ISSN No. 0976-5697

Volume 3, No. 4, July- August 2012

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

Ontology based Data Mining Approach – A Survey

S.Antoinette Aroul Jeyanthi*
Ph.d Research Scholar, Department of Computer Science
Pope John Paul II College of Education
Pondicherry, India.
jayanthijames@yahoo.com

Dr.Latha Parthiban
Department of Computer Science
Pondicherry University Community College
Pondicherry, India.
lathaparthiban@yahoo.com

Abstract: The data mining process comprises of a sequence of steps ranging from data cleaning, data selection and transformation, to pattern evaluation and visualization. One of the important challenges in data mining is to extract interesting knowledge which is useful for expert users. The use of prior knowledge may significantly enhance the discovery of interesting patterns by considering the interestingness according to expert beliefs. Ontologies are used to communicate domain knowledge and for discovering patterns. Ontology is a popular research topic in data mining, knowledge engineering, artificial intelligence, etc. In this paper, we provide a survey of the ontology based data mining approaches in the past few years..

Keywords: Data mining, domain ontology, base knowledge, association mining, interesting patterns

I. INTRODUCTION

Data mining is the process of posing various queries and extracting useful and often previously unknown, and unexpected information, interesting patterns, and trends from large quantities of data, generally stored in databases. These data could be accumulated over a long period of time or they could be large data sets accumulated simultaneously from heterogeneous sources. The goals of data mining include improving marketing capabilities, detecting abnormal patterns, and predicting the future based on past experiences and current trends and extracting knowledge or interesting data patterns by applying various techniques.

In this paper section II will give descriptions about data mining primitives. Section III will present an overview of ontology. Section IV will present the survey of recently developed ontology based data mining systems. Finally, Section V contains a brief conclusion.

II. DATA MINING PRIMITIVES

A data mining task can be specified in the form of data mining query, which is input to the data mining system. Users can communicate using a set of data mining primitives designed in order to facilitate efficient and fruitful knowledge discovery[1]. The data mining query is defined in terms of the following primitives.

A. Task-relevant data:

This is the database portion to be investigated. Rather than mining entire database, the user has to specify the data on which mining is to be performed. The task- relevant data can be specified by providing information such as database or data warehouse name, the table or data cube name, conditions for data selection, relevant attributes or dimensions and data grouping criteria.

B. Knowledge type to be Mined:

This specifies the data mining function to be performed. The kinds of knowledge include concept description, association, classification, prediction and clustering and evolution analysis. In addition, the user can be more specific and provide metapatterns that all discovered patterns must match.

C. Background Knowledge:

Background knowledge is information about the domain to be mined that can be useful in the discovery process. One of the most powerful forms of background knowledge is concept hierarchies. It allows the discovery of knowledge at multiple levels of abstraction It defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Concept hierarchy is represented as a set of nodes organized in a tree, where each node, represents a concept.

D. Interestingness Measures:

Data mining process generates a large number of patterns. Only a small fraction of these patterns will actually be of interest to the given user[2]. Thus, users need to specify interestingness measures that estimate the simplicity, certainty, utility, and novelty of patterns. Interestingness measures are either objective or subjective. Objective measures are based on the structure of discovered patterns and the statistics underlying them. Subjective measures are based on user beliefs in the data.

E. Visualization of Discovered Patterns:

Data mining systems should be able to display the discovered patterns in multiple forms, such as rules, tables, pie or bar charts, decision trees, cubes. It can help users with different backgrounds to identify patterns of interest.

III. OVERVIEW OF ONTOLOGY

An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest where formal implies that the ontology should be machine-readable and shared that it is accepted by a group or community. Ontologies are generally used to specify and communicate domain knowledge. Ontologies are very useful for structuring and defining the meaning of the metadata terms that are currently collected in a domain community.

Ontology learning is inherently multidisciplinary due to its strong connection with the Semantic Web, which has attracted researchers from a very broad variety of disciplines: knowledge representation, logic, philosophy, databases, machine learning, natural language processing, image processing, etc.

