# MINING THE WEB FOR LEARNING THE ONTOLOGY

Bassam M. Aoun and Marie Khair
*Department of Computer Sciences, Notre Dame University, Louaize, Lebanon*

Keywords: Semantic Web, Ontology, Web Mining, Text Mining, Apriori Algorithm.

Abstract: The Semantic Web is a network of information linked up in such a way as to be easily processed by machines, on a global scale. To reach semantic web, current web resources should be automatically translated into semantic web resources. This is usually performed through semantic web mining, which aims at combining the two fast-developing research areas, the Semantic Web and Web Mining. A major step to be performed is the ontology-learning phase, where rules are mined from unstructured text and used later on to fill the ontology. Making sure that all rules are found and no additional and inaccurate rules are inserted, remains a critical issue since it constitutes the basis for building the semantic web. The mostly used algorithm for this task is the Apriori algorithm, which is inherited from classical data mining. However, due to the nature of the semantic web, some important rules can be dropped. This paper presents an enhanced version of the Apriori algorithm, En_Apriori, which uses the Apriori algorithm in combination with the maximal association and the X2 test to generate association rules from web/textual documents. This provides a major refinement to the classical ontology learning approach.

## 1 INTRODUCTION

Mining association rules is an important task in data mining, which aims at extracting potential relationships among data items in databases. The main idea was proposed first by (Agrawal et Al., 1993), shortly after that it was known by the "Apriori algorithm" (Agrawal et Al., 1994), which is an influential algorithm for mining frequent itemsets for Boolean association rules. Usually most data mining algorithms are based on the Apriori algorithm where association rules are mined based on two criteria: support and confidence where the support of an association pattern is the percentage of task-relevant data transactions for which the pattern is true and Confidence is defined as the measure of certainty associated with each discovered pattern.

In this paper we will introduce an enhancement to the Apriori algorithm "En_Apriori" algorithm which is implemented by introducing and integrating the Maximal Association Rule Algorithm and the Chi-squared Test method with the Apriori algorithm. This resulted that the discovery of the association rules is more accurate and refined, in such a way that no faulty rules will be discovered and introduced into the ontology, also so that no rules would be omitted by using the classical Apriori algorithm. At the same time, the additional time needed to implement these algorithms can be considered negligible in comparison with the benefits obtained.

This paper is organized as follows: we start by giving an overview of the semantic web and data mining technologies; we then introduce the new and enhanced Apriori Algorithm, En_Apriori, by presenting previous and related work along with a step-by-step explanation of the proposed algorithm and a discussion of the output of the implementation of our method and a conclusion.

## 2 SEMANTIC WEB MINING

The idea behind the semantic web is to make the web as intelligent as possible. Therefore, the Semantic Web is about two things: first it is about common formats for interchange of data, where on the original Web we had only interchange of documents. Second, it is about language for recording how the data relates to real world objects. Therefore, the Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. It is based on the (RDF) Resource Description Framework, which integrates a variety



M. Aoun B. and Khair M. (2007).
MINING THE WEB FOR LEARNING THE ONTOLOGY.
In *Proceedings of the Second International Conference on Software and Data Technologies - Volume ISDM/WsEHST/DC*, pages 189-192
DOI: 10.5220/0001348101890192

of applications using XML for syntax and URIs (Uniform Resource Identifier) for naming. At the heart of all semantic web applications is the use of ontologies, which describe entities and relationships among entities. The concept of metadata has evolved over the years starting from data dictionaries to database schemas and now to ontologies.

Data mining aims at finding patterns and subtle relationships in data and discovering rules that allow the prediction of future results by the use of automatic or semi-automatic processes. It is an information extraction activity, whose goal is to discover hidden facts contained in databases, using a combination of machine learning, statistical analysis, modelling techniques and database technology. Mining the data on the web, however, is one of the major challenges faced by the data management and mining community, as well as those working on web information management and machine learning. The characteristic feature of Web Mining is the use of Data Mining techniques to elaborate on content, structure, and usage of Web resources.

In the Semantic Web, content and structure are strongly inter-wined. Therefore the distinction between structure and content mining vanishes. The mining algorithms can be transformed in order to deal with RDF or ontology-based data. Mining the usage can be enhanced further, if the semantics are contained explicitly in the pages by referring to concepts of ontologies.

## 3 RELATED WORK

We will firstly review the formal model of association rule as was introduced by Agrawal (Agrawal et Al., 93). Formally association rules mining can be stated as follows:

- Let I= $\{i_1, i_2, \ldots i_n\}$ be a set of items
- Let D, be a set of transactions, where each transaction T is a set of items satisfying $T \subseteq I$
- Each transaction is assigned an identifier, called TID
- Let X be a set of items, a transaction T is said to contain X if and only if $X \subseteq T$.
- An association rule is an implication of the form $X \Rightarrow Y$.

