JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 35, 787-804 (2019)
DOI: 10.6688/JISE.201907_35(4).0005

# RCUHP-SM: A Rule Generation and Clustering based Uncovering Hidden Patterns in Social Media

T. KUMARAGURUBARAN[1] AND M. INDRA DEVI[2]
[1]*Department of Computer Science and Engineering*
*Mohamed Sathak Engineering College*
*Tamil Nadu, 623806 India*
[2]*Department I/C, Department of Computer Science and Engineering*
*Kamaraj College of Engineering and Technology*
*Tamil Nadu, 626001 India*

In recent days, uncovering the hidden patterns from social media is an important and essential task. For this purpose, some of pattern mining techniques are proposed in the traditional works. But, it has some drawbacks include vagueness of termination criteria, lack of interpretability, may extract the meaningless patterns and cannot adapt any constraints within the time interval. In order to overcome these issues, this paper proposed a Rule Generation and Clustering based Uncovering Hidden Patterns in Social Media (RCUHP-SM) technique to uncover the hidden patterns. The main aim of this technique is to analyze, observe and understand the human behavior. At first, the customer review dataset is given as the input and it will be preprocessed by eliminating the irrelevant and unwanted attributes. After that, the descriptive sentences are extracted from the preprocessed data and its score is calculated by counting the tagged words. It is based on the positive, negative and neutral reviews of the user of each product. Then, a set of rules from R1 to R27 is framed to predict the category of review. Consequently, the threshold value is calculated to create the cluster groups into least similar, moderately similar and most similar. Then, it will be labeled as C1 to C6 based on its category. In the analysis phase, the features are extracted from the product description and it's corresponding score is computed. Based on the score, the features are sorted and analyzed for the recommendation. In this work, the novelty is presented in rule generation, similarity computation, threshold based cluster formation and analysis stages. In experiments, the performance of the proposed uncovering hidden pattern system is evaluated and compared in terms of Mean Absolute Precision (MAP), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) measures.

***Keywords:*** hidden pattern mining, stop words removal, parts of speech (POS) tagging, rule generation, threshold based clustering, similarity computation, filtering and feature analysis

## 1. INTRODUCTION

Social Networking sites include Google+, Facebook, Twitter, LinkedIn and *etc.*, contains an effective channel to be connected with internet subscribers for sharing the views and thoughts with others [1-3]. It has been adopted by many businesses and more companies are using these sites to offer different services. Moreover, recent research works indicate that more people are using social media for socializing, information receiving and entertaining purposes. The social media applications support sales, customer care, product development and innovation. These sites [4] comprises a massive amount

Received February 15, 2019; accepted September 12, 2018.
Communicated by Sun-Yuan Hsieh.

of data such as comments, tweets and *etc.* that is extensively distributed over different sites. This type of data hides the implicit pattern that cannot be determined by the traditional habitual analysis procedures. To identify and analyze the implicit patterns, the data mining techniques are used. Normally, the social media datasets are large in size, complex and not standardized. The main intentions of uncovering hidden patterns from large datasets are as follows:

- Analyze
- Observe
- Understand

### 1.1 Problem Description

In recent days, the process of discovering hidden pattern information [5, 6] from the social data has emerged as an interesting field of data mining. The data mining techniques provide an effective procedure to retrieve the information, preprocess the data, analyze the data and to extract the hidden patterns. The reviewers use the hidden pattern analysis to provide their opinion about various subject matters in the course of discussion. The opinion analysis [7-10] is defined as the recommendations and suggestions of the reviewers that is based on his/her own perspectives. Moreover, the opinions are categorized into positive, negative and neutral polarities. The text analysis and pattern mining techniques are the backbone of opinion analysis.

### 1.2 Motivation

In this work, a new recommender system, namely, Rule Generation and Clustering based Uncovering Hidden Patterns in Social Media (RCUHP-SM) is developed to generate an interface based on the social data by analyzing, observing and understanding the user views [11]. In addition, an improved association rule is formed to classify the reviews of the user and an exclusive similarity based clustering algorithms are proposed for data generalization. Here, a prediction estimation is generated based on a quality regression function. Then, the data validation is performed with the help of an efficient recommender system.

### 1.3 Contribution

The major contributions of this paper are as follows:

- To preprocess the customer review dataset, stop words removal and Parts of Speech (POS) tagging processes are performed in the preprocessing stage.
- To compute the score, the positive, negative and neutral reviews of each product are estimated.
- To compute the similarity between the user reviews of the product, a set of predefined rules from R1 to R27 are generated.
- To form a cluster group, the threshold value is computed based on the similarity between the reviews.

### 1.4 Organization

The remaining sections of the paper are organized as follows: Section 2 reviews some of the existing works related to hidden pattern mining and feature extraction in social media. Section 3 presents the detailed description of the proposed uncovering hidden pattern mining algorithm. Section 4 evaluates the results of both existing and proposed mining techniques. Finally, the paper is concluded and the future work to be carried out is stated in Section 5.

