ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 2, pp: (77-86), Month: April - June 2023, Available at: www.researchpublish.com

An Improved Semantic Items Generation Algorithm for Mining High Number of Significant Association Rules from Semantic Web Data

Isya Isyaku[1*], Muhammad Sirajo Aliyu[2], Abubakar Aminu Muazu[1]

[1]Department of Computer Science, Umaru Musa Yar'adua University, Katsina Nigeria

[2]Department of Computer Science, Faculty of Computing, Federal University Dutse, Jigawa State Nigeria.

DOI: https://doi.org/ 10.5281/zenodo.7936390

Published Date: 15-May-2023

Abstract: **The number of Ontologies and the Linked Open Data (LOD) on the Web is constantly increasing. This type of complex and heterogeneous semi-structured data raises new opportunities and challenges for the data mining (DM) community. Semantic Web Association Rule Mining (SWARM) algorithm have the limitations of generating many non-interesting, duplicate and equivalents rules and low algorithm performance. This research demonstrates a procedure for improving the performance of SWARM in text mining by using RDF data. The result revealed significant reduction in the number of generated rules this is significant because it helps to address the problem of discovering duplicate/equivalent Association Rules from Semantic Web Data.**

Keywords: **Semantic Web, Association Rules, Data Mining, Linked Data.**

I. INTRODUCTION

The World Wide Web has radically altered the way we share knowledge by lowering the barrier to publishing and accessing documents as part of a global information space [1]. With the advent of the Semantic Web (SW), there has been a growing popularity of Knowledge Bases (KBs). Projects like DBpedia, YAGO, NELL, Wikidata, and the Knowledge Vault strive to create and manage vast sets of data that can be processed by machines and pertain to real-world entities [2]. Numerous millions of RDF statements in large RDF-style knowledge bases serve as machine-understandable fact representations. Semantic Web is a way of representing information on the World Wide Web in a way that is machine-readable and understandable. This is done by using a structured format called Resource Description Framework (RDF), which is based on XML. RDF data is typically structured in triples, which consist of a subject, a predicate, and an object.[3]. RDF-style Knowledge Bases (KBs) typically possess an associated schema, which comprises a collection of statements or facts that establish various aspects such as the domains and ranges for relations, object and predicate equivalences, and class hierarchy. The RDF Schema (RDFS) and the Ontology Web Language (OWL) recommendation are used to provide the vocabulary needed to define the schema of an RDF KB[4].

Although the Web has been hugely significant, it's only been recently that the principles that facilitated the growth of the web of documents have been employed for data. This transforms the Web from a global information platform of linked documents to one in which both documents and data are interconnected.[5]. The objective of the process is to establish the Semantic Web or Web of Data, and Linked Data is the method to achieve this objective [1].

Data mining, which also known as knowledge discovery from databases (KDD), is described by [6] as "a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" or as the process of extracting

implicit, previously unrecognized, and potentially useful information from data [7]. Over the last few decades, significant breakthroughs in data mining methods have brought about revolutionary changes in data analytics and big data. Data mining involves the integration of techniques from a diverse range of disciplines, such as statistics, artificial intelligence, machine learning, database systems, among others, for the analysis of large datasets. According to [8] Data mining involves examining (frequently extensive) observed data collections to discover unexpected associations and present the data in innovative ways that are meaningful and beneficial to the data owner. Semantic Data Mining pertains to the data mining activities that methodically integrate domain knowledge, particularly formal semantics, into the process [9]. In pattern mining and association rule mining, domain ontologies are often used to reduce the feature space in order to get more meaningful and interesting patterns [10].

Recently, there has been an increasing interest in combining the two areas Semantic Web and Data Mining [11]. The majority of research concerning data mining for Semantic Data utilizes Inductive Logic Programming (ILP), which leverages the inherent logic contained in the data to acquire fresh concepts [12]. In this regard, the proposed methods measure the quality of the discovered rules using instance-level data only. The methods are also memory inefficient and the approaches heavily rely on domain experts.

However, Semantic Web data contains knowledge in both instance-level and schema-level. Ignoring knowledge encoded at the schema-level when mining the Semantic Web Data would negatively impacts the quality and interpretation of the discovered rules. For this reason, [13] developed an approach that automatically mines Semantic Association Rules from RDF data, the proposed approach reveals common behavioural patterns that are associated with knowledge at the instance-level and schema-level. Experimental result of the work of [13] showed that ignoring other semantic relations (such as; *owl:sameAs* & *owl:equivalentProperty*) from the schema-level during semantic items generation process negatively impacts the quality of rules mined by SWARM approach.

