Efficient Lossless Compression of Trees and Graphs *

Shenfeng Chen and John H. Reif
Department of Computer Science
Duke University
Durham, NC 27708

Abstract

Data compression algorithms have been widely used in many areas to meet the demand of storage and transfer of large size data. Most of the data compression algorithms regard the input as a sequence of binary numbers and represent the compressed data also as a binary sequence. However, in many areas such as programming languages (e.g. LISP and C) and compiler design, it is more desirable to have a compression algorithm which compresses a data structure while keeping a similar structure of the original data in the compressed form. In addition to reducing storage space, such compression also has the benefit of efficiently executing various operations (e.g. searching) in the compressed form.

In this paper, we study the problem of compressing a non-binary data structure (e.g. tree, undirected and directed graphs) in an efficient way while keeping a similar structure in the compressed form. To date, there has been no proven optimal algorithm for this problem. We use the idea of building LZW tree in LZW compression to compress a binary tree generated by a stationary ergodic source. The tree is parsed into subtrees using breadth first search and an optimal dictionary is constructed with each index pointing to a distinct subtree. We replace the parsed subtrees by dictionary indices to form the compressed tree. We also extend our tree compression algorithm to compress undirected and directed acyclic graphs.

*Email addresses: reif@cs.duke.edu and chen@cs.duke.edu. This work was supported by NSF Grant NSF-IRI-91-00681, Rome Labs Contracts F30602-94-C-0037, ARPA/SISTO contracts N00014-91-J-1985, and N00014-92-C-0182 under subcontract KI-93-01-0183.