Data Compression

Andrew C. Lee

EECS, Syracuse

Contents

Background and Motivations

Compression of Data Files

1. The frequency of occurrence of each ascii symbol is not the same
2. Estimate the probability of occurrence of each symbol
3. Minimize the average length of the codeword

Prefix Codes

Code Design

1. Each codeword corresponds to exactly one symbol
2. Decoding is simple (eg. no look ahead is required)

Question What property you wish the codewords to have ?
No codewords is a prefix of another codewords make decoding easier
Question What kinds of tree structure may help to construct the code ?

Huffman Trees I

Use a *greedy* approach:

FIGURE 11.1 Two Huffman trees created for five letters A, B, C, D, and E with probabilities .39, .21, .19, .12, and .09.

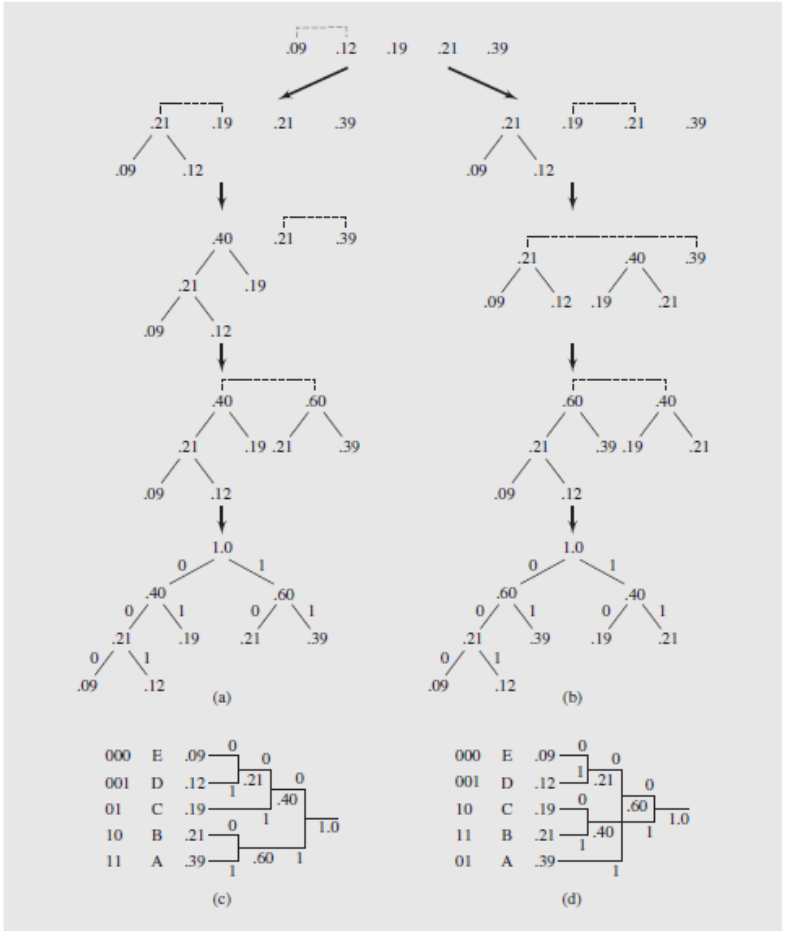

Figure : Building Huffman Trees

Huffman Trees II

FIGURE 11.2 Two Huffman trees generated for letters P, Q, R, S, and T with probabilities .1, .1, .1, .2, and .5.

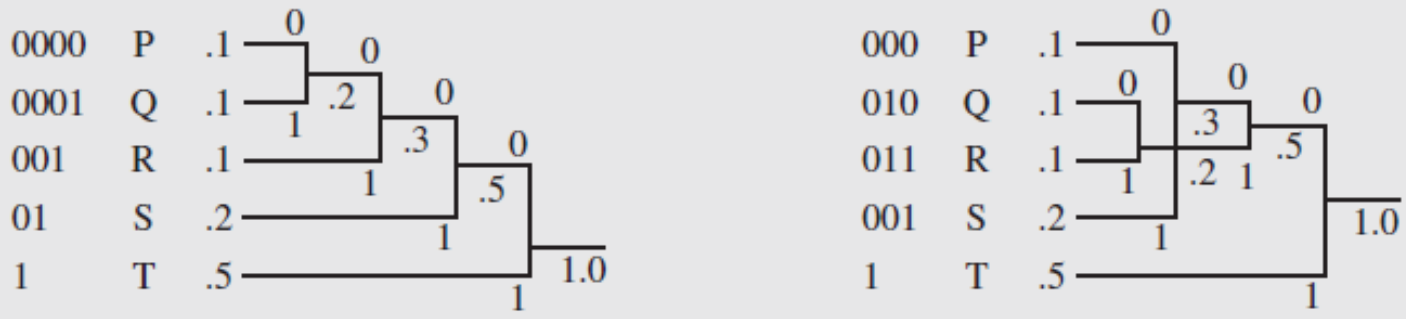

Figure : Two Huffman Trees

Question: Which one will you use ? Why ?

