Automated Creating a Data Warehouse from Unstructured Semantic Data

Shiva Talebzadeh
MS Student,
Computer Engineering, Islamic Azad University Tehran-south branch, Tehran, Iran

Mir Ali Seyyedi
Assistant Professor of Software engineering and Computer Science,
Computer Engineering, Islamic Azad University Tehran-south branch, Tehran, Iran

Afshin Salajegheh
Assistant Professor of Software engineering and Computer Science,
Computer Engineering, Islamic Azad University Tehran-south branch, Tehran, Iran

ABSTRACT
The use and analysis of unstructured semantic data such as text files, semantic web, and etc which have been included the bulk of the available data sources, such as Internet due to their high volume and the extent of their resources is very complicated and time consuming process. In this context, the ontology for modeling and better understanding this data is used. So far, various methods for creating a relational database from ontology are proposed, however, creating of data warehouse from these unstructured data due to the possibility of using OLAP, top speed of reporting and its other benefits has received much attention. Numerous approaches and implementations of creating data warehouse from semantic data have been introduced. However, they still have the following disadvantages: human-dependence, semi-automatic, losing data and relations, etc. In this paper, a solution for automated creating a data warehouse from unstructured semantic data using ontology is proposed. So that, in the absence of a business analyst, all classes and relationships and data defined in the ontology data is created in a data warehouse format. In this paper, showing ontology using Protégé, language OWL and RDF is done.

General Terms
Data warehouse, Ontology

Keywords
Ontology, Data warehouse, OLAP, Semantic Data

1. INTRODUCTION
Nowadays, creating a data warehouse from heterogeneous data sources is one of the concerns of specialists in this field. There are various strategies and techniques for creating a data warehouse based on relational data sources, but in non-relational data sources such as text files, semantic web, and ... that is one of the largest sources of available data, these techniques are not accountable and do not include its specific procedures. With the proliferation of sources immigration from structured status to unstructured status, the existence of such these data warehouses to create data cubes, data mining and business intelligence analysis (BI) based on this data is necessary and has been become a serious challenge. In this context, the semantic data analysis due to the large size of them requires a technique for its original modeling. Ontology as a powerful tool to show and express the knowledge related to a domain in a formal format and can be processed by machine, is proposed. Data on the Semantic Web have to be more intelligent to be understood by the machine. Therefore, more concepts with data must be stored. In fact, the ontology specifies relationship between the concepts in Web documents and real world. The most common languages such as RDF and OWL to creating ontology are used. Various solutions have been proposed for creating Data warehouse from ontology, but most of them are semi-automatic and existence of business analyst in the field is mandatory. In addition to the above, some ontology classes and available relations have not been considered in the design and manufacture of data warehouse and final data warehouse covers only a part of the ontology and available data in it. In this paper, an algorithm is proposed which by receiving ontology in language OWL or RDF can create desired data warehouse automatically and store the available data. During the manufacturing process of data warehouse and data storing, presence of the business analyst is not required. Then, in Section 2, the researches done in this field is evaluated. Section 3, the steps of algorithm and working procedure is described in general terms. In Section 4, the steps of doing algorithm are given in detail. In Section 5, the model provided with a sample of the work done in this area, along with examples of evaluation of its strengths and weaknesses are compared. Finally, in section 6 concludes and points of development are described.

2. RELATED WORKS
For a better understanding and modeling semantic web data, ontology creating is proposed. Different methods and tools for the automatic creating of ontology has been presented which can be named ontology automatic structure and inserting data in it using semantic web techniques [4]. In this method, first, a suitable ontology for parameters and the existing relationship between data is designed automatically. After determining all the classes and hierarchy between them, in three phases, ETL processes are done that lead to extract and integrate data from multiple data sources. Another important issue is the proper use of existing ontology which the basic method has been creating relational database of the ontology. For example, the algorithm OWL2DB [13], of existing the ontology has become to the DDL, then Running Script, the final database will be created. Since in relational databases due to the high volume of various data sources, data analysis and reporting performance is extremely difficult, in this area, extensive researches has been done which a lot of them have discussed creating data warehouse from ontology. The set of researches done has led to provide methodology and algorithms for automatic and semi-automatic doing work for data warehouse from ontology.

