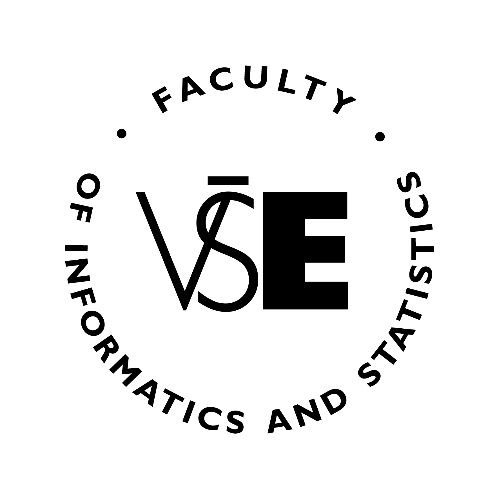
Prague University of Economics and Business

Faculty of Informatics and Statistics

****

**Mining rules from knowledge graphs**

MASTER THESIS

Study programme: Knowledge and Web Technologies

Author: Bc. Dominika Ludvíková

Master Thesis Supervisor: prof. Ing. Vojtěch Svátek, Dr.

Prague, December 2024

**Acknowledgement**

Abstract

Keywords

Contents

[Introduction 5](#_Toc176733581)

[1 Linked data 6](#_Toc176733582)

[1.1 RDF 6](#_Toc176733583)

[1.2 Knowledge graphs 7](#_Toc176733584)

[1.2.1 Ontologies 7](#_Toc176733585)

[1.2.2 Triple stores 7](#_Toc176733586)

[2 KG-Microbe 8](#_Toc176733587)

[2.1 Structure of KG-Microbe 8](#_Toc176733588)

[2.2 Experiments on KG-Microbe 9](#_Toc176733589)

[3 Association rules 10](#_Toc176733590)

[3.1 Measuring association rules 10](#_Toc176733591)

[3.1.1 Support 10](#_Toc176733592)

[3.1.2 Confidence 10](#_Toc176733593)

[3.2 AMIE algorithms 10](#_Toc176733594)

[3.2.1 Close-world assumption problem 11](#_Toc176733595)

[3.2.2 Architecture of AMIE+ 11](#_Toc176733596)

[3.2.3 RDF Rules 11](#_Toc176733597)

[4 Extraction of rules 12](#_Toc176733598)

[4.1 Data preparation 12](#_Toc176733599)

[4.1.1 Conversion to RDF 12](#_Toc176733600)

[4.1.2 Filtering data 12](#_Toc176733601)

[4.1.3 Train test splitting 12](#_Toc176733602)

[4.2 Rule mining 12](#_Toc176733603)

[4.2.1 Parameter setting 12](#_Toc176733604)

[4.2.2 Pruning 12](#_Toc176733605)

[4.3 Evaluation measures 13](#_Toc176733606)

[5 Evaluation of results 14](#_Toc176733607)

[5.1.1 Accuracy 14](#_Toc176733608)

[5.1.2 Discussion of rules 14](#_Toc176733609)

[Conclusions 15](#_Toc176733610)

[List of references 16](#_Toc176733611)

[Appendices I](#_Toc176733612)

Introduction

…

The goal for this thesis is to demonstrate the performance of an association rule mining tool in comparison with other referential approaches in the domain of knowledge graphs.

…

# Linked data

Linked data is a term for structured data on the web that are connected with other data to create one big network of knowledge instead of having parts of the knowledge in separate places. The term was introduced by Tim Berners-Lee who also defined four rules that serve as expectation of behaviour of such data in the semantic web. (Berners-Lee, 2009)

* Use URIs as names for things
* Use HTTP URIs so that people can look up those names
* When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL)
* Include links to other URIs. so that they can discover more things

## RDF

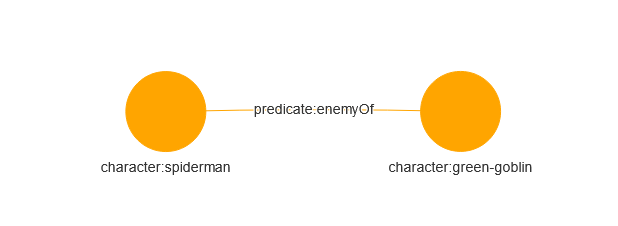
RDF is a short for Resource Description Framework and it’s a standard data model for linked data. The structure of the model can be imagined as a directed graph with nodes and edges. The basis of the RDF model are triples. A triple consists of three parts: a subject, a predicate and an object. The subject and the object are the nodes of the graph with the predicate representing the edge between the nodes. An example of a triple is shown in Figure 1.