## 2. RELATED WORK

This section reviews some of the existing works related to pattern mining, data preprocessing, stop word removal, clustering and POs tagging. Liu, *et al.* [3] developed a temporal skeletonization approach to address the problem of curse of cardinality in sequential data analysis. The main aim of this paper was to identify the temporal patterns by summarizing the temporal correlations in an undirected graph. The challenging tasks faced in this work were as follows:

- Computational Complexity – It identified the frequent sequential patterns for large symbol sets.
- Rareness – The growing cardinality of the specific sequential pattern was decreased.
- Granularity – The useful patterns were diluted due to the large number of symbols in a sequence.
- Noise – The multi-modality of events and useful patterns were not replicated, due to the nature of sequential events.

Lyons, *et al.* [12] introduced a Hidden Markov Model (HMM) fold technique for protein fold recognition. Here, the HMM profiles were extracted by using the profile-profile sequence alignment technique. The temporal clusters are identified in the low-dimensional embedding space of the temporal graph. The cardinality of the sequential data was reduced

To evaluate the performance of the suggested technique, three difference benchmark datasets such as Ding and Dubchak (DD), Extended Ding and Dubchak (EDD) and Taguchi and Gromiha (TG) were utilized in this paper. Ruiz, *et al.* [13] suggested a new approach to mine the rules based on the fuzzy rules. Moreover, a new approach was presented to represent and evaluate the fuzzy rules then the formal model was implemented for feature extraction.

The major advantages of this paper were as follows:

- It expressed some knowledge useful in different domains include fraud detection, chemical processes, medicine and anomaly deviations.
- It attained more understandable results by using the fuzzy rules.

Here, two measures include confidence and certainty factor were evaluated to reduce the number of rules. Song, *et al.* [14] introduced a cluster based feature subset selection algorithm to find a subset of features in a high dimensional data. The stages involved in this concept were as follows:

- Irrelevant features removal
- Minimum Spanning Tree (MST) construction
- MST partitioning and representative features selection

In this paper, the redundant and irrelevant features were removed by using the feature subset selection algorithm. The dimensionality of the data was reduced by considering each cluster as a single feature. Lu, *et al.* [15] developed a framework to discover various user search goals based on the feedback sessions. The performance of inferring user search goals was evaluated with the help of Classified Average Precision (CAP). This work was categorized into three classes include query classification, search result recognition and session boundary detection. Hu, *et al.* [16] recommended a clustering based collaborative filtering approach for big data applications. The main objectives of this technique were to select the similar services in the same cluster for recommending the collaborative services. The advantage of this work was, it reduced the online execution time of collaborative filtering.

Pimpale and Patel [17] conducted an experiment with POS tagging for Indian social media text. In this application, the WEKA tool was utilized to test various machine learning techniques. The features used for training and testing were as follows:

- Language of the word
- Language of the previous and next word
- POS tags
- Position of the word in sentence

Also, the authors utilized an unlabeled data for training to represent the use of distributed vectors. Pandya, *et al.* [18] compared the classification and association rule mining techniques to uncover the hidden patterns in an Indian university. The main intention of this paper was to find the fitness of the technique in an education field. Karami, *et al.* [19] suggested a new approach, namely, Fuzzy Approach Topic Model (FATM) to uncover the hidden patterns in medical text collections. The document term frequency matrix was evaluated by calculating the value of Global Term Weighting (GTM). To model an unstructured document, a fuzzy set theory and clustering techniques were utilized in this paper. Bellogin, *et al.* [20] suggested a coverage metric to uncover and compensate the precision metrics used for social, collaborative and hybrid recommenders. The main intention of this paper was to adjust each recommender's weight based on the user correctness and relevance in a social network. The main drawback of this concept was, it doesn't have the capability to produce recommendations for users. Havens, *et al.* [21] recommended a Fuzzy C-Means (FCM) clustering technique for processing very large data. In this paper, the methods were compared based on the followings:

- Sampling followed by non-iterative extension
- Kernelized version of FCM
- Incremental techniques

From the analysis, it was identified that the kernel clustering required high memory for storing the kernel matrix, which was the main drawback of this FCM. Moreover, the complexity of these algorithms were also analyzed in this paper. Thilagavathi, *et al.* [22] investigated different hierarchical clustering algorithms based on its advantages and disadvantages in data mining. From the survey, it was analyzed that the hierarchical clustering provided more informative structure compared than the unstructured set of clusters. Wu, *et al.* [23] recommended a neighborhood based collaborative filtering approach to predict the Quality of Service (QoS). The main considerations of this paper were listed as follows:

- The impact of different QoS scale was improved by calculating the adjusted cosine based similarity.
- It used a similarity fusion based approach to increase the prediction accuracy.
- It handled the data sparsity problem.

The major drawbacks of this paper were, not-scalable and it don't learn anything from the user profile. Li, *et al.* [24] developed a multidimensional clustering based collaborative filtering approach to improve the recommendation diversity. This work includes the following stages:

- Data preprocessing
- Multidimensional clustering
- Selecting the appropriate clusters for recommendation

The main aim of this paper was to increase the effectiveness and diversity of user recommendation. Pham, *et al.* [25] introduced a clustering based collaborative filtering approach for social network analysis. The authors of this paper applied this approach on the following scenarios:

- Academic venue recommendation
- Trust-based recommendation

Here, the social relationship between the users was identified based on the ratings data. Also, a complex network clustering algorithm was applied in this work to group the similar users. Renaud-Deputter, *et al.* [26] developed a new approach based on implicit recommender system by integrating both the clustering and matrix factorization. In this paper, a high dimensional, parameter free, and divisive hierarchical technique was implemented. The major advantage of this technique was, easy to implement, very effective and could be applied to any datasets. Vyas, *et al.* [27] analyzed the normalization and transliteration problems for multilingual context. From the paper, it was analyzed that the language identification, normalization and POS tagging were considered for improving the results.

