Distributed AMIE+

Preliminary Notes
of the Thesis Project

Omar Trinidad Gutiérrez Méndez

September 19, 2018

1 Introduction

Knowledge bases (KB) have the purpose of representing and store knowledge in a machine-readable format. Some well-known KBs are DBpedia [4], NELL, YAGO [6], or Freebase [1]. A usual task executed in these databases is mining logical rules, that is, find unknown relationships between entities.

However, these databases are designed under the idea of Open World Assumption (OWA), that means, if the database does not contain a fact, we are not assuming that this fact is false, as happens under the Closed World Assumption (CWA).

Finding these relations in huge datasets, and under the OWA setting is a challenging task. This problem was addressed by Galárraga et al. [3] who proposed Association Rule Mining under Incomplete Evidence (AMIE) and later suggested an improved version of the same method that they simple named AMIE+ [2].

The purpose of the current project is to explore AMIE+ and implement it in a distributed context.

1.1 Theoretical framework

Knowledge Bases KB are collections of facts; every fact is represented by a relation between a subject and object $r(s, o)$. In this work, we are focused on KBs modeled using the W3C standard Resource Description Framework (RDF). In RDF, the facts are represented as triples.

An *atom* is a fact with variables at the subject and/or object. A Horn rule is composed of a head and a body.

- The head is a single atom
- The body is a set of atoms

$$B_1 \wedge B_2 \wedge \ldots \wedge B_n \implies r(x, y) \tag{1}$$

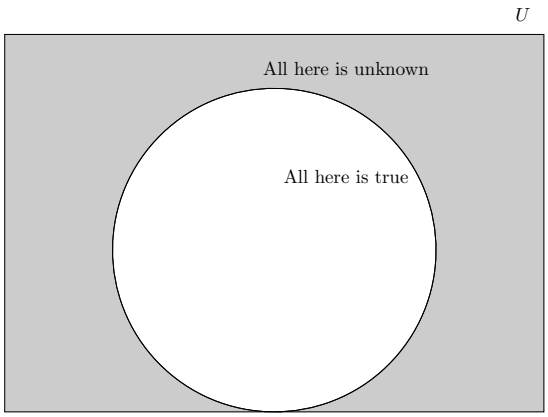

Figure 1: Flowchart of AMIE algorithm

1.1.1 Horn rules and language bias

In order to limit the size of the search space, AMIE uses constraints that are called *language bias*.

There is an aim to limit the size of the search space, with AMIE, for example, we use constraints on the structure of the mined rules, this is called language bias. The idea is to have good designed language bias to avoid to deal with an intractable search space but at the same time to generate more expressive rules.

- We aim for connectivity, two atoms in a rule are connected if they share a variable or entity. A rule is connected when every atom is connected transitively to the rest of atoms.
- The rules have to be closed.
- Also, reflexive rules are discarded.

1.1.2 Measures of significance

We want to use only rules above a given support threshold.

1.1.3 Incompleteness

As was stated before, the semantic KBs operate under the CWA, that is, we assume the facts in the database are known true facts, everything else, outside the database is asumed to be unknown.

Going beyond, we say that the unknown facts are either true or false facts, we want to predict the new ones.

1.2 Similar works

The task of finding new logical rules given a KB has been addressed from multiple angles. For example, ILP based approaches, relational machine learning or hybrid approaches.

One advantage, from AMIE over relational machine learning approaches, is that AMIE has better interpretability, which is a crucial in the Data Science world. So, with AMIE, it is possible to mine logical rules where there is a correlation in the data.

1.3 AMIE

Inside AMIE, the Partial Completeness Assumption (PCA) was used to guess the so-called counterexamples for rules.

In the PCA we say that a subject-relation pair is complete if it does exists an object that for that s,r.

Assume that if the KB knows relations for some subject then knows all the set of values.

1.4 AMIE plus

In AMIE+ it was aggregated pruning strategies and approximations that allowed to explore the search space more efficiently.

1.5 SANSA Stack

SANSA [5] is a platform whose purpose is...

References

[1] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008.

[2] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. Fast rule mining in ontological knowledge bases with AMIE + +. *The VLDB Journal*, 24(6):707–730, 2015.

[3] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*, pages 413–422. ACM, 2013.

