Information Theory over Multisets

Cosmin Bonchiș and Cornel Izbașa
Research Institute "e-Austria" Timișoara, Romania,
{cosmin, cornel}@ieat.ro

Gabriel Ciobanu
"A.I.Cuza" University, Faculty of Computer Science and
Romanian Academy, Institute of Computer Science
gabriel@info.uaic.ro

June 25, 2007

Outline

Introduction

Multiset source entropy

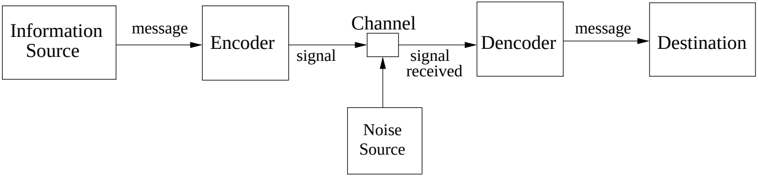
Information content

Encoding length

Channel Capacity

Short review of Shannon information theory

- Communication model [1]



Short review of Shannon information theory

- Established the fundamental natural limits on communication
- **Source entropy** [1]

$$H(X) = \sum_i P_i H_i = -\sum_{i,j} P_i p_i(j) \log p_i(j) \tag{1}$$

- **Channel capacity** [1] The *capacity* C of a discrete channel is given by

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

where $N(T)$ is the number of allowed signals of duration T.

Multiset source entropy

Consider a discrete information source which produces multiset messages:

- A message is a multiset of symbols.
- A multiset is a string equivalence class.
- The entropy rate of such a source is proved to be zero in [2]:

$$H(X_{multiset}) = \lim_{n\to\infty} \frac{1}{n} H(\{X_i\}_{i=1}^{n}) = 0$$

Information content of a multiset

The *information content* of an outcome (multiset) x is

$$h(x) = \log \frac{1}{P(x)} = \log \frac{\prod_{i=1}^{n} m_i!}{(\sum_{i=1}^{n} m_i)! \prod_{i=1}^{n} p_i^{m_i}}$$

Definition according to [3].

Proof.

$$h(x = x_1^{m_1} x_2^{m_2} \ldots x_n^{m_n}) = \log \frac{1}{P[x]} =$$

$$\log \frac{1}{\binom{k}{m_1, m_2, \ldots, m_n} \prod_{i=1}^{n} p_i^{m_i}} =$$

$$\log \left(1 / \frac{(\sum_{i=1}^{n} m_i)!}{\prod_{i=1}^{n} m_i!} \prod_{i=1}^{n} p_i^{m_i} \right) =$$

$$\log \frac{\prod_{i=1}^{n} m_i!}{(\sum_{i=1}^{n} m_i)! \prod_{i=1}^{n} p_i^{m_i}}$$

Encoding length

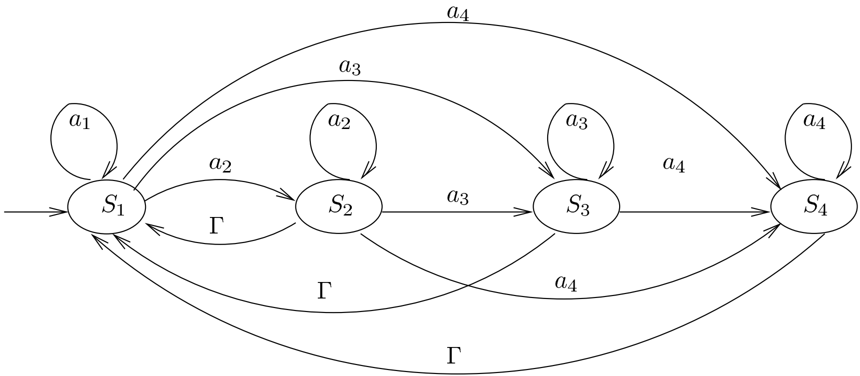
We consider a set X of N symbols, an alphabet A, and the length of encoding l, therefore:
$X = \{x_i = a_1^{n_1} a_2^{n_2} \dots a_b^{n_b} \mid \sum_{j=1}^{b} n_j = l,\ a_j \in A\}$

Theorem

Non-uniform encodings over multisets are shorter than uniform encodings over multisets.

Channel Capacity in base 4

We consider that a sequence of multisets is transmitted along the channel. The capacity of such a channel is computed for base 4, then some properties of it for any base are presented.



According to Shannons' Capacity Theorem we get $b_{ij}^{(a_k)} = t_k$ because we consider that the duration to produce a_k is the same for each $(i, j) \in E$. The determinant equation is

Channel Capacity

Theorem

The multiset channel capacity is zero, $C = 0$.

$$\begin{vmatrix} W^{-t_1}-1 & W^{-t_2} & W^{-t_3} & \cdots & W^{-t_b} \\ 0 & W^{-t_2}-1 & W^{-t_3} & \cdots & W^{-t_b} \\ 0 & 0 & W^{-t_3}-1 & \cdots & W^{-t_b} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & W^{-t_{b-1}}-1 & W^{-t_b} \\ 0 & 0 & 0 & \cdots & W^{-t_b}-1 \end{vmatrix} = 0$$

$$W = \frac{1}{\sqrt[t]{x}} \quad \Rightarrow \quad C = -\frac{1}{t}\log_b x. \tag{2}$$

$(1-x)^b = 0 \Rightarrow W = 1 \Rightarrow C = 0.$

Conclusion

- we derive a formula for the information content of a multiset
- as future work:
  - further explore the properties of multiset-based communication systems
  - compare these to similar results for string-based communication systems

References

- C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal* vol.27, pp.379-423 and pp.623-656, 1948.
- L. R. Varshney, V. K. Goyal, *Toward a Source Coding Theory for Sets*, in Proceedings of the Data Compression Conference (DCC 2006), Snowbird, Utah, 28-30 March 2006.
- David MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, England 2003.