Elysium Journal
Engineering Research & Management

PISSN:2347-4408
EISSN:2347-4734

A NOVEL MODEL FOR MINING ASSOCIATION RULES FROM SEMANTIC WEB DATA

T. Anbutamilazhagan[1], Mr.K.Selvaraj[2]

[1]M.Phil, Research Scholars, Department of Computer Science, Arignar Anna Goverment Arts College, Salem.
E-mail: tlovetamilbeauty@gmail.com
[2]Head & Associate Professor, Department of Computer Science, Arignar Anna Goverment Arts College, Salem
E-.Mail: selvaraj_kumaravel@yahoo.com

Abstract

Nowadays, there is a continuous growth in the field of ontology and semantic annotations for numerous data of wide-ranging applications. This kind of heterogeneous and complex semantic data has created new challenges in the field of data mining research. An Association Rule Mining is one of the most common data mining techniques which can be well-defined for extracting the interesting relationships among the huge amount of transactions. Additionally, the Semantic Web technologies offer solutions to efficiently use the domain information. Hence this paper proposed a novel method to provide a way to address these issues and allow to process the huge volumes of semantic data. It executes association rule discovery to store the new semantic rules using the concept of semantic richness. It exist in the ontology and apply semantic technologies during all phases of the mining process. A novel method is proposed to efficiently extract items and transactions suited for traditional association rules mining algorithms.

Index Keywords – Semantic Annotated Data, Association Rule Mining, Ontology.

1. INTRODUCTION

Recently there has been an interest in combining the two research areas: semantic web and data mining. With the help of the standardization of the ontology languages such as OWL and RDFS, the semantic web has been extended and the amount of available semantic annotations used in different applications is significantly increasing.

Mining semantic web data will provide much benefit to lots of domain-specific research communities, where their data are usually complex and heterogeneous, and a large amount of knowledge is encoded in them, in the form of semantic annotations. Therefore, by processing this kind of data and using the semantic richness of it, rules with higher semantic level can be expected. Consequently, the challenges such as homogeneousness and complexity due to the special features of semantic annotated data have to be addressed.

Thus, the semantic annotated data do not have a rigid structure. As a result, there would be structural heterogeneity problems. Moreover, traditional data mining algorithms work with homogeneous datasets which include transactions, subsets of items. However, there are ontology axioms and data instances in the form of triples (subject, predicate, and object) in a repository of semantic data expressed in OWL and RDF. In addition, there is a need to apply reasoning capabilities to make use of implicit semantic knowledge. To conclude, In this paper a novel method is presented to find semantic association rules from semantic web data using semantic web technologies. These types of rules can be applied in decision support systems to help them make more intelligent decisions.

The rest of the paper is organized as follows: Section II describes about the association rule mining algorithm and Section III describes about the semantic data mining, Section IV illustrates the Apriori Algorithm, Proposed Methodology is described in the Section V. Experimental results are illustrated in Section VI. Section VII illustrates the conclusion and Section VIII illustrates the future work.

2. ASSOCIATION RULE MINING ALGORITHM

Association rule mining algorithms usually have two main steps. The first step is finding frequent itemsets. At this phase, all of the items which meet the support threshold are discovered. The second phase is derivation of association rules. At this stage, based on frequent itemsets discovered in the first phase, association rules that apply in confidence condition will be derived. Since the second step of the association rule derivation can be done in an optimal way, the research mainly focuses on the first step, how to efficiently discover all frequent item sets [1].

Let $I = \{I_i, i = 1, 2, ..., m\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID. A transaction T contains X, a set of some items in I, if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction set D if s% of transactions in D contain $X \cup Y$ [22]. Given a set of transactions D, the problem of mining association rules is to discover all association rules with support and

PISSN:2347-4408
EISSN:2347-4734

confidence greater than the user-specified minimum support and minimum confidence respectively. In 1994, the famous association rules algorithm Apriori was presented by R. Agrawal et al. [2]. From then on, association rules were studied deeper. There are two ways to improve the algorithms to increase mining efficiency: Apriori-based algorithms and not Apriori-based algorithms [3].

3. SEMANTIC DATA MINING

Some works on semantic data mining are based on the inductive logic programming [4], which uses the underlying logic annotated in the semantic data to learn new concepts. There are other works which apply statistical machine learning methods to deal with ontologies and their annotated semantic data [7], [8], [9], [10], but their representations are not suited for association mining that requires defining the set of items and the transactions from the semantic data.

Another related field is mining tree and graph structured data. In this topic, there are methods used such as frequent sub tree [11] and graph mining [12], whose purpose is to identify frequent substructures in complex and heterogeneous data sets. However, these approaches do not pay enough attention to frequent semantically related contents.

