Association Rule Clustering Technique For User Interest Mining From Social Networking Sites

R. Umamaheswari[1]*, Dr.M. Soranamageswari[2]

[1]PhD Research Scholar, PG & Research Department of Computer Science, Government Arts College(Autonomous), Coimbatore-18.

[2]Assistant Professor, PG & Department of Information Technology, Government Arts College(Autonomous), Coimbatore-18.

*Corresponding Author: R. Umamaheswari

Abstract
Recommendation systems based on customers' interests have been broadly emerged because of a growing number of web users and online services. Over the past decades, Association Rule Mining (ARM) was applied to extract user interests from online user profiles. In addition, hybrid Competitive Swarm Optimizer and Gravitational Search Algorithm (CSO-GSA) were utilized to choose the most relevant terms for generating the rules. However, extracting the undesirable pattern of user interest needs a prior knowledge and traditional belief model which was a time-consuming process. Hence in this article, a new framework is designed that can able to automatically identify beliefs from data and expose undesirable patterns. In this framework, a clustering technique is introduced to discover undesirable patterns in the association rules. To cluster the association rules, the non-redundant association rules between user interests are defined as the numerical feature vectors before clustering. This representation is extended to summarize and visualize the association rules and their relationships. Also, a method is proposed for pruning redundant rules depending on the association among items. According to this, the users are able to analyze and choose remarkable rules interactively. Then, the non-redundant rules are classified into 2 classes: (i) beliefs and (ii) possible candidates for undesirable rules. Further, a contradiction check is applied to expose the accurate undesirable rules from the candidates. Finally, the experimental results show that the presented method achieves higher efficiency compared to the state-of-the-art methods for user interest mining.

Keywords-User interest mining, Association rules, CSO-GSA, Undesirable patterns, Belief system, Pruning, Clustering

I. Introduction

Twitter and LinkedIn have developed as important online social network platforms for real-time communications throughout the web [1]. Such social networking sites have many customers as well. The prevalence of a vast quantity of data exchange and updating on a wide range of subjects is appealing to online networking clients who are engaged in certain concepts [2]. Interests and passions reflect the cognitive characteristics of a certain individual who has an intelligence predisposition to deliberately explore the needs [3-5]. In general, an individual's cognitive interests have been used for a generalized notion that incorporates the different concepts of cognitive science, such as unexpected extent and pleasure.
The client's interests contain detailed knowledge on the customer's background, rewards, social skills, cognitive functioning, attitude, and habits. Such findings are extremely important in the fields of pedagogy and cognitive science [6]. The association rules identify the online networking user's interest, which is then exploited in numerous researches. During the last century, analysts have been focusing their efforts on identifying user profiles on social networking sites. In pedagogy, learning the client's interest is a complex issue. Many scholars completely discussed the role of user interest in interpersonal cognitive growth and so on. Numerous researches in cognitive science have shown that interest plays an important role in economic growth, likability growth, and emotional health. The user interest analysis reveals that people are interested in a variety of genres [7]. With the steady growth of apps regarding social websites and online clients in the present decade, the predictive model is formed with the help of interests derived from the online community, which is extensively used in practical situations [8]. The data collected from the networking site displays the customer's interests and experiences, which is then used to build the suggestion of structured products. The recommender systems achieve successful profitability, item suggestions, and online service [9].

Although studies on customer interests are widespread, the earlier studies are unsuccessful in identifying the association between the product and interests relying on information collection [10-12]. The information acquired from social networks contains a plethora of special characters and other short words. From this perspective, Si et al. [13] developed and analysed different hypotheses about the interests of online site customers. After that, the association rules are used to extract customers' interests from LinkedIn customers' profiles. Also, a technique is designed to extract interests for Twitter customers depending on the interest association rules and customer interest distribution on Twitter. But, an efficient pre-processing and feature selection technique is needed to get more relevant users' interest when the amount of information in the dataset is huge. So, an efficient pre-processing and feature selection strategy is introduced to enhance the more appropriate keywords for extracting customer interest [14]. The feature selection approach is quite useful in reducing the dimensionality of information collected from Twitter and LinkedIn. First, the pre-processing is achieved by stemming, lemmatization, acronyms, and negation replacement. Then, the data is utilized to generate rules using ARM. It successfully handles the occurrence of high dimensional features, which in turn generates effective rules. The interests of online networking customers are recognized, and uncertain data is removed.

