Journal of Theoretical and Applied Information Technology
15th September 2019. Vol.97. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFICIENT TREE-STRUCTURED ITEMS PRUNING FOR BOTH POSITIVE AND NEGATIVE ASSOCIATION RULES

JURYON PAIK
Pyeongtaek University, Department of Digital Information and Statistics, Gyeonggi-do 17869, South Korea
E-mail: jrpaik@ptu.ac.kr

ABSTRACT

Researchers endeavor to develop efficient association rule mining algorithms that are suitable for being stored in RDBMSs or native systems and reach a certain level of maturity. However, with rapidly developing sensor-rich environments, the volume of flexible, extensible, and cost-effective data, such as JSON, XML, and so on, also increases exponentially with a sheer diversity. Even though the researchers report impressive levels of accuracy in the discovery of association rules, widespread adoption is hardly possible. Because most proposed approaches are not designed for such flexible but complex data models. They generally focus on simple record data for structured data models. The current data model paradigm, however, is unstructured or semi-structured one, which is usually represented by graph or tree. The sophisticated tree-based models guarantee data exchangeability, heterogeneity, and granularity without consideration of data types. Tree-based data modeling underpins several disruptive data models. Representing any data requires big tradeoff, that is rigorous analyzing. It is much harder to mine hidden information from the tree-based modeling data. Under the Fourth Industrial Revolution, however, data is not just the data, but critical corporate asset. The commoditized data and its valuable analytics in unique ways enable new insights. New analytical techniques can fuel discovery and innovation. This paper targets on providing data analytics methodology, especially for tree-based data models, in order to support both positive and negative association rules. This work provides several adjusted definitions and expressions for both types of associations rules, and shows simple mathematical results applied with some constraints to decide in advance whether patterns have values to be discovered.

Keywords: *Tree-based Data Model, Tree Structured Data, Negated Tree Items, Positive Tree Items, Association Rules*

1. INTRODUCTION

Data and its analytics capabilities have made a leap forward in recent years. Computational power and storage have steadily improved more than ever. The volume of available data has grown exponentially, and more sophisticated algorithms have been developed. The convergence of these trends is fueling rapid technology advances. According to McKinsey Global Institute report [1], the cutting-edge smart environment, we call it the Internet of Things (IoT), offers a total potential economic impact of $3.9 trillion to $11.1 trillion per year by 2025. IoT has been developing in parallel to wireless communication technologies. Without smart and miniaturized sensor devices along with the vast extension of information technologies, the current IoT environment would not be possible.

Harvesting benefits from the critical endpoint fused technologies faces barriers in extracting values from data. Current data us a critical asset in all of the areas; industry, business, academy, and even retails. Data comes from the web, billions of phones, sensors, and a huge array of other sources. IoT inevitably produces tremendous volumes of stream data. It is required to handle efficiently and exchange desirably the large heterogeneous datasets. XML was originally designed to carry data, not to display data. It defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The goals of designing xml focus on generality, flexibility, and heterogeneity across the web, that makes xml is used widely for the representation of arbitrary data structures. But xml has fallen out of favor due to its parsing complexity and verbosity. Developers seek out alternatives, that is JSON. Short for JavaScript Object Notation, json is a lightweight format for data exchange, which does not require the use of xml. The simplicity of json is leading to its widespread use, especially as an

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

alternative to xml. json is now the dominant method for data exchanging and transferring format.

There is a key feature that both xml and json can rule the data interchange format. That is their flexibility. They can represent any kinds of data format and completely language independent, because they describe data in tree structures. Researchers and venders are gaining the capability to gather sufficient data better than ever. In contrast, the data structures are more complicated and harder to analyze. The leading technology users such as business managers and researchers express the frustration about being unable to harvest benefits fully from the huge amounts of tree structured data flowing. Because value of data is tied to its ultimate use. McKinsey [2] reports that data itself will become increasingly commoditized, value is likely to accrue to the owners of scarce data, to players that aggregate data in unique ways, and especially to providers of valuable analytics.

Organizational analytics to the tree structured data has been done with traditional analysis tools such as the association rules mining and the classification. However, they were faced with serious problems. First, the data is too large to process using typical on-premises database management and processing applications. It needs to be processed by a flexible, scalable compute model that evolves. Second, most data are streaming data. It differs in several properties from traditional information storage data; 1) streaming data arrives continuously with high speed rate and needs to be processed in real-time. 2) Algorithms for data streams have only a single access because random access is very expensive explained by Babcock, et.al. [3]. If raw data from sensors is analyzed properly and evaluated effectively through data mining, it can be definitely predictive insights for fully facilitating IoT environments.

One of the well-known methods is to discover interesting relations between data, called association rules mining. In recent years, there has been a significant research focused on finding interesting non-existing or infrequent parts of data leading to the discovery of negative association rules [4]. However, the discovery of non-existing data parts is far more difficult than their counterparts, that is, frequent data parts. Besides, it is the most difficult task if the data type is complex structure like json. Analyzing continuously arriving json data is intricate and complicated process, and many of the problems it presents have yet to be adequately solved. It is still in an immature stage and not fully developed to address the problem of finding negative association rules from tree-structured stream data.

In this paper, we aim to discover data analytics methodology for tree-based data models. Among many arrays of mining methods, the target is association rules mining. We focus on an efficient pruning methodology not only ordinary association rules but also negative association rules together in a single algorithm. The contributions are that 1) the approach to decide at a time whether positive or negative rules are in advance is defined, 2) two other constraints are suggested to make up for the weak points of conventional constraints, and 3) a frame algorithm is presented for pruning of both types of association rules.

2. RELATED WORK

Problem of finding associated patterns was first introduced by Agrawal, et. al [5]. The actual aim was to analyze customer behaviors and capture information from market basket transactions. The identified patterns, called rules, are such items that are very often purchased together with other certain items by meaningful percentages of the customer. Also, the patterns have significant power to decide about which item should be placed near to each other or which item should be put on sale. Discovery of such patterns has been known as the research area of mining association rules. Besides market basket analysis, association rules analysis is widely used in various domains such as bioinformatics, web mining, intrusion detection, and educations to evaluate data and support many real applications. Actually, a remarkable number of variants and improvements of association rules mining have been proposed and still actively studied by many researchers such as Han and Fu [6], Han, et al. [7], Wolff and Schuster [8], Boukerche and Samarah [9], Rashid, et al. [10], and so on.

The approach of finding association rules from stream data did not exist before 2000's. With rapidly increasing sensor network deployments and ability to generate large volumes of data, data mining communities have burst into association rules discovery from the stream data. Among the early studies on stream data from sensors, Loo, et al. [11] proposed a framework for discovering association rules from sensor networks. In their approach, sensors' values are considered mainly to generate the association rules and the time is divided into intervals. With interval list based lossy counting,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

transaction in Loo et al.'s data model, the size of data structure is significantly reduced.

With growing data volumes and increased data complexity, the importance of negative association rules is even bigger than that of positive association rules. However, there are very few research works conducted on mining negative association rules over streaming data. Most of the published articles, such as Savaere, et al. [12], Antonie and Zaïane [13], Honglei and Zhigang [14], and Sumalatha and Ramasubbareddy [15], are confined to static database environment. The reason the researches for negative association rules are much less than that of positive ones is that there are fundamental differences between them, as described by Wu, et al. [16]. While positive association rules are generated with frequently occurred itemsets, negative association rules are generated with infrequently occurred or absent itemsets. That means we must search a gigantic number of negative association rules even though the database is small. If the database becomes larger, it would be more difficult. Particularly, it is a challenge to identify which rules are beneficial or useful to applications from the enormous and rigorous size of streaming data.

