Compressing Web Graphs as Texts

Gonzalo Navarro
Center for Web Research
Dept. of Computer Science
Univ. of Chile

Summary:

Graph compression, in particular for Web subgraphs, has recently received attention for its interest when analyzing those graphs to detect communities and run other graph mining algorithms. Having a compressed graph representation that permits accessing the graph without decompressing it makes it possible to run algorithms over graphs that otherwise would not fit in main memory. At this point, there exist proposals that use about 6 bits per link and permit accessing direct and reverse neighbors of a node within the microsecond per delivered edge.

In this talk I will present a novel approach to the problem. I will regard the graph structure as a text, which can be handled with known techniques for compressed text indexing. Those techniques permit storing a text in space proportional to its k-th order entropy, so that any text substring can be efficiently retrieved, and in addition one can perform efficient substring searching on that text. I will show that this notion of entropy captures some of the regularities that appear in Web graphs, and that the two operations provided by text indexes permit retrieving direct and reverse neighbors of each node.

Finally, I will present an implementation of the idea using a specific index, which yields results that are comparable in some cases with the best existing techniques.