Received: February 1, 2019

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Semantic Approach for Extracting Medical Association Rules

Mohammed Thamer[1]* **Shaker El-Sappagh[1]** **Tarek El-Shishtawy[1]**

[1]*Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Egypt*
* Corresponding author's Email: mhmdmhmd669@gmail.com

Abstract: Healthcare sector has large amounts of data that require careful analysis in order to improve the medical service offered to the patients. Semantic data mining can play an effective rule in analyzing such amounts of data. In this paper, we propose a framework for association rule extraction based on ontology semantics. In the proposed framework, traditional medical datasets are represented using web ontology language. The medical dataset is transformed into an ontology of the form of triples (subject, object, predicate), and SPARQL is used to query the generated ontology. The Apriori algorithm is used to generate the association rules. Intensive experiments have been conducted to measure the quality and significance of the resulting association rules under different scenarios using different support and confidence values. The obtained results have shown that ontology-based Apriori algorithm is much better than the traditional Apriori algorithm. The rules generated using both algorithms have been compared in terms of several performance metrics including the number of frequent items, the number of generated rules, the computation time, the memory consumption, and the average confidence of the generated rules. The different performance metrics revealed the superiority of the proposed semantic Apriori algorithm (the ontology-based Apriori) compared to traditional Apriori algorithm.

Keywords: Semantic data mining, Association rules, Ontology, Biomedical data.

1. Introduction

Healthcare environment has huge amounts of data that are needed to be effectively analyzed. Medical knowledge discovery is the process of extracting knowledge patterns from biomedical data. The extracted patterns can play an important role in the decision-making process, which can improve the quality of services presented to patients [1]. Data mining is the field that includes approaches and techniques, which are derived from many research areas such as statistics, artificial intelligence, machine learning, database systems, etc. in order to analyze large datasets [2]. It aims at extracting implicit and potentially useful information from data [3]. Data mining approaches have been used in many areas in the healthcare sector [4]. It has been used effectively to detect fraud and abuse in healthcare insurance issues. Moreover, it has been employed to improve the treatment strategy, hospital infection control, identifying high-risk patients, etc. [4].

Semantic data mining indicates the data mining operations that comprise domain knowledge, particularly formal semantics, in a systematic manner. Several research studies have shown the positive influence of incorporating the domain knowledge altogether with data mining tasks [2]. For example, the domain knowledge could be beneficial in the preprocessing during eliminating irrelevant and erroneous data [5, 6]. In addition, the domain knowledge is employed as a prior knowledge that can be helpful in decreasing the search space and guiding the search path during the search and pattern extraction task [7, 8]. Furthermore, the detected patterns can be filtered out [9, 10] or visualized through formally encoding them using knowledge engineering approaches [11].

The foremost step to incorporate the domain knowledge into the data mining process is to represent it using representation models that are

accessible and can be processed by computers. Several techniques can be used to represent the domain knowledge in a formal way [12]. Ontology is a popular representation model for domain knowledge. It is defined as "*a formal, explicit specification of a common conceptualization where conceptualization is an abstract model of some world phenomena*" [1]. Usually, ontology is used to represent the knowledge in a certain domain such as genetics. Such ontology is known as domain ontology and it is mainly used to make sharing and reusing of knowledge a trivial issue. Several ontologies have been built within the healthcare field (e.g., the Gene Ontology (GO [13]) and Unified Medical Language System (UMLS [14])) in order to allow the interoperability and inference processes between different types of systems [15]. Moreover, it presents the relationships among the objects of a certain domain. Furthermore, ontology can be used to refine the domain knowledge of a certain field. Hence, data mining methods can be combined with ontology in order to improve the overall data mining process where ignoring the semantic representation of data leads to generating unreasonable mining models [1].

In this study, we propose a general framework for ontology-based association rule mining. In the proposed framework, traditional medical datasets are represented using Web Ontology Language (OWL). Later on, the generated ontology is transformed into triples (subject, object, predicate) through querying the generated ontology using SPARQL query language. Finally, the Apriori algorithm is used to generate the association rules. The proposed framework is validated using the chronic kidney disease dataset and the obtained have shown the effectiveness of the resulting association rules compared to the association rules generated directly from applying the Apriori algorithm on the original dataset.

The remaining sections of this study are arranged as follows: Section 2 reviews the literature of semantic data in the healthcare field. Section 3 presents the proposed framework for ontology-based association rule mining. Section 4 contains the description of the conducted experiments and results analysis. Finally, this study is concluded in Section 5.

2. Related works

In this section, two categories of research works are reviewed. The first category contains the works that attempted to improve the data mining processes such as data preprocessing, rule association, classification, clustering, etc. using ontology [7-8, 16-22]. The second category includes the works that employs the semantic data mining approaches in the biomedical field [1, 23-26].

Bernstein et al. [16] have presented an intelligent discovery assistant (IDA) that aimed at composing and choosing the appropriate data mining processes as well as the appropriate order of execution. The proposed IDA search the space of valid processes with the help of ontology. The authors have shown the effectiveness of the proposed tool in terms of cost-sensitive classification over a large and complex dataset. However, the computation time of the proposed method depends on the size of the candidate area. Another ontology-based work to improve the data mining process and to allow semantic meta-mining has been suggested by Keet et al. [17]. The authors have developed an ontology named the Data Mining OPtimization Ontology (DMOP) to help in the informed decision-making process at different selection points through the data mining process. The proposed ontology includes the descriptions of the various data mining tasks. However, the proposed work suffers from several challenges regarding meta-modeling, property chains, and handling attributes in a broader context. Panov et al. [18] have attempted to unify and formalize the definitions and concepts of the data mining field through building an ontology named OntoDM. It contains the definitions of main entities in data mining including data type and data set, data mining algorithm, data mining task, etc. In addition, more complex entities such as constrains, inductive queries, and data mining scenarios can be defined. However, they cannot determine their shortcomings of their ontology in the process and refine their structure as needed.

An updated version of OntoDM is presented in [19] that is aligned with ontology of biomedical investigation (OBI) with extend set of relations and classes. However, formalizing the knowledge and building heavy weight ontology is time and resource consuming process.

