Intelligent Automation And Soft Computing, 2020

Vol. 26, no. 1, 193–204
https://doi.org/10.31209/2019.100000140

Weighted or Non-Weighted Negative Tree Pattern Discovery from Sensor-Rich Environments

Juryon Paik

#521, 2nd Pierson Bldg., Department of Digital Information & Statistics, Pyeongtaek University, 3825, Seodong-daro, Pyeongtaek-si, Gyeonggi-do 17869, South Korea

ABSTRACT
It seems to be sure that the IoT is one of promising potential topics today. Sensors are the one that lead the current IoT revolution. The advances of sensor-rich environments produce the massive volume of raw data that is enlarging faster than the rate at which it is being handled. JSON is a lightweight data-interchange format and preferred for IoT applications. Before JSON, XML was de factor standard format for interchanging data. The common point is that their structure scheme is the tree. Tree structure provides data exchangeability and heterogeneity, which encourages user-flexibilities. Therefore, JSON sensor format is an easy to use human readable format for storing and transmitting sensor values. However, it is more challenging than ever to discover valuable and hidden information from the continuously generated tree-structured data. In the paper, we define and suggest an original method to predict and evaluate from the tree-structured sensing data.

KEY WORDS: Negative association rules, pruning constraints, tree-structured items, weighted tree items.

1 INTRODUCTION

RECENTLY the Internet of Things (IoT) has earned enormous praises and widespread adoption in not only academia but industry. Armed with cutting-edge technologies, it provides vast opportunities and revolutionizes the world. McKinsey Global Institute issued the report by J. Manyika (2015) explaining that IoT would provide a potential economic impact by 2025 of $3.9 trillion to $11.1 trillion per year. There is no doubt that IoT is the next big thing. The term was first coined by Kevin Ashton (2009). He believed IoT would turn the world into data and by processing it macro decisions could be made on resource utilizations. This is what IoT aims at. For the purpose, it harnesses and analyzes data usually from connected devices and sensors. Actually, 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.

Fused with critical endpoint technologies, the cutting-edge IoT environment inevitably produces tremendous volumes of stream data. It is required to handle efficiently and exchange desirably the large heterogeneous datasets. XML (eXtensible Markup Language) was originally designed to carry data, not to display data. It describes the rules to encode xml documents in a format which is both human and machine readable. XML is used widely for the representation of arbitrary data structures due to its generality, flexibility, and heterogeneity across the internet. 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 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. Instead, the data structures are more complicated and harder to analyze. The leading

CONTACT Juryon Paik jrpaik@ptu.ac.kr

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.

One of the biggest challenges is finding valuable but hidden information. This issue has been commonly addressed as *streaming data mining*, which is usually processed by analysis tools such as mining association rules, classification or clustering. When applying conventional methods to mine sensor-generating data, researchers were met with weighty challenges. 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, sensor data is 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. (2002). However, it is definite the streaming raw data provides the predictive insights for fully facilitating IoT environments if it is properly analyzed and evaluated. That drives data mining researchers to develop mining technologies for stream data.

Recently a significant research resented by Mahmood, et. al. (2014) focuses on discovery of interesting but non-existing or infrequent data, such researching topic is called the discovery of negative association rules. 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 tree data.

In such environments handling tree-structured streaming data must be desirable but Herculean task. It is intricate process to analyze streaming tree-structured data, that is a main reason why many of problems related to tree data cannot be adequately figured out yet. Discovery of negative association rules from streaming tree-structured data is still in an immature stage and not fully developed. The aim of this paper is to describe what the negative association rule is, suggest efficient methods to discover non-existing data parts, define a concept of weight, and present computational results, over the streaming data of tree-structured.

2 RELATED WORK

ISSUES of finding associated patterns was introduced first by Agrawal, et. al (1993). The actual aim was to analyze customer behaviors and capture information from market basket transactions. The identified patterns, called rules, are those items which are very frequently purchased together with other particular 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 tremendous number of variations and developments of mining association rules have been proposed and still actively studied by many researchers such as Han and Fu (1995), Han, et al. (2000), Wolff and Schuster (2004), Boukerche and Samarah (2008), Rashid, et al. (2014), and so on.

Since the 2000s deployments of sensor network have been rapidly increased, which caused massive volumes of data. Data mining communities have started into finding association rules for the gigantic size of streaming data. One of the beginning researchers for sensor data, Loo, et al. (2005) published a sensor data association rule mining framework. Sensors' values were mainly considered to originate association rules and intervals are made from the time by dividing. With interval list based lossy counting, 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 discovery for negative association rules is even bigger than that of positive association rules. However, there are very few research works conducted on negative association rules mining for data stream. Most of the published articles, such as Savaere, et al. (1998), Antonie and Zaïane (2004), Honglei and Zhigang (2008), and Sumalatha and Ramasubbareddy (2010), 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. (2004). 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 volume of database is small. If the database becomes larger, it would be more difficult. Particularly, it is a challenging problem to identify which one of rules is beneficial or useful to applications from the enormous and rigorous size of streaming data.

A solution for both positive and negative association rules computation was suggested by Corpinar and Gündem (2012). The important characteristic is that its data type is streaming XML data. In order to association XML stream data, the first adapted method is the original FP-Growth. They developed new pruning thresholds to reduce the search space for negative association rules. To identify frequent sets for positive association rules and

negative association rules, Corpinar and Gündem applied correlation coefficient parameter.

Another paper issued by Paik, et al. (2014) provides a scheme for mining xml stream data along with some useful definitions. The authors insist the scheme is a first approaching method for mining xml stream data in point of generating frequent tree items without redundancies. It is possible their method to apply for both individual block and whole stream.

The previous two papers commonly mention that it is expensive and complicated task to manage continuously arriving xml data, which cause many problems. It is our consideration also to extract informing tree-structured itemsets from streaming tree data for negative association rules. For the aim, pruning methods are mainly developed because negative association rules are generally configured from large numbers of infrequent tree items. The importantly discussed matter is the major constraint factors for the pruning phase. Then we apply for the first time the concept of weight for negative tree items. Then, we show computational results simply how the constraint factors draw out different results and affect the outcome of the prunings.

