ISSN:2231-0711

Lakshmi Kuncharapu et al | IJCSET |February 2013 | Vol 3, Issue 2, 59-63

Association Rules Selection using Terminological Ontologies

Lakshmi Kuncharapu,
Dept of CSIT
Christu Jyothi Institute of Technology & Sciences, Warangal

K.Shireesha
Dept of CSE
Christu Jyothi Institute of Technology & Sciences, Warangal

Abstract-Huge volume of discovered association rules from the database, limits the usefulness of it. Generally based on statistical information, all the extracted rules are not interesting to the user and it is difficult to analyze manually. To overcome this drawback, efficient post-processing task is used to integrate the user knowledge. Thus, it is crucial to help the decision-maker with an efficient reducing rule number. Hence, to prune and filter discovered rules a new interactive approach is used. In post processing step, Ontology's and Rule Schemas supervise association rule mining. At first, Terminological Ontology's are used to extract conceptual hierarchies. Second, the Rule Schema formalism is used to express the user expectations. Weighted rule mining and filtering process can be integrated with the ARIPSO scheme. The rule-mining scheme is enhanced to handle quantitative attributes. To assist the user throughout the analyzing task, interactive framework is designed. Thus by integrating domain expert knowledge over voluminous sets of rules, the number of rules are reduced to several dozens or less.

Keywords: Clustering, Classification, and Association rules, Ontology, Rule Schema, Knowledge management applications.

I. INTRODUCTION

The Data Mining (DM) results, i.e. the models, represent relations in the data and are usually employed for classifying new data or for describing correlations hidden in the data. There are several ways to reduce the computational complexity of Association Rule Mining and to increase the quality of the extracted rules: (i) reducing the search space; (ii) exploiting efficient data structures; (iii) adopting domain-specific constraints. The first two classes of optimizations are used for reducing the number of steps of the algorithm, for re-organizing the item sets, for encoding the items, and for organizing the transactions in order to minimize the algorithm time complexity. The third class tries to overcome the lack of user data-exploration by handling domain-specific constraints. In processing queries that searchers formulate, the conventional IR query languages require the searcher to state precisely what they want. Searchers need to be able to express their needs in terms of precise queries (either in Boolean form or natural languages). However, due to searchers' lack of knowledge in the search domain (anomalous state of knowledge -- An anomaly in one's state of knowledge, or lack of knowledge, with respect to a problem faced), a query syntax formulated by searchers often does not meet the searchers' information needs. In addition, a single-term query that a normal user formulates often retrieves many irrelevant articles as well as fails to find hidden knowledge or relationships buried in content of the articles. Web technology has changed the way of people's publish, access and use information dramatically since its emergence in 1990s. Especially in recent years, the emergence of a new generation of Web environment based on XML [2], could compatible with the existing Web applications well, and can better achieve information sharing and exchange in Web. The convenience and semi-structured characteristics of its text-based make XML has been widely applied in many fields, like, in information management, electronic commerce, personalized publishing, mobile communication, online education, exchange of electronic documents, and is still expanding its range of applications. XML has become the de facto standard of Data representation and exchange on the Internet [3].

To overcome this limitation with query formulation, many IR systems provide facilities for relevance feedback, with which searchers can identify documents of interest to them. IR systems can then use the keywords assigned to these desired documents to find other potentially relevant documents. Unfortunately, the lower the support is, the larger the volume of rules becomes, making it intractable for a decision-maker to analyze the mining result. Experiments show that rules become almost impossible to use when the number of rules overpasses 100. Thus, it is crucial to help the decision-maker with an efficient technique for reducing the number of rules.

However, these IR systems fail to distinguish among the attributes of the desired documents for their relative importance to the searchers' needs. However, most of the existing post processing methods are generally based on statistical information in the database. Since rule interestingness strongly depends on user knowledge and goals, these methods do not guarantee that interesting rules will be extracted. For instance, if the user looks for unexpected rules, all the already known rules should be pruned. Or, if the user wants to focus on specific schemas of rules, only this subset of rules should be selected.

The traditional data mining technology is the main face to the structured data-based relational database, transaction database and data warehouse. Thus we cannot apply the traditional relational data-based mining methods, such as Apriori, to the semi-structured data mining directly. Therefore, to develop effective methods for XML data mining become an important issue in the field of data mining and XML technology research areas.

