International Journal of Contemporary Technology and Management
Volume VII, Issue II, February - 2018 ISSN (Online) 2278-8034

User Chat Recommendation of Suggesting Images using Clustering and Classification Techniques

Yuvraj Mahajan[1], Abhay Kothari[2]
Research Scholar[1], Professor [2]
Department of Computer Science
Acropolis Institute of Technology & Research, Indore, Madhya Pradesh, India
[1]ymahajan11@gmail.com, [2]abhaykothari@acropolis.in

Abstract **-Now in these days traditional mobile phones are replaced by new generation smart phones. These smart phones are not only used for voice communication, now video calls and other text communication mediums are also being used i.e. social media. In addition of that these mobile phones are used for accessing the internet. From the different source of information it is found that the maximum mobile data is consumed during access of social media where the text messaging, image uploads and other communication is performed. That consumes a significant amount of data. In this paper, the proposed work is intended to develop a recommendation system which analyze text chat history and utilized images during the chats. Additionally during the similar text messages suggest the images which are relevant to the current chat. In order to develop such recommendation system k-means clustering algorithm and bay's classifier is used. That approach is demonstration of functional data model of proposed recommendation system design. That is implemented in Android technology and the performance of implemented system is measured in terms of accuracy and their computational complexity i.e. space and time. The computed results show the proposed model is efficient and accurate for predicting similar kinds of uploaded images.**

Keywords:Android, Chat System, Text Messages, Clustering, Bayesian Classifier, Data Consumption

I. INTRODUCTION

Smart mobile devices are loaded with sensors, Wi-Fi, and other applications which can easily downloaded and installed on phones. Due to these functions the mobile power and data consumption is also increases. Mostly, that is happening use of social media applications, because these applications are mostly used to express emotions or text communication frequently. In this context the performance scaling of mobile phones are key concern of this work. The proposed work is intended to use the concept of data mining and machine learning to learn with the user's text communication and use to reduce the effort of mobile by suggesting the most likely images from the current chat. The techniques of data mining help to understand the pattern of data and find the similar patterns among newly appeared data objects. By using this hypothesis the proposed data model is prepared.

In order to develop such a prototype the user's chat history and included images are used for training of proposed data model. Both kinds of data treated separately. First of all the text data is used with the k-means clustering algorithm. The clustering algorithm creates group of similar text or chat history. Using this chat clusters the mapping of images are performed for creating the group of similarly used images during the chat. In next created group of text messages are used with bay's classifier for performing training. By training of text data the model is enabled for classify the newly appeared text into the similar groups. Therefore a text communication is performed for testing of trained data model. Testing includes the prediction of text group and the suggestion is made for the similar group images. This section involves the overview of proposed system for designing recommendation model. The next section provides the detailed methodology of working model.

II. PROPOSED WORK

A. *Methodology*

The proposed model for recommendation of chat images is defined in figure 1. In this figure the required components of the systems are also included to understanding the flow of data and methodology of processes involved.

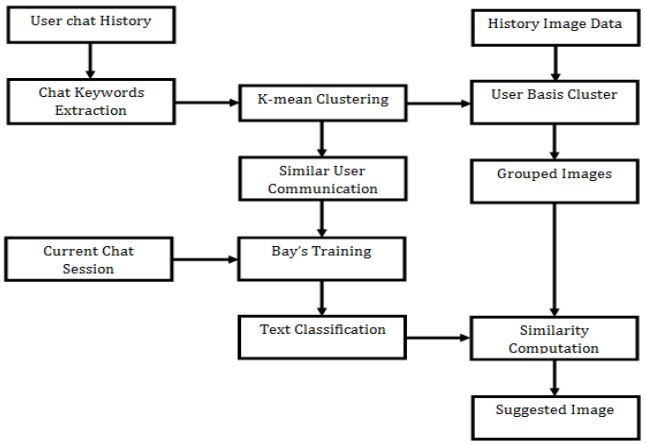

Figure 1System Architecture

User chat history: every data mining model needs initial input data to analyze and develop a mathematical model. This data model is used to identify the similar patterns of data or predict the similar classes of data. In this work the user chat history is included as initial data for analysis.

