KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

DRAFT

To remove, add 'final' to class options

Modeling Relational Data Mining

Sergey Paramonov

Supervisors:
Prof. dr. Luc De Raedt
Prof. dr. Marc Denecker

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

August 2017

Modeling Relational Data Mining

Sergey PARAMONOV

Examination committee:
Prof. dr. ir. The Chairman, chair
Prof. dr. Luc De Raedt, supervisor
Prof. dr. Marc Denecker, supervisor
Prof. dr. Gerda Janssens
Dr. Matthijs van Leeuwen
Helmut Simonis
(University College Cork, Insight Centre for Data Analytics)
Prof. dr. Christian Bessiere
(CNRS, U. Montpellier, LIRMM)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

August 2017

Preface

...

Instructies van de faculteit:

In het voorwoord wordt de algemene doelstelling van het werk samengevat in enkele regels en worden personen, diensten of firma's bedankt voor hun medewerking bij het tot stand komen van het werk.

De naam van firma's en personen uit deze firma's mogen slechts worden vermeld mits hun uitdrukkelijke toelating én na overleg met de supervisor(en)! Steeds wordt de supervisor(en) vermeld, de verantwoordelijke en eventueel de personen die rechtstreeks geholpen hebben bv. door het ter beschikking stelling van meetresultaten, faciliteiten. Ook de instantie die eventueel een doctoraatsbeurs heeft toegekend wordt bedankt (bv. FWO, IWT, ...).

Abstract

< What data mining is >
Data Mining is the process of discovering new knowledge from the data. A significant attention in the research community is devoted to data analysis in the presence of the data independence assumption, i.e., when the collected data points are independent from each other.

< What relational means >
However, it is often the case that the objects are connected and related to each other by means of relations. This setting is called *relational* and the associated task is referred as *relational reasoning*.

< Relational problems >
Electronic tables, spreadsheets, and databases are all examples of relational data that is in a wide use today. The objects are connected by means of relations and constraints. In spreadsheets these are tables and formulae, and in databases schema relations and integrity constraints.

< Key issue >
We argue that a general approach for modeling and solving data mining problems in the relational setting is missing. The goal of this thesis is to fill in this gap.

< Contribution >
Firstly, we demonstrate how the problem of learning in relation setting is different from the classical machine learning approach and propose a system named TaCLe as the first working in this setting.

Secondly, we demonstrate how the relational approach generalizes a classical problem of boolean matrix factorization into the Relational Data Factorization, which allows to model a spectrum of classical data mining problems and introduces new ones as well.

Thirdly, we demonstrate how existing relational reasoning formalisms, such as Answer Set Programming, can be enhanced by relational learning techniques

known as sketching.

Last but not least, we demonstrate how relational approach can be used to mine relational patterns, known as structured pattern mining in data mining community.

Beknopte samenvatting

...

Instructies van de faculteit:

In een beknopte tekst van maximum 2 pagina's worden de belangrijkste doelstellingen en besluiten geformuleerd, zowel in het Nederlands als in het Engels. Zulke samenvattingen kunnen worden gebruikt in wetenschappelijke verslagen van het departement of de faculteit. Het Engels moet vlekkeloos zijn.

List of Abbreviations

ASP Answer Set Programming. ix, 6

FOL First Order Logic. ix, 5

Contents

List of Figures

List of Tables

Instructies van de faculteit:

De hoofdstukken: Elk hoofdstuk is ingelast met een bepaald doel voor ogen. Dit doel wordt vermeld in de eerste paragraaf van elk hoofdstuk. Naargelang de aard van de tekst (experiment, uitvoering, theoretische ontwikkeling, ...) volgen de paragrafen elkaar op. Beweringen worden altijd gestaafd, hetzij door eigen experimenten, hetzij door een theoretische afleiding, hetzij door verwijzingen naar de literatuur. Elk hoofdstuk eindigt met een kort samenvattend besluit waarbij nagegaan wordt in hoeverre de doelstelling van het betrokken hoofdstuk verwezenlijkt is. De deelbesluiten moeten de lezer automatisch leiden naar het algemeen besluit aan het einde van het werk.