Lin, *et al.* [28] introduced a new similarity measure to estimate the similarity between two documents based on a feature. In this paper, the real time datasets were utilized to measure the effectiveness of the text classification and clustering problems. An average score of the features occurring at the least two documents was considered. The algorithms considered in this work were, single linkage, average linkage, complex linkage, *k*-means, DBSCAN and CLIQUE. Najafabadi, *et al.* [29] suggested a collaborative filtering technique to mine an implicit data based on clustering and association rules. The similar interest patterns were discovered by implementing a modified preprocessing technique. Moreover, the size of the data and the dimensionality of item space were reduced by employing the clustering technique. Lee and Yun [30] provided an efficient approach to mine an uncertain frequent patterns without false positives. The proposed list-based data structures and pruning techniques efficiently mines the frequent patterns without any pattern loss. The efficiency of the method was analyzed in terms of runtime, scalability and runtime.

From the survey, the merits and drawbacks of existing pattern mining techniques are investigated. In order to solve those issues, this paper proposed a new technique for uncovering patterns in social media. The description about the proposed technique is discussed in the next section.

## 3. PROPOSED METHOD

This section presents the detailed description of the proposed Rule Generation and Clustering based Uncovering Hidden Patterns in Social Media (RCUHP-SM) system. The overall flow of the proposed system is shown in Fig. 1, which includes the following stages:

- Preprocessing
- Score computation
- Rule generation
- Thresholding based clustering
- Overall recommendation

At first, the input customer review dataset is given as the input, which is preprocessed to eliminate the irrelevant and unwanted attributes in the dataset. After that, the stop words like *wh* words and conjunction words are removed from the sentence. Then, the Parts of Speech (POS) tagging is applied to extract the adjectives and adverbs from the sentence. Because, it describes the mode such as positive or negative of the sentence, the descriptive sentences are extracted to compute the score value. Here, we generate some rules to find the similarity between the reviews in the social media. Then, the threshold value is computed to create the cluster of reviews based on the similarity values. It is categorized into least similar, moderately similar and most similar. Then, the overall recommendation is displayed for further process. In the analysis phase, the features are extracted and its corresponding score is calculated. Based on this value, the features are analyzed for the recommendation.

### 3.1 Preprocessing

Initially, the given customer review dataset is given as the input, which is preprocessed by eliminating the unwanted attributes in the data. The main aim of preprocessing is to reduce the file size and improve the quality of the data. In this stage, the following processes are performed.

- Special characters removal
- Unwanted spaces removal
- Stemming
- Stop words removal
- POS tagging

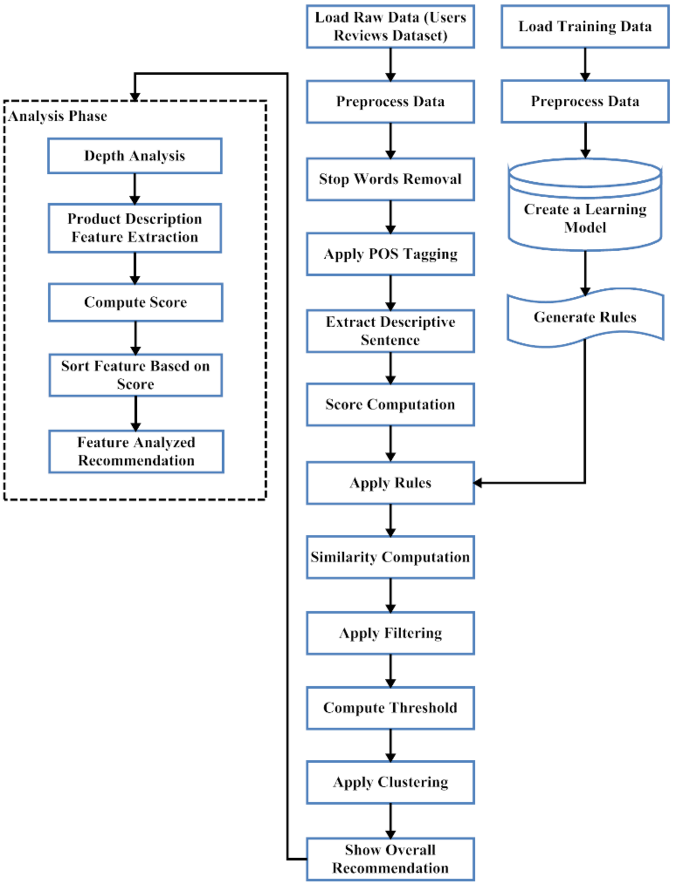

Fig. 1. Overall flow of the proposed system.