3 PRELIMINARIES

3.1 Tree-structured stream data

```
{
   "firstName": "John",
   "lastName": "Smith",
   "age": 25,
   "address": {
       "streetAddress": "21 2nd Street",
       "city": "New York",
       "state": "NY",
       "postalCode": "10021-3100"
   },
   "gender": {
       "type": "male"
    }
   "spouse": null
   "phoneNumbers": [
    {
       "type": "home",
       "number": "212 555-1234"
    },
    {
       "type": "office",
       "number": "646 555-4567"
    },
    {
       "type": "mobile",
       "number": "123 456-7890"
    }
   ]
}
```

```
<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>
   <gender>
       <type>male</type>
   </gender>
   <spouse></spouse>
   <phoneNumber>
       <type>home</type>
       <number>212 555-1234</number>
   </phoneNumber>
   <phoneNumber>
       <type>office</type>
       <number>646 555-4567</number>
   </phoneNumber>
   <phoneNumber>
       <type>mobile</type>
       <number>123 456-7890</number>
   </phoneNumber>
</person>
```

Figure 1. JSON (left) and XML (right) codes.

ENTERING the world of IoT, one of the most popular data encoding formats is JSON. At the world of the Web, it was XML. Their common feature is that they are structural documents with trees structures. Figure 1 shows representation of both json and xml describing a person. At a glance, xml is more verbose than json in general even though they represent same information. However, their structural tree schemes are similar as illustrated on the figure 2.

Because of flexibility, easy interchangeability and lightweight, json is used as an alternative to xml and is a very common data format to transmit and read data from sensors in an increasing number of IoT applications. To make better usage for the overwhelmed stream data, it is required to first understand and optimize the tree structure. Sensors continuously transmit their data to sink node, which means json documents structured with tree continuously stream with a fast speed. Several researchers published the papers related xml tree data and defined useful definitions about tree data. We use their definitions and terms in this paper because the tree of json is almost similar to that of xml.

3.2 Negative association rules

Our physical world is detected and measured changes by sensors and is turned billions of objects

into data producing things. Inevitably, massive amounts of digital data are produced unceasingly. The leading technology users such as business managers and researchers express the frustration about being unable to harvest fully the benefits from the data flowing. One of the biggest challenges is finding valuable but hidden information. This issue has been commonly addressed as *mining stream data*. Several methodologies have been proposed and still ongoing. Among them, discovery of *negative* association rules is being focused widely because its usage is rapidly growing in diverse areas. Negative association is the association that negates presence, as opposed to positive associations. More specifically defined by Yuan, et al. (2002), a negative association rule is the rule that comprises relationship between absent items and present items. The famous positive association rule "bread implies milk", expressed in *bread* ⇒ *milk*, indicates that customer's buying behavior pattern of purchasing bread and milk together. The following is another association: "customers buy Coke *do not* buy Pepsi". The association considers the absent item Pepsi with the present item Coke. The rule is expressed in *Coke* ⇒ ¬*Pepsi*. Association rules that include absent items are turning out to be as valuable as positive association rules. The discovery of negative association rules is a tricky and computationally hard problem because absent items have to be considered. Nevertheless, it is highly interesting and potentially useful.

To obtain interesting rules, various measures for constraints are applied basically. The well-knowns are the famous minimum thresholds on *support* and *confidence* values introduced by Agrawal, el al (1993). When $\mathbf{I} = \{I_1, I_2 \ldots I_n\}$ is an items set from a transaction database **D**, and $X \Rightarrow Y$, which conditions are $X \subset I \wedge Y \subset I \wedge X \cap Y = \phi$), is a positive association rule. Support of X with respect to **D** is a proportion of transactions that contains all items in X it is the function $sup(X)$,

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

Support is an indication of how frequently the itemset appears in the dataset. Hence, $sup(X \Rightarrow Y)$ is the support of an union with the items in X and Y. The value of $sup(X \Rightarrow Y)$ is written as the equation (2).

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

Confidence is the statistical measure of how often a rule $X \Rightarrow Y$ has been discovered together. It is the proportion of the transactions that have X which also have Y, written as a function $conf(X \Rightarrow Y)$ and re-expressed as,

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

Equations (1) to (3) are used in finding positive associations to the present items. However, the support and confidence cannot be used directly to find negative association rules, because negative associations encapsulate the relationships between absent and present items. Counts of non-existing items must be verified to measure support and confidence,

