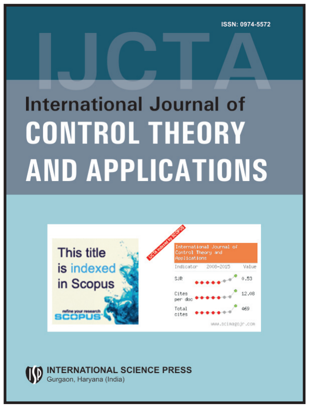

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 25 • 2017

Discovering Web Log Navigational Patterns using Negative Association Rule and Negative Selection Algorithm

R. Gobinath[a]

[a]*Assistant Professor, Department of Computer Science, School of Computing Sciences, Vels University, Chennai, Tamil Nadu, India. Email: drgobinathramar.scs@velsuniv.ac.in*

Abstract: The colossal capacity of web usage data that survives on web servers contains potentially precious information about the performance of website visitors. Pattern Mining involves applying data mining methods to large web data repositories to extract usage patterns. The emerging reputation of the World Wide Web, many websites classically experience thousands of visitors every day. Examination of who browsed what, can give imperative imminent into the buying pattern of obtainable customers. Right and timely decisions made based on this acquaintance have helped organizations in accomplishment of new heights in the market. The aim of discovering rules in Web log data is to obtain information about the navigational behavior of the users. This can be used for marketing purposes, for dynamic creation of user profiles, finding user pattern navigation, etc. Recently, mining negative association rules has received some attention and been proved to be useful in real world. This proposed work mainly concentrates on identifying efficient algorithm for mining negative association rules from the clustered weblog data. The algorithm extends traditional association rules to include negative association rules instead of using positive association rule alone. While mining negative association rules, the same minimum support threshold is used to mine frequent negative item sets. The resultant rules are optimized by applying a Negative selection algorithm applies the pruning policies of the algorithm that can discover all convincing negative association rules quickly and overcome some limitations of the traditional association rule mining methods. The experimental result reveals proposed work effectiveness.

Keywords: Web Mining, Negative Association, Negative Selection.

1. INTRODUCTION

The web usage plays a vital role for understanding the behavior of website visitors and their browsing patterns. This knowledge enhances the effectiveness of websites personalization and for web marketing. The extraction of knowledge from World Wide Web has become an essential and complicate task due to enormous amount of data which is in the form of semi structure nature. The analysis of web log files is used for applications like customer shopping sequence, web clicks, biological sequence, and dynamic users profile creation. In this paper discovery of relevant rules from the identified frequent patterns is been considered from web log files. Traditionally the web log data are in raw format. It needs to be cleaned, condensed and transformed in order to analysis, noticeable and useful information. After performing cleaning process, the relevant features which are useful for pattern recognition as to be performed for grouping similar patterns. Finally the pattern mining can be performed for extracting rules by means of association in navigation behavior [14-16].

Mining of association rules is an important research focus in web usage mining, because a web usage mining correspondingly become hotspot which uses large amount of data in the web server log and other relevant data set for mining, analysis and gains valuable knowledge model about usage of relevant web sites [17].

Association rules have been extensively used in several existing works, for their usefulness in many application domains such as disease diagnosis, decision support, telecommunication, hacker detection, marketing analysis, etc. The essential discovery of such rules has been a major focus to improve a number of remarkable variant improvements of association rule mining algorithm. Analyzing of association rules is a task of identifying association rules that occurs frequently in a given transactional data set. All the traditional association rules mining algorithms were developed to find positive associations between items and also deals with problem of generating all association rules that consist of frequent item set and the confidence greater than the user specified minimum confidence.

In the context of web personalization association rules refers to set of pages which are accessed together with the minimum support value which can help in organizing web space efficiently.

Eg: confidence is 80% of the user who accessed “News/on air/business/sports.asp” also accessed ”living/sport/travel/frontpage.asp”, but only 20% of those accessed “News/on air/business” accessed “living/sport/travel/frontpage.asp”, then it shows that some information in “sports.asp” is making the clients to access “frontpage.asp”.

With the increasing use and development of data mining patterns recently many works are focused on finding navigational patterns which can provide valuable information however mining negative association rule is difficult task. Due to the fact that there are essential differences between positive and negative association rule mining .In these papers we focus on three major key issues in negative association rule mining.