The optimization strategy such as the Competitive Swarm Optimizer (CSO) in conjunction with the Gravitational Search Algorithm (GSA) is successful in retrieving optimal features. The optimized characteristics obtained are used to generate the association rules. The association rule's interest is used to get Twitter and LinkedIn user interest. On the other hand, extracting undesirable patterns enables to detection of a failure in prior knowledge and may recommend a data feature that requires additional analysis. The undesirable patterns are often extracted using belief-driven approaches, which need an established set of beliefs. Many techniques have constructed the individual partial beliefs explicitly, but these are time-consuming.

Therefore, this study proposes a novel framework capable of automatically identifying beliefs from data and exposing undesirable behaviours. A clustering approach is included in this framework to find undesired patterns in the association rules. Before clustering the association rules, non-redundant association rules between user interests are formed as numerical feature vectors. This representation is expanded to describe and understand the association rules and their interactions. In addition, a strategy for pruning repetitive rules based on item correlation is presented. According to this, customers may constantly assess and select notable rules. The non-redundant rules are then partitioned into 2 types: (i) beliefs and (ii) potential candidates for undesired rules. Additionally, a contradiction check is performed on the candidates to discover the exact undesirable rules. Thus, this framework enhances the rule clusters to identify undesirable patterns from the set of data.

The rest of the portions of this paper are prepared as follows: The recent work linked with the extraction and classification of customer interests in online networking is presented in Section II. Section III describes the proposed method, while Section IV demonstrates its efficacy. Section V summarizes this paper and suggests its possible improvement.

II. Literature Survey

Kang & Lee [15] developed a customer modeling scheme that maps the content of texts in online networking to relevant classes in news media. In this scheme, the semantic gaps between online networks and news media were minimized by using Wikipedia as an external knowledge base. The features from a short text and news type were mapped into Wikipedia-based features like Wikipedia types and article entities. Also, the customer's micro posts were defined in a rich feature space of words. But the online site data together with the textual data was needed to identify customers' interests precisely.

Trikha et al. [16] analysed the probability of detecting customers' implicit interests depending on the topic association via frequent pattern mining with no consideration of the semantics of the topics. But it needs to consider the semantic similarity among topics for improving the extraction of customer's interest efficiency. Zarrinkalam et al. [17] designed a graph-based link prediction method that functions over a representation framework comprising customer explicit contributions of topics, the association among customers and relatedness among topics. But the temporal behaviours of customers toward topics was not analysed.

Qiu& Yu [18] developed Combining Latent Dirichlet Allocation (CLDA) that can learn the possible topics of microblog short texts and long texts concurrently. The data sparsity of the short texts was removed by aggregating long texts to assist in training short texts. Short text filtering long text was reutilized to enhance the mining accurateness, creating long texts and short texts effectively merged. However, the interference of the non-meaningful Weibo on the topic extraction was not reduced.

Deng et al. [19] developed a scheme depending on tags and bidirectional interactions to extract customer topic interests on Sina Weibo. This scheme was devised by a customer interaction graph which entirely considers the benefit of the discordance among customer interactions. Forward and back distribution was used to modify the tag distribution weights. But the other complex features like community associations and impact among customers were needed to enhance the extraction of customer topic interests.

Zhu et al. [20] designed a customer interest graph defined by the hierarchical tree structure that encloses a broad variety of topics from coarse-grained to fine-grained 3-stage interest topics. The semantic similarity was determined among the attribute terms mined from the items in a customer's profile and interest nodes of the graph. But the customer interests were inferred from positive or neutral attribute terms and the change of a customer's interest over time was a challenging problem.

Zheng et al. [21] developed a model which integrates the timeliness and interactivity of microblogs to verify the hierarchical orientation and dynamic interest trend orientation of customers' interests. Then, the 3-stages interest network was created to extract the interest of microblog customers. But its efficiency was not high.

Kavitha et al. [22] classified the customer comments shared on the YouTube video sharing platform depending on their importance to the video information provided by the description related to the video shared. Comments were evaluated for polarity and split as positive or negative. But it needs to extract multilingual and non-contiguous phrases.

Dhelim et al. [23] developed a new customer interest mining model depending on different character aspects and dynamic interests. After that, a graph-based representation scheme was applied that adaptively interpret the customer's interest and avoid the bag-of-words. But the customer's interest was modelled only based on the positive relationship between customers and topics. Also, they developed a Meta-Interest model depending on the user interest extraction and meta-path finding [24]. The customer's personality aspects were considered to estimate their topics of interest and match their personality facets with the related topics. But the total computation time complexity was slightly high.