Chat keyword extraction: the chat communication some time contains additional words which are not much appropriate for identifying the domain information from text. Therefore the unnecessary words and special characters are removed from the input text. This process is also termed as data pre-processing for optimizing the text dataset. After that the remaining words of the dataset is processed to compute the frequency of the words. The frequency of words can be computed using following formula:

$$\text{wordfrequency} = \frac{\text{wordcount}}{\text{totalwordcount}}$$

After computing the word frequency the words are sorted according to their frequency values. The higher frequently occurred values are keeping preserved and remaining data is removed.

K-means clustering: In partition-based clustering k-means algorithm is a popular technique. When application requires creating k number of partitions from input data this approach is much appropriate. Here from the input text data k number of clusters required thus first need to select k random chat objects from entire preprocessed text data as initial centroid. These initial centroids are the points to create clustered objects. After initial centroids selection distance between each object and centroid is calculated. The distance of object and centroid shows dissimilarity among them. This distance is used to identified and assigned clusters to the available data objects. In next phase optimization of centroids are taken place. In this phase average of all cluster objects are used to find new centroid. That is a cyclic process and repeated until the criterion function converged. Square error criterion for clustering can be given using the following formula:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - m_i \right\|^2$$

x_{ij}, is the sample j of i-class,$m_i$is the center of i-class,n_ii is the number of samples of i-class.K-means clustering algorithm is simply described as:

Table 1 k-Means Clustering

Input*:* N objects to be cluster $x_1, x_2, \dots, x_n$, k number of cluster

Output*:* k clusters and sum of dissimilarity between objects and nearest centroids;

Process*:*

1. Select k objects randomly as initial centroid$(m_1, m_2, \dots, m_k)$;

2. Compute distance between each object x_i and centroid m_k, Assign each object to nearest cluster, distance calculation can be performed as:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_i - m_{j1})^2}, i = 1 \dots N, j = 1 \dots k$$

$d(x_i, m_i)$is the distance between data i and cluster j.

3. Calculate mean of objects in each cluster for finding new cluster centeroid,

$$m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$$

N_i is the number of objects in cluster i;

4. Repeat 2, 3 step until function E converged, return$(m_1, m_2, \dots, m_k)$.

The clustering generates the k number of clusters as user required. The output of this phase is used to regulate two processes first for identifying the similar kinds of text messages and second for finding the images which are used in communication in grouped manner.

History image data: this the second input for the system, this data is combination of two factors the previously used image and the Meta text or descriptor. That descriptor is help to identify in which chat the image is used.

User basis cluster: the output of the k-means is used with the input image dataset that is used to identify the user, text communicated and the images.

Similar user communication: that is the second place where the k-means clustering output is used. That provides the list of similar text communication.

Grouped images: the grouped images are obtained according to the k-means clustering as described previously.

Bay's training: The standard approach of Bayesian classification uses chain rule to decompose the joint distribution:

$$\Pr(C, A_1, A_2, \dots, A_k) = \Pr(C)\Pr(A_1, A_2, \dots, A_k|C) \dots\dots\dots\dots (1)$$

The first term on the right hand side of (1) is the prior probability of the class labels. These can be directly estimated from the training data, or from a larger sample of the population. For example, we can often get statistics on the number of, say, breast cancer occurrences in the general population. The second term on the right-hand side of (1) is the distribution of attribute values given the class label. The estimation of this term is usually more complex, and we elaborate on it below.

Once we have an estimate of $\Pr(C)$and $\Pr(A_1, A_2, \dots, A_k|C)$ we can use Bayes rule to get the conditional probability of the class given the attributes:

$$\Pr(C|A_1, A_2, \dots, A_k) = \alpha \Pr(C)\Pr(A_1, A_2, \dots, A_k|C) \dots\dots\dots\dots\dots (2)$$

Whereα is a normalization factor that ensures that the conditional probability of all possible class labels sums up to 1. (In practice, we do not need to explicitly evaluate this factor because it is constant for a given instance.) Using (2) we can classify new instances by combining the prior probability of each class with the probability of the given attribute values given that class.The Naive Bayes classification algorithmic rule is a probabilistic classifier. It is based on probability models that incorporate robust independence assumptions.