The stop words include *wh* words, conjunction words, auxiliary verbs, articles, prepositions, pro-nouns and *etc*. Removing these words from the dataset can reduce the dimensionality of term space. In text mining applications, the stop words are not measured as keywords. The stop words removal is an important preprocessing step, because which are language specific functional words and has no information. After removing the stop words, the POS tagging is applied to assign each word in a sentence. Typically, the POS contains some important grammatical information. The main aim of POS tagging is to select a most probable speech sequence of the words in the sentence. Moreover, it separates the sentence into a noun, adjective, adverb, verb and all the parts of speech tag.

### 3.2 Score Computation and Rule Generation

After preprocessing, the tagged words from the database are counted to calculate the number of positive, negative and neutral words in each review. It will be compared with the trained words, based on this value, the score is computed. To identify each review as positive, negative or neutral based on the calculated score, some predefined rules are generated in this paper. In this process, there are 27 numbers of rules include R1 to R27 are formed to predict whether the review is positive, negative or neutral. If the number of positives, negatives and neutrals are 1, it is set to be rule 1 and its review is considered as neutral. Moreover, the review is categorized based on the greater value, for instance, if the number of positives is greater than the negatives and neutrals, it is considered as a positive review. Likewise, if the number of negatives is greater than the other, it is considered as negative. Similarly, the neutral is also estimated. Based on the category of reviews, the similarity is computed. The rules are tabularized in Table 1.

**Table 1. Rule generation.**

| Rules | No of positive | No of negative | No of neutral | Review |
|---|---|---|---|---|
| R1 | 1 | 1 | 1 | Neutral |
| R2 | 1 | 1 | 2 | Neutral |
| R3 | 1 | 1 | 3 | Neutral |
| R4 | 1 | 2 | 1 | Negative |
| R5 | 1 | 2 | 2 | Negative |
| R6 | 1 | 2 | 3 | Negative |
| R7 | 1 | 3 | 1 | Negative |
| R8 | 1 | 3 | 2 | Negative |
| R9 | 1 | 3 | 3 | Negative |
| R10 | 2 | 1 | 1 | Positive |
| R11 | 2 | 1 | 2 | Positive |
| R12 | 2 | 1 | 3 | Positive |
| R13 | 2 | 2 | 1 | Neutral |
| R14 | 2 | 2 | 2 | Neutral |
| R15 | 2 | 2 | 3 | Neutral |
| R16 | 2 | 3 | 1 | Negative |
| R17 | 2 | 3 | 2 | Negative |
| R18 | 2 | 3 | 3 | Negative |
| R19 | 3 | 1 | 1 | Positive |
| R20 | 3 | 1 | 2 | Positive |
| R21 | 3 | 1 | 3 | Positive |
| R22 | 3 | 2 | 1 | Positive |
| R23 | 3 | 2 | 2 | Positive |
| R24 | 3 | 2 | 3 | Positive |
| R25 | 3 | 3 | 1 | Neutral |
| R26 | 3 | 3 | 2 | Neutral |
| R27 | 3 | 3 | 3 | Neutral |

### 3.3 Similarity Computation

In this stage, the similarity values are calculated by comparing all reviews with the other reviews of the product. Then, the comparative values for the reviews with the same type of reviews are retained, where the other reviews are filtered. In this algorithm, the product Id $R$ and user id $U_n$ are given as the input. Then, for each product, the number of reviews in the list are listed. For each review of the user, the similarity is computed and the similar words of the reviews are listed. Also, the distinct of the reviews are collected and the size of two lists is calculated. After that, the common words in $N_1$ and $N_2$ are computed and the similarity between the total words and distinct words is computed. Then, this process will be repeated until computing the similarity between all the products.

**Algorithm 1: Similarity Computation**

```
Input:  Product Id $R = \{R1, R2, R3, R4, R5\}$
        User id ($U_n$), $n = 1, 2 \dots N$; // Where, $N$ represents the number of user who reviewed each product;
Output: Similarity and distance values;
Procedure:
  For each product $R$,
    Load the list of reviews in the Review List $R$list;
      For $i = 1$; // Where, $i$ represents the review of the user with user id $U_i$;
        For $j = i + 1$ is the review of the user with the user id $U_j$;
          $P_i$ = Review of the user $U_i$;
          $P_j$ = Review of the user $U_j$;
          Similarity ($P_i$, $P_j$);
          $N_1$ = Words list $P_i$;
          $N_2$ = Words list $P_j$;
          Retain distinct words in the list $N_1$ and $N_2$;
          Compute the size of both list ($S_1$, $S_2$);
          $C_{Tot} = S_1 + S_2$;
          $C_{Uniq}$ = Common words in $N_1$ and $N_2$;
          Similarity value = $2 \times (C_{Uniq}/C_{Tot})$;
          Distance = 1 – Similarity;
        End for;
      End for;
    Continue until computing the similarity for all products;
End;
```

### 3.4 Threshold based Clustering

After computing the similarity, the threshold is computed to form the cluster. In this algorithm, the number of products $n$ is given as the input. Here, the similarity score is calculated for each product based on the similarity calculation. Then, the similarity values of all the products are stored in the similarity list (simlist). This list is sorted by arranging the values as minimum to maximum, then the average is calculated. Here, the smallest value of the list is considered and the list is split into three groups. The threshold value is calculated as follows:

Threshold = average – smallest value/3;
C1 = (initial value – (initial + 1) × threshold);
Increment the initial value;
C2 = (initial value – (initial + 1) × threshold);
Increment the initial value;
C3 = (initial value – final value in the list);

It will be sorted in which the minimum and maximum similarity are estimated and the summation is performed. Based on the size of simlist, the mean is calculated and, the range is estimated as low, medium or high from the minimum and maximum values.

