

Master Thesis

Incremental Indexing of Very Large Graphs in Relational Databases

Prof. Dr. Burkhard Freitag
Chair of Information Management, University of Passau
<http://www.im.uni-passau.de/>
Tutor: Matthias Schmid (Matthias.Schmid@uni-passau.de)

March 21, 2019

Keywords: Database Systems, Very Large Graphs, Graph Databases, *SQLGraph*, Incremental Indexing

1 Introduction

Data graphs are of increasing importance in modern information systems. Therefore, an efficient method that allows to store graph data in a database is highly desirable. Unfortunately, the traditional straightforward implementation of a data graph based on edge and node tables is too inefficient. In [4] the *SQLGraph* approach to storing data graphs in a relational database has been proposed, which uses adjacency tables and can be proven to be much more efficient than the straightforward method.

In an ongoing research project at the Chair of Information Management, refined storage models and indexing schemes are investigated that are based on the *SQLGraph* proposal but can be tuned even better for some types of application.

2 Objectives

While providing competitive efficiency for reading data from a data graph - even when compared to native graph database systems -, the published *SQLGraph* approach is rendered inefficient when very large data graphs have to be loaded into the database.

As a first idea to boost efficiency, incremental loading the data graphs can be considered. However, incrementality is not easy to achieve. This is due to the fact that edges are stored redundantly in adjacency lists. Therefore all edges to or from a node must be known before the node can be processed. Moreover, all indexes used to increase access speed must be updated each time a new partial graph is loaded.

The overall objective of this thesis is the **development of a method for incremental loading and indexing data graphs based on the *SQLGraph* approach**. A derived objective is the experimental confirmation that the approach proposed in this thesis is indeed correct and efficient. To this end the LDBC social network benchmark [1, 2, 3, 5] is to be used as the data base.

The thesis consists of a conceptual part and an implementation part. The implementation can and shall be based on an existing prototypical implementation of our variant of *SQLGraph*.

3 Tasks

Among the subtasks that have to be performed are:

- Study the literature (see below);
- Study the existing system prototype;
- Implement the *SQLGraph* schema based on the ORACLE database management system;
- Implement basic graph operations such as the Star-Query;
- Develop a method for incremental loading and indexing data graphs based on the ORACLE implementation of *SQLGraph*;
- Confirm the correctness and efficiency of the proposed method using the LDBC benchmark;
- Document the proposed solution;
- Demonstrate the proposed solution.

References

- [1] P. A. Boncz. LDBC: benchmarks for graph and RDF data management. In *17th International Database Engineering & Applications Symposium, IDEAS '13, Barcelona, Spain - October 09 - 11, 2013*, pages 1–2, 2013.
- [2] O. Erling, A. Averbuch, J. Larriba-Pey, H. Chafi, A. Gubichev, A. Prat-Pérez, M. Pham, and P. A. Boncz. The LDBC social network benchmark: Interactive workload. In *Proc. of the 2015 ACM SIGMOD Intl. Conf. on Management of Data, Melbourne, Victoria, Australia, 2015*, pages 619–630, 2015.
- [3] A. Iosup, T. Hegeman, W. L. Ngai, S. Heldens, A. Prat-Pérez, T. Manhardt, H. Chafi, M. Capota, N. Sundaram, M. J. Anderson, I. G. Tanase, Y. Xia, L. Nai, and P. A. Boncz. LDBC graphalytics: A benchmark for large-scale graph analysis on parallel and distributed platforms. *PVLDB*, 9(13):1317–1328, 2016.
- [4] W. Sun, A. Fokoue, K. Srinivas, A. Kementsietsidis, G. Hu, and G. T. Xie. SQLGraph: an efficient relational-based property graph store. In *Proc. of the 2015 ACM SIGMOD Intl. Conf. on Management of Data, Melbourne, Victoria, Australia, 2015*, pages 1887–1901, 2015.
- [5] G. Szárnyas, A. Prat-Pérez, A. Averbuch, J. Marton, M. Paradies, M. Kaufmann, O. Erling, P. A. Boncz, V. Haprian, and J. B. Antal. An early look at the LDBC social network benchmark's business intelligence workload. In *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), Houston, TX, USA, June 10, 2018*, pages 9:1–9:11, 2018.