International Journal of Technology (2018) 1: 201-211
ISSN 2086-9614

# COMPARATIVE PERFORMANCE OF INTERESTINGNESS MEASURES TO IDENTIFY REDUNDANT AND NON-INFORMATIVE RULES FROM WEB USAGE DATA

Dilip Singh Sisodia[1*], Riya Singhal[1], Vijay Khandal[1]

[1]*Department of Computer Science and Engineering, National Institute of Technology Raipur, G.E. Road, Raipur, Chhattisgarh 492010, India*

(Received: April 2017 / Revised: June 2017 / Accepted: January 2018)

## ABSTRACT

Association rules are used to predict frequent web user behaviors from web usage data. These rules are formed using frequent items. The number of association rules increases as the number of frequent items increases and produces several redundant and non-informative rules. In this paper, five interestingness measures, including cosine, lift, leverage, confidence, and conviction with a constant value of support are compared based on the number of redundant and non-informative rules that they produce. Redundant and non-informative rules are a subset of rules present in the top generated rules. The experimental results suggested that leverage produced the least number of redundant rules in the top rules but also produced the least informative rules among all measures. Lift showed the highest number of redundant rules but the most informative rules among all the measures.

*Keywords:* Association rule mining; Interestingness measures; Non-informative rules; Redundant rules; Weblogs

## 1. INTRODUCTION

Association rule mining is an iterative and interactive process used to discover significant, novel, and interesting rules from a database (Li et al., 2014). The discovered rules are used to identify relationships among items with diverse applications. The most popular application is in the business field, where relationships among buying patterns are used for decision making and effective marketing. Other examples include personalization and patterns in a biological database (Sisodia & Verma 2011). URL accessing data (weblog) is highly interconnected and repetitive in nature. To identify frequent user access patterns from weblogs, frequent items are generated. The generated frequent items are used to produce association rules. Due to the inherent nature of weblog data, the number of association rules increases significantly and contains several redundant and non-informative rules. These redundant and non-informative rules do not contribute to identifying frequent user access patterns and generate considerable processing and storage overhead. Therefore, all generated rules must be evaluated by different interestingness measures to minimize redundant and non-informative rules.

Constraints and interestingness measures are used to identify interesting association rules. Association rules use two types of constraints, support, and confidence, to identify interesting rules; however, rules with a high value of support and confidence can still be uninteresting. There are approximately 20 interestingness measures that have been proposed to determine

*Corresponding author's email: dssisodia.cs@nitrr.ac.in, Tel: +91-771-2254200, Fax: +91-771-2254-600
Permalink/DOI: https://doi.org/10.14716/ijtech.v9i1.1510

the interestingness of the rules (Tan et al. 2002). These measures provide conflicting information about the interestingness of the rules, and not all measures are suitable for all domains.

Association rules are formed using frequent items. The number of association rules increases as the number of frequent items increases. The algorithms used to identify frequent items generate several redundant rules that convey the same meaning. It has been found that the number of redundant rules generated is exponential in length for the longest frequent item (Zaki, 2000), which incurs significant processing overhead.

Redundant rules are obstacles to the efficient utilization of association rules. Therefore, eliminating them is important. In this paper, the top rules are presented based on five interestingness measures, which were analyzed to remove redundant and non-informative rules. Redundant rules are rules that are either consequent or antecedent and are a subset of other rules. After their removal, a set of valid rules is obtained. These rules are then analyzed to obtain informative rules. Informative rules are the rules that contain frequent items and thus represent the relationships among frequent items. The performances of the interestingness measures were compared to generate the highest and lowest number of redundant and non-informative rules and the time required to generate these rules.

## 2. RELATED WORK

Association rules and their applications in different domains are discussed in detail in the literature. In (Sisodia et al., 2016), a fast prediction of web user browsing behaviors using the most interesting patterns was discussed using the modified parallel FP-growth algorithm. In (Dimitrijević et al., 2010), the authors applied a set of basic pruning schemes to reduce the rule set size and to remove a significant number of non-interesting rules. Their experiments confirmed that the set of generated association rules contained too many non-interesting rules before pruning, which made it difficult to find and exploit useful information. In (Ashrafi et al., 2004), the authors examined various causes for the redundancy problem in association rule mining. They proposed several methods to eliminate redundant rules. The proposed methods rigorously verify each rule and remove redundant rules, which generate a small number of rules from a given frequent itemset compared to traditional approaches. The experimental evaluation also suggested that the proposed methods not only theoretically eliminate redundant rules but also reduce redundant rules from real datasets. In (Zaki, 2004), the authors proposed a new framework based on closed itemsets that drastically reduces the rule set and presents it succinctly. In (Tan et al., 2002), the authors described several essential properties that should be considered before selecting the most appropriate measure to use for a given application domain. In this study, an algorithm was used to select a small set of rules in tabular form using interesting measures.

