Engineering Science and Technology, an International Journal 43 (2023) 101417

HOSTED BY

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Full Length Article

An interestingness measure for knowledge bases

Damla Oguz [a,*], Fatih Soygazi [b]

[a] Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey
[b] Department of Computer Engineering, Aydın Adnan Menderes University, Aydın, Turkey

ARTICLE INFO

Article history:
Received 8 November 2022
Revised 10 February 2023
Accepted 17 April 2023
Available online 30 May 2023

Keywords:
Knowledge base
Data mining
Rule mining
Interestingness measure
Confidence

ABSTRACT

Association rule mining and logical rule mining both aim to discover interesting relationships in data or knowledge. In association rule mining, relationships are identified based on the occurrence of items in a dataset, while in logical rule mining, relationships are determined based on logical relationships between atoms in a knowledge base. Association rule mining has been widely studied in transactional databases, mainly for market basket analysis. Confidence has become the most widely used interesting measure to assess the strength of a rule. Many other interestingness measures have been proposed since confidence can be insufficient to filter negatively associated relationships. Recently, logical rule mining has become an important area of research, as new facts can be inferred by applying discovered logical rules. They can be used for reasoning, identifying potential errors in knowledge bases, and to better understand data. However, there are currently only a few measures for logical rule mining. Furthermore, current measures do not consider relations that can have several objects, called quasi-functions, which can dramatically alter the interestingness of the rule. In this paper, we focus on effectively assessing the strength of logical rules. We propose a new interestingness measure that takes into account two categories of relations, functions and quasi-functions, to assess the degree of certainty of logical rules. We compare our proposed measure with a widely used measure on both synthetic test data and real knowledge bases. We show that it is more effective in indicating rule quality, making it an appropriate interestingness measure for logical rule evaluation.

1. Introduction

Association rule mining (ARM) is a well-researched data mining area which aims to find interesting rules among sets of items in a database. Although a data mining system can generate many association rules, only a few of them represent knowledge. For this reason, finding interesting association rules by using objective measures is vital. There are several objective measures of rule interestingness, among which support and confidence are the two widely used ones. Support is the percentage of occurrences of an itemset. An itemset becomes frequent when its support is equal or greater than a predefined threshold. After finding all frequent itemsets, the association rules are generated according to a predefined minimum confidence threshold. Thus, confidence assesses the degree of certainty of the detected association [1]. For an association rule of the form $X \Rightarrow Y$, the support measure represents the percentage of transactions that X and Y occur together in the database. The confidence measure is calculated by the probability that a transaction containing X also contains Y and, represents the strength of the rule. Besides transactional databases, the number of knowledge bases (KBs) increases along with the advances in information extraction. DBpedia [2], Wikidata [3], YAGO [4,5], the New York Times[1], BBC [6], NELL [7] and FreeBase[2] are some of the well-known projects that contain numerous entities and facts in different domains. KBs involve a lot of facts in different domains and these facts keep logical rules implicitly. Consider a KB which contains facts such as "London is the biggest city in England," "Paris is the biggest city in France," "Rome is the biggest city in Italy," "London is the capital of England," "Paris is the capital of France," "Rome is the capital of Italy." From these facts a logical rule that asserts "The biggest city of a country is its capital city" can be mined. This logical rule can be expressed as $biggestCity(city, country) \Rightarrow capitalCity(city, country)$. Such logical rules can be used for different purposes. First of all, new facts can be inferred by applying these rules. They can be used for reasoning

* Corresponding author.
E-mail addresses: damlaoguz@iyte.edu.tr (D. Oguz), fatih.soygazi@adu.edu.tr (F. Soygazi).

[1] http://data.nytimes.com/
[2] http://freebase.com

https://doi.org/10.1016/j.jestch.2023.101417
2215-0986/

and to understand data better. Furthermore, these logical rules can help to identify the potential errors in the KB [8].

Although there is an analogy between ARM in transactional databases and mining logical rules in KBs, an association rule and a logical rule are substantially different. Association rules present interesting rules from the frequently co-occurring items. Logical rules describe common relationships among instances through atoms in the KB. Hence, using the same measures is not possible. Another reason is related to the nature of data completeness. A frequent itemset typically refers to a set of items that often occur together in a transactional database, such as milk and bread which are frequently bought together in a market database [1]. In that case, the frequent itemsets are mined in a complete data set. On the other hand, the KBs operate under the open-world assumption (OWA) which means that a statement that is not contained in the KB just means unknown instead of false [9]. So, logical rules are mined in an incomplete data set. As a result, proposing new measures for KBs is a fundamental need.

There are a number of studies which focus on mining logical rules from KBs [8–15]. To the best of our knowledge, AMIE [9] is the first study which proposed a new confidence measure for KBs, called PCA (Partial Completeness Assumption) confidence. This new measure calculates the confidence of a rule by considering the unknown facts of KBs. Although PCA confidence works well for functions, it does not hold by definition for relations that can have several objects per subject [8]. These relations are called quasi-functions. Consider a rule stating that two persons having the same ancestor are siblings, $hasChild(x,y) \wedge hasChild(x,z) \Rightarrow sibling(y,z)$. $hasChild(x,y)$ and $sibling(y,z)$ are quasi-functions since a person can have more than one child and a person can have multiple siblings. As a downstream application example, let's assume there is a research ecosystem application that evaluates various criteria of universities and researchers. Suppose that the inbreeding rate is a matter of consideration and the rule $educatedAt(x,y) \Rightarrow employer(x,y)$ is used. To measure the rate of inbreeding effectively, a measure is required within the scope of this application. Quasi-functions are crucial for this evaluation as a person may have received education from multiple universities. The standard confidence and PCA confidence measures do not consider this issue and for this reason they undervalue the strength of a logical rule.

In this paper, we present a new measure for KBs which considers both the functions and the quasi-functions in a logical rule. The main objective of the proposed measure is to evaluate the strength of a rule more appropriately by grouping the subjects in quasi-functions. Thus, the strength of a rule is not unfairly affected when there are quasi-functions in a rule. We compare our proposed measure with the PCA confidence on a synthetic test data and on real KBs. Performance evaluation shows the effectiveness of the proposed measure when there are different types of atoms in logical rules. To the best of our knowledge, there are no studies that propose a measure to evaluate the strength of a logical rule in KBs by considering the impact of quasi-functions. Throughout the paper, the terms "measure" and "metric" are used interchangeably since some studies use measure and the others prefer metric.

The rest of the paper is organized as follows: Section 2 and Section 3 cover respectively the related work and the preliminaries. Section 4 introduces our measure for all combinations of functions and quasi-functions in the body of Horn rules. Section 5 presents the results and discussions on performance evaluation. Finally, we conclude the paper and give remarks for the future work in Section 6.

2. Related Work

As the amount of data increases, the analysis of transactional databases has become an important issue. Consequently, [16] present a new data mining functionality called association rule mining that aims to find interesting rules among items in a dataset. Since some interestingness measures are essential in discovering association rules, two objective measures, namely support and confidence, are introduced by [17]. For association rules of the form $X \Rightarrow Y$, support represents the percentage of transactions in a transactional database that the given rule satisfies, while confidence assesses the degree of certainty of the discovered association rule. Suppose an association rule like: $computer \Rightarrow mouse[support = 3\%, \ confidence = 80\%]$. This support value means that 3% of all the transactions show that *computer* and *mouse* are purchased together. On the other hand, 80% confidence value means that 80% of customers who purchased a *computer* also purchased a *mouse*. In other words, if a customer buys a *computer*, there is an 80% chance that he or she will buy a *mouse* as well. Although a huge number of association rules can be generated from a database, they are considered as interesting according to a predefined minimum support and confidence thresholds. That is, association rules are uninteresting and discarded if they do not satisfy both a minimum support threshold and a minimum confidence threshold [1].

