Hasso Plattner Institut
IT Systems Engineering
Universität Potsdam

Bachelorarbeit

User-aided Pattern Search and Analysis on Business Graphs

Nutzergestuetzte Graphanalyse und Mustersuche auf Unternehmensgraphen

Milan Gruner
milangruner@gmail.com

Eingereicht am <TBD>

Fachgebiet Informationssysteme
Betreuung: Prof. Dr. Felix Naumann, Michael Loster, Toni Gruetze

Abstract

Costructing a graph made up of thousands of businesses may be hard, but actually making sense of it is a lot harder. With huge amounts of data potentially being integrated into the data lake every day, automatic methods for finding interesting spots in the graph are needed. This paper discusses different approaches that can be taken to extract useful knowledge from such a graph.

Contents

1 Introduction

1.1 Glossary

1.2 Motivation

1.3 Understanding risk analysis on graphs

1.4 Used techniques and assumptions

2 Data structures for business entities

2.1 Graph encoding for column family storage

2.2 The *subject* data structure

2.3 A versioning scheme that stands the test of time

3 Architecture (opt.)

3.1 Job and Data Management

3.1.1 Modularizing Spark jobs

3.1.2 Coordinating Spark jobs from NodeJS

3.1.3 Managing Cassandra tables from NodeJS

3.1.4 Data flow using column family storage

3.2 Using Apache Spark and Cassandra for Graph Analysis

3.2.1 Motivation: Why Spark, Cassandra and GraphX?

3.2.2 Writing efficient Spark GraphX code

3.2.3 Optimizing Cassandra data structures for Graph Processing

4 Graph Summarization

4.1 What users actually want to see

4.2 Compressing graph information to the bare minimum

4.3 Presenting graph data appealingly

4.4 Related work

5 Pattern Search

5.1 Discerning patterns from randomness

5.2 Operating on graph diffs

5.3 Pattern types and their applications

5.4 Related work

6 Pattern Analysis

6.1 User-aided approaches for Pattern Categorization

6.2 Pattern importance measures

6.3 Machine Learning Models for analyzing user feedback

6.4 Related work

7 Lessons learned

7.1 Benchmarks and Experiments

7.2 Design decisions and trade-offs

7.3 Technical challenges

8 Literature

References