## 3. METHODOLOGY

The methodology of this paper was designed to evaluate rules generated from weblogs using five interestingness measures, including cosine, lift, leverage, confidence, and conviction, keeping support constant. The performance of all measures was compared based on the number of redundant and informative rules that they produced.

The algorithms used for association rule mining involve two phases (Zaki, 1999):

i. Identify all frequent itemsets with support greater than minimum support.
ii. Generate strong rules with minimum confidence.

A set of items present in any transaction of a database is referred to as an itemset. An itemset with k items is referred to as a k-itemset. The support of an itemset X denoted σ(X) is the number of transactions in which the itemset occurs as a subset. A k-subset is a k-length subset of an itemset. An itemset is frequent or large if its support is more than user-specified minimum support (min_sup) value.

An association rule is an expression A ⇒ B, where A and B are itemsets. The support of the rule is the joint probability of a transaction containing both A and B and is given as σ (A∪B). The confidence of the rule is the conditional probability that a transaction contains B, given that it contains A and is given as σ (A ∪ B)/σ (A). A rule is frequent if its support is greater than min_sup and strong if its confidence is more than user-specified minimum confidence (min_conf) (Zaki 1999).

### 3.1. Redundant Rules

Redundant rules are rules whose consequent or antecedents are a subset of other rules that are present. Consider a valid rule of form {x,y,z}→{a,b}. Valid rules are rules from which redundant rules are derived. There were two types of redundant rules removed in this study:

a. {x,y}→{a,b} is a redundant rule because its antecedent (LHS part of the rule) is a subset of the antecedent of the valid rule. The valid rule contains this information, and therefore these types of rules become redundant and do not convey any new information about these rules.

b. {x,y,z}→ {a} is a redundant rule because its consequent (RHS part of the rule) is a subset of a consequent part of the valid rule. This rule does not convey any new information about the associations among the itemsets.

Algorithms 1, 2, and 3 are used to identify redundant rules. The notations used in the algorithms are summarized in Table 1.

### 3.2. Non-informative Rules

Rules are considered non-informative if they are the subset of all rules. For example:

a. The rules {x,y}→{a} and {x,y}→{b} are non-informative rules because they have a common antecedent, and the consequent part of the rules can be merged to form a single rule {x,y}→{a,b}.

b. The rules {x}→{a,b} and {y}→{a,b} are non-informative rules because they have a common consequent, and the antecedent part of the rules can be merged to form a single rule {x,y}→{a,b}.

Table 1 Notations used for redundant association rule mining

| Symbol | Meaning |
|---|---|
| list_of_rules | List of all association rules |
| Antecedent | Stores all antecedents for list_of_rules |
| Consequent | Stores all consequents for list_of_rules |
| Length() | Used to find length of the list |
| Intersection() | Used to find the intersection between two lists |
| C | Length of intersection |
| removable_index | Index from list_of_rules that stores redundant rules |

The algorithm used to identify a redundant rule is presented in Appendix.

### 3.3. Interestingness Measure

For the experiment, six interestingness measures were used, which are briefly described in the following sub-sections.