In this regard, the main approaches provided, including creating data warehouse as semi-auto by analyst of business field. In solution [1], the main structure is based on the Meta-Data. Using the existing ontology, logical model of data warehouse designed by a business analyst then creating a physical model from the logical model and insert data from different data sources to physical model can be done automatically. In some ways, the needs of stakeholders were considered in order to create a better data warehouse [6], which includes an analysis of stakeholders' needs and providing opinions of business analysts and experts. The main objective of this process is breaking general pattern in order to reduce the existing complexity. Whatever Stakeholders needs is identified better, so there will be minimal changes in the final model. For this reason, "Requirements Analysis Method for designing ETL" (RAMERs) is based on the enterprise model and it focuses on conversion and transmission scheme it into developed scheme. Since, in order to create data warehouse, mining techniques presented has been non-automated and Extraction tables and inserting them relationships are not fully carried out, if the loss of some important links and not considered part of the existing data has been undeniable element in these methods.

3. STEPS OF ALGORITHM

The algorithm presented in five steps creates a data warehouse and inserting data in it. First, for all the existing classes, the corresponding table is created. Then, IS_A relationships between classes are examined and relationships of the tables are specified. Then, Object Property which is more conceptual relationships between classes is extracted and foreign key is created for classes. Then for each Data Property, according to the ontology, mapping will be as a new column in the available tables. The last step is insert and update data in created tables.

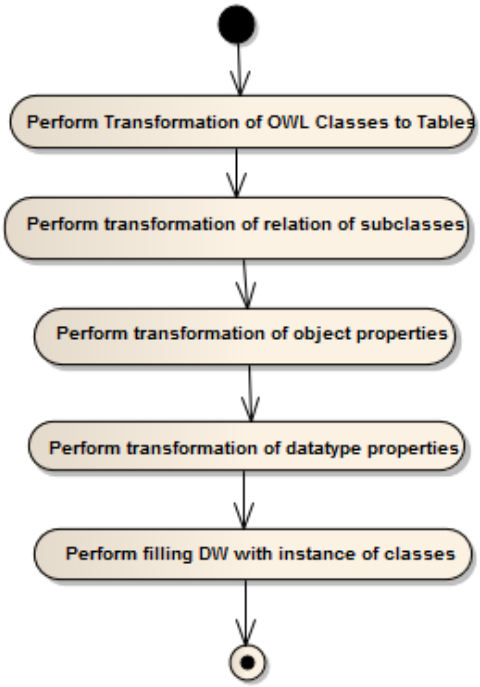

Fig 1: Steps of algorithm

4. DESCRIPTION OF THE ALGORITHM

In the proposed algorithm, ontology as input, and data warehouse including Fact and Dimension with existing relations and final data located in the tables, are considered as outputs.

4.1 Creating Tables of classes

Classes are main members of ontology. At this step, for each class that is defined in ontology with specific format, a table with the same name as the class is created. In other words, the whole ontology navigated and for available class, regardless of their relationship, a table in the data warehouse is created.

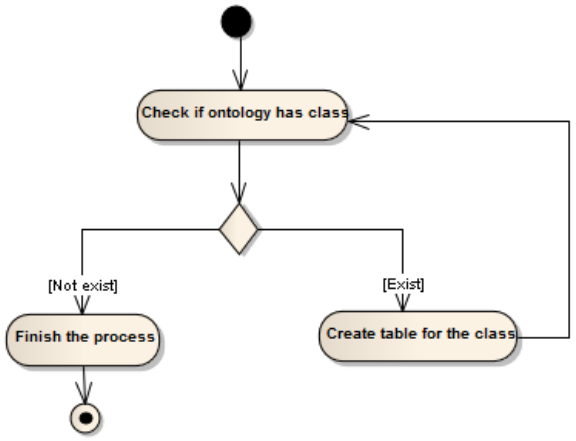

Fig 2: Creating tables of classes

4.2 Define relationships between classes

At this step, the type of SubClassOf relationship which IS_A also is called, are analyzed. This relationship is one of the limitations that it determines class c is a subclass of the class c'. Thus, each instance of c is an instance of c'. When a class can inherit from another class, all its properties can be observed in children. So, the whole ontology is navigated and whenever there is this relationship between two classes, tables related to them and relevant tables can be found in the data warehouse. Foreign key is inserted for parent and a one to many relationship is defined between these two classes.

