Formal methods for knowledge extraction and reuse from heterogeneous sources for semantic interoperability of distributed architectures

Nicolás Leutwyler

Université de Lorraine, Nancy 54000, France

Abstract. The tendency in industry, manufacturing, and agriculture nowadays goes towards adopting the Industry 4.0 practices. Additionally, Internet of Things (IoT) has seen a huge increase in its usage over the last decade, and companies are eager to profit from the advantages it has offers. Between these tendencies, the usage of data as a means to increase productivity, or similarly, to minimize loss in production is found. In those lines, Formal Concept Analysis (FCA) is a clusterization method whose output is based on patterns of concepts (sets of objects and attributes). Some extensions such as Relational Concept Analysis have arisen to tackle the use case in which there are relations between seemingly different objects, which is something FCA cannot do. However, the area of automatically using the conceptual data resulted from these methods is still immature in the sense of formalization and usage. In this Ph.D., the goal is to work in expanding the boundaries of knowledge regarding the existing algorithms, mainly looking for optimizations, and extending their current capabilities.

1 Introduction

The Ph.D., focus is on the creation of mathematical models and the implementation of intelligent sensors, or cyber-physical systems (CPS), to enrich the data layer coming from the field. One of the most relevant scientific challenges is the lack of mathematical formalization of the system models and the resulting information systems, as well as the definition of the semantics of the concepts and relationships they apply, to ensure their common understanding and facilitate their interoperability by minimizing semantic losses.

The challenge of this research project is double: on the one hand, to model data from heterogeneous sources and, on the other hand, to study the problems posed by model-based engineering in cooperative systems. It is about cooperation in “systems of actors” willing to interoperate. Cooperative systems are now organized in networks, i.e., in the form of complex systems.

A known method for knowledge extraction is the Formal Concept Analysis (FCA) [7]. Given a set of objects, with certain attributes each, the method clusterize them in pairs (X, Y) where all the objects in X have all the attributes in Y , and each attribute in Y is held by all the objects in X. The pair is called a formal concept, and the inclusion of either the first component or the second one between formal concepts forms a partially ordered set called concept lattice [2]. This is one of the most used methods for association rules mining [3]. Moreover, one of the extensions that allows to understand relations between different (heterogeneous) types of contexts is the Relational Concept Analysis, which is known to lack some capabilities such as the intuitive representation of ternary relations [12].

2 State of the art

The complex systems envisaged will be composed of CPS networks, intelligent sensors whose purpose is to retrieve data by inserting the context and thus form information networks [10, 6] which consists of relying on different types and levels of abstraction or models. Formal concept analysis (FCA) [7] is a useful and powerful method for formally describing the links between any objects (that form a context), in particular between objects that convey knowledge. This method is based on the lattice theory [2], which can be used to solve problems of interoperability assessment between information systems within companies. An extension of the FCA framework was introduced in [9] and is called Relational Concept Analysis (RCA).
RCA focuses on datasets that are compatible with Entity-Relationship Models [1] or, alternatively, Resource Description Framework (RDF) [4]. Linked open data has been recognized as a valuable source of general data mining information, and knowledge graphs are a method for formalizing this knowledge [11].
This provides a method for extracting conceptual knowledge from multirelational data. Information extraction is part of the field of study called data mining [8], information that can be related to each other can be studied through the methods of multirelational data mining (MRDM) [5] that deals with multi-contextual data. The RCA method is not limited to knowledge extraction from separate contexts, but aims to express knowledge by interoperating the semantics of different contexts, i.e., in addition to extracting knowledge from different contexts, it also extracts knowledge from a specific context, the data contained in the other contexts are used to enrich knowledge extraction.

3 Problem Statement and Contribution

The contributions of this PhD project should include: (a) a state-of-the-art research, (b) development tools, extensions and mathematical models towards the usage of knowledge extraction, (c) the inclusion of the developed tools into the ´Ecole de Ski Fran¸caise (ESF) enterprise’s software.

4 Research Methodology and Approach and Evaluation Plan

Considering the already stated problem, the methodology of the Ph.D., is divided into three main parts:

(a) systematic literature review (SLR) and the understanding of the state of the art, (b) using the outcome of the SLR to pursue a research goal and publish the advances, (c) writing the final report including a compendium of the previous works in the lines of the project, a development of each of the contributions, and a guideline to continue adding value to the field in the same topics.