II. RESEARCH OBJECTIVES

The aim of this research is to improve the work of [13] by modifying the semantic items generation algorithm and considering semantic-relations from the schema-level that has the potential of improving the number of significant mined rules.

The objectives are to;

1. Study and Implement the Semantic Web Association Rule Mining (SWARM) approach.
2. Modify the semantic items generation algorithm of SWARM approach in order to improve the quality of mined Association Rules.
3. Compare the proposed work with SWARM approach in terms of Support and Confidence quality factors.

III. REVIEW OF RELATED LITERATURE

There is already much work on the SW data mining in the fields of inductive logic programming and approaches that make use of the description logic of a knowledge base [14]. These works on data mining for SW data are based on Inductive Logic Programming (ILP) [15], which exploits the underlying logic encoded in the data to learn new concepts. [16] proposed a multi-threaded approach called AMIE for mining association (Horn) rules from RDF-style Knowledge bases. AMIE outperform the then state of the art approaches in terms of speed, precision and coverage, furthermore, AMIE mines rules orders of magnitude faster than the then state-of-the-art approaches. Similar to AMIE, [17] proposed another approach for extracting horn rules. The proposed methods measure the quality of the discovered rules using only instance-level data. However, properties of the generated rule might be having more general description. [18] extends AMIE to AMIE+ using pruning and query rewriting techniques to mine even larger KBs. Extensive experiments showed that the later approach, AMIE+, extends to areas of mining that were previously beyond reach.

Association rule mining was originally proposed for shopping basket problems [19]. Association rule mining is a technique for finding engaging relations between factors in vast databases [20]. [21] proposed a rule mining method over RDF-based medical data. Through the use of mining patterns created by SPARQL queries, transactions have been created. The approach relies heavily on domain experts and the discovered rules do not take advantage of the rich semantics encoded at the schema-level. [22] developed an approach to identify schema and value dependencies between RDF data using six different Configurations. Any part of the Subject-Predicate-Object (SPO) statement may be regarded as a context, which is used to

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 2, pp: (77-86), Month: April - June 2023, Available at: www.researchpublish.com

identify the target of mining for one of the two remaining parts of the statement. The discovered rules show the correlation among subjects, predicates or objects independently. In comparison with this approach, our proposed approach would generate rules associated with knowledge in instance-level and schema-level.

SWARM (Semantic Web Association Rule Mining) an approach developed by [23] automatically mines Semantic Association Rules from RDF data. [23] extends the approach by exposing SWARM to larger dataset and more experiments yet the approach suffers from the issue of mining duplicate Association Rules. Our suggested method differs from this approach as it would utilize additional semantic relationships at the schema level to eliminate redundant rules, which would enhance the accuracy of the discovered rules.

IV. METHODOLOGY AND ALGORITHM DESIGN

This section provides an elaborate explanation of our approach, including the definitions that support it. A schematic representation of the entire process is shown in Figure 1. The overall framework of an Improved Semantic Web Association Rule Mining (ISWARM) is shown in Figure 1. ISWARM approach has three (3) modules namely, Instance or data level, Preprocessing and Mining Module modules. The Pre-processing Module, which is divided into these two submodules, automatically processes RDF triples: Common Behavior Set Generation and Semantic Item Generation. The Mining Module uses Common Behavior Sets as input to generate Semantic Association Rules. The ISWARM approach evaluates the quality of rules by considering all the semantics at schema and instance level of the data in the ontology. Support and Confidence, two of the proposed rule quality indicators used in the study, take into account knowledge at both the instance- and schema-levels.

