Novel Algorithm For Association Rule Mining Based on Apriori Algorithm

*Author: Rishabh Johri*

*School of Computer Science and Engineering, Vellore Institute of Technology (VIT University) , Vellore , Tamil Nadu*

*Email :* [*rishabh.johri2019@vitstudent.ac.in*](mailto:rishabh.johri2019@vitstudent.ac.in)

*Abstract:*

*Apriori algorithm is the most basic yet widely used frequent pattern mining algorithm. Apriori algorithm uses two interestingness measures : Support , Confidence. My endeavor is to improve and modify this algorithm by incorporating some novel interestingness measures that are based on profit.*

*The following defects of the association rules mining methods exist: (1) Many traditional association methods generate a lot of rules, and most of them are not relevant or even rules of error. (2) Do not consider users characteristics and their changes, but users characteristics and subjectivity tend to affect relevance of several events. (3) Online trading data and user evaluation data are extremely sparse (i.e., data sparseness) due to surge of current online trading. (4) The too low threshold values of support and confidence can produce the combination explosion, but, because of data sparseness, low support rules may provide some novel knowledge that users are interested in. (5) At present some literature simply combines kinds of interestingness evaluation to measure, but this does not take rationality of various interestingness evaluation methods into account.*

*Incorporating new interestingness measures to apriori algorithm will solve these disadvantages.*

***INTRODUCTION:***

With the progress of the technology of information and the need for extracting useful information

of business people from dataset [7], data mining and its techniques is appeared to achieve the

above goal. Data mining is the essential process of discovering hidden and interesting patterns

from massive amount of data where data is stored in data warehouse, OLAP (on line analytical

process), databases and other repositories of information [11]. This data may reach to more than

terabytes. Data mining is also called (KDD) knowledge discovery in databases [3], and it includes

an integration of techniques from many disciplines such as statistics, neural networks, database

technology, machine learning and information retrieval, etc [6]. Interesting patterns are extracted

at reasonable time by KDD’s techniques [2]. KDD process has several steps, which are performed

to extract patterns to user, such as data cleaning, data selection, data transformation, data pre-

processing, data mining and pattern evaluation [4].

The architecture of data mining system has the following main components [6]: data warehouse,

database or other repositories of information, a server that fetches the relevant data from

repositories based on the user’s request, knowledge base is used as guide of search according to

defined constraint, data mining engine include set of essential modules, such as characterization,

classification, clustering, association, regression and analysis of evolution. Pattern evaluation

module that interacts with the modules of data mining to strive towards interested patterns.

Finally, graphical user interfaces from through it the user can communicate with the data mining

system and allow the user to interact.

With the progress of the technology of information and the need for extracting useful information of business people from dataset , data mining and its techniques is appeared to achieve the above goal. Data mining is the essential process of discovering hidden and interesting patterns from massive amount of data where data is stored in data warehouse, OLAP (on line analytical process), databases and other repositories of information . This data may reach to more than terabytes. Data mining is also called (KDD) knowledge discovery in databases [3], and it includes an integration of techniques from many disciplines such as statistics, neural networks, database technology, machine learning and information retrieval, etc . Interesting patterns are extracted at reasonable time by KDD’s techniques . KDD process has several steps, which are performed to extract patterns to user, such as data cleaning, data selection, data transformation, data pre-processing, data mining and pattern evaluation . The architecture of data mining system has the following main components : data warehouse, database or other repositories of information, a server that fetches the relevant data from repositories based on the user’s request, knowledge base is used as guide of search according to defined constraint, data mining engine include set of essential modules, such as characterization, classification, clustering, association, regression and analysis of evolution. Pattern evaluation module that interacts with the modules of data mining to strive towards interested patterns. Finally, graphical user interfaces from through it the user can communicate with the data mining system and allow the user to interact.

