GQS: a graph query system for pattern mining under iso- and homomorphism

Fannes, Thomas
thomas.fannes@cs.kuleuven.be

Ramon, Jan
jan.ramon@cs.kuleuven.be

September 17, 2014

Graph databases are typically used for large amounts of networked data. Recent years have seen an increasing interest in graph databases, with applications in, e.g., social networks, biological interaction networks, etc[1]. In our work, we focus on large databases where we are interested in listing and aggregating over embeddings of a pattern graph. Although these type of databases exists, efficient querying is not very deeply investigated, especially for certain aspects relevant to data mining and machine learning.

GQS, our Graph Query System, is an integration of different, state-of-the-art contributions into a reusable query system. First of all, our query system allows for the mining of (rooted) bounded treewidth patterns under homomorphism, only exponential in the treewidth and not in the network size. Secondly, GQS also allows for querying embeddings of a pattern under isomomorphism using a state-of-the-art randomization algorithm [2], only mildly exponential in the network size. Thirdly, in large networks, different embeddings can not be seen as statistically independent, which is required by many learning methods. We compute a measure for the effective sample size of a networked sample (i.e., the amount of statistical training information in a set of overlapping embeddings), which also generates weights for the training examples which then can be used by machine learning algorithms [3]. Finally, we expose a set of operators that allow different query optimizations.

Our graph query system is implemented in C++ and makes heavily use of template meta-programming.

References

[1] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Comput. Surv.*, 40(1):1:1–1:39, February 2008.

[2] AshrafM. Kibriya and Jan Ramon. Nearly exact mining of frequent trees in large networks. *Data Mining and Knowledge Discovery*, 27(3):478–504, 2013.

[3] Yuyi Wang, Jan Ramon, and Thomas Fannes. An efficiently computable subgraph pattern support measure: Counting independent observations. *Data Mining and Knowledge Discovery*, 27(3):444–477, May 2013.