Huffman Codes

Chapter 16

Data Compression

- Huffman codes are used for data compression. The motivations for data compression are obvious: reducing time to transmit large files, and reducing the space required to store them on disk or tape.
- The code was devised by Huffman as part of a course assignment at MIT in the early 1950s.

Goal

- Huffman coding is a technique for assigning binary sequences to elements of an alphabet.
- The goal of an optimal code is to assign the minimum number of bits to each symbol (letter) in the alphabet.

File Size

- Suppose that you have a file of 100K characters.
- To keep the example simple, suppose that each character is one of the 8 letters from a through h.
- How much space is required to store this file?

File Size

- Can we do better?
- Suppose that we have more information about the file: the *frequency with which each* character appears.
- Use a variable length code instead of a fixed length code
- Use fewer bits to store common characters

Example

	A	B	C	D	E	F	G	H
Frequency	45K	13K	12K	16K	9K	5K	0K	0K
Fixed length code	000	001	010	011	100	101	110	11
Variable length code								

•How much space will the file stored using variable-length code take?

•What is the saving or compression ration?

Decoding

- How can 110001001101 be decoded using the variable length code from the previous slide?

Huffman Trees

- We can represent the decoding algorithm by a binary tree, where each edge represents either 0 or 1, and each leaf corresponds to the sequence of 0s and 1s traversed to reach it, ie a particular code.

Fixed Length Code Tree

Variable Length Code Tree

HUFFMAN(C)

- C is a set of n characters and each c in C has a frequency

Example

- Find the optimal code for the following

Character	Freq	Fixed Code	Bits	Variable code	Bits
A	10	000	30		
E	15	001	45		
I	12	010	36		
S	3	011	9		
T	4	100	12		
SP	13	101	39		
NL	1	110	3		

Entropy

- Entropy is a measure of the amount of uncertainty or randomness associated with a random variable; that is, it is a measure of the amount of information on the average required to describe the variable.
- In compression, entropy is a measure of how much information is actually in the text being compressed

$$-\sum_{x_i} p(x)\log_2 p(x)$$

Example

- What is the entropy for the previous example?