1. How to effectively improvise search for negative frequent item set?
2. How to efficiently identify negative association rule?
3. In what way the optimized negative association rules are selected to improve classification accuracy.

2. RELATED WORK

The largest numbers of rules created by association rule algorithms are partially used by information analyst. Research proposal in [1] for pruning the group of rules generated may have irrelevant rules during the rule generation stage. The mining process for removing irrelevant rules from the set of rules generated by [2] is a formal approach. The mining techniques to snip off the rule set size remains abandoned and many researchers focus on reducing rule cluster size. The most common dispute in mining association rules is hand picking the perfect motivating measures as mentioned in [3]. The rule clusters over abandoned number of motivating measures is a analyzing studies of [4] for mining relevant data and conversation about association rule mining limitations are prescribed in [5]. Over a generation of rules, false findings and rule discovery which are hard to understand and minimize constant support for some of the disputes which are solved with a proposed tactic of filtering top *k*-association discoveries. Association rule mining comes under the web usage mining environment and web usage deals with the website link structure data and closely correlated items. The web pages which are closely linked may fall under same session and has very high confidence and makes user apathy. The method introduced by [6], introduces additional association rules for web usage data during the classification of association rule mining. The Generalized association rule is a basic technique used in web usage mining for analyzing patterns. The hierarchical conceptual model is used to demonstrate the relationship between attributes in different levels. The Generalized association rule permits rules in different levels of attributes [7]. The Generalized association rules are based on the URL query and cannot be used for regular website update.

3. OBJECTIVE

The substantial information about the Web server and complete history of the user accessed the web pages are maintained as web log files on the web server. The log information is useful to discover relevant data for improving the performance of web services. The patterns extracted from the web usage mining are analyzed and can improve the business by attracting the customer, increase the market and sales. Digital Libraries, e-Commerce, e-Business, e-Learning and e-Newspapers are some of the fields successfully using the web usage mining strategy.

4. MATERIALS AND METHOD

Main objectives of finding association rules is to find all co-occurrences relationship called associations [8] introduced this concept and it has attracted a great deal of attention. An association rule is an emersion $x \rightarrow y$ where x and y are set of items. In detail given a Data base D of transaction where each transaction T € D is a set of items. $X \rightarrow Y$ express that whenever transaction T contains X then T probably contains Y also. The probability is rule confidence is defined as the percentages of transactions containing Y in addition to X with regard to occurrence number of transaction contain X. In case of web log mining

Eg: News/on air/business->/sports.asp

The example represents the pages News, on air and business are accessed along with "sports" page. It shows association that the uses those who are navigating news, onair and business will also probably access "sports" webpage. The truthiness of the rule can be identified by finding the support and confidence of the above mentioned rule with the whole transaction data base. If the gained value is greater than minimum support and minimum confidence, then the rule is valid.

5. COUNTING OCCURRENCE ALGORITHM

Before support and confidence for each rules is determined, number of occurrence for each rules must be calculated. An algorithm steps for counting the number of occurrence is shown below:

```
Pseudo code: Counting Occurrences Algorithm
Input: Rules generated from web log data
Procedure:
Step 1: Read record in database
Step 2: For each record in database
Step 3: If Filter Item Level1 $\cap$ Item Level2 < > 0 then
Step 4: Counter = Counter + 1
Step 5: End if
Step 6: Next record
Output: number of occurrence of each rule
```

6. ALGORITHM FOR SUPPORT

Support measures how often the rules occur in database. To determine the support for each rules produced, several arguments have been identified in calculating the support such as Total Transaction in database and number of occurrences for each rules. The formula for support is shown in Figure 8.

```
Pseudo code: Support calculation formula
Input: Total Transaction in DB
No. of occurrences each item {x, y}
Procedure:
    $\text{Support} = \frac{\text{Number of Occurrences } \{X, Y\}}{\text{Total Transaction in DB}}$
Output: Computing support value of each rule
```

7. ALGORITHM FOR CONFIDENCE

Formula for calculating the confidence value is shown in

```
Pseudo code: Confidence calculation formula
Input: Total occurrence for item X
Total occurrence for item X and Y
Procedure:
    $\text{Confidence} = \frac{\text{Total Occurrences for item X and Y}\}}{\text{Total Occurrence for item X}}$
Output: Computing confidence value of each rule
```