The independence assumptions usually don't have an effect on reality. So they're thought of as naive. You can derive probability models by using Bayes' theorem (proposed by Thomas Bayes). Based on the nature of the probability model, you'll train the Naive Bayes algorithm program in a very supervised learning setting. In straightforward terms, a naive Bayes classifier assumes that the value of a specific feature is unrelated to the presence or absence of the other feature, given the category variable. There are two types of probability as follows:

- Posterior Probability [P (H/X)]
- Prior Probability [P (H)]

Where, X is data tuple and H is some hypothesis. According to Baye's Theorem

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)}$$

Current chat session: the trained of Bayesian classifier results the probability distribution for the classes available for text. When two users of the system communicate then this active session of chat is termed here as current chat session.

Text classification: the Bayesian probabilities are used here for computing the text classes or groups. Therefore the current text is used with trained classifier for identifying the text group.

Classified text: the text classification results the group or domain name.

Similarity computation: now on this group the text features of current text is compared with the available cluster text objects. The similarity computation is performed using the Euclidean distance. The Euclidean distance is one of the processes to compute the dissimilarity and below 50% dissimilar image objects are used with next step.

Suggested images: the image listed in previous process is final outcome of the system. The most relevant image user can select for use in the chat.

B. Proposed Algorithm

The above given methodology is described in this section using the algorithm steps. The table 2 contains the list of steps which are followed for suggesting the relevant images:

Table 2 Proposed Algorithms

Input: Image History IH, Chat History CH, Number of clusters K, Chat Session CS

Output: Suggested images R

Process:

1. $T_N = readChat(CH)$
2. $P_N = PreProcessData(T_N)$
3. $for(i = 1; i \leq N; i++)$
    a. $T_{frq} = CountFrequency(P_i)$
4. $endfor$
5. $CK_N = SelectChatKeywords(T_{frq})$
6. $[user, clusters, centroid] = Kmeans.Cluster(CK_N, K)$
7. $CreateImageGroups(IH, user, clusters)$
8. $T_{model} = Bays.Train(clusters)$
9. $C = T_{model}.classify(CS)$
10. $for(j = 1; j \leq cluster.objectLength; j++)$
    a. $D = FindDistance(CS, Cluster_j)$
    b. $if(D \leq 0.5)$
        i. $R_{list}.add(cluster_j)$
    c. $endif$

11. End for
12. Return R_{list}

III. RESULT ANALYSIS

A. *Accuracy*

The performance of the proposed user chat classification system for improvement of mobile data consumption of machine level approach is measure in terms of accuracy is calculated and demonstrated using figure 2. Fundamentally the accuracy is measurement of suitability of the classification system. The accuracy of any classification model of data mining can be figure out using the following formula:

$$\text{Accuracy} = \frac{\text{Total Correctly Classified Data}}{\text{Total Data for Classification}} \text{X100}$$

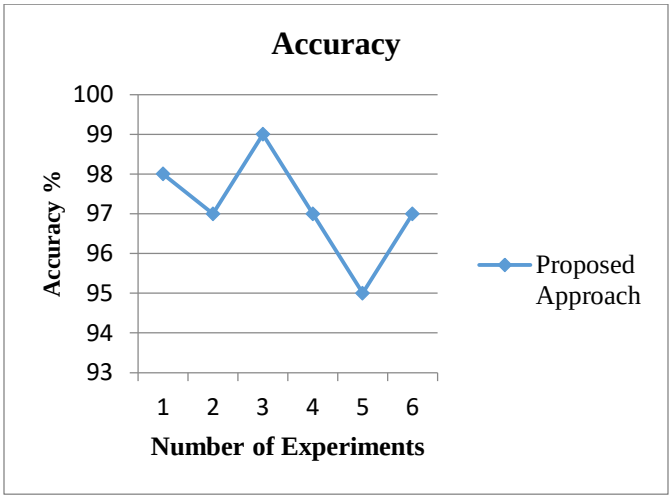

Figure 2 Accuracy

The accuracy of the proposed system for image suggestion in chat system is described using both the mode i.e. tabular and graphical manner. These achieved results of accuracy are demonstrated in percentage unit. According to the results the performance of the data model is remains consistent and not much varying with the amount of data. Thus the proposed model is acceptable for text classification and improved mobile data consumption.

