Journal of Engineering and Applied Sciences 12 (22): 5915-5920, 2017
ISSN: 1816-949X

Mining Both Positive and Negative Association Rules without Extra Database Scans

Ujwala Manoj Patil and J.B. Patil
Department of Computer Engineering, R.C. Patel Institute of Technology,
North Maharashtra University, Jalgaon, Shirpur District Dhule, Maharashtra, India

Abstract: Data mining is getting increasing acceptance in science and business areas that need to identify and represent certain dependencies between attributes. This dependency between the attributes is represented in the form of association rules. Association rule mining discovers interesting correlations between attributes in a database. All the traditional association rule mining algorithms were developed to find positive associations between attributes, i.e., $A \rightarrow B$ whereas negative association rule is an implication of the form $A \rightarrow \neg B$, $\neg A \rightarrow B$, $\neg A \rightarrow \neg B$ where A and B are database attributes, $\neg A \neg B$ are negations of database attributes. Here, we propose an apriori based algorithm to find the both positive and negative associations between attributes. Experimental results show the effectiveness and efficiency of the proposed algorithm without additional database scans.

Key words: Data mining, association rule mining, positive association rules, negative association rules, proposed, effectiveness

INTRODUCTION

Data mining is used for extraction of knowledge from large database. Data mining is broadly classified in the areas such as association rules, classifications and clustering (Agrawal and Srikant, 1995; Sujatha and Punithavalli, 2012), out of these efficient discovery of association rules has been a major focus in the data mining research. Association Rule Mining (ARM) discovers relationships from the huge amount of data by generating association rules. Association rule mining is useful in many application domains like recommender system, decision support, health care, intrusion detection, etc. (Sujathaand Punithavalli, 2012; Srivastava *et al*., 2000; Ramaraj and Venkatesan, 2008). Association rule mining was introduced by Agrawal and Srikant (1995) in terms of the apriori algorithm. After that there have been a remarkable number of variants and improvements of association rule mining algorithms (Kiran and Re, 2009; Hong *et al*., 2001; Hong and Lee, 2008; Brin *et al*., 1997; Srikant and Agrawal, 1996; Matthews *et al*., 2013; Cooley *et al*., 1997; Patil and Patil, 2016). Traditional association rule mining algorithms have been developed to find associations between items. The associations are of two types, called positive associations and negative associations. The traditional association is called positive associations which consider the presence of the item, i.e., $A \rightarrow B$ while another is negative that negates presence of the item, i.e., $\neg A \rightarrow \neg B$, $A \rightarrow \neg B$, $\neg A \rightarrow B$. Positive association rules are useful in decision making, likewise negative association rules also play important role in decision making. Mining of positive and negative rules is very expensive as it has to explore large search space. Till date very few algorithms in the literature have been proposed which use various interestingness measures to find positive as well as negative association rules.

Contribution of this study: The main contribution of this work as follows:

- We survey the current literature to discover the positive and negative association rules
- We have discussed the advantages and limitations of the existing techniques
- We have proposed an algorithm which mines both positive and negative association rules without extra database scans

Literature review: We have surveyed the literature to find what interestingness measures are used by various algorithms and how these interestingness measures are used to find Positive Association Rules (PAR) and Negative Association Rules (NAR) (Brin *et al*., 1997; Aggarwal and Han, 2014; Wu *et al*., 2004; Yang and Zhao,

Corresponding Author: Ujwala Manoj Patil, Department of Computer Engineering, R.C. Patel Institute of Technology, North Maharashtra University, Jalgaon, Shirpur District Dhule, Maharashtra, India

2009; Antonie and Zaiane, 2004; Zaiane, 2007; Tan *et al*., 2002; Ramasubbareddy *et al*., 2011). Interestingness measures used for negative association rules are computed from the relative information about positive association rules.

The algorithm proposed by Wu *et al*. (2004) extend the basic (Srikant and Agrawal, 1996) apriori algorithm. Along with support-confidence, they used Piatetsky-Shapiro's (PS) interest. The algorithm is decomposed into two steps. In first step, generate all frequent and infrequent large itemsets. Itemsets which satisfy user-specified minimum support and minimum interest are declared as frequent itemsets of potential interest, i.e., positive itemsets. Itemsets which do not satisfy user-specified minimum support and minimum interest are declared as infrequent itemsets of potential interest, i.e., negative itemsets. At the end of step one, they declare all positive large itemsets and negative large itemsets. In second step, generate all possible positive and negative association rules. They used synthetic classification datasets with approximately 100000 transactions. The good side of (Wu *et al*., 2004) algorithm is that it mines both PAR and NAR efficiently but they do not discuss how to set it and variations of Piatetsky-Shapiro's (PS) interest in the result (Wu *et al*., 2004).