Proposed Work

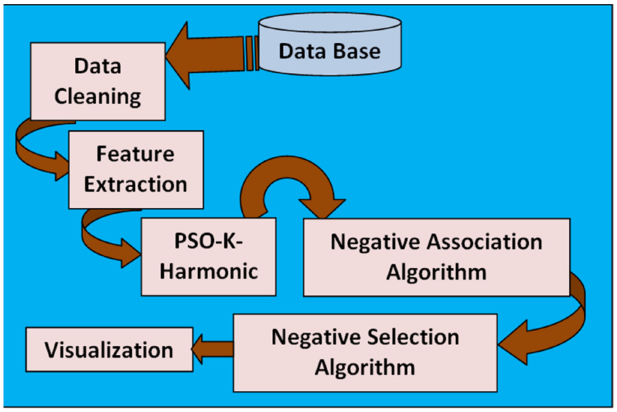

Figure 1: Architectural Diagram for Discovering Web Log Patterns

Procedure

Step 1: Collect the raw data set from web log file

Step 2: In pre-processing stage performs cleaning process by eliminating multimedia files, duplicates etc. And as a data extraction process user and session are determined

Step 3: Future extraction is done on the web log file based on the user navigation behavior.

Step 4: In order to identify similar patterns PSO-K-Harmonic mean was implied to cluster the similar user navigation patterns.

Step 5: From the clustered data set association among each item set was identified applying negative association rule mining.

Step 6: The resultant rules are justified by negative selection algorithm which prunes the resultant rules for better optimization.

Step 7: The justified resultant rules are shown clearly through visualization techniques.

The system is proposed for discovering interesting patterns in weblogs navigational sequence. Weblogs has information about accesses to various Web pages within the Web space associated with a particular server.

In case of Web transactions, association rules capture relationships among page views based on navigation patterns of users.

A. Steps Involved in the Proposed System

Our proposed system would involve the following steps:

The input is a set of Weblogs for which we have to find Positive and Negative association rules. In this paper we have chosen clustered navigational patterns of three different data sets as an input.

Now, both positive and Negative association rule mining is applied on those clusters, to obtain Negative association rules having minimum support and confidence.

As a result of negative association rule mining, interesting patterns of infrequent path traversal can be discovered and client's web usage can be evaluated.

8. NEGATIVE ASSOCIATION RULE MINING

The negation of an itemset A is indicated by ¬A, which means the absence of the itemset A. We call a rule of the form A ⇒ B a positive association rule, and rules of the other forms (A ⇒ ¬B, ¬A ⇒ B and ¬A ⇒ ¬B) negative association rules [10-13].

The support and confidence of the negative association rules can make use of those of the positive association rules [9]. The support is given by the following formulas:

$$\sup p(\neg \mathrm{A}) = 1 - \sup p(\mathrm{A}) \tag{1}$$

$$\sup p(\mathrm{A} \Rightarrow \neg \mathrm{B}) = \sup p(\mathrm{A}) - \sup p(\mathrm{A} \cup \mathrm{B}) \tag{2}$$

$$\sup p(\neg \mathrm{A} \Rightarrow \mathrm{B}) = \sup p(\mathrm{B}) - \sup p(\mathrm{A} \cup \mathrm{B}) \tag{3}$$

$$\sup p(\neg \mathrm{A} \Rightarrow \neg \mathrm{B}) = 1 - \sup p(\mathrm{A}) - \sup p(\mathrm{B}) + \sup p(\mathrm{A} \cup \mathrm{B}) \tag{4}$$

The most common framework in the association rules generation is the "support-confidence" one. Although these two parameters allow the pruning of many associations that are discovered in data, there are cases when many uninteresting rules may be produced. In this paper we consider another interesting measure called conviction that adds to the support-confidence framework. Next section introduces the measure conviction.

The confidence is given by the following formulas:

$$\mathrm{conf}(\mathrm{A} \Rightarrow \neg \mathrm{B}) = \frac{\sup\ p(\mathrm{A}) - \sup\ p(\mathrm{A} \cup B)}{\sup\ p(\mathrm{A})} \tag{5}$$

$$\mathrm{conf}(\neg \mathrm{A} \Rightarrow \mathrm{B}) = \frac{\sup\ p(\mathrm{A}) - \sup\ p(\mathrm{A} \cup B)}{1 - \sup\ p(\mathrm{A})} \tag{6}$$

$$\text{conf}(\neg\text{A} \Rightarrow \neg\text{B}) = \frac{1 - \sup\, p(\text{A}) - \sup\, p(\text{A}) + \sup\, p(\text{A} \cup \text{B})}{1 - \sup\, p(\text{A})} \quad (7)$$

The negative association rules discovery seeks rules of the three forms with their support and confidence greater than, or equal to, user-specified ms and mc thresholds respectively. These rules are referred to as an interesting negative association rule.