Association Mining is one of the most important data mining’s functionalities and it is the most popular technique has been studied by researchers. Extracting association rules is the core of data mining . It is mining for association rules in database of sales transactions between items which is important field of the research in dataset . The benefits of these rules are detecting unknown relationships, producing results which can perform basis for decision making and prediction . The discovery of association rules is divided into two phases : detection the frequent itemsets and generation of association rules. In the first phase, every set of items is called itemset, if they occurred together greater than the minimum support threshold [9], this itemset is called frequent itemset. Finding frequent itemsets is easy but costly so this phase is more important than second phase. In the second phase, it can generate many rules from one itemset as in form, if itemset {I1, I2, I3}, its rules are {I1->I2, I3}, {I2->I1, I3}, {I3->I1, I2}, {I1, I2->I3}, {I1, I3->I1}, {I2, I3->I1}, number of those rules is n2-1 where n = number of items. To validate the rule (e.g. X->Y), where X and Y are items, based on confidence threshold which determine the ratio of the transactions which contain X and Y to the transactions A% which contain X, this means that A% of the transactions which contain X also contain Y. minimum support and confidence is defined by the user which represents constraint of the rules. So the support and confidence thresholds should be applied for all the rules to prune the rules which it values less than thresholds values. The problem that is addressed into association mining is finding the correlation among different items from large set of transactions efficiency [8]. The research of association rules is motivated by more applications such as telecommunication, banking, health care and manufacturing, etc.

***Literature Survey:***

To illustrate conveniently, firstly we suppose that formal

description of association rules is as follows:

𝐴 → 𝐵. (1)

In this description, 𝐴 = {𝐴1, 𝐴2,...,𝐴𝑗}⊂𝐼 and 𝐵 =

{𝐵1, 𝐵2,...,𝐵𝑘}⊂𝐼. 𝐼 indicate itemsets, and 𝐴∩𝐵=𝜙. Rules

should meet certain support threshold 𝑠 and confidence c

Support:

Support means the frequency that the data fields 𝐴 and 𝐵 involved in association rules occur together in the data set. Only the association rules appear frequently in the itemsets, when it gets high accuracy. Support can be used to measure the usefulness of association rules. When the frequency of 𝐴 and 𝐵 occurring at the same time is equal to or greater than the designated minimum support threshold, 𝐴 and 𝐵 meet frequent itemsets. Support can be expressed as

𝑠(𝐴 → 𝐵) = 𝑃 (𝐴𝐵) = 𝑁 (𝐴𝐵) /|𝐷| -------- (2)

where 𝑁(𝐴𝐵) is the record number of 𝐴 and 𝐵 that appeared together, and |𝐷| is the total record number of transactions in data sets.

Support is classic but also has the defects of artificially controlled threshold and rare itemsets. Many infrequent itemsets in the data set may have potential value. Besides, at present in large electronic commerce system, the number of subjects (users) and the amount of projects increase exponentially. Online transaction data and user evaluation data are extremely sparse.

Confidence:

Confidence is the statistics of probability 𝑃(𝐵 | 𝐴) that subsequent events occur under the condition of occurrence of the precursor events in trading data sets. It is used to measure the reliability of the rules. Formula is

𝑐 (𝐴 → 𝐵) = 𝑃 (𝐵|𝐴) = 𝑃 (𝐴𝐵) /𝑃 (𝐴) ---------- (3)

It is used to combine confidence with support to form Support-confidence framework for mining association rules [13]. If Support is larger than the designated minimum support threshold and Confidence is larger than the designated minimum confidence threshold, the rules are called strong association rules. But strong association rules are not always effective, some are not what users are interested in, and some are even misleading.

Lift

Because of the defects of Support-confidence framework, some scholars analyze the relativity of association rules mined, namely, lift [12]. Lift means the ratio of rule’s Confidence to probability of occurrence of the consequent, which reflects positive or negative correlation of antecedent and consequence of rules. It refers to the ratio of the occurrence probability of 𝐵 under the condition 𝐴 to that without considering condition 𝐴, which reflects the relationship between “𝐴” and “𝐵”:

𝑙𝑖𝑓𝑡(𝐴 → 𝐵) = 𝑐 (𝐴→𝐵) 𝑃 (𝐵) = 𝑃 (𝐴𝐵) /(𝑃 (𝐴) 𝑃 (𝐵)) ------- (4)

The range of lift values is [0, +∞). As lift is equal to 1, it shows that 𝐴 and 𝐵 appearing at the same time belong to independent random events and have no special significance; namely, 𝐴 and 𝐵 are independent of each other with no mutual affection. We call this rule uncorrelated rules; if lift value is less than 1, it shows that the emergence of “𝐴” reduces the emergence of “𝐵,” and then we call them negative correlation rules; if Lift value is larger than 1, it shows that the emergence of “𝐴” promotes the emergence of “𝐵,” and then we call them positive correlation rules. Problems: Lift takes events 𝐴 and 𝐵 in equivalence position. According to the Lift, 𝐴→𝐵 and 𝐵→𝐴 are the same; that is to say, if we accept rule 𝐴→𝐵, 𝐵→𝐴 should be also accepted, but the fact is not like this.