Bellandi et al. [8] has proposed a framework for association rules extraction with the help of ontology. The objective of the proposed framework was to improve the quality and significance of the extracted association rules through decreasing the search space, employing efficient data structures, and using domain-dependent constraints. However, the proposed work did not model the constraints of the association rules structure. In addition, it did not integrate the evaluation of the constraints directly in the mining algorithm. Further work for extracting association rules based on ontology has been

presented by Ferraz and Garcia [20]. The authors have introduced a data preparation tool named SemPrune which constructed based on domain ontology. The objective of SemPrune is to help in the preprocessing and postprocessing phases of data mining and to produce better data mining results through ontology-enrichment of data. However, the accuracy of determining generalization/ specialization and composition/decomposition relations has a high influence on using the proposed model. In addition, an ontology-based ranking approach for the generated association rules has been suggested by Idoudi et al. [21]. The basic idea of the proposed ranking approach depends on organizing the ontology's concepts in a hierarchical manner of conceptual clusters. Then, the value of a certain association rule is assessed based on the dissimilarity of the clusters included in the items of the association rule. However, the validity of the proposed work has been proven only in a specific domain as well as they did not suggest a method to show the generality of their work. Moreover, Barati et al. [22] have introduced an automated association rule mining approach called Semantic Web Association Rule Mining (SWARM). The proposed approach can be used effectively for association rule mining from RDF data. It exploits the knowledge encoded at the instance and schema levels. However, the semantic web data suffers from incorrectness and inconsistency between the entities at the instance level and their corresponding classes in the ontology, which may lead to ambiguous interpretations. Balcan et al. [7] have presented a theoretical model to the usefulness of ontologies in learning multiple tasks using unlabeled data. They demonstrated through the proposed model that an ontology that represents the relations among multiple outputs is enough in some cases to learn a classification by utilizing a big unlabeled data. However, the proposed model works only when all categories are incident to a NAND edge.

Mohammadi et al. [23] have proposed a gene selection approach based on data mining techniques and Gene Ontology. The objective of the proposed method was to determine the disease-causing genes. It adopts the Fisher filter in addition to the support vector machine-based recursive feature elimination (SVRFE) method, with a greedy algorithm to remove the redundant genes. Another work that depends on the Gene Ontology has been introduced by Nagar et al. [24] in which they suggested an approach for finding association relationships in the annotation terms for the *Saccharomyces cerevisiae* (SGD) genome. In the proposed approach, first, a normalization algorithm is applied to make the different annotation terms have a similar level of specificity. Then, association rule mining algorithms are applied on the normalized datasets. However, the validity of their method has not been proved in a real-world application. Liu et al. [25] have presented an effective method for mining biomedical ontologies and data. The proposed method aimed at discovering the semantic associations and finding the errors exist in the ontologies using the data. It is considered a general data mining method in which the ontologies and data are represented using RDF hyper-graphs. In addition, it can suggest correction for inaccurate information found in the biomedical ontologies. However, no experiments have been conducted to show the scalability of the proposed method. In addition, the proposed method adopts only simple semantics. Hence, more complicated semantics need to be incorporated.

Mahmoodi et al. [26] have introduced an algorithm to detect gastric cancer based on rule association mining and ontology. The objective of the proposed method is to reduce the number of resulting rules. The conducted experiments over a dataset that consists of 490 cases have shown that the rules generated using the proposed algorithm are more intuitive and understandable. In addition, the time of the Apriori algorithm is reduced. However, the proposed work has been evaluated using small data set. Hence, larger data sets should be used to show the scalability of their work. Qrenawi et al. [1] have employed ontology-driven data mining techniques to determine the relationships between type II diabetes mellitus patients and their laboratory tests. The proposed method has been applied on a dataset of diabetes patients who have cardiovascular disease. The conducted experiments have shown that using ontologies reduced the number of attributes at the preprocessing level and improves other data mining stages. However, more terms and concepts need to be incorporated in their ontology in order to improve the diagnosis process.

Lakshmi et al. [27] have proposed a new method that depends on weighted association rule mining disease comorbidities prediction by employing both clinical and molecular data. However, the achieved accuracy needs to be enhanced using more datasets such as chemical-disease and drug-disease association data. Kafkas and Hoehndorf [28] have proposed a text mining system that aims to extract pathogen–disease relations from literature. The proposed system uses domain knowledge from an ontology and statistical methods to perform the extraction process. However, the proposed work can be improved by incorporating a pathogen abbreviation filter and extending the dictionaries of

their pathogens and diseases. Shen et al. [29] attempted to support rare disease differential diagnosis by enriching current rare disease sources through proposing a data-driven method. The proposed method mines the phenotype-disease associations that exist in electronic medical records. However, many suggestions can be done to improve the accuracy of the proposed work such as mining the disease-gene associations from literature. Martínez-Romero et al. [30] have proposed a recommendation system for metadata that can address several challenges that face the metadata acquisition process. The proposed system depends on association rules mining to find the associations of metadata values and ontology-based semantic mappings.

In this paper, an enhanced version of the Apriori algorithm is proposed. The proposed algorithm called semantic Apriori algorithm. The objective of the proposed algorithm is to enhance the rule mining task by representing the data using ontology and modifying the traditional Apriori algorithm to deal with the ontology-based data representation form. The efficiency and effectiveness of the proposed algorithm is evaluated using a medical data set, namely, chronic kidney disease dataset. Hence, the proposed work lies at the intersection of the two classes that have been mentioned at the beginning of this section.

3. Background and preliminaries

This section includes the basic concepts and terminologies needed to understand the proposed work. In addition, it presents the description of the dataset used to evaluate the proposed framework.

3.1 An overview on association rules

Association rule mining is the process of finding strong rules that describes the correlations among the items of a certain database. This problem was introduced for the first time in [31]. It was originally employed to address the shopping basket problem [31]. Assume that $I = \{i_1, i_2, \dots, i_m\}$ is a set of m items and $D = \{t_1, t_2, \dots, t_n\}$ is a database of n transactions. Each transaction is an itemset or subset of I. The support of an itemset S is defined in Eq. (1) [32]:

$$Sup\ (S) = \frac{Count\ of\ transactions\ in\ D\ that\ contain\ S}{n} \quad (1)$$

An association rule r is a statistical implication of the form $X \rightarrow Y$ where X and Y are itemsets of I and $X \cap Y = \emptyset$. X is called the antecedent of the rule while Y is called the consequent of the rule. The support and confidence of the association rule r are two measures can be represented by $Sup(r)$ and $Conf(r)$ and defined as in Eqs. (2) and (3):

$$Sup(r) = \frac{Count\ of\ transactions\ in\ D\ that\ contain\ (X \cup Y)}{n} = p(X \cup Y) \quad (2)$$

$$Conf(r) = \frac{\mathrm{Sup}(X \cup Y)}{Sup\ (X)} = p(X|Y) \quad (3)$$

The support of the association rule reflects the statistical significance of the rule while the confidence reflects the strength of the association rule [33]. Another useful measure of the association rule is called *lift* that is defined in Eq. (4):

$$Lift(r) = \frac{\mathrm{Conf}\ (X \rightarrow Y)}{Sup\ (Y)} = \frac{p(X \cup Y)}{(p(X) \times p(Y))} \quad (4)$$

If the value of *lift* is 1, then X and Y are independent. On the other side, if the value of the *lift* is greater than 1, this means that there is some relationship between X and Y and their existence together in some transaction is highly possible [32]. An association rule is said to be interesting of its support and confidence exceeds the user-defined thresholds Sup_{min} and $Conf_{min}$, respectively. Hence, the objective of the association rule mining is to find such interesting rules.