While association rules provide means to discover many interesting associations, they fail to discover others, no less interesting associations that are also hidden in the data. While this may not be very dangerous in classical mining procedure this seems to be a serious problem in semantic web mining since this will form the basis of the ontology that will form the semantic web. However maximal association rules are not designed to replace regular association rules, but rather to allow the discovery of the concepts, which will be included in the ontology and the relations that bind them together. For this reason, we propose in this paper enhancements to the algorithm proposed above. These enhancements will allow the discovery of new association rules to complement them (Amihood et Al., 05). Maximal associations was proposed to allow the discovery of associations pertaining to items that most often do not appear alone, but rather together with closely related items, and hence associations relevant only to these items tend to obtain low confidence in the classical algorithms, for example Apriori. In a maximal association rule we are interested in capturing the notion that whenever X appears alone then Y also appears, with some confidence (Amihood et Al., 05) and this is why it is crucial for text/web mining for learning the ontology.

In addition, some redundant, unwanted or even false strong association rules are likely to be generated because the correlation of attributes is ignored (Yong Xu et Al. 05). So the Chi-Squared test should be introduced to association rules mining since it could remove irrelevant itemsets and rules that have high support but no dependency.

## 4 ENHANCED LEARNING ALGORITHM: EN-APRIORI

The main learning algorithm that has been adopted in our paper is the one proposed in (Maedche et Al., 99)(Berendett et Al, 06). This learning algorithm is effective up to a certain level, and since it is a text/web mining approach then the techniques from text mining and web mining have been combined to achieve the learning of the ontology.

We will stress again that the learning of the ontology step is basically the most important step, since its results will allow the discovery of the concepts, which will be included in the ontology and the relations that bind them together. For this reason, we propose in this paper enhancements to the algorithm proposed above. These enhancements will allow the discovery of new association rules that have been missed by the original learning algorithm and in addition it allows the pruning of some faulty rules that appeared to be valid strong association rules. Our approach is based on the idea of introducing two new algorithms and integrating

them into the original algorithm; these two algorithms are the "Maximal Association Rule" algorithm and the "Chi-squared test" algorithm.

En_Apriori is the enhanced version of Apriori adapted for text/Web mining with one principal goal in mind: enhancing the ontology learning phase in order to bring refinement to the overall process of building the semantic web. After having applied several text processing techniques over the desired website, a set of semantically related entities is discovered. Following the Apriori based approach; these entities are used to generate the transaction database which will be processed by the Apriori algorithm in order to discover the association rules that respect the minimum support and confidence thresholds specified by the analyst.

## 4.1 En_Apriori Algorithm in Detail

After having generated the rules by using the Apriori algorithm, we proceed by running the maximal association rules algorithm (MAR). The first step of the MAR algorithm is the generation of the "M-frequent itemsets"; we proceed from there by comparing the M-frequent itemsets with the frequent itemsets generated by Apriori in order to remove the common itemsets which will allow us to avoid the generation of redundant rules and thus reduce the computation time needed. The refined M-frequent itemsets will then be processed in order to generate the M-association rules that comply with both the support and confidence thresholds specified.

After having generated the rules by means of Apriori and MAR, we proceed by testing them with the Chi-squared test, in order to evaluate the correctness of the discovered rules. The Chi-squared test will provide a way to measure the dependency between any two entities pertaining to one association rule. The purpose of this test is to provide the analyst with a way to evaluate the rules discovered before filling then into the ontology. The Chi-squared test starts by calculating the Chi-squared value of each association rule; this will generate two sets of rules: the first set contains the rules that have a chi-squared value higher than the cut-off value and the second set contains the rules that have a chi-squared value lower than the cut-off value. We are more interested in the second set of rules; these rules are presented to the analyst for revision, if he finds that some of these rules are relevant to the area of study and therefore should be present in the ontology, then these rules will be joined back with the first set of rules to constitute together the set of refined association rules, otherwise they will be dropped.

The algorithm in Figure 1 starts by running the Apriori algorithm in order to generate the frequent itemsets. Once these itemsets have been generated, the function M-frequent-sets is called, it will generate the sets of Maximal frequent itemsets; afterwards we proceed by removing the common itemsets, generated in both Apriori and maximal associations, from the M-frequent sets discovered and call the function MA_gen() to generate the maximal association rules from the remaining M-frequent itemsets. When association rules and maximal association rules have been generated, we test them using the chi-squared test: First, all rules are tested by means of chi-squared and loaded in a temporary array "*Tmp"*. Second, we load the rules that failed in the Chi-squared test in another temporary array called "*NCC"*. We proceed from there by removing the set of rules *NCC* from *Tmp* in order to keep the valid rules aside; we then present to the analyst the suspicious rules "*NCC"* for revision. The rules revised and accepted by the analyst will be put in a temporary array, *NCC'*, which will be joined in the next step with *Tmp* to form the set of trusted association rules which will be used for the ontology learning.