Yang and Zhao (2009) developed an algorithm to discover PAR and NAR based on apriori. The algorithm generates all frequent large itemsets and then generate all possible PAR and NAR from all large itemsets. Before generating PAR and NAR, it computes the correlation between itemsets. The correlation is computed with the help of Piatetsky-Shapiro's (PS) interest. For experiment they have used synthetic data with 200 transactions. The algorithm is simple and finds PAR and NAR fast as compared to Wu *et al*. (2004) algorithm because it uses only three interestingness measures, i.e., support, confidence and correlation. But this algorithm is not effective as it does not find all possible negative association rules (Yang and Zhao, 2009).

Luiza *et al*. (2014) proposed an approach for PAR and NAR by extending support-confidence approach with a correlation coefficient. In contrast to above two algorithms, it follows one step approach. The algorithm starts with large one itemsets. It gradually combines large L_{k-1} itemsets to find large two itemsets, large three itemsets, large four itemsets and so on. After combining L_{k-1} itemsets it immediately finds the support of the itemsets. The itemsets which satisfy user-specified minimum support are included in the large itemsets list. After adding it calculates the correlation coefficient for that large itemsets. Depending on the outcome of correlation coefficient for an itemset, it generates either positive or negative rule. Reuters-21578 text collection real dataset with 6488 transaction used for experimental setup. Along with simplicity in the algorithm, it automatically adjusts the value of correlation coefficient if no rule is found by the user specified minimum correlation coefficient. But again, it does not explore the search space to find all NAR (Zaiane, 2007).

Ramasubbareddy *et al*. (2011) proposed an algorithm called MPNAR. MPNAR mines, both positive and negative association rules with the support, confidence and Yule's coefficient. Like Maria-Luiza Antonie the algorithm finds all PAR and NAR in one step. They have used Synthetic dataset with 12030 transactions for experiment. The MPNAR algorithm is simple, fast and very efficient to mine PAR and NAR. While generating all possible combinations of itemsets with a negative operator it illuminate some of the combinations and hence that rule may skip (Ramasubbareddy *et al*., 2011).

MATRIALS AND METHODS

Web usage mining is the process of extraction of knowledge form user's interactions with the web. The user's interactions are represented in terms of Web server access logs, user queries, mouse-clicks, etc., to find Web access association patterns. These web access association patterns are analyzed and can be used for web personalization, business intelligence, recommendation, etc. Figure 1 shows the process of web usage mining.

Preliminaries: Suppose I = $\{i_1, i_2, ..., i_N\}$ be a set of N distinct items and data D is a set transactions over I . Each transaction T contains a set of items $i_1, i_2, ..., i_k \in I$, i.e., $T \subseteq I$. A transaction has an associated unique identifier called TID. An association rule is an implication of the form, $A \rightarrow B$ where A, $B \subseteq I$ and $A \cap B = \emptyset$. A is called the antecedent of the rule and B is called the consequent of the rule. A set of items (antecedent or consequent) is called an itemset. The number of items in the itemset is called size of an itemset. The size of item set i_1, i_2, i_3 is three. The statistical measure used for an itemset is called support, the fraction of transactions in D containing an itemset. Let us consider an item set, AB denote by $|A \cup B|$ the number of transactions that contain both A and B and |D| denote the number of transactions in the database, support (AB) = $|A \cup B|/|D|$. To measure the strength of the association rule as $A \rightarrow B$ statistical measure is used as confidence. Confidence ($A \rightarrow B$) support (AB)/support (A). We proposed an algorithm, Mining Strong Positive and Negative Association Rules (MSPNAR) which finds

