Journal of Computer Science and Cybernetics, V.33, N.3 (2017), 247-262
DOI: 10.15625/1813-9663/33/2/10683

COLLABORATIVE RECOMMENDERATION BASED ON STATISTICAL IMPLICATION RULES

NGHIA QUOC PHAN[1], PHUONG HOAI DANG[2], HIEP XUAN HUYNH[3]

[1] *Testing Office, Tra Vinh University; nghiatvnt@gmail.com*
[2] *IT Faculty, Da Nang University of Science and Technology*
[3] *College of Information & Communication Technology, Can Tho University*

Abstract. In recent researches, many approaches based on association rules have been proposed to improve the accuracy of recommender systems. These approaches are primarily based on Apriori data mining algorithm in order to generate the association rules and apply them to improving the recommendation results. However, these approaches also reveal some disadvantages of the system, such as taking a longer time for generating association rules; applying the Apriori algorithm on rating sparse matrix resulting in irrelevant information and causing poor recommendation results to target users and association rules generated primarily relying on given threshold of Support and Confidence measures leading to the focus on the majority of rules and ignoring the astonishment of rules to affect the recommendation results. In this study, we propose a new model for collaborative filtering recommender systems: The collaborative recommendation is based on statistical implication rules (IIR); Differently from collaborative recommendation based on association rules (AR), the IIR predicts the items for users based on statistical implication rules generated from rating matrix and Implication intensity measures measuring the surprisingness of rules. To evaluate the effectiveness of the model, in the experimental section, we implement the model on three real datasets and compare the results with some different effective models. The results show that the IIR has higher precision on the experimental datasets.

Keywords. Statistical implication rules, association rules, collaborative filtering recommender system, statistical implicative analysis.

1. INTRODUCTION

The widespread popularity of high-tech communication devices, such as smartphones, tablets and other intelligent terminals has attracted the number of people accessing social networks (Facebook[1], Zalo, and Twitter), e-commerce sites (Amazon [10], Netflix [4] and eBay [12]) significantly. This has brought people closer together, exchanging information with each other is more convenient and the world becomes flatter. However, the huge amount of available information on social networks and a variety of items on e-commerce sites in recent times many difficulties for people when they search for the expected information. Sometimes, they cannot receive satisfactory results even the information available on the system. Fortunately, the transaction information and feedback of users can be tracked

[1] The Facebook website was launched on February 4, 2004, by Mark Zuckerberg, along with fellow Harvard College students and roommates, Eduardo Saverin, Andrew McCollum, Dustin Moskovitz, and Chris Hughes

and recorded on the social networks and e-commerce sites. This makes it easier to analyze the preference of users. Recommender systems [30] have been used to recommend items to users and provide personalized services by analyzing behaviors of users, such as the recommendation of the film and television programs in Netflix [4], the books in Amazon [5], and videos in YouTube [3].

To find out which items suit the user's preferences, many recommendation models have been proposed. In particular, the collaborative filtering recommendation model is considered to be a successful model in many areas. This model is based on a user's rating matrix for items in the past to recommend items for current users. However, the collaborative filtering recommendation model will give inaccurate recommendation results in cases of new users (user unrated for any item), new products (item is not rated by any user), data sparsity (each item is rated by only a few users; each user only rates for a few items). In order to solve this problem, many solutions have been proposed to improve recommendation results. The Application association rules is considered as one of the effective solutions to improve the accuracy of the collaborative filtering recommendation models [1, 23, 28]. In this way, the association rule can be applied to seeking interest patterns in the process data; exploiting implicit information such as users preferences and users behaviors based on user profiles and finding out the relationship between users and items based on rating matrix.

However, the collaborative recommendation based on association rules also reveals some disadvantages. For example, it takes the system a longer time to generate association rules; Applying the Apriori algorithm on rating sparse matrix results in irrelevant information can cause poor recommendation results to target users and association rules are generated based primarily on given threshold of Support and Confidence measures, so we concentrate on the majority of rules and ignore the astonishment of rules to affect the recommendation results.

This paper proposes a new model for collaborative filtering recommender systems based on statistical implication rules. In this model we built algorithm to generate statistical implication rules from rating matrix based on statistical implicative analysis theory. Hence, these rules are an asymmetry that reflects the implication among items based on rating values of users for items. There is a difference between collaborative recommendation based on association rules and collaborative recommendation based on statistical implication rules. The collaborative recommendation based on association rules mainly relies on transaction data of users and association rules are generated based on two metrics: Support and Confidence. The collaborative recommendation based on statistical implication rules is based on user's rating data and maximizes the implication of the relationship between items based on statistical implication rules. This helps the system improve the accuracy of the recommendation's results.

This paper is divided into eight parts. Part two presents a brief review of literature related to collaborative recommender systems based on association rules mining techniques. Part three presents statistical implication analysis method. Part four presents the definition of statistical implication rules. Part five describes the required steps to build the collaborative filtering recommender model based on statistical implication rules. Part six presents the experimental results of the model on three datasets. Part seven compare the accuracy of the model with other collaborative recommender models. The last part summarizes some important achieved results.

2. RELATED WORK

In this section, we present a brief review of literature related to collaborative recommender systems based on association rules mining techniques. The association rules mining technique was applied to representing users' interests in various fields for providing recommendation models due to its ability to scale large datasets and achieve high precision [24, 26, 27]. Firstly, association rules mining technique is considered as similarity measures to determine similarity between users or between items in recommendation system. Examples of this are as follows: association rules mining technique employed in a hybrid recommendation system to compute similarity among/between users from implicit data collected in a discussion group [15]; association rules mining used for identifying similar listening history among users on the same set of songs and predicting users' unknown preferences [18]; association rules mining technique employed to find out the similarities between readers, between articles, and combine them together to generate recommendation results [29]. Secondly, association rules mining technique is considered as data mining technique to process in recommendation system. For example, association rules mining technique is employed to discover similar interest patterns among users from implicit information in data processing phase and to find similar interest patterns between users in recommendation resulting phase [18]; association rules mining technique is employed to find frequent item sets in the class of Favorite Items to find out the correlation between items. Relying on correlation between items, the system suggests new items for a particular user [2]; Thirdly, association rules mining technique is considered as a recommender approach [9, 14], such as association rules mining technique used to generate recommendation results in collaborative recommender system [20, 29]; association rule mining technique and similarity measure used to generate recommendation results [25]; association rule mining technique and content-based approach used to generate recommendation results [2]; association rule mining technique and clustering algorithms used to generate recommendation results [18].