Huffman Algorithm via a doubly linked lists

FIGURE 11.3 Using a doubly linked list to create the Huffman tree for the letters from Figure 11.1.

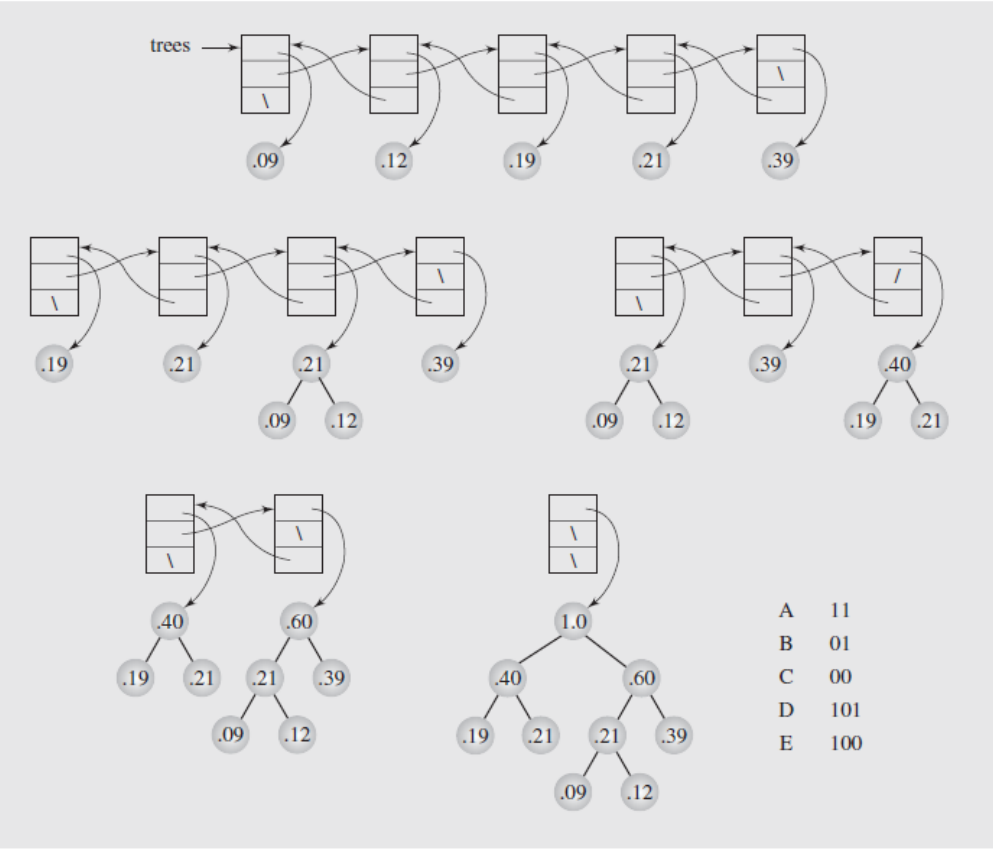

Figure : Using a doubly linked list in Huffman algorithm

Question What is the purpose of the doubly linked list ?