Milias & Psyllidis [25] studied the feature significance for the categorization of unlabelled Points-of-Interest (POI) into different classes. Also, a multi-class categorization scheme was employed to evaluate and sort the effect of POI features on the classification. It was used to estimate the location class among a particular group of POI classes or define whether the POI belongs to a particular class. But, the scalability of this scheme was not effective.

III.Proposed Methodology

In this section, the proposed framework is explained briefly.First, the tweets are collected and pre-processed based on the different pre-processing techniques such as stemming, lemmatization, expanding acronyms and negation replacing. Then, the CSO-GSA is applied to select the most relevant features before pruning the terms with the support value. By considering the selected features and terms, the association rules are generated. After that, pruning is applied to identify the repetitive rules from the tweets database. Moreover, the DBSCAN-based clustering is introduced to group the non-repetitive rules into beliefs and possible candidates for recognizing undesirable rules.An overall schematic representation of this proposed framework is illustrated in Figure 1.

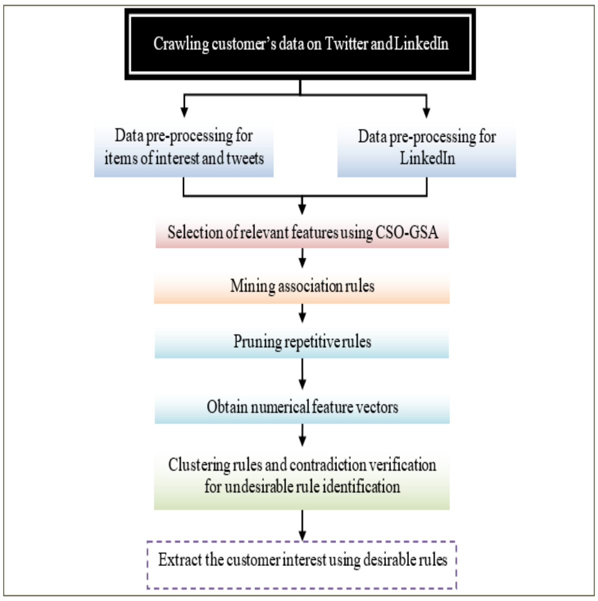
Crawling customer's data on Twitter and LinkedIn
Data pre-processing for items of interest and tweets
Data pre-processing for LinkedIn
Selection of relevant features using CSO-GSA
Mining association rules
Pruning repetitive rules
Obtain numerical feature vectors
Clustering rules and contradiction verification for undesirable rule identification
Extract the customer interest using desirable rules

Figure 1. Entire Processes in Proposed User's Interests Mining Framework

3.1 Pruning Repetitive Rules for

Contradiction Verification

For a rule$R$, $\alpha \rightarrow S$ in which α is a ubiquitous interest item in the database or α occurs wherever R occurs in the database, defines $\mathcal{P}(\alpha|R) = 1$, because of an original association. Rules with these items do not give any new knowledge about the data compared to $R \rightarrow S$. Also, these rules are longer and so difficult to understand. Such kinds of repetitive rules are filtered depending on their conditional probability $\mathcal{P}(\alpha|U), U \subseteq R\backslash\{\alpha\}$. Therefore, this redundancy inspectionprocess is applied during ARM to prune repetitive rules.

Algorithm for Contradiction Verification

```
Input:$Rule: R \rightarrow S$
Output:True when the rule is redundant; or else false
Begin$\boldsymbol{for}(n = 0: |R|)$
  $\boldsymbol{for}(every\ x\ interest\ item\ subset\ R'\ of\ R)$
    $\boldsymbol{if}(any\ item\ \alpha \in R\backslash R'\ and\ \mathcal{P}(\alpha|R') = 1)$

      Return True,
    $\boldsymbol{end\ if}$,
  $\boldsymbol{end\ for}$,
$\boldsymbol{end\ for}$
Continue these processes for $S$;
Return False
End
```

3.2Feature Vector Creation

Once the non-redundant rules are obtained, such rules are defined as the numerical feature vectors. This interpretation is based on the determination of distance among the rules and facilitates rule-based clustering. The association rule is described by 2 different interest item sets: one on the left-hand side called antecedent and another on the right-hand side called consequent. The feature vector is created by merging features for the antecedent and consequent of the rule. All features related to the distinctive item possibly exist at respective sides. Consider $r_i^{(l)}, r_j^{(r)}$ are i^{th} feature of antecedent and j^{th} feature of consequent; k_l, k_r are feature vector length for antecedent and consequent, correspondingly. In this model, a novel feature domain is considered in the range from -1 to 1 that defines how much a considered feature is associated with the antecedent or consequent of the rule. A negative range means a contradiction whereas the positive range means similarity. Particularly, all feature values are represented as a correlation score between the feature and the most correlated item in the rule.