[4] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick

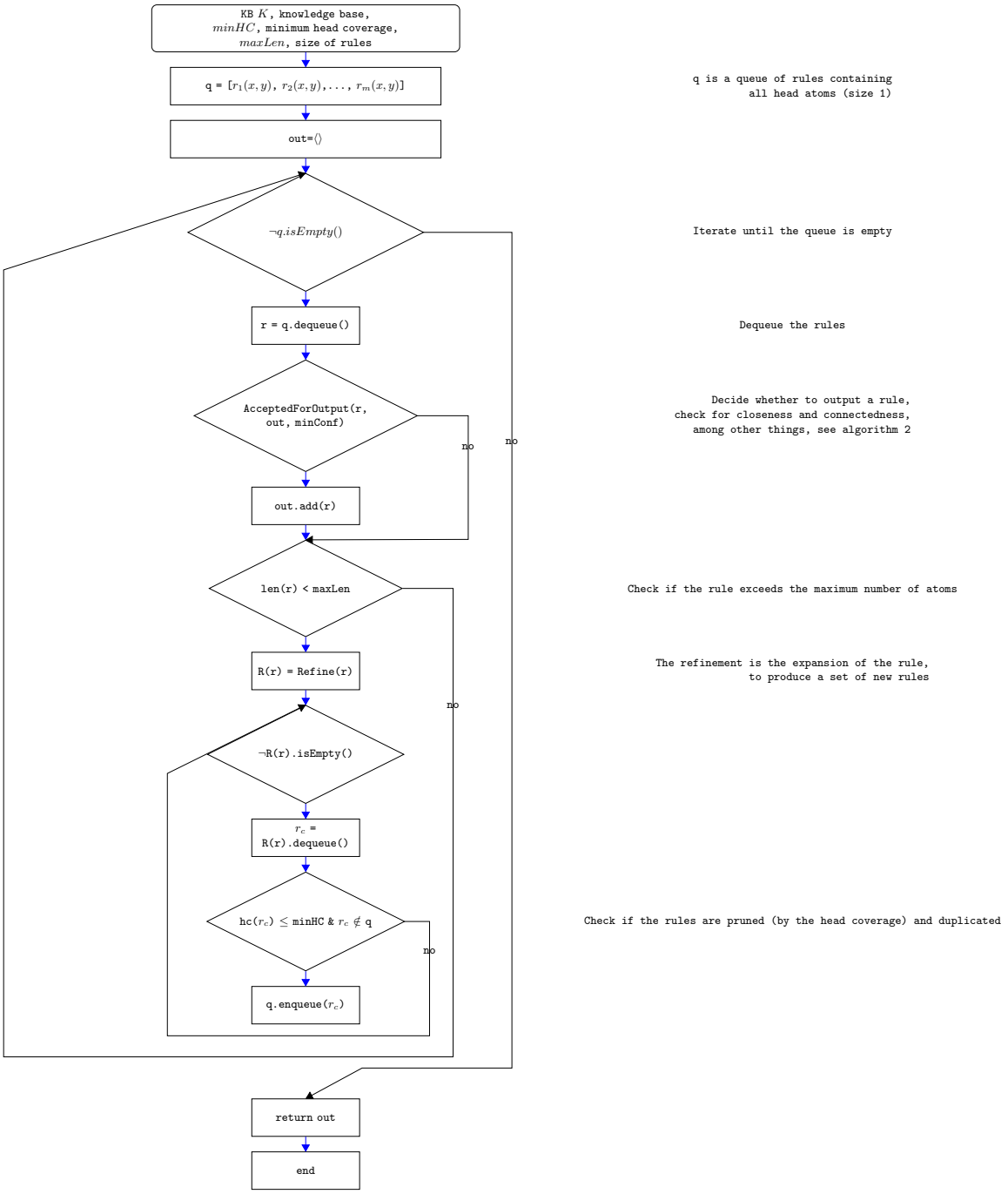

Figure 2: Flowchart of AMIE algorithm

van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.

[5] Jens Lehmann, Gezim Sejdiu, Lorenz Bühmann, Patrick Westphal, Claus Stadler, Ivan Ermilov, Simon Bin, Nilesh Chakraborty, Muhammad Saleem, Axel-Cyrille Ngomo Ngonga, and Hajira Jabeen. Distributed semantic analytics using the sansa stack. In *Proceedings of 16th International Semantic Web Conference - Resources Track (ISWC'2017)*, pages 147–155. Springer, 2017.

[6] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.