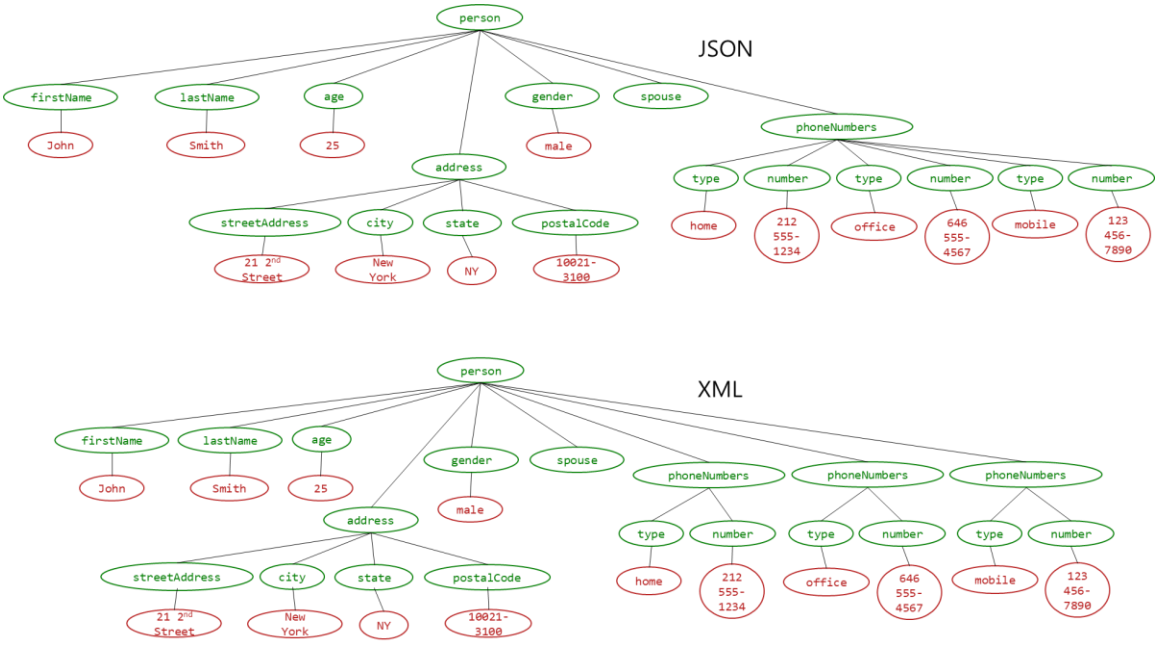

Figure 2. Tree structure of JSON (top) and XML (bottom).

nevertheless, it is hard to directly count the absent items. Instead, we derive both values from the equation (1) and (2). For any negative association rule $X \Rightarrow \neg Y$, the expressions of its support and confidence are the following.

$$sup(\neg X) = 1 - sup(X) \tag{4}$$

$$sup(X \Rightarrow \neg Y) = sup(X) - \sup(X \Rightarrow Y)$$
$$= \frac{|X|}{|\mathbf{D}|} - \frac{|X \cup Y|}{|\mathbf{D}|} = \frac{|X| - |X \cup Y|}{|\mathbf{D}|} \tag{5}$$

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

Table 1 shows an extremely small example of the market baskets. A set of items **I** = {bread, butter, diaper, jam, milk, water} and the table is a transaction database **D**. When we assume a minimum support, user specified given support, 0.75, only the items {bread, butter, milk} satisfy the constraint. Seeding from them, positive association rules are constructed. For the process required searching space is 2^3 items. The highly interesting rule 'bread ⇒ butter' is obtained by 0.75 support and 1.0 confidence. However, what if the items in the baskets have been changed; some customers take items out of their baskets or some replace a few items with others. For example, the item 'butter' was taken out of the transactions T_1 and T_2. Also, it was replaced with the item 'cheese' in T_4. Consequently, the item 'butter' is no more frequent item because its support is just 0.25 which cannot satisfy the minimum support.

Table 1. Transaction data of market basket

TID	*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 number of infrequent items is 7, such means the searching space is 2^7 items to find the infrequent but interesting item for the negative rules. As one of negative rues the following rule can be informed; the customers who put bread into their baskets usually take out the 'butter' from their baskets just before payments are made. The customers who buy 'bread' typically do not buy 'butter' together. It can be written formally as $bread \Rightarrow \neg butter$. The rule is highly interesting because its support value is 0.75 and the confidence value is approximately 0.67. This negative association rule has the high strength indicating that the rule is very reliable and helpful to market basket analysis. Analyzing negative association rules is as important as or more than that of positive association rules.

3.3 Support and confidence for tree data

The basic two constraints, support and confidence, are allowed for the absent items by the equation (4) to (6). A problem, however, still remains. The above equations are for record data stored in tables not for tree-structured data streamed from sensors. Several researchers published their papers related to xml association rules, such as Braga, et al. (2002), Paik, et al. (2007), and Feng and Dillon (2004) and they defined the counterparts of a record and an item. Based on those papers, we described a record and an item of xml stream data for the first time in the previous work by Paik, et al (2014). In this subsection the definitions are briefly stated. Full details can be found in the cited paper.

Generally, data stream is transferred in a series of blocks. Each block has its own maximum length, block size, though, we assume all blocks are of identical sizes, for simplicity. In the paper, streaming data is a continuous sequence of blocks containing same numbers of trees. Let $\mathbf{S} = (TB_1, TB_2 \ldots TB_L)$ be a given tree-structured streaming data arrived by the latest block TB_L. Each block TB_i contains a timestamp t_i and a trees set; $TB_i = (t_i, \{T_1, T_2 \ldots T_n\})$, where $n > 0$. The size of **S** is determined by a total number of trees reaching in the latest timestamp t_L.

$$|\mathbf{S}| = \sum_{i=1}^{L} |TB_i| = |TB_1| + \cdots + |TB_L|$$
$$= \left|\sum_{j=1}^{k_1} T_j\right| + \left|\sum_{j=1}^{k_2} T_j\right| + \cdots + \left|\sum_{j=1}^{k_n} T_j\right|$$
$$= \left|\sum_{i=1}^{n} \sum_{j=1}^{k_i} T_{ij}\right|. \tag{7}$$

Based on Paik et al. [12], a *fraction* and *tree-item (titem)* are defined as the equivalent roles for record and item respectively. A set of fractions is collected from all the blocks. When we assume **F** is a fraction set, it is written $\mathbf{F} = \{F_{j,k}^i \mid F_{j,k}^i \preccurlyeq T_{i,j}\}$, where $1 \le i \le L$, $1 \le j \le |T_i|$ and $1 \le k$. Each one of fractions in the set is eligible to be a titem because the entire structure is actually a collection of many fractions. Actually, fractions are all possible subtrees obtained from the set **S**. From a such reason, the equation (1), support for an item X, is rewritten with the equation (7) to use for a titem X.

$$sup(X) = \frac{|\{T_{k_1} | (X \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| +}{|\mathbf{S}|}$$
$$\frac{|\{T_{k_2} | (X \subseteq T_{k_2}) \wedge (T_{k_2} \in TB_2)| + \cdots}{|\mathbf{S}|}$$
$$\frac{+ \left|\left\{T_{k_L} \middle| (X \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\right\}\right|}{|\mathbf{S}|} \tag{8}$$