Another issue is the transaction definition according to elimination of the heterogeneity that exists in semantic web datasets. In [13], [14] it is said that XQuery is not the most suitable way for extracting data from XML data sources, because the structure of the underlying documents should be known. A better method is using the lowest common ancestor semantics to extract meaningful. Although, undesired combinations of data items might be generated sometimes [15], [16]. To solve the problem the smallest possible context data strategy was proposed and a similar approach was done in [17].

Finally, there are several works focused on integrating knowledge discovery capabilities to SPARQL by extending its grammar. Some examples are [18] that can be used with some data mining algorithms and [19], which extracts complex path relations between resources. [20] has also extended SPARQL grammar to define association rule patterns over the ontological data in a less restrictive way than the one imposed by SPARQL.

4. APRIORI ALGORITHM

Apriori is a basic algorithm for frequent item set mining and association rule mining over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to generate association rules which highlight general trends in the database.

Apriori (J, ε)
$L_1 \leftarrow$ {large 1 – itemsets}
k←2
while $L_{k-1} \neq$ emptyset
$E_k \leftarrow \{e | e = a \cup \{b\} \Lambda\ a \in L_{k-1} \Lambda\ b \epsilon \cup L_{k-1} \Lambda\ b \notin a\}$
for transactions a ∈ A
$E_j \leftarrow \{e | e \in E_k \Lambda\ e \subseteq j\}$
For candidates e ∈ E_j
count[e] ← count[e] + 1
$L_k \leftarrow \{e | e \in E_k \Lambda$ count [e] $\geq \varepsilon\}$
k← k+1

$$return \bigcup_k L_k$$

In the algorithm J denotes the transactions and e denotes the confidence. This algorithm is rewritten to deal with semantic transactions and accordingly semantic rules, with their predefined format in the ontology, will be resulted. In addition, some factors such as support and confidence should be determined for each particular data set. The achieved rules are expected to be useful in improving intelligent decision support systems.

5. PROPOSED METHODOLOGY

The first step for enabling semantic data mining and discovering new association rules efficiently using our semantic method, is the design and the implementation of the needed ontology for association rule mining concepts. This ontology is required besides the application ontologies for defining the specifications for concepts like item, transaction, and association rule and their properties to be used in the process of semantic data mining algorithm. The method will be different with other methods that have been proposed so far, according to the usage of this ontology and semantic annotated data in all parts of our mining system.

Association rule mining algorithms usually work with a dataset of transactions that each of them contains a subset of items. The semantic annotated data are transformed to semantic transactions, without losing the semantic richness of semantic web data. To implement this idea semantic query language such as SPARQL can be used. Semantic transactions and their items are formatted based on the association rule mining ontology defined earlier. Therefore, they can be linked with other semantic data, and also semantic reasoning can be performed on them to generate new more meaningful transactions.

Therefore, there are two general phases in our semantic

PISSN:2347-4408
EISSN:2347-4734

association rule mining system, semantic transaction production and running semantic association rule mining algorithm on them. The second phase is implemented based on Apriori. This algorithm is rewritten to deal with semantic transactions and accordingly semantic rules, with their predefined format in the ontology. In addition, some factors such as support and confidence should be determined for each particular data set. The achieved rules are expected to be useful in improving intelligent decision support systems.

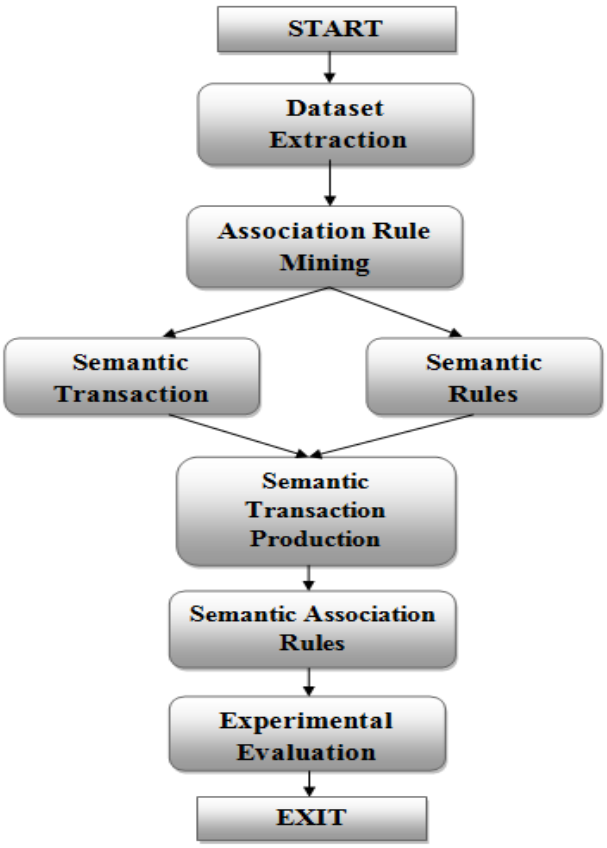

Fig.1 shows the flow diagram of the proposed methodology.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this experiment, the cross-transactions were used as a ground truth rule set with APRIORI. As said before, the evaluation is that the cross-transactions represent obvious associations because the method connects concepts with similar transactions. These associations can therefore be selected by various tested transactions.