3. STATISTICAL IMPLICATION ANALYSIS METHOD

Statistical implicative analysis [22] is the method of data analysis studying implicative relationships between variables or data attributes, allowing detecting the asymmetrical rules a → b in the form "if *a* then that almost *b*" or "consider to what extent that *b* will meet implication of *a*". The purpose of this method is to detect trends in a set of attributes (variables) by using statistical implication measures.

Let E be a set of n objects or individuals described by a finite set of binary variables (property). A ($A \subset E$) is a subset of objects that meet the property $a; B(B \subset E)$ is a subset of objects that meet the property b; $\bar{A}$ (resp. $\bar{B}$) is the complement of A (resp. B); $n_A = \text{card}(A)$ is the number of elements of set $A; n_B = \text{card}(B)$ is the number of elements of set B; and the the counter-examples $n_{A\bar{B}} = \text{card}(A \cap \bar{B})$ is the number of objects that satisfy the attribute a but does not satisfy the property b. Let X and Y be two random sets with the number n_X and n_Y respectively.

For a certain process of sampling (see [22]), the random variable $\text{card}(X \cap \bar{Y})$ follows the Poisson distribution with the parameter $\lambda = \dfrac{n_A n_{\bar{B}}}{n}$.

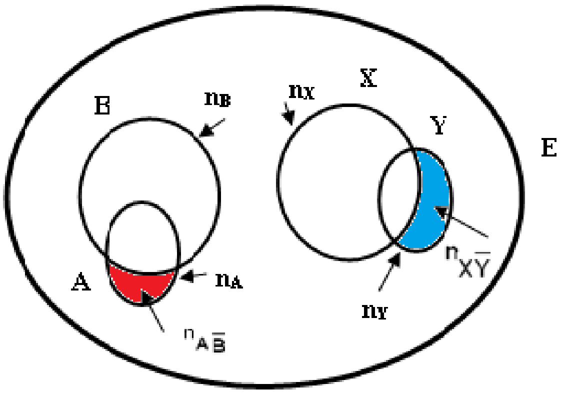


Figure 1. The model represents a statistical implication analysis method

The rule $a \to b$ is said to be admissible for a given threshold α if

$$Pr[\text{card}(X \cap \bar{Y}) \leq \text{card}(A \cap \bar{B})] \leq \alpha.$$

Let us consider the case where $n_{\bar{B}} \neq 0$. In this case, the Poisson random variable card$(X \cap \bar{Y})$ can be standardized random as

$$Q(A, \bar{B}) = \frac{\text{card}(X \cap \bar{Y}) - \dfrac{n_A(n - n_B)}{n}}{\sqrt{\dfrac{n_A(n - n_B)}{n}}}.$$

In experimental realization, the observed value $q(A, \bar{B})$ of $Q(A, \bar{B})$ is defined by

$$q(A, \bar{B}) = \frac{n_{A\bar{B}} - \dfrac{n_A(n - n_B)}{n}}{\sqrt{\dfrac{n_A(n - n_B)}{n}}}.$$

This value measures a deviation between the contingency and expected value when A and B are independent.

When the approximation is justified (e.g. $\lambda > 4$) the random variable $q(A, \bar{B})$ is approximatively $N(0, 1)$-distributed. The implication intensity $\varphi(a, b)$ of the rule $a \to b$ is defined by

$$\varphi(a, b) = 1 - Pr(Q(A, \bar{B}) \leq q(A, \bar{B})) = \begin{cases} \dfrac{1}{2\pi} \displaystyle\int_{q(A,\bar{B})}^{\infty} e^{-\dfrac{t^2}{2}} dt & \text{if } n_b \neq n, \\ 0 \quad \text{otherwise.} \end{cases}$$

This measures is used to determine the unlikehood of the counter-example $n_{A\bar{B}}$ in the set E. The implication intensity $\varphi(a, b)$ is admissible for a given threshold α if $\varphi(a, b) \geq 1 - \alpha$.

4. STATISTICAL IMPLICATION RULES

Statistical implication rules are asymmetric that have the form $a \to b$ and are selected based on statistical implication measures. Their important feature is the strong association between the properties of the left-hand side and the properties of the right-hand side. Based on this implication relationship to determine the role of individuals in the formation of rules [22].

Let $U = \{u_1, u_2, , u_n\}$ be a set of n users; $I = \{i_1, i_2, , i_m\}$ is a set of m items; $R = \{r_{j,k}\}$ is a rating matrix of n users for m items with each row representing a user u_j $(1 \leq j \leq n)$; each column represents an item i_k $(1 \leq k \leq m), r_{j,k}$ is the rating value of user u_j for item i_k; t_i is a set of items rated by u_i, t_j is a set of items rated by u_j, and t_i, $t_j \subseteq I$. A statistical implication rule is an implication of the form: $a \to b$ where $a \subseteq t_i$, $b \subseteq t_j$ and $a \cap b = \varnothing$; and is accepted with threshold α (0$\leq$ α $\leq$1) if $\varphi(a, b) \geq 1 - \alpha$.

5. THE COLLABORATIVE FILTERING RECOMMENDER MODEL BASED ON STATISTICAL IMPLICATION RULES

Suppose that $U = \{u_1, u_2, ..., u_n\}$ is a set of n users, $I = \{i_1, i_2, ..., i_m\}$ is a set of m items, $R_{\text{Train}} = \{r_{j,k}\}$ is a training dataset, with $r_{j,k}$ is the rating value of user u_j for item i_k; $R_{\text{Test}} = \{r_{i,l}\}$ is a testing dataset, with $r_{i,l}$ is the rating value of user u_i for item i_l; $SIR = \{r_1, r_2, ..., r_n\}$ is a set of statistical implication rules generated from training dataset; $MATRIX_{\text{Logical}} = \{l_{i,j}\}$ is logical matrix, with $l_{i,j}$ is logical value between rule r_i and user u_j, if user u_j has ratings for items of the left side of rule r_i then $l_{i,j}$=“TRUE”, otherwise $l_{i,j}$=“FALSE”; $SIR_{u_a} = \{r_{u_a}1, r_{u_a}2, ..., r_{u_a}k\}$ is a set of statistical implication rules selected for u_a. The recommendation results for user u_a is a set of items belonging to right side of SIR_{u_a} that user u_a has not rated $I_{u_a} = \{i_{u_a}1, i_{u_a}2, ..., i_{u_a}N\}$.

The collaborative filtering recommender model based on statistical implication rules is defined as the following phases: process experimental data, generate statistical implication rules from the training dataset, create predictive results, and evaluate the accuracy of the model. This model is presented by the following diagram.

