A Frame Work for Ontological Privacy Preserved Mining

Geetha Mary. A and Sriman Narayana Iyengar. N.Ch.

School of Computing Science and Engineering,
VIT University,Vellore-632014,Tamilnadu, INDIA
erpgeetha@gmail.com and nchsniyengar48@gmail.com

Abstract

Data Mining analyses the stocked data and helps in foretelling the future trends. There are different techniques by which data can be mined. These different techniques reveal different types of hidden knowledge. Using the right procedure of technique result specific patterns emerge.

Ontology is a specification of conceptualization. It is a description of concepts and relationships that can exist for an agent or a community of agents. To make software more user-friendly, ontology could be used to explain both the technical and domain details. In the process of analyzing a data certain important details cannot be revealed, therefore security is the most important feature dealt in all technologies and work places.

Data mining and Ontology techniques when integrated would capitulate an efficient system capable of selecting the appropriate algorithm for a data mining technique and privacy preserving techniques also by exploring the domain knowledge using ontology.

Keywords:

Ontology, Data mining, Knowledge Discovery in Databases and Privacy Preserving Data Mining

1. Introduction

Past data and present data can be used to tune the present and future environment. Different Data Mining techniques are generated to assure the needed analysis. Different steps need to be adhered before and after applying a Data Mining algorithm for precise results. These pre and post processing steps differ according to the analyst. The analysts need to have a fine knowledge of the domain and the techniques to extract the required pattern from the data. Knowledge of the domain is depicted using different means. Ontology is a method where the domain is represented using

hierarchical class structures. While performing a data mining task, certain data or pattern could be hidden from analysis or from the pattern generated from the analysis. For hiding the rules or data, privacy preserving algorithm are utilized. In this study, all these techniques are linked together for a proficient mining.

1.1 Data Mining

Kamber refers data mining as extracting or "mining" of knowledge from large amount of data [1] and the other terms which refers the process of data mining are, knowledge mining from data, knowledge extraction, data/pattern analysis, data archeology and data dredging.

Steps in data mining process:

Step 1: Data cleaning – To remove noise and inconsistent data.
Step 2: Data Integration – Multiple data sources are combined
Step 3: Data Selection – Data relevant to the analysis task are retrieved from the database
Step 4: Data Transformation – data are transformed into forms appropriate for mining for performing summary or aggregation operations
Step 5: Data Mining – Intelligent methods are applied in order to extract data patterns
Step 6: Pattern Evaluation – To identify truly interesting patterns representing knowledge based on some interestingness measures
Step 7: Knowledge Presentation - Visualization and knowledge representation techniques are used to present the mined knowledge to the user. [1]

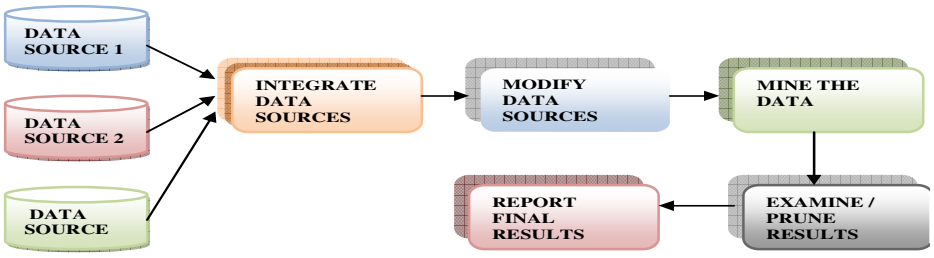

Figure 1. Steps in Data Mining Process

Different kinds of data mining techniques are specified by Bhavani Thuraisingham,

- **Classification** of group items based on a predefined attribute.
- **Association** makes correlations between items and deduces rules that define relationships. Examples include "pen and ink are purchased together" or "Aparna and Swapna travel together."

- **Clustering** groups' items based on a previously undefined attribute.
- **Prediction** forecasts trends.
- **Estimation** examines trends for clues to deduce another characteristic [2].

Different techniques help in finding out diversified types of relations among data and also suitable for different data types, so choosing the right technique for the mining is necessary.