Interestingness measures for association rule mining have been studied extensively [18–24]. Besides support and confidence, a great number of interestingness measures have been proposed for efficient discovery of association rules such as lift, correlation, collective strength, and cosine. Tan et al. [18] present a comparative study of twenty-one of these existing measures and show their strengths and weaknesses for different application domains. The main idea in [18] is that all measures are not equally successful for capturing the correlations among variables. There is not a universal measure that is always better than others in all cases. Sharma et al. [23] also focus on interestingness measures and claim that selecting optimal interestingnes measures is still an open research problem.

Although traditional researches that focus on interestingness measures have been studied for decades, the quality metrics for the rules in KBs are more recent. Galárraga et al. [8,9] and Lajus et al. [14] propose a rule mining approach called as AMIE, AMIE + and AMIE3, respectively. They also propose a quality metric called PCA confidence to measure the success of AMIE versions. This metric calculates the quality of the rules under the OWA. Although versions of AMIE outperform other rule mining systems in terms of quality of the mined rules, they ignore the impact of quasi-functions in the logical rules. In our paper, we focus on the quality of the logical rules by considering the impact of two categories of relations, namely functions and quasi-functions. Since our proposed metric and PCA confidence are related, PCA confidence is explained in Section 4.

Soft confidence [10] is another measure which also works under OWA. It can be calculated when the entity type information is available. Some studies [11–13] focus on rule quality in KBs as well and present some rule quality metrics. However, they do not consider quasi-functions in logical rules neither. To the best of our knowledge, our proposed metric is the first logical rule quality metric for KBs that considers impact of quasi-functions.

3. Preliminaries

This section explains the preliminary definitions of the concepts used in the next sections of the paper.

Atom. An atom is represented as *p(x,y)* where *x,y* are either constants or variables. For the sake of simplicity, variables are shown in lowercase letters while individuals are represented by first letters uppercase. P is the set of atoms where $p(x_n, y_n) \in P$.

Rule & **Horn Rule** & **Closed Rule.** A rule $r \in R$ is composed of a body and a head where body implies head, $B \Rightarrow H$. A Horn rule is a type of rule where the body part of the implication involves atoms in a conjunctive form, $B = p_1(x_1, y_1) \wedge p_2(x_2, y_2) \wedge \ldots \wedge p_n(x_n, y_n)$ and head is constituted with a single atom $H = q(x_k, j_l)$.

A variable is closed if it exists at least twice in a rule and for this reason a rule is defined as closed if its all variables are closed. For example, $wasBorn(x, y) \Rightarrow livesIn(x, y)$ where x and y are closed variables. A unified example might be $wasBorn(John, NewYork) \Rightarrow livesIn(John, NewYork)$ where *John* and *NewYork* are the closed variables used in the given closed rule.

Knowledge Base (KB) & **RDF KBs.** Knowledge bases create an organized set of data that includes a semantic model which differs from databases. Classes and instances for representing the knowledge and rules for inferencing on that knowledge are defined with respect to a formal model.

In this paper, we are interested in RDF KBs. A knowledge base includes a set of the *facts*, with a subject $s \in E$, a relation $p \in P$ and an object $o \in E \cup L$ where E denotes the entity set and L denotes the literal set. Then a fact is of the form *p(s,o)*.

A knowledge base is created by a set of facts (atoms) and rules, $KB = P \cup R$. When the facts show the explicit knowledge, the rules that are grounded with the facts help us to access implicit knowledge.

RDF KBs are specifically called as knowledge graphs. Although the ontologies involve the classes and the relationships among the classes to develop a general conceptual model, a knowledge graph is the representation of semantic data in instance level. In a traditional knowledge base, we need some inference algorithms to capture implicit knowledge whereas we can apply graph traversal based or graph embedding methods to capture knowledge in a knowledge graph. In this paper, we focus on RDF KBs (or knowledge graphs) that is commonly used on the Web and our method can be applied to any knowledge graphs on the web to make an analysis with respect to an interestingness measure.

Function & **Quasi-function.** A relation p is a function in KB, if p has at most one object for each subject. For example, $motherOf(Alice, Kate)$ describes that Alice is the mother of Kate. As each human has just one mother, *motherOf* relation is a function. Some relations are quasi-functions where each subject can be related with multiple objects. For example, some people can graduate from more than one university and so *graduatedFrom* relation is a quasi-function.

Support & **Confidence** Support of a Horn rule is the number of facts that B and H occur together in the knowledge base at the same time. $support(B \Rightarrow H)$=number of instances that are applied to the rules where $B \Rightarrow H$ is applied to the individualized atoms in the knowledge base.

Confidence is another important and widely used measure for Horn rules, which assesses the degree of certainty of a rule. Since we propose a confidence measure for Horn rules, the details are explained in 4.1, where we present our proposal.

4. Proposed Confidence Measure for Knowledge Bases

This section first explains the two widely used measures for KBs which are standard confidence and the PCA confidence. Second, we present our proposed metric, the COR (Categories Of Relations) confidence, which considers two categories of relations, namely the functions and quasi-functions in a Horn rule. The COR confidence measure also handles both categories of relations in the head atom. Lastly, we show the strength of the COR confidence for i) all possible combinations of functions and quasi-functions in the body of a Horn rule, and ii) functions and quasi-functions in the head of a Horn rule.

4.1. Background

We consider closed Horn rules of the form $B_1 \wedge B_2 \wedge \ldots \wedge B_n \Rightarrow r(x, y)$ where the left side denotes the body and the right side denotes the head of the rule. The head consists of only one atom while the body consists of a set of atoms, and thus, the rule can be shown as $B \Rightarrow r(x, y)$.

One of the objective measures for Horn rules is confidence, which assesses the degree of certainty of the discovered rule. Standard confidence of a rule is the ratio of the number of known facts that are implied by the rule to the facts in the KB. More formally, standard confidence is defined as in Eq. 1 where $z_1, \ldots, z_n$ are the variables of the rule apart from x and y [8]. Although standard confidence suits well for association rule mining in traditional databases, it is not appropriate for OWA because it assumes that all true facts are in the KB, but contrary to that assumption, there can be unknown facts in it. For this reason, [9,8] have proposed the PCA confidence measure as defined in Eq. 2 where $r(x, y\prime)$ are the facts that are false for the rule in the KB. Thus, PCA confidence is the ratio of number of known true facts that are implied by the rule to the number of true and false facts in the body for bounded subjects. In other words, unknown facts do not affect the PCA confidence measure. Although PCA confidence evaluates the rules under OWA, it does not consider the impact of quasi-functions in the body.