Several experiments have been carried out in this paper to prove that our method is able to generate association rule base transactions from Semantic data, which later translates into high quality association rules. Notice, the focus of the work is not on developing new mining algorithms, but on transforming and making complex semantic data amenable to existing ones. In the experiments on real world datasets, the hierarchical Semantic data are based on associativity, it achieved very good results. Apriori, on the contrary, produced some of the best results, partly because the semantic ontology based model is not used. However, its performance could be improved by multiplying by φ as proposed in [18].

Moreover, the combinations of φ with other measures produced also good results. Some of the best results, FOAF Vocabulary Specification 0.99 (It is a machine-readable ontology describing persons, their activities and their relations to other people and objects. Anyone can use FOAF to describe him- or herself. FOAF allows groups of people to describe social networks without the need for a centralized database.)

PISSN:2347-4408
EISSN:2347-4734

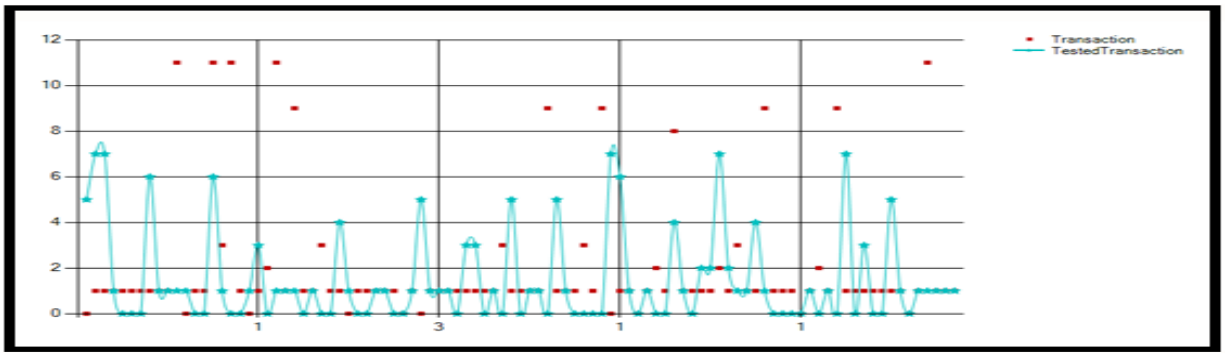

Fig.2 Shows the Tested Transactions Results.

Table.1 The table shows the comparison between the associating rule mining (ARM) and semantic web mining (SWM).

S.NO	TITLE	ARM	SWM
1	Algorithm	Association Rule Mining Algorithm	Apriori Algorithm
2	Complexity	High	Low
3	Semantic Mining	Heterogeneous Structured Data	Homogeneous Structured Data
4	Efficiency	Low	High
5	Rule Patterns	Association Rules	Semantic Rules
6	Transaction	Dataset of transaction	Semantic Transaction and Production

7. CONCLUSION

The Semantic Web technologies offer solutions to capture and efficiently use the domain knowledge. According to the challenges listed before, it is crucial to apply the knowledge of semantic annotated data based on ontology's, to produce semantic transactions efficiently. By definition of semantic transactions and their properties in the ontology, overcoming of the heterogeneity of semantic web data is achieved.

The mining process presented in this paper can be performed automatically for any kind of semantic data after extracting semantic transaction using the application ontology. Also, since all parts of the system work based on a uniform ontology, semantic integrity exists in the entire system.

Therefore, the rules and their related data are linked to other data sets and semantic technologies such as semantic reasoning can be applied to them. To conclude, In this paper a novel method is presented to find semantic association rules from semantic web data using semantic web technologies. These types of rules can be applied in decision support systems to help them make more intelligent decisions.

8. FUTURE WORK

As a future work, the generalized query patterns are applied by using the ontology axioms, as well as to automatically discover interesting contexts and their association rules. Moreover, our method could be applied in a variety of different scenarios, where the mining tasks are transaction oriented. An interesting issue for future work is to use the knowledge encoded

PISSN:2347-4408
EISSN:2347-4734

in the ontology in order to filter and prune discovered rules, and also to express the user goals. Another important direction worth exploring concerns the combination of clustering and association mining algorithms to summarize document collections.