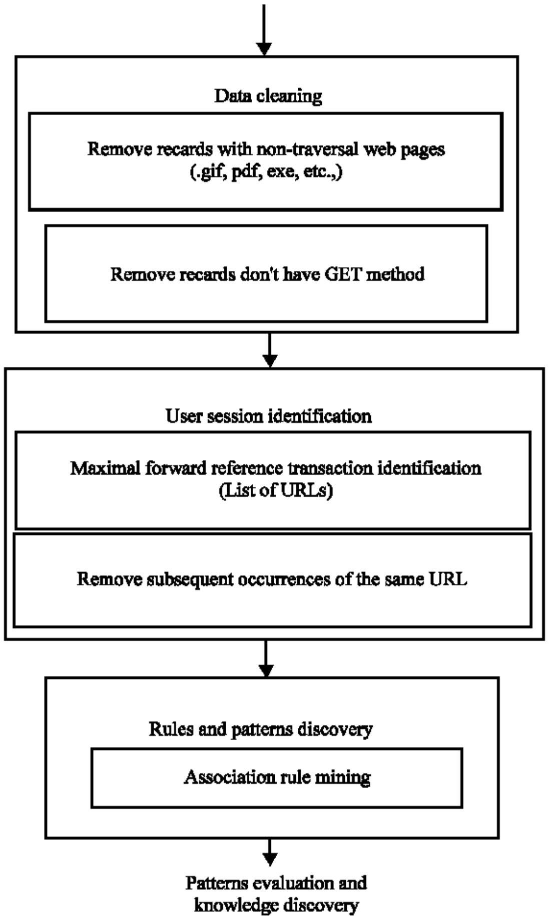
Data cleaning
Remove recards with non-traversal web pages (.gif, pdf, exe, etc.,)
Remove recards don't have GET method
User session identification
Maximal forward reference transaction identification (List of URLs)
Remove subsequent occurrences of the same URL
Rules and patterns discovery
Association rule mining
Patterns evaluation and knowledge discovery

Fig. 1: Web usage mining process

both strong positive and negative association rules by extending basic support-confidence framework with pearson's correlation coefficient.

Proposed algorithm MSPNAR:

```
Input - $L_1$: {large-1 sequences}, MS: Minimum Support; Output B L:
          Large sequences in $c_k$ $L_k$
Step 1: Generate all Large itemset:
    for (k = 2;$L_{k-1}$ = $\phi$;k++) do
    begin
        $C_k = L_{k-1} \infty L_{k-1}$
          for each i $\in$ Ck
             if $\forall$ subset of i$\notin$ $L_{k-1}$ then
                 $C_k = C_k$-{ I}
             end if
         end
           for each i$\in$ $C_k$
           begin
               s = support(i)
               if s >MS
                   $L_k = L_k \cup I$
               end if
            end
         end
Input-L: Large sequences,  MC: Minimum Confidence; Output B PAR:
Positive Association Rules, NAR: Negative Association Rules
Step 2:  Generate positive and negative association rules that have minimum
    confidence
    PAR = $\phi$, NAR = $\phi$
    for any itemset X in L do
    begin
           for any itemset $A\cup B = X$ and $A \cap B = \phi$ do
           begin
                  $\phi$ (A, B ) = calculate correlation coefficient between
             A and B using the definition 6
             if ($\phi$ (A , B ) >0 then
                   if conf ($A\rightarrow B$)$\geq$ MC then PAR = PAR $\cup$ {$A\rightarrow B$}
               end if
                   if conf ($B\rightarrow A$)$\geq$ MC then PAR = PAR $\cup$ {$B\rightarrow A$}
                end if
                    if conf ($\neg A\rightarrow \neg B$)$\geq$ MC then NAR = NAR $\cup$ {$\neg A$
              $\rightarrow \neg B$}end if
                   if conf ($\neg B\rightarrow \neg A$)$\geq$ MC then NAR = NAR $\cup$ {$\neg B\rightarrow \neg A$}
                  end if
            else if $\phi$ (A, B)<0 then
                  if conf ($A\rightarrow \neg B$)$\geq$ MC then NAR = NAR $\cup$ {$A\rightarrow \neg B$}
                 end if
                  if conf ($\neg B\rightarrow A$)$\geq$ MC then NAR = NAR $\cup$ {$\neg B\rightarrow A$}
                  end if
                  if conf ($\neg A\rightarrow B$) $\geq$ MC then NAR = NAR $\cup$ {$\neg A\rightarrow B$}
                  end if
                  if conf ($B\rightarrow \neg A$)$\geq$ MC then NAR = NAR $\cup$ {$B\rightarrow \neg A$}
                 end if
           end if
       end
   end
```

