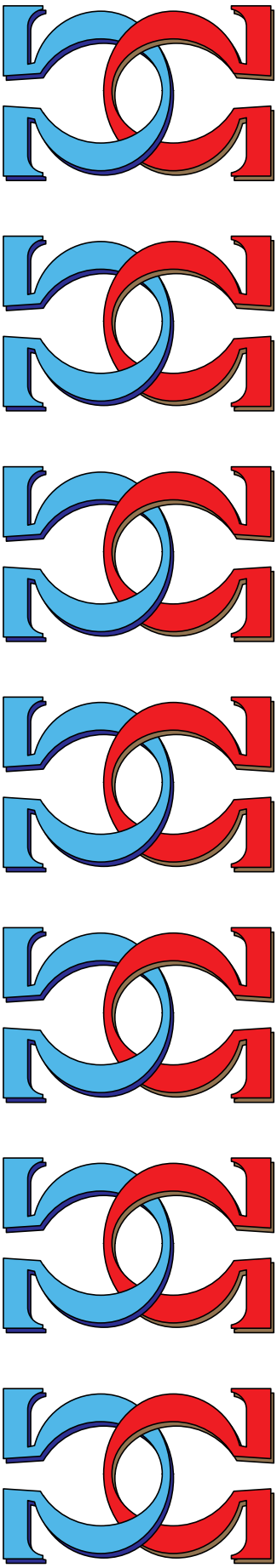

CDMTCS
Research
Report
Series

An Entropy Measure for Finite Strings based on the Shannon Entropy of a Code Set

Ulrich Günther
Department of Computer Science
University of Auckland
Auckland, New Zealand

CDMTCS-204
December 2002

Centre for Discrete Mathematics and
Theoretical Computer Science

An Improved T-Decomposition Algorithm

An Entropy Measure for Finite Strings based on
the Shannon Entropy of a Code Set

December 5, 2002

Abstract

Nicolescu and Titchener showed that every finite string could be used to generate a unique recursively constructed variable-length code set belonging to the family of the T-codes. This paper proposes a conventional entropy function for the string that is defined as the Shannon entropy of the codeset generated by the string.

1 Introduction

Consider a set T of discrete times t, a finite alphabet $S = \{a_1, \ldots, a_n\}$ with cardinality $\#S = n$, and an information source that emits a symbol $a(t) \in S$ at each time t. The probability of $a(t) = a_i$ is denoted as $P_S(i,t)$. The Shannon entropy [1, 2] of the source at time t is then given by

$$H_S(t) = -\sum_i P_S(i,t) \log_2 P_S(i,t). \tag{1}$$

More generally, the source may be seen as a system and the a_i may be regarded as discrete states, in which the system may be found at time t.
A source such as the one described above may alternatively be decoded using a prefix-free and complete code $C \subset S^+$. It may then be described as emitting codewords $x \in C$ at discrete times $\tau \in T' \subset T$. It is then possible and common to define a Shannon entropy for the source with respect to C as

$$H_C(\tau) = -\sum x \in C P_C(x,\tau) \log_2 P_C(x,\tau). \tag{2}$$

The distributions P_S and P_C are generally only indirectly accessible through the observed historical output (states) of the decoded source/system. It is thus common to use this output to obtain a best-effort estimate of P_S or P_C and hence H_S or H_C, respectively. The notion of assigning an entropy to a finite string (the historical output since observation began) is thus practically as old as Shannon theory itself. Shannon himself proposed an n-block entropy in this context, where the probabilities of individual symbols are replaced by those of blocks of n symbols from S.
T-codes are variable-length codes that can be constructed recursively from S using an algorithm called *T-augmentation* [5, 6, 13]. Nicolescu and Titchener[4, 10] showed that

