Applications of Rule Mining in Knowledge Bases

Luis Galárraga
Institute Mines Télécom
Télécom ParisTech, CNRS
luis.galarraga@enst.fr

ABSTRACT

The continuous progress of Information Extraction (IE) techniques has led to the construction of large Knowledge Bases (KBs) containing facts about millions of entities such as people, organizations and places. KBs are important nowadays because they allow computers to understand the real world and are used in multiple domains and applications. Furthermore, the discovery of useful and non-trivial patterns in KBs, known as rule mining, opens the door for multiple applications in the areas of data analysis, prediction and automatic data engineering. In this article we present an overview of our ongoing work on rule mining on KBs and some of its applications. The scale of current KBs as well as their inherent incompleteness and noise make this endevour challenging.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

knowledge bases; rule mining; ontologies

1. BIO

Luis Galárraga is a doctoral student in the DBWeb Group of the INFRES department at Télécom ParisTech. He completed his bachelor studies in Computer Engineering at Escuela Superior Politécnica del Litoral (Guayaquil, Ecuador) in 2008 and pursued a Master in Computer Science at Saarland University in 2009. For his master thesis he worked in the area of Distributed RDF Query Processing under the supervision of Katja Hose and Ralf Schenkel. In 2012 he was granted an IMPRS scholarship to pursue his PhD studies and joined the Ontologies Group, led by Fabian Suchanek, at the Max Planck Institute for Informatics in Saarbrücken to work in the area of Rule Mining on Ontological Knowledge Bases, under Fabian's supervision. His work in this topic led to the development and publication of a system called AMIE (Association Rule Mining Under Incomplete Evidence). The article published at the International World Wide Web Conference in 2013, was awarded the best student paper award. Since then, his work has been focused on improving the rule mining algorithms proposed in AMIE and apply rule mining techniques in different scenarios. His recent work on Canonicalization of Open Knowledge Bases was accepted this year at the ACM International Conference on Information and Knowledge Management. This work started during his internship at Google in the fall of 2013 and was developed in collaboration with Geremy Heitz and Kevin Murphy.

2. INTRODUCTION

In the last ten years there has been an increasing interest in Knowledge Bases (KBs), both in industry and academia. Projects such as YAGO [11], DBpedia [1], Wikidata, Wolfram Alpha, Knowledge Graph[1], Knowledge Vault [2], among others, store knowledge about millions of entities such as people, places, artifacts, etc., in a machine readable format, often RDF. This format represents knowledge as triples of the form ⟨*subject, relation, object*⟩ and, in contrast to natural language text, is easier for computers to process. KBs are therefore used to help computers understand the real world. For instance the Knowledge Graph (Google's KB) allows the Google search engine to better understand user queries by identifying real-world entities in the words of a query. This allows the search engine to deliver more semantic results instead of simply reporting pages matching the query. Furthermore, the plethora of information contained in today's KBs provides the opportunity to learn patterns and rules from the data. Such rules can help us understand the data for the sake of multiple applications. For instance, we could learn that Grammy awardees frequently play guitar or find out that often, people with children in common are married. Such rules provide a deeper understanding of the data domain and can help us, for example, predict missing information or automatically induce a schema from the data. Since current KBs contain up to billions of facts, extracting rules from such amount of data is a challenging task. The latter problem becomes worse due to the incompleteness and noise of web-extracted KBs. This happens because information extraction processes are not error-free and even reputed sources can contain information gaps.

In the next sections, we present an overview of our work on rule mining on KBs, starting with the description of a

PIKM'14, November 3, 2014, Shanghai, China.
ACM 978-1-4503-1481-7/14/11
http://dx.doi.org/10.1145/2663714.2668052 .

[1] `http://www.google.com/insidesearch/features/search/knowledge.html`

system (called AMIE) for rule mining on large, potentially incomplete KBs. We then discuss some applications of rules in tasks such as data prediction, schema alignment and KB canonicalization. We conclude with an outlook of our work.