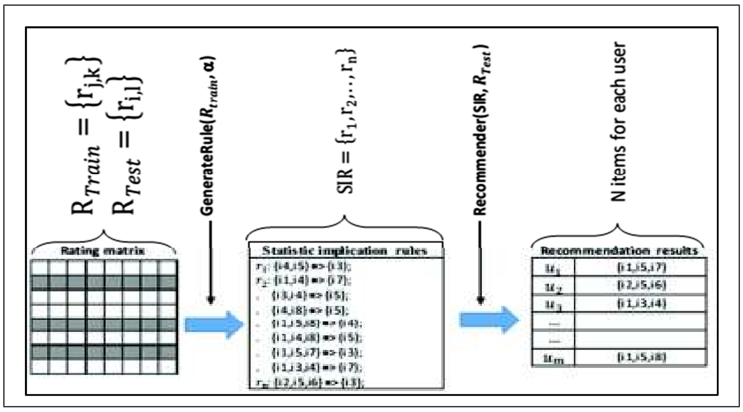
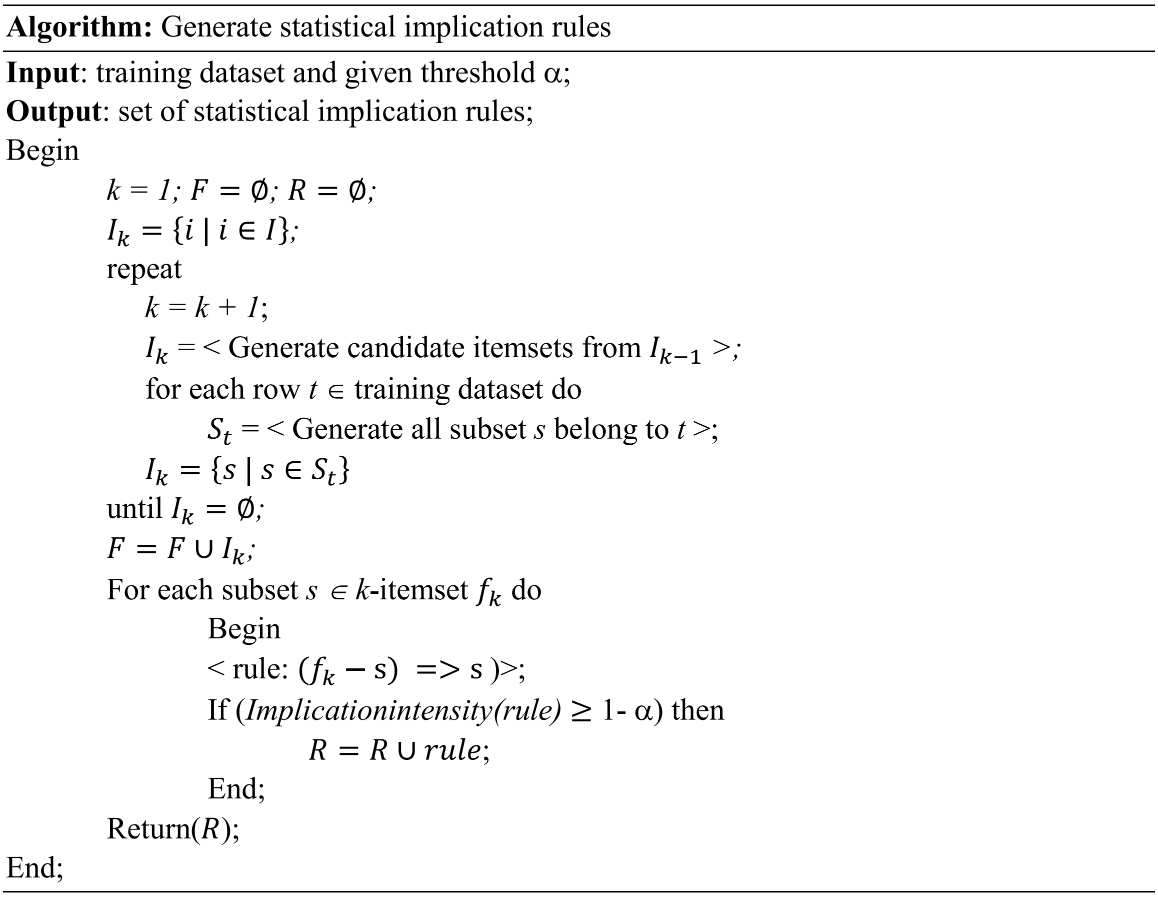


Figure 2. The collaborative filtering recommender model based on statistical implication rules

5.1. Generate statistical implication rules from the training dataset

Based on the definition of statistical implication rules presented in Section 3, the statistical implication rules is generated on the training dataset according to the following algorithm: First, the algorithm generates episodes from 1 element, set of 2 elements,..., set of k elements. Next, for each set of k elements generate its subsets (with subsets $\neq \phi$). Finally, based on subsets to generate statistical implication rules, and based on the threshold α and Implication intensity measures to choose the rules for the system.

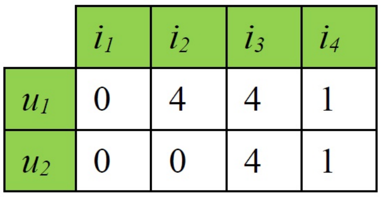


Algorithm: Generate statistical implication rules

Input: training dataset and given threshold α;
Output: set of statistical implication rules;

```
Begin
        $k = 1; F = \emptyset; R = \emptyset;$
        $I_k = \{i \mid i \in I\};$
        repeat
            $k = k + 1;$
            $I_k$ = < Generate candidate itemsets from $I_{k-1}$ >;
            for each row $t \in$ training dataset do
                  $S_t$ = < Generate all subset $s$ belong to $t$ >;
            $I_k = \{s \mid s \in S_t\}$
        until $I_k = \emptyset;$
        $F = F \cup I_k;$
        For each subset $s \in k$-itemset $f_k$ do
                  Begin
                  < rule: $(f_k - s) \Rightarrow s$ )>;
                  If ($Implicationintensity(rule) \geq 1 - \alpha$) then
                           $R = R \cup rule;$
                  End;
        Return($R$);
End;
```

Example: Let us a rating matrix of two users who rated for 4 items as follows:



	i_1	i_2	i_3	i_4
u_1	0	4	4	1
u_2	0	0	4	1

The first step, we generate candidate itemsets:

1-itemsets: $\{i_1 = 0\}, \{i_2 = 4\}, \{i_2 = 0\}, \{i_3 = 4\}, \{i_4 = 1\}$
2-itemsets: $\{i_1 = 0, i_2 = 4\}, \{i_1 = 0, i_2 = 0\}, \{i_1 = 0, i_3 = 4\}, \{i_1 = 0, i_4 = 1\}$
$\{i_2 = 4, i_2 = 0\}, \{i_2 = 4, i_3 = 4\}, \{i_2 = 4, i_4 = 1\}$
$\{i_2 = 0, i_3 = 4\}, \{i_2 = 0, i_4 = 1\}$
$\{i_3 = 4, i_4 = 1\}$
3-itemsets: $\{i_1 = 0, i_2 = 4, i_2 = 0\}, \{i_1 = 0, i_2 = 4, i_3 = 4\}, \{i_1 = 0, i_2 = 4, i_4 = 1\}$
$\{i_1 = 0, i_2 = 0, i_3 = 4\}, \{i_1 = 0, i_2 = 0, i_4 = 1\}$
$\{i_1 = 0, i_3 = 4, i_4 = 1\}$
$\{i_2 = 4, i_2 = 0, i_3 = 4\}, \{i_2 = 4, i_2 = 0, i_4 = 1\}$
$\{i_2 = 4, i_3 = 4, i_4 = 1\}$
$\{i_2 = 0, i_3 = 4, i_4 = 1\}$
4-itemsets: $\{i_1 = 0, i_2 = 4, i_2 = 0, i_3 = 4\}$
$\{i_1 = 0, i_2 = 4, i_2 = 0, i_4 = 1\}$
$\{i_1 = 0, i_2 = 4, i_3 = 4, i_4 = 1\}$
$\{i_1 = 0, i_2 = 0, i_3 = 4, i_4 = 1\}$
$\{i_2 = 4, i_2 = 0, i_3 = 4, i_4 = 1\}$
5-itemsets: $\{i_1 = 0, i_2 = 4, i_2 = 0, i_3 = 4, i_4 = 1\}$

The second step, we generate statistical implication rules with threshold $\alpha = 0.6$ and obtain 36 rules as follows:

Statistical implication rules
{V1=0} ⇒ {V2=4}, {V2=4} ⇒ {V1=0}, {V1=0} ⇒ {V4=1}, {V4=1} ⇒ {V1=0}, {V2=4} ⇒ {V4=1}, {V4=1} ⇒ {V2=4}, {V1=2} ⇒ {V2=0}, {V2=0} ⇒ {V1=2}, {V1=2} ⇒ {V4=0}, {V4=0} ⇒ {V1=2}, {V2=0} ⇒ {V4=0}, {V4=0} ⇒ {V2=0}, {V1=0,V2=4} ⇒ {V4=1}, {V1=0,V4=1} ⇒ {V2=4}, {V2=4,V4=1} ⇒ {V1=0}, {V1=0,V3=4} ⇒ {V2=4}, {V2=4,V3=4} ⇒ {V1=0}, {V1=0,V3=4} ⇒ {V4=1}, {V3=4,V4=1} ⇒ {V1=0}, {V2=4,V3=4} ⇒ {V4=1}, {V3=4,V4=1} ⇒ {V2=4}, V1=2,V2=0} ⇒ {V4=0}, {V1=2,V4=0} ⇒ {V2=0}, {V2=0,V4=0} ⇒ {V1=2}, {V1=2,V3=4} ⇒ {V2=0}, {V2=0,V3=4} ⇒ {V1=2}, {V1=2,V3=4} ⇒ {V4=0}, {V3=4,V4=0} ⇒ {V1=2}, {V2=0,V3=4} ⇒ {V4=0}, {V3=4,V4=0} ⇒ {V2=0}, {V1=0,V2=4,V3=4} ⇒ {V4=1}, {V1=0,V3=4,V4=1} ⇒ {V2=4}, {V2=4,V3=4,V4=1} ⇒ {V1=0}, {V1=2,V2=0,V3=4} ⇒ {V4=0}, {V1=2,V3=4,V4=0} ⇒ {V2=0}, {V2=0,V3=4,V4=0} ⇒ {V1=2}

5.2. Create recommendation results

The recommendation results are generated based on statistical implication rules set and testing dataset by the following algorithm: First, the algorithm is based on the number of statistical implication rules and the number of users in the testing set to produce a logical matrix with n rows, m columns (n is the number of rules, m is the number of users). This matrix is created as follows: if user j has ratings for items of the left side of rule i and then cell $i; j$ of the matrix is assigned the value "TRUE", otherwise assigns the value "FALSE". Next, for each column j, if the value of cell i, j is "TRUE", then select the corresponding statistical implication rule in line i for user j. Finally, ascendingly sort the selected rules of each user by the value of the statistical implication intensity and select the N items from the right-hand side of rules with the highest possible value that the user has not yet rated for recommendation results.

```
Algorithm: Generate recommendation results
Input: set of statistical implication rules and testing dataset;
Output: recommendation result matrix;
Begin
        $n$ = Number of statistical implication rules;
        $m$ = Number of users in testing dataset;
           $MATRIX_{Logical} = \begin{array}{|c|c|c|c|c|} \hline & u_1 & u_2 & \dots & u_m \\ \hline r_1 & \text{TRUE} & \text{FALSE} & & \text{TRUE} \\ \hline r_2 & \text{FALSE} & \text{TRUE} & & \text{FALSE} \\ \hline . & & & & \\ \hline . & & & & \\ \hline r_n & \text{TRUE} & \text{FALSE} & & \text{TRUE} \\ \hline \end{array}$
        For ($i=0; i<n; i$++) do
                For ($j=0; i<m; j$++) do
                        If ($Set\ of$ items of the left side of rule $on\ row\ i$) $\cap$
                         ($Set\ of\ items\ rated\ by\ user\ on\ column\ j$) $\neq \emptyset$ ()) then
                                $MATRIX_{Logical}[i,j] = "TRUE"$;
                        Else
                                $MATRIX_{Logical}[i,j] = "FALSE"$;
        For ($j=0, j<m; j$++) do
                For ($i=0, i<n; i$++) do
                        If($MATRIX_{Logical}[j,i]$ = "$TRUE$") then
                                < Choose statistical implication rules on row $i$ for user $j$>;
        For ($j=0, j<m; j$++) do
          Begin
                < Sort ascending the rules according to the value of Implication intensity>;
                Result[j] = <N items from the right side of the rules with the highest value
                of Implication intensity that user $j$ has not rated >;
          End;
        MatrixResult = matrix(Result, $N$, $m$);
        Return(MatrixResult);
End;
```