This technique was formerly introduced through the Frequent Item set based Hierarchical Clustering (FIHC). Basically, the FICH algorithm generates clusters from frequent item sets, which in turn constitute the cluster descriptors. Several enhancements of this algorithm have been proposed since then. Recently, proposed a novel approach also based on frequent item pairs that provides more homogeneous clusters and better descriptions than those obtained with FIHC. Alternative research lines, which are out of the scope of the present work, consist in applying more sophisticated data mining algorithms to the generated transactions and study their performance. Equally interesting is to devise new data mining algorithms that take profit from the semantically enriched items of the generated transactions.

REFERENCES

[1] Mala A., Ramesh Dhanaseelan F., 2011: DATA STREAM MINING ALGORITHMS – A REVIEW OF ISSUES AND EXISTING APPROACHES, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 7, Pages 2726-2732.
[2] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.
[3] Mingzhu Zhang, Changzheng He, "Survey on Association Rules Mining Algorithms", in Advancing Computing, Communication, Control and Management Lecture Notes in Electrical Engineering Volume 56, 2010, pp 111-118.
[4] S. Muggleton, L.D. Raedt, Inductive logic programming: theory and methods, J.Log. Program. 19 (20) (1994) 629–679.
[5] F.A. Lisi, F. Esposito, Mining the semantic web: a logic-based methodology, in: ISMIS, LNCS, vol. 3488, Springer, 2005, pp. 102–111.
[6] J. Hartmann, Y. Sure. A knowledge discovery workbench for the semantic web, in: Workshop on Mining for and from the Semantic Web at the ACM SIGKDD, August 2004.
[7] S. Bloehdorn, Y. Sure, Kernel methods for mining instance data in ontologies, in: ISWC/ASWC, LNCS, vol. 4825, Springer, 2007, pp. 58–71
[8] S. Bloehdorn, Y. Sure, Kernel methods for mining instance data in ontologies, in: ISWC/ASWC, LNCS, vol. 4825, Springer, 2007, pp. 58–71. R. Dánger, J. Ruiz-Shulcloper, R.B. Llavori, Objectminer: a new approach for mining complex objects, ICEIS 2 (2004) 42–47.
[9] N. Fanizzi, C. d'Amato, F. Esposito, Metric-based stochastic conceptual clustering for ontologies, Inf. Syst. 34 (8) (2009) 792– 806.
[10] A. García, R. Berlanga, R. Dánger, A description clustering data mining technique for heterogeneous data, Commun. Comput. Inform. Sci., Softw. Data Technol. 10 (2008) 361–373.
[11] Y. Chi, R.R. Muntz, S. Nijssen, J.N. Kok, Frequent subtree mining – an overview, Fundam. Inform. 66 (1–2) (2005) 161–198.
[12] M. Kuramochi, G. Karypis, Frequent subgraph discovery, in: N. Cercone, T.Y. Lin, X. Wu (Eds.), ICDM, IEEE Computer Society (2001) 313–320.
[13] Y. Li, C. Yu, H.V. Jagadish, Schema-free XQuery, in: VLDB '04: Proceedings of the 30th International Conference on Very Large Data Bases, VLDB Endowment, 2004, pp. 72–83.
[14] Y. Xu, Y. Papakonstantinou, Efficient keyword search for smallest LCAs in XML databases, SIGMOD '05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, 2005. pp. 527–538.
[15] T. Näppilä, K. Järvelin, T. Niemi, A tool for data cube construction from structurally heterogeneous xml documents, JASIST 59 (3) (2008) 435–449.
[16] T. Niemi, T. Näppilä, K. Järvelin, A relational data harmonization approach to xml, J. Inf. Sci. 35 (5) (2009) 571–601.
[17] A. Tagarelli, S. Greco, Semantic clustering of xml documents, ACM Trans. Inf. Syst. 28 (1) (2010).
[18] C. Kiefer, A. Bernstein, A. Locher, Adding data mining support to SPARQL via statistical relational learning methods, in: S. Bechhofer, M. Hauswirth, J. Hoffmann, M. Koubarakis (Eds.), ESWC, Lecture Notes in Computer Science, vol. 5021, Springer, 2008, pp. 478–492.
[19] K. Kochut, M. Janik, SPARQLeR: extended SPARQL for semantic association discovery, in: ESWC, LNCS, vol. 4519, Springer, 2007, pp. 145–159.
[20] Nebot V., Berlanga R., 2012: Finding association rules in semantic web data, Knowledge-Based Systems, Vol. 25, Pages 51-62.
[21] Berners-Lee, Tim; James Hendler and Ora Lassila. "The Semantic Web". Scientific American Magazine. Retrieved March 26, 2008.
[22] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far. In: IJSWIS, Vol. 5, Issue 3, Pages 1-22, 2009.
[23] Samad Paydar, Mohsen Kahani, Behshid Behkamal, Mahboobeh Dadkhah, Elaheh Sekhavaty, Publishing Data of Ferdowsi University of Mashhad as Linked Data, the 2010 International Conference on Computational Intelligence and Software Engineering (CiSE 2010).