Corpinar and Gündem [17] suggest a rule mining system that provides solution to positive and negative association rules computation. However, their type of data stream is different to that of other approaches. The data is xml data stream. To achieve the goal, they first adapt the original FP-Growth method to support stream data mining and negative rules. To decrease the search space for negative association rules, they devised new pruning thresholds along with adding correlation coefficient parameter into their methodology to separate the frequent sets for positive and negative rules.

Another paper published by Paik et al. [18] presents several new definitions and scheme related to association rule mining over xml data streams in wireless sensor networks. The authors′ proposed scheme is the first approach to mining association rules from xml stream data in the sense that it generates frequent tree items without any redundancy. The overall methodology can be applied to any individual block, as well as the whole stream.

The previous two papers commonly mention that managing continuously arriving xml data is expensive and complicated task. Many problems caused by the stream data structured tree have yet to be discussed fully. In this paper, we consider pruning techniques because positive and negative association rules are built from a huge number of candidate tree items. Without pruning such useless tree items, the algorithm of mining both types of association rules will definitely have the worst time and space cost. To accelerate and leverage the mining process, we mainly discuss two major methods, *interestingness vs. correlation coefficient*, for the pruning phase. Then, we show different results when two more measuring factors are applied to simple examples, along with a frame algorithm to extract useful tree items both for positive and negative association rules.

3. ASSOCIATION RULES FROM TREE-BASED DATA

3.1 Sections and Subsections

In the IoT revolution, one of the popular data formats is JSON. In the previous era, the world of the Web, it was XML. The common characteristics of both data formats are heterogeneity, exchangeability, and especially tree structure. In this paper is mainly the structure focused. Figure 1 provides data encodings of json and xml about a partial information for a person 'John Smith'. The data by xml is more verbose than that by json, although they represent same information. Therefore, JSON is often used as an alternative to XML, due to its flexibility, easy interchangeability and lightweight. Also, it is a very common data format to transmit and read data from sensors in an increasing number of IoT applications.

```
{
   "firstName": "John",
   "lastName": "Smith",
   "age": 25,
   "address": {
       "streetAddress": "21 2nd Street",
       "city": "New York",
       "state": "NY",
       "postalCode": "10021-3100"
   },
                  . . .
}
```

Figure 1: Partial Data Encoding of JSON(upper) and XML(bottom)