This correlation score between a feature and an item is represented as the Spearman rank-order correlation coefficient of their qualified appearances in the database. After that, such coefficients are represented as a similarity, or a contradiction based on their sign. This representation is depending on the hypotheses that items appearing together may be identical and items that exist in distinct tweets may be contradictory. To avoid complex correlations, the coefficients are maintained with $\rho \leq 0.05$ and another coefficient is assigned as 0. Small correlation coefficients represent tiny when any relationship. A magnitude threshold is assigned which is database-dependent since more tweets enable minor correlation to be important. In this study, the empirical correlation coefficient threshold is assigned as 0.1. The algorithm for determining the feature values for a rule is presented below.

Algorithm for Feature Value Determination

```
Input:$Rule: R \rightarrow S$, Correlation matrix $M$
Output: Feature vector of $Rule$: $\phi(rule) = \left(r_1^{(l)}, \dots, r_{k_l}^{(l)}, r_1^{(r)}, \dots, r_{k_r}^{(r)}\right)$

Begin
Initialize $\phi(rule) \leftarrow \vec{0}, P \leftarrow Antecedent\ items, Q \leftarrow Consequent\ items$;
$\boldsymbol{for}(i = 0: |P|)$
  $\alpha \leftarrow item\ related\ to\ the\ feature\ r_i^{(l)}$;
  $r_i^{(l)} \leftarrow M(\alpha, \beta) s.t. \beta \in P, \forall \beta' \in P, |M(\alpha, \beta)| \geq |M(\alpha, \beta')|$;
$\boldsymbol{end\ for}$
$\boldsymbol{for}(i = 0: |Q|)$
  $\alpha \leftarrow item\ related\ to\ the\ feature\ r_i^{(r)}$;
  $r_i^{(r)} \leftarrow M(\alpha, \beta) s.t. \beta \in Q, \forall \beta' \in Q, |M(\alpha, \beta)| \geq |M(\alpha, \beta')|$;
$\boldsymbol{end\ for}$
Obtain the numerical feature vectors for a given rule.
End
```

Such feature vectors are compared to analyze how the rules are correlated to every other. This can be used to cluster the rules and find the undesirable rules.

3.3 Rule Clustering

To identify undesirable rules, the association rules are clustered and verified for belief contradiction. Before clustering, a Principal Component Analysis (PCA) is applied to the feature vectors defining the association rules. It minimizes the clustering computation cost on large-scale rule sets. The clustering is performed based on the DBSCAN algorithm using the Euclidean distance on the numerical feature vectors. The DBSCAN categorizes rules as core points, reachable points or outliers. The rules that are core or reachable points are considered as beliefs since they are supported by many analogous adjacent. The outliers are candidate undesirable rules. To identify whether they are undesirable or not, such outliers are finally compared to the beliefs by contradiction verification.

A rule $R \rightarrow S$ is undesirable regarding a belief $R' \rightarrow S'$ on the database D when the below criteria satisfy:

- S and S' rationally disagree with every other.
- R and R' co-exist in a considerable sub-collection of tweets in D. A sub-collection is important when it satisfies the customer-defined support threshold.
- R and R' are analogous to every other.

By integrating these criteria, the knowledge base,$\{R \rightarrow \sim S', R' \rightarrow S', R, R'\}$ is obtained. This is unsatisfactory and proved by the resolution. The occurrence of $R \rightarrow S$ becomes undesirable since $R' \rightarrow S'$ is a belief. The discrepancy between 2 interest itemsets is estimated by determining the cosine similarity of their feature vectors. When the cosine value is greater than the given threshold, such item sets are considered as different or analogous to each other. A similarity threshold for antecedent feature vector is represented as ε_1 and a contradiction threshold for consequent feature vector is represented as ε_2, where ε_1 is a positive integer and ε_2 is a negative integer for an undesirable rule to exist.

