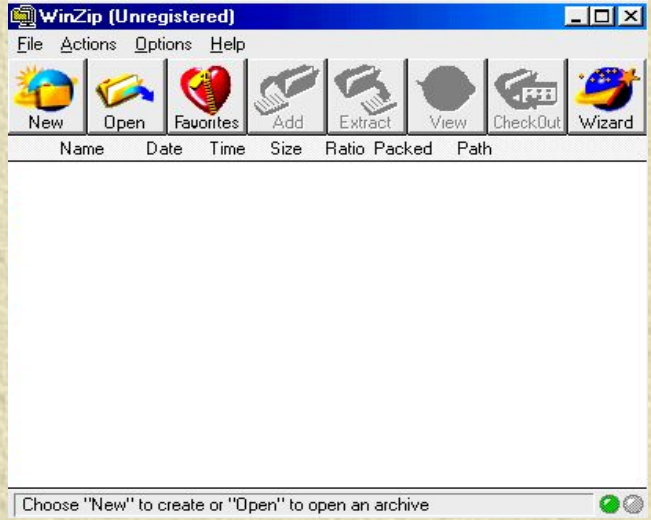
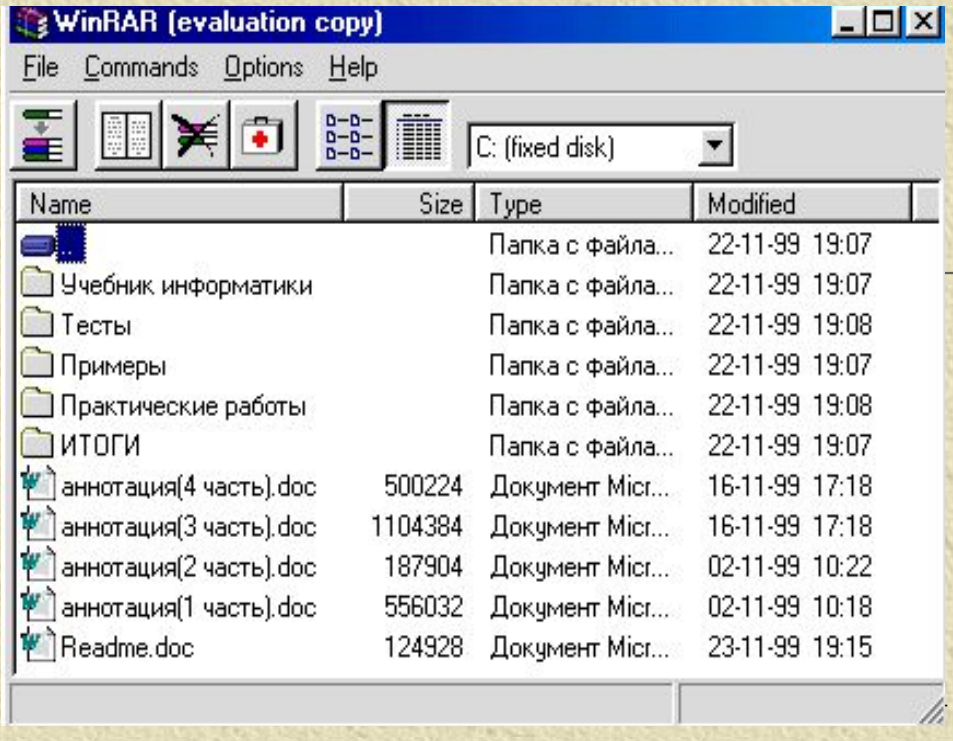
Методы сжатия информации





ПРОГРАММЫ – АРХИВАТОРЫ

Работая с компьютером, пользователь постоянно сталкивается с проблемой больших объемов информации, подлежащих, как хранению, так и перемещению.

Основные ситуации, требующие уменьшения объема данных это:

- экономия места на носителях информации
- перенос информации с компьютера на компьютер
- передача информации по электронной почте

Для чего и были созданы специальные средства, позволяющие уменьшить объем дискового пространства, занимаемого файлами. Эти средства используют различные методы сжатия информации, основанные на избыточности данных, записанных в файлах. Функцию сжатия информации выполняют специальные **программы – архиваторы.**

Под **архивацией** понимают слияние нескольких файлов и даже папок в единый файл - архив.

Упаковка (сжатие) - это уменьшение объема исходных файлов за счёт устранения избыточности информации.

Цель сжатия – размещение информации на носителях внешней памяти в более компактном виде, что требует меньших объёмов памяти.

Современные архиваторы обеспечивают выполнение **обоих** указанных функций.

Архиватор (упаковщик) – программа, создающая копии файлов меньшего объема, используя специальные методы сжатия, и объединяющая их в один архивный файл.

Степень сжатия зависит от:

- **используемого архиватора;**
- **метода сжатия;**
- **типа исходного файла.**

И характеризуется коэффициентом

$$K_c = \frac{V_c}{V_и} 100\% ,$$

где V_c – объём сжатого файла;
$V_и$ – объём исходного файла.

Методы сжатия информации

делятся на две группы:

1. **Сжатие с потерями (lossy compression)**
2. **Сжатие без потерь (lossless compression);**

1. Методы сжатия с потерями обычно используются для передачи изображения и звука и позволяют достичь коэффициента десятикратного сжатия. При сжатии с потерями часть информации утрачивается, но при этом усеченная часть информации не будет заметна для человеческого глаза и уха, т.к. это не окажет существенного влияния на восприятие информации в целом.

Данный метод имеет два существенных **недостатка:**

1. Невозможность достоверности анализа графической информации
2. Повторная компрессии и декомпрессия приводит к эффекту накопления погрешностей.