A. Design Criteria:

Gruber (1995) suggests [3]five design criteria for ontologies which are considered in more detail below:

- ***a)*** Clarity: *D*efinitions should be formal, complete, objective and independent of social or computational context. This results in restricting the number of possible interpretations of a concept, thereby contributing to the effectiveness of communication between agents.
- ***b)*** Coherence: only inferences consistent with existing definitions should be allowed.
- ***c) Extendibility:*** the ontology should be designed to serve as a conceptual foundation for a range of anticipated tasks. It should be possible to extend the ontology without altering the existing definitions.
- ***d) Minimal encoding bias*:** conceptualizations should be specified at a knowledge-level. Representation choices should not be based on convenience of notation or implementation issues at a symbol-level.
- ***e) Minimal ontological commitment*:** the minimum ontological commitment sufficient to support the intended knowledge sharing activities should be allowed.

B. Basic elements of Ontology:

Ontology includes the following basic elements.

- ***a) Individuals*:** are the ”ground level” components of an ontology.Individuals can be concrete objects of a domain i.e.people, animals, automobiles, molecules... or abstract individuals i.e. numbers and words.
- ***b) Concepts*:** are collections of objects. They may contain individuals, other classes, or a combination of both.
- ***c) Attributes*:** describe the objects in the ontology.
- ***d) Relationships*:** make explicit the links between objects. A relationship can be modelled as an attribute whose value is another object in the ontology, a mathematical relation.

C. Different kinds of Ontologies:

Ontologies can differ in levels of hierarchy and the variation in the relationships between concepts and concept scope and purpose i.e. whether domain specific or describing types of concepts and relations possible in any domain

- ***a) Domain Ontology*:** models a specific domain. It represents the particular meanings of terms as they apply to that domain
- ***b) Upper Ontology (or foundation ontology):*** is a model of the common objects that are generally applicable across a wide range of domain ontologies.It contains a core glossary in whose terms can be used to describe a set of domains.
- ***c) Object Ontology*:** An ontology on “things” and “events”
- ***d) Task Ontology*:** An ontology on “doing”
- ***e) Heavy-weight ontology:*** fully described ontology including concept definitions and relations, in particular in a logical way.
- ***f) Light-weight ontology*:** partially described an ontology including typically only is-a relation

D. Ontology-Assisted Data Mining Technology:

There are two ways for ontology to assist data mining, the first is a domain ontology based on mining objects; the second is task ontology based on mining tasks.[5].

Domain ontology is a sort of special ontology to describe the specified domain knowledge. As we know, there are many reusable ontology repositories on the web, and you can also build them through the guidance of field experts.

a. Definition 1. Domain ontology:

Domain ontology is D=(C, A, R) in which C is a class set, A is an attribute set, R is a set of relationship among classes.

b. Definition 2. The depth of relationship:

Depth(C1, C2), where C1 and C2 belong to the classes of the ontology, reveals the depth of the relationship between C1 and C2. If C1 is the parent class of C2, Depth(C1, C2)=0, if both of C1 and C2 generate from the same parent class, Depth(C1, C2)=1, if there is a two-layer indirect relationship between them, Depth(C1, C2) =2, and so on.

c. Definition 3. Mining object:

Mining object is written as T=(C1, C2, ... Ci), in which Ci is the class that user requirement belongs to and it is derived from the class set of domain ontology.

Task ontology describes the information of data mining task, it can help workers to choose the most appropriate mining algorithm, so some important mining algorithms that probably contribute to knowledge discovery will not be ignored.

d. Definition 4. Task ontology:

Task ontology is denoted as S=(F, C, T, R, E). Where F is the mining function, described the type of data mining algorithms, such as classification, regression, and clustering; C is the mining conditions; T is the data type; R is the result type; E is the execution environment to provide the application scope of data mining algorithms, including the scale of the issue, complexity and so on.

IV. SURVEY OF ONTOLOGY BASED APPROACH

In this section, a comprehensive survey of recently developed ontology based approach is presented. Our focus is on the use of ontology in data mining, not on the data mining process.