#### *3.3.1. Support*

Support (Equation 1) indicates the number of transactions that contain both X and Y (Agrawal et al., 1993). Its value lies in the range [0, 1].

$$supp(X \rightarrow Y) = P(XUY) \qquad (1)$$

### *3.3.2. Confidence*

Confidence (Equation 2) provides the fraction of a total number of transactions that contain Y, given that the transaction contains X (Agrawal et al., 1993). Its value lies in the range [0, 1].

$$conf\,(X \rightarrow Y) = P(Y|X) = \frac{P\,(XUY)}{P\,(X)} \tag{2}$$

### *3.3.3. Cosine*

Cosine (Equation 3) is used to determine how X and Y are related. If closer to 0, transactions that contain X do not contain Y, and vice versa. If closer to 1, transactions that contain X also contain Y, and vice-versa (Azevedo & Jorge, 2007). Its value lies in the range [0, 1].

$$cosine\,(x, y) = \frac{P(XUY)}{\sqrt{P(X).P(Y)}} \tag{3}$$

### *3.3.4. Lift*

The lift (Equation 4) value explains how the occurrence of one item "lifts" the occurrence of another item (Brin et al., 1997). Its value lies in the range [0, +∞].

$$lift(X \rightarrow Y) = \frac{P(X,Y)}{P(X).P(Y)} = \frac{P(Y|X)}{P(Y)} = \frac{conf\,(X \rightarrow Y)}{P\,(Y)} \tag{4}$$

### *3.3.5. Leverage*

Leverage (Equation 5) is the difference between X and Y appearing together in a dataset and whether they are independent (Azevedo & Jorge, 2007). Its value lies in the range [-0.25, 0.25].

$$leve\,(X \rightarrow Y) = P(Y|X) - P(X).P(Y) = supp\,(XUY) - supp\,(X).supp\,(Y) \tag{5}$$

### *3.3.6. Conviction*

Conviction (Equation 6) is the ratio of expected frequency that X occurs without Y, if X and Y are independent, divided by the observed frequency of incorrect prediction (Brin et al., 1997). Its range is [0.5, +∞].

$$onv(X \rightarrow Y) = \frac{P(X)P\,(Y')}{P(XUY')} = \frac{(1 - supp(Y))}{(1 - conf\,(X \rightarrow Y))} \tag{6}$$

## 4. RESULTS AND DISCUSSION

The experiments were performed using Apache Spark 1.6.0 on a personal computer equipped with an Intel Core i3 processor, 4GB RAM, 313 GB hard disk, and Ubuntu 14.04 operating system. The proposed algorithm was implemented in Python language using pySpark API (Spark, 2015). For the experiments, the NASA weblog was used (NASA_SeverLog, 1995). These weblogs were recorded in the Apache weblog format and had large-scale temporal data. The summary of the raw log datasets is given in Table 2.

Table 2 Summary of statistical information computed from raw web server log files

| Parameters | NASA_Access_Log_Jul95 | NASA_Access_Log_Aug95 |
|---|---|---|
| Web access log durations | 00:00:00 1st July 1995 to 23:59:59 31st July 1995 | 00:00:00 1st August 1995 to 23:59:59 31st August 1995 |
| Size of uncompressed log file | 205.2 MB | 167.8 MB |
| Number of log records (total hits) | 1,891,715 | 1,569,898 |
| Number of unique users | 81,969 | 75,042 |
| Number of unique page URLs | 21,192 | 15,337 |
| Number of sessions | 162,362 | 141,443 |

The web log files may have some incomplete or irrelevant data that must be removed from the log files to produce a clean weblog to apply various mining algorithms. All the records that have missing data, URLs with an image, audio or video extensions, and records with an exception status code were removed during data cleaning to generate clean data as suggested in (Sisodia et al., 2015b; Sisodia et al., 2015a). The major challenge in the generation of sequence data is identifying the user sessions. The user sessions are extracted based on a time-oriented sessionization scheme (Mobasher & Liu, 2007). Some security related websites define the user session for ten minutes only. Other general sites have user sessions of one hour, and some websites do not identify user sessions. In this study, a session of one hour was chosen (Sisodia et al., 2016a; Sisodia et al., 2016b). A summary of the pre-processed log datasets is provided in Table 3.