APRIORI ALGORITHM:

Apriori Property –

All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. Apriori assumes that all subsets of a frequent itemset must be frequent.

If an itemset is infrequent, all its supersets will be infrequent.

Step-1: K=1

(I) Create a table containing support count of each item present in dataset – Called C1(candidate set)

(II) compare candidate set item’s support count with minimum support count(here min\_support=2 if support\_count of candidate set items is less than min\_support then remove those items). This gives us itemset L1.

Step-2: K=2

Generate candidate set C2 using L1 (this is called join step). Condition of joining Lk-1 and Lk-1 is that it should have (K-2) elements in common.

Check all subsets of an itemset are frequent or not and if not frequent remove that itemset.

Now find support count of these itemsets by searching in dataset.

(II) compare candidate (C2) support count with minimum support count(here min\_support=2 if support\_count of candidate set item is less than min\_support then remove those items) this gives us itemset L2.

Step-3:

Generate candidate set C3 using L2 (join step). Condition of joining Lk-1 and Lk-1 is that it should have (K-2) elements in common. So here, for L2, first element should match.

Check if all subsets of these itemsets are frequent or not and if not, then remove that itemset.

find support count of these remaining itemset by searching in dataset.

(II) Compare candidate (C3) support count with minimum support count

Step-4:

Generate candidate set C4 using L3 (join step). Condition of joining Lk-1 and Lk-1 (K=4) is that, they should have (K-2) elements in common. So here, for L3, first 2 elements (items) should match.

Check all subsets of these itemsets are frequent or not.

We stop here because no frequent itemsets are found further

Step-5:

Generate association rules and filter based on confidence threshold

***Proposed Model:***

Read Data

Using frequent itemset elements, generate association rules.

Result : frequent itemset

For K =1 to n( a very large number){

Generate K-itemsets;

Filter based on support threshold;

If(k-1 set == k item set){

Break;}

}

Read support threshold S

Compute Set of distinct items

Find confidence of each rule using

Conf(A->B) = Sab/Sa \*100

Find lift of each rule using

Lift(A->B) = Sab/(Sa\*Sb) \*100

Select rules with maximum confidence,lift in case 1 and 2 respectively.

Let those filtered rules belong to sets s1 and s2 respectively.

Final resultant set(S) = Union(s1,s2)

***Empirical study***

We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called Apriori property which helps by reducing the search space.

All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. As Support is used as one of the interestingness measure, the appriori property holds true for this algorithm too. All subsets of a frequent itemset must be frequent.If an itemset is infrequent, all its supersets will be infrequent.

Lift is the ratio of Confidence to Expected Confidence. A lift ratio larger than 1.0 implies that the relationship between the antecedent and the consequent is more significant than would be expected if the two sets were independent. The larger the lift ratio, the more significant the association.

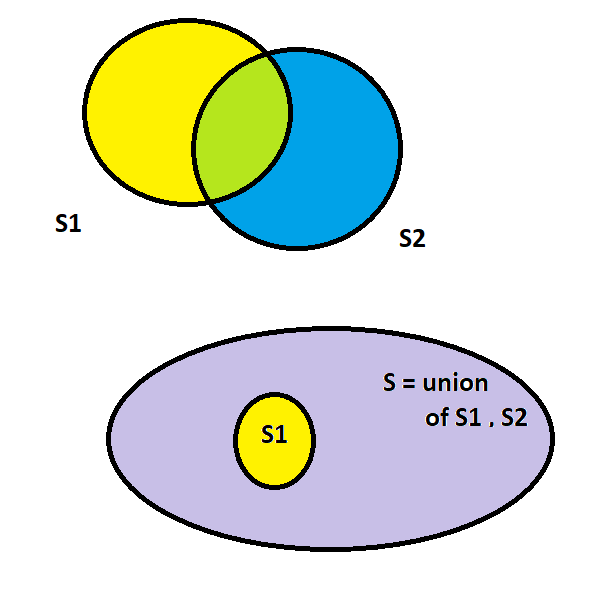
The algorithm uses a combination of confidence and lift as an interestingness measure and instead of a threshold based approach, it uses the maximum values to identify strong association rules that standard appriori algorithm might have missed. Due to the incorporation of the apriori approach through the means of finding union in last step, the valuable rules generated by apriori are not missed.