The DBSCAN algorithm utilized to detect the candidate undesirable rules needs 2 factors: $\min Ps$ andds, where $\min Ps$ denotes the least number of data points needed to create a dense area and ds denotes the highest distance a data point and the nearby data point is taken as aspect of the identical cluster. The selection of such factors directly influences the belief model and the detected outliers. But a belief model must not comprise any undesirable rules. It is utilized to discard particular belief model solutions that are unacceptable.

So, this analysis is utilized as a no-contradiction restraint on the $\max ds$ factor for the DBSCAN clustering. This restraint defines that there must be no contradiction enabled among clusters or as a factor that can be adjusted, no contradiction among clusters that have a huge quantity of rules.

The highest dimension of a contradicting cluster is $M_\tau = \tau|N|$ rules, where N denotes the set of all rules and τ denotes the small fraction. After that, ds is reduced gradually until DBSCAN discovers a set of clusters that do not break the no-contradiction restraint.Thus, the undesirable rules are obtained from the database and sorted them based on the similarity in feature vectors between undesirable rules and their opposing beliefs. Moreover, the user's interests are mined with the help of desirable rules accurately. The entire algorithm to identify undesirable rules is presented below.

<u>Algorithm for Finding Undesirable Rules based on Rules Clustering</u>

Input: Set of rules N, minimum support $(minSupport)\varepsilon_1 \in (0,1]$, $\varepsilon_2 \in [-1,0)$, $\min Ps$, ds, a reduction step for $ds(ds\_Step)$ and τ
Output: Set of undesirable rules E

```
Begin
//Determineoptimal $ds$
$M_\tau = \tau|N|$;
$optimal\_ds \leftarrow \max ds$;
$\boldsymbol{while}(optimal\_ds > 0)$
    $clusters, outliers \leftarrow DBSCAN(N, ds, min\,Ps)$;
    $\boldsymbol{if}\begin{pmatrix} exists\ 2\ clusters\ C_1, C_2 \in clusters\ s.t. \\ |C_1| \geq M_\tau\ and\ |C_2| \geq M_\tau\ and\ C_1\ is\ contradict\ C_2 \end{pmatrix}$
        $optimal\_ds \leftarrow optimal\_ds - ds\_Step$;
    $\boldsymbol{else}$
        Break.
    $\boldsymbol{end\ if}$
$\boldsymbol{end\ while}$
//Contradiction verification
$\boldsymbol{for}(every\ rule\ N: R \rightarrow S\ in\ outliers)$
    $\boldsymbol{for}(every\ cluster\ C\ in\ clusters)$
        $\boldsymbol{if}\begin{pmatrix} exists\ N': R' \rightarrow S'\ in\ C\ s.t. sim(\phi(N)^{(l)}, \phi(N')^{(l)}) > \varepsilon_1 \\ and\ sim(\phi(N)^{(r)}, \phi(N')^{(r)}) \leq \varepsilon_2\ and\ Support(RR') \geq minSupport \end{pmatrix}$
            $E \leftarrow E \cup \{N\}$;$\boldsymbol{end\ if}$
    $\boldsymbol{end\ for}$
$\boldsymbol{end\ for}$
End
```

IV. Experimental Results

In this section, the efficiency of the proposed method is analysedby implementing it in JAVA and evaluated with the conventional methods. For this experiment, a set of records having 9215 profiles with 221 interest items of high frequency is considered. According to the analysis of interest items, the association rules are extracted among interest items and their characteristics, namely confidence, lift, support and expectation degrees are analyzed. Table 1 presents the number of robust association rules retrieved based on the minimum thresholds.

Table 1. Number of Association Rules Retrieved based on Different Minimum Threshold Ranges

Minimum ConfidenceThreshold (%)	Minimum Support Threshold					
	0.2 %	0.6 %	1 %	1.4 %	1.8 %	2.2 %
10	1528	295	101	44	20	9
20	536	93	49	25	13	7
30	271	51	30	14	7	4
40	155	17	4	1	0	0
50	69	4	2	1	0	0
60	27	0	0	0	0	0
70	9	0	0	0	0	0
80	1	0	0	0	0	0
90	0	0	0	0	0	0

From Table 1, it is viewed that various counts of desired association rules are detected under the different minimum confidence thresholds and the minimum support thresholds. So, a set of a desirable association rules is obtained by using different minimum thresholds. In addition to this, few association rules are detected and the relationship between the customer'sinterests is shown in Table 2.