According to the above equation, the support expressions for a negated titem X and a negative

association $X \Rightarrow \neg Y$, as well as the confidence equation (6) are suited for titems.

$$sup(X \Rightarrow \neg Y) =$$

$$\frac{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots + |\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}| -}{|\mathbf{S}|}$$

$$\left(\frac{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (Y \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots}{|\mathbf{S}|}\right.$$

$$\left. + \frac{|\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (Y \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}|}{|\mathbf{S}|}\right) \quad (9)$$

$$conf(X \Rightarrow \neg Y) = \frac{sup(X \Rightarrow \neg Y)}{sup(X)}$$

$$= \mathbf{1} - \frac{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (Y \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots + |\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (Y \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}|}{|\{T_{k_1}|(X \subseteq T_{k_1}) \wedge (T_{k_1} \in TB_1)\}| + \cdots + |\{T_{k_L}|(X \subseteq T_{k_L}) \wedge (T_{k_L} \in TB_L)\}|}$$

(10)

The titem X is called frequent if and only if the value of $sup(X)$ is equal to or greater than a user specified minimum support (*ms*). Otherwise, X is infrequent. For the positive 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. Filtering titems just by applying *ms* may lead to erroneous results in negative associations because *ms*-unsatisfying titem sets can have high values of support, confidence, or both, when they are negative. In addition to support-confidence approach which fundamentally bases the occurring frequency counts, other constraints are required to supplement pruning titems.

For the purpose, we suggest two key points. First one is a concept of *weight*. Most earlier researches conduct the frequency counts over the whole stream data itself. Therefore, the characteristics of deducted titem sets are decided by the stream, even though some titems are not included in some blocks. On the contrary, some fractions are not eligible to be titems because they do not satisfy *ms* in spite of often occurring within some blocks. Due to apply weight, we plan to decide titems from two aspects: stream vs. block.

3.4 Weighted or non-weighted negative titems

The strength and reliability for a rule $X \Rightarrow \neg Y$ is determined by its $sup(X \Rightarrow \neg Y)$ and $conf(X \Rightarrow \neg Y)$ respectively. In this paper are specified both constraints with or without *weight*. Stream data is a sequence of blocks and a block is a set of tree data. Some titems appears often within a block, but some occurs often within the whole stream. The former indicates that those titems are meaningful information only for some blocks, however, the latter informs that the titems are meaningful information for the stream itself. We name the titems considered in the entire stream are *weighted titems* and the titems in a block are *non-weighted titems*. Based on the weight with a given **S**, the support and confidence between titems $X \Rightarrow \neg Y$, equations (9) and (10), are specified as two separate constraints:

1. **weighted support & confidence**, $wsup(X \Rightarrow \neg Y)$ and $wconf(X \Rightarrow \neg Y)$ respectively. They are actually same as the original support and confidence because the equation (9) and (10) computes over the whole stream. From now on, we use *wsup*/*wconf* instead *sup*/*conf*.
2. **non-weighted support & confidence**, $bsup(X \Rightarrow \neg Y)$ and $bconf(X \Rightarrow \neg Y)$ respectively. 'b' indicates 'block' because it is important that how many times the titem appears in a given block. Stream-focused two constraints are modified for block dependency. For a tree block TB_i,

$$bsup(X \Rightarrow \neg Y, TB_i) = \frac{\left|\left\{\left\{T_{k_i}\right|(X \subseteq T_{k_i})\right\}\right| - \left|\left\{\left\{T_{k_i}\right|(X \subseteq T_{k_i}) \wedge (Y \subseteq T_{k_i})\right\}\right|}{|TB_i|} \quad (11)$$

$$bconf(X \Rightarrow \neg Y, TB_i) = 1 - \frac{\left|\left\{\left\{T_{k_i}\right|(X \subseteq T_{k_i}) \wedge (Y \subseteq T_{k_i})\right\}\right|}{\left|\left\{\left\{T_{k_i}\right|(X \subseteq T_{k_i})\right\}\right|} \quad (12)$$

Generally, a rule discovery is to find the form $X \Rightarrow \neg Y$ by applying the equations (9) and (10) and comparing to given thresholds *ms* and *mc* respectively, which is the way for weighted titems. In the proposed approach we consider non-weighted titems for block dependency. Given block TB_i, any rule $X \Rightarrow \neg Y$ is virtue to unreveal from the block if and only if its *bsup* and *bconf* are equal to or greater than the given block thresholds bms_i and bmc_i, respectively. Let us consider streaming tree data on Figure 3. We assume the entire stream data **S** consists of two blocks, **S** = {(t_1,TB_1), (t_2,TB_2)}, and identically each block size is 4. The size of **S**, |**S**|, is 8. Figure 4 shows any two fractions selected from the set **F**. We compute its constraints with respect to **S**, TB_1, and TB_2.

The following is the computational results obtained by (4).

$$wsup(F_X) = \frac{2}{8} = 0.25$$

$$wsup(F_Y) = 1 - \frac{3}{8} = 0.625$$

$$bsup(F_X, TB_1) = \frac{2}{4} = 0.5$$

$$bsup(F_X, TB_2) = \frac{0}{4} = 0$$

$$bsup(F_Y,\ TB_1) = 1 - \frac{2}{4} = 0.5$$

$$bsup(F_Y,\ TB_2) = 1 - \frac{1}{4} = 0.75$$

When we set each threshold *ms*=0.3 and *bms*=0.3, the fraction F_X is not eligible to be a weighted titems, but is a non-weighted titem in both blocks. On the other hand, F_Y is enough to be a weighted titem but is only non-weighted titem in the block TB_2. By using weight, the selection of titems is more concentrated to the applicability. All constrains presented in the paper can be applied to both weighted or non-weighted way together. For simplicity, we explain for weighted titems.