S1 : Rules generated by apriori algo

S : Rules generated by the proposed model

As S = **S1**  **∪ S2**

Therefore S1 **⊂ S2**

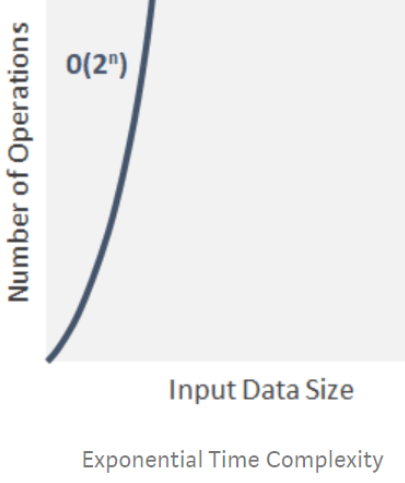


The time complexity and space complexity of generating frequent itemset is O(2D)(in standard apriori) , as that part is incorporated in this model , the complexity remains the same.

The only difference is that instead on computing 1 measure , 2 measures are being computed in the proposed model.

Complexity of apriori = O(2D+ n) = O(2D) { for large input size}

Complexity of proposed = O(2D+2n) = O(2D) { for large input size}

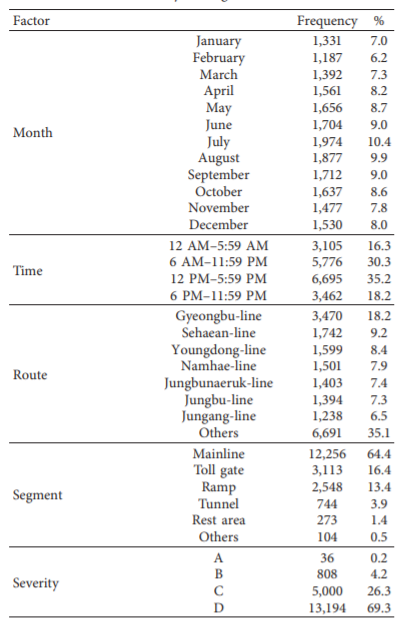


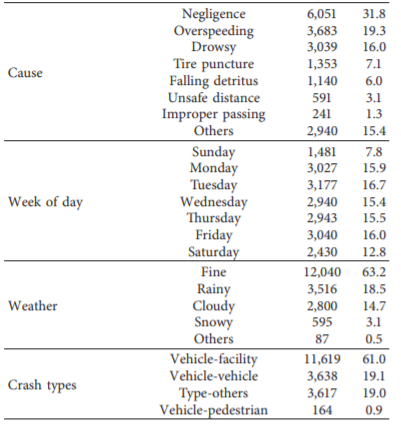
***Experimental Analysis***

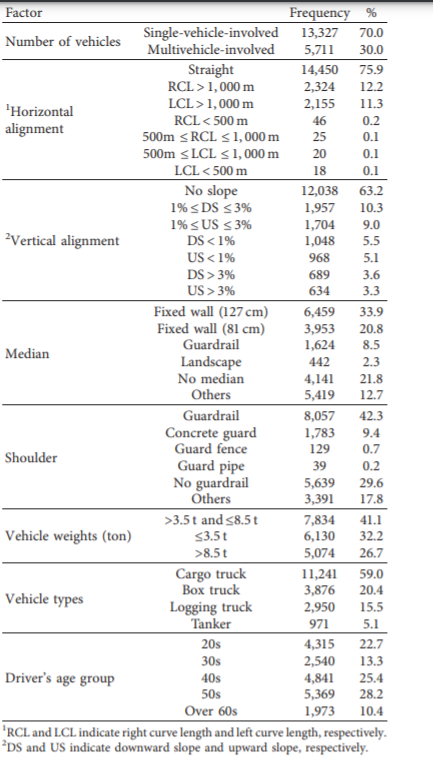
Data set description:

Traffic crash records are collected and stored in the Korean Expressway Corporation (KEC) crash database system and managed by specially trained crash investigators from the KEC.

