), то последовательность из 1000 таких символов займет **2x1000** = **2000** бит.

Эффект сокращения длины кода при использовании энтропийного кодирования с 2000 бит до **1600** (на 20%) очевиден, а с учетом обычно используемого для представления каждого символа длины кода в 8 бит тем более очевиден (с 8000 бит до 1600 — то есть в 5 раз).

Другой подход к сжатию данных используется, например, при кодировании изображений. Изображение представляет собой последовательность точек (пикселей), каждая из которых кодируется несколькими байтами. При этом, как правило, велика вероятность нахождения рядом точек одного цвета. Поэтому целесообразно запись последовательности цветов точек указывать в виде пар чисел (количество рядом расположенных одноцветных точек и число, определяющее их цвет).
В реальных программах-архиваторах процедура поиска и перекодировки намного сложнее.

Типовые функции программ-архиваторов состоят в:

- помещении исходных файлов в архив,
- извлечении файлов из архива,
- удалении файлов из архива,
- просмотре оглавления архива,
- верификации (проверки) архива.

Первые программы-архиваторы были ориентированы на работу под управлением MS DOS: ARJ, PKZIP/PKUNZIP, PAK, LHA, RAR. Эти программы отличались форматом архивных файлов, скоростью работы, степенью сжатия файлов в архиве, интерфейсом пользователя. Общим их недостатком являлся недостаточно удобный интерфейс. Пользователю необходимо было помнить форматы команд и постоянно указывать их в командной строке. Этого недостатка лишены программы-архиваторы,

ориентированные на работу под управлением ОС Windows. Среди современных программ-архиваторов выделяют: WinRAR (разработка Е. Рошал), WinZip фирмы Niko Mak Computing и др.

В пособии рассмотрено использование архиватора WinRAR, отличающегося большой степенью сжатия, работой с длинными именами файлов, удобным интерфейсом. Этот архиватор поддерживает обработку многих архивных форматов и использует оригинальный алгоритм упаковки, особенно эффективный для исполняемых и текстовых файлов. К важным дополнительным возможностям программы относят: защиту архива при помощи пароля, восстановление поврежденных архивов, создание многотомных и самораспаковывающихся архивов, сохранение комментариев к архивам. Пользовательский интерфейс WinRAR содержит основное меню, панель инструментов и рабочую область, в которой показаны все файлы текущей папки, (рис. 1.4 *а).* При работе с WinRAR архивы воспринимаются как папки, содержимое которых можно просмотреть традиционными способами.

Основное меню архиватора состоит из пунктов *Файл, Команды, История, Избранное, Параметры* и *Справка,* содержащих сгруппированные по функциональному назначению команды архиватора (см. рис. 1.2).

Команды пункта меню *Файл* выполняют операции над файлами, содержащимися в архиве или помещаемыми в архив.

Команда *Пароль* применяется при установке пароля на вновь создаваемый архивный файл. При выборе этой команды пользователю следует в появившемся окне набрать и подтвердить пароль. Впоследствии без знания этого пароля невозможно будет получить доступ к содержимому хранящихся в архиве файлов.

Следующая группа команд используется для выделения нескольких файлов. Так, команда *Выделить все* автоматически выделяет все файлы текущей папки. Команда *Выделить группу* активизирует маску ввода шаблона файлов, удовлетворяющих некоторому критерию. Аналогично,

команда *Снять выделение* вызывает маску шаблона для отмены выделенных файлов.

Второй пункт основного меню *Команды* содержит команды обработки содержимого архива.

Для создания архива или добавления данных в существующий архив используется команда *Добавить файлы в архив* или пиктограмма с изображением стопки книг. При этом в архив помещаются предварительно выделенные файлы. При выборе этой команды на экране появляется диалоговое окно *Имя и параметры архива* (рис. 1.4).