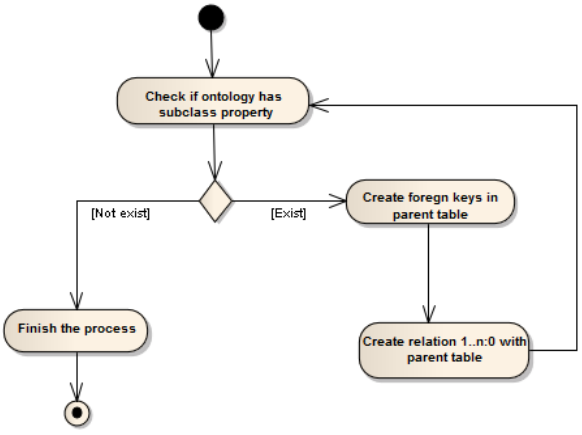

Fig 3: Define relationships between classes

4.3 Defining of Object Property relations

ObjectProprty is relationship between instances of two classes. At this step, the ontology in terms of having ObjectProperty is

navigated. For each of the ObjectProperty, Domain and Range are discussed. Domain and Range are the source and destination classes which have been converted into a table in the data warehouse. Each ObjectProperty which its Domain and Range, is not specified, in terms of having sub Property is discussed, then the Domain and Range of its parent are considered. If such a relationship to be defined for it, one to many relationship between tables of Domain and Range classes is defined.

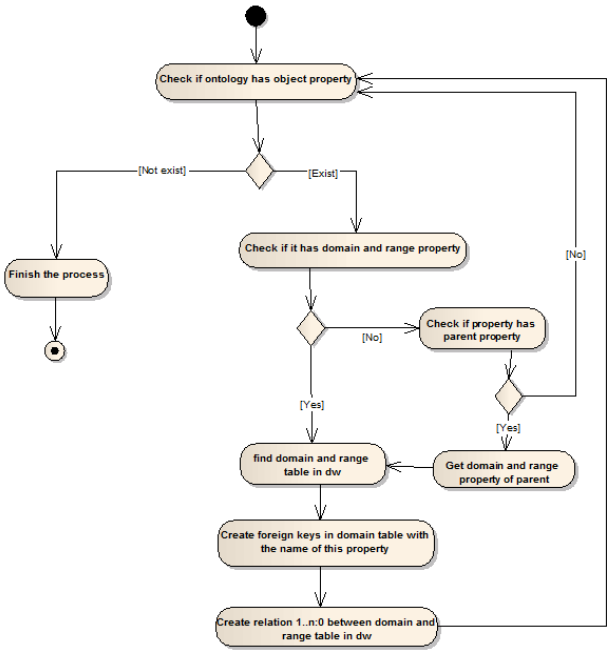

Fig 4: Defining of Object Property relations

4.4 Definition of Data Property

DataProperty or Data type in each ontology, may be is defined for each class and like ObjectProprty has Domain and Range. that Domain defines relevant class and range defines its data type. The algorithm, transforming ontology data type properties into columns of existing tables. according to Domain value it finds table and create column with the name of the property. Column data type is a set according to Range value. Sometimes Domain and Range of a Data Property is not defined. In these circumstances, if a DataProperty be inherited from another Data Property, its Domain and Range will be equal to the parents Domain and Range property.

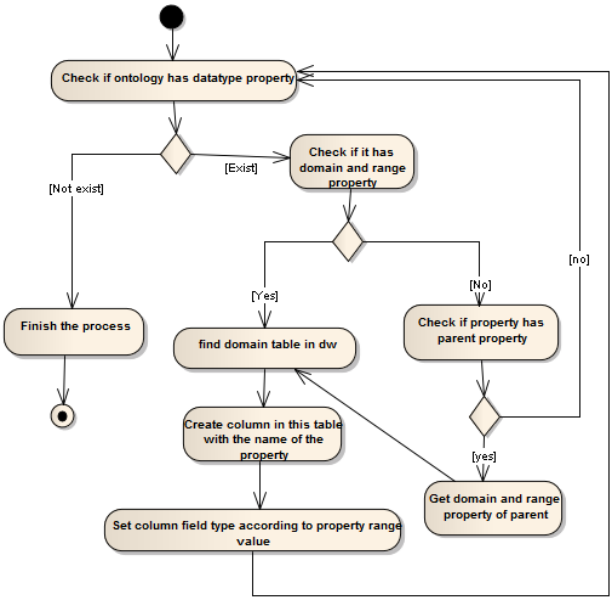

Fig 5: Defining of Data Property

4.5 Inserting data in tables

The members or instances of each class are defined in Ontology as INDEVIDUAL. Each individual has type which defines the class related to that member. At this step, for each extracted individual, class corresponding to it is specified and data with new id insert in a table with the same name of its class. Then, presence of parent for the class related to the individual is examined, if any, data and new id and foreign key are inserted in the parent tables too.