Figure 1 Depiction of an RDF triple (Source: Author)

The subject can have the form of an URI to identify a specific instance. A subject without a URI is called a *blank node* and it simply specifies the existence of a subject without any identifying attribute. The predicate defines the relationship between the subject and the object and can only be represented by a specific URI. The object describes additional information about the connected subject and in addition to a URI and a blank node can be also represented by a literal value.

There are many ways to express the abstract RDF model. Identical information can be conveyed in various formats with a different syntax called serializations. Each serialization can be converted into another. The most common RDF serializations are:

* Turtle
* N-Triples
* JSON-LD
* RDF/XML

## Knowledge graphs

Large sets of linked data focused on representing real-life entities and their relations usually in one area of expertise, are called knowledge graphs. A knowledge graph uses the interconnected nature of a graph to map the relationships between entities to create a web of knowledge, sometimes linking data from multiple sources. Knowledge graphs can be used in multiple domains. One example is the Google Knowledge Graph which is used to display additional information connected to a requested search. Other uses of knowledge graphs are in financial services, medical and biological research and other relationship-based domains.

### Ontologies

### Triple stores

# KG-Microbe

KG-Hub is an initiative which serves as a platform that creates, maintains and distributes several projects focused on biological and biomedical knowledge graphs. Each graph is produced using many standardized patterns including the usage of OBO (Open Biological and Biomedical Ontologies) ontology. One of the knowledge graphs that is a part of this project is KG-Microbe which focuses on collecting data about microbial traits and other biological descriptors classified by taxonomy. Biomedical data has a wide range of ontologies and models that can be used in knowledge graphs such as KG-Microbe that elevate the standardization of information across the domain. Examples of used schemas in this specific knowledge graph are the aforementioned OBO ontology, the NCBO Bioportal and the Biolink Model which serves as a model for biological entities. Natural Language Processing (NLP) methods have been used as well to annotate new terms from raw data. (Joachimiak et al., 2021)

## Structure of KG-Microbe

KG-Microbe is a large knowledge graph that consists of 6 794 053 triples. Below in Table 1 are the 26 predicates that connect the nodes in the graph:

|  |  |
| --- | --- |
| Predicate | Occurrence |
| biolink:subclass\_of | 1 383 873 |
| <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> | 1 318 552 |
| rdfs:label | 1 289 513 |
| dc1:identifier | 1 158 879 |
| biolink:synonym | 587 173 |
| biolink:has\_participant | 154 642 |
| biolink:has\_phenotype | 138 549 |
| biolink:has\_part | 123 439 |
| biolink:consumes | 118 429 |
| biolink:capable\_of | 118 080 |
| biolink:related\_to | 117 016 |
| biolink:has\_output | 79 862 |
| biolink:has\_input | 79 862 |
| biolink:assesses | 50 357 |
| biolink:occurs\_in | 32 311 |
| biolink:location\_of | 25 950 |
| biolink:enabled\_by | 7 636 |
| biolink:enables | 4 443 |
| biolink:produces | 3 493 |
| biolink:has\_chemical\_role | 1 062 |
| biolink:part\_of | 509 |
| biolink:subPropertyOf | 213 |
| biolink:is\_assessed\_by | 112 |
| biolink:inverseOf | 46 |
| biolink:type | 45 |
| biolink:associated\_with | 7 |

Table 1 List of predicates occurring in KG-Microbe

Some predicates have been used to denote relationships between entities while others are simply descriptive.

## Experiments on KG-Microbe

One of the focus points of research on KG-Microbe has been classification of the media that is connected to a microbiological entity by the predicate *biolink:occurs\_in*. A medium in a microbiological context is used to supply the microbe with all necessary nutrients for growth in a laboratory environment.