Journal of Theoretical and Applied Information Technology
15th September 2019. Vol.97. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
<person>
   <firstName>John</firstName>
   <lastName>Smith</lastName>
   <age>25</age>
   <address>
       <streetAddress>21 2nd Street</streetAddress>
       <city>New York</city>
       <state>NY</state>
       <postalCode>10021-3100</postalCode>
   </address>
                    . . .
</person>
```

Figure 1: Partial Data Encoding of JSON(upper) and XML(bottom)

Based on their encoding data, their data models are depicted on the figure 2. They are rarely different than encoded data. On the contrary, they show the similarity as a tree structure. The target dataset of the paper is such tree structured data not just a few but also a huge number of trees. Several researchers published the papers related xml tree data and defined useful definitions about tree data. We use and adjust their definitions and terms in this paper.

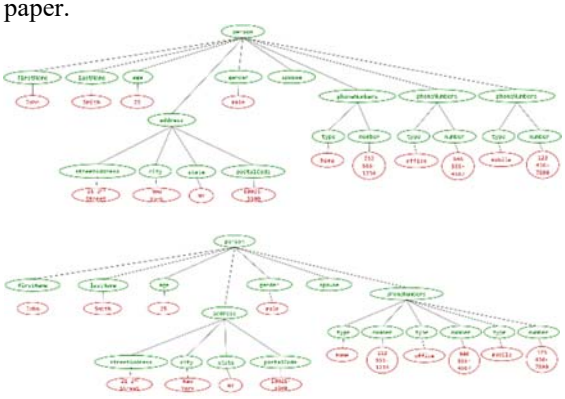

Figure 2: Tree Structure of JSON(upper) and XML(bottom)

3.2 Association Rules Mining

Based on the fundamental papers [19, 20], we briefly review some key definitions important facts for the association rules mining. Actually, the term 'association rules mining' is for positive association rules mining. For simplicity and genericity, we use the term 'association rules' without 'positive'.

Let $\mathbf{I} = \{I_1, I_2, \ldots, I_n\}$ be a set of items from a transaction database ***D***, such that is a RDBMS or native system. An (positive) association rule has the form X ⇒ Y, where X is called antecedent or left-hand-side (LHS) and Y consequent or right-hand-side (RHS). Both X and Y are subsets of **I**, and they are disjoint each other, X ∩ Y = ϕ. The rule X ⇒ Y states that a transaction T (T ∈ ***D***) containing the items in X (X ⊂ T) is likely to contain also the items in Y (Y ⊂ T).

To select meaningful rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best well-known constraints are the *support* and the *confidence*. The support of X with respect to ***D*** is defined as the proportion of transactions that contains all items in X. In formula is *sup*(X). Note that *sup*(X ∪ Y) means the support of the union of the items in X and Y; the proportion of the transactions in ***D*** that contains all the items in X and Y, because the argument of *sup*( ) is a set of preconditions, which makes rules more restrictive, instead of more inclusive. The following equations are for the constraint support:

$$sup(X) = \frac{|X|}{|\boldsymbol{D}|} \quad (1)$$

$$sup(X \Rightarrow Y) = \frac{|X \cup Y|}{|\mathbf{D}|} \quad (2)$$

The confidence value of a rule X ⇒ Y with respect to ***D***, measures the proportion of the transactions that contains X which also contains Y, written as *conf*(X ⇒ Y). It is the probability of finding the consequent part in transactions under the preconditions. Formally, measuring the confidence is defined by the support values.

$$conf(X \Rightarrow Y) = \frac{sup(X \Rightarrow Y)}{sup(X)} = \frac{|X \cup Y|}{|X|} \quad (3)$$

Table1 shows an extremely small example of a market basket domain. The set of items **I** = {bread, butter, diaper, jam, milk, water} and the table is a transaction database ***D***. Each record T_i ($1 \le i \le 4$) contains several items in **I**. Suppose there is a test rule, bread ⇒ milk, such means if bread is purchased, customers also buy milk. The support value, *sup*(X ⇒ Y), is the support of itemset {bread, milk}, which has a support of 3/4 = 0.75. The factor is that the test rule is occurred with 75% frequency ratio in all transaction data. The confidence value of the test rule is determined by the equation (3), and its value is 3/4 = 0.75. It clearly explains that for 75% of the transactions purchasing the item bread, milk is purchased as well.

The next test rule is {bread, butter} ⇒ milk. Its support and confidence are 2/4 = 0.5 and 2/3 ≈ 0.66, respectively, which means customers who buy

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

bread and butter also buy milk in more than 66% of the cases. Also, the rule holds for 50% of the entire transactions. Adding the item butter reduces the support value from 0.75 to 0.5 because it makes the rule more restrictive. With the useful relationships among the underlying data, market managers will put bread, milk, and butter together, and this may increase their profits because the relationships identify new opportunities for cross-selling their products.

Table 1: Record Data of Market Baskets

T_{ID}	*Basket Items*
T_1	{bread, butter, milk}
T_2	{bread, butter, diaper}
T_3	{bread, milk, jam, water}
T_4	{bread, butter, milk, water}

The discovery of association rules is formally stated as the following: given a set of transactions database ***D***, discover all the rules having support ≥ *ms* and confidence ≥ *mc*, where *ms* and *mc* are corresponding minimum support and minimum confidence thresholds, respectively. A naive approach for finding association rules is to compute the support and confidence values for every one of possible rules. This approach is, however, too expensive to apply data sets, which is mainly caused by the fact that there are exponentially many rules extracted from a data set. In order to avoid unnecessary computations, it is required to prune early the rules which support and confidence values do not satisfy the given conditions, without computing their support and confidence.

3.3 Negative Association Rules Mining

With the development of mining techniques and rapid growth of their usage, an alternative approach has been presented that considers **negative** association rules [16, 17, 21]. In the form of a negative associat ion rule, the positive association rule bread ⇒ milk can be expressed by one of the followings: 1) bread ⇒ ¬ milk implies the customers who buy bread usually do not purchase milk. 2) ¬ bread ⇒ milk, the customers who do not buy bread usually purchase milk. 3) ¬ bread ⇒ ¬ milk, the customers who do not buy bread usually do not purchase milk, either. For convenience, negative rules with the form X ⇒ ¬ Y are considered from now on.

The definition of a negative association rule is almost similar to that of a positive association rule, except that negative association rules comprise the relationship between absent and present items defined by Yuan, et al [22]. Association rules that include absent items are turning out to be as valuable as positive association rules. Even though interesting and potentially useful negative information exists in a dataset, it is not simply attempted to find them, because absent items should be considered.

As explained in the subsection 3.2, *sup*( ) and *conf*( ) values for positive association rules measure the count how many times itemsets **appear** in a dataset. However, negative association rules mining is a little different to apply two constraints. Due to its negated items, support and confidence of the rule X ⇒ ¬ Y have to count **non-existing** items in transactions, that is rarely possible to count the absent items actually. Instead directly counting them, we compute the measures via the support and confidence values for positive rules, as shown in the following:

$$sup(\neg X) = 1 - sup(X) \quad (4)$$

$$sup(X \Rightarrow \neg Y) = sup(X) - \sup(X \Rightarrow Y) \quad (5)$$

$$conf(X \Rightarrow \neg Y) = \frac{\sup(X \Rightarrow \neg Y)}{\sup(X)} = 1 - \frac{\sup(X \Rightarrow Y)}{\sup(X)} = 1 - conf(X \Rightarrow Y) \quad (6)$$

Table 1 has four record data configured by several items. When we assume *ms* = 0.75, the satisfying items are only {bread, butter, milk}, such that is typically used for derivation of positive association rules. The required searching space is 2^3 items. For example, a highly interesting rule bread ⇒ butter is obtained with 0.75 support and 1.0 confidence. However, what if the baskets were rechecked just before payments were made? There are slight changes in customers market baskets; some customers take items out of their baskets or some replace a few items with others. The changes made by customers are presented on Table 2.

Table 2: Changed Market Baskets

T_{ID}	*Basket Items*
T_1	{bread, milk}
T_2	{bread, butter, diaper}
T_3	{milk, jam, water}
T_4	{bread, cheese, apple, water}