Table 3 Summary of statistical information computed from cleaned and pre-processed server log files

| Parameters | NASA_Access_ Log_Jul95 | NASA_Access_ Log_Aug95 |
|---|---|---|
| Number of log records after removing missing values | 1,890,851 | 1,569,003 |
| Number of log records removed with multimedia extensions | 657,993 | 49,5028 |
| Number of log records removed with an unsuccessful status code | 649,956 | 487,874 |
| Number of unique users | 75,601 | 69,594 |
| Number of unique URLs | 17,043 | 11,843 |
| Number of sessions | 145,195 | 127,418 |

## 4.1. Frequent Sequence Pattern Mining

The most popular frequent-sequence pattern mining algorithm, FP-Growth (Han et al., 2004), was applied to obtain the association rules. After performing data pre-processing and generating the sequence data, a sample record of 1000, 10,000, 100,000, and 1,000,000 sizes were selected for the experiment from both weblogs. Each sequence consisted of the URLs that were accessed by the user during the user session. The URLs sequence is represented in the number form for better understanding. These sequence numbers are unique in a particular session. There were several frequent sequences produced. An example of one such sequence that was produced is:
4529, 4026, 15693, 13885

Where 4529 – “/shuttle/missions/sts-70/mission-sts-70.html”,
4026 – “/shuttle/countdown/”,
15693 – “/shuttle/resources/orbiters/discovery.html”,
13885 – “/shuttle/missions/sts-70/images/images.html”.

These sequences were then used to identify association rules using the FP-Growth algorithm.

## 4.2. Association Rules

The five interestingness measures, confidence, conviction, lift, leverage, and cosine were considered to identify the association rules while keeping support constant. After finding all the rules with min_sup 0.009, the top10, 50, 100, 500, and 1000 rules for each measure were determined by considering the maximum value for each interestingness measure. For the top rules, the redundant rules were computed using an automated script.

Table 4 Redundant rule evaluation

| No. of Top Rules | Number of Redundant Rules for Various Interesting Measures | | | | |
|---|---|---|---|---|---|
| | Confidence | Lift | Conviction | Leverage | Cosine |
| 10 | 2 | 3 | 2 | 1 | 3 |
| 50 | 24 | 22 | 21 | 16 | 29 |
| 100 | 56 | 59 | 43 | 38 | 67 |
| 500 | 286 | 391 | 287 | 266 | 364 |
| 1000 | 606 | 805 | 607 | 538 | 771 |

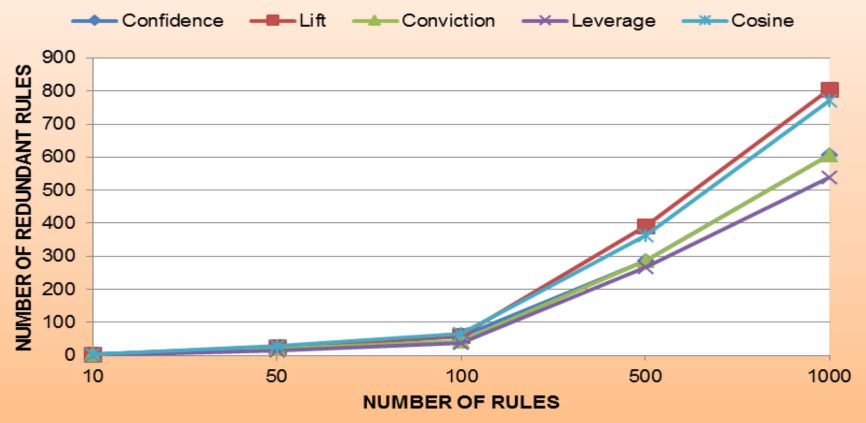

Figure 1 Comparative performance of various interestingness measures used to generate a number of redundant rules

Table 4 shows the number of top rules according to the five parameters. The numbers of redundant rules that occurred in these top rule sequences were used to identify the association rules using the FP-Growth algorithm. Figure 1 shows the comparative performance of each interestingness measure used to generate the number of redundant rules. Table 5 shows the number of top 10 rules generated according to the cosine measure.

Table 5 Top 10 rules of cosine

| Top 10 Rules | Cosine |
|---|---|
| { 4644,11399,6123 } => { 5688,11577,6161,9471 } | 0.917617 |
| { 5688,11577,6161,9471 } => { 4644,11399,6123 } | 0.917617 |
| { 5688,11577,6161 } => { 4644,11399,6123 **}** | 0.915584 |
| { 4644,11399,6123 } => { 5688,11577,6161 } | 0.915584 |
| { 4644,11399,6123,9471 } => { 5688,11577,6161 } | 0.914553 |
| { 5688,11577,6161 } => { 4644,11399,6123,9471 } | 0.914553 |
| { 4644,6123,6161 } => { 5688,11577,11399,9471 } | 0.911520 |
| { 5688,11577,11399,9471 } => { 4644,6123,6161 } | 0.911520 |
| { 4644,6123,6161 } => { 5688,11577,11399 } | 0.906795 |
| { 4644,6123,6161,9471 } => { 5688,11577,11399 } | 0.905773 |