After inserting all the data in the tables, the Data Properties are checked for individuals. Here is an example of a data property for an individual:

```
<DataPropertyAssertion>
   <DataProperty IRI="#Address"/>
   <NamedIndividual IRI="#David"/>
   <Literal
datatypeIRI="&rdf;PlainLiteral">NortonPark</Literal>
</DataPropertyAssertion>
```

As already mentioned, the Data Properties have been created as a column in extant tables. For each data property, according to its domain property, it finds the table with the same name of domain property and update the row in column with the same name of data property for a given individual with specified value. if domain property is not defined, domain property of its parent is considered.

Finally Object Property for data is checked. In this case we have an object property and two individuals that is shown in this example:

```
<ObjectPropertyAssertion>
   <ObjectProperty IRI="#Drives"/>
   <NamedIndividual IRI="#David"/>
   <NamedIndividual IRI="#Q748"/>
</ObjectPropertyAssertion>
```

It finds the table of the first individual. Then if there is a column with the same name of object property in this table or its parents, it will be update this table in the column with the same name of object property with id of second individual for first individual.

If there is no column with the same name of object property, we should create a column in the table of the first individual and then update that column. This occurs because sometimes, between the individuals of each class in the ontology object property also is defined. In other words, we have an individual which based on an object property is connected to an other individual without having any connection between their classes.