II. MOTIVATION

The so-called information society demands for complete access to available information, which is often heterogeneous and distributed. In order to establish efficient information sharing, many technical problems have to be solved. First, a suitable information source must be located that might contain data needed for a given task. Finding suitable information sources is a problem addressed in the areas of information retrieval and information filtering . Once the information source has been found, access to the data therein has to be provided. This means that each of the information sources found in the first step have to work together with the system that is querying the information. The problem of bringing together heterogeous and distributed computer systems is known as *interoperability problem*. In order to achieve semantic interoperability in a heterogeneous information system, the *meaning* of the information that is interchanged has to be understood across the systems. Semantic conflicts occur whenever two contexts do not use the same interpretation of the information. Goh identifies three main causes for semantic heterogeneity

• *Confounding conflicts* occur when information items seem to have the same meaning, but differ in reality, e.g. due to different temporal contexts.

• *Scaling conflicts* occur when different reference systems are used to measure a value. Examples are different currencies.

• *Naming conflicts* occur when naming schemes of information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms.

The use of ontology's for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity. Uschold and Gr¨uninger mention interoperability as a key application of ontology's and many ontology based to information integration in order to achieve interoperability have been developed. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. The main tasks of Data mining are generally divided in two categories: Predictive and Descriptive. The objective of the predictive tasks is to predict the value of a particular attribute based on the values of other attributes, while for the descriptive ones, is to derive patterns (correlations, trends, clusters, ...) that summarize the relationships in the data. The Association rule mining is one of the major techniques of data mining and it is perhaps the most common form of local-pattern discovery in unsupervised learning systems. A further criterion is the focus of the approach on the integration of information sources. We therefore do not consider approaches for the integration of knowledge bases. We evaluate the remaining approaches according to four main criteria:

Use of Ontology: The role and the architecture of the ontology's influence heavily the representation formalism of an ontology.

Ontology Representation: Depending on the use of the ontology, the inference capabilities differ from approach to approach.

*Use of Mappings***:** In order to support the integration process the ontology's have to be linked to actual information. If several ontology's are used in an integration system, mapping between the ontology's are also important.

Ontology Engineering: How does the integration system support the reuse or acquisition of ontology's?

Initially, ontology's are introduced as an ”explicit specification of a conceptualization” .Therefore, ontology's can be used in an integration task to describe the semantics of the information sources and to make the content explicit. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts. However, in several projects ontology's take over additional tasks.

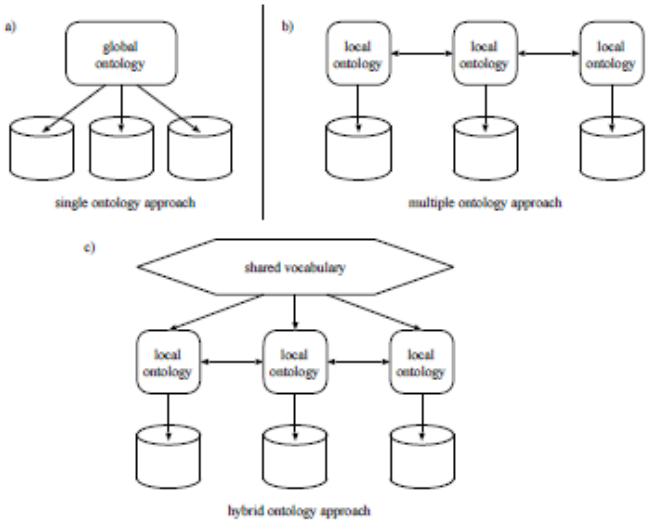

Figure 1: The three possible ways for using ontology's for content explication

2.1 Content Explication

In nearly all ontology–based integration approaches ontology's are used for the explicit description of the information source semantics. But the way, how the ontology's are employed, can be different. In general, three different directions can be identified: *single ontology approaches*, *multiple ontology's approaches* and *hybrid approaches*. Figure 1 gives an overview of the three main architectures. The integration based on a single ontology

seems to be the simplest approach, because it can be simulated by the other approaches. Some approaches provide a general framework where all three architectures can be implemented. The following paragraphs give a brief overview of the three main ontology architectures.