Using knowledge encoded at the schema-level and instance-level

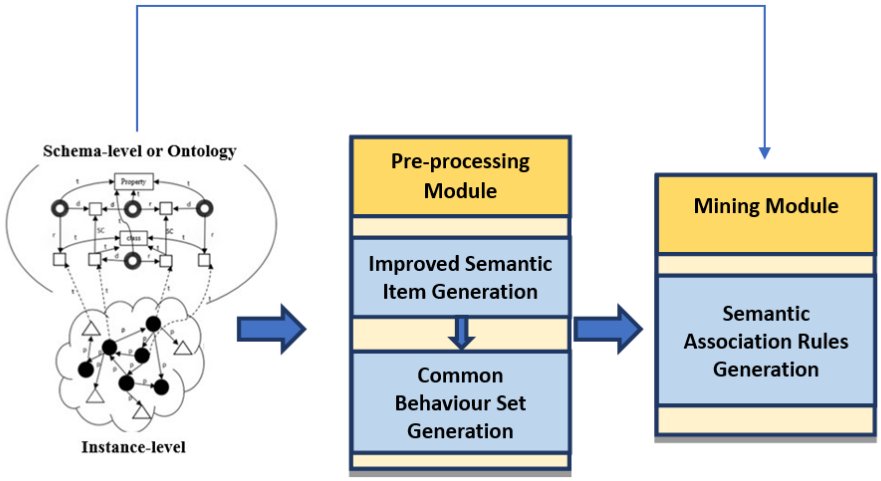

Figure 1. Improved Semantic Web Association Rule Mining

Pre-Processing Module

Finding intriguing connections between things in a dataset can be done using the approach of association rule mining. This idea was first stated by [5]. Let $I = \{i_1,i_2,...,i_n\}$ be a set of items and $D = \{t_1,t_2,...,t_m\}$ be a set of transactions. One subset of the items in I is present in every transaction. An association rule is a consistent pattern shows displays how certain objects are present in transactions, and typically uncovers patterns of behaviour of specific entities. For instance, the shopping basket problem demonstrates a customer behaviour pattern with the rule {CPU, Monitor} ⇒ {Keyboard}, which implies that if a customer purchases CPU and Monitor together, they are also likely to buy Keyboard. Traditional algorithms for mining association rules are primarily appropriate for homogeneous repositories, where items and transactions are vital components of the mining process.[5]. The majority of Semantic Web data is not in the form of transactional data, and thus there are no transactions or items to work with. Therefore, it is necessary to create models that represent such concepts in order to derive associations from Semantic Web data.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 2, pp: (77-86), Month: April - June 2023, Available at: www.researchpublish.com

Table 1: Some RDF triples from the DBpedia dataset (3.8)

Triples	Subject	Predicate	Object
t_1	Michael Jackson	instrument	Guitar
t_2	Michael Jackson	spouse	Deborah Jeanne
t_3	Michael Jackson	occupation	Musician
t_4	Yoko Ono	birthplace	Tokyo
t_5	Bob Marley	instrument	Guitar
t_6	Bob Marley	occupation	Musician
t_7	Barack Obama	office	President of the USA
t_8	Barack Obama	party	Democratic
t_9	Bill Clinton	office	President of the USA
t_{10}	Bill Clinton	party	Democratic
t_{11}	George W. Bush	office	President of the USA
t_{12}	George W. Bush	party	Republic
t_{13}	Fela Kuti	device	Guitar
t_{14}	Fela Kuti	job	Musician
t_{15}	Joe Biden	office	President of the USA
t_{16}	Joe Biden	caucus	Democratic

Definition 1 *(Entity Behaviour).* An Entity Behaviour can be derived from an RDF triple t_i and is represented as a 2-tuple, $eb_i = (e_i, pa_i)$. Here, e_i refers to an Entity that can be either the subject *(s)* or the object *(o)* of the triple. pa_i is a Pair associated with eb_i that denotes a behavior exhibited by e_i. Based on the content in e_i, pa_i consists of either the combination of predicate-object or predicate-subject, i.e., *(p,o)* or *(p,s)*, respectively.

Based on the provided explanation, our aim is to identify the behavioural patterns demonstrated by different entities. To achieve this goal, we introduce two concepts, namely Semantic Item and Common Behaviour Set. These concepts help to demonstrate the behaviours that are shared among the entities.

Semantic Item Generation

As seen in Table 1, the subjects in the t_1, t_5 and t_{13} i.e., *Michael Jackson, Bob Marley and Fela Kuti*, share a semantically common activity, i.e., *(instrument/device Guitar)*. i.e., playing guitar is a common behavior taken by this group of entities, i.e., *Michael Jackson, Fela Kuti* and *Bob Marley*; due to the fact that the predicate/behavior *device* can as well be used in place of the predicate/behavior *instrument* as stated in the Schema level (A-Box) Knowledge. This is how our work distinguished from the previous works by further utilizing the knowledge in the Schema Level in order to mine semantically rich association rules from RDF data.

Definition 2 *(Semantic Item).* The definition of a Semantic Item, denoted by si_j, is a 2-tuple consisting of es_j and pa_j. Here, es_j is an Element Set that comprises a list of subjects or objects, i.e., $\{s_1, s_2, ..., s_n\}$ or $\{o_1, o_2, ..., o_n\}$, respectively. The Pair of si_j, denoted by pa_j, is associated with the content in es_j and contains a combination of predicate-object or predicate-subject, i.e., *(p, o)* or *(p, s)*.