RESULTS AND DISCUSSION

Experimental setup and evaluation: To study the effectiveness of MPNAR, we have performed several experiments on intel core I 3 processor of 2.40 GHz and RAM 4.00 GB with XAMPP Server V3.2.2 and NetBeans IDE 8.0. All the experiments are conducted on real dataset, derived from United States Environmental Protection Agency (EPA). The EPA dataset contains a 24 h period of Hypertext Transfer Protocol (HTTP) requests to a web server (Cooley *et al*., 1997). Table 1 shows sample records from EPA dataset.

Effectiveness and efficiency of proposed algorithm: To evaluate the effectiveness of a proposed algorithm, we compare our approach, firstly with the support-confidence framework proposed by Agrawal and Srikant (1995) and then with mining PAR NAR by Yang and Zhao (2009). The comparison is based on ability to find positive association rules and negative association rules. The number of frequent itemsets generated by the Agrawal algorithm (Agrawal and Srikant, 1995), Yang and Zhao (2009) algorithm (Yang and Zhao, 2009) and our proposed algorithm MSPNAR are same as shown in Fig. 2.

It means the seed for all the algorithm is same but the number of association rules generated by each method is different.

When we compare the Fig. 3-5, it is clear that the number of association rules with the same

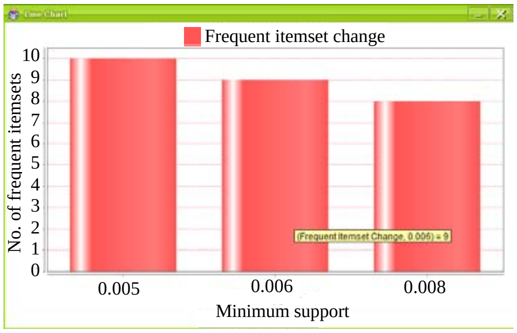

Fig. 2: Support vs. frequent itemsets

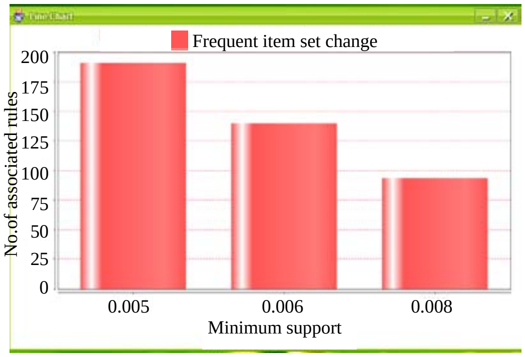

Fig. 3: Support vs. total number of rules for basic apriori algorithm (Maximum association rule count vs. support)

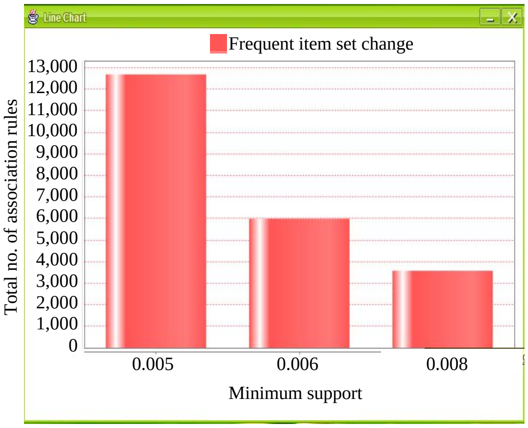

Fig. 4: Support vs. total number of rules for algorithm (Yang and Zhao, 2009) (Total PAR in NAR rule count vs. support)

support and confidence is different. In case of simple support-confidence framework proposed by Agrawal and Srikant (1995) it generates only few positive association rules.