4 DOSCOVERY FOR NEGATIVE TITEMS WITH FOUR CONSTRAINTS

PRUNING must be done with care in negative associations. Two statistical methods, support and confidence, of course, prune many unnecessary titems quietly well, but they have the nature of the problem that is they 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. Weight explained in the subsection 3.4 helps to choose the range of pruning to select proper titems, however, it also depends the frequency counts.

Many studies have been conducted to complement the weak point but, there is no widespread agreement. Instead, they can be grouped into two types: interestingness vs. correlation. Interestingness plays an important role in data mining. So far there is no universally accepted formal definition. Nevertheless, interestingness is intended for the patterns in ranking and selecting which is explained by Geng and Hamilton (2006).

The second type is Correlation coefficient value that describes 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. For reliable and trustworthy pruning, the interestingness used by Wu, et al. (2004) and the correlation coefficient measure by Antonie and Zaïane (2004) are adjusted for applying to the target titems.

4.1 Interestingness vs. Correlation-coefficient values for titems

When a preposition expressed by Piatetsky-Shapiro (1991) is applied to a possible positive association rule $X \Rightarrow Y$, $sup(X \Rightarrow Y) \approx sup(X) \times sup(Y)$, the rule is not interesting if its itemset X and itemset Y are independent. Based on the proposition the function *interest* with a threshold *mi*, minimum interest, was defined by Wu, et al. (2004). It computes a numerical value of a potential rule interest. If the produced value is less than *mi*, the input itemsets do not provide interesting information. Using the idea, the tailored function interest covers titems:

$$interest(X \Rightarrow Y)$$

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

The equation (13) cannot be used directly for a possible negative association rule $X \Rightarrow \neg Y$ because of the counting difficulty for $\neg Y$. Instead, it is derived using Y. The modified expression is the equation (14).

$$interest(X \Rightarrow \neg Y)$$

$$= |wsup(X \Rightarrow \neg Y) - wsup(X) \cdot wsup(\neg Y)|$$

$$= |\mathrm{w}sup(X) - \mathrm{w}sup(X \Rightarrow Y) - \mathrm{w}sup(X) \cdot \mathrm{w}sup(\neg Y)|$$

$$= |\mathrm{w}sup(X) \cdot \mathrm{w}sup(Y) - \mathrm{w}sup(X \Rightarrow Y)| \quad (14)$$

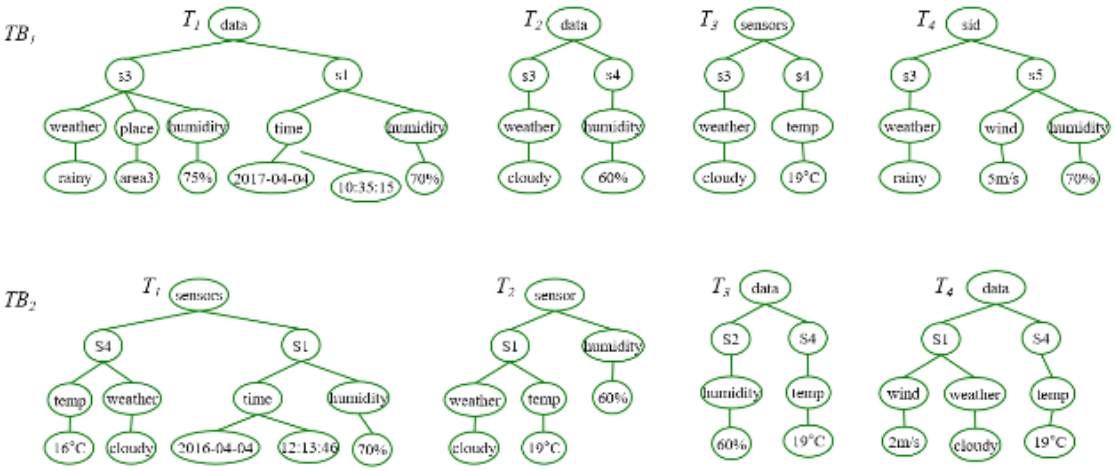

Figure 3. **Stream data with 2 blocks.**

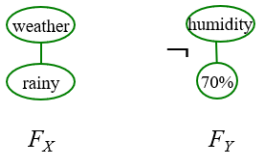

Figure 4. Two fractions in the set **F**.

The rule has rarely interesting information if *interest*($X \Rightarrow \neg Y$) ≈ 0. However, it is worth to discover if the value is greater than or equal to *mi*, nevertheless the support and confidence values are low. The function *interest* is used for non-weighted titems by *bsup* instead of w*sup*. In this case, the value is limited to a certain block.

There is another measurement to prune uninteresting items, that is Correlation-coefficient value. Explained by Cohrn (1988) applying it, the statistical relationship between two variables is measured. Originally it is the degree of linear dependency between random variables X and Y, known as the covariance of the two variables, ρ_{XY}. The range of ρ_{XY} is from -1 to +1. If $\rho_{XY} > 0$, those two variables are positively correlated. On the contrary, they are negatively correlated each other, if $\rho_{XY} < 0$. There is a strong correlation between X and Y if ρ_{XY} is close to either -1 or $+1$. But, if $\rho_{XY} = 0$, X and Y are independent each other. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, the value of one increase as the value of the other decreases.