As mentioned before, the process of mining negative association rules can be decomposed into the following three sub problems, in a similar way to mining positive rules only.

1. Generate the set PL of frequent itemsets and the set NL of infrequent itemsets.
2. Extract positive rules of the form A ⇒ B in PL.
3. Extract negative rules of the forms A ⇒ ¬B and ¬A ⇒ B in NL.

Let DB be a database, and ms, mc, dms and dmc given by the user. Our algorithm for extracts negative association rules with a correlation coefficient measure and pruning strategies is designed. Negative Association rule mining generates only all negative association rules are extracted from association rule mining. When mining negative association rules, we use same threshold to improve the usability of the frequent negative item sets. With a Yule's correlation coefficient measure and pruning strategies, the algorithm can find all valid association rules quickly. An example of mining negative itemsets is given below for illustrative purposes.

Algorithm: Negative Association Rules

Input: TDB-Transactional Database

MS-Minimum Support

MC-Minimum Confidence

Output: Negative Association Rules

Method:

```
N ← Φ
Find F1 ← Set of frequent 1-itemsets
for ($k$ = 2; Fk − 1! = Φ; $k$++)
{
Ck = Fk − 1 join Fk − 1
// Prune using Apriori Property
for each $i$ ε Ck, any subset of i is not in
    Fk-1 then Ck = CK − {$i$}
for each $i$ ε CK find Support($i$)
    {
for each A, B (A ∪ B = $i$)
{
```

```
$QA, B = Association(A, B);$
if $Q < 0$
{
if(supp($A \leftarrow \neg B$) $\geq$ MS&&conf($A \leftarrow \neg B$) $\geq$ MC)
then
$N \leftarrow N \cup \{A \leftarrow \neg B\}$
if(supp($\neg A \cup B$) $\geq$ MS&&conf($\neg A \cup B$) $\geq$ MC)
then
$N \leftarrow N \cup \{\neg A \leftarrow B\}$
}
}
}
$AR \leftarrow P \cup N$
```

9. OPTIMIZATION OF RULE GENERATION USING NEGATIVE SELECTION ALGORITHM

The Negative Selection algorithm is inspired by the self-non self discrimination behavior observed in the mammalian acquired immune system. The clonal selection theory of acquired immunity accounts for the adaptive behavior of the immune system including the ongoing selection and proliferation of cells that select-for potentially harmful (and typically foreign) material in the body. An interesting aspect of this process is that it is responsible for managing a population of immune cells that do not select-for the tissues of the body, specifically it does not create self-reactive immune cells known as auto-immunity. This problem is known as 'self-non self discrimination' and it involves the preparation and on going maintenance of a repertoire of immune cells such that none are auto-immune. This is achieved by a negative selection process that selects-for and removes those cells that are self-reactive during cell creation and cell proliferation. This process has been observed in the preparation of T-lymphocytes, naive versions of which are matured using both a positive and negative selection process in the thymus.

```
          Input: SelfData
Output: Repertoire
Repertoire
While (StopCondition())
  Detectors GenerateRandomDetectors()
  For (Repertoire)
    If (Matches(, SelfData))
      Repertoire
    End
  End
End
Return (Repertoire)
          Input: InputSamples, Repertoire
For ( InputSamples)
    "non-self"
  For ( Repertoire)
    If (Matches(, ))
        "self"
      Break
    End
  End
End
```

- The Negative Selection Algorithm was designed for change detection, novelty detection, intrusion detection and similar pattern recognition and two-class classification problem domains.
- Traditional negative selection algorithms used binary representations and binary matching rules such as Hamming distance, and -contiguous bits.
- A data representation should be selected that is most suitable for a given problem domain, and a matching rule is in turn selected or tailored to the data representation.
- Detectors can be prepared with No. prior knowledge of the problem domain other than the known (normal or self) dataset.
- The algorithm can be configured to balance between detector convergence (quality of the matches) and the space complexity (number of detectors).
- The lack of dependence between detectors means that detector preparation and application is inherently parallel and suited for a distributed and parallel implementation, respectively.
- In our proposed method the rules generated using negative association rule mining are pruned using the negative selection algorithm
- This negative selection algorithm optimizes the performance of the negative association rule for improving the accuracy of classification.