Zheng describes the phases for building the typical application for Data Mining and they are,

- Identifying the need of Data Mining in some specific areas.
- Communication with domain experts and do request analysis.
- System design, which includes data preparation.
- System implementation, which includes model training and building.
- System evaluation [3].

Issues addressed by Zheng are,

- The common divergence between users' understanding of Data Mining techniques.
- Choice of the core Data Mining algorithm.
- Design guidelines for Data Mining system architecture.
- Methods to incorporate domain experts' knowledge.

All these concerns of Zheng are addressed in this study are stated below,

- An ontology inference rule deduces which algorithm to be selected for data mining and privacy preserving.
- Ontology is used to explain the structure of the system.
- Domain knowledge is explained using Domain Ontology.

1.2 Privacy Preserving Data Mining:

Recently, a new class of data mining methods, known as privacy preserving data mining (PPDM) algorithms, has been developed by the research community working on security and knowledge discovery. The purpose of these algorithms is to extract relevant knowledge from large amount of data, while protecting at the same time sensitive information. Several data mining techniques, incorporating privacy protection mechanisms, have been developed that facilitate one to hide sensitive item sets or patterns, before the data mining process is executed.

Privacy preserving classification methods prevent a miner from building a classifier which is able to predict sensitive data. In appendage to this privacy preserving clustering techniques have been recently proposed, which distort sensitive numerical attributes, while preserving general features for clustering analysis. A crucial issue is to determine which ones among these privacy-preserving techniques better protect sensitive information. It is also important to assess the quality of the data resulting from the modifications applied by each algorithm, as well as the performance of the algorithms. Thus there is a need for identifying a comprehensive set of criteria with respect to assessing the existing PPDM algorithms and determine which algorithm meets specific requirements.

Wang specifies two types of privacy,

- Output Privacy
- Input Privacy[4]

Input Privacy: Data is manipulated so that the mining result is not affected or minimally affected.

Output Privacy: Data is altered so that the mining result will preserve certain privacy.

Some of the techniques for output privacy are perturbation, blocking, aggregation or merging, projection and swapping. Zhang classified output privacy techniques into,

- **Perturbation based approach:** adding noise directly to the original data values.
- **Aggregation based approach:** Data are generalized according to the domain hierarchy.
- **Blocking based approach:** Sensitive attributes are truncated and not passed on.
- **Projection based approach:** Reduces the dimension, but retains the minimum information for creating data mining model [5].
- **Swapping based approach:** transferring the data between two or more fields.

Two Party Computations: Benny specifies yet another type of privacy techniques which could be applied when there are two parties. Data mining algorithm is run among two or more databases in order to find certain knowledge without revealing certain unnecessary information. Two parties should share some common data which is necessary and the data is encrypted and shared among them. A trusted third party could be used to analyze the data. A common encrypted key is shared among the parties which is unknown to the third party can be used so that the data is not revealed to the third party. A loop hole in this model is that, one of the parties and the third party could come to an understanding and the secret crypt key is shared with third party so that other party's data could be got by the other party [6].

Secure Multi party Computation: Chris accepts a computation as secure only when multi party computation at the end of the computation none of the party involved knows anything except its own input and the results [7].

1.3 Ontology

Natalya defines ontology as, a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them [8].

Reasons for the need of Ontology are,

- To share common understanding of the structure of information among people or software agents.
- To enable domain knowledge reuse.
- To make explicit domain assumptions.
- To separate domain knowledge from the operational knowledge.
- To analyze domain knowledge.

Natalya Specifies the ways ontology in which ontology is defined and used and thereby utilizing the application becomes easy and comfortable, though there are no certain specific rules to create ontology.

Developing ontology includes:

- Defining classes in the ontology,
- Arranging the classes in a taxonomic (subclass–super class) hierarchy,
- Defining slots and describing allowed values for these slots,
- Filling in the values for slots for instances.

Protégé is an ontology tool, developed by Stanford University. It works with both classes and instances at same time [9]. The top-level design of protégé performs,

- The modeling of ontology of classes describing a particular subject.
- The creation of a knowledge-acquisition tool for collecting knowledge.
- The input of specific instances of data and creation of a knowledge base.
- The execution of applications.