In the RDF data model, any object from one triple can serve as a subject of another triple. Definition 2 takes this feature into account, where an Element Set es can contain a list of subjects or objects. It's worth noting that in this work, ISWARM uses the $\{s_1, s_2, ..., s_n\}(p,o)$ structure to mine Semantic Association Rules, indicating that an Element Set es can only contain a list of subjects. Additionally, it's important to mention that we haven't defined an Element Set es as a list of predicates. As mentioned earlier, the predicate in each Pair (*p*, *o*) is crucial in representing the concept of a behaviour for a specific Semantic Item.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 2, pp: (77-86), Month: April - June 2023, Available at: www.researchpublish.com

Algorithm 1: Semantic Item Generation

```
1  SI ← Ø;
2  foundFlag ← false;
3  predicateSynonyms ← new HashMap<String, HashSet<String>>();

4  for each predicate p in predicateSynonyms.keySet() do
       synonyms ← new HashSet<String>();
       synonyms.add(p);
       for each synonym in predicateSynonyms.get(p) do
           synonyms.add(synonym);
       predicateSynonyms.put(p, synonyms);
5  end for

6  for each triple tj in Triples do
       pa0 ← the Pair of tj

       for each sii in SI do
           if predicateSynonyms.get(sii.predicate).contains(pa'.predicate) then
               sii.subjects.add(tj.subject);
               foundFlag ← true;
               break;
           end if
       end for

       if foundFlag = false then
           newSI ← new Semantic Item containing tj;
           SI.add(newSI);
       end if

       foundFlag ← false;
7  end for
8  return SI;
```

According to Definition 2, triples in a triple dataset can be transformed into a set of Semantic Items i.e., $SI=\{si_1, si_2,...,si_n\}$. Semantic Items can be generated from triples dataset using the Algorithm 1. The algorithm accepts triples as input and creates a set of Semantic Items as the output. Step 1: Create an empty set SI to hold the generated Semantic Items and set foundFlag to false. Step 2: Create a HashMap called *predicateSynonyms* to group synonyms of each predicate. For each predicate in the *predicateSynonyms* map, create a new HashSet called *synonyms* and add the predicate itself and all its synonyms to this set. Then, add the synonyms set to the *predicateSynonyms* map with the predicate as the key. Step 3: Iterate over each triple t_j in the given set of Triples. Step 4: Extract the subject-predicate pair *pa'* from the current triple t_j. Step 5: Iterate over each Semantic Item si_i in the *SI* set. Step 6: Check if the predicate of the current triple tj matches any of the synonyms of the predicate in the Semantic Item si_i. To do this, retrieve the synonyms set for the predicate in the *predicateSynonyms* map, using the get method. If the synonyms set contains the predicate of the current triple, then it is considered a match. Step 7: If a matching Semantic Item si_i is found, add the subject of the current triple tj to the subjects set of the Semantic Item si_i and set *foundFlag* to true. Then, exit the loop since there is no need to check for further matches. Step 8: If no matching Semantic Item is found, create a new Semantic Item newSI containing the current triple t_j and add it to the SI set. Step 9: Reset the *foundFlag* to false for the next iteration. Step 10: Return the final *SI* set.

Example 1: Consider the triples shown in Table 1 Some of the Semantic Items generated by Algorithm 1 have been shown in Table 2 For example, the Element Set of si_2 contains three entities *Michael Jackson*, *Bob Marley* and *Fela Kuti*. The Pair of si_2 is (*occupation*, *Musician*). si_2 indicates that (*occupation*, *Musician*) is a particular behaviour of *Michael Jackson*, *Bob Marley* and *Fela Kuti*.

Table 2: Some examples of semantic items

Semantic Items	
si_1	{Michael Jackson, Bob Marley, Fela Kuti} (instrument, Guitar)
si_2	{Michael Jackson, Bob Marley, Fela Kuti} (occupation, Musician)
si_3	{Barack Obama, Bill Clinton, George W. Bush, Joe Biden} (office, President of the USA)
si_4	{Barack Obama, Bill Clinton, Joe Biden} (party, Democratic)

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 2, pp: (77-86), Month: April - June 2023, Available at: www.researchpublish.com

Common Behavior Set Generation

The main aim of this research is to reveal the shared behavioral patterns in RDF triples. As previously mentioned in the preceding section, a Semantic Item displays a shared behavior (explained in the Pair) exhibited by a set of entities (i.e., subjects) in the Element Set. In this subsection, we introduce a novel concept called the Common Behavior Set, which represents all the shared actions taken by groups of similar entities in the Element Sets.