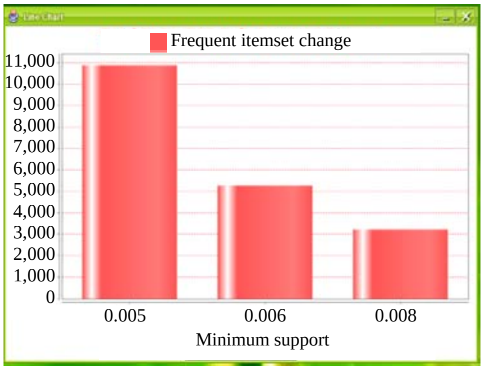

Fig. 5: Support vs. total number of rules for MSPNAR algorithm (Total PAR in NAR rule count vs. support)

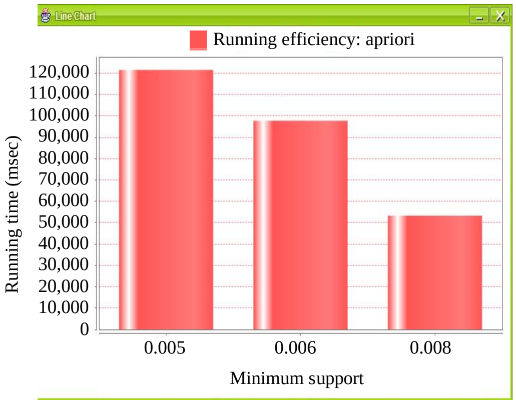

Fig. 6: Support vs. time for basic apriori algorithm

Table 1: Sample of records from EPA dataset

Session ID	Web pages
0	0, 1, 2, 3, 4
1	6, 7, 8, 9, 10
2	6, 7, 8
3	13, 14
4	15, 16, 17

Yang and Zhao (2009) generates a very large number of rules where as our proposed algorithm generates a moderate number of strong positive and negative association rules.

When the comparison is based on execution time, Fig. 6-8 show support versus time for basic Apriori, Jingrong Yang *et al.* and MSPNAR algorithms, respectively. Basic apriori takes very short time to find all the positive association rules due to only support-confidence measures. Yang and Zhao (2009) algorithm takes more time as compare to both basic apriori and MSPNAR algorithm. It is clear that our algorithm finds strong positive and negative association rules with minimum time.

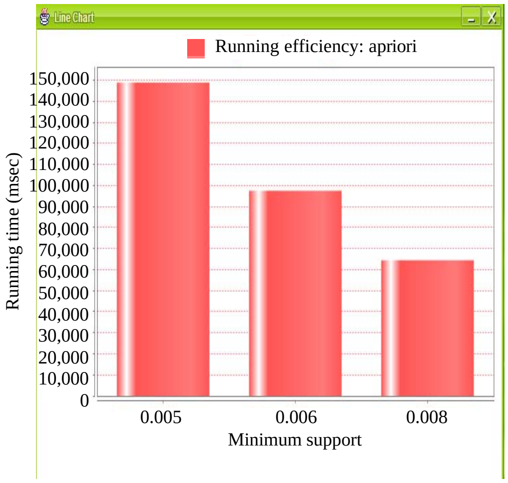

Fig. 7: Support vs. time for Yang and Zhao (2009) algorithm

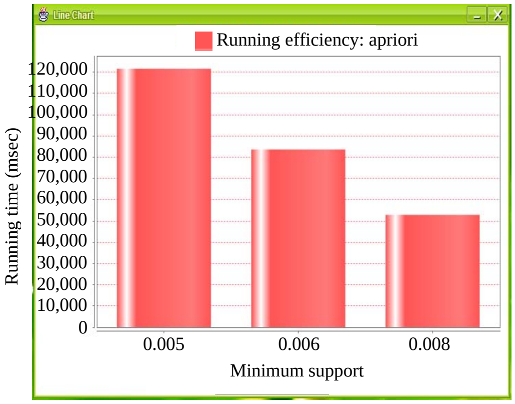

Fig. 8: Support vs. time for MSPNAR algorithm

CONCLUSION

We have proposed an algorithm called MSPNAR which mines, both PAR and NAR. Mining of PAR and NAR is very interesting and challenging because of the complexity and size of the search space. Still, very few researchers have proposed algorithms to mine both PAR and NAR. Our proposed algorithm explored all search space to find both PAR and NAR. We compared our results with well-known algorithms proposed in the literature. From experimental results it is clear that our proposed algorithm is effective and efficient without any extra database scans.

ACKNOWLEDGEMENTS

Research is supported by Avice Chancellor Research Motivation Scheme (VCRMS), NMU/11A/VCRMS/budget-2015-16/Engg. AND Tech.-13/66/2016@, to college teachers through a university fund of North Maharashtra University, Jalgaon.