The value of the denominator of any confidence formula represents the total number of facts that should be considered to evaluate the strength of a rule. This value increases as the number of objects belonging to a relation increases. It can also be said that a good measure should consider the categories of relations in the body of a Horn rule. Standard confidence and PCA confidence formulas ignore the effect of quasi-functions in the body. For this reason, we propose a confidence measure that considers different types of atoms, namely categories of relations, in a rule. The value of the denominator of the proposed measure is calculated by considering both the functions and the quasi-functions in the body of a rule. With this approach, the proposed measure is not affected by the number of objects belonging to a relation. Furthermore, the proposed measure considers whether the head atom is a function or a quasi-function.

$$standard\_conf(B \Rightarrow r(x, y)) = \frac{support(B \Rightarrow r(x, y))}{\#(x, y) : \exists z_1, \ldots, z_n : B} \quad (1)$$

$$PCA\_conf(B \Rightarrow r(x, y)) = \frac{support(B \Rightarrow r(x, y))}{\#(x, y) : \exists z_1, \ldots, z_n, y\prime : B \wedge r(x, y\prime)} \quad (2)$$

4.2. The COR Confidence Measure

We propose Eq. 3, namely the COR confidence formula, to evaluate the strength of a Horn rule. This confidence formula is the ratio of the number of known facts that are implied by the rule to the number of groups of subjects according to the quasi-functions in the rule. This confidence measure not only evaluates the rules under OWA, but also considers the impact of quasi-functions in the rule. The algorithm for calculating the number of groups of subjects is depicted in Algorithm 1. First, we bind the atoms in the body to generate all possible combinations which are called *n-variable bindings*. Then, we decide the grouping set subjects according to the quasi-functions. We order the atoms according to the symmetric property in such a way that asymmetric ones are listed first. We check each atom respectively if it is a quasi-function or not, and add its subject to the grouping set if it is a quasi-function. We also add an atom's object to the grouping set if i) it is a quasi-function and symmetric, and ii) its subject

has been already added to the grouping set. Third, we group the generated *n-variable bindings* in the first step according to the grouping set found in the second step. Then, the groups are reconsidered if the atom in the head of the rule is a quasi-function. The groups are decomposed according to the objects in the head. Finally, the number of groups are counted, and is used as the denominator of the COR confidence formula. In the worst case, when the head atom is a quasi-function, the time complexity of COR confidence is $O(n * m)$ where n is the total size of the atoms in the body of the rule and m is the generated groups according to the subjects in the head atom.

$$COR\_conf(B \Rightarrow r(x,y)) = \frac{support(B \Rightarrow r(x,y))}{\#groups\_acc\_to\_quasifunctions} \quad (3)$$

Algorithm 1: Calculating number of groups according to the quasi-functions

Input: Head of a Horn rule $r(x,y)$where $x \in X$ and $y \in Y$, Body of a Horn rule $B = \{B_1 \wedge B_2 \wedge B_3 \wedge \ldots B_n\}$, Grouping set $G = \{\}$, Bound set $S = \{\}$
Output: The denominator of the COR confidence formula, c

```
Generate all possible combinations of n-tuples by binding the atoms in B and add to S
Order B_i in B according to the symmetric property that asymmetric atoms are listed first and set this list to B'
for (B_i in B') do
  if B_i is quasi function and symmetric then
    if The subject of B_i exists in G then
      Add the object of B_i to G
    else
      Add the subject of B_i to G
    end if
  end if
  if B_i is quasi function and not symmetric then
    Add the subject of B_i to G
  end if
end for
Group the members of S according to G
Count the number of groups and assign to c
if r(x,y) is a quasi-function then
  for(x_i in X) do
    if ((Count of y_j in Y) > 1) then
      c = c + 1
    end if
  end for
end if
Return c    # of groups according to the quasi functions
```

Table 1
An example KB containing one quasi-function (R_1).

livesIn	wasBornIn
(Adam, Paris)	(Adam, Paris)
(Adam, Rome)	(Carl, Rome)
(Adam, Siena)	(Bob, Rome)
(Adam, Toulouse)	
(Adam, Istanbul)	
(Bob, Zurich)	

4.2.1. The COR Confidence when There is One Quasi-Function in the Body

Consider $R_1 : livesIn(x,y) \Rightarrow wasBornIn(x,y)$ as an example rule. A person can live in multiple cities, that is, $livesIn(x,y)$ is a quasi-function since there can be multiple y's for each x. Table 1 shows an example KB that contains two relations, namely *livesIn* and *wasBornIn*, and nine facts. In this case, $standard\_confidence(R_1) = 1/6$, because i) there is only one true fact for the rule, *wasBornIn(Adam, Paris)*, and ii) there are six facts in the body. $PCA\_confidence(R_1) = 1/6$ as well, because there are not any unknown facts for any x in the body.

In this example, there is only one atom in the body that holds and it is a quasi-function. According to the proposed algorithm, the 2-variable bindings in the body must be grouped according to the subject of the relation, x. There are only two different x's in the body that are Adam and Bob, and thus the number of groups is equal to 2 and $COR\_confidence(R_1) = 1/2$.

In brief, standard confidence and PCA confidence ignore the quasi-functions and for this reason they calculate the strength of the given rule less than it should be. Therefore, we propose to use COR confidence measure to evaluate the strength of a Horn rule when the single body atom is a quasi-function.

4.2.2. The COR Confidence when There are n Quasi-Functions in the Body

Consider a rule which states that a student's advisor works at the same institution that the student is educated: $R_2 : educatedAt(x,y) \wedge hasAcademicAdvisor(x,z) \Rightarrow worksAt(z,y)$ where $educatedAt(x,y)$ and $hasAcadamicAdvisors(x,z)$ are quasi-functions.

The following 3-variable bindings are generated according to the given KB in Table 2: $educatedAt(x,y), hasAcademicAdvisor(x,z)$: *(John, Yale University, Kate), (Clara, Boston University, Dave), (Bob, Creshire Academy, Alice), (Bob, Creshire Academy, Amy), (Bob, University of Michigan, Alice), (Bob, University of Michigan, Amy), (Brad, Cats College Academy, Ann), (Brad, Cats College Academy, Sue), (Brad, Yale University, Ann), (Brad, Yale University, Sue), (Brad, Indiana University, Ann), (Brad, Indiana University, Sue).*

Table 2
An example KB containing quasi-functions in the body (R_2).

educatedAt	hasAcademicAdvisor	worksAt
(John, Yale University)	(John, Kate)	(Kate, Yale University)
(Clara, Boston University)	(Clara, Dave)	(Sue, Boston University)
(Bob, Creshire Academy)	(Bob, Alice)	(Alice, University of Michigan)
(Bob, University of Michigan)	(Bob, Amy)	(Ann, Stanford University)
(Brad, Cats College Academy)	(Brad, Ann)	(Amy, Yale University)
(Brad, Yale University)	(Brad, Sue)	
(Brad, Indiana University)		

Table 3
An example KB containing quasi-functions and symmetric relationship in the body (R_3).

livesIn	marriedTo	wasBornIn
(Brandon, Berlin)	(Brandon, Amber)	(Amber, Berlin)
(Brandon, Amsterdam)	(Brandon, Siena)	(Siena, Chicago)
(Brandon, Toulouse)	(Bobby, April)	(April, Amsterdam)
(Bobby, Berlin)	(Bobby, Jane)	(Kori, Chicago)
(Bobby, Zurich)	(John, Kori)	
(John, Chicago)	(John, Foster)	
(John, New York)		

Thus, $standard\_confidence(R_2) = 2/12$, because there are twelve generated facts in the body and two facts in the head follow the rule: *worksAt(Kate, Yale University)* and *worksAt(Alice, University of Michigan)*. However, $PCA\_confidence(R_2) = 2/11$. Since the working place of Dave is unknown, *(Clara, Boston University, Dave)* is disregarded. COR confidence does not only disregard it but also considers the impact of quasi-functions.

Since both $educatedAt(x, y)$ and $hasAcademicAdvisor(x, y)$ atoms in the body are quasi-functions and not symmetric, x is the only grouping member. According to the proposed algorithm, the 3-variable bindings in the body must be grouped according to x: Group 1: *(John, Yale University, Kate)*, Group 2: *(Bob, Creshire Academy, Alice), (Bob, Creshire Academy, Amy), (Bob, University of Michigan, Alice), (Bob, University of Michigan, Amy)*, Group 3: *(Brad, Cats College Academy, Ann), (Brad, Cats College Academy, Sue), (Brad, Yale University, Ann), (Brad, Yale University, Sue), (Brad, Indiana University, Ann), (Brad, Indiana University, Sue)*.