The crash database used contains 17 explanatory items with 98 subitems. The table shows that truck frequency is lowest on Sundays but is almost similar throughout the other days of the week. Also, truck-involved crashes are persistent during the day (6 AM to 11:59 PM: 30.3%; 12 PM to 5:59 PM: 35.2%). Considering the variable for the months, it shows that the majority of the crashes occurred during summer (June, July, and August).



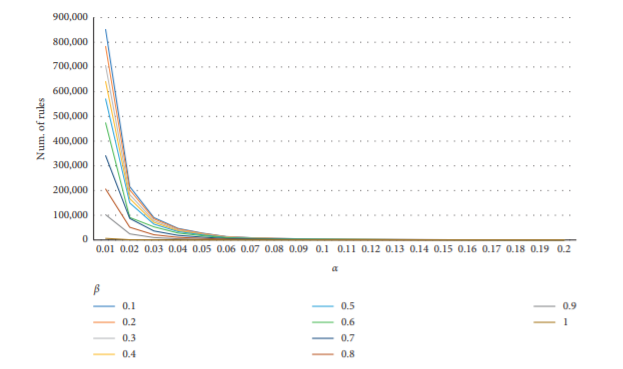


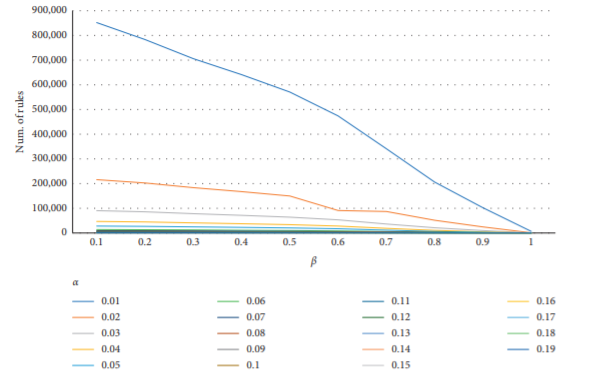


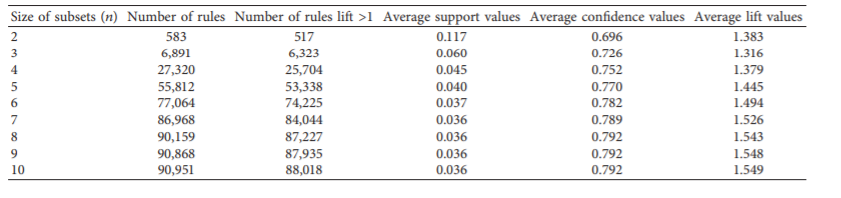
Results:

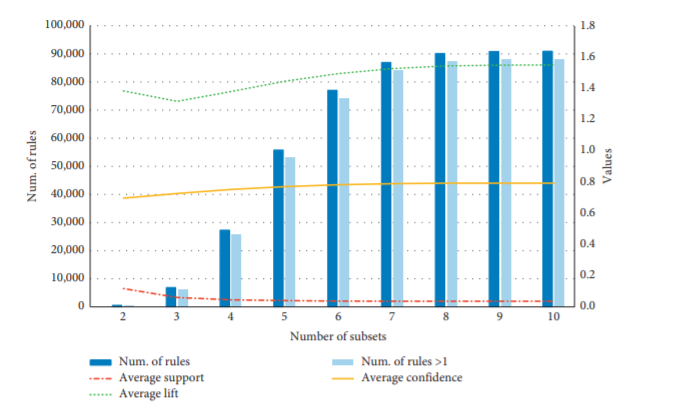
In the real world, many factors contribute to each distinct crash. Hence, it is likely to have several items in either the antecedent or consequent. Overall, the top ten frequent items in the truck-involved crash database are {Median: Guardrail}, {Horizontal Alignment: Straight}, {Number of vehicles involved: Single vehicle-involved}, {Segment: Mainline}, {Weather: Fine}, {Vertical Alignment: No slope}, {Crash type: vehicle-facility}, {Vehicle type: Cargo truck}, {Median = Fixed wall}, and {Cause: Negligence}, in that order.