10. EXPERIMENTAL RESULTS

For conducting experimental result in this proposed work we have used the three different log server files for feature extraction Kdlog, Online shopping server log file and msnbc.com datasets.

The below table shows the sample rule generation based on positive association rule with the transaction support and confidence.

Table 1
Positive Association rule Support and Confidence

Positive Association rule	*Support*	*Confidence*
Opinion and misc and travel → on air	90.26	92.37
Living and sports and business and bbs → front page	90.00	91.72
News and tech and living and business and sports → front page	89.00	90.81
Front page and tech and opinion and living and news	87.60	88.59
Front page and tech and living and business → sports	87.87	88.01
News and living and business and sports → front page	87.18	88.14
Misc and living and travel → on air	86.55	87.12
News and tech and misc and bbs → front page	85.99	84.52
Misc and business and travel → on air	85.65	85.39
Misc and business and bbs → front page	85.57	85.41
Tech and living and sports and bbs → news	85.49	86.09
Local and misc and business and sports → frontpage	85.32	87.35
News and on air and business → sports	85.01	85.69
Front page and opinion and living and sports → news	87.81	87.94
News and misc and business and sports → front page	86.22	86.84

The Table 1 shows the support and confidence obtained using positive association rules was shown. In this the "on air " page was mostly associated when accessing " Opinion\misc\travel " with 90.26% support and 92.37% confidence. It was leastly associated with " misc\business\travel" with 85.65% as support and 85.39% as confidence.

The below table shows the sample rule generation based on Negative association rule with the transaction support and confidence.

Table 2
Negative Association rule Support and Confidence

Negative Association rule	*Support*	*Confidence*
Not Opinion and misc and Not business → on air	89.21	91.53
Living and not sports and not travel → front page	91.27	91.45
Not News and not tech and living and not business and sports → front page	93.56	93.94
Not Front page and not tech and not opinion and living → news	94.28	94.63
Front page and not tech and not living and business → sports	91.20	94.51
Not News and not living and not business and sports → front page	89.49	89.53
Not Misc and not living and not travel → on air	88.53	88.09
News and tech and misc and not bbs → front page	87.28	87.05
Misc and not business and travel → on air	86.14	86.13
Not Misc and business and not bbs → front page	85.94	85.37
Not Tech and living and not sports and bbs → news	86.31	86.25
Not Local and misc and not business and not sports → frontpage	85.58	85.63
News and not onair and business → sports	85.32	85.29
Front page and not opinion and not living and not sports → news	85.36	85.41
News and not misc and business and not sports → front page	85.39	85.04

The Table 2 shows the support and confidence obtained using Negative association rules. In this the "news" page was mostly associated when accessing the pages which are not navigated through front page, tech, opinion but through living, which was represented as "Not Front page and not tech and not opinion and living "with 94.28% support and 94.63% confidence. It was leastly associated while accessing the pages through front page but not with opinion, living, sports. The negative association rule such as "Front page and not opinion and not living and not sports" contains 85.36% as support and 85.41% as confidence.