{ 5688,11577,6161,9471 } => { 4644,11399,6123 }, { 5688,11577,6161 } => { 4644,11399,6123 }, { 4644,11399,6123 } => { 5688,11577,6161 } and { 4644,6123,6161 } => { 5688,11577,11399 } are redundant, as they are a subset of rules { 5688,11577,6161,9471 } => { 4644,11399,6123 }, { 4644,11399,6123,9471 } => { 5688,11577,6161 }, and { 4644,6123,6161,9471 } => { 5688,11577,11399 }, respectively. The list of valid rules for the cosine measure after removing the redundant rules is shown in Table 6.

Table 6 Valid rules of cosine

| Valid Rules | Cosine |
|---|---|
| { 4644,11399,6123 } => { 5688,11577,6161,9471 } | 0.917617 |
| ['5688', '11577', '6161', '9471'] => ['4644', '11399', '6123'] | 0.917617 |
| ['4644', '11399', '6123', '9471'] => ['5688', '11577', '6161'] | 0.914553 |
| ['5688', '11577', '6161'] => ['4644', '11399', '6123', '9471'] | 0.914553 |
| ['4644', '6123', '6161'] => ['5688', '11577', '11399', '9471'] | 0.911520 |
| ['5688', '11577', '11399', '9471'] => ['4644', '6123', '6161'] | 0.911520 |
| ['4644', '6123', '6161', '9471'] => ['5688', '11577', '11399'] | 0.905773 |

### 4.3. Time Required to Identify the Top Rules

Table 7 presents the list of top rules and the time required to identify the top rules. Figure 2 provides a graphical representation of the results.

Table 7 Number of rules vs. Time required to identify top rules (in ms)

| Number of Top Rules | Time Required to Identify Top Rules (in ms) |
|---|---|
| 10 | 0.355 |
| 50 | 3.957 |
| 100 | 13.498 |
| 500 | 362.913 |
| 1000 | 1206.623 |

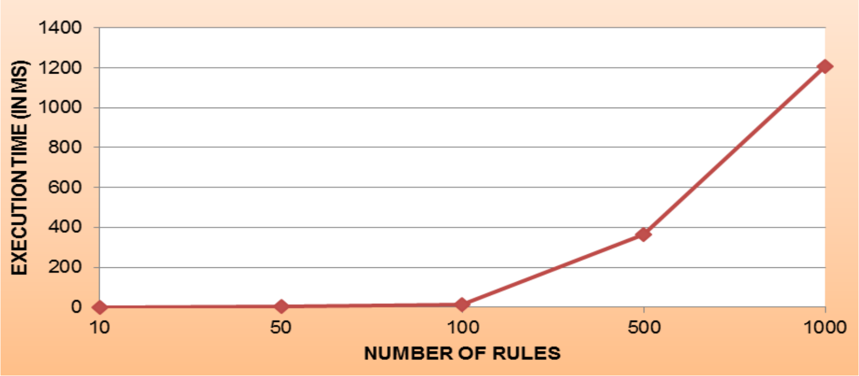

Figure 2 Number of top rules vs. execution time

### 4.4. Interestingness Measures with a Similar Number of Redundant Rules

Based on Figure 1, it can be observed that the value of redundancy for confidence and conviction are similar, and that of list and cosine are similar as well. Leverage had the lowest redundancy. The separate graphs are shown in Figures 3 and 4 for lift and cosine and confidence and conviction, respectively.