**Algorithm 2: Threshold Computation**

```
$N$ = 5 (number of products);
For $N$ = 1 to 5;
    Simlist ← Similarity values of the product
    // Where distance ! = 1;
    Sort (Simlist);
    Minval ← Simlist ($l$);
    Simsum = Sum (Sorted Simlist);
    Let $S_z$ be the size of the Simlist;
    Maxval = Mean (SimSum);
    Let $M$ = 3;
    Calculate the range as low, medium or high from Minval and Maxval;
End;
```

After calculating the threshold value, the reviews are grouped into the clusters based on the similarity values with respective to the type of reviews of all the products. Then, the clusters are comes under the category of least similar, moderately similar and most similar in the cases of positive and negative reviews.

Furthermore, it will be labeled by using the following table:

**Table 2. Clustering based on the reviews.**

| Review | Cluster Name |
|---|---|
| Positive and Most Similar | C1 |
| Positive and Moderately Similar | C2 |
| Positive and Least Similar | C3 |
| Negative and Most Similar | C4 |
| Negative and Moderately Similar | C5 |
| Negative and Least Similar | C6 |

### 3.5 Analysis Phase

After clustering, the processes include depth analysis, feature extraction for product description, score computation, feature sorting and recommendation are performed during the analysis phase. Here, the number of positive and negative comments are identified, based on this value, the score is calculated for the descriptive words. In this algorithm, the number of products *N* that are reviewed, total number of products $P_i$, number of positive reviews *NOP*, number of negative reviews *NON* and number of neutral reviews *NOU* are considered. If the value of NOP is greater than the NON & NOU or if the NOP is less than the NON & NOU values, it is considered as a positive review. If the value of NON is greater than the NOP & NOU or if the NON is less than the NOP & NOU values, it is considered as a negative review. If the review is positive, the reviews are sorted based on the minimal verge score. Then, the tagging is applied to extract the product features and the relevancy is computed for those features. Then, it will be displayed in the descending order. If the review is negative, the reviews are sorted based on the minimal verge score. Then, the same process mentioned in the positive review is applied for the negative review. Finally, it will be recommended for further analysis.

**Algorithm 3: Deep Analysis**

```
Let, N – Number of products that are reviewed, P_i – Product, where i = 1 to N, NOP –
        number of positive reviews, NON – number of negative reviews, NOU – number of neutral reviews;
Begin
        If (NOP > (NON && NOU)) || (NOP == NOU) > NOP)
        Type (P_i) → Positive;
        Identify the overall dimension of the review based on NON, NOP and NOU;
            If (NON > (NOP && NOU) || (NON == NOU) > NOP)
            Type (P_i) → Negative;
            If (Type (P_i) → Positive)
                Extract Plist ← sort (Positive reviews based on the minimal verge score);
                P_feat ← Apply tagging to extract the features;
                Compute relevance (P_feat);
                Rank P_feat based on relevancy;
                Display feature that are in descending order;
            End if;
            If (Type (P_i) → Negative)
                Extract Plist ← Sort (Negative reviews based Minimal verge vector);
                N_feat ← Apply tagging to extract features;
                Compute relevancy (N_feat);
                Rank N-feat based on relevancy;
                Display features that are in descending order;
            End if;
End;
```

The major advantages of this pattern mining system are as follows:

- Highly efficient
- Provides the best similarity matching results

- Extracts only the meaningful patterns
- Accurate score calculation

The superiority of the proposed RCUHP-SM technique is proved in the next section.

## 4. PERFORMANCE ANALYSIS

This section evaluates the performance results of both existing and proposed techniques in terms of execution time, accuracy, mean average precision, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The dataset used in this analysis is customer review dataset [31], which contains the details of customer reviews of multiple usage products.

### 4.1 Dataset

Fig. 2 compares the existing [32] and proposed clustering techniques based on the dataset purity value, where the x-axis represents the algorithms and the y-axis represents the dataset purity value. The existing techniques considered in this analysis are, Kl-FCM-GM, IWKM, EKP, OCIL and ACC-FSFDP. Here, the purity is calculated for the Apex DVD player, Canon G3 Camera, Zen MP3 Player, Nikon Camera and Nokia Phone products. From the analysis, it is observed that the proposed RCUHP-SM technique provides high dataset purity value, when compared to the other techniques.

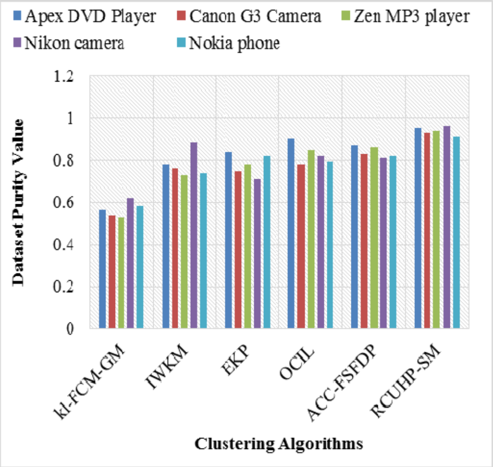

Fig. 2. Dataset purity value.