Chapter 1

Introduction

In his seminal work Sir Bob Kowalski (1979) proclaimed:

$$\text{Algorithm} = \text{Logic} + \text{Control}.$$

1.1 Structure of the Text

Sergey: talk about chapters here

1.2 Datasets, code and experimental results

Sergey: talk about github here: TaCLe, SkASP, etc

Chapter 2

Background

> From now on you shall be called Brian that is called Brian.
>
> The Life of Brian

In this chapter, we introduce commonly used formalisms and definitions.

2.1 First Order Logic

In this section, we describe the syntax and semantics of FOL, for an extensive overview of FOL, we refer to Enderton (2001).

A formal language is a triple: *vocabulary*, *syntax* and *semantics*. Vocabulary is the set of symbols that can be used. Syntax is the set of rules on how these symbols can be combined together. And semantics is the way to interpret the statements.

For each predicate p and each function symbol f in the vocabulary, we assume a natural number n called *arity* to be given, written as p/n and f/n. This number indicates the number of parameters it takes, we often omit the arity if it is clear from the context. Propositional symbols are zero-arity predicates and constants are zero-arity functions. We assume propositional symbols *true*, $\top$ and *false*, $\bot$ to be always in the vocabulary.

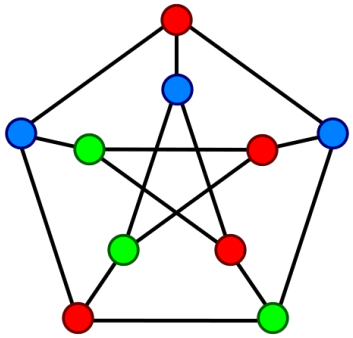

Figure 2.1: Graph coloring of the Petersen's graph using three colors

Example 2.1 (Predicates and functions)**.** Consider the famous problem of coloring a map, as in Figure 2.1. The vocabulary would contain a predicate symbol *border/2* and a function *coloring/1*, together with a set of constants representing countries, such as *belgium*, *netherlands*, etc.

Syntax

Semantics

2.2 Answer Set Programming

2.3 FO(·) and the IDP system

2.4 Inductive Logic Programming and Relational Pattern Mining

Chapter 3

This is conclusion

...

Instructies van de faculteit:

Algemene besluiten: Verwijzend naar de inleiding en naar de besluiten van de afzonderlijke hoofdstukken worden op het einde van het proefschrift de voornaamste besluiten gebundeld. Hier wordt de nadruk gelegd op de eigen inbreng, de verworven resultaten, de 'stellingen' van het proefschrift en de originele bijdragen tot het onderzoeksdomein. De onopgeloste problemen worden aangestipt en suggesties voor eventueel verder onderzoek worden gemaakt.

Appendix A

This is myappendix

...

Instructies van de faculteit:

De appendices: ze omvatten alle gedeelten uit de tekst die weliswaar essentieel zijn voor het proefschrift, maar waarvan de inlassing in de tekst de leesbaarheid ervan nadelig zouden beïnvloeden bv. omwille van hun lengte. Zo kunnen bv. de brute meetresultaten of een computerprogramma met zijn bron, commentaar en voorbeelden beter thuishoren in een appendix dan in de tekst zelf. De appendices kunnen desgevallend worden gebundeld in een apart boekdeel.

Bibliography

Enderton, H. B. (2001). *A mathematical introduction to logic*. 2nd ed. Harcourt/Academic Press.

Kowalski, R. (1979). “Algorithm = Logic + Control”. In: *Commun. ACM* 22.7, pp. 424–436. DOI: 10.1145/359131.359136.

Instructies van de faculteit:

De bibliografie. Departementale richtlijnen terzake te volgen.

This is curriculum

...

Instructies van de faculteit:

Beknopt CV van de doctorandus.

List of publications

Input file chapters/publications/publications.tex does not exist. Make sure its starts with “\chapter{List of publications}”. To not include this chapter in the table of contents, use the starred version of the \chapter command...

Instructies van de faculteit:

Lijst van de publicaties door de doctorandus/a (auteur of co-auteur).

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
DTAI
Celestijnenlaan 200A
B-3001 Leuven
sergey.paramonov@kuleuven.be
sergey-paramonov.com