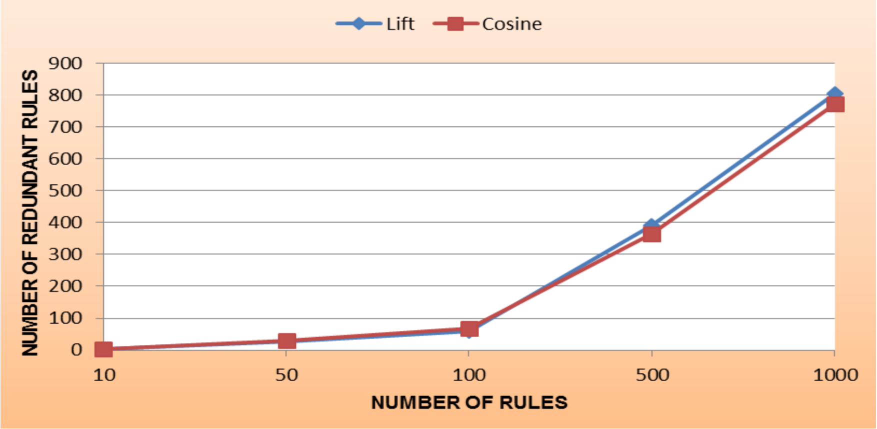

Figure 3 Number of redundant rules vs. Number of top rules for lift and cosine

Based on the formula for the interestingness measures, it can be verified that lift and cosine and confidence and conviction produced similar top rules. As the value of confidence increased, the value of conviction also increased. Similarly, if the value of lift was high, the value of cosine was also high.

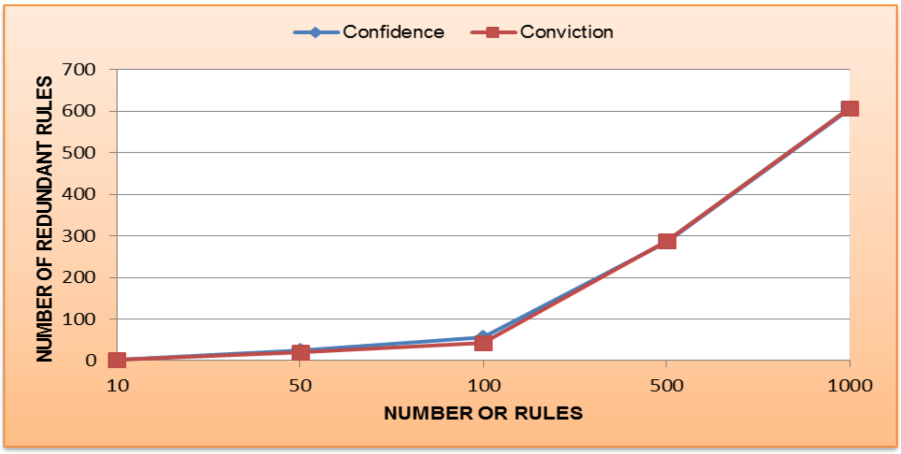

Figure 4 Number of redundant rules vs. Number of top rules for confidence and conviction

### 4.5. Interestingness Measure that Produced More Informative Rules

When the rules were evaluated for redundancy, it was observed that lift produced rules that were more informative. The rules were considered more informative if the valid rules found could not be combined to form a new rule. The examples of top rules for the lift are: {/history/mercury/ma-8/ma-8.html,/history/mercury/mr-4/mr-4.html,/history/mercury/ma-6/ma-6.html, , /history/mercury/mercury.html} → {/history/mercury/ma-7/ma-7.html, /history/mercury/ma-9/ma-9.html, /history/mercury/mr-3/mr-3.html}.

These URLs were replaced with unique numbers for a better understanding. They are represented as follows:
{4644, 11399, 6123, 9471} → {5688, 11577, 6161} (Lift: 769.945)

This rule had a high value of lift at 769.945. The top 100 rules also contained rules for which the above rule is a superset. Some of the rules are shown in Table 8.

Table 8 Top rules according to lift

| Rules | Lift |
|---|---|
| {4644, 11399, 6123} → {5688, 11577, 6161} | 760.739 |
| {4644, 6123, 9471} → {5688, 11577} | 671.0284 |
| {4644, 6123} → {5688, 11577} | 669.71 |
| {4644, 11399} → {5688, 11577} | 665.732 |

Several more rules were also present in the top rules as a subset of the valid rules. These rules were redundant and increased the number of redundant rules for the interestingness measures. The redundancy for lift was approximately 80%. Cosine also produced top rules similar to those of lift. The redundancy for cosine was approximately 77%.

Examples of top rules according to leverage are shown in Table 9.