It can be known that the item butter is taken out of the transactions T_1 and T_2. Also, it has been replaced with the item cheese in T_4. Consequently, the item butter is no more frequent item because its

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

support value is just 0.25, which does not satisfy the given *ms*. Instead, it is a member of 7 infrequent items, which indicates the searching space is 2^7 items to find some helpful items from the infrequent, specifically absent items; from the table 2, a meaningful rule can be inferred such that customers who put bread into their baskets usually take out the butter from their baskets just before they cash. Simply stating is that the customers who buy bread typically do not buy butter together when they actually pay written formally as 'bread ⇒ ¬ butter'. The rule is highly interesting because its support value is 0.75 and the confidence value is approximately 0.67 according to the equation (4) to (6). The example negative association rule has a quite high strength indicating that the rule is very reliable and helpful to market basket analysis. Analyzing negative associa0tion rules is as important as or more than that of positive association rules.

3.4 Items for Tree Structure Data

The fundamental two constraints for association rules, support and confidence, are applicable for absent items via the equation (4) to (6). A problem, however, still remains. The previously mentioned equations are for the record data stored in tables not for tree-structured data. Several researchers published their papers related to xml association rules, such as Braga, et al. [23], Paik, et al. [24], and Feng and Dillon [25] and they defined the counterparts of a record and an item. A record-like and an item-like of xml stream data have been described for the first time also in the work by Paik, et al [18]. In this subsection we briefly review main definitions. Full details can be found in the cited paper.

Let $\boldsymbol{T} = (T_1, T_2 \ldots T_n)$ be a given tree data with n numbers, where $n > 0$. The size of $\boldsymbol{T}$ depends on a total number of trees n, notated as $|T| = \sum_{i=1}^{n} T_i$. The equivalent part of a record and an item are defined as a fraction and titem (tree-item) respectively. When $\boldsymbol{F}$ is a set of fractions, the entire fractions for the given trees data can be expressed as $\boldsymbol{F} = \{F_{j,k} \mid F_{j,k} \preccurlyeq T_j\}$, where $1 \le j \le |\boldsymbol{T}|$ and $1 \le k$. Once fractions are collected from the original dataset $\boldsymbol{T}$, each one of fractions is eligible to be a titem. Like traditional association rules mining, the antecedent and consequent of tree data association rules are sets of titems that do not have any titem in common. The differentiation between positive and negative association rules comes from the way how to discover and apply titems to generate the rules.

As the first step to discover any kinds of rules is to obtain the smallest items by applying two factors, *sup*( ) and *conf*( ). The problem is the presented equations (1) to (6) cannot be used directly to our target dataset which is stored in tree structure. But the measures are for record data stored in tables. They should be appropriately suited to handle tree structured items, because the basic units of the association rules mining are not items but titems.

3.5 Constraints for Tree Structure Data

An positive association rule is the form X ⇒ Y, where the rule body X and head Y satisfy following two conditions: (1) X ∪ Y ⊂ ***F***, (2) X ∩ Y = ϕ. A support of X in a tree data set ***T*** is notated by the function *sup*(X) is rewritten with a titem in the equation (7). Also the support of a rule X ⇒ Y, *sup*(X ⇒Y), the support of an union of titems in X and Y, is expressed in (8).

$$sup(\mathrm{X}) = \frac{|X|}{|T|} = \sum_{i=1}^{n} \frac{T_i \mid X \subseteq T_i}{T_i} \quad (7)$$

$$sup(\mathrm{X} \Rightarrow \mathrm{Y}) = \frac{|X \cup Y|}{|T|} = \sum_{i=1}^{n} \frac{T_i \mid (X \subseteq T_i) \wedge (Y \subseteq T_i)}{T_i} \quad (8)$$

The confidence a rule X ⇒ Y to ***T*** determines how frequently the titems Y appears in trees that contain the titems X. The higher the confidence, the more likely it is for Y to be present in trees that contain X, which also provides an estimate of the conditional probability of Y given X. Because confidence is closely related to support, *conf*(X ⇒ Y) is rewritten with the equation (8),

$$conf(\mathrm{X} \Rightarrow \mathrm{Y}) = \frac{sup(X \Rightarrow Y)}{sup(X)} = \frac{|X \cup Y|}{|X|} = \sum_{i=1}^{n} \frac{T_i \mid (X \subseteq T_i) \wedge (Y \subseteq T_i)}{T_i \mid X \subseteq T_i} \quad (9)$$

According to the above equations along with equations (4) through (6), the support and confidence expressions for a negated titem X and a negative association X ⇒ ¬Y are suited as in the following:

$$sup(\neg \mathrm{X}) = 1 - \frac{|X|}{|T|} = 1 - \sum_{i=1}^{n} \frac{T_i \mid X \subseteq T_i}{T_i} \quad (10)$$

$$sup(X \Rightarrow \neg Y) = sup(\mathrm{X}) - sup(\mathrm{X} \Rightarrow \mathrm{Y}) = \sum_{i=1}^{n} \frac{T_i \mid X \subseteq T_i}{T_i} - \sum_{i=1}^{n} \frac{T_i \mid (X \subseteq T_i) \wedge (Y \subseteq T_i)}{T_i} \quad (11)$$

$$conf(X \Rightarrow \neg Y) = 1 - conf(X \Rightarrow \neg Y) = 1 - \sum_{i=1}^{n} \frac{T_i \mid (X \subseteq T_i) \wedge (Y \subseteq T_i)}{T_i \mid X \subseteq T_i} \quad (12)$$

A titem X is called frequent if the value of *sup*(X) is greater than or equal to the user specified ms. Otherwise, X is infrequent. For the positive

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

association rules, the set of infrequent titem is all pruned before any mining process is operated because they are useless. However, for negative association rules, infrequent titem sets are importantly considered due to their usefulness as shown in the previous page. Definitely the negated support, *sup*(¬X), satisfies usually because it inverses *sup*(X).

Figure 3 depicts a simple sensor web and presents examples of JSON sensor data. We can notice some data are frequently shown; the keys "weather", "humidity" and the value "rainy". It is informed that sensors are sending mainly the information that humidity is always checked whenever the status of weather is given. When we assume *ms* = 1, the discovered information is whenever the weather is rainy, the humidity information can be obtained. However, the specific humidity values cannot be known because each value of the humidity is different. Therefore, this giving information is quite ambiguous and less help. But what if we change the information to "whenever the weather is rainy, the value of humidity is **NOT** 60? By using the negated value, the specific values 80, 90, 77are able to be mined, and from the figure we can derive one fact that every location sensed is very humid when it is rainy day.

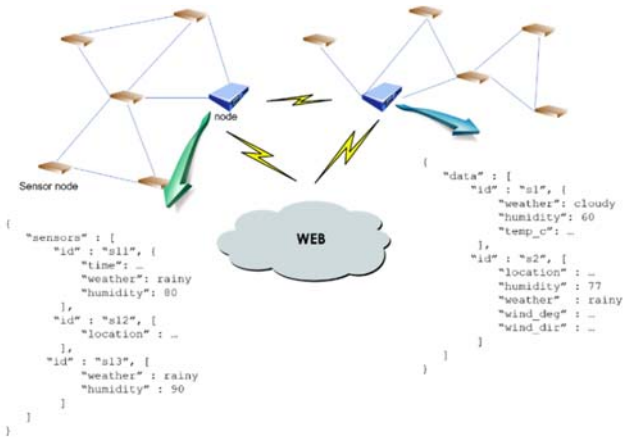

Figure3: Example of Simple JSON Data

Pruning must be done with care because *sup*(¬X) can be a high value if *sup*(X) is low according to the equation (4). The problem of how to efficiently prune uninteresting patterns and generate high quality candidates is an active and important part of data mining researches. Many studies have been conducted in this area, however, there is no widespread agreement as to how to compute the interest. The most common method in the association rules generation is the "support-confidence" approach [5, 6]. Although these two statistical methods perform the pruning of many unnecessary associations, they have the nature of the problem that is both approaches basically rely on frequency counts of patterns. Furthermore, there is a fundamental critique in that the same support threshold is being used for rules containing a different number of patterns. Therefore, it tends to miss necessary associations just with the support and confidence methods as presented on the simple examples.

4. PRUNING FRACTIONS FOR BOTH TYPES OF ASSOCIATION RULES

In addition to the support-confidence framework, other frameworks that add some measures have been suggested. They can be classified into two major types according to what is analyzed: interestingness vs. correlation. Interestingness measures play an important role in data mining, regardless of the kind of patterns being discovered. So far there is no universally accepted formal definition of interestingness, but generally these measures are intended for selecting and ranking patterns according to their potential interest [26]. Some popular measures are:.