Since the project involves both mathematical and software contributions, the evaluation plan has to consider them differently. On the one hand, the mathematical contributions would be evaluated by their proofs and the level of maturity they reach. On the other hand, the software would be evaluated by the users, whom can later provide insight on how useful or complete the software is to their expectations.

5 Preliminary results, Conclusions and Learned Lessons

So far, the project contains a small contribution, which is an obvious extension to the RCA algorithm to allow mining data represented with ternary relations. This is something that was already done in different ways, the novelty of the contribution would be that the way in it is implemented is more intuitive and natural than the other ones.

The generalization of the method (from ternary to n-ary) is still a work in progress, and has already been rejected in a conference.

In terms of learned lessons, I could say that, firstly, I learned the importance of having a plan when reading scientific papers. Secondly, to be clear about each contribution is way more important when writing.

Lastly, that it is also important to have a plan for publications.

References

1. Peter Pin-Shan Chen. “The entity-relationship model&#x2014;toward a unified view of data”. In: ACM Transactions on Database Systems 1.1 (Mar. 1976), pp. 9–36. issn: 0362-5915. doi: 10.1145/320434.320440.
2. Rudolf Wille. “Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts”. In: Ordered Sets. Ed. by Ivan Rival. Dordrecht: Springer Netherlands, 1982, pp. 445–470. isbn: 978-94-009-7798-3. doi: 10.1007/978-94-009-7798-3_15.
3. Rakesh Agrawal, Tomasz Imieli´nski, and Arun Swami. “Mining association rules between sets of items in large databases”. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data. SIGMOD ’93. New York, NY, USA:

Association for Computing Machinery, June 1993, pp. 207–216. isbn: 978-0-89791-592-2. doi: 10.1145/170035.170072.

4. Eric Miller. “An Introduction to the Resource Description Framework”. In: Bulletin of the American Society for Information Science and Technology 25.1 (1998). eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bult.105, pp. 15–19. issn: 1550-8366. doi: 10.1002/bult.105.
5. Saso Dzeroski. “Multi-relational data mining: an introduction”. In: ACM SIGKDD Explorations Newsletter 5.1 (July 2003), pp. 1–16. issn: 1931-0145. doi: 10.1145/959242.959245.
6. Gérard Morel et al. “Manufacturing Enterprise Control and Management System Engineering: paradigms and open issues”. en. In: Annual Reviews in Control 27.2 (Jan. 2003), pp. 199–209. issn: 13675788. doi: 10.1016/j.arcontrol.2003.09.003.
7. Uta Priss. “Formal concept analysis in information science”. In: Annual Review of Information Science and Technology 40.1 (Sept. 2007), pp. 521–543. issn: 00664200. doi: 10.1002/aris.1440400120.
8. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. ISBN: 9780511809071 Publisher: Cambridge University Press. July 2008. doi: 10.1017/CBO9780511809071.
9. Mohamed Rouane-Hacene et al. “Relational Concept Analysis: Mining Concept Lattices From Multi-Relational Data”. In: Annals of Mathematics and Artificial Intelligence 67 (Jan. 2013). doi: 10.1007/s10472-012-9329-3.
10. Olivier Cardin. “Contribution a la conception, l’evaluation et l’implémentation de systémes de production cyber-physiques”. Habilitation a diriger des recherches. Université de Nantes, Dec. 2016. https://tel.archives-ouvertes.fr/tel-01443318.
11. Petar Ristoski and Heiko Paulheim. “RDF2Vec: RDF Graph Embeddings for Data Mining”. In: The Semantic Web – ISWC 2016. Ed. by Paul Groth et al. Cham: Springer International Publishing, 2016, pp. 498–514. isbn: 978-3-319-46523-4. doi: 10.1007/978-3-319-46523-4_30.
12. Priscilla Keip et al. “Effects of Input Data Formalisation in Relational Concept Analysis for a Data Model with a Ternary Relation”. In: Formal Concept Analysis. Ed. by Diana Cristea, Florence Le Ber, and Baris Sertkaya. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 191–207. isbn: 978-3-030-21462-3. doi: 10.1007/978-3-030-21462-3_13.