Thus ontology helps in understanding the concepts of a domain. The concepts are classified into classes and sub classes. Each class can have an instance and each class is related to other classes.

2. Related Work

Mon Fong has discussed about different data types and mining algorithms [10]. He says that all mining algorithms are not best suited for all data types. Selecting a mining algorithm not only depends on the pattern to be extracted but also on the data types of the data on which the mining algorithm has to be applied. Data is transformed into the required format. For data transformation 'data type's generalization process' is used.

Sridharan has developed ontology for web based learning. The flow followed contains the steps,

- Knowledge creation.
- Knowledge extraction.
- Knowledge classification.
- Knowledge retrieval.
- Knowledge sharing and use[11].

According to Luigi, many different experts can work together and each expert may use a particular vocabulary (a precise common terminology does not exist), since no rules were provided to help in the use of each term. Due to this, the vulnerabilities are,

- Synonyms may exist.
- Some term can be used in different disciplines with similar, but not identical, meanings (semantic differences appear using the same term in different disciplines) [12].

All these reasons suggest the need for the creation of a unified, complete and consistent terminology, which can be used in different formal contexts and related applications. Ontology is a practical way to achieve this goal. So, these domain concepts could be used along with data mining to get a clear picture of the application.

Lin has applied ontology to the whole data mining process itself so that it keeps track of the methods(techniques) used for preprocessing, post processing and also for mining[13].

Yen Ting talks about the ontology driven data mining on a medical driven database[14]. The database contains information about patients undergoing treatment for chronic kidney disease. Ontology was used as an aid to provide facts about the attributes of the database and also used for controlled vocabulary and the attributes with more impact are selected manually. Association Rule mining is applied on these selected attributes and antecedents are selected. Then the antecedents are

taken as new class variables and association rule is applied. The final output was good when compared to that of naive method of association rule mining.

Pawel [15] has done an Clustering analysis based on the ontology. Ontology is used to describe all the compared objects. Three dimensions were taken into consideration to calculate the similarity between individual objects and they are taxonomy similarity, relationship similarity and attribute similarity. Similarity was calculated through a formula using these dimensions. Using protégé tool, owl file is generated to describe the objects. The similarity measures were calculated and a matrix is produced. This matrix is further processed using external tools.

Wang explores the data generalization concept from data mining as a way to hide detailed information, rather than discover trends and patterns. Once the data is masked, standard data mining techniques can be applied without modification. This method used in this paper not only hides the sensitive data but also provides a appropriate mining result [16].

Jafari reveals that certain specific sensitive association rules are hidden by decreasing its support or confidence than the pre-defined minimum support and minimum confidence. To decrease the confidence of a rule $X \Rightarrow Y$, either increases the support of X, i.e., the left hand side of the rule, but not support of $X \cup Y$, or decrease the support of the item set $X \cup Y$. For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of $X \cup Y$ [17].

3. Our Approaches

New patterns can be extracted by Data mining, using different techniques an efficient and effective mining can be done. In order to achieve this goal, data mining is connected with different technologies. When an analyst does a mining, some irrelevant patterns would also be produced which are filtered in pattern evaluation phase. Irrelevant patterns possibly are minimized by selecting suitable attributes for our pattern. Proper attributes could be selected with assistance of domain expert and understanding of domain knowledge could be gained with the aid of Ontology. Technical details could also be included in Ontology to explain the whole process of mining and also to describe about various algorithms used in mining. While mining certain rules need to be hidden and also some data need to be sealed. For this reason privacy preserving algorithms are used. In my work, I use these technologies, Ontology for domain knowledge and intelligent decision making, and privacy preserving methods to preserve data leads to a proficient system.