There are several association rule mining algorithms such as Apriori [34], Eclat [35], Declat [35], FP-growth [35], etc. However, the Apriori algorithm [34] is considered the most popular and widely adopted algorithm for extracting the association rules. The objective of this work is to compare the quality and significance of the association rules extracted by the Apriori algorithm from traditional database and the association rules extracted by the same algorithm using ontology. The Apriori algorithm involves two main stages, as shown in Algorithm 1. In the first stage, the itemsets that have a support value higher than Sup_{min} are extracted. In the second stage, the association rules that have support and confidence higher than Sup_{min} and $Conf_{min}$ are obtained from the itemsets produced in the first stage. The pseudocode of the Apriori algorithm is shown below:

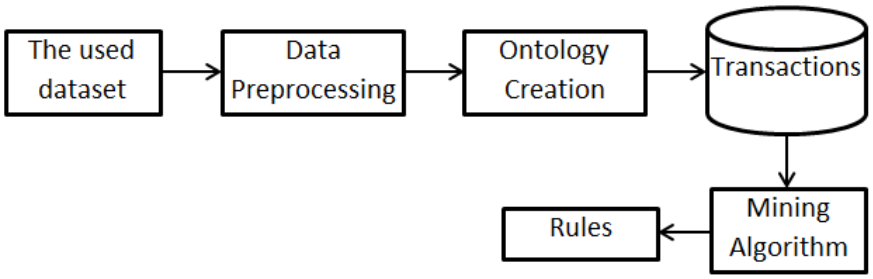

Figure. 1 The block diagram of the proposed framework for ontology-based rule mining

Algorithm 1: The Apriori Algorithm

1 C_K: Candidate itemset of size k
2 L_K: Frequent itemset of size k
3 L_1 = {Frequent items}
4 K = 1
5 **While** ($L_K \neq \emptyset$) **Do**
6 C_{K+1}= candidates generated from L_K
7 **For each** transaction t in database D **Do**
① Increment the count of all candidates
② in C_{K+1} that are contained in t.
③ L_{K+1} = candidates in C_{K+1} with Sup_{min}
8 **End**
9 k = k + 1
10 **End**
11 Return $\cup_k L_k$

3.2 An overview on ontology

Ontology languages and their corresponding query languages perform a vital role to represent information about the real world for the evolving semantic web. Several ontology languages have been developed including RDF, OWL, DAML + OIL, etc. However, they do not have the same expressive power or the same computing complexity for reasoning [36]. One definition of ontology is expressed as "An ontology is a formal, explicit specification of a shared conceptualization." In this definition, the term "conceptualization" indicates an abstract model of some phenomenon or topic in the world. The term "explicit" means that both the type of the utilized concepts and the constraints that control their usage are explicitly defined. The term "formal" means that the ontology should be understandable by the machine [37]. In the proposed framework, the traditional datasets are represented using the Web Ontology Language (OWL). OWL was developed on top of RDF and borrowed from DAML+OIL. OWL is the standard recommended by W3C for semantic web. OWL has high expressivity power as well as high computational complexity. To provide a balance between the expressivity power and the computational complexity, three OWL-based sublanguages are presented namely, OWL Lite, OWL DL, and OWL Full [36].

3.3 Dataset description

The proposed framework is evaluated using a dataset called Chronic Kidney Disease (CKD) dataset which is obtained from the UCI machine learning repository [38]. This dataset can be used to predict the chronic kidney disease. It consists of 25 attributes (13 nominal attributes, 11 numerical attributes and 1 class attribute). It contains 400 records (150 CKD and 250 NotCKD).

4. The proposed system

In this section, a general framework for ontology-based association rule mining is presented. As shown in Fig. 1, the proposed framework consists of a number of steps including data preprocessing, ontology building, encoding the constructed ontology into triples of the form "subject-predicate-object", and applying the semantic Apriori algorithm in order to extract the association rules. A detailed description for each step is given in a separate subsection.

4.1 Data preprocessing

In this step, the used dataset is preprocessed and formatted to get the best results. First, the missing values are addressed by replacing the missing values of an attribute in the dataset with the median value of that attribute. Second, the noisy data is handled by applying the normalization and balancing the data. Finally, the extracted rules are compared to the rules generated by directly applying the Apriori algorithm on the used dataset to evaluate the quality and significance of extracted association rules. However, the Apriori algorithm cannot process the

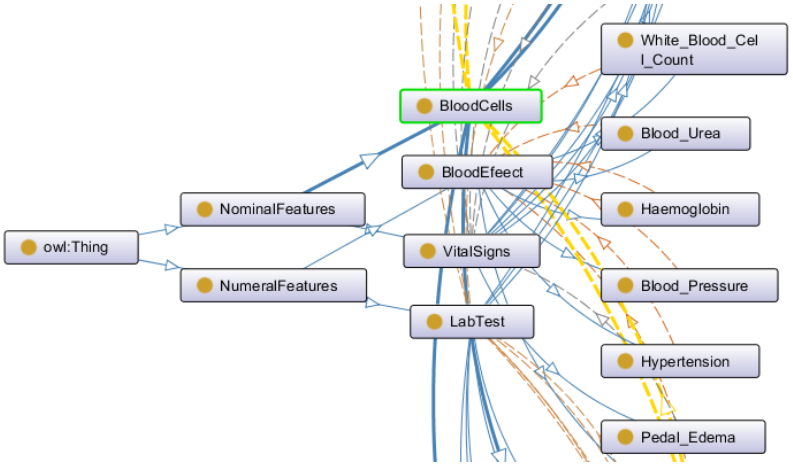

Figure. 2 Example of ontology graph fragment

numeric data without discretization. Hence, all numeric attributes in the used dataset such as age, blood pressure, blood glucose random, etc. are discretized before applying the traditional Apriori algorithm.

4.2 Ontology construction

The objective of this step is to represent the preprocessed dataset using OWL ontology. Generally, there is no single correct way to represent the domain knowledge; hence, there is no one correct way to build ontology. However, the quality of the constructed ontology highly depends on the skills of the person who is responsible for creating the ontology [39]. Many approaches have been suggested for ontology building such as *Cyc*, *Uschold and King's method*, *KACTUS*, *Methontology*, *SENSUS*, *On-to-Knowledge*, *Grüninger and Fox*, *TOVE*, *CommonKADS*, and *DILIGENT* [40-42]. In the proposed work, we have the followed a manual ontology building technique, namely *Noy and McGuinness* [43] which consists of 7 steps:

1. *Determine the domain and scope of the ontology:* This step involves answering a set of questions such as what is the domain of the ontology we are intending to construct? What the purpose of the ontology we are intending to construct? What kind of questions that ontology we are intending to construct should answer? and so on.
2. *Consider reusing existing ontologies:* Rather than building the ontology from scratch, the literature is examined to determine if there exist other ontologies that can be extended or modified.
3. *Enumerate important terms in the ontology:* The purpose of this step is to determine the concepts and the terms that the ontology we are intending to construct should cover.
4. *Define the classes and the class hierarchy:* The purpose of this step is to determine the classes that should be included in the ontology as well as the class hierarchy that can be achieved using the top-down approach, the bottom-up approach, or the middle-out approach.
5. *Define the properties of classes (slots):* The internal structure of the classes is determined including the attributes or properties of these classes. These attributes are defined as the slots of the models.
6. *Define the facets of the slots:* Slots can have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take.
7. *Create instances:* The last step is creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires (1) choosing a class, (2) creating an individual instance of that class, and (3) filling in the slot values.

