Compressing multisets using tries

Vincent Gripon, Michael Rabbat, Vitaly Skachek and Warren J. Gross (Télécom Bretagne, Brest - McGill University, Montréal)

1. Motivation

How to efficiently encode binary words produced by the source while disregarding order?

2. About order and sequences

2.1. Toy example

Encode sequence: 

2.1.1. With order

2.1.2. Without order

Cannot do better!

2.2. Previous work ([Varshney & Goyal], [Reznik])

Multiset \( M = \{ M_1, M_2, \ldots M_m \} \) obtained by drawing \( m \) elements of \( n \) bits with repetition (contains at most \( m \) distinct elements).

2.2.1. \( 2^m = o(m) \)

Encode histogram of cardinalities: 

\[ \#(M) \]

2.2.2. \( m = o(2^n) \)

Ignore repetitions \( \Rightarrow \) elements are drawn uniformly.

Problem: under these conditions, asymptotic entropies of sequences with or without order are the same.

\( \Rightarrow \) disregarding order has a negligible impact on compression.

2.3. Our contribution

We derive a lower bound and present an algorithm that asymptotically achieves this bound within a constant factor for the regime where 

\[ c \geq 2 \text{ and fixed} \]

3. Lower bound

As a consequence of the Kraft inequality, the expected minimum number of bits to encode such a multiset is its entropy:

\[ H(M) \sim mn(n - \log_2(m)) \sim \frac{2^m \log_2(e)}{c} . \]

4. An algorithm based on tries

4.1. Tries


5. Encoding & Decoding

5.1. Encoding

1. List all branches in lexicographic order.

\[
\begin{array}{cccccccc}
0000001 & 1000001 & 0100001 & 1010000 & 1001000 & 1100000 & 1100010 & 1100100 \\
\end{array}
\]

2. Remove duplicate consecutive prefixes.

3. Replace all 01 with 010.

4. Add 01 at the ends.

5. Add as many 0 as the degree of the branch (0 for degree 1).

6. Concatenate.

5.2. Decoding

0 \[ \begin{array}{cccccccc}
0000001 & 1000001 & 0100001 & 1010000 & 1001000 & 1100000 & 1100010 & 1100100 \\
\end{array} \]

1. Split after maximal occurrences of \( (01)^2 = 0 \). \text{ \text{\textbackslash n}}

2. Count ending 0’s and make it the degree (degree 1 for 0).

3. Remove last 01 for each branch.

4. Replace maximum (0101’ \text{ by } 01’).

5. Duplicate missing prefixes.

5.3. Remarks

\text{\textbullet} \text{ \text{Lossless encoding. \textbullet \text{ Limited complexity: \textbullet \text{ Encoding: } O(m(n + \log(m))) \textbullet \text{ Decoding: } O(mn).}}

6. Performance & Conclusions

6.1. Asymptotic analysis

Let \( L \) be the expected length of the encoding of a multiset using the trie technique.

6.1.2. Comparison to encoding ordered sequence

Let \( H(M) \) be the entropy of the ordered sequence obtained using the source \( (H(M) = mn) \), then, for \( c \) fixed:

\[ \frac{L}{H(M)} \leq \frac{10c}{3n} + \frac{4c^2}{3n^2} + \frac{c}{3n^2} + \frac{2c}{3n^2} + 0 . \]

6.1.3. Comparison to the lower bound

Let \( \ell \) be a non-negative integer, and \( c = 2^\ell \). Then for any \( \epsilon > 0 \), there exists a positive integer \( n_0 \) such that for any \( n \geq n_0 \) we have

\[ \frac{L}{H(M)} \leq \frac{L}{H(M)} \leq \frac{2}{3} \left( \frac{1}{\log_2(e)} + c(1 - e^{-\frac{1}{2}}) \right) + \epsilon \to \frac{5}{3} + \epsilon . \]

6.2. Simulations

6.3. Conclusions

\text{\textbullet \text{ Fast algorithm to encode multisets. \textbullet \text{ At a constant factor of } 5/3 \text{ from the lower bound.}}}

6.4. Open questions

\text{\textbullet \text{ What if the source is not uniform? \textbullet \text{ What about non-Bernoulli sources? \textbullet \text{ Can the encoded version be efficiently used to apply set operations on multisets (union, intersection, . . . )? \textbullet \text{ How to get closer to the lower bound?}}}}

References