$\textit{isMarriedTo}(x, y) \wedge \textit{livesIn}(x, z) \Rightarrow \textit{livesIn}(y, z)$
$\textit{isCitizenOf}(x, y) \Rightarrow \textit{livesIn}(x, y)$
$\textit{hasAdvisor}(x, y) \wedge \textit{graduatedFrom}(x, z) \Rightarrow \textit{worksAt}(y, z)$
$\textit{wasBornIn}(x, y) \wedge \textit{isLocatedIn}(y, z) \Rightarrow \textit{isCitizenOf}(x, z)$
$\textit{hasWonPrize}(x, \textit{G. W. Leibniz}) \Rightarrow \textit{livesIn}(x, \textit{Germany})$
$\textit{hasWonPrize}(x, \textit{Grammy}) \Rightarrow \textit{hasMusicalRole}(x, \textit{Guitar})$

Figure 1: Some Rules mined by AMIE on YAGO2

3. RULE MINING WITH AMIE

AMIE [5], which stands for Association Rule Mining Under Incomplete Evidence, is a system that efficiently learns closed Horn rules from (potentially) incomplete RDF KBs. A closed Horn rule has the form:

$$B_1, B_2 \ldots B_n \Rightarrow r(x, y), \text{written as } \vec{B} \Rightarrow r(x, y)$$

where each B_i and $r(x, y)$ are atoms. Atoms are triples where the relation is fixed and at least one of its arguments is a variable. The left-hand side of the rule is a conjunction of atoms and is called the *body*, whereas the right-hand side is the *head*. In a closed Horn rule all the variables occur at least twice, that is, no variable is existentially quantified. This constraint allows our rules to make concrete predictions when binding the variables to values in the KB. For instance, given the closed Horn rule:

$$\textit{hasChild}(x, y), \textit{isCitizenOf}(x, z) \Rightarrow \textit{isCitizenOf}(y, z)$$

and the facts *hasChild*(Barack Obama, Sasha Obama) and *isCitizenOf*(Barack Obama, USA), our example rule will predict *isCitizenOf*(Sasha Obama, USA).

Inductive Logic Programming (ILP) is the field that studies the methods to learn logical rules from a set of positive and negative examples. The goal of ILP is to find hypotheses that cover as many positive examples as possible and avoid the negative examples. Conceptually, ILP is not applicable to KBs because they do not store negative information. For this reason, different approaches resort to different ways to simulate negative evidence. An approach based on standard association rule mining assumes that any fact predicted by a rule and missing in the KB is false. This is known as the Closed World Assumption (CWA), which contrasts with the Open World Assumption (OWA) that KBs make. Under the OWA, if a fact is not present in the KB, it is labeled as unknown. The approach presented in [9] constructs counterexamples in a random fashion. To account for the incompleteness of some relations, AMIE simulates counterexamples by means of the Partial Completeness Assumption (PCA). The PCA states that if a KB knows some r-values for an entity x, then it knows all its r-values and any other value is assumed false. If for instance a KB knows the birthday of a person, any prediction that differs from the value in the KB is assumed as counter-evidence. In contrast, if the birthday of the person is unknown, the PCA will not use predictions as counter-examples. In practice the PCA is a reasonable assumption. It is perfectly sound for functions (birthday, place of birth) and works well for quasi-functions (e.g. citizenship) and relations extracted from well-documented sources.

Rules have scores associated to them. The support of a rule is defined as the number of positive examples covered by the rule:

$$\textit{support}(\vec{B} \Rightarrow r(x, y)) := \#(x, y) : \exists z_1, ..., z_m : \vec{B} \wedge r(x, y)$$

The standard confidence is the ratio of positive examples with respect to all positive and negative examples according to the CWA:

$$\textit{stdconf}(\vec{B} \Rightarrow r(x, y)) := \frac{\textit{support}(\vec{B} \Rightarrow r(x, y))}{\#(x, y) : \exists z_1, ..., z_m : \vec{B}}$$

Since this metric uses missing information as counter-evidence, AMIE resorts to the PCA to define an improved confidence metric, the PCA confidence:

$$\textit{pcaconf}(\vec{B} \Rightarrow r(x, y)) := \frac{\textit{support}(\vec{B} \Rightarrow r(x, y))}{\#(x, y) : \exists z_1, ..., z_m, y' : \vec{B} \wedge r(x, y')}$$

This new metric normalizes the support over the number of positive examples and the negative examples assumed by the PCA.