A sample of the ontology graph is shown in Fig. 2.

After creating the ontology, we specify the kind of patterns we interested in to obtain from the ontology. Since the domain knowledge is represented using OWL ontology, we have extended the SPARQL with a statement to specify the patterns we are interested in. The code snippet in SPARQL is shown in Fig. 3 as well as the obtained results.

The objective of SPARQL code is to obtain the knowledge exist in the ontology as triples in the form (*subject, predicate, object*) where the subject refers to the attribute name, the predicate indicates

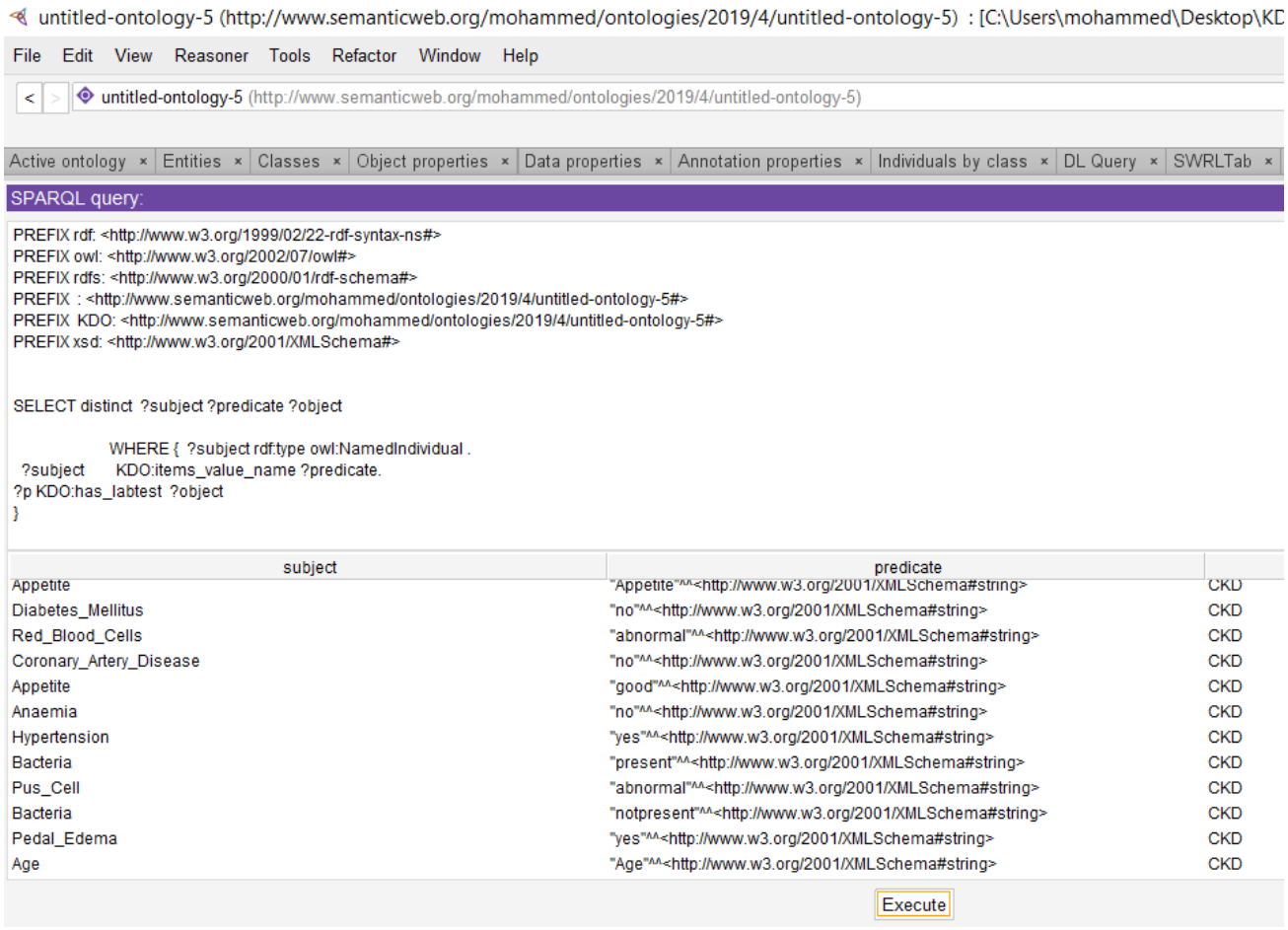

Figure. 3 The SPARQL code snippet and the obtained results

the attribute value in the current instance, and the object indicate the class name (CKD or NotCKD).

4.3 Association rule mining

In this section, the Apriori algorithm is used to extract the frequent itemset from the file that contains the triples of (*subject*, *predicate*, *object*). The pseudocode of the semantic Apriori algorithm is shown below.

Algorithm 2: The proposed Semantic Apriori Algorithm

1 C_K: Candidate itemset of size k
2 L_K: Frequent itemset of size k
3 Set the value of Sup_{min}
4 Load the (Subject, Predicate, Object) file.
5 Preprocess the loaded file.
6 Find the frequent itemset (L_K) from C_K that has support value >= Sup_{min}
7 Generate the strong rules of support and confidence values greater than or equal to Sup_{min} and $Conf_{min}$, respectively.

```
Pus_Cell_Clumps,present,NORMAL
Haemoglobin,16,CKD
Diabetes_Mellitus,no,CKD
Pus_Cell,normal,NORMAL
Pedal_Edema,no,NORMAL
Serum_Creatinine,1,CKD
Serum_Creatinine,18,NORMAL
Sugar,4,CKD
Blood_Urea,60,CKD
Packed_Cell_Volume,42,NORMAL
Potassium,7,CKD
Pus_Cell,abnormal,CKD
Bacteria,notpresent,CKD
Sugar,5,CKD
Coronary_Artery_Disease ,yes,NORMAL
Potassium,2,CKD
White_Blood_Cell_Count ,12000,CKD
Bacteria,present,NORMAL
Diabetes_Mellitus,yes,NORMAL
Red_Blood_Cells,abnormal,CKD
Packed_Cell_Volume,22,CKD
White_Blood_Cell_Count,10000,NORMAL
Red_Blood_Cell_Count,4,CKD
Sugar,3,NORMAL
Sodium,125,CKD
Blood_Urea,160,NORMAL
Serum_Creatinine,12,CKD
Appetite,Appetite,NORMAL
Appetite,poor,CKD
Diabetes_Mellitus,no,NORMAL
Sugar,2,NORMAL
Potassium,2,NORMAL
```