A. Chin-Ang Wu et al. Approach:

In Chin-Ang Wu et al. approach [6], the authors focus on intelligent data warehouse mining. The authors proposed and implemented an ontology-integrated approach for multidimensional association mining. The proposed mining system incorporates various ontologies to fulfill the function

of intelligent assistance in mining processes such as schema ontology, schema constraint ontology domain ontology and user preference ontology. Schema ontology is used to construct the relationships between the dimensions or attributes not shown in the multidimensional model, yet can benefit the data mining process, including concept hierarchical relationships and different additive characteristics of fact measures. Schema constraint ontology is helpful for checking mining model settings. User specification of t_G and t_M can be verified with the knowledge presented in this ontology. Domain ontology is used to construct the domain expert knowledge related to the mining subject and it helps the system to find concept extended rules from the existing primitive data warehouse.

The authors employed the association rule mining technique over the mining log to find a surrogate patterns representative of frequently used queries in the mining history and then construct into the user preference ontology. The proposed system helps the users in building better mining models by providing semantic checking and model recommendations, avoids the generation of useless patterns, discover concept extended rules and provide an active knowledge re-discovering system through the support of ontologies.

B. Carlos M. Toledo et al. Approach:

In Carlos M. Toledo et al. approach [7], the authors proposed document retrieval strategy based on domain ontologies for the Organizational Knowledge Management system. This approach integrates information retrieval technology, domain ontologies and organizational management system. In this approach, the knowledge storage and retrieval strategies are implemented by domain ontologies. They enable the query refinement and reasoning processes. An additional benefit offered by ontologies is context representation.

C. Clement Jonquet et al. Approach:

Clement Jonquet et al. [8] developed the Resource Index – a growing, large-scale ontology-based index of more than twenty heterogeneous biomedical resources. The National center for Biomedical Ontology (NCBO) – BioPortal, is an open library of more than 200 ontologies in biomedicine. The authors used the NCBO as the source of ontologies for the Resource Index and used the BioPortal REST services to traverse the ontologies and to create a dictionary of terms to use for direct annotations of data elements in biomedical resources. Then, expanded annotations are created using the ontology IS-A hierarchy Finally, all the annotations are aggregated and scored taking into consideration their frequency and context. The authors proposed and build the main Resource Index user interface, a search-based interface geared towards biomedical end-users. Users can retrieve annotations in several formats such as text; tab delimited, XML, RDF and OWL. The authors adopted a horizontal approach. The use of ontologies significantly enhances recall of searches without affecting precision of the top results and improves the quality of the results and to simplify the user interaction.

D. Amal Zouaq et al. Approach:

Amal Zouaq et al. in [9] proposed an approach towards open ontology learning and filtering. The proposed approach uses the OntoCmaps , a domain-independent and open ontology learning tool, which is described in details in [ ], to extract deep semantic representations from corpora. The authors presented an experimental study that relies on a set of hypotheses to rank terms and relationships based on the structure of concept maps. OntoCmaps generates rich conceptual representations in the form of concept maps and proposes an innovative filtering mechanism based on metrics from graph theory. The authors proposed voting schemes that provide a good performance in relationship identification. The approach is evaluated against a gold standard and is compared to the ontology learning tool Text2Onto. The OntoCmaps generated ontology is more expressive than Text2Onto ontology especially in conceptual relationships and leads to better results in terms of precision, recall and F-measure.

E. Shu-Hsien Liao et al. Approach:

Shu-hsien Liao et al. in [10] and in [11] proposed an identical ontology-based data mining approach for mining customer and product knowledge from the database. The proposed approach uses the Apriori algorithm of association rules and clustering analysis. Knowledge extracted from data mining results is illustrated as knowledge patterns, rules, and maps in order to propose a suggestion and solutions to the case's firm for possible product promotion. In this approach, Portégé(2000) developed by Stanford Medical Informatics was chosen as the ontology design tool and developed ontology diagrams of consumer, sports product, the case company and endorser. This ontology diagram was employed to develop a questionnaire and through that five sorts of information about consumers collected.