AMIE runs orders of magnitude faster than state-of-the-art ILP solutions. It can run on a subset of 1M facts from YAGO2 in less than 4 minutes, whereas other approaches could not even run or took hours. To achieve this performance, AMIE relies on a highly concurrent implementation, a set of mining operators and an tailored in-memory database to efficiently explore the search space of closed Horn rules. AMIE's mining operators produce new rules by adding atoms in different ways to existing rules. Since every new atom cannot increase support, i.e., this metric is monotonic, the AMIE algorithm stops adding atoms once a rule pattern is below a given minimum support threshold. This is a way to prune the search space and has the biggest impact in runtime. Although confidence is not used for pruning in this way (it is not monotonic), it can be used to reduce the number of rules output by the system. If the refinement of a rule has lower confidence than its parent rule, then the more specific rule is not output. Table 1 shows some rules mined by AMIE on YAGO2.

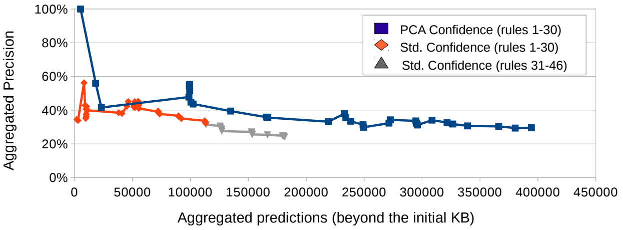
4. APPLICATIONS OF RULE MINING

In this section we discuss some applications of Rule Mining on Knowledge Bases using our system AMIE.

4.1 Data Prediction

Web-extracted KBs often contain gaps, either because they inherit those gaps from their sources or because the information extraction process contains bugs. Albeit not 100% precise, the rules found by AMIE encode regularities that are frequently true and can therefore be used as signals for data prediction. In [5], we carried out an experiment to compare the suitability of the PCA confidence as a confidence metric, in comparison with the standard confidence. The experiment was aimed to determine whether the rules with high PCA confidence are good predictors of missing information in KBs. To show this, we ran AMIE on YAGO2 and took the top 30 rules ranked by PCA confidence and standard confidence. We then used the rules to make predictions. We verified a random sample of such predictions, automatically (in a newer version of the KB) or manually in Wikipedia. We then plotted the cumulative number of predictions (shown in Figure 2) against the estimated cumulative precision and found that the top 30 rules ranked by PCA confidence produce many more predictions at a higher precision that even the top 45 rules ranked by standard confidence. This happens because the standard confidence is a conservative metric that penalizes rules with incomplete head relations (due to the CWA) whereas the PCA is not affected by this incompleteness, leaving room for more predictions. Our results suggest that the standard confidence metric is more suitable for data description, while the PCA confidence is more suitable for prediction in KBs operating under the OWA.

Figure 2: Std. Confidence vs. PCA Confidence



It is important to notice that the cumulative precision of our rules is around 40%. Predicting concrete facts is a very difficult task, so it is clear that our rules cannot be directly used to push more facts into KBs. The predictions made by rules are, however, plausible statements that could be sent to human assessors for verification or serve as input for a statistical model to learn confidence scores for predictions. Recall that a statement can be predicted by multiple rules, therefore those rules can be used as signals for a machine learning classifier to learn an aggregate confidence score based on multiple evidence in the data. We leave the exploration of such methods as future work.