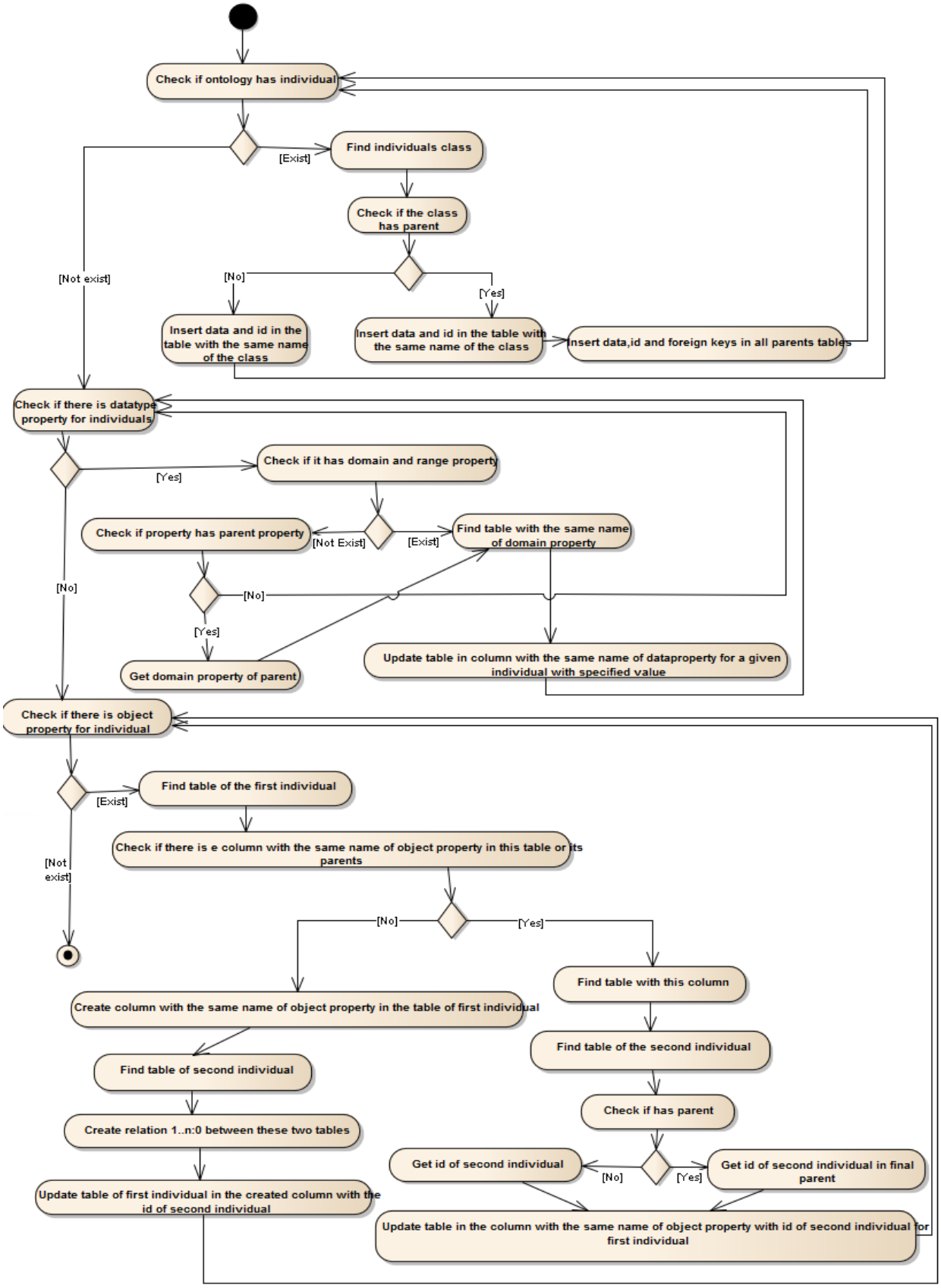

Fig 6: Inserting data in tables

5. EVALUATION

To view the performance of the proposed algorithm , all steps performed on a sample ontology and final data warehouse to be displayed. To compare the performance of this algorithm, the algorithm OWL2RDB [13] which automatically creates a relational database from ontology, is performed on the same sample and the following results are compared.