F. Nelson K.Y. Leung et al. Approach:

Nelson K.Y. Leung et al. in [12] proposed an ontology-based Inter-organizational Knowledge Management. The authors proposed a selection framework to assist organizations in selecting a suitable ontology mediation approach. The knowledge reusability and mismatches reconcilability of ontology and its related mediation methods enable organizational knowledge managements to collaborate and communicate with each other. \\\\\

G. Claudia Marinica et al. Approach:

In Claudia Marinica et al. approach [14], the authors proposed and implemented a interactive approach to prune and filter discovered rules. The goal of this approach is to overcome the problem of selecting interesting association rules from the huge volumes of discovered rules. The author used ontologies in order to improve the integration of user knowledge in the postprocessing task and tested the approach over voluminous sets of rules to reduce the number of rules to several dozens or less.

V. CONCLUSION

Ontology-integrated approach have attracted significant attention over the past few years, many researchers have integrated ontology in the data mining process and proposed and implemented many ontology based approaches.. In this paper we highlighted the use of ontology in data mining on the existing approaches.

VI. REFERENCES

[1] Jiawei Han et a., "Data Mining Concepts and Techniques", Morgan Kaufmaan Seires, 2000

[2] A. Silberschatz, A. Tuzhilin., "What makes patterns interesting in Knowledge Discovery Systems", IEEE Transactions on Knowledge and Data Engineering,8(6):970-974,1996.

[3] Gruber T., "Towards Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal of Human and Computer Studies, 43(5/6), 907-928, 1995.

[4] Zhao, G., Gao, Y. and Meersman, R. (2004) Ontology-Based Approach to Business Modelling. Proceedings of the International Conference of Knowledge Engineering and Decision Support (ICKEDS2004).

[5] Lurent Brisson, Martine Colland, Nicolas Pasquier., " Improving the Knowledge Discovery Process Using Ontologies", IEEE International Workshop on Mining Complex Data ,2005

[6] Chin-Ang Wu, Wen-Yang Lin, Chang-Long Jiang., "Toward Intelligent data warehouse mining : An ontology-integrated approach for multi-dimensional association mining", Expert Systems with Applications 38 (2011) 11011-11023

[7] Carlos M. Toledo, Mariel A. Ale, Omar Chiotti, maria R.Galli, "An Ontology-driven Documant Retrieval Strategy for Organisational Knowledge Management Systems", Electronics Notes in Theoretical Computer Science 281 (2011) 21-34.

[8] Clement Jonquet et al., " NCBO Resource Index: Ontology-based search and mining of biomedical resources", Web semantics: Science, Services and Agents on the World Wide Web 9 (2011) 316-324.

[9] Amal Zouaq, Dragan Gasevic, Marek Hatala., "Towards open ontology learning and filtering", Information Systems 36 (2011) 1064-1081

[10] Shu-Hsien Liao, Jen-Lung Chen, Tze-Yuan Hsu., "Ontology-based data mining approach implemented for sport marketing" ,, Expert Systems with Applications 36 (2009) 11045-11056.

[11] Shu-Hsien Liao, Jen-Lung Chen, Tze-Yuan Hsu., "Ontology-based data mining approach implemented on exploring product and brand spectrum" ,, Expert Systems with Applications 36 (2009) 11730-11744.

[12] Nelson K.Y. Leung et al.," Ontology-based Collaborative Inter-organizational knowledge Management Network", Interdisciplinary Journal of Information, Knowledge and Management, Volume 4 , 2009.

[13] Khalid M. Albarrak and Edgar H. Sibley., " A survey of Methods that transform Data Models into Ontology Models", IEEE IRI 2011, August 2011

[14] Claudia Marinica and Fabrice Guillet, " Knowledge-based Interactive Postmining of Association Rules Using Ontologies", IEEE Transactions on Knowledge and Data Engineering, Vol 22, No.6, June 2010.