- all-conference [27]: with this measure, an association is deemed interesting if all rules that can be produced from that association have a confidence greater than or equal to the minimum all-confidence value. This indicates that there is a dependency between all of the items in the association. The degree of the dependency is based on the threshold value.
- coverage [13, 28]: it measures how often a rule X ⇒ Y is applicable in a database. More specifically, coverage measures the comprehensiveness of a rule, that is, the part of all transactions in the dataset that matches the rule. If a rule characterizes more information in the dataset, it tends to be more interesting. Sometimes it is called antecedent support.
- lift (originally called interest) [29]: it measures how many times more often X and Y occur together than expected if they were statistically independent. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence of a rule is defined as the product of the support values of the rule body and the rule head divided by the support of the rule body. With the lift value, it can be interpreted the importance of a rule.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- conviction [29]: it was developed as an alternative to confidence which was found to not capture direction of associations adequately. Conviction compares the probability that X appears without Y if they were dependent with the actual frequency of the appearance of X without Y. In that respect it is similar to lift, however, it contrasts to lift it is a directed measure since it also uses the information of the absence of the consequent. An interesting fact is that conviction is monotone in confidence and lift.

The second alternative framework is to add the correlation coefficient approach. Correlation coefficient is a coefficient value that illustrates a quantitative measure of some type of correlation and dependence, meaning statistical relationships between two or more random variables or observed data values. It is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, one of the values increases as the other decreases.

In this section, we consider the alternative frameworks that add the measures of interestingness and correlation coefficient to efficiently prune fractions *for both positive and negative association rules mining*. And, we show how much those two approaches produce different analysis and result. For the aim, we adjust the interestingness measure used in [16] and the correlation coefficient measure applied in [13].

4.1 Interestingness Measure for Titems

With respect to the proposition argued in [30], $sup(X \cup Y) \approx sup(X) \times sup(Y)$, it can be known that the rule is not interesting if its antecedent and consequent are more or less independent. Based on the proposition, Wu et al. defined the function *interest*( ) in the paper [16] with a threshold minimum interest, *mi*,. The function computes a numerical value of a potential rule interest. If the produced value is less than *mi*, the input itemsets do not provide interesting association rules. Using the idea, a tailored function *interest*( ) covers titems. For a possible positive rule $X \Rightarrow Y$, a value of its interestingness is obtained by the equation (13).

$$interest(X \Rightarrow Y) = |sup(X \Rightarrow Y) - sup(X) \cdot sup(Y)| \quad (13)$$

The above equation cannot be used directly for a possible negative association rule $X \Rightarrow \neg Y$. Such as other measuring factors, it is derived by use of Y, that is expressed in the equation (14).

$$\begin{aligned} &interest(X \Rightarrow \neg Y) \\ &= |sup(X \Rightarrow \neg Y) - sup(X) \cdot sup(\neg Y)| \\ &= |sup(X) - sup(X \Rightarrow Y) - sup(X) \cdot sup(\neg Y)| \\ &= |sup(X) - sup(X \\ &\qquad \Rightarrow Y) - sup(X) \cdot (1 - sup(Y)| \\ &= |sup(X) \cdot sup(Y) - sup(X \Rightarrow Y)| \end{aligned} \quad (14)$$

The rule has rarely interesting information if $interest(X \Rightarrow Y) \approx 0$ or $interest(X \Rightarrow \neg Y) \approx 0$. However, it is worth to discover if the value is greater than or equal to *mi* even though its support and confidence are low. That means the fractions have no chance to be titems without applying interestingness. However, the factor interestingness itself also has weak point. It will be discussed in later.

4.2 Correlation-Coefficient Measure for Titems

Correlation Coefficient is another measurement to prune uninteresting items. It measures a strength of association between two variables [31]. The correlation coefficient value between random variables *a* and *b* is the degree of linear dependency, which is known as the covariance of the two variables, divided by their standard deviations (σ):

$$\rho_{ab} = \frac{Cov(a,b)}{\sigma_a \sigma_b} = \frac{E(ab) - E(a)E(b)}{\sigma_a \sigma_b}$$

where, the values $E(a)$, $E(ab)$ are the expected values. The range of ρ_{ab} is from -1 to +1. If $\rho_{ab} > 0$, those two variables are positively correlated. On the contrary, they are negatively correlated each other, if $\rho_{ab} < 0$. There is a strong correlation between *a* and *b* if ρ_{ab} is close to either -1 or +1. But, if $\rho_{ab} = 0$, *a* and *b* are independent each other, which means there is no reason to consider both variables together. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, the value of one increases as the value of the other decreases.

By Karl Pearson ϕ coefficient was introduced to evaluate the association between two itemsets X and Y. It measures the association for two binary values, 1 or 0. The reason why it can be used for itemsets is that the fundamental point of mining association rules is to consider the existence of an itemset in transactions; if an itemset exists it is regarded as 1, otherwise 0. Simply assumed *a* and *b*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

are two binary variables. The association between two binary variables can be summarized in a 2 × 2 contingency table given in Table 3. In the table, n_{11}, n_{10}, n_{01}, n_{00} are positive counts of numbers representing the existence of *a*, *b*, or both. And *n* is the total number of a data set. With the counts, the association of *a* and *b* is evaluated by the ϕ correlation coefficient. It is the equation (15) and the range of ϕ_{ab} is $-1 \le \phi \le 1$.

Table 3: 2 x 2 Contingency Table of two binary variables

	b = 1	*b = 0*	*total*
a = 1	n_{11}	n_{10}	n_{1+}
a = 0	n_{01}	n_{00}	n_{0+}
total	n_{+1}	n_{+0}	*n*

$$\emptyset_{ab} = \frac{n_{11}n_{00}-n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}}. \qquad (15)$$

By transposing n_{00} to $(n - n_{11} - n_{10} - n_{01})$, the terms of the expression are all identified to be 1s which mean 'exist';

$$\begin{aligned}\emptyset_{ab} &= \frac{n_{11}(n-n_{11}-n_{10}-n_{01})-n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}} \\ &= \frac{n_{11}n-n_{11}n_{11}-n_{11}n_{10}-n_{11}n_{01}-n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}} \\ &= \frac{nn_{11}-(n_{11}^2+n_{11}n_{10}+n_{11}n_{01}+n_{10}n_{01})}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}} \\ &= \frac{nn_{11}-(n_{11}+n_{01})(n_{11}+n_{01})}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}} \\ &= \frac{nn_{11}-n_{1+}n_{+1}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}} \\ &= \frac{nn_{11}-n_{1+}n_{+1}}{\sqrt{n_{1+}(n-n_{1+})n_{+1}(n-n_{+1})}} \qquad (16)\end{aligned}$$

Titems X and Y can configure a contingency table provided on Table 4. Each cell is a possible combination of X and Y with occurrence count, *sup*( ). Based on Table 4, the equation (16) is identified with support values of X and Y, that is the equation (17).

Table 4: 2 x 2 Contingency Table for Titems

	Y	¬*Y*	*sum*
X	*sup*(*X*⇒*Y*)	*sup*(*X*⇒¬*Y*)	*sup*(*X*)
¬*X*	*sup*(¬*X*⇒*Y*)	*sup*(¬*X*⇒¬*Y*)	*sup*(¬*X*)
sum	*sup*(*Y*)	*sup*(¬*Y*)	1

$$\emptyset_{XY} = \frac{sup(X \cup Y)-sup(X)\cdot sup(Y)}{\sqrt{sup(X)\cdot(1-sup(X))\cdot sup(Y)\cdot(1-sup(Y))}} \qquad (17)$$

