Huffman coding

Johan Montelius

October 25, 2013

Getting started

In this seminar session we will look at different ways to represent data, using lists, trees and tuples to find the best representation. The *best representation* could of course mean many things, we might need a representation that gives us efficient code or we might want a representation that is easy to explain, implement and maintain. We will start by using quite simple representations and then refine them to gain better performance.

To have something to work with we will implement the *Huffman* encoding and decoding functions. You should do some reading on Huffman coding, this text will not explain the algorithm but how to implement it.

1 Huffman overview

Huffman coding can divided into two parts, one part is how to construct the coding table and the other, much simpler, is how to encode or decode a text using the table.

The idea behind Huffman coding is of course to encode frequent characters with few bits and infrequent characters with more bits. To keep things simple we will represent sequences of bits as lists of zeros and ones but this could of course be changes if we intend to do a real implementation that reads and writes to files. For our experiments it is sufficient.

The table should give a one to one maping from characters to codes but we might use one representation when we encode text and another when we decode text; the information it holds is the same but we might want to do this for efficiency.

Once we are done we will have a module that defines the following functions:

- table(Sample): create a table containing the maping from characters to codes
- encode(Text, Table): encode the Text using the maping in Table, return a sequence of bits
- decode(Sequence, Table): decode the bit Sequence using the maping in Table, return a Text

Start by defining the module, some compile directives, things that are good to have and dummy code for the functions.

```
-module(huffman).

-compile(export_all).

sample() -> "the quick brown fox jumps over the lazy dog
     this is a sample text that we will use when we build
     up a table we will only handle lower case letters and
     no punctuation symbols the frequency will of course not
     represent english but it is probably not that far off".

text() -> "this is somthing that we should encode".

test() ->
    Sample = sample(),
    Table = table(Sample),
    Text = text(),
    Seq = encode(Text, Table),
    Text = decode(Seq, Table).

table(_Sample) -> na.

encode(_Text, _Table) -> na.

decode(_Seq, _Table) -> na.
```

2 the table

In order to create the table we need first to find out the frequency distribution in our sample text. Once we have the frequency distribution we can start building a Huffman tree and using the tree we will extract the codes.

```
table(Sample) -> Freq = freq(Sample),
                 Tree = huffman(Freq),
                 codes(Tree).
```

The sample is of course a list of characters (`[102,111,111]` or `[$f,$o,$o]` as we can write in Erlang), you should run through this list and collect the frequencies of the characters. If "foo" was the sample text we should have the frequencies *f/1*, *o/2*. How would you represent this information? Note that you need not know beforehand which characters that will occur in the sample.

You will probably end up with a structure that looks like this, but how you represent the frequencies is up to you.

```
freq(Sample) -> freq(Sample, ...).

freq([], Freq) ->
      Freq;
freq([Char|Rest], Freq) ->
      freq(Rest, ...).
```

Now once we have the frequencies we will create a Huffman tree. This is simpler than you might think but before you read further you must understand why we create a tree and what properties it should have. If you started to read this with out understanding how Huffman coding works this is the time to stop reading.

2.1 the Huffman tree

OK, so a Huffman tree is a tree with the characters in the leafs but the low frequency characters have long branches and high frequency characters have short branches. Assume we represent a leaf with the tuple {`leaf, Char, Freq`} and a node by {`node, Freq, Left, Right`}. The frequency in the node is the combined frequency of the left and right tree.

If you turn your table into an ordered sequence of leafs where each leaf represents a character and its frequency. The "foo" example above would correspond to the sequence `[{$f, 1},{$o, 2}]`.

Now, assuming we have such a sequence, what would happen if we took the two smallest elements and combined them into a new node and added the node to the remaining sequence. Can we repeat this process, what will the final result be?

How do we represent the sequence so that it is easy to find the two smallest elements? How do we keep this representation?

Implement something that works and then look at this: what happens if we represents a leaf by the tuple {`node, Char, Freq, na, na`} and a node by {`node, na, Freq, Left, Right`}, would things be easier?

2.2 the Huffman codes

I assume now that you have a tree representation of the table and it is time to find the codes. The codes are of course hidden in the tree in the branches and the code of a character is the path to the leaf holding the character (*left*, *left*, *right*, *left* or *0,0,1,0*).

Traverse the tree, and collect the characters in the leafs. Keep track of the path to the leaf and record this path as a sequence of zeros and ones.

When you're done you should have something like `[{$f, [ 1,1,0]}, {$r, [1,0,1,0]}`, or whatever the tree looks like.

Start by writing a function that only collects the characters, once this is mastered you can start to keep track of the path.

2.3 half way

Half-way there might be an exaggeration but at lest you're now done with the first part, you have a maping from characters to Huffman codes. It's represented by a list of tuples `{$f, [1,1,0]}`, one for each characters. Time to use this table in the encoding and decoding.

3 Huffman encoding

This is simple, we have a text represented as a list of characters and for each character we have a sequence of bits found in the table. You could probably create something very simple that works

If you manage to implement the encoder you should be able to turn the text "this is somethin..." into a list of bits `{[1,1,0,1,0,1,0, ...]`.

Stop here an ponder what the time complexity is. You will of course have a linear factor, depending on the length of the text, since the text is encoded character by character but you might have other factors. What is the the time complexity of looking up a character in the table? What is the complexity of producing the final sequence of bits?

I'm quite sure that your original code is open for improvements but let's leave it for now.

4 Huffman decoding

Decoding is slightly more tricky since we do not know exactly how many bits are used to code each character. If we have a sequence `[1,1,0,1,0,1,0, ...]` it could be that the first four bits is a t and the following three is an i but we do not know; what we do know is that thanks to Huffman it is only possible to decode it in one way given a table with Huffman codes.

We could of course assume that one bit is used in the coding and then search the table for a character with this pattern (`{[1]`). If a character is found, problem solved, if not, look for a character using two bits (`{[1,1]`). Let's implement this function first and see what we have, it will work and will probably not take that much time.

Your solution might look something like this:

```
decode([], _Table) ->
    [];
```

```
decode(Seq, Table) ->
     {Char, Rest} = decode_char(Seq, 1, Table),
     decode(Rest, Table).

decode_char(Seq, N, Table) ->
     {Code, Rest} = lists:split(N, Seq),
     case lists:keysearch(Code, 2, Table) of
       ... ->
             ...;
       false ->
             decode_char(..., ..., Table)
      end.
```

I'm using some functions from the *lists library*, you should look these up and understand how they work. If you fill in the blanks you're up and running, you have your first Huffman encoder/decoder.

Make sure that it works correctly, by testing the smaller functions and make sure that they behave correctly, then work your way up. Always test the corner cases i.e. when lists are empty etc before trying more complicated tasks.

5 Performance

Let's now run some performance tests. You need to find a large text that you can use to benchmark the program. You would also need to find a sample text of the given language that gives you the correct frequencies but we can cheat and use the text it self to do the frequency analysis.

Do some benchmark of your system and determine how well it performs, try to estimate the time to encode or decode a text given the length of the text. Also do some experiments where you change the size of the alphabet, for example using only eight characters, the regular alphabet, all ascii characters, or something even larger.

You could generate a list of codes by interpreting a string of 8-bit characters as 16-bit values. Try the code below:

```
unicode:characters_to_list(<<"this must be an even number.">>).
```

These codes are of course only faked UTF16 codes but it could serve the purpose as a benchmark for a larger character set.

How does your implementation scale with larger texts and larger alphabets?