B. *Error Rate*

The error rate of the proposed multi-class text classification system of user chat system is estimated in terms of percentage. The error rate of the proposed classification model is also depicted in tabular form. The error rate of any classification model demonstrates the amount of incorrectly recognized patterns over the total samples produced for classification or recognition. The error rate of the system is evaluated using the ratio of both the parameters, which is denoted as:

$$\text{Error Rate} = \frac{\text{Incorrectly Identified Pattern}}{\text{Total Pattern to Identify}} \text{X100}$$

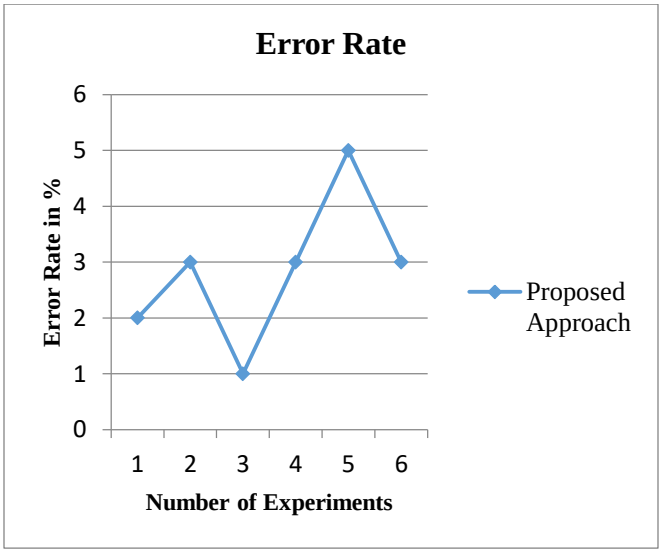

Figure 3 Error Rate

The error rate percentage of the system is demonstrated using figure 3. The error rate demonstrates the misclassification rate of the system. The misclassification rate of model is described in Y axis and number of experiments is denoted using X-axis. The achieved misclassification rate of the approach is not much fluctuating according variation of experiments. Proposed classification approach is shown using blue line. Thus the proposed model for text classification of user chat system of machine level approach is acceptable for end user applications.

C. *Time Consumption*

In order to process designed algorithm for user chat classification that the time required by the system for processing the algorithm is termed as time consumption. The time requirement of the algorithm is directly depends on the amount of data supplied for processing. This is also termed as the time complexity of the system

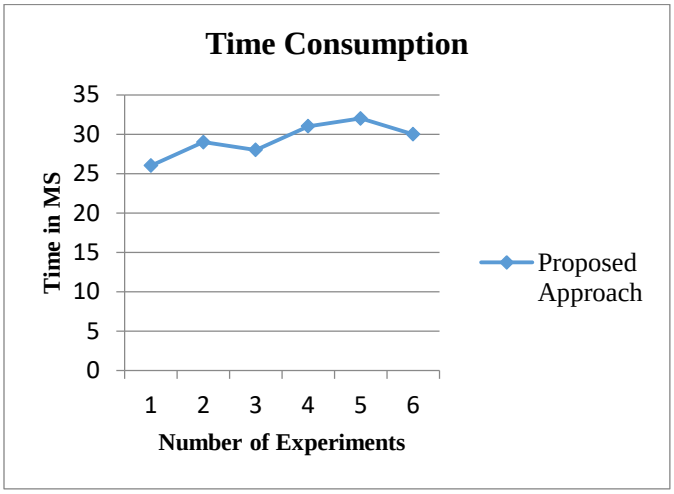

Figure 4Time Consumption

The time is measured here in terms of milliseconds (MS) and reported using figure 4. Generated graph of the proposed system express the increasing line graph where the X axis includes different experiments and the Y axis show the respective amount of consumed time. The line graph demonstrates the similar amount of time is required to process an instance of data therefore as the data increases the amount of time requirement is also increases. Thus model demonstrates the steady behavior for classification time and acceptable for the real time data classification applications.

