DOI: 10.17117/na.2016.12.02.212

<http://ucom.ru/doc/na.2016.12.02.212.pdf>

Поступила (Received): 19.12.2016

Гасангусейнов К.А. Архивация данных методом Шеннон-Фано

Gasanguseinov K.A.
Backup method Shannon-Fano

В данной статье автор ставит цель дать описание алгоритму, написанному двумя американскими учеными Клодом Шенноном и Робертом Фано. Основная мысль на которой акцентирует внимание автор заключается в том, чтобы показать проблему сжатия информационных данных для передачи по сети интернет или для переноса на другие устройства

Ключевые слова: архивация, метод Шеннон-Фано, алгоритм, код, информация

In this article, the author aims to give a description of the algorithm, written by two American scientists Claude Shannon and Robert Fano. The basic idea on which the author emphasizes is to show the problem of compression for transmission of information data on the Internet or for transfer to other devices

Key words: archiving, Shannon-Fano method, algorithm, code, information

Гасангусейнов Кади Арсланович

Студент

Дагестанский государственный университет
г. Махачкала, ул. М. Гаджиева, 43 А

Gasanguseinov Kadi Arslanovich

Student

Dagestan state university
Makhachkala, M. Gadzhieva st., 43 A

Алгоритмы сжатия:

Они бывают двух видов, с потерями и без потерь. Первый зачастую применяют для сжатия аудио, видео форматов, а также изображений, благодаря потери качества можно существенно уменьшит размер, при этом сохранив “читабельность”.

А алгоритмы без потерь применяют для сжатия текста, данных и вообще любого рода данных, которые впоследствии требуется в точности восстановить.

В данной статье автор ставит цель дать описание алгоритму, написанному двумя американскими учеными Клодом Шенноном и Робертом Фано. Основная мысль на которой акцентирует внимание автор заключается в том чтобы показать проблему сжатия информационных данных для передачи по сети интернет или для переноса на другие устройства.

Алгоритм Шеннона-Фано, теория:

Алгоритм использует коды переменной длины: часто встречающийся символ кодируется кодом меньшей длины, редко встречающийся – кодом большей длины. Это алгоритм префиксного неоднородного кодирования, относится к вероятностным методам сжатия.

Код строится с помощью дерева, построение этого дерева начинается от корня. Всё множество кодируемых элементов соответствует корню дерева. Оно разбивается на два подмножества с примерно одинаковыми суммарными вероятностями. Ветви кодового дерева размечаем символами 1 и 0.

При построении кода Шеннона – Фано разбиение множества элементов может быть произведено. Выбор разбиения на уровне n может ухудшить варианты разбиения на следующем уровне $(n + 1)$ и привести к не оптимальности кода в целом. Оптимальное поведение на каждом шаге пути ещё не гарантирует оптимальности всей совокупности действий. Поэтому код Шеннона – Фано не является оптимальным в общем смысле, хотя и дает оптимальные результаты при некоторых распределениях вероятностей. Теперь описав алгоритм можно построить дерево.

Пример кодового дерева

Предположим, что у нас есть предложение где у нас буква А встречается 50 раз, В-39, С-18, В-49, Е35, F-24 раза.

Отсортировав по убыванию, мы получим:

А – 50 D – 49 В – 39 Е – 35 F – 24 С – 18

№1. Теперь просуммировав количество букв и разделив их на 2.

№2. Большую часть помещаем в левое поддерево, а меньшую часть в правую.

№3. Если на ветке больше одной буквы возвращаемся к шагу #1.

В итоге у нас получится вот -такое дерево.

Теперь пройдясь по всему дереву можно получить коды букв.

А – 11, В – 101, С – 100, D – 00, Е – 011, F – 010.

Теперь вычислим $P(i)$, вычисляется она

По формуле: $P(i) = \frac{i}{\sum_{k=1}^n word}$.

Проще говоря количество конкретной буквы делится на суммарное количество букв в слове. Нарисуем таблицу.

Буква	Двоичное число	P(i)
A	11	$\approx 0,2325581$
D	00	$\approx 0,2279069$
B	101	$\approx 0,1813953$
E	011	$\approx 0,1627906$
F	010	$\approx 0,1116279$
C	100	$\approx 0,0837209$

Давайте теперь убедимся, что составленный нами код при отсутствии ошибок в самом деле является оптимальным. Для этого найдем сначала среднюю информацию, содержащуюся в одной букве текста по формуле [2, 127].

$$H = - \sum_{i=1}^n p_i \log p_i$$

Получим:

$$H = 0.232 \log 0.232 \dots + 0.0837 \log 0.0837 \approx 2.5013699 \text{ бит}$$

Теперь надо получить среднее число элементарных символов на букву по формуле:

$$k = - \sum_{i=1}^n k_i p_i$$

Где k – это длина кода буквы в дереве.

Получим:

$$k = 2 * 0.232 \dots \dots 3 * 0.0837 = 2.539588 \text{ бит}$$

Теперь получаем информацию на элементарный символ по формуле:

$$I_s = \frac{H}{k}$$

И получим:

$$I_s = \frac{2.501369}{2.539588} = 0,9849 \text{ бит}$$

Таким образом, информация на один символ близка к своему верхнему пределу 1, и, из этого следует что выбранный нами бинарный код весьма близок в оптимальному.

Кодирование Шеннона – Фано является достаточно старым методом сжатия, и на сегодняшний день оно не представляет особого практического интереса.

Список используемых источников:

1. Ватолин Д., Ратушняк А., Смирнов М., Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. Кодирование Шеннона-Фано. 2002.

URL: http://www.compression.ru/download/articles/huff/tiger_shannon-fano.html

2. Кошкин Г.М. Энтропия и информация // Соросовский образовательный журнал. Т.7. №11. 2001. С. 127.

© 2016, Гасангусейнов К.А.

Архивация данных методом Шеннон-Фано

© 2016, Gasanguseinov K.A.

Backup method Shannon-Fano