Assignment 2

Authors: Akos Engelmann, Gergely Parady, Ipek Cakin, Simon Jasansky

# Task 1

In the first task, we want to handcraft rules for relations between subjects and objects.

Figure 1:

Diagram

Description automatically generated

Before and after every rule, we determined the confidence by calculating support and body coverage (without PCA Assumption)

## 1. Rule

R1: If a paper that appears in a conference has a domain, also the conference has this domain

Reason why it could not be a good rule: it could be that the domain of a paper is not the main domain of the conference, but its just somehow related.

Quality: low (confidence: 0)

Horn Rule:

:conferenceA hasDiscipline :domainX <= :somepaper appearsInConferenceSeries :conferenceA, :somepaper hasDiscipline :domainX

Before Insertion: Support: 0; Body coverage: 5151, Confidence: 0.0

After Insertion: Support: 5151; Body coverage: 5151, Confidence: 1.0

That means 5151 triples were inserted

## 2. Rule

R2: An affiliation has a discipline, if an author has created a paper, and the paper has a discipline

Reasoning why it could not be a good rule: it could be that the domain of a paper is not the main domain of the affiliation (university), but its just somehow related.

Quality: medium low (Confidence: 0)

Horn Rule

:affiliationQ hasDiscipline: domainX <= paperA hascreator: authorA, paperA hasDiscipline: domainX, authorA memberOf: affiliationQ

Before insertion: Support: 0; Body coverage: 18815, Confidence: 0.0

After Insertion: Support: 18815; Body coverage: 18815, Confidence: 1.0

That means 18815 triples were inserted

## 3. Rule

R3: An author has a discipline, if an author has created a paper, and the paper has a discipline

Reasoning why it could not be a good rule: It could be that the domain of a paper is not the main domain of the author, but the author just supervised the paper or only gave comments or improvement points to the main authors of the paper but still appears as an author.

Quality: low (Confidence: 0)

Horn Rule

:authorA hasDiscipline domainX <= :paperA hasDiscipline domainX, :paperA hascreator :authorA

Before Insertion: Support: 0; Body coverage: 41929, Confidence: 0.0

After Insertion: Support: 41929; Body coverage: 41929, Confidence: 1.0

That means 41929 triples were inserted

## 4. Rule

R4: A paper is published by an affiliation, if the paper has a creator author, and the author is a member of the affiliation

Quality: low (Confidence: 0)

Horn Rule:

:paperA hasAffiliation :affiliationQ <= :paperA hascreator :authorA, :authorA memberOf :affiliationQ

Before Insertion: Support: 0; Body coverage: 8302, Confidence: 0.0

After Insertion: Support: 8302; Body coverage: 8302, Confidence: 1.0

That means 8302 triples were inserted

## 5. Rule

R5: An author attends a conference, if a paper appears in conference and the paper has creator author

Quality: low (Confidence: 0)

Horn Rule:

:authorA attendsConference conferenceZ <= :paperC appearsInConferenceSeries conferenceZ, :paperC hascreator authorA

Before Insertion: Support: 0; Body coverage: 8349, Confidence: 0.0

After Insertion: Support: 8349; Body coverage: 8349, Confidence: 1.0

That means 8349 triples were inserted

## 6. Rule

R6: Two authors are colleagues, if they are the members of the same affiliation, and if they are coauthors, and if they work in the same domain

Quality: low (Confidence: 0)

Horn Rule:

:authorA isColleagueOf :authorB <= :authorA memberOf :affiliationZ, :authorB memberOf :affiliationZ, :authorA coauthor ?authorB, :authorA hasDiscipline :domainA, :authorB hasDiscipline :domainA

Before Insertion: Support: 0; Body coverage: 9017, Confidence: 0.0

After Insertion: Support: 9017; Body coverage: 9017, Confidence: 1.0

9110 triples were inserted

NOTE: After only having rules with confidence 0, we decided to add two more rules with higher confidence, as required for the assignment

Also, these rules are not included in the sketch above, as we only added them after finding out all our previous rules are low confidence.

## 7. Rule

R7: Paper A has the same domain as paper B, if paper A cites paper B.