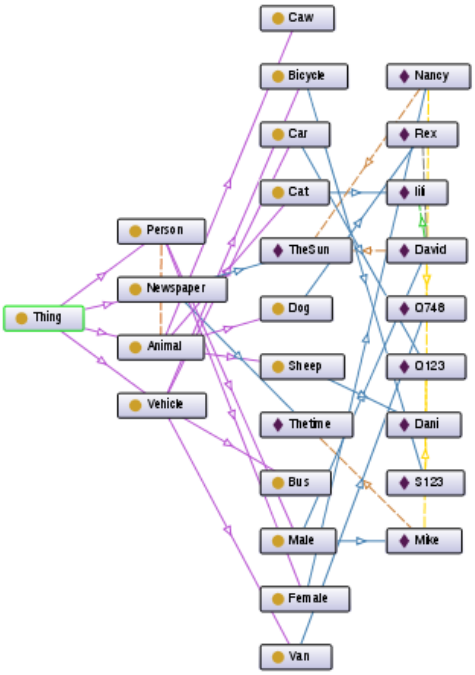

Fig 7: Example of ontology

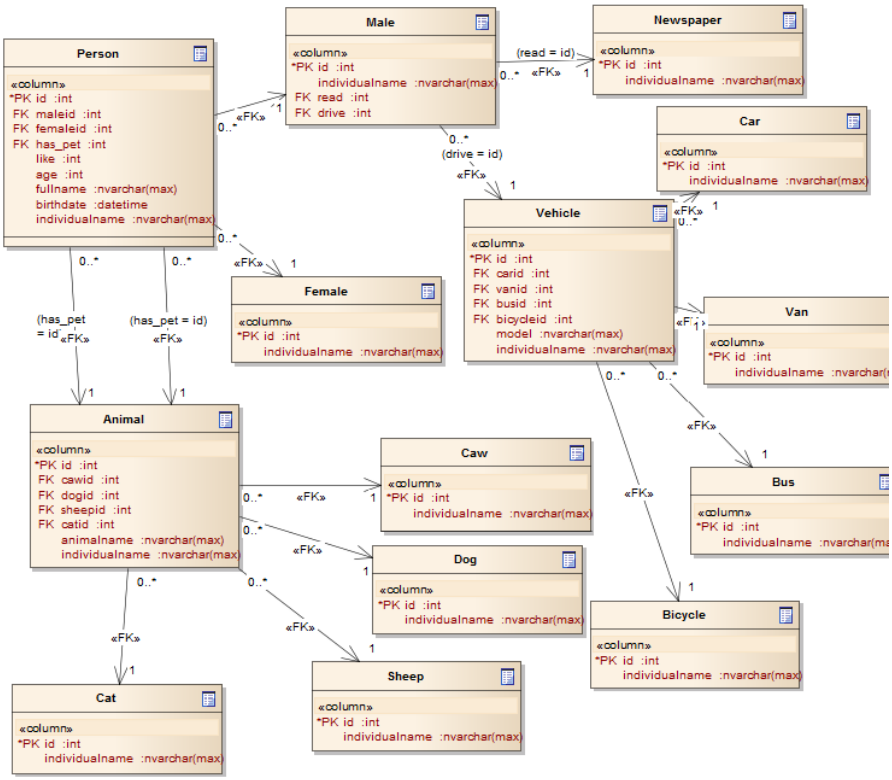

Fig 8: Final data warehouse created with OWL2DW

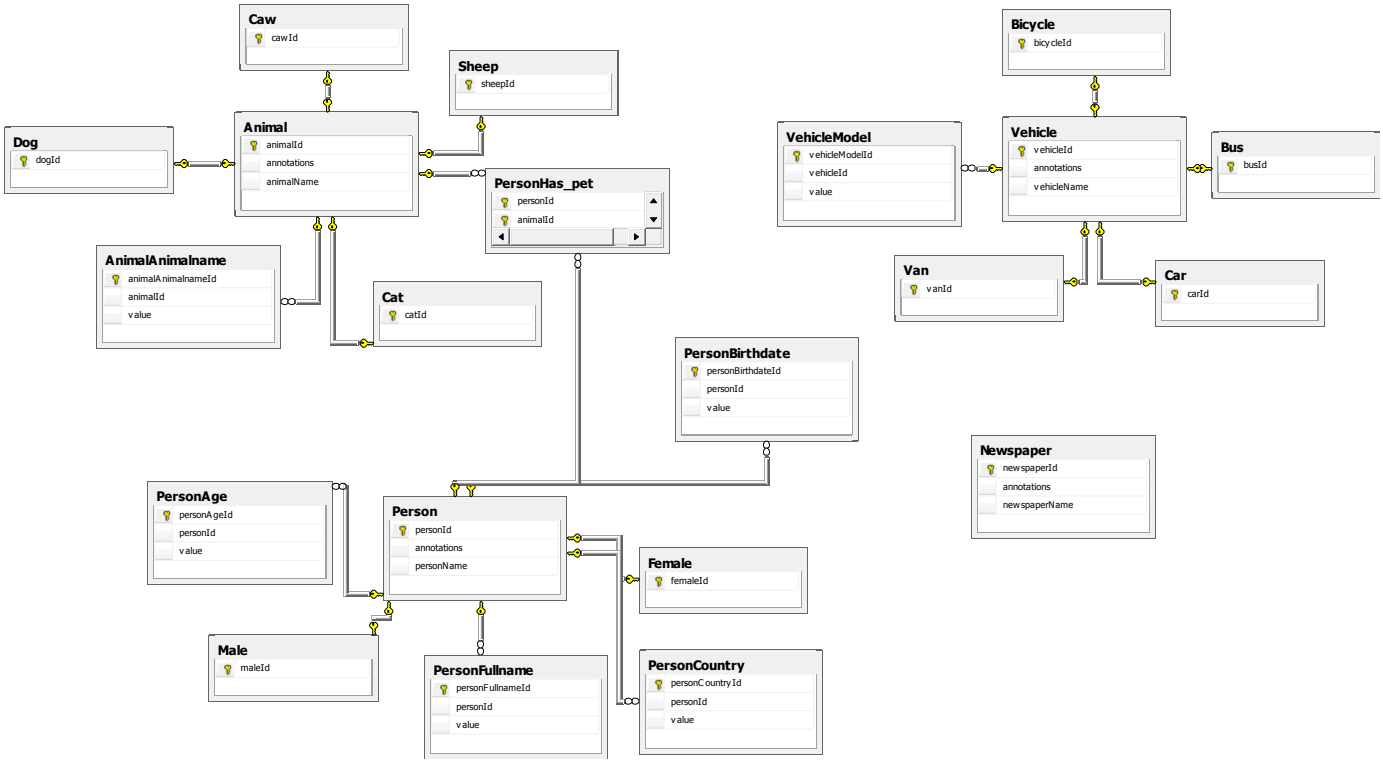

Fig 9: Final relational database created with OWL2RDB

Comparing the results obtained from the output, both the algorithm of evaluation results are specified and briefly in the following table is displayed.