4.2 Ontology Alignment

Some of the publicly available KBs in the Semantic Web overlap in their entities. Initiatives such as the Linked Open Data are the first step towards the integration of different RDF data sources. This integration takes the form of instance alignments, expressed as *owl:sameAs* links between entities of different KBs. However, instance alignments are not enough for a full data integration in the way it was envisioned in the Semantic Web. For example, Freebase can express that two people are siblings with a single relation, e.g., *sibling*(Michael Jackson, Janet Jackson), whereas YAGO needs two facts to express the same idea, e.g., *hasChild*(Joseph Jackson, Janet Jackson), *hasChild*(Joseph Jackson, Michael Jackson). An interoperable web of data would require an alignment of the schema of Freebase and YAGO, e.g., state that $F{:}sibling(x, y) \equiv Y{:}hasChild(z, y)$, $Y{:}hasChild(z, x)$. While it seems feasible to manually write schema mappings for a given pair of KBs, this task becomes impractical at the scale of all KBs in the Semantic Web. The approach introduced in [6] makes use of rule mining techniques in combination with the plethora of instance alignments available in the Semantic Web to tackle the problem of automatic schema aligment. Given two RDF KBs $\mathcal{K}$ and $\mathcal{K}'$, a ROSA[2] rule is defined as a cross-ontology rule, where the body contains atoms of relations in $\mathcal{K}$ and the head is a relation in $\mathcal{K}'$. Table 1 includes the list of ROSA alignments proposed in [6]. This definition is asymmetric in the sense that we can mine ROSA rules from $\mathcal{K}$ to $\mathcal{K}'$ and from $\mathcal{K}'$ to $\mathcal{K}$. ROSA rules define one class of cross-ontology alignments that are easy to learn with our rule mining machinery. To mine ROSA rules from a pair of ontologies, we coalesce the KBs that we want to align and run AMIE on the new coalesced KB. Given KBs $\mathcal{K}$ and $\mathcal{K}'$, we construct such coalesced KB as follows:

$r(x, y) \Rightarrow r'(x, y)$	(R-subsumption)
$r(x, y) \Leftrightarrow r'(x, y)$	(R-equivalence)
$type(x, C) \Rightarrow type'(x, C')$	(C-subsumption)
$r_1(x, y), r_2(y, z) \Rightarrow r'(x, z)$	(2-Hops alignment)
$r(z, x), r(z, y) \Rightarrow r'(x, y)$	(Triangle alignment)
$r_1(x, y), r_2(x, V) \Rightarrow r'(x, y)$	(Specific R-subsumption)
$r(y, V) \Rightarrow r'(x, V')$	(Attr-Value translation)
$r_1(x, V_1), r_2(x, V_2) \Rightarrow r'(x, V')$	(2-Value translation)

Table 1: ROSA Rules ($r \in rel(\mathcal{K}), r' \in rel(\mathcal{K}')$).

Rule	Conf.
$D{:}musicalArtist(x, y) \Rightarrow Y{:}created(y, x)$	90%
$Y{:}directed(x, y) \Leftrightarrow D{:}director(y, x)$	98%
$Y{:}type(x, Y{:}Site) \Rightarrow D{:}type(x,\ D{:}PopulatedPlace)$	97%
$Y{:}wasBornIn(x, y),\ Y{:}label(y, z) \Rightarrow I{:}bornIn(x, z)$	37%
$Y{:}child(x, y),\ Y{:}child(x, z) \Rightarrow F{:}sibling(y, z)$	37%
$Y{:}graduated(x, y),\ Y{:}type(y, Univ) \Rightarrow F{:}institute(x, y)$	98%
$Y{:}locatedIn(x, Italy) \Rightarrow D{:}timeZone(x,\ CET)$	100%
$F{:}type(x, Royal),\ F{:}gender(x, fem) \Rightarrow Y{:}type(y, Princess)$	55%

Table 2: Examples of ROSA mappings between YAGO, DBpedia, Freebase and IMDb.

$$\hat{\mathcal{K}} = \{\hat{r}(\sigma(x), \sigma(y)) \mid r(x, y) \in \mathcal{K} \wedge \hat{r} = t(r)\} \cup \mathcal{K}'$$

σ is a substitution function, that maps the instances of $\mathcal{K}$ to the *sameAs* counterparts from $\mathcal{K}'$ if any, or to themselves otherwise. It leaves literals unchanged. As pointed out in [10], different KBs may use the same relation (as given by a URI) in different ways. Therefore, we use a substitution t that substitutes all relation names in $\mathcal{K}$ so as to make sure they are different from the relation names in $\mathcal{K}'$. Our coalesce definition entails that we have two sets of relations: $rel(\mathcal{K}) = t(\pi_{relation}(\mathcal{K}))$ and $rel(\mathcal{K}') = \pi_{relation}(\mathcal{K}')$. Our coalesced KB subsumes both KBs. Table 2 provides examples of ROSA rules between YAGO, DBpedia, Freebase and a crawl of IMDb(used in [10]). The value in the second column is the PCA confidence of the rule. Since equivalences rules are the combination of two subsumptions rules, we report the minimum PCA confidence of the individual subsumptions as their score.