5.3. Evaluate the accuracy of the model

To test the accuracy of the model, we use the evaluation method based on model recommendations [7, 8]. This method evaluates the accuracy of the model by comparing the model's recommendations with the choice of users purchase by using three metrics: *Precision*, *Recall*, and *Fmeasure* to measure the accuracy of the model [11, 19]. The model is evaluated good if three indices gain high value. The following is a formula for calculating the value of measures

$$Precision = \frac{Correctlyrecommendeditem}{Totalrecommendeditems},$$

$$Recall = \frac{Correctlyrecommendeditems}{Totalusefulrecommendations},$$

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall}.$$

For example, to see more clearly the process of implementation of the model, assuming that the system recommends users to select 8 items (from i_1 to i_8) and the system has 10 users (from u_1 to u_{10}) with rating matrix as follows

u/i	$i1$	$i2$	$i3$	$i4$	$i5$	$i6$	$i7$	$i8$
$u1$	0	1	1	1	1	1	0	0
$u2$	1	0	0	0	1	1	0	0
$u3$	1	0	0	1	1	1	0	1
$u4$	1	0	1	1	1	0	1	1
$u5$	1	1	0	0	0	0	0	1
$u6$	0	1	0	0	0	1	0	1
$u7$	1	1	0	1	0	1	1	0
$u8$	1	0	0	1	0	1	0	0
$u9$	0	1	1	0	0	0	1	0
$u10$	1	0	1	0	0	0	0	1

In particular, the training dataset consists of the first seven users (u_1 to u_7); the testing dataset contains the last three users (u_8 to u_{10}), with each user selecting two items to recommend. The first step, the model generates statistical implication rules based on training dataset with $\alpha = 0.5$ (49 rules). The next step, it relies on sets of the statistical implication rules and testing dataset to build a logical matrix. The final step, from the logical matrix, the model generates the recommendation results as follows Table 1.

6. EXPERIMENT

6.1. Process experimental data

To evaluate the effectiveness and performance of our proposed model, we use k-fold crossvalidation method [19] with $k = 5$. k-fold cross validation refers to dividing the examples into k equally sized subsets and using one subset for testing and the rest for training. This is done repeatedly so that each subset acts as a test set once and is part of the training set $k - 1$ times. The evaluation results of this method are average value of k evaluations.

6.2. Datasets

The three datasets of MovieLense [6], MSWeb [13] and Jester5k [16] are used in our experiments. First, MovieLense dataset is collected from rating results of 943 users for

Table 1

Users	**Statistic implication rules**	**Implication intensity**	**Recommendation items**
u8	{i4,i5} ⇒ {i3}; {i1,i4} ⇒ {i7}; {i3,i4} ⇒ {i5}; {i4,i8} ⇒ {i5}; {i1,i5,i8} ⇒ {i4}; {i1,i4,i8} ⇒ {i5}; {i1,i4,i5} ⇒ {i8}; {i4} ⇒ {i3}; {i4} ⇒ {i7}; {i4} ⇒ {i5}; {i1,i3} ⇒ {i7}; {i3,i4,i8} ⇒ {i7}; {i4,i7,i8} ⇒ {i3}; {i1,i3,i8} ⇒ {i7}; {i1,i7,i8} ⇒ {i3}; {i4,i5,i7} ⇒ {i3}; {i1,i3,i5} ⇒ {i7}; {i1,i5,i7} ⇒ {i3}; {i1,i3,i4} ⇒ {i7}; {i2,i4,i5} ⇒ {i3}; {i2,i5,i6} ⇒ {i3}; {i1,i2,i4} ⇒ {i7}; {i1,i2,i6} ⇒ {i7}; {i3,i4,i5,i8} ⇒ {i7}; {i4,i5,i7,i8} ⇒ {i3}; {i1,i3,i5,i8} ⇒ {i7}; {i1,i5,i7,i8} ⇒ {i3}; {i1,i3,i4,i8} ⇒ {i7}; {i1,i4,i7,i8} ⇒ {i3}; {i1,i3,i4,i5} ⇒ {i7}; {i1,i4,i5,i7} ⇒ {i3}; {i2,i4,i5,i6} ⇒ {i3}; {i1,i2,i4,i6} ⇒ {i7}; {i1,i3,i4,i5,i8} ⇒ {i7}; {i1,i4,i5,i7,i8} ⇒ {i3}	[0.631238, 0.510458]	i3, i7
u9	{i3,i4} ⇒ {i5}; {i3} ⇒ {i5}; {i3} ⇒ {i4}; {i7} ⇒ {i4}; {i3,i5} ⇒ {i4}; {i1,i7} ⇒ {i4}; {i1,i3} ⇒ {i7}; {i3,i4,i8} ⇒ {i7}; {i4,i7,i8} ⇒ {i3}; {i1,i3,i8} ⇒ {i7}; {i1,i7,i8} ⇒ {i3}; {i4,i5,i7} ⇒ {i3}; {i1,i5,i7} ⇒ {i3}; {i1,i3,i4} ⇒ {i7}; {i2,i4,i5} ⇒ {i3}; {i2,i5,i6} ⇒ {i3};{i1,i2,i4} ⇒ {i7}; {i1,i2,i6} ⇒ {i7}; {i3,i4,i5,i8} ⇒ {i7}; {i4,i5,i7,i8} ⇒ {i3}; {i1,i3,i5,i8} ⇒ {i7}; {i1,i5,i7,i8} ⇒ {i3}; {i1,i3,i4,i8} ⇒ {i7}; {i1,i4,i7,i8} ⇒ {i3}; {i1,i3,i4,i5} ⇒ {i7}; {i1,i4,i5,i7} ⇒ {i3}; {i2,i4,i5,i6} ⇒ {i3}; {i1,i2,i4,i6} ⇒ {i7}; {i1,i3,i4,i5,i8} ⇒ {i7}; {i1,i4,i5,i7,i8} ⇒ {i3}; {i3,i8} ⇒ {i7}; {i7,i8} ⇒ {i3}; {i5,i7} ⇒ {i3}; {i3,i5,i8} ⇒ {i7}; {i5,i7,i8} ⇒ {i3}; {i2,i5} ⇒ {i3}	[0.575627, 0.510458]	i5, i4
u10	{i1,i4} ⇒ {i7}; {i3,i4} ⇒ {i5}; {i3} ⇒ {i5}; {i3} ⇒ {i4}; {i3,i5} ⇒ {i4}; {i4,i8} ⇒ {i5}; {i1,i5,i8} ⇒ {i4}; {i1,i4,i8} ⇒ {i5}; {i1,i4,i5} ⇒ {i8}; {i5,i8} ⇒ {i4}; {i1,i3} ⇒ {i7}; {i3,i4,i8} ⇒ {i7}; {i4,i7,i8} ⇒ {i3}; {i1,i3,i8} ⇒ {i7}; {i1,i7,i8} ⇒ {i3}; {i1,i5,i7} ⇒ {i3}; {i1,i3,i4} ⇒ {i7}; {i1,i2,i4} ⇒ {i7}; {i1,i2,i6} ⇒ {i7}; {i3,i4,i5,i8} ⇒ {i7}; {i4,i5,i7,i8} ⇒ {i3}; {i1,i3,i5,i8} ⇒ {i7}; {i1,i5,i7,i8} ⇒ {i3}; {i1,i3,i4,i8} ⇒ {i7}; {i1,i4,i7,i8} ⇒ {i3}; {i1,i3,i4,i5} ⇒ {i7}; {i1,i4,i5,i7} ⇒ {i3}; {i1,i2,i4,i6} ⇒ {i7}; {i1,i3,i4,i5,i8} ⇒ {i7}; {i1,i4,i5,i7,i8} ⇒ {i3}; {i3,i8} ⇒ {i7}; {i7,i8} ⇒ {i3}; {i3,i5,i8} ⇒ {i7}; {i5,i7,i8} ⇒ {i3}; {i1,i3,i5} ⇒ {i7}	[0.631238, 0.510458]	i7, i5