В поле *Архив* этого окна указывается местонахождение помещаемых в архив файлов. По умолчанию архив создается в текущем каталоге. Для указания другой папки следует использовать кнопку *Обзор.*

Содержимое поля *Метод сжатия* определяет степень сжатия создаваемого архива. По умолчанию установлен вариант *Обычный.* Если установить наилучшую степень сжатия, то архивный файл будет занимать меньший объем, однако время его создания возрастет. В поле *Размер словаря* задается длина фрагмента, в рамках которого алгоритм сжатия ищет повторения для кодирования и сжатия.

Параметр *Размер тома* применяется при создании многотомных архивов. Его установка позволяет создавать архив в виде нескольких файлов, размер которых не превосходит заданного размера тома. Это особенно актуально при необходимости последующего переноса архива на гибких дискетах, если архив превышает емкость имеющихся носителей, или при пересылке по электронной почте. На рисунке 1.4 б в этом поле указан размер 3,5" дискеты.

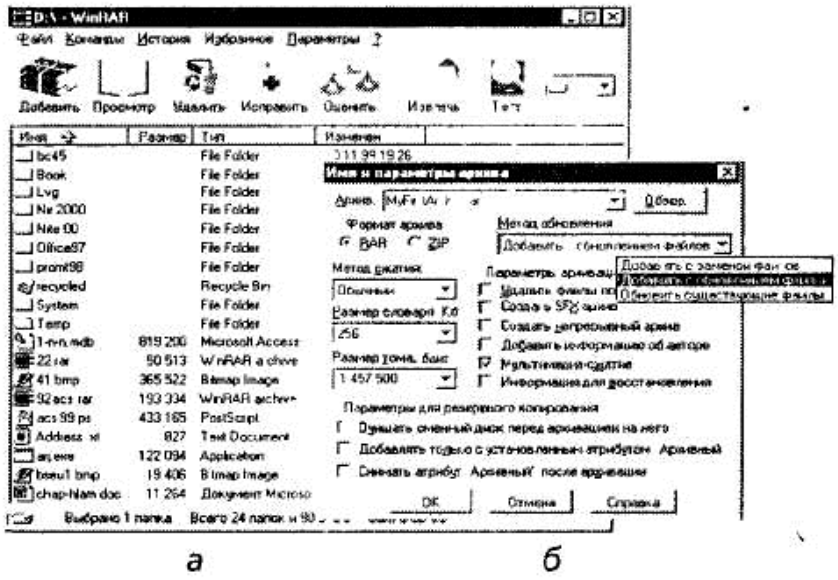

Рис. 1.4. Пользовательский интерфейс архиватора WinRAR *(а) и диалоговое окно команды* ***Добавить файлы в архив (б)***

В поле со списком *Метод обновления* определяются варианты помещения файлов в архив. Устанавливаемый по умолчанию режим *Добавить с заменой файлов* определяет архивирование всех выделенных пользователем файлов. Указание режима *Обновить существующие файлы* позволяет добавить в архив лишь те файлы, старые версии которых уже находятся в архиве. Режим *Добавить с обновлением файлов* помещает в архив файлы, копии которых в архиве отсутствуют.

Группа команд *Параметры архивации* позволяет выбрать алгоритм сжатия данных. По умолчанию программа настроена на базовый вариант. Однако, например, для мультимедийных данных целесообразно использовать вариант *Мультимедиа-сжатие.*

Когда настройка завершена, следует нажать кнопку *ОК*. Появившееся динамическое окно иллюстрирует текущее состояние процедуры архивации (степень обработки очередного файла набора и общее состояние выполнения всей процедуры).

На этом процедура создания архива или добавления данных в существующий архив закончена.

Команда *Восстановить архив* используется при нарушении целостности архива, возникающем, например, в результате его длительного

хранения.

Архиватор WinRAR позволяет удалять ненужные файлы, как это делается в приложении *Проводник.* Для этого используется команда *Удалить файлы,* нажатие клавиши <Delete> или пиктограммы с изображением корзины.

Остальные команды этого меню относятся только к файлам, содержащимся в архиве, и становятся доступными, если в рабочую область WinRAR загружен файл архива.