Table 1. Evaluation of OWL2DB and OWL2DW

Type of evaluation	OWL2RDB	OWL2DW
Transformation of object property	Sub Property between object properties are not defined. Object property between instances are not defined.	All relations are defined
Transformation of data property	Sub Property between data type properties are not defined. All data type properties are defined as tables.	All relations are defined as column of tables
Relation between tables	Some relationships are not defined	All relations are defined
Number of tables	Because of part 2 a lot more than number of classes	Equal to the number of classes
Inserting data	Some data were inserted	All data were inserted
Runtime	Normal	Due to further steps is high
Complexity of the final model	Due to the large number of tables is high	Normal

6. CONCLUSION

In this paper, an algorithm for creating a data warehouse from semantic data using ontology presented. The proposed algorithm as fully automatic surveys the ontology and based on classes, available relationships, and the data stored in them as individuals, creates the tables of data warehouse and inserts data in it. In order to better understand the process of working on an ontology sample, all the performing algorithm steps and the final output is presented and is compared with a similar algorithm that tries to create a relational database. In this algorithm, it is attempted that all important and effective relationships in making tables are considered. Some relationships, such Equivalent To or Disjoint With, due to the lack of need to apply in the data warehouse or not impact on it, are not considered. Restriction relationship which in general defines the cardinality of the classes, due to the nature of data warehouse which is non-normal, is not considered. To promote the proposed method in improving the efficiency of the information registration step in the high volume of data, some solutions are presented.

7. REFERENCES

[1] Joel Villanueva1,Chávez, Xiaoou Li1," Ontology based ETL process for creation of ontological data warehouse", IEEE Intelligent Systems, 16:1, 2011.

[2] Azman Taa, Mohd Syazwan Abdullah, Norita Md. Norwawi," A goal-ontology approach to analyse the requirements for data warehouse systems", ISSN: 1790-0832- Issue 2, Volume 7, February 2010

[3] O. Romero and A. Abelló, "Automating Multidimensional Design from Ontologies", in DOLAP'07, Lisboa, Portugal, 2007.

[4] Skoutas, D., & Simitsis, A. (2006). "Designing ETL Processes Using Semantic Web Technologies". In Proc.

of the 9th ACM International Workshop on Data Warehousing and OLAP.

[5] Dimitrios Skoutas, Alkis Simitsis, and Timos Sellis," Ontology-driven Conceptual Design of ETL Processes using Graph Transformations", Springer Journal on Data Semantics(JoDS), Special issue on \Semantic Data Warehouses" (JoDS XIII), LNCS 5530, pp.119-145, 2009."

[6] Azman Ta'a, Mohd Syazwan Abdullah," Goal-ontology approach for modeling and designing ETL processes", Science Direct Procedia Computer Science 3 (2011) 942–948

[7] Oscar Romero , Alberto Abelló ," A framework for multidimensional design of data warehouses from ontologies", Science Direct Data & Knowledge Engineering 69 (2010) 1138–1157

[8] Michael Granitzer , Vedran Sabol , Kow Weng Onn, Dickson Lukose and Klaus Tochtermann," Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques", Future Internet 2010, 2, 238-258; doi:10.3390/fi2030238

[9] Wache. H, Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hubner, S.: Ontology-based integration of information - a survey of existing approaches. In: Proc. of IJCAI-01 Workshop: Ontologies and Information Sharing. (2001) 108-117

[10] Azman Ta'a1, Mohd. Syazwan Abdullah2 and Norita Md. Norwawi3,"Goal-ontology ETL processes specification", Journal of ICT, 2010, pp: 15–43

[11] Victoria Nebot, Rafael Berlanga," Building Data Warehouses with Semantic Data", EDBT 2010, March 22–26, 2010, Lausanne, Switzerland

[12] Anirban Sarkar,Sankhayan Choudhury," Conceptual Level Design of Object Oriented Data Warehouse:Graph Semantic Based Model", Department of Computer Science, University of Calcutta, Kolkata, India,2009

[13] Ernestas Vysniauskas,Lina Nemuraite," Transforming Ontology Representation From owl to Relational DataBase", Department of Information System, Kaunas University of Technology, Studentu st. Kaunas, Lithuania,2006