Lempel-Ziv Welch (LZW) Compression

We will now look at the Lempel-Ziv Welch compression algorithm, which is a lossless compression algorithm that does particularly well on data with repetitions.

A useful feature of LZW compression is that the dictionary is built adaptively during encoding. The dictionary does not need to be passed with the compressed text — the decoding algorithm produces the same dictionary from the compressed text.

The original paper that describes the LZW algorithm is:


This paper describes an improvement to a compression method introduced by Ziv and Lempel in 1977 and 1978.

LZW and variants have been used in popular software such as Unix compress and GIF compression.
Sources

There is a lot of information about LZW on the web. See, for example, Wikipedia, or the nice animation at


Also, see Dave Marshall’s notes.

http://www.cs.cf.ac.uk/Dave/Multimedia

Compression

The dictionary is initialised so that there is a codeword for every extended ASCII character.

<table>
<thead>
<tr>
<th>character</th>
<th>code word</th>
</tr>
</thead>
<tbody>
<tr>
<td>null</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>A</td>
<td>65</td>
</tr>
<tr>
<td>B</td>
<td>66</td>
</tr>
<tr>
<td>C</td>
<td>67</td>
</tr>
<tr>
<td>D</td>
<td>68</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>□</td>
<td>255</td>
</tr>
</tbody>
</table>

The compression algorithm

```
Initialise dictionary
w ← NIL.
while there is a character to read
  k ← next character in text
  if wk is in the dictionary
    w ← wk
  else
    Add wk to the dictionary
    Output the code for w
    w ← k
Output the code for w
```

Example

Let $T = ABACABA$

<table>
<thead>
<tr>
<th>w</th>
<th>k</th>
<th>text</th>
<th>dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIL</td>
<td>A</td>
<td>BACABA</td>
<td>A 65</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>ACABA</td>
<td>AB 256 65</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>CABA</td>
<td>BA 257 66</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>ABA</td>
<td>AC 258 65</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>BA</td>
<td>CA 259 67</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>AB</td>
<td>A</td>
<td>ABA</td>
<td>ABA 260 256</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>ABA 260 65</td>
</tr>
</tbody>
</table>

Initialise dictionary
w ← NIL.
while there is a character to read
  k ← next character in text
  if wk is in the dictionary
    w ← wk
  else
    Add wk to dictionary
    Output code for w
    w ← k
Output the code for w
The Decompression Algorithm

The basic decompression algorithm is as follows.

Initialise dictionary
\( c \leftarrow \text{first codeword} \)
output the translation of \( c \)
\( w \leftarrow c \)

While there is a codeword to read
\( c \leftarrow \text{next codeword} \)
output the translation of \( c \)
\( s \leftarrow \text{translation of } w \)
\( k \leftarrow \text{first character of translation of } c \)
Add \( sk \) to dictionary
\( w \leftarrow c \)

Example

Code = 65, 66, 65, 67, 256, 65

<table>
<thead>
<tr>
<th>( c )</th>
<th>( w )</th>
<th>( s )</th>
<th>( k )</th>
<th>( \ldots )</th>
<th>( \text{output} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>65</td>
<td>A</td>
<td>B</td>
<td>\ldots</td>
<td>A</td>
</tr>
<tr>
<td>66</td>
<td>66</td>
<td>B</td>
<td>A</td>
<td>AB</td>
<td>256</td>
</tr>
<tr>
<td>65</td>
<td>65</td>
<td>A</td>
<td>C</td>
<td>BA</td>
<td>257</td>
</tr>
<tr>
<td>67</td>
<td>67</td>
<td>C</td>
<td>A</td>
<td>AC</td>
<td>258</td>
</tr>
<tr>
<td>256</td>
<td>256</td>
<td>A</td>
<td>AB</td>
<td>ABA</td>
<td>259</td>
</tr>
<tr>
<td>65</td>
<td>65</td>
<td>AB</td>
<td>A</td>
<td>ABA</td>
<td>260</td>
</tr>
</tbody>
</table>

Initialise dictionary
\( c \leftarrow \text{first codeword} \)
output the translation of \( c \)
\( w \leftarrow c \)

While there is a codeword to read
\( c \leftarrow \text{next codeword} \)
output the translation of \( c \)
\( s \leftarrow \text{translation of } w \)
\( k \leftarrow \text{first character of translation of } c \)
Add \( sk \) to dictionary
\( w \leftarrow c \)

Reﬁnement

The decoding algorithm as stated does not always work as it fails if \( c \) is not in the dictionary.

Example

Initialise dictionary
\( w \leftarrow \text{NIL} \)
while there is a char to read
\( k \leftarrow \text{next char in text} \)
If \( wk \) is in the dictionary
\( w \leftarrow wk \)
Else
Add \( wk \) to dictionary
Output code for \( w \)
\( w \leftarrow k \)
Output the code for \( w \)
The problem arises when the dictionary has $cs$ in the dictionary for a character $c$ and a string $s$ and then the input contains $cscsc$.

The decompression algorithm can be modified to deal with this case.
There are lots of interesting implementation issues. For example, what if the dictionary runs out of space?

Also, if we start re-using dictionary space, what data structure do we use to make dictionary access efficient?

GIF compression solves the problem of dictionary overflow by having variable-length codes. We will not cover the details.