1,664 movies through the MovieLens website (movielens.umn.edu) during 7 months (from 09/19/1997 to 22/04/1998). This dataset is organized in a matrix format consisting of 943 rows, 1.664 columns, and 1.569.152 cells containing rating values. Of which there are more than 93 percent cells having rating values equal 0 and nearly 7 percent remaining cells having rating values from 1 to 5. Second, MSWeb dataset is a dataset of users visiting Microsoft sites during one week in February 1998. It is sampled and processed from the log file of the address www.microsoft.com [13]. This dataset includes 38,000 anonymous users getting access to 285 original web addresses, and is processed and organized into binary rating matrix with 32,710 rows, 285 columns, and 98,653 rating values. Third, Jester5k dataset is obtained from rating results of 5000 users through Jester Online Joke Recommender System in the period from April 1999 to May 2003. This dataset is organized in a matrix format consisting of 5,000 rows, 100 columns, and 362.106 rating values. In particular, each user has rated at least 36 jokes. The rating values for jokes are real values ranging from -10.00 to 10.00 and the value "99" corresponds to "null". It means that a user does not rate for jokes.

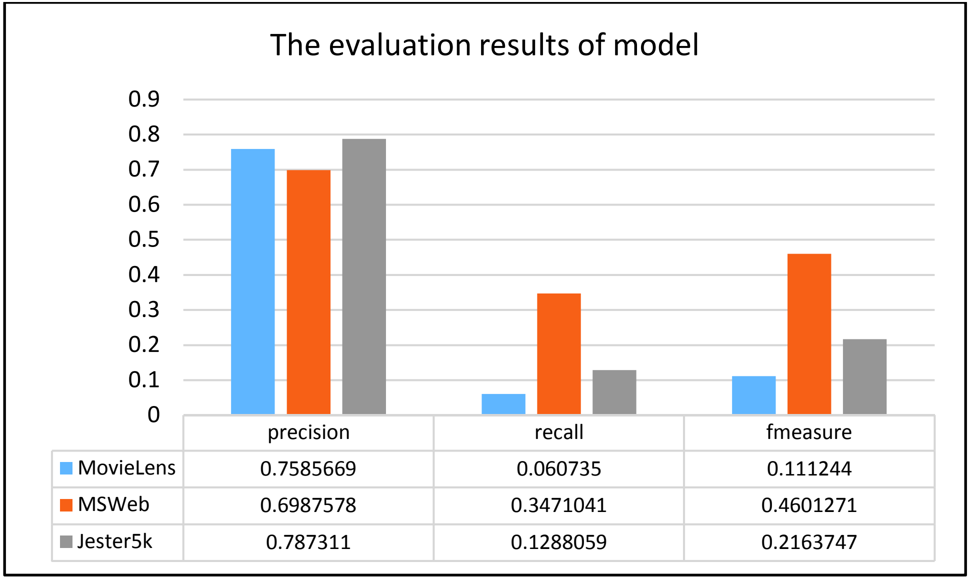
For evaluating the accuracy of the model, each dataset is divided into two parts, 20 percent used for testing and the rest used for training. Since MovieLens is numerical rating matrix and Jester5k real rating matrix, we transform them into binary format before applying the model. With MovieLens, if the rating value is greater than or equals 3 on numerical rating matrix, then transform equals 1 on binary matrix, otherwise transform equals 0. With Jester5k, if the rating value is greater than or equals 0 on real rating matrix, then transform equals 1 on binary matrix, otherwise transform equals 0.

6.3. Experimental tools

In our experiments, we use ARQAT tool [17] which is developed on language R by our team. This is a tool developed for implementing recommender models based on statistical implication analysis. It includes the following functions: processing data, generating statistical implication rules, designing and evaluation recommender models [19].

6.4. The results of model

In the experiment, we examine models on three different datasets respectively. Then, we evaluate the accuracy of the model based on its recommendation results on each experimental dataset. The evaluation results based on three metrics $Precision, Recall$ and $Fmeasure$ on three experimental datasets are presented in Figure 3. The Figure 3 shows that the recommendation results on MSWeb dataset are higher than the recommendation results on the other two datasets: Jester5k and MovieLens. In particular, the values of *Recall* and Fmeasure on MSWeb are much higher than the value of these metrics on the two remaining datasets. This result reflects that the IIR produces good recommendation results on binary dataset. The Figure 3 also shows that the value of *Precision* of the model is higher than those of *Recall* and $Fmeasure$ on three experimental datasets. This means that the recommendation results of the model are relatively good and the ability to find the recommendation items of the model is not really effective.



	precision	recall	fmeasure
🟦 MovieLens	0.7585669	0.060735	0.111244
🟧 MSWeb	0.6987578	0.3471041	0.4601271
◼ Jester5k	0.787311	0.1288059	0.2163747

Figure 3. The evaluation results of the model on three experimental datasets (with $\alpha = 0.5$)

7. COMPARE THE ACCURACY OF THE MODEL WITH OTHER COLLABORATIVE RECOMMENDER MODELS

In order to evaluate the effectiveness of the model, we compare the accuracy of the proposed model with the precision of the collaborative filtering recommender models: User-based collaborative filtering (UBCF) [19], Item-based collaborative filtering (IBCF) [19] and Collaborative filtering recommender based on association rules mining techniques (AR) [21, 29]. To accomplish this, we conduct experiments in the models with the same training dataset and testing dataset on all three experiment datasets. After that, we evaluate the models in each experimental dataset. The evaluation results of the models on three datasets are shown in Table 2.