Figure. 4 A sample of the file that contains the (Subject, Predicate, Object) triples

```
Subject   Pus_Cell_Clumps           Predicate   present      Object   NORMAL
Subject   Haemoglobin               Predicate   16           Object   CKD
Subject   Diabetes_Mellitus         Predicate   no           Object   CKD
Subject   Pus_Cell                  Predicate   normal       Object   NORMAL
Subject   Pedal_Edema               Predicate   no           Object   NORMAL
Subject   Serum_Creatinine          Predicate   1            Object   CKD
Subject   Serum_Creatinine          Predicate   18           Object   NORMAL
Subject   Sugar                     Predicate   4            Object   CKD
Subject   Blood_Urea                Predicate   60           Object   CKD
Subject   Packed_Cell_Volume        Predicate   42           Object   NORMAL
Subject   Potassium                 Predicate   7            Object   CKD
Subject   Pus_Cell                  Predicate   abnormal     Object   CKD
Subject   Bacteria                  Predicate   notpresent   Object   CKD
Subject   Sugar                     Predicate   5            Object   CKD
Subject   Coronary_Artery_Disease   Predicate   yes          Object   NORMAL
Subject   Potassium                 Predicate   2            Object   CKD
Subject   White_Blood_Cell_Count    Predicate   12000        Object   CKD
Subject   Bacteria                  Predicate   present      Object   NORMAL
Subject   Diabetes_Mellitus         Predicate   yes          Object   NORMAL
Subject   Red_Blood_Cells           Predicate   abnormal     Object   CKD
Subject   Packed_Cell_Volume        Predicate   22           Object   CKD
Subject   White_Blood_Cell_Count    Predicate   10000        Object   NORMAL
Subject   Red_Blood_Cell_Count      Predicate   4            Object   CKD
Subject   Sugar                     Predicate   3            Object   NORMAL
Subject   Sodium                    Predicate   125          Object   KD
Subject   Blood_Urea                Predicate   160          Object   NORMAL
Subject   Serum_Creatinine          Predicate   12           Object   CKD
Subject   Appetite                  Predicate   Appetite     Object   NORMAL
Subject   Appetite                  Predicate   poor         Object   CKD
Subject   Diabetes_Mellitus         Predicate   no           Object   NORMAL
Subject   Sugar                     Predicate   2            Object   NORMAL
Subject   Potassium                 Predicate   2            Object   NORMAL
```

Figure. 5 Associating the different values with the corresponding attributes

As shown in Algorithm 2, the value of Sup_{min} is set and the file that contains the triples of (*subject, predicate, object*) is load and preprocessed. A sample of the original file is shown in Fig. 4. In the preprocessing step, the file is scanned line by line to remove any URL values. In addition, any triple that contain a *null* value is removed.

After applying the preprocessing step, the frequent itemset is determined through a number of operations. One of these operations is to associate the subject term with the first value, the predicate term with the second value and the object term with the third value, as shown in Fig. 5. Finally, the triples that satisfy the value of Sup_{min} are used to generate the frequent itemset (L_K).

5. Implementation and results evaluation

In this section, the proposed semantic Apriori algorithm is implemented using JAVA programming language with the help of Jena API. In order to evaluate the quality and significance of the rules generated using the semantic Apriori algorithm, the traditional Apriori algorithm is applied directly on the used dataset through its implementation using JAVA Programming language. The association rule mining process involves two main stages: extraction of frequent itemsets and rule generation.

In this section, a set of experiments are conducted to evaluate the performance of Apriori algorithm and the proposed semantic Apriori during the two stages. All the experiments have been done on a machine with intel(R) core i5-2430m CPU @ 2.40GHz 2.40 GHz and 8 GB RAM. In the first experiment, the performance of both algorithms is evaluated during the extraction of the frequent itemsets in terms of the number of frequent itemsets with different values of Sup_{min}. The obtained results of this experiment are shown Table 1 as well as visualized in Fig. 6.

Based on Table 1 and Fig. 6, it noticed for both algorithms that the number of generated frequent itemsets decreases when the value of Sup_{min} increases. This notice is rational because the set of generated frequent itemsets at certain value of Sup_{min} is a subset of the generated frequent itemsets of smaller Sup_{min}. Also, it is noticed that the traditional Apriori algorithm produces a lager

Table 1. The performance of both algorithms during the frequent itemsets extraction

Sup_{min}	No. of Frequent Itemsets	
	Apriori [44]	S. Apriori
0.1	74524	2207
0.2	9122	737
0.3	8350	187
0.4	30	130
0.5	5	61
0.6	0	46
0.7	0	43
0.8	0	38
0.9	0	36

Table 2. The performance of both algorithms during the association rule generation

Sup_{min}	$Conf_{min}$	No. of Generated Rules		Computation Time		Memory Consumption	
		Apriori [44]	S. Apriori	Apriori [44]	S. Apriori	Apriori [44]	S. Apriori
0.1	0.1	74524	3017	608	2.6	355	10
0.1	0.4	72692	2584	610	2.5	343	9
0.2	0.1	9122	462	70	0.6	38	3
0.2	0.3	9122	428	70	0.77	38	3
0.2	0.7	7768	254	62	0.54	26	3
0.3	0.3	8350	59	65	0.35	29	1
0.3	0.9	2841	13	27	0.31	17	1
0.4	0.1	30	60	1.9	0.31	11	4
0.5	0.2	5	28	1.6	0.3	4	2
0.6	0.5	0	16	0	0.32	0	1
0.7	0.6	0	15	0	0.29	0	1
0.8	0.1	24	29	1.56	0.34	7	1
0.9	0.1	0	29	0	0.28	0	1

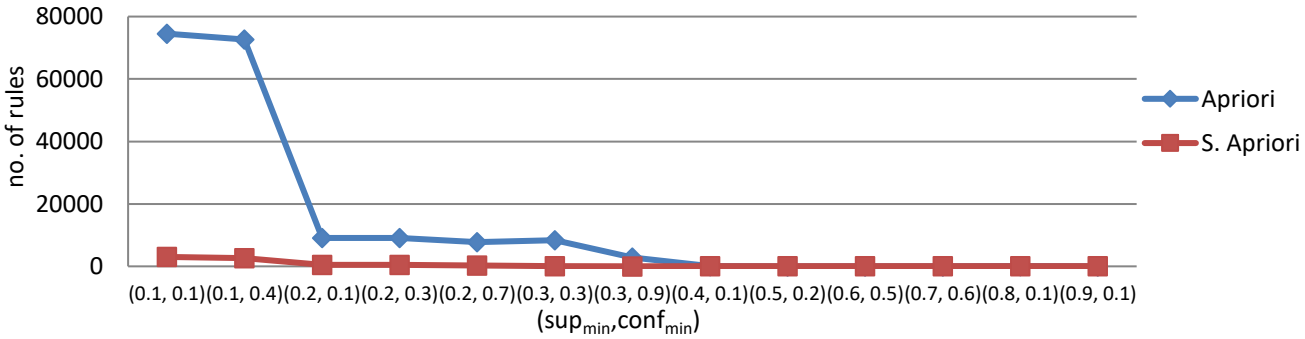

Figure. 7 The number of generated rules of each algorithm using different values of Sup_{min} and $Conf_{min}$