Single Ontology approaches- Single Ontology approaches use one global ontology providing a shared vocabulary for the specification of the semantics (see fig. 1a). All information sources are related to one global ontology. A prominent approach of this kind of ontology integration is SIMS . The SIMS model of the application domain includes a hierarchical terminological knowledge base. Each source is simply related to the global domain ontology. Existing systems that are currently in use, such as Axis Log Miner and Web Miner, will be analyzed. Zhu et al. [6] propose a new vector space retrieval algorithm based on association diagram extension of key words. By using key words and the related words appearing simultaneously in a large scale, the algorithm allows generating the association diagram which indicates the simultaneous relationship between key words. The degree of association between any two key words is represented by mutual information. In addition, the algorithm derives the weight of key words in retrieval vector via association diagram, thus the vector space retrieval based on association diagram extension of key words is realized. Martinez-de-Pison et al. [7] propose an experience based on the use of association rules from multiple time series captured from industrial processes. The main goal is to seek useful knowledge for explaining failures in these processes. An overall method is developed to obtain association rules that represent the repeated relationships between pre-defined episodes in multiple time series, using a time window and a time lag. First, the process involves working in an iterative and interactive manner with several pre-processing and segmentation algorithms for each kind of time series in order to obtain significant events. In the next step, a search is made for sequences of events called episodes that are repeated among the various time series according to a pre-set consequent, a pre-established time window and a time lag. Extraction is then made of the association rules for those episodes that appear many times and have a high rate of hits. Finally, a case study is described regarding the application of this methodology to a historical database of 150 variables from an industrial process for galvanizing steel coils. Ho et al. [8] propose artificial intelligence methodology provides investors with the ability to learn the association among different parameters.

III. HYPOTHESIS

A question that arises from the use of ontology's for different purposes in the context of information integration is about the nature of the used ontology's. Investigating this question we mainly focus on the kind of languages used, and the general structures that can be found. We do not discuss ontology content, because we think that the content strongly depends on the kind of information that has to be integrated. We further restrict the evaluation to object-centered knowledge representation systems that form the core of the languages used in most applications.

3.1 The System Architecture

The system architecture of our semantic query expansion system, SemanQE, is illustrated in Fig. 2. SemanQE consists of three major components: 1) core association rule-based query expansion 2) feature selection, and 3) ontologies-based expansion components. We use the Lemur IR system as a backend engine for SemanQE in that Lemur is robust and achieves high accuracy in terms of precision[10]. Lemur is developed by collaboration between the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University. Lemur is designed to facilitate research in language modeling and information retrieval.

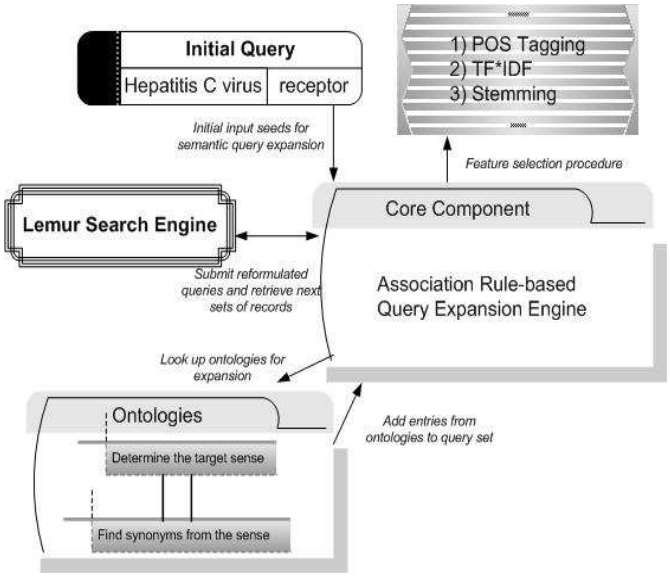

Fig. 2. System architecture of SemanQE

Ontologies component expands queries selected from the core component. Word Net is used as ontologies for our system. With a set of terms and phrases, we first disambiguate word senses based on formula proposed Word Net is then referenced to find relevant entries semantically and syntactically. The outline of the approach described in Figure 2 is as follows:

Step 1: Starting with a set of user-provided seed instances (the seed instance can be quite small), our system retrieves a sample of documents from the backend indexes via a search engine. At the initial stage of the overall document retrieval process, we have no information about the documents that might be useful for extraction. The only information we require about the target answer sets is a set of user-provided seed instances. We use some simple queries (just use the attribute values of the initial seed instances) to extract the document sample of pre-defined size from the search engine.

Step 2: On the retrieved document set, we parse each document into sentences and apply IR and natural language

processing techniques to select important terms and phrases from the input documents.

Step 3: Applying a hybrid querying expansion algorithm that combines association rules and ontologies to derive queries targeted to match and retrieve additional documents similar to the positive examples.