any finite string $\sigma \in S^+$ defines a unique T-code set $S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$, in which any string of the form σy, where $y \in S$, is one of the longest codewords. The algorithm that is used to derive the T-prefixes $p_1, \ldots, p_n$ and the T-expansion indices $k_1, \ldots, k_n$ for the construction of $S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$ from is called T-decomposition. A detailed explanation of T-decomposition may also be found in [13].
Titchener [12, 9, 11, 15] further used T-decomposition to derive a T-complexity measure, and from it a T-information and T-entropy measure, the latter of which was shown experimentally to be closely related to the Kolmogorov-Sinai/Pesin entropy of the logistic map [16]. This paper does not follow Titchener's route, rather, it explores a classical entropy approach – that of Shannon.

2 Entropy

Based on the existing results, we are thus able to map an arbitrary string σ uniquely to a state machine - the decoder for the T-code set $C = S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$, for which σy is one of the longest codewords. It is thus possible to compute a classical (Shannon) entropy for the string, provided we are able to state the probability with which the state machine may be found in any one of its states. This poses two problems:

- We need to decide whether to regard the internal nodes of the decoding tree as valid states (i.e., whether we permit them to have a non-zero probability). This corresponds to the question whether we permit observations to be taken in these states, i.e., at times when the decoder is mid-way through the decoding of a codeword. The obvious alternative is to permit leaf node (codeword) states only, and treatment in this paper will be rectricted to this case.

- The probability of finding the decoder in any particular state obviously depends on the semi-infinite string that it decodes. If we do not know the string (or at least some of its properties), then how can we derive the probability? If we apply the principle of maximum entropy, as seems prudent, then we need to demand that each of the symbols in S ought to be equally probable and that the probability of occurrence of any finite string $s \in S^\star$ is $P(s) = \#S^{-|s|}$. In other words: The semi-infinite string should be presumed to be random. This answers the second question.

If we only consider leaf nodes, the probability of occurrence of a symbol x in $S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$ is $P_C(x,\tau) = \#S^{-|x|}$. In this case, we may define an entropy function (here called H_1) from Eqn. (2) and the lengths of the codewords of the codeset that the decomposed string gives rise to, as:

$$H_1(\sigma) = -\sum_{x \in S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}} \#S^{-|x|} \log_2 \#S^{-|x|}, \tag{3}$$

where σy for any $y \in S$ is one of the longest codewords in $S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$.
The recursive construction of T-codes suggests that $H_1(\sigma)$ may also be derived recursively. For any $p_{n+1} \in S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$, we may indeed derive H_1 for the T-augmented

set/string by adding terms for the newly added codewords and by subtracting the term for p_{n+1}, which is no longer in the new codeset $S^{(k_1,k_2,\ldots,k_{n+1})}_{(p_1,p_2,\ldots,p_{n+1})}$. We obtain $H_1(p_{n+1}^{k_{n+1}}\sigma)$ from $H_1(\sigma)$ as follows:

$$\begin{aligned} H_1(p_{n+1}^{k_{n+1}}\sigma) &= H_1(\sigma) \\ &\quad - \sum_{k'_{n+1}=1}^{k_{n+1}} \sum_{x\in S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}} \#S^{-|x|-k'_{n+1}|p_{n+1}|} \log_2 \#S^{-|x|-k'_{n+1}|p_{n+1}|} \\ &\quad + \sum_{k'_{n+1}=1}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} \log_2 \#S^{-k'_{n+1}|p_{n+1}|} \end{aligned} \tag{4}$$

where we have used the identity

$$P(p_{n+1}^{k'_{n+1}}) = P(p_{n+1})^{k'_{n+1}} = \#S^{-k'_{n+1}|p_{n+1}|}. \tag{5}$$