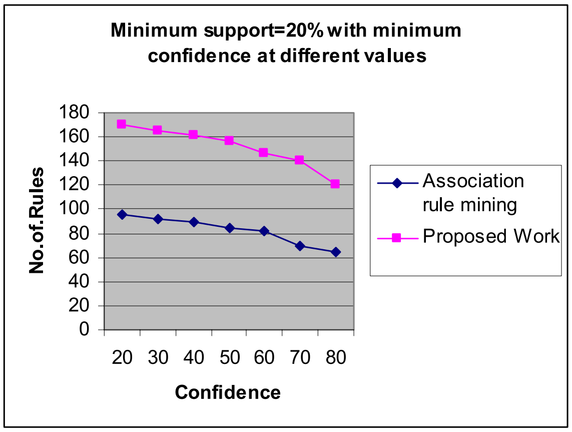

Figure 2: Minimum support = 20% and different minimum confidences

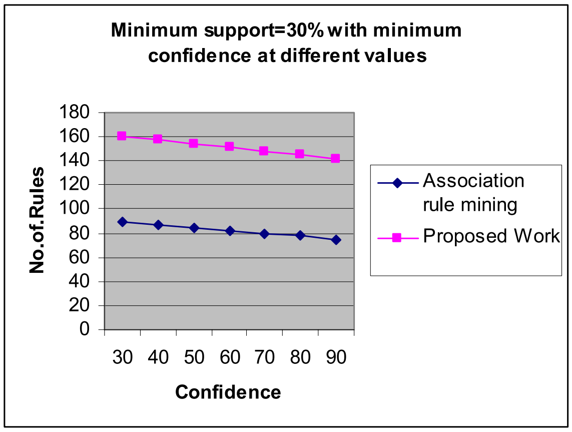

Figure 3: Minimum support = 30% and different minimum confidences

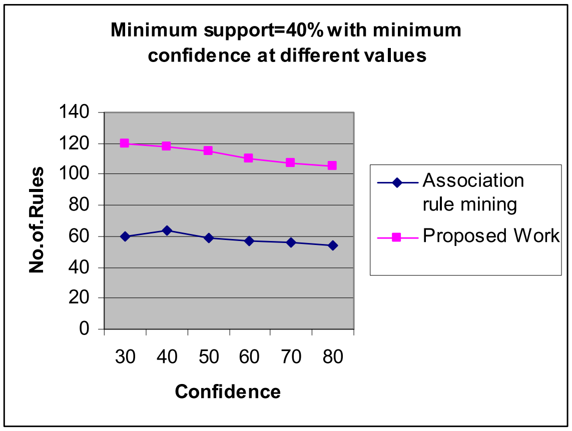

Figure 4: Minimum support = 40% and different minimum confidences

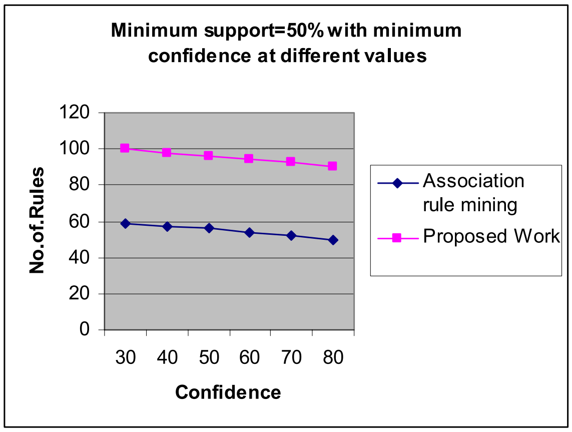

Figure 5: Minimum support = 50% and different minimum confidences

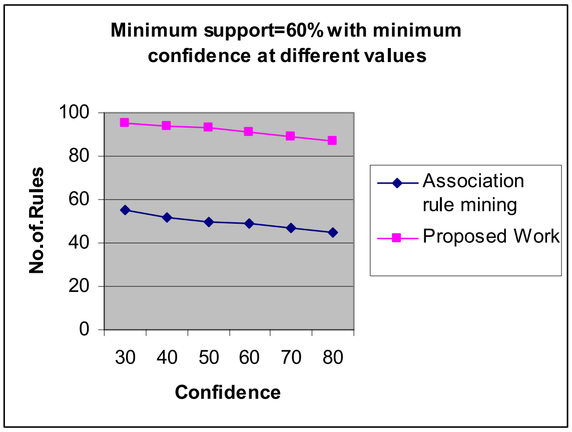

Figure 6: Minimum support = 60% and different minimum confidences

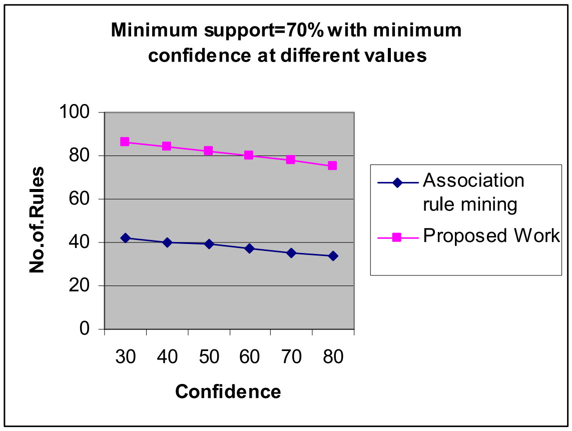

Figure 7: Minimum support = 70% and different minimum confidences

The Figure 2, 3, 4, 5, 6 and 7 shows, that the performance of association rule mining with the proposed approach using different values of support and confidence. When the value of confidence started increasing the number of rules get decreased to sustain the minimum support and confidence.