REFERENCES

Aggarwal, C.C. and J. Han, 2014. Frequent Pattern Mining. Springer, Berlin, Germany, ISBN: 978-3-319-07820-5, Pages: 469.

Agrawal, R. and R. Srikant, 1995. Mining sequential patterns. Proceedings of the 11th International Conference on Data Engineering, March 6-10, 1995, Taipei, Taiwan, pp: 3-14.

Antonie, M.L. and O.R. Zaiane, 2004. An associative classifier based on positive and negative rules. Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, Junary 13, 2004, ACM, Paris, France, ISBN:1-58113-908-X, pp: 64-69.

Brin, S., R. Motwani, J.D. Ullman and S. Tsur, 1997. Dynamic itemset counting and implication rules for market basket data. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data Vol. 26, May 11-15, 1997, ACM, Tucson, Arizona, USA., ISBN:0-89791-911-4, pp: 255-264.

Cooley, R., B. Mobasher and J. Srivastava, 1997. Web mining: Information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence 1997, November 3-8, 1997, IEEE, Newport Beach, California, ISBN:0-8186-8203-5, pp: 558-567.

Hong, T.P. C.S. Kuo and S.C. Chi, 2001. Trade-off between computation time and number of rules for fuzzy mining from quantitative data. Intl. J. Uncertainty Fuzziness Knowl. Based Syst., 9: 587-604.

Hong, T.P. and Y.C. Lee, 2008. An Overview of Mining Fuzzy Association Rules. In: Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, Humberto, B., H. Francisco and M. Javier (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-73722-3, pp: 397-410.

Kiran, R.U. and P.K. Re, 2009. An improved multiple minimum support based approach to mine rare association rules. Proceedings of the IEEE International Symposium on Computational Intelligence and Data Mining CIDM'09, March 30-April-2, 2009, IEEE, Nashville, Tennessee, USA., ISBN:978-1-4244-2765-9, pp: 340-347.

Luiza, A., L. Jundong and Z. Osmar, 2014. Negative Association Rules. In: Frequent Pattern Mining, Aggarwal, C.C. and J. Han (Eds.). Springer, Berlin, Germany, pp: 135-145.

Matthews, S.G., M.A. Gongora, A.A. Hopgood and S. Ahmadi, 2013. Web usage mining with evolutionary extraction of temporal fuzzy association rules. Knowl. Based Syst., 54: 66-72.

Patil, U.M. and J.B. Patil, 2016. Mining positive and negative association rules: A survey. Intl. J. Control Theory Appl., 9: 1112-1118.

Ramaraj, E. and N. Venkatesan, 2008. Positive and negative association rule analysis in health care database. Intl. J. Comput. Sci. Network Secur., 8: 325-330.

Ramasubbareddy, B., A. Govardhan and A. Ramamohanreddy, 2011. Mining Indirect Positive and Negative Association Rules. In: Advances in Computing and Communications. Abraham, A., L. Jaime, F.B. John, S. Junichi and M.T. Sabu (Ed.). Springer, Berlin Heidelberg., pp: 581-591.

Srikant, R. and R. Agrawal, 1996. Mining quantitative association rules in large relational tables. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data Vol. 25, June 04-06, 1996, ACM, Montreal, Quebec, Canada, ISBN:0-89791-794-4, pp: 1-12.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorat., 1: 12-23.

Sujatha, V. and Punithavalli, 2012. Improved user navigation pattern prediction technique from web log data. Procedia Eng., 30: 92-99.

Tan, P.N., V. Kumar and J. Srivastava, 2002. Selecting the right interestingness measure for association patterns. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases, July 23-26, 2002, Edmonton, Canada, pp: 32-41.

Wu, X., C. Zhang and S. Zhang, 2004. Efficient mining of both positive and negative association rules. ACM Trans. Inform. Syst., 22: 381-405.

Yang, J. and C. Zhao, 2009. Study on the data mining algorithm based on positive and negative association rules. Comput. Inf. Sci., 2: 103-106.

Zaiane, M.L.A.O.R., 2007. Mining positive and negative association rules: An approach for confined rules. Master Thesis, Department of Computing Science, University of Alberta, Edmonton, Alberta.