Definition 3 *(Common Behavior Set).* A Common Behavior Set (cbs) is comprised of a group of Semantic Items that share similar Element Sets, i.e., $cbs = \{si_1,si_2,...,si_n\} = (ES,PA)$, where $ES = \{si_1.es \cup si_2.es \cup ... \cup si_n.es\}$ and $PA = \{si_1.pa \cup si_2.pa... \cup si_n.pa\}$. Items are grouped together into the same *cbs* if the similarity level of their Element Sets meets or exceeds the *SimTh* (Similarity Threshold). The level of similarity between Element Sets can be calculated using Equation 1.

$$SD(es_a, es_b, \dots, es_m) = \frac{|es_a \cap es_b \cap \dots \cap es_m|}{|es_a \cup es_b \cup \dots \cup es_m|} \quad (1)$$

Definition 3 states that a *cbs* is a set of Semantic Items combined based on the similarity of entities within their Element Sets. A *cbs* displays a set of shared occurrences of certain actions performed by entities within their Element Sets. A Total Common Behavior Set (*TCBS)* can be created, containing all Common Behavior Sets, as per Algorithm 2. The algorithm takes Semantic Items $SI=\{si_1,si_2,...,si_n\}$ and Similarity Threshold *SimTh* as inputs and returns *TCBS* as output. For each si_i in SI, the algorithm calculates the *SD* of si_i's Element Sets and stores the result in *SD* (Lines 3-7). If the *SD* is equal to or greater than *SimTh*, then the algorithm adds si_i to cbs_j (Lines 8-14). If no existing Common Behavior Set has a similar Element Set, the algorithm generates a new *cbs*, cbs^0_j, and adds it to TCBS (Lines 11-14). Finally, the updated *TCBS* is returned by the algorithm (Line 15).

```
1    TCBS ← an empty set
2    foundFlag ← false

3    for each si in S do:
4        for each cbsj in TCBS do:
5            seta ← intersection of esi for all si in cbsj
6            Setb ← union of esi for all si in cbsj
7           SD ← size of intersection of esi and seta divided by size of
             union of es_i and set_b
8            if SD ≥ SimTh then:
9                add si to cbsj
10               foundFlag ← true

11       if foundFlag = false then:
12           cbsj ← a new set containing only si
13           add cbsj to TCBS

14       foundFlag ← false

15   return TCBS
```

Example 2: Table 3 displays the Common Behavior Sets that were produced by the Semantic Items listed in Table 2. By setting the Similarity Threshold *(SimTh)* to 50%, two Common Behavior Sets, cbs1 and cbs2, were created from si_1, si_2, and si_3, si_4, respectively (as listed in Table 2). For instance, *cbs2* depicts that over 50% of entities within its Element Sets exhibit two common behaviors, namely, *(office, President of the USA) and (party, Democratic).*

Table 3: Common Behavior Sets

Common Behaviour Sets	
cbs_1	{Michael Jackson, Bob Marley, Fela Kuti} (instrument, Guitar)
	{Michael Jackson, Bob Marley, Fela Kuti} (occupation, Musician)
cbs_2	{Barack Obama, Bill Clinton, George W. Bush} (office, President of the USA)
	{Barack Obama, Bill Clinton} (party, Democratic)

Mining Module

Association rule mining aims to identify frequently co-occurring associations among a group of items. Its goal is to discover rules that can predict the occurrence of a specific item based on a set of transactions. To apply association rule mining in the context of Semantic Web (SW) data, we need to consider the notion of frequency. As discussed earlier, each Common Behavior Set *(cbs)* represents a unique set of activities that are commonly performed by the subjects in the Element Sets. This can be seen as a type of transaction. Therefore, we explore how to generate Semantic Association Rules from *cbs*. To incorporate semantics into the generated rules, we introduce quality rule factors such as Support, Confidence, and Lift. These factors utilize both instance-level and schema-level knowledge.