Команда *Извлечь файлы из архива* обеспечивает распаковку предварительно выделенных пользователем файлов из данного архива (рис. 11.5). Если необходимо развернуть архивные данные не в текущую папку, то следует воспользоваться командой *Извлечь в другую папку* и указать путь к этой папке-получателю.

Тестирование отдельных файлов в архиве на предмет возможных повреждений их структуры производится с помощью команды *Протестировать файлы в архиве.* Эту команду следует применять для проверки целостности файлов при их длительном хранении, особенно на ненадежных магнитных носителях.

Архив можно снабдить комментарием, воспользовавшись командой *Добавить архивный комментарий* или соответствующей пиктограммой. Выбор команды *Добавить информацию для восстановления* вызывает специальную процедуру, которая вносит в текущий архив дополнительные данные, повышающие его устойчивость к сбоям. Однако это оборачивается небольшим увеличением объема архива.

Архиватор WinRAR позволяет создавать самораспаковывающиеся архивы, разворачивающиеся при запуске их на исполнение. Режим становится доступным, если в среду WinRAR загружен какой-нибудь архивный файл. Для этого требуется выполнить команду *Преобразовать архив в SFX.*

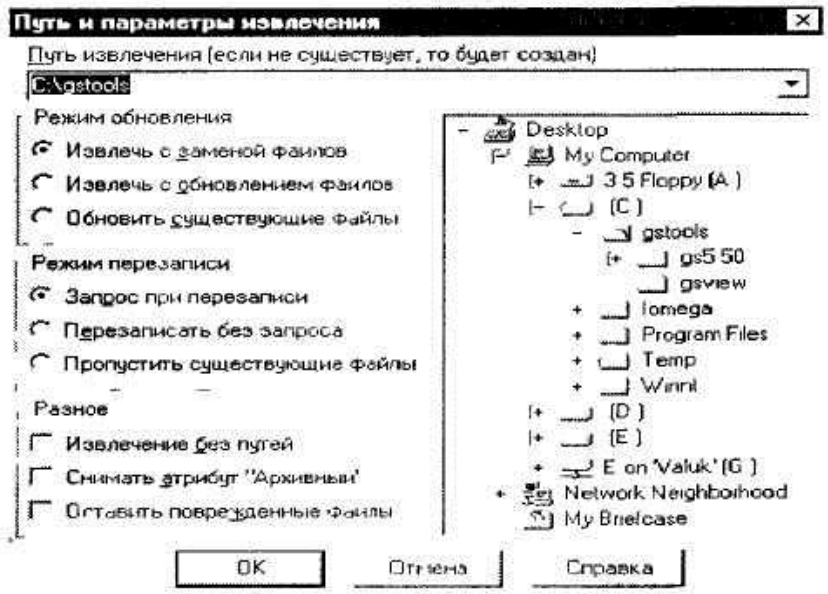

РИС. 1.5. Пользовательский интерфейс диалогового окна ***Извлечь файлы из архива***

Выбор команды *Информация об архиве* позволяет получить сведения об архиве, загруженном в данный момент в рабочую область (размер и количество файлов в архиве; коэффициент сжатия архива; наличие комментариев; наличие пароля; операционная система, для которой этот архив создан).

Команды меню *История* обеспечивают доступ к последним обрабатываемым архивам, с которыми работал пользователь.

Группа команд меню *Параметры,* предназначена для настройки основных параметров архиватора WinRAR и регистрации пользователей через Интернет.

Выбрав в диалоговом окне команды *Установки* соответствующие вкладки, пользователь имеет возможность: определить интерфейс архиватора; задать значения по умолчанию основных параметров архиватора (метода сжатия, размера словаря); определить папку, в которую следует помещать файл архива и др.

Команды последнего пункта основного меню *Помощь* описывают возможности работы архиватора, поясняют технику работы с ним и содержат информацию о разработчике и процедуре, приобретения архиватора.