The first term on the RHS of Eqn. (4) corresponds to the entropy contribution made by the codewords from $S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$, all of which (except p_{n+1}, whose contribution we still include in this term) are still present in the new set with unchanged probabilities of occurrence. The second term adds the entropy contributions from the k_{n+1} new copies of the original tree, the probabilities weighted by the respective distance from the root, which is given by $P(p_{n+1}^{k'_{n+1}})$, i.e., the length of the T-prefix chain. The third term subtracts the contributions of the T-prefix chain $p_{n+1}^{k'_{n+1}}$, $1 \le k'_{n+1} \le k_{n+1}$, which are "mistakenly" included in the first two terms, i.e., leaf nodes on the original tree or its copies that are now used to link each tree to the subsequent copy. Using Eqn. (3), Eqn. (4) may be rewritten as:

$$\begin{aligned} H_1(p_{n+1}^{k_{n+1}}\sigma) &= - \sum_{k'_{n+1}=0}^{k_{n+1}} \sum_{x\in S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}} \#S^{-|x|-k'_{n+1}|p_{n+1}|} \log_2 \#S^{-|x|-k'_{n+1}|p_{n+1}|} \\ &\quad + \sum_{k'_{n+1}=1}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} \log_2 \#S^{-k'_{n+1}|p_{n+1}|}, \end{aligned} \tag{6}$$

where we have integrated the existing entropy $H_1(\sigma)$ into the second term by extending the range of the outer sum to include a term for $k'_{n+1} = 0$. Rewriting the first term in terms of $H_1(\sigma)$, we get

$$\begin{aligned} H_1(p_{n+1}^{k_{n+1}}\sigma) &= - \sum_{k'_{n+1}=0}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} \sum_{x\in S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}} \#S^{-|x|} \log_2 \#S^{-|x|-k'_{n+1}|p_{n+1}|} \\ &\quad + \sum_{k'_{n+1}=1}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} \log_2 \#S^{-k'_{n+1}|p_{n+1}|} \\ &= - \sum_{k'_{n+1}=0}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} \log_2 \#S^{-k'_{n+1}|p_{n+1}|} \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{k'_{n+1}=0}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} H_1(\sigma) \\
&+ \sum_{k'_{n+1}=1}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} \log_2 \#S^{-k'_{n+1}|p_{n+1}|} \\
= \;& H_1(\sigma) \sum_{k'_{n+1}=0}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|}. \qquad (7)
\end{aligned}
$$

Here, we have made use of the fact that $\sum_{x \in S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}} \#S^{-|x|} = 1$ and that the term for $k'_{n+1} = 0$ in the first sum after the second equal sign is zero.

3 Observations on H_1

The result of the previous section permits a few observations:

- Since $k_{n+1} \geq 1$, the entropy function H_1 increases monotonously as we T-augment $S^{(k_1,k_2,\ldots,k_n)}_{(p_1,p_2,\ldots,p_n)}$, i.e., as σ has additional T-prefix strings added to it.
- As $\lim_{n\to\infty} |p_{n+1}| = \infty$, we can expect that for large n the terms for $k'_{n+1} > 1$ will either be zero (because $k_{n+1} = 1$) or very small compared to those for $k'_{n+1} = 1$. Hence we get:

$$
\begin{aligned}
H_1(p_{n+1}^{k_{n+1}}\sigma) &= H_1(\sigma) \sum_{k'_{n+1}=0}^{k_{n+1}} \#S^{-k'_{n+1}|p_{n+1}|} \\
&\approx H_1(\sigma)(1 + \#S^{-|p_{n+1}|}). \qquad (8)
\end{aligned}
$$

Note that the overwhelming majority of sufficiently long strings σ do not begin with a repetition of two or more long patterns, which makes $= 1$ the by far most common T-expansion parameter in the T-decomposition of long strings.