### 4.2 Execution Time

Execution time is defined as the amount of time taken to execute the process. Fig. 3 illustrates the execution time of existing and proposed clustering techniques. Here, the products such as an Apex DVD player, Canon G3 camera, Zen MP3 player, Nikon camera and Nokia phones are considered for comparison. When compared to the existing techniques, the proposed RCUHP-SM provides the best results. Also, the execution time with respect to number of user reviews is analyzed in Fig. 4, where the x-axis represents the reviews from 100 to 300 and the y-axis represents the execution time. In this analysis, the techniques such as StrAP, StrDenAP, Str-FSFDP and proposed RCUHP-SM are considered. From the analysis, it is observed that the proposed technique provides requires minimum execution time, when compared to the other techniques.

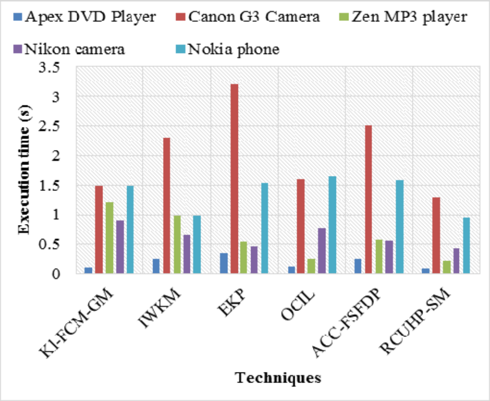

Fig. 3. Average execution time of existing and proposed clustering techniques.

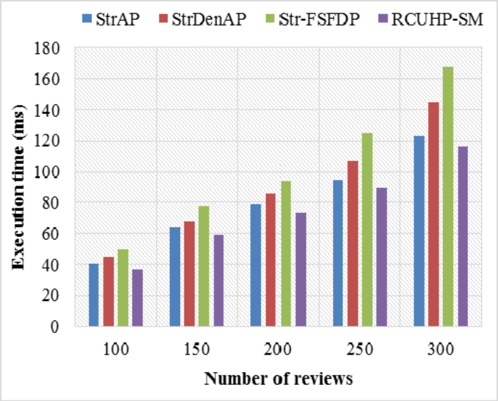

Fig. 4. Execution time vs. Number of reviews.

### 4.3 Mean Average Precision (MAP)

The Mean Average Precision (MAP) is a ranked precision metric that provides a larger credit to correctly recommended items in the top ranks. It is calculated for a given top-$N$ recommendation list, which is shown in below:

$$Precision@N = \frac{|U_{N,rec} \cap U_{adopted}|}{N}. \tag{1}$$

Where, $N$ represents the number of recommendations received and $C_{adopted}$ indicates the items that a user adopted in the test data. Fig. 5 shows the MAP of both existing [33] and proposed techniques, where the x-axis represents the training set and the y-axis represents the MAP. For this analysis, the canon G3 dataset is used. Also, the MAP is analyzed with respect to the % of dataset as shown in Fig. 6, where the Nokia 6610 dataset is used. From these results, it is observed that the proposed RCUHP-SM provides high precision value, when compared to the other techniques.

### 4.4 Root Mean Squared Error (RMSE)

The RMSE is one of the widely used error measures in data mining applications. In this measure, the sum of the individual squared errors is obtained, where each error indicates the sum of individual errors. Moreover, it can be varied with respect to the error magnitude. The RMSE value is calculated as follows:

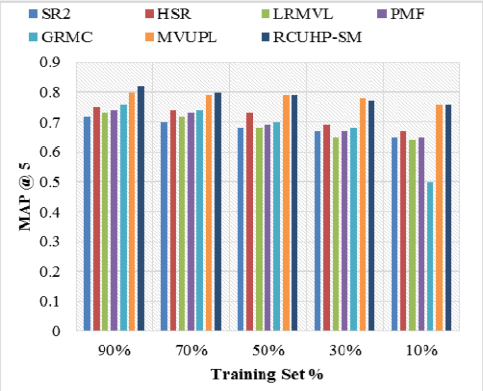

Fig. 5. MAP for Canon G3 dataset.

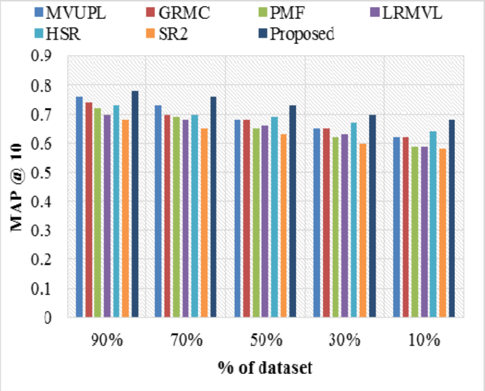

Fig. 6. MAP for Nokia 6610 dataset.

$$RMSE = \sqrt{\tfrac{1}{n}\sum\nolimits_{i=1}^{n} e_i^2}. \tag{2}$$

Where, the $n$ represents the number of samples of model error $e_i$. Fig. 7 shows the RMSE value of proposed RCUHP-SM technique with respect to different alpha values, where the canon G3 dataset is used. Moreover, the RMSE values for the Nokia 6610 dataset is analyzed in Fig. 8. From the results, it is analyzed that the proposed technique provides the minimized RMSE value for both datasets.