Table 9 Top rules according to leverage

| Rules | Leverage |
|---|---|
| {11579, 4622} → {9375} | 0.0214 |
| {1941, 4622} → {9375} | 0.0123 |
| {11579, 6494} → {9375} | 0.0112 |
| {7035, 4622} → {9375} | 0.0107 |
| {2293, 4622} → {9375} | 0.0104 |

Several more rules had the same consequent (RHS part of the rule). Their antecedents (LHS part of the rule) could be combined to form a single valid rule, such as: {11597, 4622, 1941, 6494, 7035, 2293} → {9375}.

On the other hand, the value of leverage, according to the leverage formula from Table 5, was subtle. Hence, it did not occur in the top rules. Thus, the redundancy for the leverage measure was approximately 54%, which is quite low compared to other measures.

Confidence and conviction produced rules that contained rules similar to both leverage and lift. Examples of the top rules are shown in Table 10.

Table 10 Top rules according to confidence

| Rules | Confidence |
|---|---|
| {23, 6264, 1693} → {8411, 5720} | 1 |
| {23, 2795, 7526, 1693} → {8411, 5720} | 1 |
| {23, 6264, 1693} → {8411} | 1 |
| {23, 6264, 1693} → {5720} | 1 |

In Table 10, for the rules that have the same consequent, i.e., the 1st and 2nd rule, their antecedents can be combined to form a single valid rule. The 3rd and 4th rules are the subsets of the 1st rule and hence are the redundant rule. Therefore, confidence and conviction contained a combination of both types of rules. The redundancy for the confidence measure and the conviction measure was approximately 61%.

## 5. CONCLUSION

In this paper, the number of redundant and non-informative rules was determined using five interestingness measures, including confidence, conviction, cosine, leverage, and lift, for a constant value of support. The top rules are listed according to the respective values of the interestingness measures, and the number of redundant rules was determined. A rule is considered redundant if it is a subset of a valid rule. It was observed that leverage had the least redundant but also the least informative rules, as the antecedent or consequent could be combined to form a superset of the rule. Lift and cosine showed a similar type of top rules and showed the maximum redundant but more informative rules compared to other measures. Lift showed maximum redundant rules, as the valid rule that was an informative rule contained several subsets of the top rules. Confidence and conviction showed a similar type of top rules and contained rules that had valid rules that could be combined. This study experimentally confirmed that no measure is consistently better than others for all circumstances; however, there are situations in which many of these measures are highly correlated with each other. The presented algorithm was used to select a small set of rules in the form of tables so that experts can select the most appropriate measure by examining the small set of tables. The scope of present work was to identify redundant and non-informative rules from the total generated rules using interestingness measures and to compare the performance of the different measures used for the same purpose. This work can be utilized to identify the relationships among buying patterns of online users, decision making in effective e-marketing strategies, designing web-based personalized systems, and other types of patterns.

## 6. REFERENCES

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining Association Rules between Sets of Items in Large Databases. *In:* ACM SIGMOD Record, Volume 22(2), pp. 207–216

Ashrafi, M.Z., Taniar, D., Smith, K., 2004. A New Approach of Eliminating Redundant Association Rules. *In:* International Conference on Database and Expert Systems Applications, pp. 465–474

Azevedo, P.J., Jorge, M., 2007. Comparing Rule Measures for Predictive Association Rules. *In:* Machine Learning: ECML 2007, pp. 510–517

Brin, S., Motwani, R., Ullman, J.D., Tsur, S., 1997. Dynamic Itemset Counting and Implication Rules for Market Basket Data. *In:* ACM SIGMOD Record, Volume 26(2), pp. 255–264

Dimitrijević, M., Bošnjak, Z., Subotica, S., 2010. Discovering Interesting Association Rules in the Web Log Usage Data. *Interdisciplinary Journal of Information, Knowledge, and Management*, Volume 5, pp. 191–207

Han, J., Pei, J., Yin, Y., Mao, R., 2004. Mining Frequent Patterns without Candidate Generation: A Frequent-pattern Tree Approach. *Data Mining and Knowledge Discovery*, Volume 8(1), pp. 53–87

Li, M., Yu, X., Ryu, K.H., 2014. MapReduce-based Web Mining for the Prediction of Web-user Navigation. *Journal of Information Science,* Volume 40(5), pp. 557–567

Mobasher, B., Liu, B., 2007. Chapter 12: Web Usage Mining. *In:* Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, pp. 449–483

