Compressing Big Data: when the rate of convergence to the entropy matters

Filippo Mignosi

Computer Science Department, University of L'Aquila, Italy
Filippo.Mignosi@univaq.it

Abstract

In this talk we discuss of the rate of convergence to the entropy of dictionary based compressors. A faster rate of convergence to the theoretical compression limit should correspond to better compression in practice, but constants also matters. Therefore in the analysis of the rate of convergence one must also analyse the "transient" phase.

Concerning dictionary based compressors, it is known that LZ78-alike compressors have a faster convergence than LZ77-alike compressors, when the texts to be compressed are generated by a memoryless source. In practice instead it seems that LZ77-alike performs better. This seems due to the effect of a strategy of Optimal Parsing (that can be applied in both LZ77 and LZ78 cases) rather then to the fact that the texts are generated by a memoryless source. To our best knowledge there are no theoretical results concerning the rate of convergence to the entropy of both LZ77 and LZ78 case when it is used a strategy of Optimal Parsing.

We discuss some experimental results on LZ78 that show that the rate of convergence to the entropy presents a kind of *wave* effect that become bigger and bigger as the entropy of the memoryless source decrease. It can be a *tsunami* for a zero entropy source.

Compressing Big Data: when the rate of convergence to the entropy matters