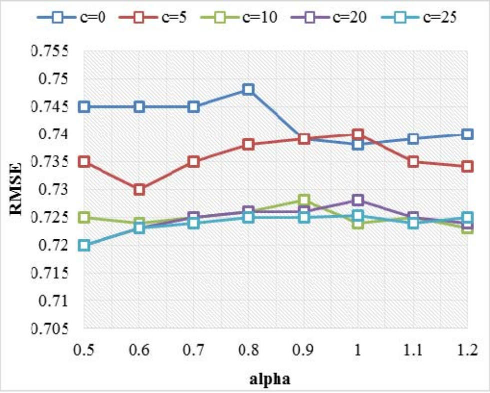

Fig. 7. RMSE for canon G3.

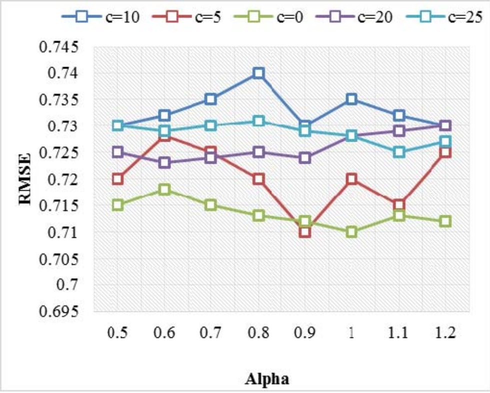

Fig. 8. RMSE for Nokia 6610 dataset.

### 4.5 Mean Absolute Error (MAE)

The MAE is also an important measure that is widely used in many model evaluation applications. It obtains the total error by summing the magnitudes of the errors and dividing the total error. The MAE is calculated as follows:

$$MAE = \tfrac{1}{n}\sum\nolimits_{i=1}^{n} | e_i |. \tag{3}$$

Where, the *n* represents the number of samples of model error $e_i$. Fig. 9 shows the MAE of both existing [16] and proposed techniques, where the x-axis represents the k values and the y-axis represents the MAE. From the analysis, it is observed that the proposed RCUHP-SM provides the minimized error value, when compared to the other techniques.

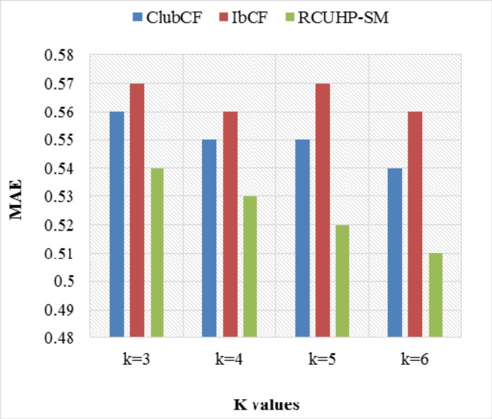

Fig. 9. MAE of existing and proposed techniques.

## 5. CONCLUSION AND FUTURE WORK

This paper proposed a RCUHP-SM technique for uncovering the hidden patterns from social media. Here, the customer reviewer dataset is used to validate the proposed pattern mining system. The data preprocessing performed on the dataset by eliminating the unwanted attributes. In this stage, the procedures that include special characters removal, unwanted spaces removal, stemming, stop words removal and Parts of Speech (POS) tagging are applied for preprocessing data. The descriptive sentence extraction and its score calculation are performed based on the tagged words. Moreover, a set of rules are formed to categorize the review as positive, negative and neutral. Furthermore, the cluster is formed based on the obtained threshold value. In the analysis module, the feature extraction and score computation processes are performed for the recommendation. The major advantages of this paper are, easy to implement, minimized error value, highly efficient and low cost. In experiments, the performance of the proposed mining technique is evaluated in terms of execution time, MAP, MAE and RMSE measures. Furthermore, these results are compared with the existing pattern mining techniques for proving the superiority of the proposed technique. From this analysis, it is analyzed that the proposed RCUHP-SM provides the best results compared than the other techniques.

In future, this work will be enhanced by implementing the proposed technique in observing and understanding phases.

## REFERENCES

1. B. C. Fung, *et al.*, "Anonymizing social network data for maximal frequent-sharing

pattern mining," in *Recommendation and Search in Social Networks*, Springer, 2015, pp. 77-100.

