MSIAM – 2nd year internship

Lempel-Ziv compression of tree-structured data

Supervisors: François Cayre(1) and Jean-Baptiste Durand(2)
E-mail: |Francois.Cayre@gipsa-lab.grenoble-inp.fr, Jean-Baptiste.Durand@imag.fr
Location of internship: GIPSA-lab

(1) GIPSA-lab, 11 rue des Mathématiques, Saint Martin d'Hères. +33(0) 4 76 82 63 78
(2) Laboratoire Jean Kuntzmann & Inria, Equipe Projet Mistis, Saint Martin d'Hères. +33(0)4 76 63 57 09

Context:

In the last years, collections of tree-structured data have emerged, either for the purpose of communications or data analysis (big data). This has occurred independently in several fields, and particularly in XML data and in plant growth analysis. In both fields, the data of interest can be represented as tree graphs. Sharing such data (for example in the purpose of distributed data mining) requires efficient data compression algorithms, which requires the particular structure of the data to be taken into account. A lossless compression algorithm was proposed by Choi and Szpankowski (2012) to compress Erdös-Rényi graphs up to graph isomorphic mappings, but this assumption is too restrictive to include tree graphs. Some algorithms have been proposed for ordered tree-structured data (Chen & Reif, 1996 and Itokawa *et al.*, 2009). However: 1) the desired properties of tree compression algorithms have been poorly formalized and 2) these algorithms are dedicated to ordered trees and obviously embed more information than necessary for the compression of unordered trees. Moreover, Chen & Reif (1996) claim that their algorithm belongs to the family of Lempel-Ziv compression algorithms. However the dictionary built while traversing the tree must be transmitted, which contradicts the Lempel-Ziv principle of online reconstruction of the dictionary while uncompressing the tree.

Tasks:

The aim of the internship is to define what desirable properties a lossless compression algorithm for labeled or unlabeled unordered trees should satisfy. Then, the internship consists in developing a true Lempel-Ziv algorithm for labeled or unlabeled unordered trees, and to assess its compression performance experimentally based on simulated trees under various assumptions of distributions. Finally, its performance on real datasets will be assessed, in particular using plant measurements.

Prerequisites:

Basic knowledge in lossless compression theory and advanced programming skills are required.

Related Master tracks:

Data Science, Statistics

This work is intended to be continued as a PhD thesis.

References:

S. Chen and J. Szpankowski. Compression of graphical structures: fundamental limits, algorithms, and experiments. *IEEE Transactions on Information Theory* **58**(2), 620-638 (1996).

Y. Choi and W. Reif. Efficient lossless compression of trees and graphs. In: *Proceedings of Data Compression Conference* (1996).

Y. Itokawa, K. Katoh , T. Uchida and T. Shoudai. Dictionary-based compression algorithms for tree structured data. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists* IMECS 2009 Vol I, March 18-20, Hong-Kong (2009).