By Karl Pearson ϕ coefficient was introduced. The statistical association for two binary values, 1 or 0, is measured by using ϕ coefficient. Usually, it is easily applied to identify existence (1) or non-existence (0) of any itemset in transactions. When we assume simply X and Y are two binary variables, the associated relationships between them are summarized in 4 cases which are X=Y=1, X=Y=0, X=1∧Y=0, and X=0∧Y=1. According to each case count number (n_{11}, n_{00}, n_{10}, n_{01}), the association is evaluated as

$$\emptyset_{XY} = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}}$$

The above ϕ_{XY} is composed of only those terms their binary values are just 1 because meaning of 1 is 'existence'. It is the following equation (15).

$$\emptyset_{XY} = \frac{nn_{11} - n_{1+}n_{+1}}{\sqrt{n_{1+}(n-n_{1+})n_{+1}(n-n_{+1})}} \qquad (15)$$

Equation (15) is used and modified for titems X, Y, and $\neg Y$ by specifying contingency values in Table 2. Each cell represents possible combinations of titems X and Y with frequency counts with respect to the size of the whole data set. Using the table, equation (15) is re-written in the following;

$$\emptyset_{XY} = \frac{sup(X \cup Y) - sup(X) \cdot sup(Y)}{\sqrt{sup(X) \cdot (1 - sup(X)) \cdot sup(Y) \cdot (1 - sup(Y))}} \qquad (16)$$

Table 2. Contingency table 2×2 for titems X and Y

	Y	¬Y	sum
X	*wsup*($X \Rightarrow Y$)	*wsup*($X \Rightarrow \neg Y$)	*wsup*(X)
¬X	*wsup*($\neg X \Rightarrow Y$)	*wsup*($\neg X \Rightarrow \neg Y$)	*wsup*($\neg X$)
sum	*wsup*(Y)	*wsup*($\neg Y$)	1

Hopkins (2000) described details for the strength of correlation coefficient in his articles. The author thought about carefully only positive values. Regarding his articles the statistical level of ϕ is redefined for the aim of this paper. Those are 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 titems 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*.

4.2 *Simple example*

With respect to the simple dataset **S** on Figure 3, 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 derived from **S**, which has enormous numbers of fraction due to the subtree calculation. In order to show the differentiation between for weighted titems and for non-weighted titems, two types of computation are provided. Each type has the same computational steps; (1) Over the given fractions, support constraint is applied at first to ensure that one of fractions is eligible to be a titem. (2) With the titems, some candidate association rules are generated. (3) To verify the reliability of a rule, confidence constraint is applied. (4) The constraint interestingness is applied to find out missing negative rules. (5) Finally, correlation-coefficient factor is used to discover the negatives that are less interesting but have strong correlation between titems. We choose the following five fractions and assume the values in Table 3 are given thresholds.

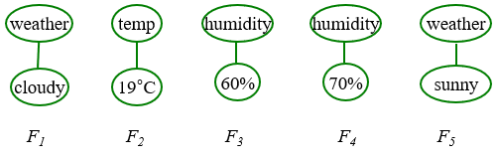

Figure 5. Five sampling fractions included in the set **F**.

Table 3. **Thresholds for 4 constraints**

Threshold	Weight	Non-weight
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

Case 1. Weighted negative titems

$$wsup(F_1) = \frac{5}{8} = 0.65\ ,\quad wsup(F_2) = \frac{4}{8} = 0.5,$$

$$wsup(F_3) = \frac{2}{8} = 0.25\ ,\ wsup(F_4) = \frac{3}{8} = 0.375,$$

$$wsup(F_5) = \frac{2}{8} = 0.25$$

According to *ms*, the fractions F_1, F_2, F_4 are eligible to be titems. The other two would be pruned if it is for positive association rules. Based on equations (4) and (8), the following is computed.

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

Without pruning those two important (their negated support values are the highest) fractions total five frequent titems, $\{F_1, F_2, F_3, F_4, F_5\}$ are obtained. Figure 6 presents 4 possible candidate association rules generated from the frequent titems set. For the each rule its support and confidence values together are computed to constraint its strength and reliability. Equations (5), (9), (6), (10) are used.

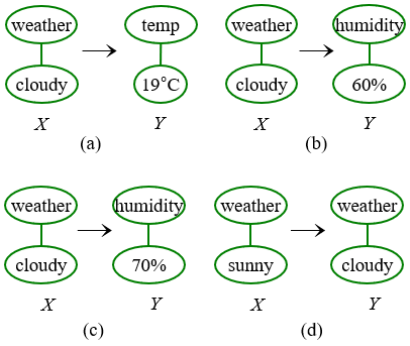

Figure 6. **4 possible candidate association rules.**

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

From the point of positive association rules, the rule (a) is the only proper candidate rule which can be a positive association rule. When we apply negative titem $\neg Y$ instead Y, the result is completely changed. Especially, the rule (c) and (d) provides useful and important knowledge such that is not came out in positive approaches. According to the negative association rule (d) we can informed that if weather is sunny, it is hardly a cloudy day.

$$\text{(b) } wsup(X \Rightarrow \neg Y) = \frac{3}{8} = 0.375$$
$$wconf(X \Rightarrow \neg Y) = \frac{3}{5} = 0.6,$$
$$\text{(c) } wsup(X \Rightarrow \neg Y) = \frac{4}{8} = 0.5$$
$$wconf(X \Rightarrow \neg Y) = \frac{4}{5} = 0.8,$$
$$\text{(d) } wsup(X \Rightarrow \neg Y) = \frac{2}{8} = 0.125$$
$$wconf(X \Rightarrow \neg Y) = 1.$$

With a negative titems set the obtained association rules support excellent strength and reliability. However, there is no such an algorithm that can directly determine the conjunction of presence and absence of titems. The most difficult to fulfill is to evaluate not only all fractions but also all negated fractions. It is a challenge to identify such fractions can be potentially valuable titemsets no matter frequency is high or not, that is the aim of this work. We take the interestingness and correlation coefficient for the purpose.