Huffman Algorithm via a heap

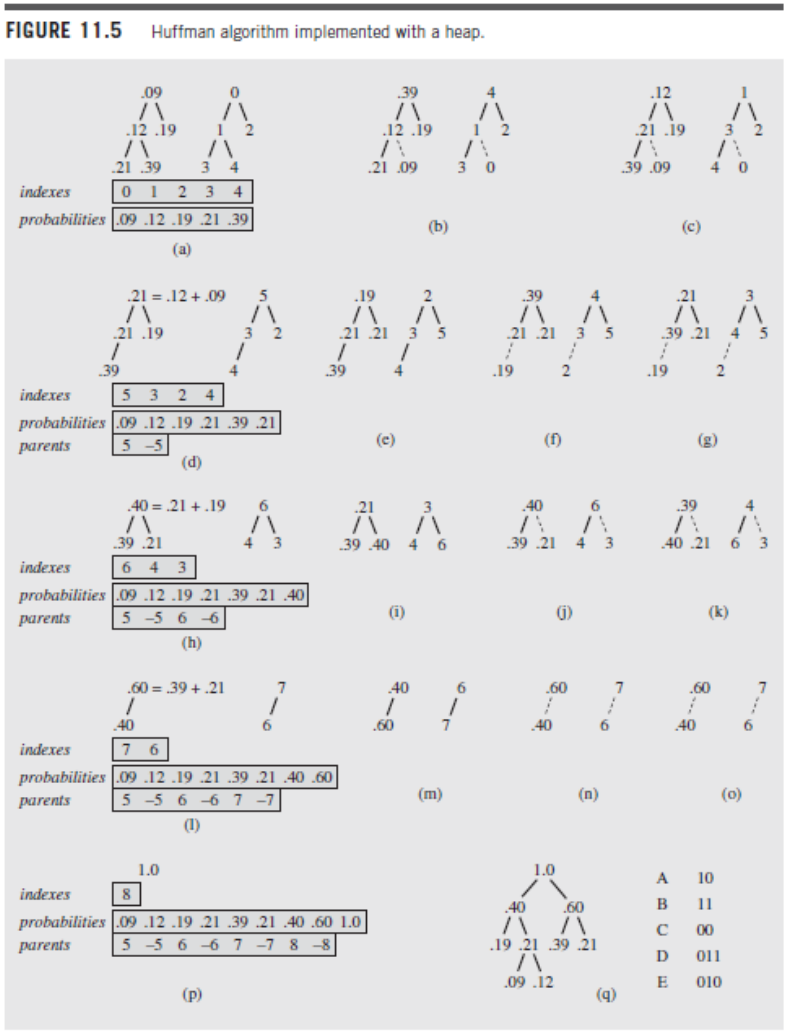

Figure : Using a heap in Huffman algorithm

Adaptive Huffman coding I

FIGURE 11.7 Doubly-linked list nodes formed by breadth-first right-to-left tree traversal.

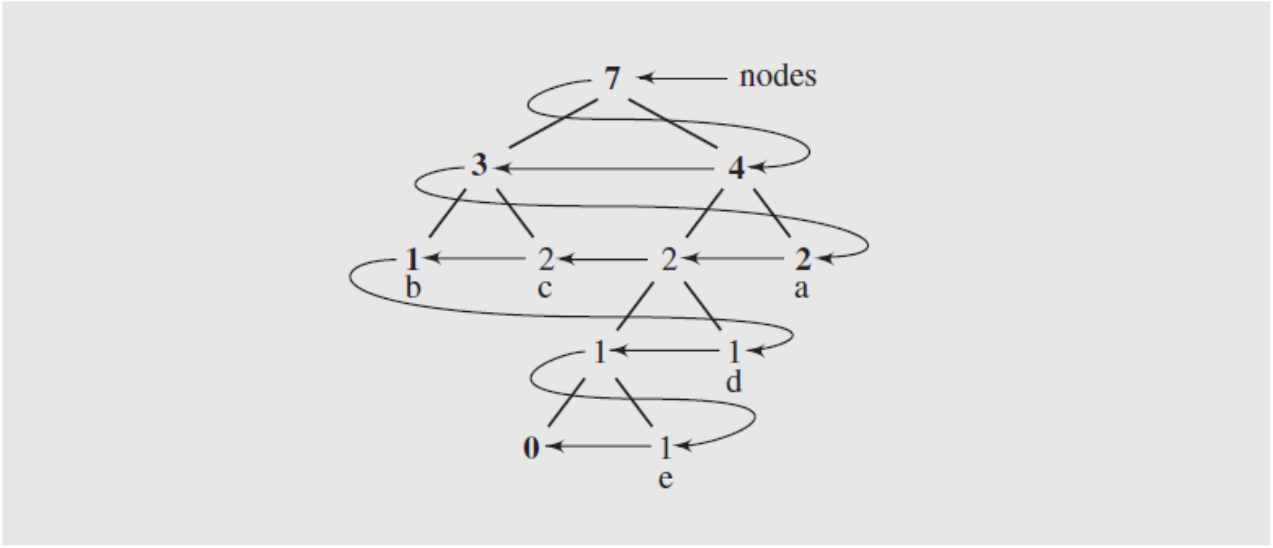

Figure : A linked structure for supporting adaptive Huffman coding

Adaptive Huffman coding II

FIGURE 11.8 Transmitting the message "aafcccbd" using an adaptive Huffman algorithm.