Semantic Association Rules Generation

Definition 4 *(Semantic Association Rule).* A Semantic Association Rule consists of two parts, namely pa_{ant} and pa_{con}. pa_{ant}, also known as Antecedent Pairs, contains some of the Pairs of a given cbs_j, denoted as $\{pa_1,...,pa_n\}$On the other hand, pa_{con}, also known as Consequent Pairs, contains the remaining Pairs of *cbsj*, denoted as $\{pa_{n+1},...,pa_m\}$. A rule r is characterized by a common Rule's Element Set, denoted as res, which is the union of Element Sets in the cbsj, denoted as {es1∪...∪esm}. Each Element Set, denoted as esi, is a set of instances, i.e., $es_i = \{ins_1,ins_2,...,ins_k\}$. The antecedent and consequent of a rule r can be expressed using the implication sign.

$$res: pa_{ant} \Longrightarrow pa_{con}$$

Table 4: Some discovered Improved Semantic Association Rules Semantic Association Rules

Semantic Association Rules	
r_1	$\{Michael\ Jackson, Bob\ Marley, Fela\ Kuti\}$: $(instrument, Guittar) \Rightarrow (occupation, Songwriter)$
r_2	$\{Jimmy\ Carter, Bill\ Clinton, George\ W.Bush\}$: $(office, President\ of\ the\ USA) \Rightarrow (party, Democratic)$

Table 4 illustrates two instances of rules that were generated from the Common Behaviour Sets listed in Table 3. For instance, rule r_1 comprises of a common Rule's Element Set *(res)* created by uniting the Element Sets in *cbs1*, i.e., *{Michael Jackson, Bob Marley and Fela Kuti}*. The antecedent and consequent of r_1 contain the Pairs *(instrument/device, Guitar) and (occupation, Musician)*, correspondingly. It is crucial to note that the order of antecedents and consequents in rules can be interchanged.

Quality Factors for Rules

Most current methods used for association rule mining from RDF data focus solely on the instance-level when assessing the quality of rules. However, in this study, we introduce two new quality factors, Support, Confidence, and Lift, which consider knowledge not only at the instance-level but also at the schema-level. In traditional association rule mining, transactions typically involve recording the behaviour of a single class of actors, such as shopping customers. However, in the context of SW, instances often belong to different types or classes in various ontologies. The method we present takes into account this feature of SW data, as instances in the Rule's Element Sets reflect this diversity.

Take, for instance, the instances in the Rule's Element Set of rule r_1, which includes *Michael Jackson*, *Fela Kuti*, and *Bob Marley*. Table 1 shows a small portion of the DBpedia ontology, where *Bob Marley* is classified as a Musical Artist and *Michael Jackson* is a Person. In terms of the ontology hierarchy, if the Musical Artist class is a subclass of the Artist class and the Artist class is a subclass of the Person class, then instances of the Musical Artist class also belong to the Person class. However, *Michael Jackson*, being a Person, does not belong to the Musical Artist class. The interpretation of association rules in the context of SW data poses a significant challenge that fundamentally depends on the ontology structure. It is inadequate to evaluate the quality of rules by merely considering knowledge at the instance-level.

SUPPORT: Suppose we have a Semantic Association Rule *r* that has the form of $res: pa_{ant} \Longrightarrow pa_{con}$. The calculation of Support *sup (r)* can be done by applying Equation 2:

$$Sup(r) = \frac{frequency(pa_{ant})}{Total\ Number\ of\ Transactions} \quad (2)$$

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 11, Issue 2, pp: (77-86), Month: April - June 2023, Available at: www.researchpublish.com

The numerator of the fraction is the total number of instances of Class c_i that holds pa_{antk} as the Pairs. The denominator of the fraction is the total number of instances of Class c_i.

Example 3: To calculate the value of Support for the r_1 shown in Table 2, we use following fraction:

$$Sup(r_1) = \frac{|frequency(instrument\ Guitar)|}{|Total\ Number\ of\ Instances\ in\ Person\ Class|}$$

The numerator of the Support fraction indicates the total number of instances of *Person* class that hold *instrumentGuitar* as a Pair. As shown in Table 3, there are only three instances which share *instrumentGuitar and deviceGuitar* as a Pair. The denominator of the fraction also shows the total number of instances of *Person* class which is seven in this example (*sup*=0.42).

The numerator part of the Support fraction represents the sum of instances belonging to the *Person* class that have a Pair containing *instrumentGuitar* and *deviceGuitar*. In Table 3, only three instances have both *instrumentGuitar and deviceGuitar* as Pairs. The denominator part of the fraction represents the total number of instances in the *Person* class, which in this example is 7. Therefore, the resulting Support value is:

$$sup(r_1) = \frac{3}{7} = 0.42$$

CONFIDENCE: Suppose we have a Semantic Association Rule r that has the form of $res: pa_{ant} \Longrightarrow pa_{con}$. The calculation of Confidence *con(r)* can be done by applying Equation 2:

$$Sup(r) = \frac{frequency(pa_{ant} \wedge pa_{con})}{frequency(pa_{ant})} \quad (3)$$

Example 4: The numerator part of the Confidence fraction for rule r_1 indicates the total count of instances categorized as Person class, which have both i*nstrumentGuitar/deviceGuitar* and occupation Musician as their Pairs. As for the bottom part of the fraction, it represents the total count of instances categorized as Person class, which have been assigned *instrumentGuitar* as a Pair. Thus, the Confidence value for this rule is 1.0. In other words, the rule implies that most people who play Guitar also work as Musicians. Additionally, at least 60% of the instances in the Rule's Element Set follow the rule.