The evaluation results of the models show that the value of *Precision* of IIR is superior to the rest on three datasets (MovieLense: UBCF: 0.5364286, IBCF: 0.5342857, AR: 0.6951952, **IIR: 0.7585669;** MSWeb: UBCF: 0.4245, IBCF: 0.40125, AR: 0.5010753, **IIR: 0.6987578;** Jester5k: UBCF: 0.75028, IBCF: 0.44376, AR: 0.7794214, **IIR: 0.787311**). This result shows that the IIR has practical applicability. In particular, on binary data (MSWeb), the model has far more precision than the rest of the models. Figure 4 shows the results of the comparison of the indicators of the accuracy of the models.

To compare the accuracy of IIR with UBCF, IBCF and AR, we build ROC chart to present the evaluation results (plot the *Precision* and *Recall* ratios for each model) on MSWeb dataset. In particular, we give the parameter specifying the number of items to recommendation varies from 1 to 10 (each user is recommended from 1 to 10 websites). The models have the following parameters: The UBCF and IBCF models use Cosine as similarity measure, the AR model selects the association rules with $Support = 0.3$ and $Confidence = 0.5$, the IIR selects the parameter $\alpha = 0.5$. Figure 5 shows that IIR has the best results

Table 1. The evaluation results of the models on three datasets (Average result of k-fold = 5)

Datasets	Models	Precision	Recall	Fmeasure
MovieLens	UBCF	0.5364286	0.0854183	0.1473701
	IBCF	0.5342857	0.0851085	0.1468281
	AR	0.6951952	0.0518046	0.0979258
	IIR	0.7585669	0.0607350	0.1112440
MSWeb	UBCF	0.4245000	0.5299460	0.4713982
	IBCF	0.4012500	0.5019791	0.4459979
	AR	0.5010753	0.1179903	0.2747900
	IIR	0.6987578	0.3471041	0.4601271
Jester5k	UBCF	0.7502800	0.2685480	0.3955254
	IBCF	0.4437600	0.1306010	0.2018086
	AR	0.7794214	0.1259575	0.2133795
	IIR	0.7873110	0.1288059	0.2163747

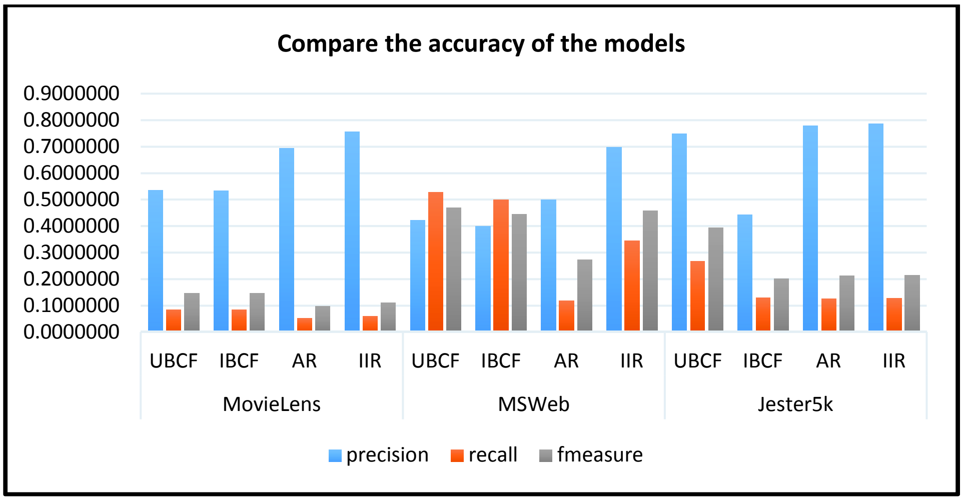


Figure 4. Compare the accuracy of the models: UBCF, IBCF, AR, and IIR

in the four models studied. This confirms again that IIR gives good results on binary datasets. The result shows that statistical implication rules have increased the accuracy of the collaborative filtering model compared to the association rules. The main reason for this increase is that statistical implication rules are generated based on the implicative value between items instead of Support and Confidence as association rules.

8. CONCLUSION

This paper presents the research on recommender systems based on association rules and analyses of some disadvantages of this approach. To overcome the shortcomings, we propose a new model for collaborative filtering recommender systems based on statistical

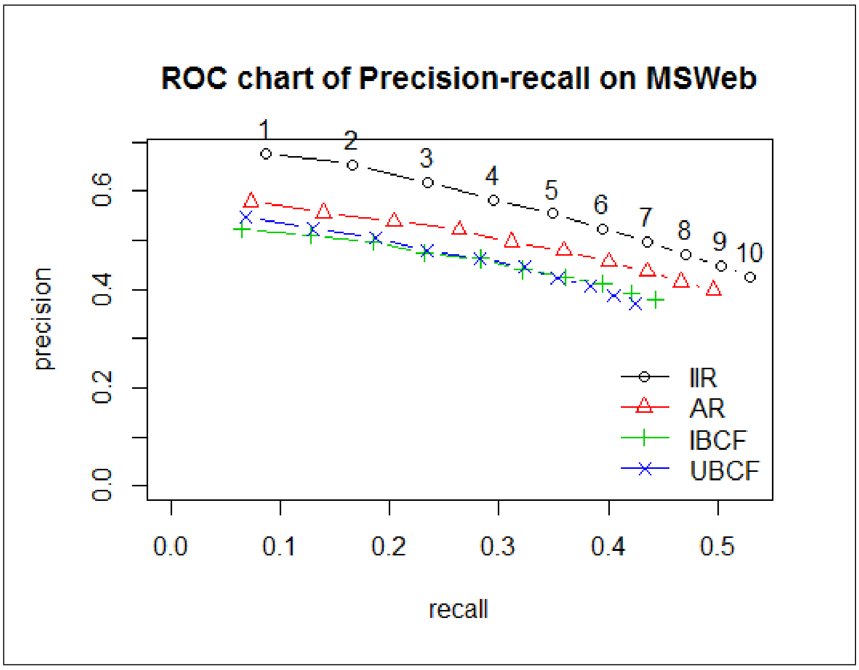


Figure 5. Compare the ROC chart of Precision/Reall of the models: UBCF, IBCF, AR, and IIR

implication rules. The article also defines how to determine a statistical implication rule for a given threshold based on statistical implicative analysis and describes the steps of building the IIR. To demonstrate the effectiveness of the proposed model, the experimental part was carried out on three datasets MovieLens, MSWeb, and Jester5k. The experimental results show that the recommendation results of the proposed model are quite accurate on MSWeb dataset. The evaluation results of the models on three datasets show that the value of Precision of IIR is higher than those of the other models. This result is a testament to the applicability in practice to improve the accuracy of the recommendation results.

