Лекция

Код Хаффмана

- Имеются сообщения, состоящие из последовательности символов. В каждом сообщении символы появляются с известной вероятностью, не зависящей от позиции в сообщении.
- Пусть есть сообщение из символов *a*, *b*, *c*, *d*, *e*, которые появляются в сообщениях с вероятностями 0.12, 0.4, 0.15, 0.08 и 0.25 соответственно.
- Задача: закодировать каждый символ последовательностью 0 и 1 так, чтобы код любого символа являлся префиксом кода сообщения.
- Префиксное свойство позволяет декодировать строку из 0 и 1 путём удаления префиксов (кодов символов) из этой строки.

Два двоичных кода

- Две возможные кодировки:

Символ	Вероятность	Код 1	Код 2
a	0.12	000	000
b	0.40	001	11
c	0.15	010	01
d	0.08	011	001
e	0.25	100	10

- Первый и второй коды обладают префиксным свойством: любая префиксная последовательность однозначно идентифицируется символом.

Два двоичных кода

- Алгоритм декодирования для Кода 1:

1. Взять три бита
2. Преобразовать их в символ
3. Отбросить их. Перейти к шагу 1.

001010011 – bcd.

- Алгоритм декодирования для Кода 2:

1. Определить последовательность битов, однозначно кодирующих символ.
2. Преобразовать их в символ
3. Отбросить их. Перейти к шагу 1.

1101001 – bcd.

Задача

- Задача конструирования кодов Хаффмана: имея множество символов и значения вероятностей их появления в сообщениях, построить такой код с префиксным свойством, чтобы средняя длина кода (в вероятностном смысле) была минимальной.
- Т.е. требуется минимизировать среднюю длину кода, чтобы уменьшить длину сообщения (т.е. сжать сообщение).
- Чем короче длина кода символов, тем короче закодированное сообщение. Таким образом, символы с большими вероятностями появления должны иметь более короткие коды.

Алгоритм Хаффмана

- Алгоритм Хаффмана – способ нахождения оптимального префиксного кода. В этом алгоритме находятся 2 символа *a* и *b* с наименьшими вероятностями появления и заменяются одним фиктивным символом *x*, который имеет вероятность появления, равную сумме вероятностей появления символов *a* и *b*. Затем, использую процедуру рекурсивно находится оптимальный префиксный код для меньшего множества символов (где символы *a* и *b* заменены одним символом *x*).

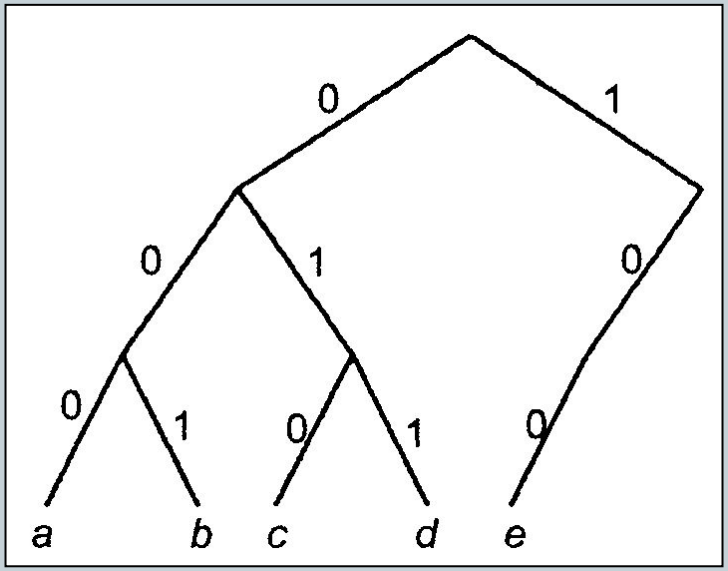
Алгоритм Хаффмана

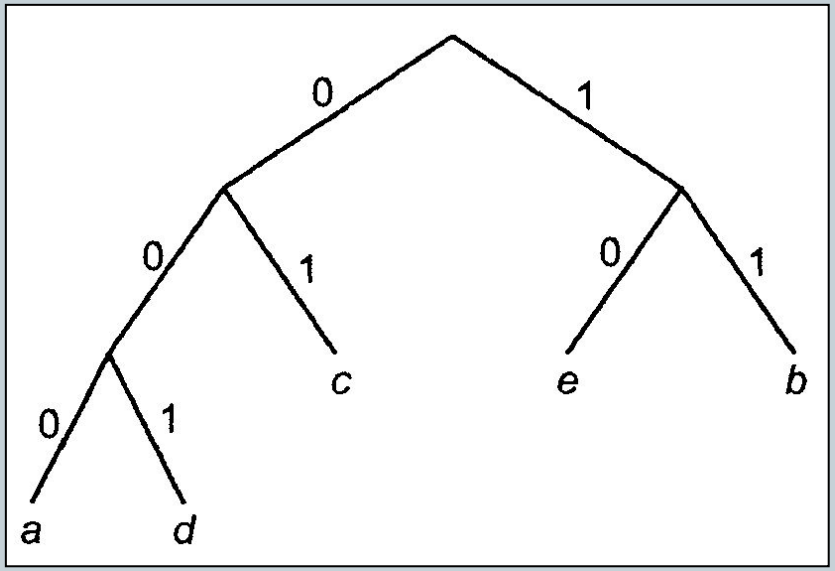
- Код для исходного множества символов получается из кодов замещающих символов путем добавления 0 и 1 перед кодом замещающего символа, и эти два новых кода принимаются как коды заменяемых символов. Например, код символа *a* будет соответствовать коду символа *x* с добавленным нулем перед этим кодом, а для кода символа *b* перед кодом символа *x* будет добавлена единица.

Пути на двоичном дереве

- Можно рассматривать префиксные коды как пути на двоичном дереве: прохождение от узла к его левому сыну соответствует 0 в коде, а к правому сыну — 1. Если мы пометим листья дерева кодируемыми символами, то получим представление префиксного кода в виде двоичного дерева.
- Префиксное свойство гарантирует, что нет символов, которые были бы метками внутренних узлов дерева (не листьев), и наоборот, помечая кодируемыми символами только листья дерева, мы обеспечиваем префиксное свойство кода этих символов.

Двоичные деревья для кодов 1 и 2





Реализация алгоритма Хаффмана

- Для реализации алгоритма Хаффмана используется лес – деревья, чьи листья помечены кодируемыми символами.
- Важный этап работы алгоритма – выбор из леса двух деревьев с наименьшими весами. Эти два дерева комбинируются в одно с весом, равным сумме весов составляющих деревьев. При слиянии деревьев создается новый узел, который становится корнем объединённого дерева и который имеет в качестве левого и правого сыновей корни старых деревьев. Этот процесс продолжается до тех пор, пока не получится только одно дерево.
- Это дерево соответствует коду, имеющему минимальную длину.

Реализация алгоритма Хаффмана