V. RESULT

The result discovered in this research provides solution to the first part of the aim, which is to eliminate duplicate or equivalent rules, which contribute semantically to the problem being solved.

Improved Semantic Web Association Rule

Comparing the rules generated from applying the improved Semantic Items generation algorithm on both the SWARM and ISWARM, we discovered that there was a significant rule reduction which translates to semantically enriched rules (strong and quality rules). The number of rules generated at different support thresholds is presented in Table 5.

Table 5: Comparison of SWARM vs ISWARM

	SWARM	***ISWARM***
Support	***Number of rules***	***Number of rules***
10	2478	862
20	1831	811
30	1461	759
40	1302	692
50	1322	601
60	1071	521

70	981	468
80	903	301
90	821	141
100	802	98

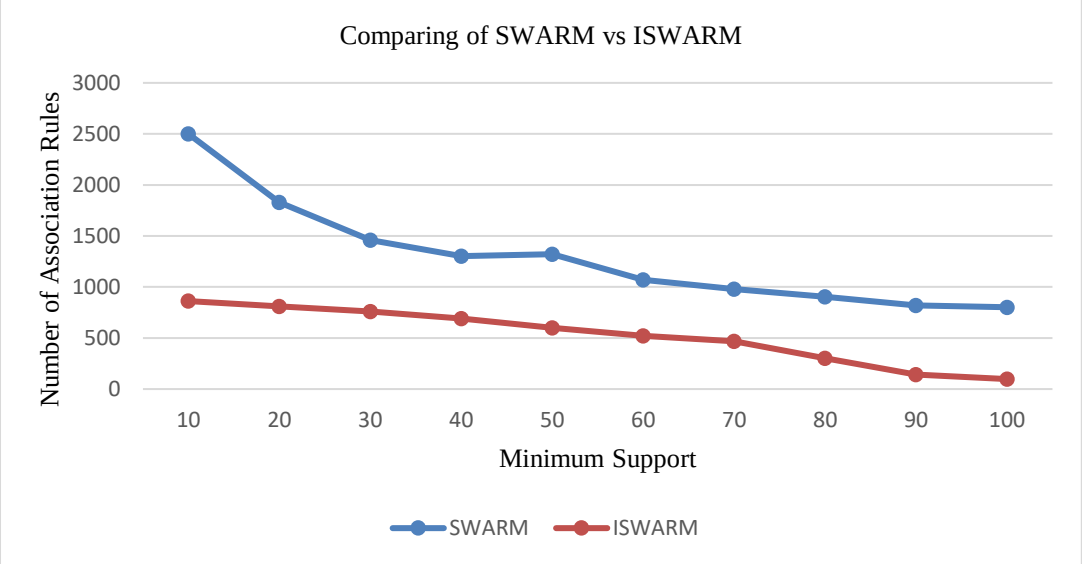

Fig. 2 Comparison of SWARM vs ISWARM

Figure 2 illustrates a comparison between the SWARM and ISWARM approaches by presenting the values obtained from both methods at different support thresholds. The figure reveals that the ISWARM approach generated a smaller number of rules (98) compared to the SWARM approach (802), indicating a significant reduction in the number of generated rules. This reduction is essential because it addresses the issue of identifying duplicate or equivalent Association Rules in Semantic Web Data.

VI. CONCLUSION

The main contribution of this work is to eliminate mining non interesting and duplicate rules from RDF data. To solve this issue, ISWARM is an improved approach that attaches semantics to the rules by utilizing knowledge encoded at the schema-level. ISWARM is able to automatically mine Semantic Association Rules from RDF-style KBs. We demonstrated that ignoring semantic relations at the schema-level negatively impacts the interpretation of rules. The results obtained showed that the proposed approach is more promising and effective in generating strong Association Rules from Semantic Web Data in terms of Support and Confidence quality factors. The time complexity of SWARM algorithm belongs to the $O(n^2)$ class (including the time for generating the Semantic Items, Common Behavior Sets, and Semantic Association Rules). In future research, the execution time of ISWARM should be improved to deal with larger datasets. ISWARM should also be exposed and tested using data from different knowledge bases.