With equations (13) and (14), each value of interestingness for original possible positive associations and their negatives are computed as the following;

$$\text{(a) } interest(X \Rightarrow Y) = \left|\frac{3}{8} - \frac{5}{8} \cdot \frac{4}{8}\right| = 0.062$$
$$interest(X \Rightarrow \neg Y) = \left|\frac{5}{8} \cdot \frac{5}{8} - \frac{3}{8}\right| \approx 0.14,$$
$$\text{(b) } interest(X \Rightarrow Y) = \left|\frac{2}{8} - \frac{5}{8} \cdot \frac{3}{8}\right| = 0.0156$$
$$interest(X \Rightarrow \neg Y) = \left|\frac{5}{8} \cdot \frac{2}{8} - \frac{2}{8}\right| \approx 0.09\ ,$$
$$\text{(c) } interest(X \Rightarrow Y) = \left|\frac{1}{8} - \frac{5}{8} \cdot \frac{3}{8}\right| \approx 0.109$$
$$interest(X \Rightarrow \neg Y) = \left|\frac{5}{8} \cdot \frac{5}{8} - \frac{1}{8}\right| \approx 0.109\ ,$$
$$\text{(d) } interest(X \Rightarrow Y) = \left|0 - \frac{2}{8} \cdot \frac{5}{8}\right| \approx 0.84$$
$$interest(X \Rightarrow \neg Y) = \left|\frac{2}{8} \cdot \frac{5}{8} - 0\right| \approx 0.16\ ,$$

Based on the above results we can roughly decide which type of association rules is proper to a purpose. In the example, the rule (a) and (b) have more interesting if they are provided as negative associations, while (d) is for the positives. The rule (c) cannot be decided that it is worth to discover when interestingness is applied to it. Such decisions cannot

be made just by using the support and confidence methods. There are certain rules which interest values are quite high in spite that their support and confidence values are not sufficient. Using interestingness is beneficial for finding negative association rules, but there is a problem that is it depends on how appropriately give *mi* to find satisfying titems. Therefore, we lastly determine how strongly two titems are related each other according to the equation (16). Since correlation coefficient expression applies the non-existence of titems in nature, we do not need to compute separately.

$$\text{(a) } \emptyset_{XY} = \frac{1\cdot\frac{3}{8}-\frac{5}{8}\cdot\frac{4}{8}}{\sqrt{\frac{5}{8}\cdot\left(1-\frac{5}{8}\right)\cdot\frac{4}{8}\cdot\left(1-\frac{4}{8}\right)}} = \frac{1}{\sqrt{15}} \approx 0.258$$

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

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

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

With respect to the strength of correlation coefficient explained in the previous, we decide that the candidate rule (b) is not valuable to be mined due to its correlation coefficient value, which is less than +0.1. The statistics means that those two titems configuring the rule exist almost independently each other. Therefore, the associated relationship between them is rarely made. In other measures, it has been determined as a frequently occurred but less reliable rule by the support/confidence and it has been determined as a not much interesting rule in both positive and negative. Clearly, correlation-coefficient factor determines that their correlation is almost independent, therefore, the rule is very rare. However, the titems used in the rules (c) and (d) have strong negative association between them, which implies that the negative titems are mostly appear together with their positive titems even though their support values are less than the given *ms*.

Case 2. Non-Weighted negative titems

It is omitted because the process is identical to Case 1, except it is applied to some blocks not the whole stream.

4.3 Advantages

Applying two more constraints, interestingness and correlation coefficient values, is helpful and useful in finding association rules in when 1) there is any negative relationship between titem sets, 2) the statistical relationship which is correlation coefficient value's strength is reasonably strong enough to be useful information, and 3) although the support value is less than a given *ms*, thus it is not counted in the positive association rules, it can be valuable negative association rule providing many predictive insights for further mining process.

By using interestingness and correlation coefficient values together, hidden association rules present benefits especially when it is mined for negative rules, such are not revealed by support/confidence or even interestingness alone.

In addition, the concept weight specifies the range of obtaining titem, such provides to choose and decide the most appropriate dataset for negative titems.

The following algorithm broadly outlines the procedure explained in previous pages. It determines the way how to apply four measuring factors and uncover informative negative tree-structured items along with the weight usage.

INP: **S** OUTP: W*NT or N*- W*NT*

1. IF (*weight*)
2. FOR EACH block $TB_i \in \mathbf{S}$ $(1 \le i \le k)$
3. FOR $j \leftarrow 1$ to n
4. IF $freq(X_{ij}, \mathbf{S}) \ge |\mathbf{S}| \times \delta$
5. THEN $FT = FT + \{ X_{ij} \}$;
6. FOR titemset $X \subset FT$, $Y \subset FT$, $X \cap Y = 0$
7. IF $sup\ (X \Rightarrow Y) < ms$
8. or $conf\ (X \Rightarrow Y) < mc$
9. THEN
10. IF $interest(X, \neg Y) < mi$
11. THEN
12. IF $\phi_{(X, \neg Y)} \le -0.3$ or $\phi_{(X, \neg Y)} \ge +0.3$
13. THEN W*NT* ← W*NT* + $\{X \Rightarrow \neg Y\}$;
14. ELSE THEN
15. W*NT* ← W*NT* + $\{X \Rightarrow \neg Y\}$;
16. ELSE THEN
17. W*NT* ← W*NT* + $\{X \Rightarrow \neg Y\}$;
18. ELSE THEN
19. FOR SOME block $TB_i \in \mathbf{S}$ $(1 \le i \le k)$
20. FOR $j \leftarrow 1$ to n
21. IF $freq(X_{ij}, TB_i) \ge |TB_i| \times \delta$
22. THEN $FT = FT + \{ X_{ij} \}$;
23. FOR titemset $X \subset FT$, $Y \subset FT$, $X \cap Y = 0$
24. IF $bsup\ (X \Rightarrow Y) < ms$
25. or $bconf\ (X \Rightarrow Y) < mc$
26. THEN
27. IF $binterest(X, \neg Y) < mi$
28. THEN
29. IF $b\phi_{(X, \neg Y)} \le -0.3$ or $b\phi_{(X, \neg Y)} \ge +0.3$
30. THEN *N*-W*NT* ← *N*-W*NT*
31. + $\{X \Rightarrow \neg Y\}$;
32. ELSE THEN
33. *N*-W*NT* ← *N*-W*NT*
34. + $\{X \Rightarrow \neg Y\}$;
35. ELSE THEN
36. *N*-W*NT* ← *N*-W*NT* + $\{X \Rightarrow \neg Y\}$;
37. RERURN *NTS or N-WNT*
38. END

5 CONCLUSIONS

IN this work, we considered how to efficiently obtain negative tree-structured items for association rules from streaming tree format data. For the purpose, the primarily considered part was to verify fractions from the stream 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 titemsets. Besides, it was supported by weight choosing the range of fractions, by which the characteristics of titems were decided and the usage of them could be more specific.