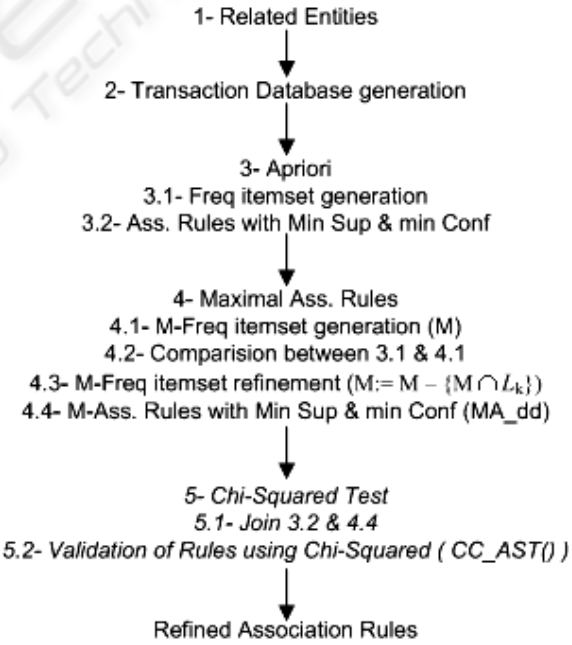


Figure 1: The main steps of the Apriori algorithm.

### 4.2 Main Steps of the EN_Apriori Algorithm

Let us proceed by giving some definitions of the functions and variables introduced and used in the algorithm in figure 1.

- MA_gen(): is the function that calls the MAR procedure, this function takes as input the M-frequent itemsets and outputs a set of rules that meets the support and confidence thresholds specified by the analyst.
- MA_dd: is the container that will hold the set of M-association rules discovered by MA_gen().
- M: is the set of M-frequent itemsets
- CC_Test(): is the procedure that tests the dependency between two entities of a given association rule
- CC_AST(): is the function that calls the CC_Test() procedure and then re-sorts the itemsets according to the analyst's decisions.
- Tmp, NCC, NCC': are temporary arrays used to perform various computations.

## 5 EXPERIMENTATION

In order to simulate our idea, an unstructured text has been chosen. After analyzing this text, semantically related pairs of words emerge along with their frequencies of occurrence and weight. To build our transaction database, we need to annotate each entity first; the entities will be annotated in the following form: "Entity 1 → I1", "Entity 2 → I2", etc. Each sentence in the text is considered to be a transaction "Ti"; we proceed by filling our transaction database. After comparing the number of rules discovered by both Apriori and En_Apriori, we can clearly see that En_Apriori brings refinement to the process of "learning the ontology".

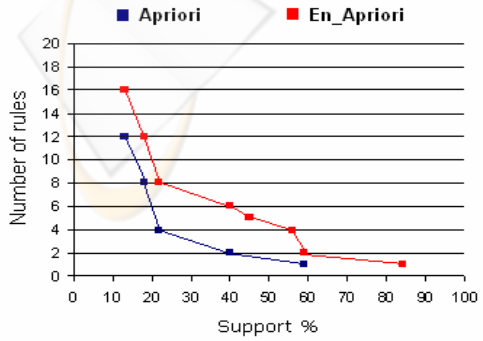


Figure 2: Number of rules for the different support.

Figure 2 shows the number of rules that are discovered using Apriori and En_Apriori at various support thresholds.

## 6 CONCLUSION

The main objective of our research is to obtain a set of rules that is complete and minimal as much as possible. For this purpose, we enhanced the Apriori algorithm by introducing the maximal association rules that may find some association rules missed by the Apriori algorithm and we proceeded by introducing the Chi-Square test to validate the obtained itemset. Further refinements and experimentations are still needed for the proposed algorithm and for our future work we intend to compare the obtained association rules with already agreed standard rules to validate the results.

## REFERENCES

Agrawal A., Srikant R., 1993. Fast algorithms for mining association rules in large databases. In *Proc. Of the 20th Intel. Conf. on Very Large databases.*

Agrawal A., Tomasz Imielinski, Arun Swami, 1993. Mining Association rules between sets of items in large databases. In *Proc. Of ACM SIGMOD Conference on Management of Data, Wahington D.C., pp. 207-216.*

Amihood Amir, Yonatn Auman, Ronen Feldman, Moshe Fresko, 2005. Maximal association Rules: A tool for mining Associations in text, *Journal of Intelligent information systems,* 25:3, 333-345.

Bettina Berendt, Andreas Hotho, Gerd Stumme, 2006. Semantic web mining-state of the art and future directions, *Journal of Web Semantics.*

Maedche, A., Staab, S., 1999. Discovering Conceptual Relations from Text, *Institute AIFB,* Karlsruhe University Germany.

Yong Xu, Sen-Xin Zhou, Jin-Hua Gong. 2005. Mining Association Rules With New Measure Criteria. In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetic*.