We tested our proposed algorithm of Negative association rule mining after pruning the generated rules with negative selection algorithm and the comparison was performed with existing traditional association rule mining which considers only the positive rules. We consider a web log file transactions for three different datasets. We tested our algorithm with different minimum supports and minimum confidences. Our algorithm is performing well than existing one. It has focuses the important fact that when the support and confidence level increases its minimum values the number of rules dropped considerably.

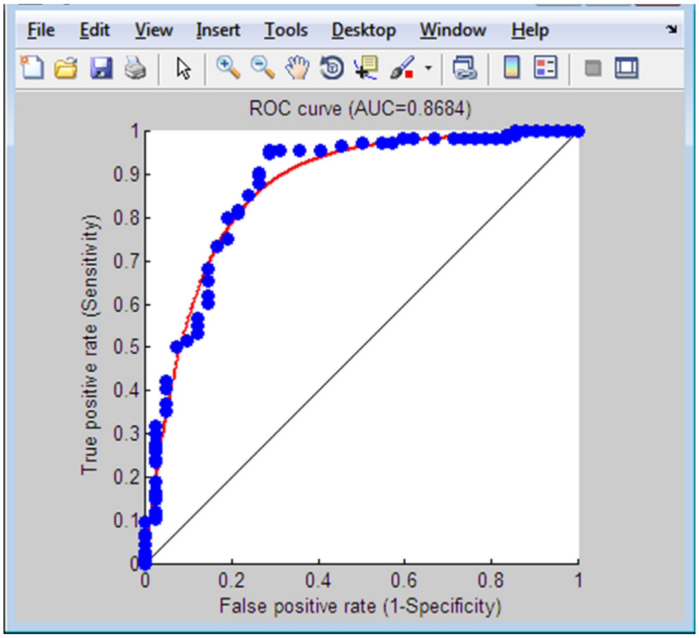

Figure 8: ROC curve with performance

The result of ROC curve with performance of proposed work is displayed with the AUC 0.8684 The maximum specificity and efficiency is shown. The Max Sensitivity Cut-off point = 194.00, Max Specificity Cut-off point = 122.00, Cost effective Cut-off point = 152.00 and Max Efficiency Cut-off point = 161.00