Table 2. Examples of Extracted Desirable Association Rules

Antecedent	Consequent	Confidence	Lift	Support	Expectation
Food	Travel	59.45%	231.52%	2.95%	29.52%
Friends	Family	67.18%	999.18%	2.37%	10.18%
Marketing	Media	40.32%	629.64%	3.52%	27.25%
Culture	Travel	56.81%	250.33%	2.64%	29.34%
read; photography	Travel	60.43%	275.46%	1.75%	30.07%
read; music	Movie	37.95%	370.55%	1.68%	15.49%
read; movie	Travel	41.28%	200.17%	2.81%	28.56%
read; movie	Music	45.84%	272.21%	1.27%	24.91%
read; cooking	Travel	56.52%	250.31%	1.93%	32.04%
sport; music	Travel	47.97%	220.97%	1.74%	29.62%

In the association rule, the read; photography→Travel for the degree of confidence is 60.43%, the degree of support is 1.75%, lift degree is 275.46% and the expectation is 30.07%. In the other instance,the friends→family for the degree of confidence is 67.18%, the degree of support is 2.37%, lift degree is 999.18% and the expectation is 10.18%. From this empirical relationship analysis, it is observed that the correlation between the interest and rule attains the maximum lift, support, confidence and expectation. It defines the necessary inherent association among the customer's interests.

- **Accuracy:**It is the proportion of exactly identified undesirable rules for user's interest mining over the total number of data analysed.

$$Accuracy = \frac{True\ Positive + True\ Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + FalseNegative\ (FN)}$$

- **Recall:**It is the fraction of properly identified undesirable rules by the clustering method.

$$Recall = \frac{TP}{TP + TN}$$

- **F-measure:**It is the harmonic mean of precision and recall.

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The weights are adjusted depending on the desirable rules and then the interests are recreated to the weights by reducing the degree that is illustrated in Table 3.

Table 3. Statistical Analysis of Empirical Outcomes for Recall, Accuracy and F-Measure

Intervals	Recall		Accuracy		F-Measure
	No. of members	%	No. of members	%	%
100-90%	321	40.56	2	2.61	1.36
90-80%	75	11.28	1	2.95	1.92
80-70%	69	12.41	1	3.14	2.04
70-60%	109	20.15	2	3.22	1.94
60-50%	114	21.63	5	4.31	2.08
50-40%	23	6.84	3	2.53	1.63
40-30%	55	10.36	1	2.09	2.12
30-20%	63	12.09	14	6.95	2.41
20-10%	27	6.91	103	20.46	11.26
10-0%	101	16.72	823	93.68	79.35

The user's interests are extracted properly from the considered database and the efficiency of the ARM is displayed in Figure 2.

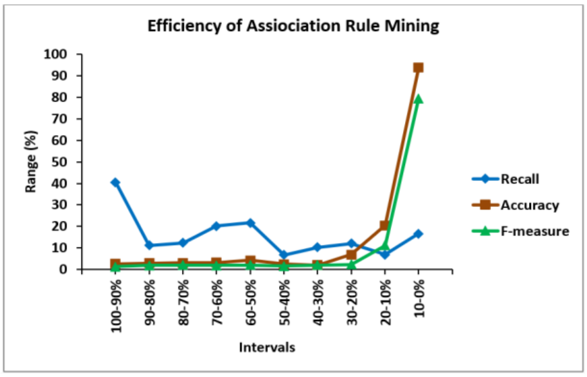

Figure 2. Efficiency of ARM for Different Intervals

The efficiency of the proposed method is depicted in Figure 2 and it achieves93.68% accuracy, 16.72% recall and 79.35% f-measure for 0-10% intervals. According to this, it is concluded that the proposed rule-based clustering is highly effective for detecting undesirable rules.

Table 4. Interest of Hits and Its Proportion with n^{th} Frequencies

Frequency Order	User count	Hits Count	Hits Proportion (%)
1.	942	175	22.48
2.	931	163	21.34
3.	927	158	25.51
4.	919	147	22.13
5.	921	120	19.26
6.	909	100	20.05
7.	905	96	18.68
8.	903	88	17.16
9.	898	87	16.02
10.	891	86	15.11
11.	886	86	14.04
12.	867	94	14.92
13.	861	62	13.75
14.	855	65	14.36
15.	861	74	13.43
16.	834	60	13.87
17.	821	70	11.09
18.	809	59	9.76
19.	796	57	8.81
20.	791	51	8.26

From Table 4, it is observed that the actual interest of user probability is typically raised regarding the increase in the frequency of the data in the database. The data collected for various hit ratios is portrayed in Figure 3.