Thus, $COR\_confidence(R_2) = 2/3$, because there are three groups according to the quasi-functions. To sum up, COR confidence is a more informative rule quality metric when there are quasi-functions in the body of a rule. Standard confidence and PCA confidence quantify the strength of a given rule by ignoring the effect of quasi-functions in the body.

Consider another rule $R_3 : livesIn(x, y) \wedge marriedTo(x, z) \Rightarrow wasBornIn(z, y)$. A person can live in multiple cities and can get married more than once. So both the atoms in the body are quasi-functions, and marriage is symmetric relationship.

Fourteen 3-variable bindings are generated according to the given KB in Table 3: $livesIn(x, y) \wedge marriedTo(x, z)$: *(Brandon, Berlin, Amber), (Brandon, Berlin, Siena), (Brandon, Amsterdam, Amber), (Brandon, Amsterdam, Siena), (Brandon, Toulouse, Amber), (Brandon, Toulouse, Siena), (Bobby, Berlin, April), (Bobby, Berlin, Jane), (Bobby, Zurich, April), (Bobby, Zurich, Jane), (John, Chicago, Kori), (John, Chicago, Foster), (John, New York, Kori), (John, New York, Foster)*.

$wasBornIn(Amber, Berlin)$ and $wasBornIn(Kori, Chicago)$ are the two facts which follow the rule. So, $standard\_confidence(R_3) = 2/14$. However, $PCA\_confidence(R_3) = 2/10$. This is because, *wasBornIn* relation of Jane and Foster are unknown, hence *(Bobby, Berlin, Jane), (Bobby, Zurich, Jane), (John, Chicago, Foster)*, and *(John, New York, Foster)* bindings are disregarded.

The COR confidence also disregards the mentioned bindings for the same reason and considers quasi-functions in the body. $livesIn(x, y)$ is a quasi-function so its subject, x, is added to the grouping set. $marriedTo(x, z)$ is also a quasi-function and additionally it is a symmetric relation. Since its subject, x has been already added to the grouping set, its object, z is added to the grouping set, too. Hence, the generated facts are divided into four groups according to x and z: Group 1: *(Brandon, Berlin, Amber), (Brandon, Amsterdam, Amber), (Brandon, Toulouse, Amber)*, Group 2: *(Brandon, Berlin, Siena), (Brandon, Amsterdam, Siena), (Brandon, Toulouse, Siena)*, Group 3: *(Bobby, Berlin, April), (Bobby, Zurich, April)*, Group 4: *(John, Chicago, Kori), (John, New York, Kori)*. So, in this case $COR\_confidence(R_3) = 2/4$.

Table 4
An example KB containing both function and quasi-function (R_4).

wasBornIn	isCitizenOf	country
(Martin, Rome)	(Martin, Italy)	(Berlin, Germany)
(Sam, Berlin)	(Sam, Germany)	(Chicago, US)
(Michael, Chicago)	(Michael, US)	(Amsterdam, Netherlands)
(Jane, Chicago)	(Michael, Italy)	(Rome, Italy)
(Grace, Zurich)	(Michael, Germany)	
	(Grace, US)	

4.2.3. The COR Confidence when There are Functions and Quasi-Functions in the Body

Consider a rule stating that a person's city of birth belongs to one of the countries that he or she is citizen of, $R_4 : wasBornIn(x, y) \wedge isCitizenOf(x, z) \Rightarrow country(z, y)$. It is clear that $wasBornIn(x, y)$ is a function while *isCitizenOf* (x, z) is a quasi-function.

Six 3-variable bindings are generated by binding the atoms in the body according to the given KB in Table 4: $wasBornIn(x, y) \wedge isCitizenOf(x, z)$: *(Martin, Rome, Italy), (Sam, Berlin, Germany), (Michael, Chicago, US), (Michael, Chicago, Italy), (Michael, Chicago, Germany), (Grace, Port Moresby, US)*. There are three facts in the head which follow the rule: *country(Berlin, Germany), country(Chicago, US)* and *country(Rome, Italy)*. So, $standard\_confidence(R_4) = 3/6$. However, *country* relation of Port Moresby is unknown - it does not exist. So, *(Grace, Port Moresby, US)* is disregarded and $PCA\_confidence(R_4) = 3/5$. COR confidence does not only disregard the mentioned triple, but also considers the quasi-function in the body that $isCitizenOf(x, z)$. According to the proposed Algorithm 1, the bound triples must be grouped according to x: Group 1: *(Martin, Rome, Italy)*, Group 2: *(Sam, Berlin, Germany)*, Group 3: *(Michael, Chicago, US), (Michael, Chicago, Italy), (Michael, Chicago, Germany)*. Thus, $COR\_confidence(R_4) = 3/3$. This example shows once again the importance of considering different categories of relations for assessing rule quality in KBs.

4.2.4. The COR Confidence when There are only Functions in the Body

We propose to use the PCA confidence measure when there are not any quasi-functions in the body. Consider the rule stating that people were born and passed away in the same city, $R_5 : wasBornIn(x, y) \Rightarrow diedIn(x, y)$. For each x, there is at most one y for *wasBornIn* relation, so this atom is a function. The same principle does hold when there are multiple atoms which are functions in the body of a rule. To conclude, we propose to use PCA confidence for the Horn rules in the form of $B \Rightarrow r(x, y)$ where the atoms in B is a set of functions.

4.2.5. COR Confidence when the Head Atom is a Quasi-Function

Consider a rule stating that a person's spouse lives in the same city with him or her: $R_6 : livesIn(x, y) \wedge marriedTo(x, z) \Rightarrow livesIn(z, y)$. Fourteen 3-variable bindings are generated according to the given KB in Table 5: $livesIn(x, y) \wedge marriedTo(x, z)$: *(Brandon, Berlin, Amber), (Brandon, Berlin, Siena), (Brandon, Amsterdam,*

Table 5
An example KB containing quasi-functions both in the body and in the head atom (R_6).

livesIn	marriedTo	livesIn
(Brandon, Berlin)	(Brandon, Amber)	(Amber, Berlin)
(Brandon, Amsterdam)	(Brandon, Siena)	(Amber, Amsterdam)
(Brandon, Toulouse)	(Bobby, April)	(Siena, Chicago)
(Bobby, Berlin)	(Bobby, Jane)	(April, Amsterdam)
(Bobby, Zurich)	(John, Kori)	(Kori, Chicago)
(John, Chicago)	(John, Foster)	(Kori, Istanbul)
(John, New York)		

Amber), (Brandon, Amsterdam, Siena), (Brandon, Toulouse, Amber), (Brandon, Toulouse, Siena), (Bobby, Berlin, April), (Bobby, Berlin, Jane), (Bobby, Zurich, April), (Bobby, Zurich, Jane), (John, Chicago, Kori), (John, Chicago, Foster), (John, New York, Kori), (John, New York, Foster).

There are three facts which follow the rule: *livesIn(Amber, Berlin)*, *livesIn(Amber, Amsterdam)* and *livesIn(Kori, Chicago)*. So, $standard\_confidence(R_6) = 3/14$. Since *livesIn* relation of Jane and Foster are unknown, *(Bobby, Berlin, Jane), (Bobby, Zurich, Jane), (John, Chicago, Foster)*, and *(John, New York, Foster)* bindings are disregarded when PCA confidence and COR confidence values are calculated. For this reason, $PCA\_confidence(R_6) = 3/10$.