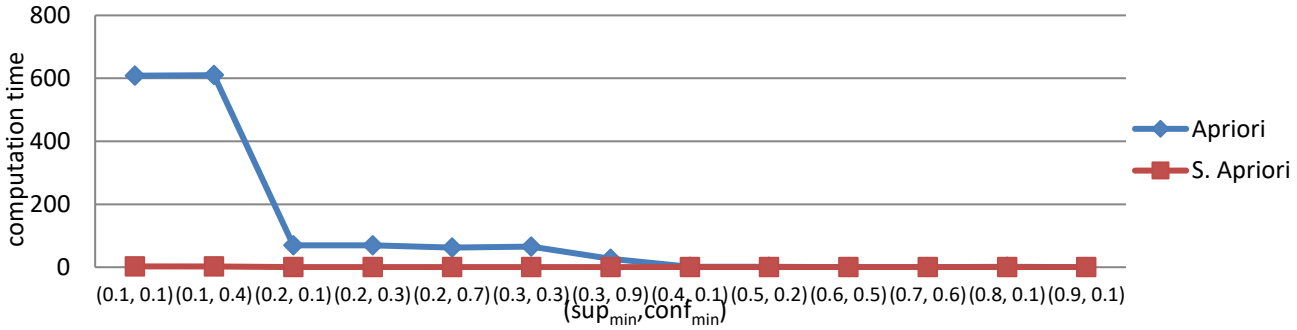

Figure. 8 The computation time of each algorithm using different values of Sup_{min} and $Conf_{min}$

set frequent itemsets compared to the semantic Apriori algorithm until the value of Sup_{min} becomes 0.4, then, the semantic Apriori algorithm generates a lager set frequent itemsets compared to the Apriori algorithm.

In the second experiment, both algorithms are assessed during the rule generation stage in terms of the number of generated rules using different values of Sup_{min} and $Conf_{min}$, computation time, and memory consumption. The obtained results of this experiment are shown in Table 2 as visualized in Figs. 7-9.

Based on Table 2 and Figs. 7-9, it is noticed that increasing the support and confidence values reduces the number of extracted rules for both algorithms. This notice is rational where many rules can fulfill the small support and confidence values while a small group of them can fulfill the higher support and confidence values. However, this small group of rules is more trusted based on the high number of records that confirm them. Also, it is noticed the proposed semantic Apriori algorithm

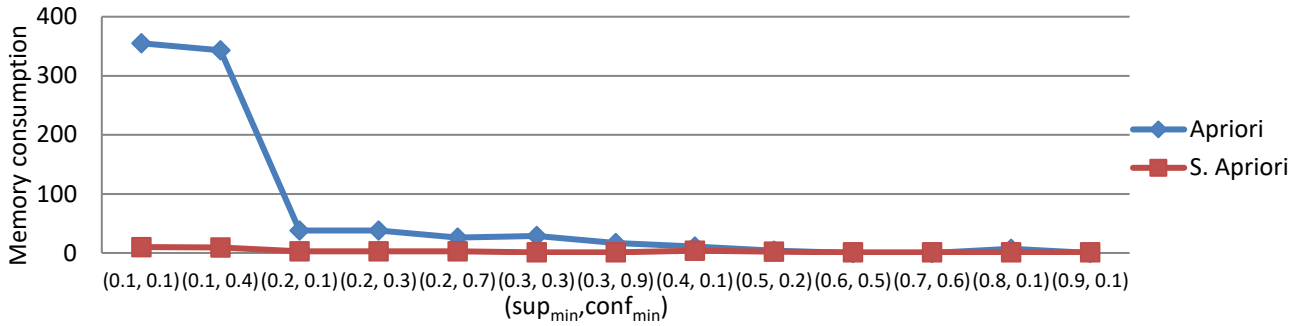

Figure. 9 The memory consumption of each algorithm using different values of Sup_{min} and $Conf_{min}$

Table 3. The $Conf_{avg}$ for the generated rules under different scenarios

Algorithm	Class	Sup_{min}				
		0.1	**0.2**	**0.3**	**0.4**	**0.5**
Apriori [44]	**CKD**	0.72	0.6	0.5	0.55	0.58
	Not-CKD	0.78	0.44	0.48	0.66	0.63
S. Apriori	**CKD**	0.96	0.89	0.87	0.93	0.79
	Not-CKD	0.74	0.91	0.85	0.95	0.86

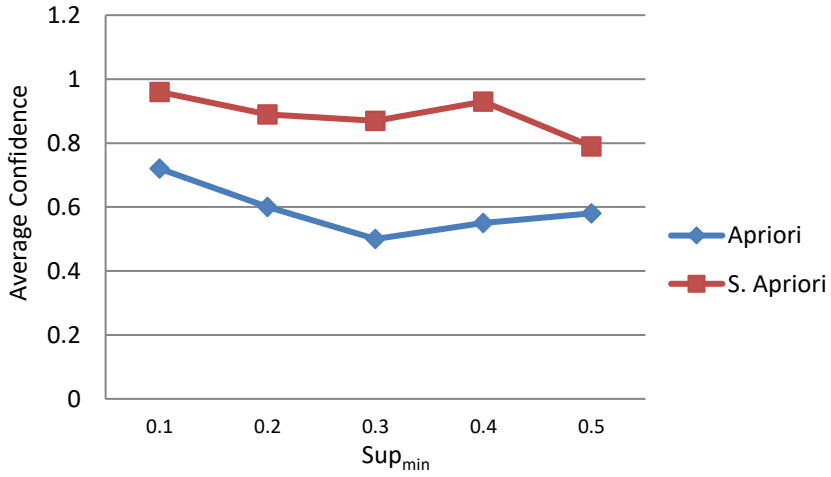

Figure. 10 The average confidence for the generated rules for both algorithms using different minimum support for the CKD class

generates a smaller set of rules compared to the traditional Apriori algorithm when the minimum support value is less than 0.4. This notice is considered an advantage for the proposed semantic Apriori algorithm form the diagnosis point of view where decreasing the number of possibilities is desired during the diagnosing of a certain case. In addition, the proposed semantic Apriori algorithm is better than the traditional Apriori algorithm in terms of computation time and memory consumption.

In the third experiment, the quality of the generated rules is assessed for both algorithms in terms of $Conf_{avg}$ which measures the average strength of the generated rules for the different classes (i.e., CKD and Not-CKD) in the used dataset. The obtained values for $Conf_{avg}$ under different scenarios are shown in Table 3.

Based on Table 3, it is noticed that the $Conf_{avg}$ of the rules generated using the proposed semantic Apriori algorithm for both classes (CKD and Not-CKD) is higher than the $Conf_{avg}$ of the corresponding rules generated using the traditional Apriori. That means that the rules generated using the proposed semantic Apriori algorithm are of higher quality than the corresponding rules generated using the traditional Apriori. The obtained results of the last experiment are visualized in Figs. 10 and 11. In addition, the best association rules that are generated using the proposed semantic Apriori algorithm are shown in Fig. 12.