4 Conclusions

This paper proposed a Shannon-like entropy for finite strings, based on the T-decomposition of strings. The T-decomposition algorithm maps each finite string σ into a unique recursive variable-length code of the T-code family. The proposed entropy function $H_1(\sigma)$ of the string is the Shannon entropy of the associated T-code under the assumption that the code set decodes a source with maximum uncertainty, i.e., a random string. A simple recursive formula for the derivation of this entropy function H_1 was derived. It was also shown that $H_1(\sigma)$ exhibits properties one would expect from an entropy function as additional information is added to σ by prefixing.
As this paper reports on work in progress, it remains to be investigated in how far H_1 is compatible with Titchener's measures. However, due to its conventional derivation, it provide a useful tool in investigating links between the Titchener and Shannon entropy concepts.

5 Acknowledgements

My thanks go to Mark Titchener, Cris Calude and Monica Dumitrescu for useful discussions, hints and encouragment.

References

[1] C. E. Shannon: *A Mathematical Theory of Communications*. Bell Systems Technical Journal, 27:379, July 1948

[2] C. E. Shannon: *A Mathematical Theory of Communications*. Bell Systems Technical Journal, 27:623, October 1948

[3] A. Lempel and J. Ziv: *On the Complexity of Finite Sequences*. IEEE Trans. Inform. Theory", 22(1), January 1976, pp. 75-81.

[4] R. Nicolescu: *Uniqueness Theorems for T-Codes*. Technical Report. Tamaki Report Series no.9, The University of Auckland, 1995.

[5] M. R. Titchener: *Generalized T-Codes: an Extended Construction Algorithm for Self-Synchronizing Variable-Length Codes*, IEE Proceedings – Computers and Digital Techniques, 143(3), June 1996, pp. 122-128.

[6] U. Guenther: *Data Compression and Serial Communication with Generalized T-Codes*, Journal of Universal Computer Science, V. 2, N 11, 1996, pp. 769-795. `http://www.iicm.edu/jucs_2_11`

[7] U. Guenther, P. Hertling, R. Nicolescu, and M. R. Titchener: *Representing Variable-Length Codes in Fixed-Length T-Depletion Format in Encoders and Decoders*, CDMTCS Research Report no.44, Centre of Discrete Mathematics and Theoretical Computer Science, The University of Auckland, August 1997. `http://www.cs.auckland.ac.nz/research/CDMTCS/docs/pubs.html`.

[8] U. Guenther, P. Hertling, R. Nicolescu, and M. R. Titchener: *Representing Variable-Length Codes in Fixed-Length T-Depletion Format in Encoders and Decoders*, Journal of Universal Computer Science, 3(11), November 1997, pp. 1207–1225. `http://www.iicm.edu/jucs_3_11`.

[9] M. R. Titchener: *A Deterministic Theory of Complexity, Information and Entropy*, *IEEE Information Theory Workshop*, February 1998, San Diego.

[10] R. Nicolescu and M. R. Titchener, *Uniqueness Theorems for T-Codes*, Romanian Journal of Information Science and Technology, 1(3), March 1998, pp. 243–258.

[11] M. R. Titchener, *A novel deterministic approach to evaluating the entropy of language texts*, *Third International Conference on Information Theoretic Approaches to Logic, Language and Computation*, June 16-19, 1998, Hsi-tou, Taiwan.

[12] M. R. Titchener, *Deterministic computation of string complexity, information and entropy*, *International Symposium on Information Theory*, August 16-21, 1998, MIT, Boston.

[13] U. Guenther: *Robust Source Coding with Generalized T-Codes*. PhD Thesis, The University of Auckland, 1998. `http://www.tcs.auckland.ac.nz/~ulrich/phd.ps.gz`

[14] M. R. Titchener, P. M. Fenwick, and M. C. Chen: *Towards a Calibrated Corpus for Compression Testing*, Data Compression Conference, DCC-99, Snowbird, Utah, March 1999.

[15] M. R. Titchener: *A measure of Information*, IEEE Data Compression Conference, Snowbird, Utah, March 2000.

[16] W. Ebeling, R. Steuer, and M. R. Titchener: *Partition-Based Entropies of Deterministic and Stochastic Maps*, *Stochastics and Dynamics*, 1(1), p. 45.