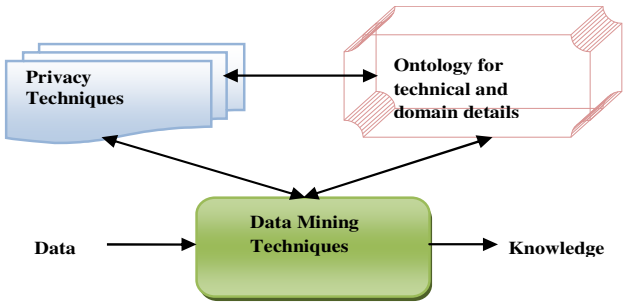

Figure 2. Ontological Privacy Preserved Mining: High Level

The various components of the system are,

Preprocessor: Preprocessor handles the data and transforms data into the desired format. The preprocessor consists of sub units and they are data cleaner, data integrator and data transformer.

Data Cleaner: Data cleaner fills the missing values and smooth out the noise if an outlier has been identified.

Data Integrator: Data Integrator merges data from several data sources and forms a solitary data source. Issues while integration like, entity identification problem and redundancy problem will be taken into consideration and correction methods will also be done by data integrator.

Data Transformer: Data Transformer converts data into required format for mining.

Domain Ontology Generation: From Domain experts', domain knowledge is gained and is expressed as an ontological structure.

Data Selector: The pertinent data for the analysis are chosen and picked from the data source at the data selector.

Concept Hierarchy: Gained domain knowledge is represented using concept hierarchy by Ontological structures. Knowledge is symbolized as classes, sub classes, slots, objects and instances.

Ontology Based Decision Taker: Based on the data selected for mining, suitable algorithm among a data mining technique and privacy preserving algorithm is offered by the Ontology based Decision Taker. It contains task ontology, object ontology, physical ontology, ontology parser, inference rule generator and mission generator as sub units.

Task Ontology: Task ontology illustrates about the choice analysis done.

Object Ontology: Object ontology depicts about the various mining and privacy preserving algorithms.