# Association rules

Association rule mining is a method of data mining that focuses on finding patterns and discovering relationships between entities. Association rule can be imagined as an implication in the form of an if-then statement. Such statements have two parts. First part is called the antecedent and it represents the hypothesis of a statement or the *if* side. Second part is the consequent which is the conclusion or the *then* side of the rule. A statement in the form of an association rule specifies that if the antecedent is true, then the consequent is true as well.

An example of an association rule can be the following statement:

cloudy rain

It can be roughly translated into common language as “If it is cloudy, it will rain”. Such statement might be true sometimes, but there are also days where it will not rain even with a cloudy sky. That does not necessarily mean that we cannot use this rule to predict future occurrences of rain. It might have a better chance at recognizing a rainy day than other rules for example:

sunny rain

To differentiate the rules that result in better findings various measures have been established. These measures focus on different evaluations of the importance of a rule and help us make the distinction which rules we can use for better predictions. The most common measures are support and confidence.

## Measuring association rules

### Support

### Confidence

## AMIE algorithms

### Close-world assumption problem

The main problem of mining association rules is figuring out how to find the best possible rules in an acceptable computational time. One of the first algorithms for association rule mining was the Apriori algorithm. This algorithm was created for mining rules from transactional data in relational databases, using the closed-world assumption (CWA). CWA assumes that all true statements are present in the system and therefore every statement that is not explicitly stated in the system is considered false.

Knowledge graphs of all kinds are often considered incomplete as they are usually created to describe data from areas that are not fully explored. A different type of a system other than the CWA is needed to deal with incomplete data. The opposite of CWA is called the open-world assumption (OWA). A system operating under the OWA does not treat the missing statements automatically as false and instead they are classified as unknown. It is possible that such statement is missing for other reason than that it is false and it might be added to the system in the future after we acquire the knowledge.

An example of a system operating under CWA would be a database at a veterinary clinic. Every single pet that is being treated at this facility is recorded in the database. If we have the following information in Table 2 about the pets owned by John, the statement “John owns a rabbit” would be considered false. For the purposes of one veterinary clinic the CWA system is sufficient as it is only necessary to have data about the pets treated there.

|  |  |  |
| --- | --- | --- |
| Name | Species | Owner |
| Alex | Dog | John |
| Marty | Cat | John |

Table 2 Example of a database

In contrast with the previous example, if we considered the same data under the OWA we would have to concede that there is not enough information to consider the statement “John owns a rabbit” false. It is possible, that John has a rabbit that he takes to a different veterinary clinic.

It is obvious that for smaller enclosed systems that we have control over, the CWA is the right system to use. For observational data and more abstract general knowledge where we know that our data is incomplete, we might consider using OWA to avoid classifying missing statements as negative examples.

### Architecture of AMIE+

### RDF Rules

# Extraction of rules

## Data preparation

KG-Microbe is distributed as a pair of TSV files, where one file carries the information about the nodes and the other describes the edges of the knowledge graph. The two files were converted by a script into one RDF collection that could be processed in a tool specialising on analysing RDF data (*KIZI/KgMicrobeToRdf*, 2024).

### Conversion to RDF

The most common prefixes had been defined to make results easier to read and interpret. The file containing nodes of the graph has been used to extract identifying information and labels whereas the file comprising the edges has been used to create connections between nodes. The Turtle serialization has been used since it supports prefix definition and is human-readable, which can be an advantage when interpreting resulting rules.

### Filtering data

### Train test splitting

## Rule mining

### Parameter setting

### Pruning

## Evaluation measures

# Evaluation of results

### Accuracy

### Discussion of rules

Conclusions

List of references

Berners-Lee, T. (2009). *Linked Data—Design Issues*. Linked Data. https://www.w3.org/DesignIssues/LinkedData

Joachimiak, M. P., Hegde, H., Duncan, W. D., Reese, J. T., Thessen, A. E., & Mungall, C. J. (2021). *KG-Microbe: A Reference Knowledge-Graph and Platform for Harmonized Microbial Information*.

*KIZI/KgMicrobeToRdf*. (2024). [Jupyter Notebook]. KIZI. https://github.com/KIZI/KgMicrobeToRdf

Appendices