NASA_SeverLog, 1995. *NASA Kennedy Space Center's Www Server Log Data*, Available online at http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

Sisodia, D., Verma, S., 2011. Application of Weblogs to Construct Smart Web Servers to Handle User Traffic Efficiently. *International Journal of Advanced Computer Engineering and Architecture*, Volume 1(1), pp. 141–152

Sisodia, D.S., Khandal, V., Singhal, R., 2016. Fast Prediction of web User Browsing Behaviours Using the Most Interesting Patterns. *Journal of Information Science*, pp.1–19

Sisodia, D.S., Verma, S., Vyas, O.P., 2016a. A Conglomerate Relational Fuzzy Approach for Discovering Web User Session Clusters from Web Server Logs. *International Journal of Engineering and Technology*, Volume 8(3), pp. 1433–1443

Sisodia, D.S., Verma, S., Vyas, O., 2016b. A Discounted Fuzzy Relational Clustering of Web Users using Intuitive Augmented Sessions Dissimilarity Metric. *IEEE Access*, Volume 4(1), pp. 6883–6893

Sisodia, D.S., Verma, S., Vyas, O.P., 2015a. A Comparative Analysis of Browsing Behavior of Human Visitors and Automatic Software Agents. *American Journal of Systems and Software,* Volume 3(2), pp. 31–35

Sisodia, D.S., Verma, S., Vyas, O.P., 2015b. Agglomerative Approach for Identification and Elimination of Web Robots from Web Server Logs to Extract Knowledge about Actual Visitors. *Journal of Data Analysis and Information Processing*, Volume 3(2), pp. 1–10

Spark, 2015. *Python Programming Guide for Spark 0.9.0 Documentation*. Available online at https://spark.apache.org/docs/0.9.0/python-programming-guide.html, Accessed on December 25, 2015

Tan, P.N., Kumar, V., Srivastava, J., 2002. Selecting the Right Interestingness Measure for Association Patterns. *In:* Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '02, pp. 32–41

Zaki, M.J., 1999. Parallel and Distributed Association Mining: a Survey. *IEEE Concurrency,* Volume 7(4), pp. 14–25

Zaki, M.J., 2000. Generating Non-redundant Association Rules. *In:* Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 34–43

Zaki, M.J., 2004. Mining Non-redundant Association Rules. *Data Mining and Knowledge Discovery*, Volume 9(3), pp. 223–248

# Appendix

## Algorithm 1: Antecedent_redundancy_removal (list-of-rules)

**Input:** Top rules <list_of_rules>
**Output**: Set of valid rules

```
$antecedent = parse(list\_of\_rules)$
$consequent = parse(list\_of\_rules)$
for $idx_i\ from\ 0\ to\ length(antecendent)$
for $idx_j\ from\ 0\ to\ length(antecendent)$
if $idx_i\ != idx_j$
        $c = length(intersection($
            $antecendent[idx\_i], antecendent[idx\_j]))$
        if $c == length(antecedent[idx\_j])$
        if $consequent[idx\_i] == consequent[idx\_j]$
$removable\_index.add(idx\_j)$
        end
        end
end
end
end
return removable_index
```

## Algorithm 2: Consequent_redundancy_removal (list-of-rules)

**Input:** Top rules <list_of_rules> and removable index from antecedent_redundancy_removal ()
**Output:** Set of valid rules

```
$antecedent = parse(list\_of\_rules)$
$consequent = parse(list\_of\_rules)$
for $idx_i\ from\ 0\ to\ length(consequent)$
for $idx_j\ from\ 0\ to\ length(consequent)$
if $idx_i\ != idx_j$
        $c = length(intersection($
                $consequent\ [idx\_i], consequent\ [idx\_j]))$
        if $c == length(consequent\ [idx\_j])$
        if $antecedent\ [idx_i] == antecedent\ [idx_j]$
$removable\_index.add(idx\_j)$
        end
        end
end
end
end
return removable_index
```

## Algorithm 3: Consequent_redundancy_removal (list-of-rules)

**Input:** Top rules <list_of_rules>
**Output:** Set of indexes that contains valid rules

```
$removable\_index = Antecedent\_redundancy\_removal(list - of - rules)$
$removable\_index = Consequent\_redundancy\_removal(list - of - rules,\ removable\_index)$

$delete\,(removable\_index)$
```