An Ontology Based Approach to Data Mining

[1]Atiya Kazi, [2]Prof. D.T. Kurian,
[1]M.E student, [2]Head of Department,
[1]Information Technology,
[1]RMD Sinhgad School of Engineering, Pune, India

***Abstract* - Data Mining is the process of extracting potentially useful knowledge from raw data. This is usually referred to as the Knowledge discovery process in the context of databases. Recently, Ontologies have come into picture, as being an integral part of knowledge structuring, to create knowledge-intensive systems. An ontology is defined as an explicit formal conceptualization of some domain of interest which helps in the portrayal of concepts and their relationships for that domain. To build any ontology, one needs a domain expert who declares all the domain concepts and the relationships between them for a specialized domain. This paper presents the problem of assessing a given ontology for a particular criterion of an application, by typically determining which of several ontologies would best suit the current application domain. It also focuses on techniques, which incorporate ontology during the data mining process. It further proposes a methodology for building an ontology on the basis of the output of data mining result. The effects of the generated ontology are studied on improving the data mining process.**

***Index Terms* - Data Mining, ETL process, Ontology, Semantic Web Mining**

I. Introduction

The task of data mining usually involves prediction by using some variables or fields in the database which help in the analysis of future values of interest. This is followed by description, where the focus is on finding human-interpretable patterns which describe the data. To shed some light on ontology, one can say it is a formal, explicit specification of a shared conceptualization. Here the word Conceptualization refers to an abstract model of some real world phenomena. Also, ontology should be machine-readable and it should capture the shared knowledge related to the data stored within the warehouse. Ontology plays a major role in the format of Semantic Web integration[6], where in information is given well-defined meaning and the Search engines deploy ontology to find pages with words that are syntactically different but semantically similar.

A data warehouse is the crux of any decision support system which stores cleaned and integrated data for knowledge discovery during the data mining systems. To enhance the overall data warehouse mining process, one needs to adopt an intelligent data warehouse mining approach incorporated with an user preference ontology. This ontology benefits the mining process by providing intelligent assistance through the support of the ontologies. This can help users in building useful data mining models, which prevent ineffective pattern generation. They help in discovering of concept extended rules, and provide an active mechanism for rediscovering knowledge. A typical data mining process is depicted using figure 1. It shows the steps involved right from collection of raw data from several heterogeneous sources to the final output in the form of knowledge discovery.

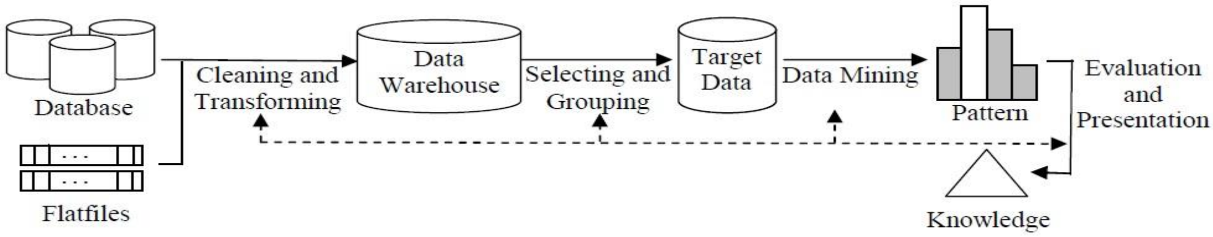

Figure 1: Steps in Data Mining

As in any data mining process, the data within the data warehouse is extracted from several heterogeneous data sources. This extracted data then undergoes a complex cleaning and integration process. Finally, it is loaded into the data warehouse where it acts as a consistent and integrated data repository. The relation between Ontologies and data mining is studied in two manners:

- From ontologies to data mining, by incorporating knowledge in the process through the use of ontologies. This helps in finding how experts comprehend and carry out the analysis tasks.
- From data mining to Ontologies, by including domain knowledge in the input information. The final analysis is done over these ontologies. Thus by including knowledge in the process through ontologies, one can transform data mining into knowledge mining.

II. Related work

Significant research has been conducted on various information system architectures such as generic domain independent systems[5], where no assumptions are made about the ontology during compile-time. There are also domain specific systems[5], where a part of the ontological schema is compiled into the object model. Such approaches simplify the work of a programmer, who can now focus on the object-model counterparts of ontological entities directly. Ontologies describe an approach for knowledge representation which can express a set of entities and their relationships along with their constraints, axioms and the vocabulary of a given domain. A survey of several ontology evaluations provides a discussion on existing ontology learning techniques [7][8]. An ontology must exist in a machine readable format so that the information systems can use them to represent and share the knowledge about the application domain. They are a powerful tool in supporting natural language processing, information filtering, information retrieval and data access. One of the greatest application of ontologies is the Semantic Web [6], in which the semantic of documents is expressed only in natural language using ontologies. This way, the Semantic Web provides an outlet for enhancing the effectiveness of Web information access. However, the manual construction of ontologies is an expensive and time consuming task handled by certain domain experts. Hence, fast and cheap ontology development is crucial for the success of knowledge based applications and the future of Semantic Web. A suggested solution for this problem is to provide an automatic or semi-automatic support for ontology construction[1]. The introduction part expresses the data mining process along with ontology definition and relevance. Section 3 introduces the ontology building concept. Section 4 discusses the overall process of combining ontology learning and data mining. Section 5 concludes this paper with a final discussion on the approached topic.