The strength of correlation coefficient was described in the articles by Hopkins [32]. The author thought about carefully only positive values. Based on his arguments we re-define the statistical level of ϕ as; correlation of ±0.5 is large, ±0.3 is moderate, and ±0.1 is small, where anything which is smaller than ±0.1 is not worth to be considered. The given value, ±0.5, ±0.3, or ±0.1, called correlation threshold, is set by an input value or default value ±0.5. By adopting the correlation coefficient measure, the titemsets X and Y negatively correlated and leveled more than certain reliable strength are uncovered and used to generate informative negative association rules, even in the situation where their confidence values are reasonably high, but support values are less than a given *ms*.

4.3 Discovery of Titems for Two Types of Associations Rules

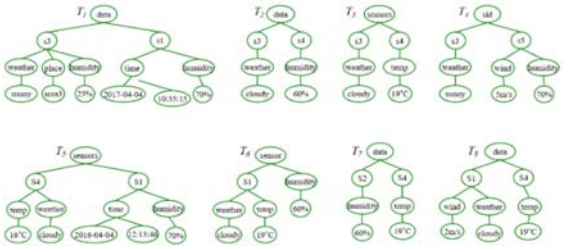

*Figure5: Dataset **T** with 8 Tree-Structured Data*

With an example dataset ***T*** on Figure 5, four constraints – support, confidence, interestingness, and correlation coefficient – are taken to verify their different influence for a set of titems. A fraction set ***F*** is built from ***T***, which has enormous numbers of fraction because it follows the subtrees calculation. For simplicity's sake, we assume that each fraction $F \in \boldsymbol{F}$ has been already organized. We choose following five sample fractions for further examples. Table 5 provides all thresholds that are applied to the fractions.

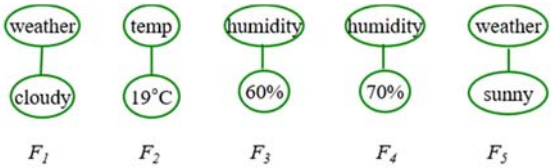

Figure 6: Five Sampling Fractions

Journal of Theoretical and Applied Information Technology
15th September 2019. Vol.97. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Four Constraints Threshold Values

threshold	vale	implication
minimum support	*0.3*	NA
minimum confidence	*0.5*	NA
minimum interest	*0.1*	NA
correlation coefficient	± 0.1	no association
	± 03	moderate association
	± 0.5	strong association

$$sup(F_1) = \frac{5}{8} = 0.65,\ sup(F_2) = \frac{4}{8} = 0.5$$
$$sup(F_3) = \frac{2}{8} = 0.25,\ sup(F_4) = \frac{3}{8} = 0.375,$$
$$sup(F_5) = \frac{2}{8} = 0.25\ .$$

As an initial step a threshold *ms* is applied to every fraction itself. The fractions F_1, F_2, F_4 are eligible to be titems and the other two F_3, F_5 are pruned for a generation of positive association rules. However, those two are not pruned because their negated support values satisfy *ms*.

$$sup(\neg F_3) = sup(\neg F_5) = 1 - \frac{2}{8} = 0.75$$

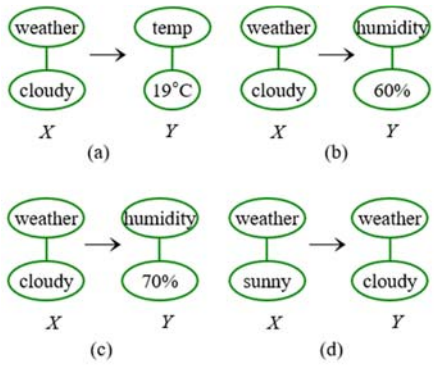

Figure 7: Four Candidate Association Rules

Titems are classified into two groups: titems for positive association rules and titems for negative association rules. Their notations are P_T and N_T respectively: $P_T = \{F_1, F_2, F_4\}$, $N_T = \{F_3, F_5\}$. Based on the titems obtained, Figure 7 provides four possible candidate association rules.

For the first computational factor, the ordinary support-confidence measure is applied to each rule, equations (7) ~ (9) with *ms* = 0.3 and *mc* = 0.5;

(a) $sup(X \Rightarrow Y) = \frac{3}{8} = 0.375$
$conf(X \Rightarrow Y) = \frac{3}{5} = 0.6,$
(b) $sup(X \Rightarrow Y) = \frac{2}{8} = 0.25$
$conf(X \Rightarrow Y) = \frac{2}{5} = 0.4,$
(c) $sup(X \Rightarrow Y) = \frac{1}{8} = 0.125$
$conf(X \Rightarrow Y) = \frac{1}{5} = 0.2,$
(d) $sup(X \Rightarrow Y) = \frac{0}{8} = 0$
$conf(X \Rightarrow Y) = \frac{0}{2} = 0.$

As a result, the rule (a) only is qualified to be a positive association rule and its titems X (F_1) and Y (F_2) are together processed for the further step. The others are no longer used even F_4. For not qualified three rules, we compute their interestingness and correlation factors. The correlation coefficient constraint, equation (17), is first applied to decide titems X and Y are related positively or negatively.

(b) $\emptyset_{XY} = \frac{1 \cdot \frac{2}{8} - \frac{5}{8} \cdot \frac{3}{8}}{\sqrt{\frac{5}{8} \cdot \left(1 - \frac{5}{8}\right) \cdot \frac{3}{8} \cdot \left(1 - \frac{3}{8}\right)}} = \frac{1}{15} \approx 0.07$

(c) $\emptyset_{XY} = \frac{1 \cdot \frac{1}{8} - \frac{5}{8} \cdot \frac{3}{8}}{\sqrt{\frac{5}{8} \cdot \left(1 - \frac{5}{8}\right) \cdot \frac{3}{8} \cdot \left(1 - \frac{3}{8}\right)}} = -\frac{7}{15} \approx -0.47$

(d) $\emptyset_{XY} = \frac{1 \cdot 0 - \frac{2}{8} \cdot \frac{5}{8}}{\sqrt{\frac{2}{8} \cdot \left(1 - \frac{2}{8}\right) \cdot \frac{5}{8} \cdot \left(1 - \frac{5}{8}\right)}} = -\frac{10}{\sqrt{180}} \approx -0.74$

It is informed that the candidate rule (b) is not worth to be considered because its value is less than +0.1, because two titems are nearly independent each other and the association between them is seldom made. It assists the result of support-confidence in which turns out (b) is less reliable rule. In comparison, the other two reveal totally different outcomes. Their correlation coefficient values are close to -0.5 or even more: the relation between titems are strongly associated, especially negative way. The values direct the rule (c) and (d) mostly gives strongly reliable information together and recommend it is worthy to discover them as negative association rules. This fact was already proven in the previous page, because their support-confidence constraints were seriously less than *ms*. For more trusty results, we identify how much the rules (c) and (d) are interested to aid the correlations. According to the equation (14), following results are drawn;

(c) $interest(X \Rightarrow \neg Y) = \left|\frac{5}{8} \cdot \frac{5}{8} - \frac{1}{8}\right| \approx 0.109\ ,$
(d) $interest(X \Rightarrow \neg Y) = \left|\frac{2}{8} \cdot \frac{5}{8} - 0\right| \approx 0.16.$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

When the rules are mined as negative association rules, their interestingness satisfy a given mi. Such a result cannot be derived by the support-confidence approach. Even worse, they would be pruned if support-confidence constraints were only applied. For more evidence, interestingness of (b) is computed in two ways. It reveals a rule (b) has no interesting information as a positive rule nor negative rule.

$$\text{(b) } interest(X \Rightarrow Y) = \left|\frac{2}{8} - \frac{5}{8} \cdot \frac{3}{8}\right| = 0.0156,$$
$$interest(X \Rightarrow \neg Y) = \left|\frac{5}{8} \cdot \frac{2}{8} - \frac{2}{8}\right| \approx 0.09$$

The next example explains why correlated coefficient factor is used before the factor interestingness.

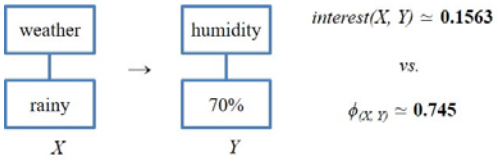

Figure 8: Gap of interestingness vs. correlation coefficient value

The support-confidence values of the rule on Figure 8 are 0.25 and 1, respectively. The association between X and Y does not satisfy the condition and is pruned from the positive association rule generation, even though two titems have the strong tandem. Under the support/confidence framework, there is no chance to consider them. To prevent such erroneous outcome, the constraint interestingness is used to verify how much the rule has interesting information. When the given *mi* is 0.3, the candidate rule is still pruned because it is determined as 'uninteresting'. The rule could be interest to be mined if *mi* would be set less than 0.15. The determination of interestingness is highly dependable on setting up *mi*. But, the value of correlation coefficient gives a more objective result. Its obtained value is approximately 0.75, which means there is a strong positive correlation between X and Y and cannot be ignored. As proved by the confidence value, the titems Y always occurs if the titems X occurs. Concerning the coefficient determination first is more helpful to decide what type of association rule is appropriate to candidate rules.

The following algorithm broadly outlines the procedure explained in previous pages. It determines the way how to apply four measuring factors in order to discover appropriate titems used to generate positive and negative associations in a straight line. In use of two more constraints, interestingness and correlation coefficient values, verifies that 1) the relation between titems is positively correlated or negatively, 2) the strengths of their coefficients are quite strong enough to give valuable information, and 3) the generated association rule will provide many opportunities for further mining, even though its support value is less than *ms* and the rule are not attracted in the positive type. With the frame of correlation coefficient, the hidden association provides benefits when it is mined for a negative rule, which is not caught by support/confidence or even interest.