A person can live in multiple cities and a person can get married more than once. So the atoms in the body of the rule are quasi-functions. For this reason, the grouping set members are x and z. Since the relations and their instantiations in the body are the same in Table 3 and Table 5, the generated facts are divided into the same four groups. Group 1: *(Brandon, Berlin, Amber), (Brandon, Amsterdam, Amber), (Brandon, Toulouse, Amber)*, Group 2: *(Brandon, Berlin, Siena), (Brandon, Amsterdam, Siena), (Brandon, Toulouse, Siena)*, Group 3: *(Bobby, Berlin, April), (Bobby, Zurich, April)*, Group 4: *(John, Chicago, Kori), (John, New York, Kori)*. However, unlike other examples the head atom is a quasi-function as well. There are an additional one y values for *"Amber"* and *"Kori"* in the head atom, *livesIn (Amber, Amsterdam)* and *livesIn (Kori, Istanbul)*. Thus, according to the Algorithm 1, the number of groups is increased by 2 and assigned to 6.

The main idea of COR confidence measure is based on grouping the instances of atoms according to the subjects in quasi-functions. When the head atom is also a quasi-function, the COR confidence measure aims to decompose the already generated groups according to the subjects in the head atom. In this example, the decomposed groups according to y with all instances of body and head atoms in R_6 are: Group 1: *(Brandon, Berlin, Amber, Berlin), (Brandon, Amsterdam, Amber, Berlin), (Brandon, Toulouse, Amber, Berlin)*, Group 2: *(Brandon, Berlin, Amber, Amsterdam), (Brandon, Amsterdam, Amber, Amsterdam), (Brandon, Toulouse, Amber, Amsterdam)*, Group 3: *(Brandon, Berlin, Siena, Chicago), (Brandon, Amsterdam, Siena, Chicago), (Brandon, Toulouse, Siena, Chicago)*, Group 4: *(Bobby, Berlin, April, Amsterdam), (Bobby, Zurich, April, Amsterdam)*, Group 5: *(John, Chicago, Kori, Chicago), (John, New York, Kori, Chicago)*, Group 6: *(John, Chicago, Kori, Istanbul), (John, New York, Kori, Istanbul)*. To conclude, the number of groups is 6, and 3 instances follow the rule, for this reason $COR\_confidence(R_6) = 3/6$.

5. Evaluation

In this section, the COR confidence and the PCA confidence measures are compared in terms of assessment of the Horn rule quality on synthetic test data, and on the real KBs Wikidata and DBpedia. The PCA confidence is the first measure which assesses the Horn rules under OWA. Although it is capable of providing more interesting rules than the standard confidence due to this property, it does not consider quasi-functions. Since the COR confidence measure works under OWA as well and it is the first measure which considers quasi-functions in rule mining, we compare our proposed measure with PCA confidence.

This section is divided into three subsections. In the first subsection, we present a confidence weighting method which is used in the evaluation phase. In the second subsection, the properties of the generated KB are introduced. In the third subsection, the COR confidence is compared with the PCA confidence when i) there are only quasi-functions in the body, and ii) there are both functions and quasi-functions in the body. In the last subsection, the measures are compared when there are quasi-functions in both the body and the head of a Horn rule.

5.1. Confidence Weighting Method

The COR confidence groups the instances of atoms according to the subjects when there are quasi-functions in a Horn rule. For this reason, COR confidence measures the quality of a Horn rule more appropriately as will be explained in 5.3. Although the interestingness of the rule is affected with the number of objects per each subject when there are quasi-functions both in the body and the head of a Horn rule, this effect is not proportional. It is obvious to represent the effect of this approach with a weight assignment method.

There are different weight assignment methods in ARM literature like weighted association rule mining (WARM) [25] and weighted classification based on association rules algorithm (WCBA) [26]. Tao et al. [25] focus on assigning weights to each item while Alwidian et al. [26] are interested in weighting the attributes of the items to find the most efficient ones for extracting more accurate rules. In each of these approaches, the items in the itemset are the core components. We propose a contribution to the quality assessment of Horn rules by handling the rule as a whole with a weighting factor. This weighting factor, which is inspired by the Natural Language Processing (NLP) literature scores a term with respect to a document corpus, $tf - idf$ (term frequency–inverse document frequency) [27]. It indicates the relevancy of a word to a document in a collection of documents and are used in association rule mining [28,29]. Apart from these studies, we propose a weighting factor, $ir - gre$ (*implied rule-grouping rule effect*), to measure the efficiency of COR confidence. We have adapted the formula of $tf - idf$ to obtain the COR confidence interestingness weight to the PCA confidence as shown in Eq. 4. When $tf - idf$ weights the relevancy of a word in a document inside the whole corpus, $ir - gre$ evaluates the impact of grouping instances of atoms according to the quasi-functions in a Horn rule.

Eq. 4 denotes the proposed confidence weighting measure, $ir - \text{gre}$. ir and gre are shown in Eq. 5 and Eq. 6, respectively. ir denotes the number of known facts that are implied by the rule. gre indicates the impact of grouping when there are quasi-functions in a Horn rule. Although the interestingness of the rule is affected according to the number of objects per each subject in quasi-functions, it is not proportional. A mathematical function is needed to dampen this effect as done in $tf - \text{idf}$. As a consequence, log function is used in each part of $ir - \text{gre}$ formula to prevent unfair boosting. If $ir - \text{gre}$ is significantly greater than 0, the rule quality is measured more effectively. We use log scale since we propose an interestingness measure to interpret the increase in the number of subjects in quasi-functions in a nonlinear way.

$$ir - gre = ir * gre \tag{4}$$

$$ir = log(support(B \Rightarrow r(x, y))) \tag{5}$$

$$gre = log\left(\frac{\#(x, y) : \exists z_1, \ldots, z_n, y\prime : B \wedge r(x, y\prime)}{\#groups\_according\_to\_the\_quasifunctions}\right) \tag{6}$$

5.2. Evaluation Setup

RDF KBs mostly involve incomplete knowledge that might not easily be processed during evaluation. To deal with the evaluation process in a controlled way, we generated synthetic RDF data involving triples using LUBM (The Lehigh University Benchmark) that facilitates the evaluation of Semantic Web repositories in a standard way [30]. Since LUBM serves a sample ontology for the

```
<ub:FullProfessor rdf:about="http://www.Department0.University0.edu/FullProfessor0">
  <ub:name>FullProfessor0</ub:name>
  <ub:teacherOf rdf:resource="http://www.Department0.University0.edu/Course0" />
  <ub:teacherOf rdf:resource="http://www.Department0.University0.edu/GraduateCourse0" />
  <ub:teacherOf rdf:resource="http://www.Department0.University0.edu/GraduateCourse1" />
  <ub:undergraduateDegreeFrom>
    <ub:University rdf:about="http://www.University84.edu" />  </ub:undergraduateDegreeFrom>
  <ub:mastersDegreeFrom>
    <ub:University rdf:about="http://www.University875.edu" />  </ub:mastersDegreeFrom>
  <ub:doctoralDegreeFrom>
    <ub:University rdf:about="http://www.University241.edu" />  </ub:doctoralDegreeFrom>
  <ub:worksFor rdf:resource="http://www.Department0.University0.edu" />
  <ub:emailAddress>FullProfessor0@Department0.University0.edu</ub:emailAddress>
  <ub:telephone>xxx-xxx-xxxx</ub:telephone>
  <ub:researchInterest>Research13</ub:researchInterest>
  <ub:researchInterest>Research14</ub:researchInterest>
  <ub:researchInterest>Research2</ub:researchInterest>
  <ub:topicOfPhdDegree>Research2</ub:topicOfPhdDegree>
</ub:FullProfessor>
```

Fig. 1. Sample Fragment of The Generated Knowledge Base.

university domain, we generated our triples according to this example domain and created the related rules. The benchmark ontology[3] includes the classes in our customized KB. Fig. 1 represents a sample fragment of the generated KB using the LUBM Data Generator. There is a sample full professor instance and their related properties such as the taught courses, masters degree information, doctoral degree information, occupation, email address, phone number, research interests and topic of PhD degree. We changed the number of triples in a controlled way to compare the PCA confidence and the COR confidence measures in different cases.