III. STEPS IN ONTOLOGY BUILDING

The Ontology building from data mining can be formulated into two phases, the data mining phase and the ontology building phase as explained below[3][4].

- The data mining phase which includes data preparation, selection, and extraction of knowledge.
- The ontology building phase which helps in building an ontology from the extracted knowledge, that acts as the output of the data mining process.

The Data mining Phase

The first step of the data mining phase is data preparation, followed by data mapping. After this data mining techniques are applied for knowledge discovery from the mapped data.

- **Data preparation** is essential for the knowledge engineer to understand the semantics of the data and sort out which tables and attributes will be used during the mining process.
- **Data Mapper** transforms the input data into an ARFF format which can be used by any machine learning tool such as WEKA. This module helps in the conversion of input data into a nominal format to suit the ontology builder's requirements. The user specifies the database table or view that needs to be mined and selects the attributes which will be used during the data mining process. The output of this module is an ARFF file containing the mined data.
- **Data mining techniques** help in discovering knowledge from the pre-processed data. The classification algorithm used could be any Neural Network based or a decision tree based algorithm. The decision tree algorithm is favoured here as it helps in the discovery of knowledge in a machine readable format.

The Ontology Building Phase

This phase automatically generates the ontology from the data mining output. It parses the output of the data mining result and generates ontology in two languages, XML & OWL. The ontology generation described above can be depicted using figure 2.

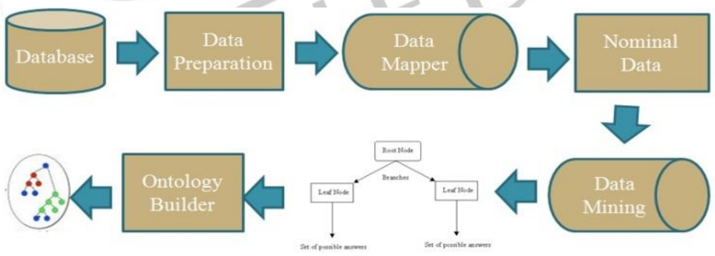

Figure2 : Steps in Ontology building

IV. COMBINING ONTOLOGY WITH DATA MINING

The fundamental principles on which the method for knowledge discovery is based on says that the knowledge discovery process is dominated by pre-existing data and the ontologies relevant to the considered domain[2]. Both data and ontologies evolve over a period of time by interacting with each other. The ontologies are enriched with knowledge from the patterns extracted with the help of the data mining tools, while the data is enriched through new inferences which are derived from the ontologies. Data mining techniques are used to produce suitable patterns that can be filtered out and selected on the basis of their integration with the ontologies. Ontologies are used to select the input of the data mining techniques, based on their common relevance. New ontological models help in abstracting and validating the existing ones on their consistency. They help in consolidating the available data leading to multiple versions of ontologies and data. They can branch over multiple iterations. The proposed data warehouse mining system framework helps in supporting the system's intelligence by incorporating ontologies in the data mining framework. It includes the characteristics of data warehouse schema, along with attributes based constraint relationships, domain specific knowledge and the user preference based ontologies.

- The mining process as depicted using figure 3 begins with the user deciding on a specific mining model[1].

- The target data will be prepared as per the user preferences specified in the mining model. This signals the launch of the mining engine.
- The user keeps tuning the mining model repetitively until the results are as accurate as possible.
- These settings of a satisfactory mining process are preserved in the mining log. It helps in further analysis to group together closely related model patterns.
- The analyzed results will be utilized in the construction of user preference ontology.
- The settings of a user’s mining model is a highly interactive process between the user and the system.
- New users might not know exactly what they want or how to initiate mining models.
- In such cases, the system provides the users with intelligent assistance in setting the mining models closer to their preferences.
- The knowledge of experienced mining users is utilized during the mining model settings, which provides recommendations for selecting the grouping attributes and the interested mining-items or judging minimal support and minimal confidence.

Specifically, the history of mining models settings satisfying the users is logged periodically, and distilled into the structure of the user preference ontology. In short, the association rule mining technique is applied over the mining log to find surrogate patterns which represent frequently used queries in the mining history. In retrospect, the mining model settings are assisted by intelligent functions which eliminate the possibilities of illegal settings. Also, appropriate recommendations of the mining model elements avoid the execution of redundant mining processes. The users are guided towards approaching the mining models closer to their mining intentions. Hence, more precise knowledge can be discovered. It provides the system with knowledge browsing capabilities that simultaneously compares a mining model with a user preference ontology for any sort of duplication or similarities. This ultimately results in saving the system’s resources.