2. G. Bello-Orgaz, *et al.*, "Social big data: Recent achievements and new challenges," *Information Fusion*, Vol. 28, 2016, pp. 45-59.
3. C. Liu, *et al.*, "Temporal skeletonization on sequential data: patterns, categorization, and visualization," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, 2016, pp. 211-223.
4. M. Injadat, *et al.*, "Data mining techniques in social media: A survey," *Neurocomputing*, Vol. 214, 2016, pp. 654-670.
5. S. Qiao, *et al.*, "A self-adaptive parameter selection trajectory prediction approach via hidden Markov models," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, 2015, pp. 284-296.
6. D. J. Peuquet, *et al.*, "A method for discovery and analysis of temporal patterns in complex event data," *International Journal of Geographical Information Science*, Vol. 29, 2015, pp. 1588-1611.
7. L. Zhang and B. Liu, "Aspect and entity extraction for opinion mining," in *Data Mining and Knowledge Discovery for Big Data*, Springer, 2014, pp. 1-40.
8. K. Khan, *et al.*, "Mining opinion components from unstructured reviews: A review," *Journal of King Saud University-Computer and Information Sciences*, Vol. 26, 2014, pp. 258-275.
9. M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," in *Proceedings of the 6th International Conference on Communication Systems and Networks*, 2014, pp. 1-8.
10. J. P. Verma, *et al.*, "Big data analysis: recommendation system with Hadoop framework," in *Proceedings of IEEE International Conference on Computational Intelligence and Communication Technology*, 2015, pp. 92-97.
11. J. Bao, *et al.*, "Recommendations in location-based social networks: a survey," *Geoinformatica*, Vol. 19, 2015, pp. 525-565.
12. J. Lyons, *et al.*, "Advancing the accuracy of protein fold recognition by utilizing profiles from hidden Markov models," *IEEE Transactions on Nanobioscience*, Vol. 14, 2015, pp. 761-772.
13. M. D. Ruiz, *et al.*, "Discovering fuzzy exception and anomalous rules," *IEEE Transactions on Fuzzy Systems*, Vol. 24, 2016, pp. 930-944.
14. Q. Song, *et al.*, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, 2013, pp. 1-14.
15. Z. Lu, *et al.*, "A new algorithm for inferring user search goals with feedback sessions," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, 2013, pp. 502-513.
16. R. Hu, *et al.*, "ClubCF: A clustering-based collaborative filtering approach for big data application," *IEEE Transactions on Emerging Topics in Computing*, Vol. 2, 2014, pp. 302-313.
17. P. B. Pimpale and R. N. Patel, "Experiments with POS tagging code-mixed Indian social media text," arXiv preprint arXiv:1610.09799, 2016.
18. S. D. Pandya and P. V. Virparia, "Comparing the application of classification and association rule mining techniques of data mining in an Indian university to uncover

hidden patterns," in *Proceedings of International Conference on Intelligent Systems and Signal Processing*, 2013, pp. 361-364.

19. A. Karami, *et al.*, "A fuzzy approach model for uncovering hidden latent semantic structure in medical text collections," in *Proceedings of iConference*, 2015.
20. A. Bellogín, *et al.*, "An empirical comparison of social, collaborative filtering, and hybrid recommenders," *ACM Transactions on Intelligent Systems and Technology*, Vol. 4, 2013, p. 14.
21. T. C. Havens, *et al.*, "Fuzzy c-means algorithms for very large data," *IEEE Transactions on Fuzzy Systems*, Vol. 20, 2012, pp. 1130-1146.
22. G. Thilagavathi, *et al.*, "A survey on efficient hierarchical algorithm used in clustering," *International Journal of Engineering Research and Technology*, Vol. 2, 2013, pp. 2553-2556.
23. J. Wu, *et al.*, "Predicting quality of service for selection by neighborhood-based collaborative filtering," *IEEE Transactions on Systems*, *Man*, *and Cybernetics: Systems*, Vol. 43, 2013, pp. 428-439.
24. X. Li and T. Murata, "Using multidimensional clustering based collaborative filtering approach improving recommendation diversity," in *Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012, pp. 169-174.
25. M. C. Pham, *et al.*, "A clustering approach for collaborative filtering recommendation using social network analysis," *Journal of Universal Computer Science*, Vol. 17, 2011, pp. 583-604.
26. S. Renaud-Deputter, *et al.*, "Combining collaborative filtering and clustering for implicit recommender system," in *Proceedings of IEEE 27th International Conference on Advanced Information Networking and Applications*, 2013, pp. 748-755.
27. Y. Vyas, *et al.*, "POS tagging of English-Hindi code-mixed social media content," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 974-979.
28. Y.-S. Lin, *et al.*, "A similarity measure for text classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, 2014, pp. 1575-1590.
29. M. K. Najafabadi, *et al.*, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," *Computers in Human Behavior*, Vol. 67, 2017, pp. 113-128.
30. G. Lee and U. Yun, "A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives," *Future Generation Computer Systems*, Vol. 68, 2017, pp. 89-110.
31. "Opinion mining, sentiment analysis, and opinion spam detection," https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html, 2004.
32. J.-Y. Chen and H.-H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data," *Information Sciences*, Vol. 345, 2016, pp. 271-293.
33. H. Lu, *et al.*, "Social recommendation via multi-view user preference learning," *Neurocomputing*, Vol. 216, 2016, pp. 61-71.

**T. Kumaragurubaran** in Department of Computer Science and Engineering, Mohamed Sathak Engineering College, Kilakarai, India. E-mail: kumaragurubarancse@hotmail.com

**M. Indra Devi** in Department of Computer Science and Engineering, Kamaraj College of Engineering and Technology, India. E-mail: indradevimdr@outlook.com