DS 200 - Module 5 - Summary of Research Problem

Shriram R.
M Tech (CDS)
06-02-01-10-51-18-1-15763

November 3, 2019

1 Problem Statement

The research problem revolves around the design of a distributed graph querying system which can provide interactive response to linear path queries over large temporal graphs at scale. The emphasis is on scalability to large graphs and addressing the temporal aspects of the graph and proposed model of queries using novel techniques.

2 Introduction

Graphs are a specific kind of data structure which has nodes (entities) and links/edges (relationships) between them. They are found in real-world in the form of social networks, transportation, financial transactions, Internet of things etc. My work focuses on temporal graphs [1] where the structure of the graph can change with time. I model such graphs in the form of interval graphs where intervals of validity for nodes and edges are provided. In addition, the graphs can also contain other properties like name, country etc. in each nodes/edges. The path query that is being worked deals with finding a fixed length linear path in the graph satisfying predicates on properties and temporal relationships between the elements in the path. Example: Finding a chain of friends who follow each other in a specific temporal order, likes from two friends on a post in a specific temporal order etc.

3 Existing Gaps

There are existing graph database systems that can handle large graphs. However, they do not have first class support for temporal dimension thereby limiting the expressiveness in the query language and do not take advantage of the temporality to improve the response time. Existing algorithms for static graph query processing cannot be trivially extended to answer temporal queries [3]. This applies to indexing techniques for graphs as well. There are more recent works [4] that does support temporal graph querying by retrofitting existing graph databases but they do not scale well for large graphs that do not fit in a single compute node.

4 Solution

The solution is built on vertex centric model adopted by various distributed graph processing systems [2]. Here, each node in the graph performs a user-defined computations and sends messages to their adjacent nodes. This happens in a synchronous manner and is globally coordinated by a master. The first aspect of our solution is query planing and optimization. A path query can executed in several ways by splitting it into smaller subqueries and then joining the results of these into the final resultset.

There is a cost model associated with the system which gives the cost to execute a query for a given split point in the linear path. It makes use of use of several statistics of the temporal graph like node/edge counts for different property types, degree of nodes at different time points in the graph, frequency distribution of property values for different time intervals etc.

To address the scalability and load balancing challenge, I partition the input graph based on their semantic types and temporal validity and then only perform computation on a subset of partitions based on the predicates in the query. There is also a plan for building an out of core execution mechanism for dynamically loading only required partitions into memory which decrease memory pressure and allow us to scale for large graphs.

5 References

1. Holme, Petter, and Jari Saramäki. "Temporal networks." Physics reports 519.3 (2012): 97-125.

2. Malewicz, Grzegorz, et al. "Pregel: a system for large-scale graph processing." Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010.

3. Wu, Huanhuan, et al. "Path problems in temporal graphs." Proceedings of the VLDB Endowment 7.9 (2014): 721-732.

4. Byun, Jaewook, Sungpil Woo, and Daeyoung Kim. "ChronoGraph: Enabling temporal graph traversals for efficient information diffusion analysis over time." IEEE Transactions on Knowledge and Data Engineering (2019).