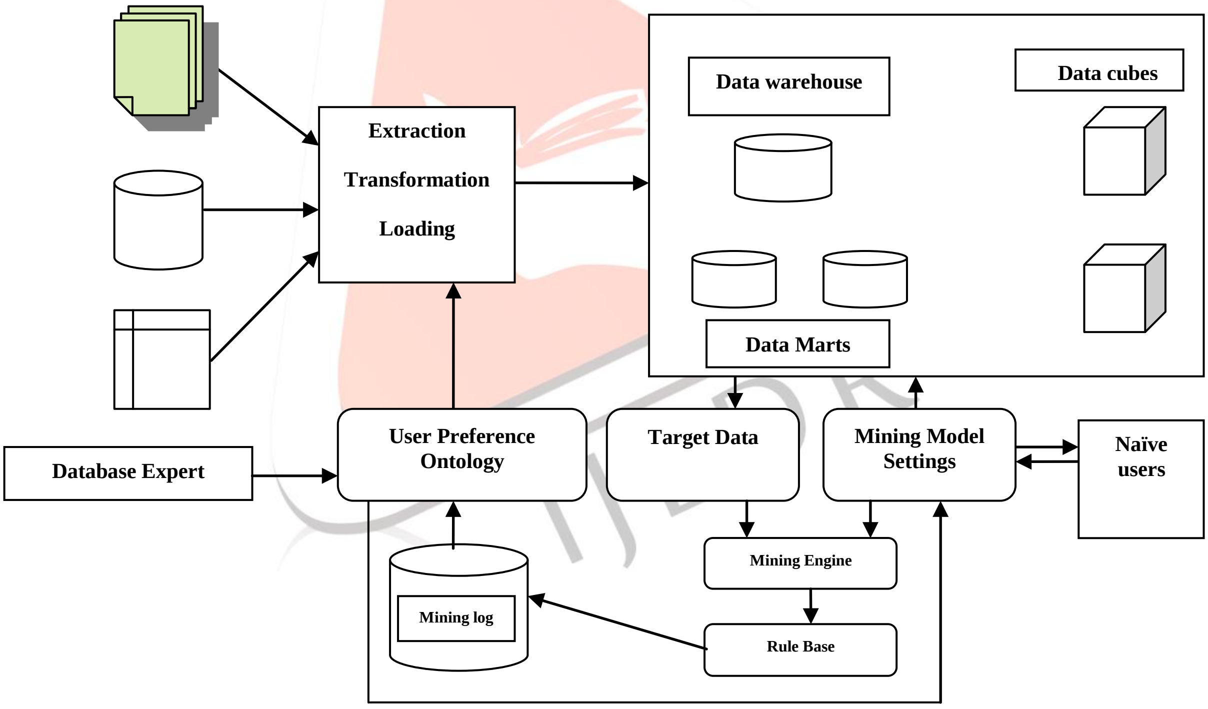

Figure 3 : Data mining using User Preference Ontology

V. Conclusion

It was observed that the focus of data mining process is to find real and useful knowledge that users actually want to pursue. This paper dealt with a data warehouse mining system framework incorporating a user preference ontology which improves the effectiveness and efficiency of mining. The methodology proposed for ontology building helps in fabricating an expert system based on the data mining result. The knowledge engineer grasps the required knowledge from the domain expert, as a guidance for data mining. As seen here, Ontologies have gained most importance in various fields such as knowledge management, information extraction, as well as the semantic web. The domain expert validates the extracted knowledge, and remembers the missed knowledge. For future work one can innovate the existing methodology, for building an ontology from unstructured data such as web pages by including side information.

REFERENCES

[1] Chin-Ang Wu et al., “Toward Intelligent Data Warehouse Mining: An Ontology-Integrated Approach for Multi-Dimensional Association Mining”, Information Technology and Management Science, Expert Systems with applications, volume 38, Issue 9, pp 11011-11023, sept-2011.

[2] Mathieu d’Aquina, Gabriel Kronbergerb, and Mari Carmen Suárez-Figueroa, “Combining Data Mining and Ontology Engineering to enrich Ontologies and Linked Data”, Proc. first International workshop on knowledge discovery and Data Mining , pp 19-24, 2012.

[3] Abd-Elrahman Elsayed, Samhaa R. El-Beltagy, Mahmoud Rafeal, Osman Hegazy, “Applying data mining for ontology building”, proc. of ISSR, 2007.

[4] Henrihs Gorskis, Yuri Chizhov, “Ontolog Building Using Data Mining Techniques”, Information technology and management science, vol 15, pp 183-188, 2013.

[5] Martin Ledvinka, Petr Kremen, “JOPA:Developing Ontology-Based Information Systems”, proc. the 13th Annual Conference Znalosti 2014, *pp 108-117, 2014.*

[6] Natalya F. Noy, “Semantic Integration: A Survey of Ontology-Based Approaches”, ACM SIGMOD Record, vol 33, issue 4, pp 65-70, 2004.

[7] Lucas Drumond, Rosario Girardi, “A Survey of Ontology Learning Procedures”, proc. 3rd workshop on ontologies and their applications, 2008.

[8] Janez Brank, Marko Grobelnik, Dunja Mladenić, “A Survey of ontological evaluation techniques”, proc. the Conference on Data Mining and Data Warehouses, pp 166-170, 2005.