The example results of each constraint with weight 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. For the first time the analyses of both interestingness and correlation coefficient methods with weight or non-weight have been suggested over tree-structured stream data. Future work includes presenting a full mining algorithm and experimental results of negative association rules for it, that is proven to work with the four constraints as well as the influence of the weight.

6 ACKNOWLEDGMENT

THIS research was supported by Basic the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1A2B1007015) Science Research Program through.

7 REFERENCES

R. Agrawal, T. Imielinski, and A. N. Swami, (1993). Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 207-216.

M. L. Antonie and O. R. Zaïane, (2004). 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*, *LNCS 3202*, 27-38.

K. Ashton, (2009) That `Internet of Things' thing. In the real world, things matter more than ideas, *RFID Journal*. 22 Jun. Available: http://www.rfidjournal.com/articles/view?4986.

B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, (2002). Models and issues in data stream systems, *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 1-16.

A. Boukerche and S. Samarah, (2008). A novel algorithm for mining association rules in wireless ad hoc sensor networks, *IEEE Transactions on Parallel and Distributed Systems*, 19, 865-877.

D. Braga, A. Campi, S. Ceri, M. Klemettinen, and P. L. Lanzi, (2002). A tool for extracting xml association rules, *Proceedings of the 14th IEEE International Conference on Tools with artificial Intelligence*, 57-64.

J. Cohrn, (1988). Statistical power analysis for the behavioral sciences, *Lawrence Erlbaum*, 109-143.

S. Corpinar and T. Í. Gündem, (2012). Positive and negative association rule mining on xml data streams in database as a service concept, *Expert Systems with Applications*, 39(8), 7503-7511.

L. Feng and T. Dillon, (2004). Mining interesting xml-enabled association rules with templates, *Proceedings of the 3rd International Workshop on Knowledge Discovery and Inductive Databases*, *LNCS 3377*, 66-88.

L. Geng and H. J. Hamilton, (20060. Interestingness measures for data mining: a survey, *ACM Computing Survey*, 38(1), article no. 9.

J. Han and Y. Fu, (1995). Discovery of multiple-level association rules from large databases. *Proceedings of the 21st International Conference on Very Large Data Bases*, 420-431.

J. Han, J. Pei, and Y. Yin, (2000). Mining frequent patterns without candidate generation, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 1-12.

Z. Honglei and X. Zhigang, (2008). An effective algorithm for mining positive and negative association rules, *Proceedings of International Conference on Computer Science and Software Engineering*, 455-458.

W. Hopkins, (2000). A new view of statistics, *Electronic edition, Available: http://www.sportsci.org/resource/stats/*.

K.K. Loo, I. Tong, B. Kao, and D. Chenung, (2005). 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, LNCS 3518*, 143-149.

J. Manyika and M. Chui, (2015). By 2025, Internet of things applications could have $11 trillion impact, *Available:http://www.mckinsey.com/insights/mgi/in the news/by 2025 internet of things applications could have 11 trillion impact*.

S. Mahmood, M.Shahbaz, and A. Guergachi, (2014). Negative and Positive Association Rule Mining from Text Using Frequent and Infrequent Itemsets. *The Scientific World Journal*, ID 973750.

J. Paik, J. Nam, U. Kim, and D. Won, (2014). Association rule extraction from xml stream data for wireless sensor networks, *Sensors*, 14, 12937-12957.

J. Paik, J. Nam, S. Lee, and U. Kim, (2007). A framework for data structure-guided extraction of xml association rules, *Proceedings of the 7th International Conference on Computational Science*, *LNCS 4489*, 709-716.

G. Piatetsky-Shapiro, (1991). Discovery, analysis, and presentation of strong rules, *Knowledge Discovery in Databases*, *AAAI Press*, 229-248.

M. M. Rashid, I. Gondal, and J. Kamruzzaman, (2014). Mining associated patterns from wireless sensor networks, *IEEE Transactions on Computers*, 64, 1998-2011.

A. Savasere, E. Omiecinski, and S. Navathe, (1998). Mining for strong negative associations in a large database of customer transactions,. *Proceedings of the 14th International Conference on Data Engineering*, 494-502.

R. Sumalatha and B. Ramasubbareddy, (2010). Mining positive and negative association rules, *International Journal on Computer Science and Engineering*, 2(09), 2916-2910.

R. Wolff and A. Schuster, (2004). Association rule mining in peer-to-peer systems, *IEEE Transactions on Systems, Man, and Cybernetics*, *Part B: Cybernetics*, 34, 2426-2438.

X. Wu, C. Zhang, and S. Zhang, (2004). Efficient mining of both positive and negative association rules, *ACM Transaction on Information Systems*, 22(03), 381-405.

X. Yuan, B. P. Buckles, Z. Yuan, and J. Zhang, (2002). Mining negative association rules, *Proceedings of the 7th International Symposium on Computers and Communications*, 623-628.

8 NOTES ON CONTRIBUTORS

J. Paik achieved B.E. degree in Information Engineering from Sungkyunkwan University, Korea, in 1997. She worked at the Samsung SDS Company for about one year, after she graduated the undergraduate. She achieved her degrees of M.E. and Ph.D. in Computer Engineering from the same university, Sungkyunkwan University in 2005 and 2008, respectively. She was a Post-Doc and Research Professor at the Department of Computer Engineering, Sungkyunkwan University from March 2008 to February 2016. During that time, she was devoted to the research of xml data mining and tree databases. From the year 2016, she is an Assistant Professor at the Department of Digital Information & Statistics in Pyeongtaek University, South Kore. Currently, her research interests include tree mining, big data mining, semantic mining, information retrieval, and web search engines.