```
INP: F, ms, mc, mi, ϕ
OUTP: PT, NT

1. for each fraction F ∈ F
2.     if sup(F) ≥ ms
3.       PT ← PT + {F}
4.     else
5.       NT ← NT + {F}
6. For titems X, Y ∈ PT
                        form X ⇒ Y
7. If sup(X⇒Y) < ms or
                        conf(⇒Y) < mc
8.     Then switch(ϕXY)
9.     Case (ϕXY ≤ -0.3):
10.      If interest(X ⇒ ¬Y) ≥ mi
11.         NT ← NT + {X, Y}
12.      else
13.         NT ← NT - {Y}
14.    Case (ϕXY ≥ 0.3):
15.      If interest(X ⇒ ¬Y) ≥ mi
16.         PT ← PT + {X, Y}
17.      else
18.         PT ← PT - {Y}
19. else
20.     then switch(ϕXY)
21.     Case (ϕXY ≤ -0.3):
22.       If interest(X ⇒ ¬Y) ≥ mi
23.          NT ← NT + {X, Y}
24.     Case (ϕXY ≤ 0.1):
25.       If interest(X ⇒ Y) < mi
26.          PT ← PT - {Y}
24. Return PT, NT
```

4.4 Advantage

Compared to previously suggested methods, applying two more measuring factors to both types of association rules improves the quality of titemsets

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and makes association rules be more accurate and useful. On Figure 7, the candidate rules (c) and (d) would have been pruned definitely if they had been applied by support and confidence only for the positive association rules. In particular, the values of (d) have all 0s, which means there is no chance to find it. (d) would be pruned definitely. As a result, the rules (c) and (d) won't be able to be obtained even though it has high reliability and strong association *negatively*. However, the proposed algorithm prevents those important titemsets from being pruned and supports to discover valuable negative association rules consequently, as well as the original positive association rules. It is the main advantage of the proposed algorithm.

5. CONCLUSION

In this work, the author considered how to efficiently obtain negated tree itemsets for negative association rules from xml-based stream data. For the purpose, the primarily considered part was to evaluate fractions of xml-based data whether they could generate informative negative rules or not, even if their support and confidence values were not enough to the given conditions. Only with the support-confidence framework tended to mistakenly prune useful titems, thus, other frameworks that added some measures were suggested as the alternatives; interestingness and correlation coefficient. We adjusted both measures for our data to determine non-existing but important titemsets. The example results of interestingness and correlation coefficient were presented and compared with a few illustrations based on the algorithm *DNTS*. We drew out it would be more efficient and reliable to prune fractions with the correlation determination than that of interestingness, too. We presented for the first time the analyses of both interestingness and correlation coefficient methods over tree structured stream data. Future work includes presenting a full mining algorithm and experimental results of negative association rules for tree structured stream data that is proven to work with the four measurements.

ACKNOWLEDGMENTS

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A2B1007015).

REFERENCES:

[1] M. Botts, G. Percivall, C. Reed, and J. Davidson, "OCG® Sensor Web Enablement: Overview and High Level Architecture", Proceedings of the 2nd International Conference on GeoSensor Networks, October 1-3, 2006, USA, LNCS 4540, pp. 175-190.

[2] A. Bröring, J. Echterhoff, S. Jirka, I. Simonis, T. Everding, C. StSCH, s. Kiang, and R. Lemmens, "New Generation Sensor Web Enablement", Sensors 2011, 11, pp. 2652-2699.

[3] S. Mahmood, M.Shahbaz, and A. Guergachi, "Negative and Positive Association Rule Mining from Text Using Frequent and Infrequent Itemsets", The Scientific World Journal, Vol. 2014, ID 973750, 2014, 11 pages.

[4] K.K. Loo, I. Tong, B. Kao, and D. Chenung, "Online Algorithms for Mining Inter-Stream Associations from Large Sensor Networks", Advances in Knowledge Discovery and Data Mining, Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 18-20, 2005, Vietnam, LNCS 3518, pp. 143-149.

[5] A. Savasere, E. Omiecinski, and S. Navathe, "Mining for Strong Negative Associations in a Large Database of Customer Transactions", Proceedings of the 14th International Conference on Data Engineering, February 23-27, 1998, USA, pp. 494-502.

[6] M. L. Antonie and O. R. Zaïane, "Mining Positive and Negative Association Rules: An Approach for Confined Rules", Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 20-24, 2004, Italy, LNCS 3202, pp. 27-38.

[7] Z. Honglei and X. Zhigang, "An Effective Algorithm for Mining Positive and Negative Association Rules", Proceedings of International Conference on Computer Science and Software Engineering, December 12-14, 2008, China, pp. 455-458.

[8] R. Sumalatha and B. Ramasubbareddy, "Mining Positive and Negative Association Rules", International Journal on Computer Science and Engineering, Vol. 2, No. 09, 2010, pp. 2916-2910.

[9] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules", ACM Transaction on Information Systems, Vol. 22, No. 03, 2004, pp. 381-405.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[10] W. Ouyang, "Mining Positive and Negative Association Rules in Data Streams with a Sliding Window", Proceedings of the Fourth Global Congress on Intelligent Systems, December 3-4, 2013, Hong Kong, pp. 205-209.

[11] S. Corpinar and T. Í. Gündem, "Positive and Negative Association Rule Mining on XML Data Streams in Database as a Service Concept", Expert Systems with Applications, Vol. 39, No. 8, 2012, pp. 7503-7511.

[12] J. Paik, J. Nam, U. Kim, and D. Won, "Association Rule Extraction from XML Stream Data for Wireless Sensor Networks", Sensors, Vol. 14, 2014, pp. 12937-12957.

[13] X. Yuan, B. P. Buckles, Z. Yuan, and J. Zhang, "Mining Negative Association Rules", Proceedings of the 7th International Symposium on Computers and Communications, July 1-4, 2002, Italy, pp. 623-628.

[14] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of the ACM SIGMOD international conference on Management of Data, May 25-28, 1993, USA, pp. 207-216.

[15] D. Braga, A. Campi, S. Ceri, M. Klemettinen, and P. L. Lanzi, "A Tool for Extracting XML Association Rules", Proceedings of the 14th IEEE International Conference on Tools with artificial Intelligence, November 4-6, 2002, USA, pp. 57-64.

[16] J. Paik, J. Nam, S. Lee, and U. Kim, "A Framework for Data Structure-Guided Extraction of XML Association Rules", Proceedings of the 7th International Conference on Computational Science, May 27-30, 2007, China, LNCS 4489, pp. 709-716.

[17] L. Feng and T. Dillon, "Mining Interesting XML-Enabled Association Rules with Templates", Proceedings of the 3rd International Workshop on Knowledge Discovery and Inductive Databases, September 20, 2004, Italy, LNCS 3377, pp. 66-88.

[18] L. Geng and H. J. Hamilton, "Interestingness Measures for Data Mining: A Survey", ACM Computing Survey, Vol. 38, No. 1, 2006, article no. 9.

[19] G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules", Knowledge Discovery in Databases, Edited G. Piatetsky-Shapiro and W. J. Frawley, AAAI Press, 1991, pp.229-248.

[20] J. Cohrn, "Statistical Power Analysis for the Behavioral Sciences", Lawrence Erlbaum), New Jersey, 1988, pp.109-143.

[21] W. Hopkins, "A New View of Statistics", Electronic edition, Available: http://www.sportsci.org/resource/stats/.

