Albert Chieu

CSC 220

7/28/11

Huffman Coding Discussion

**What is the problem?**

Use Huffman code to encode a string of English words (Chapter 1 of *A Tale of Two Cities* by Charles Dickens) and calculate the efficiency or compression ratio that the coding of this given string can achieve.

**What is Huffman Coding?**

Huffman Coding is an algorithm to define a binary code with optimal code length based on the frequency of characters appearing in the text input.

**How does Huffman Coding work?**

Huffman Coding is based on the Huffman tree or trie. The trie is built based on the frequency of characters. All characters are defined as leaf nodes, the smaller the frequency, the lower the level in the trie. The final code is determined by concatenating all codes while traversing the trie down to the leaf nodes that holds the character. Traversing to the left will lead to a code of 0 while traversing to the right will lead to a code of 1.

**How is the program implemented?**

An input string is read from a file. A hashtable is used to keep track of all the character count for each character. The all character nodes that hold the character and its frequency will be put into an array to build the trie using Huffman’s coding algorithm.

Starting with the initially array of nodes, the following steps will be applied to the array:

1. Sort the array in increasing order according to frequency value.
2. Take the first two lowest frequency nodes as children of a new node that holds the sum of their frequencies.
3. Remove the two lowest frequency nodes from the array and add the new parent node to the array.
4. Sort the array in increasing order according to frequency value and repeat step 2 until the array size reduces to 1 node.

**Generally how well does it do?**

The compression ratio between the total message bits and the total encoded bits is 1.8365135453474677. This shows that Huffman coding saves approximately 50% space during file compression. The resulting Huffman code is not unique for each character. However, the output code length should be the same number of bits.