**Исследование оптимальных кодов. Код Шеннона-Фано. код Хаффмена**

**Методические указания**

к лабораторным работам по дисциплине

«Теория информации»

**1 Общие положения**

Оптимальные коды применяются в тех случаях, когда символы алфавита встречаются в сообщениях с различной вероятностью. В этом случае применение оптимальных кодов позволяет минимизировать избыточность кода, а, следовательно, сократить время передачи сообщений. Код Шеннона-Фано и код Хаффмана относятся к множеству оптимальных кодов. Алгоритм Шеннона-Фано ‒ один из первых алгоритмов сжатия, который впервые сформулировали американские учёные [Шеннон](https://ru.wikipedia.org/wiki/%D0%A8%D0%B5%D0%BD%D0%BD%D0%BE%D0%BD,_%D0%9A%D0%BB%D0%BE%D0%B4_%D0%AD%D0%BB%D0%B2%D1%83%D0%B4) и [Фано](https://ru.wikipedia.org/w/index.php?title=%D0%A4%D0%B0%D0%BD%D0%BE,_%D0%A0%D0%BE%D0%B1%D0%B5%D1%80%D1%82_%D0%9C%D0%B0%D1%80%D0%B8%D0%BE&action=edit&redlink=1" \o "Фано, Роберт Марио (страница отсутствует)). Данный метод сжатия имеет большое сходство с [алгоритмом Хаффмана](https://ru.wikipedia.org/wiki/%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC_%D0%A5%D0%B0%D1%84%D1%84%D0%BC%D0%B0%D0%BD%D0%B0), который появился на несколько лет позже. Алгоритм использует коды переменной длины: символы, имеющие большую вероятность появления, кодируется кодом меньшей длины, символы с малой вероятностью появления ‒ кодом большей длины. Рассмотрим каждый из алгоритмов в отдельности.

**Алгоритм оптимального кодирования Шеннона-Фано:**

1. Все символы алфавита[[1]](#footnote-1) упорядочиваются в порядке убывания вероятности их появления.
2. Кодируемые символы делятся на две равновероятные или приблизительно равновероятные подгруппы.
3. Каждому символу из верхней подгруппы присваивается код «0», а каждому символу из нижней подгруппы ‒ код «1».
4. Каждая из подгрупп снова делится на две равновероятные или приблизительно равновероятные подгруппы. При этом каждому символу из верхней подгруппы присваивается код «0», а из нижней ‒ «1».
5. Деление на подгруппы проводится до тех пор, пока в подгруппе не останется по одному символу.
6. Результирующие кодовые слова записываются слева направо по кодам подгрупп, соответствующих кодируемому символу.

Пример оптимального кодирования по методу Шеннона-Фано приведен в таблице 1.1.

Таблица 1.1 ‒ Кодирование по методу Шеннона-Фано

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Символ алфавита | Вероятность | Шаги алгоритма | | | | | Количество элементарных символов | Кодовое слово |
| 1 | 2 | 3 | 4 | 5 |
|  | 0,30 | 0  0,55 | 0  0,3 |  |  |  | 2 | 00 |
|  | 0,25 | 0 | 1  0,25 |  |  |  | 2 | 01 |
|  | 0,15 | 1 | 0  0,25 | 0  0,15 |  |  | 3 | 100 |
|  | 0,1 | 1 | 0 | 1  0,1 |  |  | 3 | 101 |
|  | 0,1 | 1  0,45 | 1  0,2 | 0  0,1 |  |  | 3 | 110 |
|  | 0,05 | 1 | 1 | 1  0,1 | 0  0,05 |  | 4 | 1110 |
|  | 0,04 | 1 | 1 | 1 | 1  0,05 | 0  0,04 | 5 | 11110 |
|  | 0,01 | 1 | 1 | 1 | 1 | 1  0,01 | 5 | 11111 |

Отметим, что рассмотренная методика построения оптимального кода имеет некоторую неоднозначность для случаев, когда невозможно разбить символы алфавита на подгруппы с равными вероятностями. В таких случаях для одного и того же распределения вероятностей появления символов алфавита могут быть получены коды различной длины. Этой неоднозначности можно избежать, если для построения эффективного кода использовать алгоритм кодирования Хаффмана.

**Алгоритм оптимального кодирования Хаффмана:**

1. Все символы алфавита упорядочиваются в порядке убывания их вероятностей появления.
2. Проводится «укрупнение» символов. Для этого два последних символа «укрупняются» в некоторый вспомогательный символ с вероятностью, которая равняется сумме вероятностей символов, которые были «укрупнены».
3. Образовавшаяся новая последовательность вновь сортируется в порядку убывания вероятностей с учетом вновь образованного за счет «укрупнения» символа.
4. Процедура повторяется до тех пор, пока не получится один «укрупненный» символ, вероятность которого равна 1.

*Замечание*: На практике не используют многократное выписывание символов алфавита и их упорядочивание. Обходятся построением кодового дерева так как это показано в таблице 1.2. Верхние ветви кодового дерева кодируются как 0, нижние ‒ 1. Кодирование каждой буквы проводится путем считывания значений (0 или 1) ответвлений, которые находятся на пути от корня кодового дерева (точки с вероятностью 1), до вершин, соответствующих каждому символу алфавита.

Таблица 1.2 ‒ Кодирование по методу Хаффмана

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Символ алфавита | Вероятность | Кодовое дерево | | | | | | | Количество элементарных символов | Кодовое слово |
|  | 0,30 |  |  | 0 |  | F (0,55) |  |  | 2 | 00 |
|  | 0,25 |  |  | 1  1 |  |  | 0 | G (1) | 2 | 01 |
|  | 0,15 |  | 0 | D (0,25) |  | 0 |  | 1 | 3 | 100 |
|  | 0,1 |  | 1 |  | C (0,2) |  | E (0,45) |  | 3 | 101 |
|  | 0,1 |  | 0  0 |  | 1 | 1 |  |  | 3 | 110 |
|  | 0,05 |  | 0 | 1 | B (0,1) |  |  |  | 4 | 1110 |
|  | 0,04 |  |  | A (0,05) |  |  |  |  | 5 | 11110 |
|  | 0,01 |  | 1 |  |  |  |  |  | 5 | 11111 |

**2 Задание на работу**

1. Выбрать из своей фамилии имени и отчества первые десять букв. В результирующей последовательности каждая буква должна встречаться только один раз.
2. Составить таблицу, состоящую из двух столбцов. В первый столбец должна быть записана буква из последовательности, выбранной в п.1, во втором столбце ‒ соответствующая вероятность появления символа. Вероятность последней, десятой буквы последовательности вычислить с помощью соотношения

.

Это делается для того, чтобы суммарная вероятность для группы из выбранных десяти символов алфавита равнялась 1.

1. Для последовательности символов из п.1. построить оптимальный код Шеннона-Фано.
2. Для последовательности символов из п.1. построить оптимальный код Хаффмана.
3. Рассчитать энтропию источника с помощью соотношения

.

1. Рассчитать среднюю количество элементарных символов на один символ первичного алфавита с помощью соотношения

.

Расчеты выполнить отдельно для метода Шеннона-Фано  и отдельно для метода Хаффмана .

1. Рассчитать недогруженность по методу Шеннона-Фано и по методу Хаффмана, используя соотношения

 и .

1. Рассчитать избыточность, которая показывает сколько процентов букв может быть удалено из текста без утраты смысла сообщения, используя соотношения

 и .

1. Суммарная вероятность появления символов алфавита должна равняться 1 [↑](#footnote-ref-1)