5. CONCLUSION

In this work, we considered how to appropriately prune tree structured items, called fraction, both for positive and negative association rules mining. For the purpose, the primarily considered part was to verify fractions from the tree modeled dataset whether they could generate informative negative rules or not, even if their support and confidence values were not enough to the given constraints. Only with the support-confidence framework tended to mistakenly prune useful titems, thus, other frameworks that added some measures were suggested as the alternatives; interestingness and correlation coefficient. We adjusted both measures for our data to determine non-existing but important titems. Besides, it was supported in the discovery of positive association rules.

The example results of each constraint were presented and compared. We drew out it would be more efficient and reliable to prune fractions with the correlation determination than that of interestingness, too. Future work includes presenting a full mining algorithm and experimental results, that is proven to work with the four constraints.

ACKNOWLEDGMENTS

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A2B1007015).

REFRENCES:

[1] J. Manyika and M. Chui, "By 2025, Internet of Things Applications Could Have $11 Trillion Impact",Available:http://www.mckinsey.com/insights/mgi/, 2015.

[2] J. Bughin, J. Manyika, and J. Woetzel, "The Age of Analytics: Competing in a Data-Driven World", December 2016, McKinsey Global Institute

[3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and Issues in Data Stream

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Systems", Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, June 2-6, 2002, USA, pp. 1-16.

[4] S. Mahmood, M.Shahbaz, and A. Guergachi, "Negative and Positive Association Rule Mining from Text Using Frequent and Infrequent Itemsets", The Scientific World Journal, Vol. 2014, ID 973750, 2014, 11 pages.

[5] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, 1993, USA, pp. 207-216.

[6] J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases", Proceedings of the 21st International Conference on Very Large Data Bases, September 11-15, 1995, pp. 420-431.

[7] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns Without Candidate Generation", Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 15-18, 2000, USA, pp. 1-12.

[8] R. Wolff and A. Schuster, "Association Rule Mining in Peer-to-Peer Systems", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 34, 2004, pp. 2426-2438.

[9] A. Boukerche and S. Samarah, "A Novel Algorithm for Mining Association Rules in Wireless Ad Hoc Sensor Networks", IEEE Transactions on Parallel and Distributed Systems, 19, 2008, pp. 865-877.

[10] M. M. Rashid, I. Gondal, and J. Kamruzzaman, "Mining Associated Patterns from Wireless Sensor Networks", IEEE Transactions on Computers, 64, 2014, pp. 1998-2011.

[11] K.K. Loo, I. Tong, B. Kao, and D. Chenung, "Online Algorithms for Mining Inter-Stream Associations from Large Sensor Networks", Advances in Knowledge Discovery and Data Mining, Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 18-20, 2005, Vietnam, LNCS 3518, pp. 143-149.

[12] A. Savasere, E. Omiecinski, and S. Navathe, "Mining for Strong Negative Associations in a Large Database of Customer Transactions", Proceedings of the 14th International Conference on Data Engineering, February 23-27, 1998, USA, pp. 494-502.

[13] M. L. Antonie and O. R. Zaïane, "Mining Positive and Negative Association Rules: An Approach for Confined Rules", Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 20-24, 2004, Italy, LNCS 3202, pp. 27-38.

[14] Z. Honglei and X. Zhigang, "An Effective Algorithm for Mining Positive and Negative Association Rules", Proceedings of International Conference on Computer Science and Software Engineering, December 12-14, 2008, China, pp. 455-458.

[15] R. Sumalatha and B. Ramasubbareddy, "Mining Positive and Negative Association Rules", International Journal on Computer Science and Engineering, Vol. 2, No. 09, 2010, pp. 2916-2910.

[16] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules", ACM Transaction on Information Systems, Vol. 22, No. 03, 2004, pp. 381-405.

[17] S. Corpinar and T. Í. Gündem, "Positive and Negative Association Rule Mining on XML Data Streams in Database as a Service Concept", Expert Systems with Applications, Vol. 39, No. 8, 2012, pp. 7503-7511.

[18] J. Paik, J. Nam, U. Kim, and D. Won, "Association Rule Extraction from XML Stream Data for Wireless Sensor Networks", Sensors, Vol. 14, 2014, pp. 12937-12957.

[19] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", Proceedings of the 20th International Conference on VLDB, September 12-15, 1994, pp. 478–499.

[20] A. Savasere, E. Omiecinski, and S. Navathe, S. "Mining for Strong Negative Associations in a Large Database of Customer Transactions", Proceedings of the 14th International Conference on Data Engineering, February 23-27, 1998, USA, pp. 494-502.

[21] F. Hussain, H. Liu, E. Suzuki, and H. Lu, "Exception Rule Mining with a Relative Interestingness Measure", Proceedings of the 3rd Pacific Asia Conference on Knowledge Discovery and Data Mining, April 18-20, 2000, Japan, LNCS 1805, pp. 87-97.

[22] X. Yuan, B. P. Buckles, Z. Yuan, and J. Zhang, "Mining Negative Association Rules", Proceedings of the 7th International Symposium on Computers and Communications, July 1-4, 2002, Italy, pp. 623-628.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[23] D. Braga, A. Campi, S. Ceri, M. Klemettinen, and P. L. Lanzi, "A Tool for Extracting XML Association Rules", Proceedings of the 14th IEEE International Conference on Tools with artificial Intelligence, November 4-6, 2002, USA, pp. 57-64.

[24] J. Paik, J. Nam, S. Lee, and U. Kim, "A Framework for Data Structure-Guided Extraction of XML Association Rules", Proceedings of the 7th International Conference on Computational Science, May 27-30, 2007, China, LNCS 4489, pp. 709-716.

[25] L. Feng and T. Dillon, "Mining Interesting XML-Enabled Association Rules with Templates", Proceedings of the 3rd International Workshop on Knowledge Discovery and Inductive Databases, September 20, 2004, Italy, LNCS 3377, pp. 66-88.

[26] L. Geng and H. J. Hamilton, "Interestingness Measures for Data Mining: A Survey", ACM Computing Survey, 38, 2006, pp. 1-32.

[27] E. R. Omiecinski, "Alternative Interest Measures for Mining Associations in Databases", IEEE Transactions on Knowledge and Data Engineering, 15, 2003, pp. 57-69.

[28] G. I. Webb and D. Brain, "Generality is Predictive of Prediction Accuracy", Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop, 2002, Japan, LNCS 3755, pp. 117-130.

[29] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", Proceedings of the ACM SIGMOD International Conference on Management of Data, May 11-15, 1997, USA, pp.~255-264.

[30] G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules", Knowledge Discovery in Databases, Edited G. Piatetsky-Shapiro and W. J. Frawley, AAAI Press, 1991, pp.229-248.

[31] J. Cohrn, "Statistical Power Analysis for the Behavioral Sciences", Lawrence Erlbaum), New Jersey, 1988, pp.109-143.

[32] W. Hopkins, "A New View of Statistics", Electronic edition, Available: http://www.sportsci.org/resource/stats/.