Step 4: Reformulate queries based on the results of Step 3 and query the search engine again to retrieve the improved result sets matched to the initial queries.

The proposed technique can be applied to ontology structures forming a directed acyclic graph.

Thus, it supports multiple inheritance. The required formal definition of input ontology's contains two core items shared by most formal definitions of an ontology in the literature: concepts and a hierarchical IS-A relation. Thus, we define a core ontology as: a pair G = (C, r), where C is a set of concepts and r is a partial order on C, i.e. a binary relation r ε C x C which is reflexive, transitive, and anti symmetric. After numbering the Ontology, the result is a hash table that includes all the nodes of the ontology with their respective integers. Then, to extract all the possible paths, with the objective to quick reach and examine with priority terminal nodes its paths, a Depth First Search (DFS) is run, that provides all possible paths number-named in a list type format.

IV. Experimental Results

For evaluation purposes, a total of 400 combinations of (s,c) are examined. Consider three Ontology G1, G2 and G3, in such a way that G2 is directly derived from G1 and G3 directly derived from G2 (see Figures 3-5). Therefore G1 is considered as the Reference Ontology that we run tests against for G2, G3.

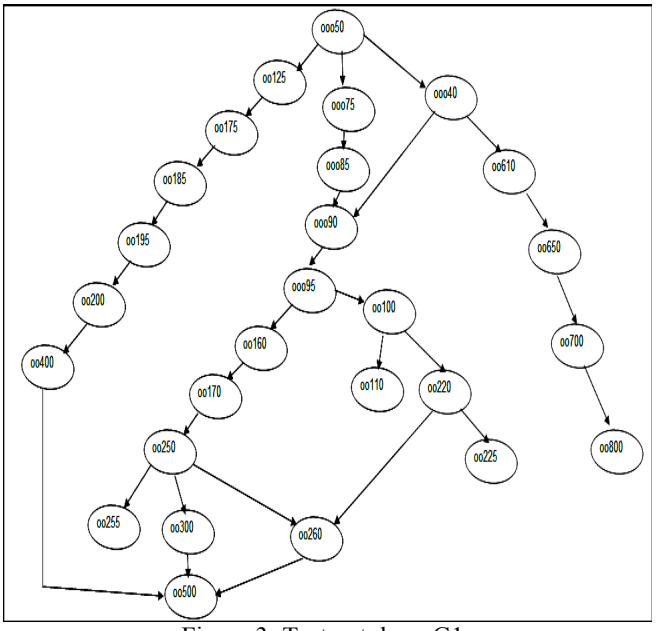

Figure 3: Test ontology G1.

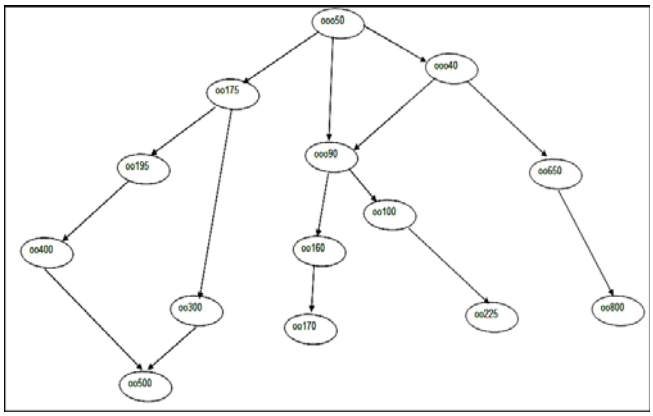

Figure 4: Test ontology G2.

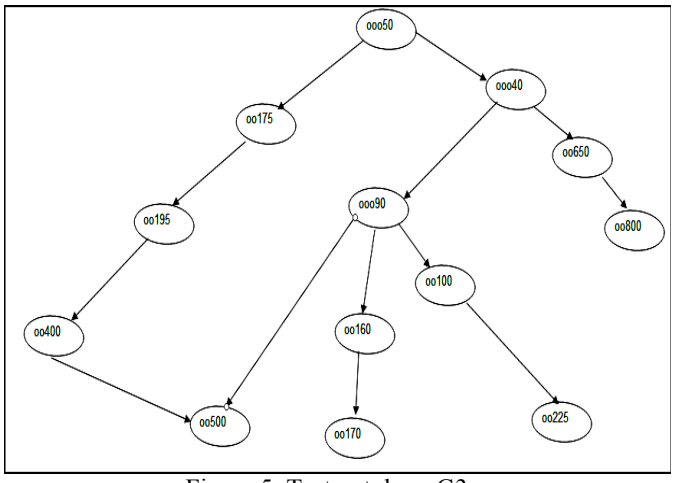

Figure 5: Test ontology G3

After applying ONARM to them in a manner of G2 against G1 and G3 against G1 we obtain the following summarized results:

a) For the comparison and mapping of the G1 and G2 Ontology, ONARM found 76 correct matching between them. The best cases were found in the areas where the minimum support was between 5% - 40% and the minimum confidence 5% - 65%. Thus, minimum support and c confidence values are not critical. The only requirement is to set low values. Theoretically, this is true because of the small number of paths of an ontology.

b) In the last column is presented the number of matches per case, as non-zero-values [NonZeroVal].

c) The average, variance and standard deviation analysis is based upon the score that has been assigned to each case, as Kt, as described in the theoretical background section.

V. Conclusions

This paper discusses the problem of selecting interesting association rules throughout huge volumes of discovered rules. The major contributions of our paper are stated below. First, we propose to integrate user knowledge in association rule mining using two different types of formalism: ontologies and rule schemas. On the one hand, domain

ontologies improve the integration of user domain knowledge concerning the database field in the post processing step. The meaning of the concepts is also taken into consideration, , by applying any linguistic analysis. Thus, it is important to note that input ontologies might have different label domains for node naming, without reducing the efficiency of the proposed methodology. The main advantages of the proposed framework can be summarized in terms of extensibility and flexibility. Other important future works are the possibility of modeling the antecedent and the consequent of an association rule as ontology concepts in order to express constraints on the association rules structure. Furthermore we could improve the system by integrating the constraints evaluation directly in the mining algorithm.

REFERENCES

[1] Y. Sebastian, H. H. Then Patrick, "Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses", Knowledge-Based Systems, Vol. 24, No. 5, 2011, pp. 609-620.

[2] J. H. Feng, G. L. Li, "Efficient fuzzy type-ahead search in XML data", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 5, 2012, pp. 882-895.

[3] Y. Lou, Z. H. Li, Q. Chen, "Semantic relevance ranking for XML keyword search", Information Science, Vol. 190, 2012, pp. 127-143.

[4] Y. H. Chang, C. Y. Wu, C. C. Lo, "Processing XML queries with structural and full-text constraints", Journal of Information Science and Engineering, Vol. 28, No. 2, 2012, pp. 221-242.

[5] Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In Ian Horrocks and James Hendler, editors, Proceedings of the first International Semantic Web Conference (ISWC 2002), number 2342 in Lecture Notes in Computer Science, pages 5468, Sardinia, Italy, June 9 12, 2002. Springer Verlag, Heidelberg Germany.

[6] Chen, X., Zhou, X., Scher, R., and Geller, J.: Using an interest Ontology for Improved Support in Rule Mining. In Proceedings of the 5th International Conference of Data Warehousing and Knowledge Discovery (DaWaK 2003), Prague, Czech Republic, pp. 320–329

[7] Han, J., and Fu, Y.: Discovery of multiple-level association rules from large databases. In Proceedings of the 21st International Conference on Very Large Data Bases (VLDB '95) San Francisco, CA, pp. 420–431

[8] Hou, X., Gu, J., Shen, X., and Yan, W.: Application of Data Mining in Fault Diagnosis Based on Ontology. In Proceedings of the 3rd Conference on Information Technology and Applications (ICITA '05), Sydney, Australia, pp. 260–263.

[9] Ng, R., T., Lakshmanan, L., V., S., Han, J., and Pang A.: Exploratory mining and pruning optimizations of constrained associations rules. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data (SIGMOD Seattle, WA, pp. 13–24.

AUTHORS

Lakshmi Kuncharapu working as an Assistant Professor in department of IT, Christu Jyothi Institute of Technology & sciences, Jangaon,Warangal,A.P, received her Master of Technology(compter science and Engineering) from Balaji Institute of Technology &Science, Narsampet, affiliated to JNTU. Her Research area includes Data Mining and Networking

K. Shireesha had received her Master of technology (computer science and engineering) from Swarnandhra College of Engineering and Technology, Narasapur, affiliated to JNTU. Currently working at Christu Jyothi Institute of Technology &sciences, Jangaon, Warangal, A.P, as an Assistant Professor in department of CSE. Her Research area includes Data Mining..