The results show that driving cargo trucks on a straight and flat mainline section on a fine weather is likely to result in a single-vehicle (SV) crash where the truck runs into a fixed wall or any roadway facility









Conclusion:

The novel approach was able to identify patterns that apriori algorithm might have missed and it is fair to conclude (from a purely application based point of view) that for this particular dataset the novel algorithm works better than the standard algorithm.

***References:***

1. P. N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Pearson—AddisonWesley, Boston, Mass, USA, 1st edition, 2005.

2. J. H. Fowler and M. Laver, “A tournament of party decision rules,” Journal of Conflict

Resolution, vol. 52, no. 1, pp. 68–92, 2008.View at: Publisher Site | Google Scholar

3. R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in Proceedings

of the 20th International Conference on Very Large Data Bases (VLDB '94), pp. 487–499,

1994.View at: Google Scholar

4. R. Srikant and R. Agrawal, “Mining generalized association rules,” in Proceedings of the

21st International Conference on Very Large Data Bases, pp. 407–418, Zurich, Switzerland,

1995.View at: Google Scholar

5. M.-C. Tseng, W.-Y. Lin, and R. Jeng, “Incremental maintenance of generalized association

rules under taxonomy evolution,” Journal of Information Science, vol. 34, no. 2, pp. 174–

195, 2008.View at: Publisher Site | Google Scholar

6. M.-C. Tseng and W.-Y. Lin, “Maintenance of generalized association rules with multiple

minimum supports,” Intelligent Data Analysis, vol. 8, no. 4, pp. 417–436, 2004.

7. M.-C. Chen, “Ranking discovered rules from data mining with multiple criteria by data

envelopment analysis,” Expert Systems with Applications, vol. 33, no. 4, pp. 1110–1116, 2007.

8. M. Toloo, B. Sohrabi, and S. Nalchigar, “A new method for anking discovered rules from data

mining by DEA,” Expert Systems with Applications, vol. 36, no. 4, pp. 8503–8508, 2009.

9. L. Geng and H. J. Hamilton, “Interestingness measures for data mining: a survey,” ACM

Computing Surveys, vol. 38, no. 3, pp.1–32, 2006.

10.N. Hoque, B. Nath, and D. K. Bhattacharyya, “A new approach on rare association rule

mining,” International Journal of Computer Applications, vol. 53, no. 3, pp. 1–6, 2012.

11. J. Zhang, Y. Wang, and J. Feng, “Attribute index and uniform design based multiobjective

association rule mining with evolutionary algorithm,” The Scientific World Journal, vol.

2013,Article ID 259347, 16 pages, 2013.

12. J. Han and M. Kamber, Data Mining Concept and Technology, China Machine Press, Beijing,

China, 2007.

13. G. Piatetsky-Shapiro, “Discovery, analysis, and presentation of strong rules,” in Knowledge

Discovery in Databases, G. Piatetsky Shapiro and W. Frawley, Eds., pp. 229–238, AAAI/MIT

Press,1991.

14. L. Ma andW. Jie, “Research on judgment criterion of association rules,” Control and Decision, vol.

18, no. 3, pp. 277–280, 2003.

[15] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, “Dynamic

itemset counting and implication rules for market basket data,”

in Proceedings of the ACM SIGMOD International Conference

on Management of Data (SIGMOD ’97), pp. 255–264, ACM,

Tucson, Ariz, USA, May 1997.

[16] P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the right objective measure for association

analysis,” Information Systems, vol.29, no. 4, pp. 293–313, 2004.

[17] Y. Li, C. Wu, and K. Wang, “A new interestingness measures for

Ming association rules,” Journal of the China Society for Scientific

and Technical Information, vol. 30, no. 5, pp. 503–507, 2011.

[18] J. Chen and Y. Gao, “Evaluation criterion for association rules

with influence degree,” Computer Engineering and Application,

vol. 45, no. 8, pp. 141–143, 2009.

[19] F. Berzal, I. Blanco, D. Sanchez, and M.-A. Vila, “Measuring the

accuracy and interest of association rules: a new framework,”

Intelligent Data Analysis, vol. 6, no. 3, pp. 221–235, 2002.

[20] D. H. Choi, B. S. Ahn, and S. H. Kim, “Prioritization of

association rules in data mining: multiple criteria decision

approach,” Expert Systems with Applications, vol. 29, no. 4, pp.

867–878, 2005