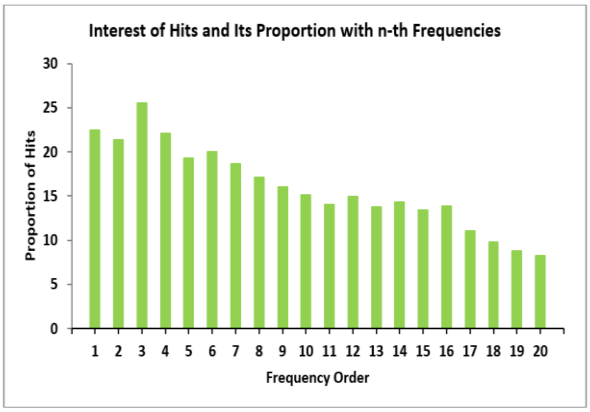

Figure 3. Proportion of Hits vs. Frequency Order

Table 5 gives the performance of the proposed rule-based clustering method compared to the other existing methods in terms of recall, accuracy and f-measure.

Table 5. Comparison of Performance

Performance Metric	CSO-GSO-ARM without Clustering	CSO-GSO-ARM with Rule-based Clustering
Recall	100	100
Accuracy	6.33	15.21
F-measure	71.54	76.3

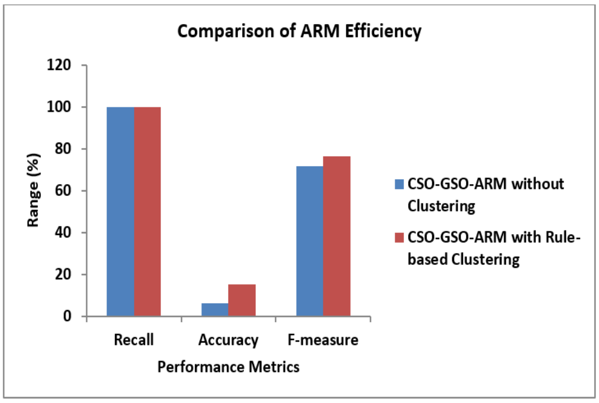

Figure 4. Comparison of Performance of the ARM

From Figure 4, it is observed that the proposed rule-based clustering method achieves high recall, f-measure and accuracy compared to the CSO-GSA-ARM without clustering for identifying the undesirable rules.

V. CONCLUSION

In this study, a new framework based on rule clustering was developed which identifies beliefs from data and finds undesirable rules in an automated manner. First, the collected tweets were pre-processed, and the most relevant features were chosen using the CSA-GSO algorithm to create the association rules. Then, the non-redundant association rules between user interests were defined as the numerical feature vectors before clustering. This was extended for summarizing and visualizing the association rules and their correlation. In addition, the repetitive rules were pruned based on the similarity among items. Moreover, the non-redundant rules were categorized into beliefs and possible candidates to identify the undesirable rules for effective user interest mining. At last, the test findings proved that the presented method has betteraccuracy compared to the state-of-the-art methods to mine the customer's interests.

REFERENCES

[1]. Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: past, present and future. *Information Systems Frontiers*, *20*(3), 531-558.

[2]. Madakam, S., & Tripathi, S. (2021). Social media/networking: applications, technologies, theories. *JISTEM-Journal of Information Systems and Technology Management*, *18*, 1-19.

[3]. Das, K., & Sinha, S. K. (2016). A survey on user behaviour analysis in social networks. *International Journal of Computer Science and Information Security*, *14*(11), 895-908.

[4]. Waheed, H., Anjum, M., Rehman, M., & Khawaja, A. (2017). Investigation of user behavior on social networking sites. *PloS one*, *12*(2), 1-19.

[5]. Mican, D., Sitar-Tăut, D. A., & Mihuţ, I. S. (2020). User behaviour on online social networks: relationships among social activities and satisfaction. *Symmetry*, *12*(10), 1-16.

[6]. Burbach, L., Halbach, P., Ziefle, M., & Calero Valdez, A. (2020). Opinion formation on the internet: the influence of personality, network structure, and content on sharing messages online. *Frontiers in Artificial Intelligence*, *3*, 45.