D. Memory Usages

The main memory required to process the algorithm is known as memory usages or space complexity of the system.

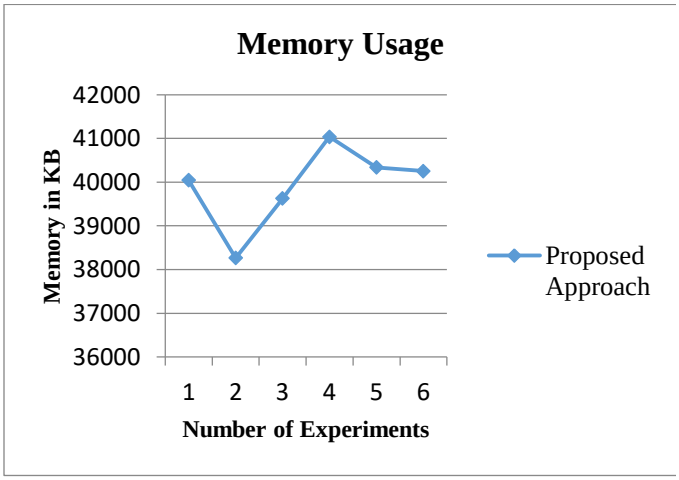

Figure 5Consumed Memories

The computation of memory is given here in terms of kilobytes (KB). Figure 5 shows the memory usages of the algorithm for classifying text of chat system. The memory requirements of the algorithm are demonstrate in Y axis of the graph and the X axis contains the different experiments. The memory requirement of the system is varies when we figure out number of experiments of similar size of data but acceptable because it is not much increasing as the amount of data.

IV. CONCLUSION AND FUTURE WORK

A. Conclusion

The main aim of the proposed work is to reduce the mobile data consumption and also reduce the mobile power consumption. In this context a data mining based chat image recommendation model is proposed for design and implementation. This model is usage the user's mobile chat history in both formats text and image. Additionally using the data mining techniques the accurate or relevant images are predicted as recommendation for the current chat session. The proposed technique first usages the K-means clustering technique to make clusters of text chat data. That is further used to create similar group of images which are closely belongs to text chat. Finally the Bays classifier is used to train or learn with the text patterns and then it is used to classify the text chat. The chat data is compared with the similar domain or clusters chat text to find the closer images for the current chat session.

B. Future Work(s)

The main objective of the proposed work to accurately predict or suggest the images for the chat sessions is accomplished successfully. In near future the following extensions are proposed for work:

- ✓ The current system only predicts the images from the history of mobile device in near future the web based image suggestion are also included.
- ✓ The current system is also extended for pre-fetching of the web pages to optimize the performance of mobile browsers.

REFERENCES

[1] Anand, M., Nightingale, E. B., And Flinn, J. Ghosts in the machine: Interfaces for better power management. In Proceedings of the 2nd International Conference on Mobile Systems, Applications and Services (Boston, MA, June 2004), pp. 23–35

[2] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H Hong, and A. K. Dey, "Factors Influencing Quality of Experience of Commonly Used Mobile Applications,"

IEEE Communications Magazine, vol. 50, pp. 48–56, 2012

[3] Ericsson. Ericsson Mobility Report: On the Pulse of the Networked Society, Stockholm, Sweden; 2014.

[4] Bing Liu, “Sentiment Analysis and Opinion Mining”, Morgan & Claypool, Synthesis Lectures on Human Language Technologies, 2012

[5] Mukherjee, Subhabrata, and Pushpak Bhattacharyya. "Sentiment analysis: A literature survey", arXiv preprint arXiv: 2013, pp. 1304.4520.

[6] Thenmalar V, Tamilselvi R, Sandhiya S3 and Dhivya Shree M, “Social Recommendation for Interactive Online System”, International Journal of Science and Research (IJSR), Volume 6 Issue 2, February 2017

[7] Loh, Stanley, et al. "Recommendation of complementary material during chat discussions." Knowledge Management & E-Learning 2.4, 2010, pp. 385-399.