We created 21394 class instances and 88257 property instances in the generated KB. Organizational classes (University, Department, Faculty), academic positions (Professor, FullProfessor, AssociateProfessor, AssistantProfessor, Lecturer, TeachingAssistant, ResearchAssistant), type of students (Student, UndergraduateStudent, GraduateStudent), course information (Course, GraduateCourse), other academic entities (Publication, Chair, Research, ResearchGroup), and the properties related to these classes are stored in the generated KB. Since our objective is to assess the quality of the proposed rule metric, we consider the properties in our KB as the atoms in the rule evaluation task. For example, we especially inserted "topicofPhdDegree" property to the core ontology specification in the ontology class information of data generator[4] to create a sample rule such that: $researchInterest(x, y) \Rightarrow topicofPhdDegree(x, y)$.

As a result, a complete KB is generated in order to evaluate the impact of quasi-functions on the PCA confidence and the COR confidence measures. If the confidence value of a rule is equal to 1 (the maximum confidence value), this rule is satisfied by all instances of atoms. In the experiments, we calculate the two confidence measures when the rule is satisfied by different percentage of instances. We refer to this parameter as the "distribution of rule strength". The maximum possible distribution of rule strength is equal to 100% and it means that all the instances of atoms satisfy the rule.

[3] http://swat.cse.lehigh.edu/onto/univ-bench.owl
[4] http://swat.cse.lehigh.edu/projects/lubm/uba1.7.zip

5.3. Evaluation Results

5.3.1. Impact of Quasi-Functions in the Body of Horn Rules

In this section, we first aim to show how the PCA confidence is unfairly affected by the quasi-functions. Consider a rule stating that a professor's topic of PhD degree is one of their research interest topics, $R_7 : researchInterest(x, y) \Rightarrow topicofPhdDegree(x, y)$. In this case, we change the number of research interests from 1 to 10 for each professor. Fig. 2 shows the PCA confidence when rule strength distributions are 60%, 80% and 100%. As the number of research topics increases, the PCA confidence decreases for all rule strength distributions. The reason for this decrease is due to calculating the additional number of instances of atoms in the body. Since all professors' topic of PhD degree is one of their research interest topics, a good confidence measure should be equal to 1 when the distribution of rule strength is 100%. However, the PCA confidence is less than 1 in this case and it decreases as the number of research interest topics increases. To conclude, the PCA confidence calculates the confidence value of a rule less than it should be when there is a quasi-function in the body of a Horn rule. The gap between the confidence measures increases as the number of associated objects to a given subject increases.

We also aim to show how the COR confidence can handle quasi-functions in the body of a Horn rule. In this case we fix the number of research interest to 2, for the rule $R_7 : researchInterest(x, y) \Rightarrow topicofPhdDegree(x, y)$. Fig. 3 shows the COR confidence and the PCA confidence values with different rule strength distributions. The COR confidence values are equal to the rule strength distribution because the generated KB is complete and the proposed measure groups the 2-bindings of $researchInterest(x, y)$ according to the x which denotes the professors. However, the PCA confidence ignores the impact of quasi-functions and counts a professor's various research interest topics as different facts in the body.

We also compare the COR confidence with the PCA confidence for various Horn rules in real KBs, Wikidata and DBpedia. Consider a rule which states that a person's native language is one of their spoken or written languages: $R_8 : languages\_spoken\_written(x, y) \Rightarrow native\_language(x, y)$. One can speak and write in more than one language even though one can only have a single native language. The relation

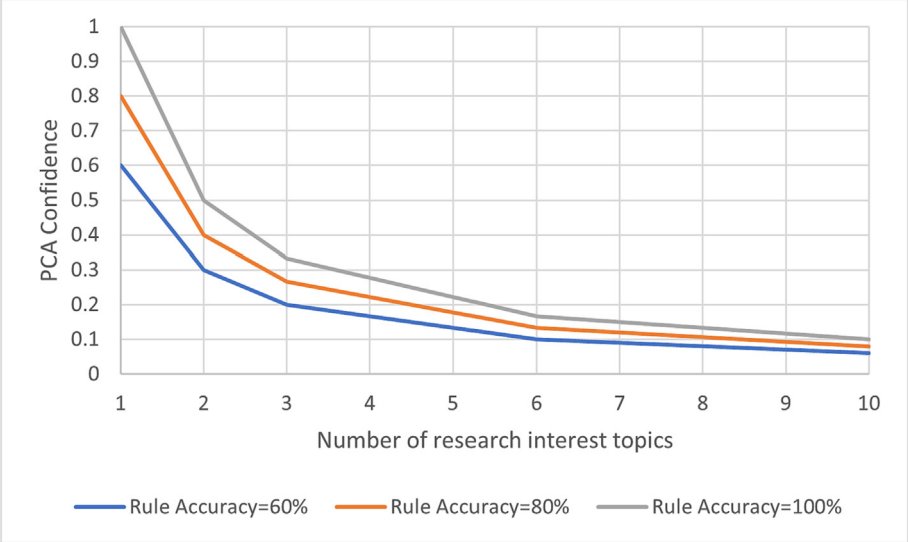

Fig. 2. Impact of number of research interest topics in R_7.

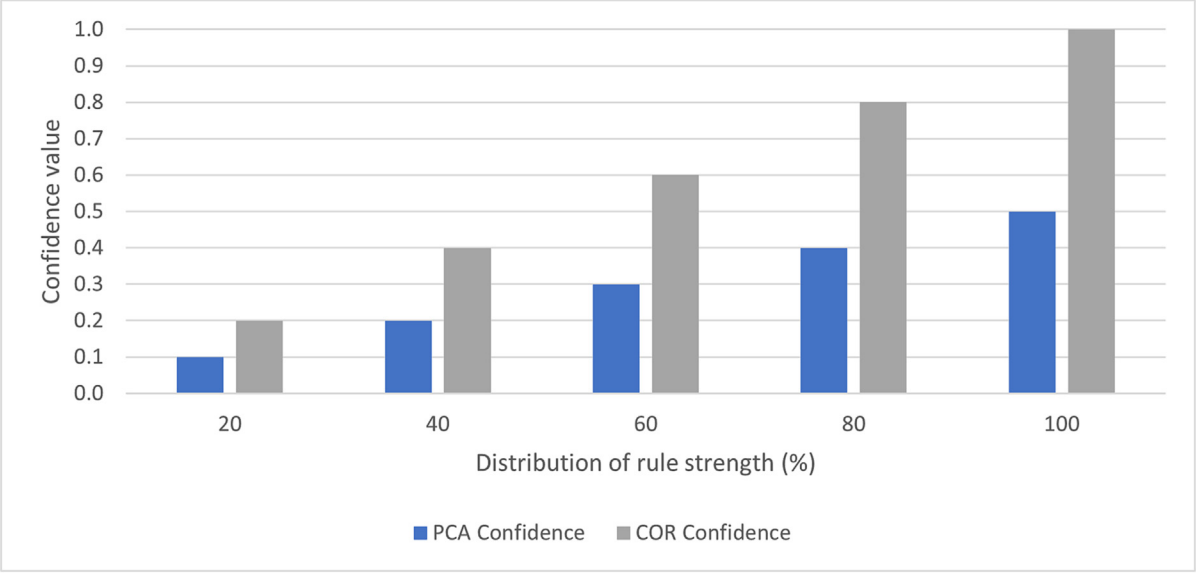

Fig. 3. Comparison of PCA Confidence and COR Confidence in R_7.