[7]. Uday Kumar, S., Senadeera, D. C., Yamunarani, S., & Cheon, N. J. (2018). Demographics analysis of twitter users who tweeted on psychological articles and tweets analysis. *Procedia computer science*, *144*, 96-104.

[8]. Makki, R., Soto, A. J., Brooks, S., & Milios, E. E. (2016). Twitter message recommendation based on user interest profiles. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 406-410.

[9]. Khattak, A. M., Batool, R., Satti, F. A., Hussain, J., Khan, W. A., Khan, A. M., & Hayat, B. (2020). Tweets classification and sentiment analysis for personalized tweets recommendation. *Complexity*, *2020*, 1-11.

[10]. Yang, C., Zhou, Y., & Chiu, D. M. (2016). Who are like-minded: mining user interest similarity in online social networks. In *Proceedings of the International AAAI Conference on Web and social media*, *10*(1), 1-11.

[11]. Chao, F. Y., Xu, J., & Lin, C. W. (2016). Mining user interests from social media by fusing textual and visual features. In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1-8.

[12]. Zarrinkalam, F., Fani, H., & Bagheri, E. (2019). Social user interest mining: methods and applications. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3235-3236.

[13]. Si, H., Zhou, J., Chen, Z., Wan, J., Xiong, N. N., Zhang, W., & Vasilakos, A. V. (2019). Association rules mining among interests and applications for users on social networks. *IEEE Access*, *7*, 116014-116026.

[14]. Umamaheswari, Optimal feature selection for improving user interest identification in social media.

[15]. Kang, J., & Lee, H. (2017). Modeling user interest in social media using news media and wikipedia. *Information Systems*, *65*, 52-64.

[16]. Trikha, A. K., Zarrinkalam, F., & Bagheri, E. (2018). Topic-association mining for user interest detection. In *European Conference on Information Retrieval*, Springer, Cham, pp. 665-671.

[17]. Zarrinkalam, F., Kahani, M., & Bagheri, E. (2018). Mining user interests'overactive topics on social networks. *Information Processing & Management*, *54*(2), 339-357.

[18]. Qiu, L., & Yu, J. (2018). CLDA: An effective topic model for mining user interest preference under big data background. *Complexity*, *2018*, 1-10.

[19]. Deng, L., Jia, Y., Zhou, B., Huang, J., & Han, Y. (2018). User interest mining via tags and bidirectional interactions on Sina Weibo. *World Wide Web*, *21*(2), 515-536.

[20]. Zhu, Z., Zhou, Y., Deng, X., & Wang, X. (2019). A graph-oriented model for hierarchical user interest in precision social marketing. *Electronic Commerce Research and Applications*, *35*, 1-24.

[21].Zheng, W., Ge, B., & Wang, C. (2019). Building a TIN-LDA model for mining microblog users' interest. *IEEE Access*, *7*, 21795-21806.

[22].Kavitha, K. M., Shetty, A., Abreo, B., D'Souza, A., & Kondana, A. (2020). Analysis and classification of user comments on YouTube videos. *Procedia Computer Science*,*177*, 593-598.

[23].Dhelim, S., Aung, N., & Ning, H. (2020). Mining user interest based on personality-aware hybrid filtering in social networks. *Knowledge-Based Systems*, *206*, 106227, 1-42.

[24].Dhelim, S., Ning, H., Aung, N., Huang, R., & Ma, J. (2020). Personality-aware product recommendation system based on user interests mining and metapath discovery. *IEEE Transactions on Computational Social Systems*, *8*(1), 86-98.

[25].Milias, V., & Psyllidis, A. (2021). Assessing the influence of point-of-interest features on the classification of place categories. *Computers, Environment and Urban Systems*, *86*, 1-12.

R. Umamaheswari is doing her Full Time Ph.D Programme in Computer Science. She has obtained her M.Phil, PG and UG Degree in Bharathiar University, Coimbatore. She has published Research Papers in various Research Journals. Her Area of Specialization is Data Mining.

Dr.M.Soranamageswari is an Assistant professor in PG & Department of Information Technology, Government Arts College, Coimbatore. She has obtained her PG Degree in Avinashilingam University, Coimbatore and Mphil Degree in Manonmanium sundaranar University, Thirunelveli. She completed her Doctorate degree from Avinashilingam University, Coimbatore, India, in the year 2011. Her Specialization is Data mining, Network security and Image processing.