The best rules the semantic Apriori algorithm are obtained using Sup_{min} value of 0.6 and $Conf_{min}$ value of 0.5 where the number of rules is nearly the same as the number of generated frequent itemsets [45].

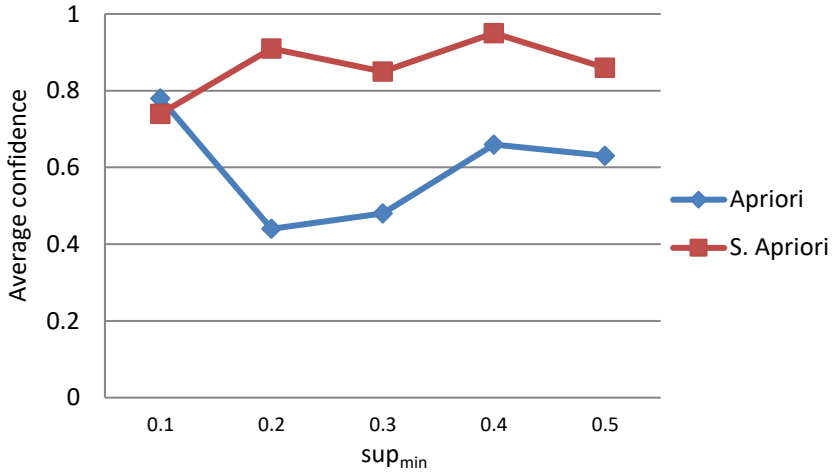

Figure. 11 The average confidence for the generated rules for both algorithms using different minimum support for the Not-CKD class

```
Best rules found:

 1. subject= SerumCreatinine=6.5 predicate= Pus Cell Clumps=notpresent ==> object= Class=notckd 200    conf:(0.57)
 2. subject= Sugar=5 predicate= WhiteBloodCellCount=19100 ==> object= Class=ckd 228    conf:(0.6)
 3. subject= Coronary Artery Disease=no predicate= Hemoglobin=6.1 ==> object= Class=notckd    conf:(0.59)
 4. subject= Pus Cell Clumps=notpresent predicate= RedBloodCellCount=4.6 ==> object= Class=ckd 208    conf:(0.58)
 5. subject= Red Blood Cells=normal predicate= RedBloodCellCount=4.7 ==> object= Class=notckd 203    conf:(0.58)
 6. subject= Sodium=113 predicate= WhiteBloodCellCount=7300 ==> object= Class=ckd 200    conf:(0.57)
 7. subject= Pus Cell Clumps=notpresent predicate= Bacteria=notpresent 346 ==> object= Class=ckd 196    conf:(0.57)
 8. subject= Anemia=no predicate= albumin=5 ==> object= Class=ckd 190    conf:(0.56)
 9. subject= Pus Cell=normal predicate= Pus Cell Clumps=notpresent ==> object= Class=ckd conf:(0.52)
10. subject= SerumCreatinine=10 predicate= Bacteria=notpresent 339 ==> object= Class=notckd conf:(0.56)
11. subject= PackedCellVolume=45 objClass= Coronary Artery Disease=no 334 ==> predicate= Class=ckd 184    conf:(0.55)
12. subject= Hemoglobin=14 predicate= Pedal Edema=no 311 ==> objClass= Class=notckd    conf:(0.52)
13. subject= Sugar=2 predicate= Sodium=163 predicate Coronary Artery Disease=no 327 ==> object= Class=notckd conf:(0.54)
14. subject= RedBloodCellCount=3.7 predicate= Coronary Artery Disease=no 325 ==> object= Class=ckd    conf:(0.54)
15. subject= Coronary Artery Disease=no predicate= Anemia=no 313 ==> object Class=ckd 163    conf:(0.52)
16. subject= BloodGlucoseRandom=268 predicate= Pus Cell Clumps=notpresent 320 ==> object= Class=notckd    conf:(0.53)
```

Figure. 12 The best association rules generated using the semantic Apriori algorithm

6. Conclusion

Semantic data mining can be used effectively to improve the provided medical service through analyzing the huge amounts of data that exist in the healthcare field. In this study, we have presented a general framework for association rule mining based on OWL ontology and Apriori algorithm. The proposed framework is evaluated using a dataset called Chronic Kidney Disease (CKD) dataset. Additionally, different experiments have been conducted to assess the performance of the proposed semantic Apriori algorithm and the traditional Apriori algorithm during frequent items extraction and rule generation in terms of the number of generated items, computation time, and memory consumption. The obtained results have shown that the proposed semantic Apriori algorithm produces a smaller set of frequent items and association rules compared to that generated by the traditional Apriori algorithm. However, based on the average confidence of the rules generated by both algorithms, the rules generated by the proposed semantic Apriori are more trusted and more effective. In addition, the computation time and memory consumption of the proposed semantic Apriori algorithm are much lesser than those of the tradition Apriori algorithm. In the future, the proposed framework can be evaluated using larger and more challenging datasets.

References

[1] M. I. Qrenawi and W. Al Sarraj, “Identification of Cardiovascular Diseases Risk Factors among Diabetes Patients Using Ontological Data Mining Techniques”, In: *Proc. of International Conf. on Promising Electronic Technologies (ICPET)*, pp. 129-134, 2018.

[2] D. Dejing, H. Wang, and H. Liu, “Semantic Data Mining: A Survey of Ontology-based Approaches”, In: *Proc. of International Conf. on semantic computing (ICSC)*, pp. 244-251, 2015.