Table 6
Some Facts for $languages\_spoken\_written(x, y) \Rightarrow native\_language(x, y)$ in Wikidata (R_8).

languages_spoken_written	native_language
(Buddy Van Horn, English)	(Buddy Van Horn, English)
(Erik Vliegen, German)	(Erik Vliegen, German)
(Erik Vliegen, English)	(Stephen Harper, English)
(Erik Vliegen, French)	(Ugur Sahin, Turkish)
(Erik Vliegen, Dutch)	(Wolfgang Amadeus Mozart, German)
(Erik Vliegen, Italian)	
(Stephen Harper, English)	
(Stephen Harper, French)	
(Ugur Sahin, Turkish)	
(Ugur Sahin, German)	
(Ugur Sahin, English)	
(Wolfgang Amadeus Mozart, Italian)	
(Wolfgang Amadeus Mozart, English)	
(Wolfgang Amadeus Mozart, French)	
(Wolfgang Amadeus Mozart, Latin)	
(Wolfgang Amadeus Mozart, German)	

in the body of this rule is a quasi-function while the head atom is a function. However, there are some instances with multiple native languages in Wikidata such as English and British English. We ignore these facts and consider the facts where each person has got only one native language when the rule consists *native_language* relation. Table 6 shows some related facts about the rule R_8 in Wikidata. As shown, there can be multiple y's for each x for *languages_spoken_written_or_signed (x, y)* since it is a quasi-function. PCA confidence does not consider this reality while the COR confidence groups the subjects and assesses the rule quality accordingly. So, $PCA\_confidence(R_8) = 0.86$ while $COR\_confidence(R_8) = 0.98$. Also, $ir - gre(R_8) = 0.32$. When we think the rule semantically, a person's native language must be one of their spoken or written languages. These results indicate that the COR confidence measures this rule's quality more properly.

Consider a rule stating that a person's spouse's native language is one of their spoken/written/signed languages: $R_9 : languages\_spoken\_written(x, y) \wedge spouse(x, y) \Rightarrow native\_language(z, y)$. Both rela-

Table 7
Some Facts for $languages\_spoken\_written(x,y) \wedge spouse(x,y) \Rightarrow native\_language(x,y)$ in Wikidata (R_9).

languages_spoken_written	spouse	native_language
(JM Jarre, English)	(JM Jarre, Charlotte Rampling)	(Anne Parillaud, French)
(JM Jarre, French)	(JM Jarre, Anne Parillaud)	(Michelle Obama, English)
(Barack Obama, English)	(JM Jarre, Gong Li)	
(Barack Obama, Indonesian)	(Barack Obama, Michelle Obama)	

Table 8
An example KB from Wikidata when body and the head atoms are quasi-functions (R_{11}).

educatedAt	employer
(Michel Ibrahim, Uni. Iberoamericana)	(Michel Ibrahim, Boston Medical Center)
(Saira Butt, Uni. Tecnológica de Santiago)	(Michel Ibrahim, Jackson Memorial H.)
(Saira Butt, Lahore College for Women Uni.)	(Michel Ibrahim, Miami VA Healthcare S.)
	(Michel Ibrahim, University of Miami H.)
	(Saira Butt, Indiana Uni. – Purdue U. I.)
	(Saira Butt, Uni. of Mississippi Medical Center)

tions in the body of this rule are quasi-functions. Table 7 shows the related facts about two subjects, Jean-Michel Jarre (JM Jarre) and Barack Obama. Although both the PCA confidence and the COR confidence evaluate the rule quality under OWA, only the COR confidence considers quasi-functions. Accordingly, $PCA\_confidence(R_9) = 0.76, COR\_confidence(R_9) = 0.86$, and $ir-gre(R_9) = 0.23$. When we think about the meaning of the rule, we can also intuitively say that the COR confidence measures the rule more appropriately.

Consider another rule from Wikidata which states a music album's title is one of its tracks contained on this album: $R_{10}: instance\_of(x, Album) \wedge tracklist(x,y) \wedge title(x,z) \Rightarrow instance\_of(y, Music\_Track\_with\_Vocals) \wedge title(y,z)$. A music album contains several tracks except singles, so $tracklist(x,y)$ is a quasi-function. $PCA\_confidence(R_{10}) = 0.05$ while $COR\_confidence(R_{10}) = 0.15$. The proposed weighting factor, $ir-gre(R_{10}) = 1.26$ shows that the rule quality is measured more effectively by the COR confidence since $ir-greR_{10})$ is significantly greater than 0. As other examples, the COR confidence measures the rule quality more properly since it considers quasi-functions.

5.3.2. Impact of Quasi-Functions in the Body and Head of the Horn Rules

Consider the rule $R_7: researchInterest(x,y) \Rightarrow topicofPhdDegree(x,y)$ describes that a professor's topic of PhD degree is in the list of her research interests, and Fig. 1 shows a sample fragment of the generated KB as explained previously. *FullProfessor0* is an instance of this rule that holds because her topic of the PhD degree, namely *Research2*, is also her research interest. There are thirty professors in the generated KB. In this case, each professor has four research interests and the distribution of rule strength is equal to 1. The proposed confidence weighting factor, $ir-gre$ values for the rule R_7 are respectively 0.89, 0.54, 0.24 and 0 when the number of topic of Phd degree for each professor is changed from one to four. Correspondingly, $ir-gre$ value decreases as the number of objects for each subject in the head atom increases. When the number of atoms' instances in $topicofPhdDegree(x,y)$ per each x is equal to the number of atoms' instances in $researchInterest(x,y)$, denominators of the COR and PCA confidence measures become the same and as a result $ir-gre$ value is equal to 0. Briefly, COR confidence is more efficient than PCA confidence when $ir-gre$ is greater than 0.

We also compared the two confidence measures on Wikidata when there are quasi-functions in the body and the head of a Horn rule. Consider the rule $R_{11}: educatedAt(x,y) \wedge occupation(x, Researcher) \wedge country(y, DominicanRepublic) \Rightarrow employer(x,y)$. Table 8 shows some related facts about the rule R_{11} in Wikidata. As shown, the $educatedAt(x,y)$ atom in the body and the head atom $employer(x,y)$ are quasi-functions. For this rule, $PCA\_confidence(R_{11}) = 0.14, COR\_confidence(R_{11}) = 0.18$ and the $ir-gre(R_{11}) = 0.032$. The COR confidence measures the rule strength of R_{11} more effective since $ir-gre$ value is greater than 0. Although $educatedAt(x,y)$ is a quasi-function, there are not any people who are educated at multiple universities in the Dominican Republic in R_{11}. It is important to note that the $ir-gre(R_{11})$ would increase if there were some people who are educated at multiple universities in Dominican Republic.

Consider another rule from Wikidata, $R_{12}: child(x,y) \wedge academicDegree(x, DoctorofPhilosophy) \wedge academicDegree(y, DoctorofPhilosophy) \wedge educatedAt(x,z) \Rightarrow educatedAt(y,z)$. In this rule, both *child* and *educatedAt* are quasi-functions since a person can have multiple children and a person can be educated at multiple schools. The confidence values and the confidence weighting measure are $PCA\_confidence(R_{12}) = 0.08, COR\_confidence(R_{12}) = 0.25$, and $ir-gre(R_{12}) = 0.036$. The $ir-gre$ is greater than 0, so the COR confidence measure is more effective.

6. Conclusion

In this paper, we have presented a new confidence measure called the COR confidence for Horn rules on RDF KBs. The COR confidence measure aims to assess the degree of certainty of a Horn rule under the OWA by considering the impact of quasi-functions both in the body and the head atoms. Since each subject can be related with multiple objects in quasi-functions, The COR confidence measure is based on grouping the subjects. Besides, it considers symmetric relations and includes the objects of relations to grouping members when required.