REFERENCES

[1] A. Kumar, "A fast and new collaborative web recommendation system using fast adaptive association rule mining," *International Journal of Computer Science and Information Technologies*, vol. 5, no.6, pp. 6992-6995, 2014.

[2] Ahmed Mohammed K. Alsalama, "A hybrid recommendation system based on association rules," Engineering and Technology *International Journal of Computer, Electrical*, Automation, Control and Information Engineering, vol. 9, no.1, pp. 55-62, 2015.

[3] B. Shumeet, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, M. Aly, "Video suggestion and discovery for YouTube: taking random walks through the view graph," in: *International Conference on World Wide Web*, pp. 895-904, 2008.

[4] Carlos A. Gomez-Uribe and Neil Hunt, "The netflix recommender system: algorithms, business value, and innovation," *ACM Transactions on Management Information Systems*, vol. 6, no. 4, Article 13, pp. 1-19, 2015.

[5] E. Brynjolfsson, Y.J. Hu, M.D. Smith, "Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers manage," *Sci.*, vol. 49, no.11, pp.1580–1596, 2003.

[6] F. Maxwell Harper and Joseph A. Konstan, "The movielens datasets: history and context," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no.4, Article 19, pp. 1–19, 2015.

[7] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh, "Recommendation systems: principles, methods and evaluation," *Egyptian Informatics Journal*, pp. 261–273, 2015.

[8] Feng Zhang, Ti Gong, Victor E. Lee, Gansen Zhao, Chunming Rong, and Guangzhi Qu, "Fast algorithms to evaluate collaborative filtering recommender systems," *Knowledge-Based Systems*, volume 96, pp. 96–103, 2016.

[9] Gabroveanu Mihai, "Recommendation system based on association rules for distributed e-learning management systems," *ACTA Universitatis Cibiniensis*, [Online]. DOI: https://doi.org/10.1515/aucts-2015-0072, 2015.

[10] Greg Linden, Brent Smith, and Jeremy York, "Amazon.com recommendations item-to-item collaborative filtering," *IEEE Computer Society*, pp.76–80, 2003.

[11] JL. Herlocker, JA. Konstan, LG. Terveen and JT. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, ISSN 1046-8188, pp. 5–53, 2004.

[12] J. Ben Schafer, Joseph A. Konstan, and John Riedl, "E-commerce recommendation applications," *Data Mining and Knowledge Discovery*, vol. 5, pp. 115153, 2001.

[13] Jack S. Breese, David Heckerman and Carl M. Kadie, "Anonymous web data from www.microsoft.com," *Microsoft Research, Redmond WA*, 98052-6399, USA. [Online]. Available: https://kdd.ics.uci.edu/databases/msweb/msweb.html, 1998.

[14] JinHyun Jooa, SangWon Bangb, and GeunDuk Parka, "Implementation of a recommendation system using association rules and collaborative filtering, " *Procedia Computer Science*, volume 91, pp. 944–952, 2016.

[15] A. A. Kardan, and M. Ebrahimi, "A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups," *Information Sciences*, vol. 219, pp. 93–110, 2013.

[16] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins, "Eigentaste: A constant time collaborative filtering algorithm," *Information Retrieval*, vol. 4, no. 2, pp. 133–151, 2001.

[17] Lan Phuong Phan, Nghia Quoc Phan, Ky Minh Nguyen, Hung Huu Huynh, Hiep Xuan Huynh, and Fabrice Guillet, "Interestingnesslab: A framework for developing and using objective interestingness measures," in *ICTA 2016: International Conference on Advances in Information and Communication Technology*, pp. 302–311, 2016.

[18] Maryam Khanian Najafabadi, Mohd Naz'ri Mahrin, Suriayati Chuprat, and Haslina Md Sarkan, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data, " *Computers in Human Behavior*, volume 67, pp. 113–128, 2017.

[19] Michael Hahsler, “Lab for developing and testing recommender algorithms,” Copyright (C) Michael Hahsler (PCA and SVD implementation (C) Saurabh Bathnagar). [Online]. Available: http://R-Forge.R-project.org/projects/recommenderlab/, 2015.

[20] N. Bendakir, and E. A. Imeur, “Using association rules for course recommendation,” *Proceedings of the AAAI Workshop on Educational Data Mining*, pp. 31-40, 2006.

[21] Phan Quoc Nghia, Nguyen Minh Ky, Nguyen Tan Hoang, Huynh Xuan Hiep, “Recommender system based on statistical implicative analysis,” *Proceedings of the 8th National Conference on Fundamental and Applied Information Technology Research (FAIR’8)*, ISBN: 978-604-913-397-8, pp. 297–308, 2015

[22] R. Gras and P. Kuntz, “An overview of the Statistical Implicative Analysis (SIA) development,” *Statistical Implicative Analysis Studies in Computational Intelligence*, vol. 127, Springer-Verlag, pp. 11–40, 2008.

[23] Rakesh Agrawal and Ramakrishnan Srikant, “Fast algorithms for mining association rules,” *Proc. of the 20th VLDB Conference*, pp. 487-499, 1994.

[24] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, “Mining association rules between sets of items in large databases,” *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 207-216, 1993.

[25] T.Chellatamilan, and R. Suresh, “An e-learning recommendation system using association rule mining technique,” *European Journal of Scientific Research*, vol. 64, no. 2, pp. 330–339, 2011.

[26] Thang Mai, Bay Vo, and Loan T.T. Nguyen, “A lattice-based approach for mining high utility association rules,” *Information Sciences*, volume 399, pp. 81–97, 2017.

[27] Tyagi, S., & Bharadwaj, K. K, “Enhancing collaborative filtering recommendations by utilizing multi-objective particle swarm optimization embedded association rule mining,” *Swarm and Evolutionary Computation*, vol. 13, pp. 1–12, 2013.

[28] Ujwala H. Wanaskar, Sheetal R. Vij, Debajyoti Mukhopadhyay, “A hybrid web recommendation system based on the improved association rule mining algorithm,” *Journal of Software Engineering and Applications*, vol. 6, pp. 396–404, 2013.

[29] Weiyang Lin, Sergio A. Alvarez, Carolina Ruiz, “Collaborative recommendation via adaptive association rule mining,” *Proceedings L in 2000 Collaborative RV*, pp.1–7, 2000.

[30] Xiaoyuan Su and Taghi M. Khoshgoftaar, “A survey of collaborative filtering techniques,” *Hindawi Publishing Corporation, Advances in Artificial Intelligence*, vol. 2009, Article ID 421425, pp. 1–9, doi:10.1155/2009/421425, 2009.

Received on September 18, 2017
Revised on December 17, 2017