Характерные форматы сжатия **с потерей** информации:

- **MPG** для видеоданных
- **JPEG** для графических данных (неподвижных изображений)
- **MP3** для звуковых данных

Методы сжатия без потерь используются для любой информации, т. к. обеспечивают абсолютно точное восстановление данных после кодирования и декодирования. Данный метод основан на принципе преобразования данных из одной группы символов в другую, более компактную.

Характерные форматы сжатия **без потери** информации:

- **GIF, TIF, PCX** и др. для графических данных
- **AVI** для видеоданых
- **ZIP, ARJ, RAR, LH, CAB** для любых данных

Принцип работы любого метода основан на поиске избыточной информации и последовательной ее кодировке с целью уменьшения объема. Размер файла и объем информации измеряются в одних и тех же единицах - **байтах**. Но это не одно и то же. Обычно объем информации, заключенный в файле меньше размера самого файла. Согласно теории информации необходимо различать **семантику** (смысл) и **синтаксис** сообщения (набор правил, выражающих информацию).

Задача сжатия информации без потерь заключается в том, чтобы представить ее в виде набора непредсказуемых чисел, и тогда размер файла соответственно приблизится к размеру содержащейся в ней информации.

Метод Хаффмана (Huffman method, 1952 г.) - один из ранних методов сжатия информации и заключается в кодировании символов алфавита кодами различной длинны. Чем чаще встречается символ, тем короче код, по аналогии с азбукой Морзе.

Так обычный текстовый файл содержит алфавитно-цифровые символы и непечатные коды управления. Каждый символ в таблице ASCII (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange) представлен одним байтом, что составляет 8 бит. В технических системах на каждый символ приходится 1 бит, т.к. буквы в сообщениях встречаются с различной частотой (вероятностью). Так, например, в русском языке самыми распространенными буквами считаются буквы: **О, Е, А**, а буквы: **Ф, Ц, Щ, Э** встречаются редко. Логично разные буквы кодировать различным количеством нулей и единиц. Причем, наиболее часто встречающемуся символу ставится в соответствие самый короткий код.

Существует несколько реализаций метода Хаффмана, причем не только для текстовых файлов, т.к. неравномерность частоты появления тех или иных байтов информации характерна практически для любого типа файла. Метод Хаффмана дает достаточно высокую скорость и хорошее качество сжатия информации.

Метод Лемпеля-Зива (Lempel-Ziv, 1977 г.) - заключается в кодировании последовательностей символов, в отличие от алгоритма Хаффмана, где кодируются отдельные символы. Он основан на замене одинаковых строк ссылкой на аналогичную строку, ранее встречающуюся в тексте. Т.е. чем длиннее строка, и чем чаще она встречается в тексте, тем больше будет степень сжатия файла. Задача состоит в том, чтобы найти как можно больше одинаковых строк и наиболее эффективно их закодировать. Метод **LZ** достаточно эффективен для различных видов информации и обладает высокой скоростью декодирования.

Впоследствии появился метод, названный **LZH** (Lempel-Ziv-Huffman), который объединил алгоритмы вышеописанных методов.
Методы сжатия информации, основанные на алгоритмах Лемпеля-Зива, используются в прграммах-архиваторах PKZIP, WinZip, ARJ и RAR.

Современные средства архивации данных используют и более сложные алгоритмы, основанные на комбинации нескольких теоретических методов. Общим принципом в работе таких **синтетических** алгоритмов является предварительный просмотр и анализ исходных данных для индивидуальной настройки алгоритма на особенности обрабатываемых данныхю.

Примеры программ-архиваторов

- *ARJ*

- *WinZip (ZIP)*

- *WinRAR (RAR)*

RAR - это мощное средство создания архивов и управления ими. Существует несколько версий RAR для разных операционных систем, в частности, RAR для **Windows**, **Linux, DOS**, **OS/2**, **UNIX**. RAR для Windows поставляется в двух вариантах:

- **WinRAR.exe –** 32-разрядная версия с графическим интерфейсом пользователя (GUI)
- **Rar.exe** - 32-разрядная консольная версия, работающая из командной строки в текстовом режиме. Консольную версию RAR удобно использовать для вызова из пакетных файлов (BAT и CMD), для запуска из приглашения DOS и др. Она поддерживает больше команд и ключей в командной строке, чем WinRAR..

Характеристики Программ-Архивторов:

- **Степень сжатия – K_A** (коэффициент сжатия)
- **Время сжатия – T_A** (скорость архивации)
- **Объем Архива – V_A**
- **Число файлов в Архиве – N_A**
- **Управление архивами других форматов** (универсальность)
- **Дополнительные параметры**

(тестирование и восстановление поврежденных архивов, добавление комментариев и др.)

Основные параметры архивации WinZip:

- **Имя Архива** (по умолчанию - Имя файла)
- **Адрес (папка) Архива** (по умолчанию - текущая папка)
- **Формат** (7z, Tar, Zip)
- **Уровень сжатия** (без сжатия, скоростной, быстрый, нормальный, максимальный, ультра)
- **Режим изменения Архива** (добавить и заменить, обновить и добавить, обновить, синхронизировать)
- **Опции** (создание SFX-архива, сжимать открытые для записи файлы)
- **Шифрование Архива:** (ввод пароля)

Общие параметры архивации WinRar:

- **Имя Архива** (по умолчанию - Имя файла)
- **Адрес (папка) Архива** (по умолчанию - текущая папка)
- **Формат** (Rar или Zip)
- **Метод сжатия** (без сжатия, обычный, хороший, максимальный, быстрый, скоростной)
- **Метод обновления Архива**
- **Параметры архивации** (удаление файла после упаковки, создание SFX-архива, создание непрерывного архива и др.)
- **Комментарии Архива:** (из файла или вручную)