The primary objective of this study is to address the issue of mining non-interesting and redundant rules from RDF data. To address this problem, the authors proposed an improved approach called ISWARM that utilizes semantic knowledge encoded at the schema-level to attach semantics to the rules. The ISWARM approach can automatically mine Semantic Association Rules from RDF-style knowledge bases. We demonstrated that ignoring semantic relations at the schema-level can have a negative impact on rule interpretation. The results of the study indicate that the proposed approach is more promising and effective in generating strong Association Rules from Semantic Web Data in terms of Support and Confidence quality factors. The time complexity of the SWARM algorithm is $O(n^2)$ which includes the time required to generate the Semantic Items, Common Behavior Sets, and Semantic Association Rules. In future studies, the execution time of ISWARM should be improved to deal with larger datasets, and to test the approach using data from various knowledge bases.

REFERENCES

[1] Bizer, C., Berners-Lee, T., & Heath, T. (2011). Linked data : The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22.

[2] Lius Galárraga, E. (2016). Knowledge Bases. In Encyclopedia of Database Systems (pp. 1–7). Springer US.

[3] Molood, B., Quan, T. T., & Qing, L. (2016). Resource Description Framework (RDF). In Encyclopedia of Database Systems (pp. 2524–2528). Springer US.

[4] Galárraga, L. (2016). Ontology. In Encyclopedia of Database Systems (pp. 1904–1910). Springer US.

[5] Mathieu, C., Sabrina, K., & Serena, V. (2018). Data on the Web Best Practices. World Wide Web Consortium (W3C).

[6] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. AAAI Press/MIT Press.

[7] Han, J., Pei, J., & Kamber, M. (2005). Data mining: concepts and techniques. Elsevier.

[8] Bellandi, A., Grossi, V., & Romei, A. (2007). Ontology-Driven Data Mining for Adaptive Web Sites. Journal of Data Mining and Knowledge Discovery, 15(2), 225–249.

[9] Dou, Z., Wang, B., & Liu, J. (2015). Semantic Data Mining: A Survey of the Field and Current Directions. Journal of Intelligent Information Systems, 45(3), 491–506.

[10] Ristoski, P., & Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics, 36, 1–22.

[11] Chomboon, N., Kaoungku, J., Kerdprasop, N., & Kerdprasop, K. (2014). Data Mining on Semantic Web for Business Intelligence Application. Procedia Computer Science, 35, 1035–1042.

[12] Nebot, V., & Berlanga, R. (2010). Ontology-based Data Mining. In Advances in Intelligent Data Analysis IX (pp. 9–20). Springer Berlin Heidelberg.

[13] Molood, B., Quan, T. T., & Qing, L. (2017). SWARM: Mining semantic association rules from RDF data. Expert Systems with Applications, 76, 116–126.

[14] Abedjan, Z., & Naumann, F. (2013). Data fusion and mining on linked data. IEEE Data Engineering Bulletin, 36(1), 59-67.

[15] Muggleton, S., & Raedt, L. D. (1994). Inductive logic programming: Theory and methods. Journal of Logic Programming, 19/20, 629-679.

[16] Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. M. (2013). AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. Proceedings of the 22nd International Conference on World Wide Web, 413-422.

[17] Yang, J., Shao, B., & Zhou, Y. (2014). Mining horn rules with confidence from ontological knowledge bases. Proceedings of the 7th International Conference on Knowledge Science, Engineering and Management, 258-269.

[18] Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. M. (2015). Fast rule mining in ontological knowledge bases with AMIE+. The VLDB Journal, 24(6), 707-731.

[19] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, 207-216.

[20] Shridhar, M., & Parmar, M. (2017). Association rule mining: A comprehensive review. International Journal of Computer Science and Information Security, 15(5), 113-121.

[21] Nebot, V., & Berlanga, R. (2012). Rule mining over RDF-based medical data. Expert Systems with Applications, 39(15), 12176-12188.

[22] Abedjan, Z., Zeuch, S., Naumann, F., & Lehner, W. (2013). Dependency discovery in RDF data. Proceedings of the 12th International Semantic Web Conference, 1-16.

[23] Molood, B., Aliasghari, S., & Rahgozar, M. (2016). SWARM: Semantic Web Association Rule Mining. Journal of Web Semantics, 41, 18-30.