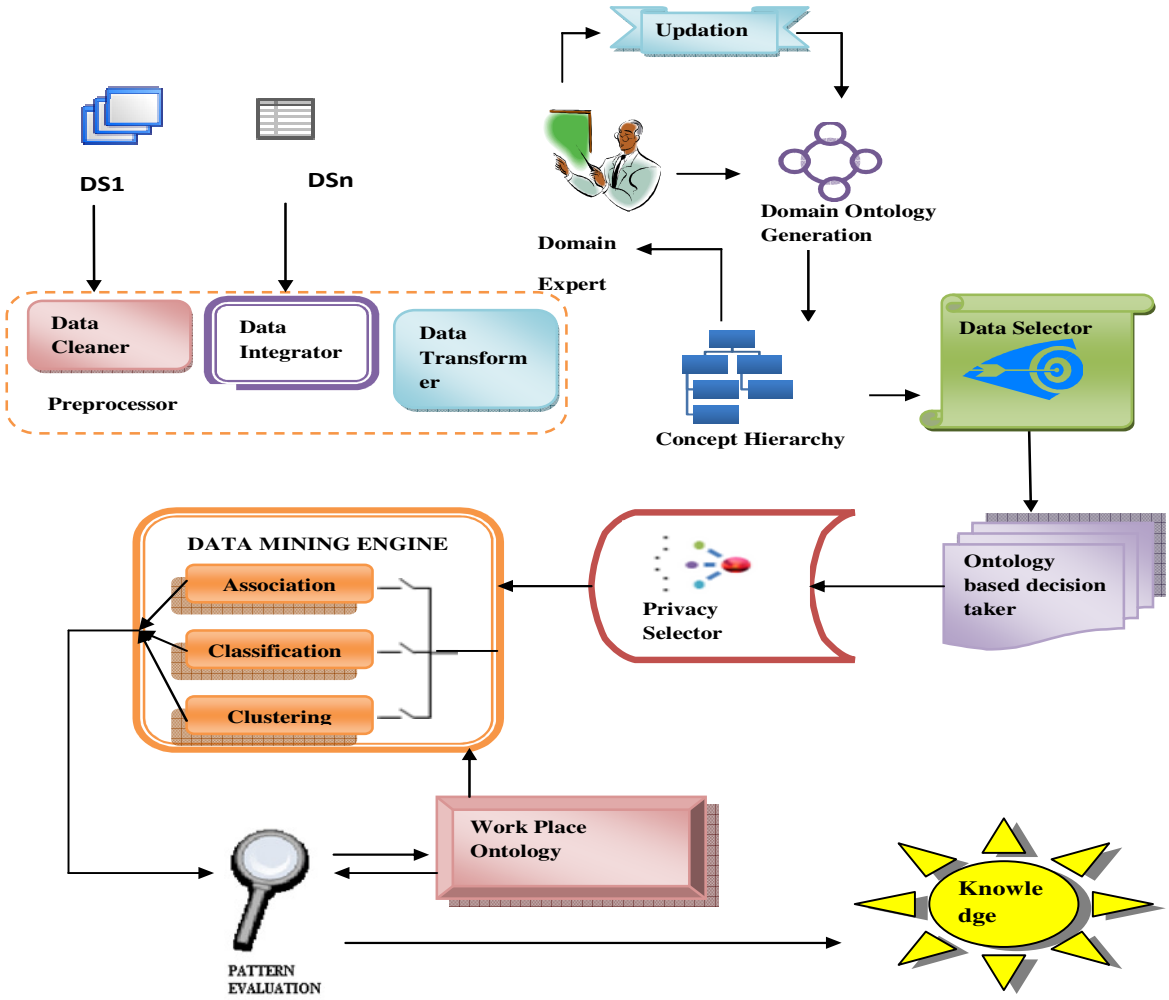

Figure 3. Architecture Diagram: Ontological Preserved Mining.

Ontology Parser: Ontology parser breaks up the diverse ontologies.

Physical Ontology: Physical ontology portrays about the flow of data across the various components.

Inference Rule Generator: Inference Rule Generator builds the rules that suggest appropriate mining and privacy preserving algorithms.

Mission Generator: Instead of recommended algorithms, the analyst could prefer algorithms other than the advised algorithms and set it as the mission in Mission Generator.

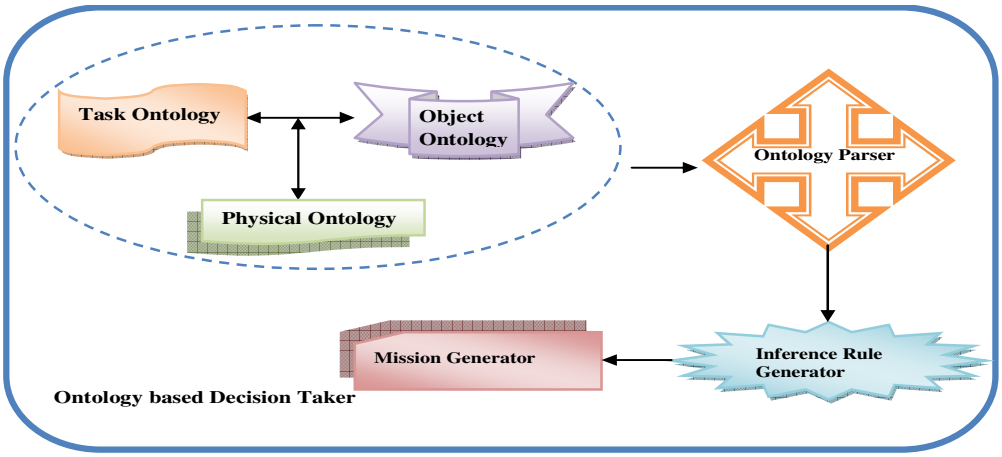

Figure 4. Ontology based decision maker.

Privacy Selector: The Chosen privacy preserving algorithm is selected and executed among the offered algorithms of Privacy Selector.

Data Mining Engine: The preferred data mining algorithm is employed in the Data Mining Engine.

Pattern Evaluation: Interesting Patterns are identified in the Pattern Evaluation.

Work Place Ontology: The measures for the patterns to be picked from the various patterns extracted, is set in the Work Place Ontology.

Knowledge: The extracted pattern is the knowledge generated from the data, which is displayed in the opted format.

