Label Hierarchy Inference in Property Graph Databases

Bachelor thesis

vorgelegt von

Fabian Klopfer

an der

Universität Konstanz

Sektion Naturwissenschaften

Fachbereich Informatik und Informationswissenschaften

1.Gutachter: Prof. Dr. Michael Grossniklaus
2.Gutachter: Prof. Dr. Tatjana Pertov

Konstanz, 2019

Contents

Chapter 1

Introduction

1.1 Motivation

1.2 Overview

1.3 Contributions

1.4 Outline

Chapter 2

Background

2.1 Theory

2.1.1 Property Graph Model

2.1.2 Clustering

2.1.3 Formal Concept Analysis

2.1.4 Conceptual Clustering

2.2 Related Work

2.2.1 Cobweb

2.2.1.1 Category Utility

2.2.1.2 Concept Representation

2.2.2 Search and Operators

2.2.3 Cobweb/3

2.2.4 Extended Category Utility

2.2.5 Labyrinth

2.2.6 Incremental Self-Organizing Memory

2.2.7 Subdue

Chapter 3

Method

3.1 Problem Statement

3.2 Proposed Method

3.2.1 Clustering Property Containers

3.2.2 Extracting Structural Features

3.2.3 Clustering summarized Node

3.2.4 Complexity Analysis

3.2.4.1 Computation

3.2.4.2 Memory

3.3 Implementation

Chapter 4

Evaluation

4.1 Setup

4.1.1 Datasets

4.1.2 Ground Truth

4.2 Results

4.2.1 Yelp

4.2.1.1 Object-based

4.2.1.2 Graph-based

4.2.2 LDBC SNB

Chapter 5

Conclusion

5.1 Applications

5.2 Future Work

5.3 Summary

Bibliography

[Asghar and Ghenai, 2015] Asghar, N. and Ghenai, A. (2015). Automatic discovery of functional dependencies and conditional functional dependencies : A comparative study.

[Bohannon et al., 2007] Bohannon, P., Fan, W., Geerts, F., Jia, X., and Kementsietsidis, A. (2007). Conditional functional dependencies for data cleaning. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 746–755. IEEE.

[Fan et al., 2011] Fan, W., Geerts, F., Li, J., and Xiong, M. (2011). Discovering conditional functional dependencies. *IEEE Transactions on Knowledge and Data Engineering*, 23(5):683–698.

[Gullo et al., 2017] Gullo, F., Ponti, G., Tagarelli, A., and Greco, S. (2017). An information-theoretic approach to hierarchical clustering of uncertain data. *Information Sciences*, 402:199–215.

[Heller and Ghahramani, 2005] Heller, K. A. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pages 297–304. ACM.

[Huhtala et al., 1999] Huhtala, Y., Kärkkäinen, J., Porkka, P., and Toivonen, H. (1999). Tane: An efficient algorithm for discovering functional and approximate dependencies. *The computer journal*, 42(2):100–111.

[Jain et al., 1999] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

[Karypis et al., 1999] Karypis, G., Han, E.-H., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75.

[Martin-Löf, 1975] Martin-Löf, P. (1975). An intuitionistic theory of types: Predicative part. In *Studies in Logic and the Foundations of Mathematics*, volume 80, pages 73–118. Elsevier.

[Norell, 2007] Norell, U. (2007). *Towards a practical programming language based on dependent type theory*, volume 32. Citeseer.

[Papenbrock et al., 2015] Papenbrock, T., Ehrlich, J., Marten, J., Neubert, T., Rudolph, J.-P., Schönberg, M., Zwiener, J., and Naumann, F. (2015). Functional dependency discovery: An experimental evaluation of seven algorithms. *Proceedings of the VLDB Endowment*, 8(10):1082–1093.

[Robinson et al., 2013] Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph databases.* " O'Reilly Media, Inc.".

[Thompson, 1991] Thompson, S. (1991). *Type theory and functional programming.* Addison Wesley.

[Voevodsky et al., 2013] Voevodsky, V. et al. (2013). Homotopy type theory: Univalent foundations of mathematics. *Institute for Advanced Study (Princeton), The Univalent Foundations Program*, pages 2007–2009.

[Warrell, 2016] Warrell, J. H. (2016). A probabilistic dependent type system based on non-deterministic beta reduction. *CoRR*, abs/1602.06420.