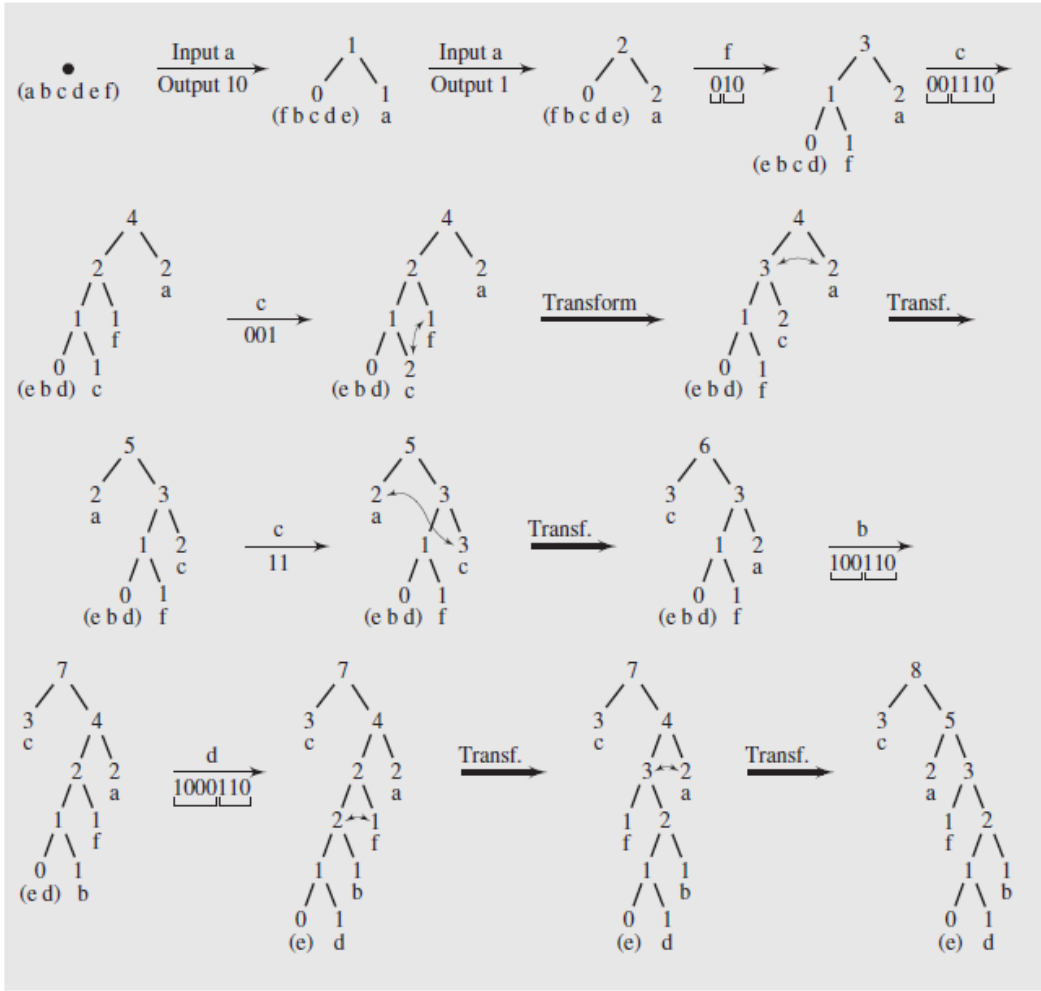

Figure : Adaptive Huffman coding procedure: An example

Run Length Encoding Methods

1. A run is defined as a sequence of identical characters. For example, the string s = “aaabba” has three runs: a run of three “a”s followed by runs of two “b”s and one “a.” The run-length encoding technique takes advantage of the presence of runs and represents them in an abbreviated, compressed form.
2. Ziv Lempel Algorithms (eg. LZ77, LZW)

LZ77

FIGURE 11.9 Encoding the string "aababacbaacbaadaaa . . ." with LZ77.

Input	Buffer	Code Transmitted
aababacbaacbaadaa . . .	aaaa	a
aababacbaacbaadaa . . .	aaaaaaba	22b
abacbaacbaadaaa . . .	aaababac	23c
baacbaadaaa . . .	abacbaac	12a
cbaadaaa . . .	cbaacbaa	03a
daaa . . .	cbaadaaa	30d
aaa . . .	. . .	

Figure : Example: LZ77

LZW

FIGURE 11.10 LZW applied to the string "aababacbaacbaadaaa . . . ."

Encoder		Table		
Input	Output	Index (Codeword)	Full String	Abbreviated String
		1	a	a
		2	b	b
		3	c	c
a		4	d	d
a	1	5	aa	1a
b	1	6	ab	1b
ab	2	7	ba	2a
a	6	8	aba	6a
c	1	9	ac	1c
ba	3	10	cb	3b
ac	7	11	baa	7a
baa	9	12	acb	9b
d	11	13	baad	11d
aa	4	14	da	4a
a	5	15	aaa	5a
	...			

Figure : Illustration: LZW algorithm