4. Conclusion

In this paper, a framework is proposed for an efficient mining by qualifying the user in domain and also in technical information. Suggestion is given by the aid of ontologies for suitable data mining algorithm and privacy preserving algorithm. With this suggestion, the user could select the best data mining and privacy preserving algorithm for the data. The selection is based on the data types of the task relevant data.

5. References

[1] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2006.

[2] Bhavani Thuraisingham, "A primer for understanding and applying data mining", IT PRO, IEEE January/February 2000.

[3] Fan W, Zheng Qin, Xiao-Ling Jia, "Data Mining Application Issues In Fraudulent Tax Declaration Detection", feb 2003, IEEE.

[4] Shyue-Liang Wang and Ayat Jafari, "Using Unknowns for Hiding Sensitive Predictive Association Rules", IEEE, August 2005.

[5] Nan zhang, Wei zhao, "Privacy preserving data mining systems", IEEE 2007.

[6] Benny Pinkas, "Cryptographic techniques for privacy preserving data mining", SIGKDD Explorations, issue 2, page 12.

[7] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, Michael Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining", SIGKDD Explorations, Volume 4, Issue 2 - page 1.

[8] Natalya F. Noy and Deborah L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[9] http://protege.stanford.edu

[10] Mon-Fong Jiang Shian-Shyong Tseng Shan-Yi Liao, "Data Types Generalization for Data Mining Algorithm", IEEE, October 1999.

[11] Sridharan B., Tretiakov A. & Kinshuk (2004)," Application of Ontology to Knowledge Management in Web based Learning", Proceedings of the 4th IEEE International Conference on Advanced learning Technologies 2004.

[12] Luigi Ceccaroni, Ulises Cort´es, Miquel S`anchez-Marr, "WaWO - An ontology embedded into an environmental decision-support system for wastewater treatment plant management", Proceedings of the workshop ECAI 2000 - W09: Applications of ontologies and problem-solving methods, Berlin, Germany.

[13] Mao-Song Lin, Hui Zhang, and Zhang-Guo Yu, "An Ontology for Supporting Data Mining Process", IMACS Multiconference on Computational Engineering in Systems Applications (CESA), October 4-6, 2006, Beijing, China.

[14] Yen-Ting Kuo, Andrew Lonie, Liz Sonenberg and Kathy Paizis," Domain Ontology Driven Data Mining" , SIGKDD Workshop on Domain Driven Data Mining, (DDDM2007), August 12 2007, San Jose, California, USA, ACM, 2007.

[15] Pawel lula and Grazyna Paliwoda Pekosz, "An Ontology based cluster analysis framework", ISWC'08 October 26-30, 2008, karlruhe, Germany, ACM 2008.

[16] Ke Wang, Philip S. Yu and Sourav Chakraborty, "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection", Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04).

[17] Shyue-Liang Wang and Ayat Jafari, "Using Unknowns for Hiding Sensitive Predictive Association Rules", IEEE, August 2005.

Authors

Geetha Mary A received her M.Tech. from Vellore Institute of Technology,Vellore, India in year 2008 and B.E. from University of Madras,Tamil Nadu,India in 2000. She is working for VIT University as Assistant Professor. She is currently doing her Ph.D at VIT University.Her field of interest spans and is not limited to Computer Science and Health care management. Her research interests include Security for Data Mining, Databases and Ontology.

Dr.N.Ch.S.N.Iyengar is a Senior Professor at the School Of Computing Science and Engineering at VIT University, Vellore, Tamil Nadu India. He received M.Sc (Applied Mathematics) & PhD from Regional Engineering College Warangal (Presently known as NIT, Warangal). Kakatiya University, Andhra Pradesh, India, & M.E.(Computer Science and Engineering) from Anna University, Chennai, India. His research interests include Fluid Dynamics (Porus Media),Agent based E-Business Applications, Data Privacy, Image Cryptography, Information security, Mobile Commerce and cryptography. He has authored several textbooks and had research Publications in National, International Journals & Conferences. He is also Editorial Board member for many National and International Journals. He chaired many international conferences' and delivered invited , technical lectures along with keynote addresses beside being International programme committee member.