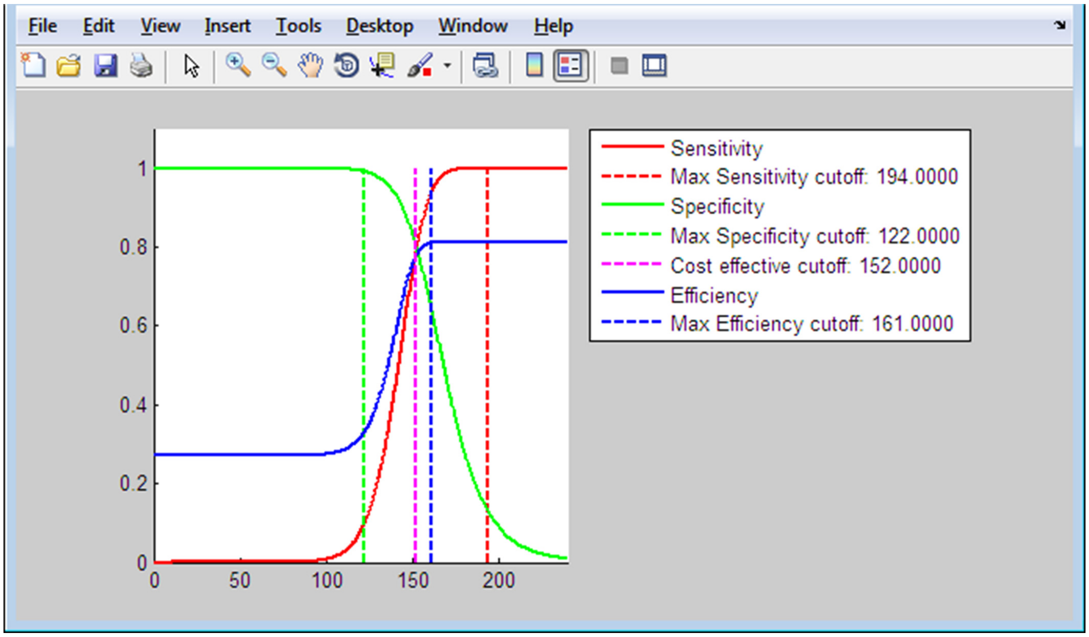

Figure 9: Sensitivity cutoff

11. CONCLUSION

In this proposed paper, we have designed a new architecture for efficiently mining negative association rules in clustered weblog dataset. To cluster the similar navigation pattern we adopted from our earlier work PSO-K harmonic mean. Our approach is novel and unique from existing traditional association rule mining research work on weblog dataset. We have intended pruning strategies for reducing the search space and improving the usability of mining rules, and have used the Negative selection algorithm which is an inspiration of artificial immune system to judge which selects the promising negative association rule should be used for better optimization of weblog data classificaiton. The result shows that instead of considering the frequent itemset in transaction, the infrequent itemset should also be taken into the consideration for better managerial decision support in designing the websites based on the category of user's behavior.

REFERENCES

[1] Sahaaya, A.M., and Malarvizhi M. (2010). Improving web navigation technique using weighted order representation. International Journal of Research and Reviews in Computer Science, 1(2), 55-60.

[2] Balcázar J.L. (2010). Redundancy, reduction schemes, and minimum-size bases for association rules, logical methods. Computer Science, 6(2:3), 1-33.

[3] Huang, X. (2007). Comparison of interestingness measures for web usage mining: An empirical study, International Journal of Information Technology & Decision Making (IJITDM), 6(1), 15-41.

[4] Webb, G.I. (2011). Filtered-top-k association discovery. WIREs Data Mining Knowl Discov 2011, 1, 183-192. doi: 10.1002/widm.28

[5] Dimitrijevic M., & Bosnjak Z. (2011). Association rule mining system. Interdisciplinary Journal of Information, Knowledge, and Management, 6, 137-150.

[6] Kosala, R., & Blockeel, H. (2000). Web mining research: A survey, SIGKDD Explorations, 2(1), 1-15.

[7] Tan, P., Kumar, V., & Srivastava, J. (2004). Selecting the right interestingness measure for association patterns. Information Systems, 29(4), 293-313.

[8] R. Agrawal, T. IMIELINSKI, and A. SWAMI, "Mining association rules between sets of items in massive databases," In Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, Washington D.C., 1993, pp. 207-216.

[9] M.L. Antonie and O.R. Za¨ıane, "Mining Positive and Negative Association Rules: an Approach for Confined Rules", Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, 2004, pp. 27-38.

[10] Wu, X., Zhang, C., Zhang, S.: efficient mining both positive and negative association rules. ACM Transactions on Information Systems, Vol. 22, No. 3, July 2004, Pages 381-405.

[11] Wu, X., Zhang, C., Zhang, S.: Mining both positive and negative association rules. In: Proc. of ICML. (2002) 658-665

[12] Yuan, X., Buckles, B., Yuan, Z., Zhang, J.:Mining Negative Association Rules. In: Proc. of ISCC. (2002) 623-629.

[13] Honglei Zhu, Zhigang Xu: An Effective Algorithm for Mining Positive and Negative Association Rules. International Conference on Computer Science and Software Engineering 2008.

[14] Goethals, B., Zaki, M., eds.: FIMI'03: Workshop on Frequent Item set Mining Implementations. Volume 90 of CEUR Workshop Proceedings series. (2003) http://CEUR-WS.org/Vol-90/.

[15] Teng, W., Hsieh, M., Chen, M.: On the mining of substitution rules for statistically dependent items. In: Proc. of ICDM. (2002) 442-449.

[16] Tan, P., Kumar, V.: Interestingness measures for association patterns: A perspective. In: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining. (2000)

[17] Gourab Kundu, Md. Monirul Islam, Sirajum Munir, Md. Faizul Bari ACN: An Associative Classifier with Negative Rules 11th IEEE International Conference on Computational Science and Engineering, 2008.

[18] Gobinath, R and Hemalatha, M. (2014). A Negative Association Rules For Web Usage Mining Using Negative Selection Algorithm. Journal of Theoretical and Applied Information Technology, 64(3), 687-695.