We have shown through several examples in 5 that the COR confidence measures the strength of Horn rules more effectively due to consideration of categories of relations in the atoms of the rule. The COR confidence calculates the confidence value of a rule greater than the PCA confidence when there are quasi-functions in the atoms. The reason of that is the grouping property of the COR confidence which prevents considering related facts separately. In conclusion, the COR confidence measure takes a step further on the current research of confidence measures for KBs by considering categories of relations for the first time.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd edition., Morgan Kaufmann, 2011.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A Nucleus for a Web of Open Data, in: Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 722–735.

[3] D. Vrandecic, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85, https://doi.org/10.1145/2629489.

[4] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: C.L. Williamson, M.E. Zurko, P.F. Patel-Schneider, P.J. Shenoy (Eds.), Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8–12, 2007, ACM, 2007, pp. 697–706. doi: 10.1145/1242572.1242667.

[5] T. Pellissier Tanon, G. Weikum, F. Suchanek, Yago 4: A reason-able knowledge base, in: A. Harth, S. Kirrane, A.C. Ngonga Ngomo, H. Paulheim, A. Rula, A.L. Gentile, P. Haase, M. Cochez (Eds.), The Semantic Web, Springer International Publishing, Cham, 2020, pp. 583–596.

[6] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, R. Lee, Media meets semantic web–how the bbc uses dbpedia and linked data to make connections, in: European semantic web conference, Springer, 2009, pp. 723–737.

[7] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka, T.M. Mitchell, Toward an architecture for never-ending language learning, in: in: Twenty-Fourth AAAI conference on artificial intelligence, 2010.

[8] L.A. Galárraga, C. Teflioudi, K. Hose, F.M. Suchanek, Fast rule mining in ontological knowledge bases with AMIE+, VLDB J. 24 (6) (2015) 707–730, https://doi.org/10.1007/s00778-015-0394-1.

[9] L.A. Galárraga, C. Teflioudi, K. Hose, F.M. Suchanek, AMIE: association rule mining under incomplete evidence in ontological knowledge bases, in: D. Schwabe, V.A.F. Almeida, H. Glaser, R. Baeza-Yates, S.B. Moon (Eds.), 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13–17, 2013, International World Wide Web Conferences Steering Committee/ ACM, 2013, pp. 413–422. doi: 10.1145/2488388.2488425.

[10] Z. Wang, J. Li, Rdf2rules: Learning rules from RDF knowledge bases by mining frequent predicate cycles, CoRR abs/1512.07734. arXiv:1512.07734, doi: 10.48550/arXiv.1512.07734.

[11] T.P. Tanon, D. Stepanova, S. Razniewski, P. Mirza, G. Weikum, Completeness-aware rule learning from knowledge graphs, in: C. d'Amato, M. Fernández, V.A. M. Tamma, F. Lécué, P. Cudré-Mauroux, J.F. Sequeda, C. Lange, J. Heflin (Eds.), The Semantic Web - ISWC 2017–16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I, vol. 10587 of Lecture Notes in Computer Science, Springer, 2017, pp. 507–525. doi: 10.1007/978-3-319-68288-4_30.

[12] T.P. Tanon, D. Stepanova, S. Razniewski, P. Mirza, G. Weikum, Completeness-aware rule learning from knowledge graphs, in: J. Lang (Ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13–19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 5339–5343. doi: 10.24963/ijcai.2018/749.

[13] K. Zupanc, J. Davis, Estimating rule quality for knowledge base completion with the relationship between coverage assumption, in: Proceedings of the 2018 World Wide Web Conference, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 1073–1081. doi: 10.1145/3178876.3186006.

[14] J. Lajus, L. Galárraga, F.M. Suchanek, Fast and exact rule mining with AMIE 3, in: A. Harth, S. Kirrane, A.N. Ngomo, H. Paulheim, A. Rula, A.L. Gentile, P. Haase, M. Cochez (Eds.), The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings, vol. 12123 of Lecture Notes in Computer Science, Springer, 2020, pp. 36–52. doi: 10.1007/978-3-030-49461-2_3.

[15] V. Zeman, T. Kliegr, V. Svátek, Rdfrules: Making RDF rule mining easier and even more efficient, Semantic Web 12 (4) (2021) 569–602, https://doi.org/10.3233/SW-200413.

[16] R. Agrawal, T. Imielinski, A.N. Swami, Mining association rules between sets of items in large databases, in: P. Buneman, S. Jajodia (Eds.), Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, May 26-28, 1993, ACM Press, 1993, pp. 207–216. doi: 10.1145/170035.170072.

[17] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, pp. 487–499.

[18] P.N. Tan, V. Kumar, J. Srivastava, Selecting the right interestingness measure for association patterns, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02, Association for Computing Machinery, New York, NY, USA, 2002, p. 32–41. doi: 10.1145/775047.775053.

[19] L. Geng, H.J. Hamilton, Interestingness measures for data mining: A survey, ACM Comput. Surv. 38 (3) (2006) 9, https://doi.org/10.1145/1132960.1132963.

[20] C.V. Tew, C.G. Giraud-Carrier, K.W. Tanner, S.H. Burton, Behavior-based clustering and analysis of interestingness measures for association rule mining, Data Min. Knowl. Discov. 28 (4) (2014) 1004–1045, https://doi.org/10.1007/s10618-013-0326-x.

[21] M. Kirchgessner, V. Leroy, S. Amer-Yahia, S. Mishra, Testing interestingness measures in practice: A large-scale analysis of buying patterns, in: 2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17–19, 2016, IEEE, 2016, pp. 547–556. doi: 10.1109/DSAA.2016.53.

[22] M. Shaikh, P.D. McNicholas, M.L. Antonie, T.B. Murphy, Standardizing interestingness measures for association rules, Stat. Anal. Data Min. 11 (6) (2018) 282–295, https://doi.org/10.1002/sam.11394.

[23] R. Sharma, M. Kaushik, S.A. Peious, S.B. Yahia, D. Draheim, Expected vs. unexpected: selecting right measures of interestingness, in: International Conference on Big Data Analytics and Knowledge Discovery, Springer, 2020, pp. 38–47.

[24] R. Somyanonthanakul, T. Theeramunkong, Scenario-based analysis for discovering relations among interestingness measures, Information Sciences 590 (2022) 346–385, https://doi.org/10.1016/j.ins.2021.12.121.

[25] F. Tao, F. Murtagh, M. Farid, Weighted association rule mining using weighted support and significance framework, in, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 661–666, https://doi.org/10.1145/956750.956836.

[26] J. Alwidian, B.H. Hammo, N. Obeid, Wcba: Weighted classification based on association rules algorithm for breast cancer disease, Applied Soft Computing 62 (2018) 536–549, https://doi.org/10.1016/j.asoc.2017.11.013.

[27] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (5) (1988) 513–523, https://doi.org/10.1016/0306-4573(88)90021-0.

[28] S. Mahmood, M. Shahbaz, A. Guergachi, Negative and positive association rules mining from text using frequent and infrequent itemsets, The Scientific World Journal (2014), https://doi.org/10.1155/2014/973750.

[29] H.J. Kim, J.W. Baek, K. Chung, Optimization of associative knowledge graph using tf-idf based ranking score, Applied Sciences 10 (13) (2020) 4590, https://doi.org/10.3390/app10134590.

[30] Y. Guo, Z. Pan, J. Heflin, Lubm: A benchmark for owl knowledge base systems, Journal of Web Semantics 3 (2) (2005) 158–182, selcted Papers from the International Semantic Web Conference, 2004. doi: 10.1016/j.websem.2005.06.005.