Engineering Wavelet Tree Implementations for Compressed Web Graph Representations

Meng He, and Chen Miao
Faculty of Computer Science, Dalhousie University, Canada

We study compressed representations of web graphs. Among previous work, the solution by Hernández and Navarro [1] supports more queries than alternative approaches, including in-neighbour queries, out-neighbour queries and a set of mining queries. Their main strategy is to extract dense subgraphs from the given graph, and encode them using succinct data structures such as wavelet trees. Previous experimental studies on wavelet trees, however, test performance using textual data, and more engineering work is needed for the data generated from web graphs.

Our strategy is to use different implementations to encode bit vectors at different levels of the wavelet trees constructed for dense subgraphs, based on the observation that bit vectors at top levels are more compressible than the rest. These implementations are considered: `RRR`, practical implementations [2] of the structure by Raman et al. [3]; `RLEG`, a bit vector structure based on run-length and Elias gamma codes [4]; and `Plain`, an uncompressed representation with low overheads [5]. Two specific approaches are used to combine them: The first approach encodes bit vectors using `RRR` starting from the root of a wavelet tree, until a level for which `Plain` uses less space is reached. Then, starting from this level downwards, `Plain` is used to encode bit vectors. The second approach uses `RLEG`, `RRR` and `Plain` in a similar top-down fashion, and different tradeoffs can be achieved by using different block sizes for `RLEG`.

We implemented these approaches with code from [1, 4] and the compact structures library libcds (`http://recoded.cl/`), to encode data sets from the WebGraph Framework project (`http://webgraph.di.unimi.it/`). We obtained a rich set of time/space tradeoffs that can not be achieved using a single bit vector structure for all levels. The following three tradeoffs are particularly interesting: A new encoding scheme that decreases the space cost of Hernández and Navarro's structure by 9% to 19% (more than 13% for all but one graph), while only doubling query time; a new scheme that decreases the space cost by 4% to 12% (10% or more for most graphs), with roughly the same query time; and a new scheme that decreases the space cost and the query time by about 2% and $1\%-9\%$ (5% or more for most graphs), respectively.

References

[1] C. Hernández and G. Navarro, "Compressed representations for web and social graphs," *Knowledge and Information Systems*, vol. 40, no. 2, pp. 279–313, 2014.

[2] F. Claude and G. Navarro, "Practical rank/select queries over arbitrary sequences," in *SPIRE*, 2008, pp. 176–187.

[3] R. Raman, V. Raman, and S. R. Satti, "Succinct indexable dictionaries with applications to encoding k-ary trees, prefix sums and multisets," *ACM Transactions on Algorithms*, vol. 3, no. 4, p. 43, 2007.

[4] R. Grossi, J. S. Vitter, and B. Xu, "Wavelet trees: From theory to practice," in *CCP*, 2011, pp. 210–221.

[5] G. Navarro and E. Providel, "Fast, small, simple rank/select on bitmaps," in *SEA*, 2012, pp. 295–306.