[3] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.
[4] H. C. Koh and G. Tan, "Data mining applications in healthcare", *Journal of Healthcare Information Management*, Vol. 19, No. 2, pp. 64-72, 2011.
[5] N. Khasawneh and C. Chan, "Active User-based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining", In: *Proc. of International Conf. on Web Intelligence*, pp. 325-328, 2006.
[6] D. Perez-Rey, A. Anguita, and J. Crespo, "Ontodataclean: Ontology-based Integration and Preprocessing of Distributed Data", In: *Proc. of International Symposium on Biological and Medical Data Analysis*, pp. 262-272, 2006.
[7] M. F. Balcan, A. Blum, and Y. Mansour, "Exploiting Ontology Structures and Unlabeled Data for Learning", In: *Proc. of International Conf. on Machine Learning*, pp. 1112-1120. 2013.
[8] A. Bellandi, B. Furletti, V. Grossi, and A. Romei, "Ontology-Driven Association Rule Extraction: A Case Study", In: *Proc. of International Workshop Contexts and Ontologies Representation and Reasoning*, pp. 657-668 2007.
[9] C. Marinica and F. Guillet, "Knowledge-based Interactive Postmining of Association Rules Using Ontologies", *IEEE Transactions on Knowledge and Data Engineering*, Vol.22, No.6, pp. 784-797, 2010.
[10] G. Mansingh, K. M. Osei-Bryson, and H. Reichgelt, "Using Ontologies to Facilitate Post-Processing of Association Rules by Domain Experts", *Information Sciences*, Vol.181, No.3, pp. 419-434, 2011.
[11] D. C. Wimalasuriya and D. Dou, "Components for Information Extraction: Ontology-based Information Extractors and Generic Platforms", In: *Proc. of International Conf. on Information and Knowledge Management*, pp. 9-18, 2010.
[12] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education Limited, Malaysia, 2016.
[13] Gene Ontology Consortium, "Creating the Gene Ontology Resource: Design and Implementation", *Genome Research*, Vol.11, No. 8, pp. 1425-1433, 2001.
[14] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray, "The Unified Medical Language System", *Yearbook of Medical Informatics*, Vol.2, No.1, pp. 41-51, 1993.
[15] C. A. Wu, W. Y. Lin, C. L. Jiang, and C. C. Wu, "Toward Intelligent Data Warehouse Mining: An Ontology-Integrated Approach for Multi-Dimensional Association Mining", *Expert Systems with Applications*, Vol.38, No.9, pp. 11011-11023, 2011.
[16] A. Bernstein, F. Provost, and S. Hill, "Toward Intelligent Assistance for A Data Mining Process: An Ontology-based Approach for Cost-Sensitive Classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.4, pp. 503-518, 2005.
[17] C. M. Keet, A. Ławrynowicz, C. d'Amato, A. Kalousis, P. Nguyen, R. Palma, R. Stevens, and M. Hilario, "The Data Mining Optimization Ontology", *Journal of Web Semantics*, Vol.32, pp. 43-53.
[18] P. Panov, S. Džeroski, and L. N. Soldatova, "OntoDM: An Ontology of Data Mining", In: *Proc. of International Conf. on Data Mining*, pp. 752-760, 2008.
[19] P. Panov, L. N. Soldatova, and S. Džeroski, "Towards An Ontology of Data Mining Investigations", In: *Proc. of International Conf. on Discovery Science*, pp. 257-271, 2009.
[20] N. Ferraz and A. C. B. Garcia, "Ontology in Association Rules", SpringerPlus, Vol. 2, No. 1, 2013.
[21] R. Idoudi, K. S. Ettabaa, B. Solaiman, and K. Hamrouni, "Ontology Knowledge Mining Based Association Rules Ranking", *Procedia Computer Science*, Vol.96, pp. 345-354, 2016.
[22] M. Barati, Q. Bai, and Q. Liu, "Mining Semantic Association Rules from RDF Data", *Knowledge-Based Systems*, Vol.133, pp. 183-196, 2017.
[23] A. Mohammadi, M. H. Saraee, and M. Salehi, "Identification of Disease-Causing Genes Using Microarray Data Mining and Gene Ontology", *BMC Medical Genomics*, Vol.4, No.1, p.12, 2011.
[24] Nagar, M. Hahsler, and H. Al-Mubaid, "Association Rule Mining of Gene Ontology Annotation Terms for SGD", In: *Proc. of International Conf. on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1-7, 2015.
[25] H. Liu, D. Dou, R. Jin, P. LePendu, and N. Shah, "Mining Biomedical Ontologies and Data Using RDF Hypergraphs", In: *Proc. of International Conf. on Machine Learning and Applications*, pp. 141-146, 2013.
[26] S. A. Mahmoodi, K. Mirzaie, and S. M. Mahmoudi, "A New Algorithm to Extract Hidden Rules of Gastric Cancer Data Based on Ontology", *SpringerPlus*, Vol.5, No. 1, p.312, 2016.

[27] S. Lakshmi, and G. Vadivu, "A Novel Approach for Disease Comorbidity Prediction Using Weighted Association Rule Mining", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-8, 2019.
[28] S. Kafkas and R. Hoehndorf, "Ontology Based Mining of Pathogen–Disease Associations from Literature", *Journal of biomedical semantics*, Vol.10, 2019.
[29] F. Shen, Y. Zhao, L. Wang, M.R. Mojarad, Y. Wang, S. Liu, and H. Liu, "Rare disease knowledge enrichment through a data-driven approach", *BMC Medical Informatics and Decision Making*, Vol. 19, 2019.
[30] M. Martínez-Romero, M.J. O'Connor, A.L. Egyedi, D. Willrett, J. Hardi, J. Graybeal, and M. A. Musen, "Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases", *Database*, Vol. 2019, 2019.
[31] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", *ACM Sigmod Record*, Vol. 22, No. 2, pp. 207-216, 1993.
[32] V. Nebot and R. Berlanga, "Mining Association Rules from Semantic Web Data", *Knowledge-Based System*, Vol. 25, pp. 51-62, 2012.
[33] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Education India, 2016.
[34] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", In: *Proc. of International Conf. on Very Large Data Bases*, pp. 487-499. 1994.
[35] M. J. Zaki, and K. Gouda, "Fast Vertical Mining Using Diffsets", In: *Proc. of International Conf. on Knowledge Discovery and Data Mining*, pp. 326-335. 2003.
[36] Z. Zhang and J. A. Miller, "Ontology Query Languages for the Semantic Web: A Performance Evaluation", *PhD diss., University of Georgia*, 2005.
[37] D. Kalibatiene and O.Vasilecas, "Survey on Ontology Languages", In: *Proc. of International Conf. on Business Informatics Research*, pp. 124-141, 2011.
[38] https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease, Last Access: 28/9/2019.
[39] Kapoor and S. Sharma, "A Comparative Study Ontology Building Tools for Semantic Web Applications", *International Journal of Web & Semantic Technology (IJWesT)*, Vol. 1, No. 3, pp. 1-13, 2010.
[40] M. Fernández-López and A. Gómez-Pérez, "A Survey on Methodologies for Developing, Maintaining, Evaluating and Reengineering Ontologies", *Deliverable 1.4 of the OntoWeb project*, Available Online: http://ontoweb.aifb.uni-karlsruhe.de/About/Deliverables, 2002.
[41] O. Corcho, Oscar, M. Fernández-López, and A. Gómez-Pérez, "Methodologies, Tools and Languages for Building Ontologies. Where Is Their Meeting Point?", *Data & Knowledge Engineering*, Vol. 46, No. 1, pp. 41-64 , 2003.
[42] S. Staab and R. Studer, *Handbook on ontologies*, Springer Science & Business Media, 2010.
[43] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", 2001.
[44] C. Aggarwal, M. A. Bhuiyan, and M. Al Hasan., "Frequent pattern mining algorithms: A survey", *Frequent Pattern Mining*, pp. 19-64, 2014.
[45] N. Ali, F. Mohammed, and A. A. M. Hamed, "Usage Apriori and Clustering Algorithms in WEKA Tools to Mining Dataset of Traffic Accidents", *Journal of Information and Telecommunication*, Vol. 2, No. 3, pp. 231-245, 2018.