4.3 Canonicalization of Open KBs

KBs such as YAGO, DBpedia or Freebase contain information extracted from well-structured sources. For instance, YAGO combines information from the Wikipedia in-

[2] Rule for Ontology Schema Alignment

Verb phrases	Freebase relation
be an abbreviation for, be known as, stand for, be an acronym for	-
be spoken in, be the official language of, be the national language of	`location.country.official_language`
be bought, acquire	`organization.organization.acquired_by`
be the birth place of, be the hometown of	`ns:location.location.people_born_here`
be a borough located in, be a borough in	`ns:location.hud_county_place.county`

Table 3: Examples of clusters of verbal phrases. Some of them were mapped to Freebase.

foboxes and categories, Wordnet and Geonames. While this paradigm for Information Extraction (IE) guarantees high precision for the extracted facts, its recall is limited since a lot of factual information in Wikipedia is only available in the articles' text. This standard way to extract information is often known as Closed IE. Standard Closed IE extractors collect facts from the sources according to the schema of the target KB and therefore require a set of hand-crafted rules and mappings as additional input. These can be regular expressions or simple mappings from labels to relations. If the sources change or new sources are included, those mappings must be updated accordingly.

On the other side of the spectrum, open IE projects such as Reverb [3] can extract triples from natural language text without additional input from the user. For example, given the sentence "McCain fought hard against Obama, but finally lost the election", an Open IE system will extract two triples, ⟨*McCain, fought against, Obama*⟩ and ⟨*McCain, lost, the election*⟩. The triples can contain arbitrary named entities, such as *Obama*, but also common noun phrases, such as *the election*. The predicate can be any sequence of words that appear between the two arguments. This basic approach can harvest a huge number of triples from Web corpora, even though some will be uninformative like in ⟨*McCain, lost, the election*⟩ (which election?). Nevertheless, Open IE extractors offer an attractive opportunity to either construct new KBs or improve the recall of the existing ones. The major disadvantage of open IE is that extractions are "dirty", i.e., noun and verbal phrases are not canonicalized. This is a problem for any application relying on the data. For example, if a user requires all the information about "Barack Obama", any triple referring to him as "President Obama" or "Obama" will be omitted in the result. This particular scenario requires to all possible ways to refer to "Barack Obama". Existing approaches for entity linking and synonym resolution [7] can solve this problem by linking noun phrases to (already canonicalized) entities in a KB or identifying synonyms from features taken from the extractions and the sources. These solutions still leave the problem of canonicalizing the verbal phrases. A user querying the places of residence of a person will have trouble in formulating the query since open IE triples may express such relation in multiple ways, e.g., "lives in", "resides in", etc.

Rule mining techniques offer a simplistic and elegant alternative to canonicalize verbal phrases in open KBs. In [4], we canonicalize the relations of a significant percentage (up to 33%) of the triples of a set of 1.3M Reverb extractions from ClueWeb09[3]. Our method assumes subjects and objects have been canonicalized somehow, for example by linking them to Freebase using the method described [7]. Our approach then runs AMIE on the semi-canonicalized KB in order to extract a set of highly confident verbal phrase mappings. Examples are:

stand-for(x, y) ⇔ *be-an-acronym-for(x, y)*
be-short-for(x, y) ⇔ *be-an-acronym-for(x, y)*
refer-to(x, y) ⇔ *be-short-for(x, y)*