Quality: low (confidence: 0.12113991612657263)

Horn Rule:

:paperA hasDiscipline :domainX <= :paperA cites :paperB, :paperB hasDiscipline :domainX

Before Insertion: Support: 2542; Body coverage: 20984, Confidence: 0.12113991612657263

After Insertion: Support: 21640; Body coverage: 41360, Confidence: 0.52321083172147

That means 18442 triples were inserted

## 8. Rule

R8: If authorA is coauthor of authorB, authorB is coauthor of authorA

Quality: high (confidence: 0.9891)

Horn Rule:

:authorA coauthor :authorB <= :authorB coauthor :authorA

Before Insertion Support: 24552; Body coverage: 24822, Confidence: 0.9891225525743292

After Insertion: Support: 25092; Body coverage: 25092, Confidence: 1.0

That means 270 triples were inserted

# Task 2: Association Rule Mining

This is a widely used technique for discovering interesting relationships between different entities in a dataset. It involves identifying frequent itemsets (combinations of entities that co-occur frequently) and then generating association rules between them. In the context of the given knowledge graph, this technique could be used to identify frequent co-occurrences of authors, papers, affiliations, domains, and conferences, and then generate rules based on these co-occurrences.

We used the java implementation of EMIE (https://github.com/dig-team/amie) to mine rules from the graph. In total, we got 19 rules, of which the best 5 will be shown here.

## 1. Rule

If a is coauthor of b, b is coauthor of a

This simply makes the coauthorship bidirectional, and is the same as the 8th handcrafted rule.

Confidence: High (0,989840348 PCA confidence)

Horn Rule:

?b  <http://lsdis.cs.uga.edu/projects/semdis/opus#coauthor>  ?a   => ?a  <http://lsdis.cs.uga.edu/projects/semdis/opus#coauthor>  ?b

Before insertion: Support: 24552; Body coverage: 24822, Confidence: 0.9891225525743292

After Insertion: Support: 25092; Body coverage: 25092, Confidence: 1.0

That means 270 triples were inserted

## 2. Rule

If g cites a and g appears in converence b then a also appears in conference b

Quality: Medium (0,363336334 PCA confidence)

Horn Rule:

?g  <http://purl.org/spar/cito/cites>  ?a  ?g  <https://makg.org/property/appearsInConferenceSeries>  ?b   => ?a  <https://makg.org/property/appearsInConferenceSeries>  ?b

Support: 1211; Body coverage: 3909, Confidence: 0.30979790227679715

Support: 4088; Body coverage: 5876, Confidence: 0.6957113682777399

That means 2698 triples were inserted

## 3. Rule

a is a member of b if a is a coauthor of h and h is a member of b

Quality: Medium (0,407460821 PCA confidence)

Horn Rule:

?a  <http://lsdis.cs.uga.edu/projects/semdis/opus#coauthor>  ?h  ?h  <http://www.w3.org/ns/org#memberOf>  ?b   => ?a  <http://www.w3.org/ns/org#memberOf>  ?b

Support: 4258; Body coverage: 13974, Confidence: 0.30470874481179333

Support: 15149; Body coverage: 38227, Confidence: 0.3962905799565752

That means 9716 triples were inserted

## 4. Rule

R4: a has discipline b, if a cites h and h has discipline b

Quality: Low (0,12311716 PCA confidence)

Horn Rule:

:a <http://purl.org/spar/cito/cites> :h :h <http://purl.org/spar/fabio/hasDiscipline> :b => :a <http://purl.org/spar/fabio/hasDiscipline> :b

Support: 2542; Body coverage: 20984, Confidence: 0.12113991612657263

Support: 21640; Body coverage: 41360, Confidence: 0.52321083172147

That means 18442 triples were inserted

## 5. Rule

a is of type b if a has coauthor g and g is of type b

This links the type Author to other Authors, if they have a coauthor

Quality: High (1 PCA confidence)

Horn Rule:

?g  <http://lsdis.cs.uga.edu/projects/semdis/opus#coauthor>  ?a  ?g  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  ?b   => ?a  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  ?b

Support: 7650; Body coverage: 7650, Confidence: 1.0

Support: 7650; Body coverage: 7650, Confidence: 1.0

That means 0 triples were inserted

# Task 4

For this task we used SHACL. After examning some examples and reading some of the documentation, we started putting together our own shapes.

We decided to use kind of an open word assumption in a sense, that we did not require a subject to have every possible predicate (that makes sense fot that subject), but if it has some predicates, then they should be of the correct class (that were specified in the assingment).

This way we found 3 errors:

- <http://MAGexample.org/093C4716> is referenced as a domain in a paper, but its type is not specified as a domain (paper: <http://MAGexample.org/7E65D12B>)

- <http://MAGexample.org/03DDEE22> suffers from the same issue (paper: <http://MAGexample.org/7E7BB75D>)

- <http://MAGexample.org/45701BF3 is referecnced as a conference but is not specified so (paper: <http://MAGexample.org/8115C904>)

If we restrict for example the ProperPapers class to have at least one author we find 140 errors instead of 3. This lead to us thinking that this graph is incomplete, and this is the reason that we went with our "open world assumption".