Since the equivalence relation is transitive, the mappings are then transitively grouped, producing clusters of verbal phrases with close meaning. From the example, our approach would group the phrases *stand-for*, *be-an-acronym-for*, *short-for* and *refer-to* into a single cluster. With this simple method (using a confidence threshold of 0.8 PCA confidence), we found approximately 500 clusters of synonym verbal phrases with 90% macro-precision[4] covering 15% of the triples of the KB. The clusters can be canonicalized by selecting a representative relation and replacing all occurrences of the phrases in the cluster with the representative. An alternative for canonicalization is to find the representative in a KB by linking the clusters to existing relations. This is an ontology alignment task, which we solve using the ROSA alignments as described in Section 4.2. We coalesce our semi-canonicalized KB with Freebase and mine cross-ontology equivalences of the form *rv(x, y)* ⇔ *rf(x, y)* where *rv* is a Reverb verbal phrase and rf is a Freebase relation. Then, for each cluster of synonym verbal phrases, we gather all Freebase mappings of verbal phrases in the cluster. We could unambiguously map to Freebase, up to 25% of the clusters of verbal phrases. Some clusters and their Freebase representative relations are listed in Table 3.

Our approach also accounts for the polysemy of some verbal phrases. For example, the phrase *belongs-to* conveys different meanings in the sentences "The Wii belongs to Nintendo" (*invention created by organization*) and "Mallorca belongs to Spain" (*island belongs to country*). Since polysemous phrases hurt the precision of our clusters, we implemented a variant of our verbal phrase canonicalization, inspired on the work described in [8], where we configured AMIE to mine equivalence mappings augmented with data types. This trick allows to treat the relations *belongs-to: invention* ⇒ *organization* and *belongs-to: island* ⇒ *country* as separate relations. Since each verbal phrase can produce multiple type-augmented relations, we restricted this feature to the most common Freebase relations: person, organization, location and word. This enhacement increased all our precision metrics significantly, e.g., macro-precision increased from 90% to 95%.

5. CONCLUSIONS AND OUTLOOK

In this summary, we have described how to mine logical rules from RDF KBs with AMIE and how to use such rules for different applications. Closed Horn rules provide

[3] `http://www.lemurproject.org/clueweb09.php/`

[4] A cluster is pure if all its verbal-phrases have a close meaning according to our human evaluators. The macro precision measures the percentage of pure clusters.

an insight of the patterns that govern the data. For example, confident rules can be used to predict missing or future facts. They can be used for data integration and maintanance tasks such as ontology alignment and canonicalization of KBs. Note however that our work has been restricted to closed Horn rules, which are only one class of logical rules, and have therefore their limitations. Applications in the area of automatic schema induction (in the context of OWL statements) rely on constraints with existentially quantified variables, negations, cardinality and inequality constraints, which are beyond AMIE's language bias. Applications such as the analysis of trends in history or recommendation systems may require to account for the temporal dimensions of facts, something our rules do not consider so far. Finally, our work in data prediction has potential applications in probabilistic databases since (as suggested in Section 4.1), they can serve as the building block for a model to accurately estimate the likelihood of predictions drawn from statistical evidence.

6. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *ISWC*, 2007.

[2] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *KDD*, 2014.

[3] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, 2011.

[4] L. Galárraga, G. Heitz, K. Murphy, and F. Suchanek. Canonicalizing Open Knowledge Bases. In *CIKM*, 2014.

[5] L. Galárraga, C. Teflioudi, K. Hose, and F. Suchanek. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In *WWW*, 2013.

[6] L. A. Galárraga, N. Preda, and F. M. Suchanek. Mining rules to align knowledge bases. In *AKBC*, 2013.

[7] T. Lin, Mausam, and O. Etzioni. Entity linking at web scale. In *AKBC-WEKEX*, 2012.

[8] B. Min, S. Shi, R. Grishman, and C. Y. Lin. Ensemble semantics for large-scale unsupervised relation extraction. In *EMNLP-CoNLL*, 2012.

[9] S. Muggleton. Learning from positive data. In *ILP*. Springer-Verlag, 1997.

[10] F. M. Suchanek, S. Abiteboul, and P. Senellart. Paris: Probabilistic alignment of relations, instances, and schema. *PVLDB*, 5(3), 2011.

[11] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A Large Ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 2008.