International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Enhanced Classification of Service Usages with Human Trajectory Data for Location Recommendation Systems

Suryakumar B, Ramadevi E

Abstract*: The rapid growth of mobile messaging apps has led to an important process to manage social networks based on the localization of internet traffic in different types of use of in-app services. In the past researches, Improved Multi-Context Trajectory Embedding Model with Service Usage Classification Method (IMC-TEM-SUCM) has been proposed to recommend the locations based on the trajectory data of individuals and their service usage types. In this model, the traffic features were classified by using Random Forest (RF) classifier whereas the outlier was detected by clustering Hidden Markov Model (HMM). However, the RF was supervised classifier which requires knowledge about the class label of data. Also, a huge amount of data was needed to train a clustering HMM. Therefore, in this article, an IMC-TEM with Enhanced SUCM (IMC-TEM-ESUCM) is proposed in which an unsupervised classifier, namely K-means clustering is proposed to classify the service usage types. Initially, traffic flows are split into different sessions and dialogs using a combined hierarchical clustering and thresholding heuristics technique. Then, the traffic features are extracted based on the packet length and time delay. After that, K-means classification is proposed to classify the service usage types and also DBSCAN is proposed to detect the outliers. Finally, the experimental results on two different datasets show that the proposed model achieves higher performance than the existing model in terms of precision, recall, f-measure and accuracy.

Keywords: Location-based social networks, Service usage classification, Random forest, Clustering-HMM, K-means, DBSCAN .

I. INTRODUCTION

With the advent of mobile devices capable of being placed in a digitized environment, the social networks focused on positions have expanded. Location-based social networks are generally a digital mirror of human mobility in the physical world as they provide an opportunity to fully understand the spatial and temporal activities / behaviors of lifestyles of people [1].

Manuscript published on November 30, 2019.
* Correspondence Author
Suryakumar B*, Ph.D., in Computer Science at Bharathiar University, Coimbatore.
Ramadevi E , Associate Professor in Computer science at NGM College, Pollachi, India

This has made the success of instant messaging apps like Whatsapp, Facebook and WeChat the heart of several mobile user behaviors [2]. Since, they can help people chat, exchange images, details and link with business activities, such as bill payment, booking tickets and shopping.Various mobile businesses utilize messaging Apps to monetize their products. The study of the consumer use in messaging apps and social networks is therefore important to
companies because it facilitates awareness of end-user in-app habits, allowing a number of applications. Although the analysis of end user performance and app performance is detailed, the main process of in-app usage analytics is the classification of internet traffic for messaging applications into various types of usage, including text, images, audio, stream video calls, sharing locations, short videos, news feeds and outer and unknown combinations.Many methods of traffic classifications were developed through the analysis of the IP packet TCP / UDP port number or the reconstruction of the signature protocol in its payload. [3-4]. Nonetheless, the problems for the analysis of IP packet contents are discussed because communication systems use volatile port numbers. Additionally, many mobile applications use the Secure Sockets Layer (SSL) as a building block of secure communications or their Transport Layer Security (TLS) counterpart. The creation of data mining tools for the identification of coded internet traffic information generated by communication software for various types of network use addressed these challenges.The study of human path information has been proposed in previous research by IMC-TEM [5]. In this method, the parameters needed to increase the reliability of both the contextual system and the social connections were calculated through a frog leaping optimization algorithm [6]. Nevertheless, only the characterization of contexts for various applications was considered. Also, it needs to analyze the service usage characteristics to enhance the location recommendation. As a result, a Service Usage Classification Model (SUCM) has been proposed including IMC-TEM based on the encrypted Internet traffic analysis [7]. Here, IMC-TEM and CUMMA model were combined to classify human trajectory data with internet traffic characteristics. Four major processes were performed such as traffic partition, feature extraction, service usage type classification and outlier's detection. The network traffic was split into time periods and dialogs based on the combined hierarchical clustering thresholding heuristics. Then, the traffic features were extracted and classified by using the random forest classifier to classify the service usage types. Also, the outliers are detected and removed by using the clustering-Hidden Markov Model (HMM).

However, the RF was supervised classifier which requires prior knowledge about the class label of data. Also, a huge amount of data was needed to train a clustering HMM.

Therefore in this article, IMC-TEM-ESUCM is proposed based on the K-means clustering which is unsupervised classifier to classify the service usage types. Initially, traffic flows are split into different sessions and dialogs using a combined hierarchical clustering and thresholding heuristics technique.

Then, the traffic features are extracted based on the packet length and time delay. After that, K-means classification is proposed to classify the service usage types and also DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is proposed to detect the outliers. Thus, the location recommendation is enhanced by using both human trajectory data and internet traffic features.The rest of the article is organized as follows: Section II describes the works related to the methods of classification of Internet traffic and use of services. Section III explains the methodology of the proposed model. Section IV describes the performance efficiency of the proposed model. Section V concludes the research work.

II.LITERATURE SURVEY

Identification of the different user patterns of smartphone apps [8] was suggested on the basis of network measurements from a United States (US) regional tier-1 cellular network provider. In addition, correlations are found throughout the apps in terms of regional reach, trends of diurnal usage, etc. In addition, to classify user behavior, the diurnal patterns of different genres and app mobility have been discovered. Much traffic may be lacking, though, as it does not utilize generic application URL API. Detection inaccuracy may also occur due to the time difference between fields of the user-agent.Regarding user content and actions on the mobile Internet, an empirical study [9] was suggested. The main objective was to analyze whether the two activities had a positive or negative interdependence. A unique panel dataset was then used, consisting of mobile internet usage data at the individual level that includes the generation of individual multimedia content and usage behavior. This knowledge was then combined with user calling pattern data to build their social network and compute their geographic mobility. In addition, a simultaneous equation panel model of individual levels was developed to control the various sources of social network endogeneity. However, the output was not successful due to lack of data as it does not have uploaded or downloaded details about the particular type of content and the destination websites.Using correlation data, the classification of network traffic [10] was suggested. A new, non-parametric method has been suggested by integrating a comparison of traffic flows in this methodology of traffic identification. On the basis of this method, the quality of the graded specimens has been enhanced and evaluated. Examples of the conceptual and experimental features were also given. However, this approach requires some previous knowledge in order to classify traffic flows.The unknown flow detection was proposed to provide an effective network transport classification [11]. This method's main purpose was to avoid unknown flow problems in a semi-monitored network. The flux comparison has been integrated in this approach into a semi-monitored system of uncertain flow detection capabilities. In addition, to address the problem of a limited controlled training set, the stream tag propagation was introduced for automatically marked suitable flows from a broad unlabelled data set. In addition, the classification of compounds was proposed to mutually identify the correlated flows to improve the accuracy of classification. However, this approach was not suitable for traffic classification across the network.

A study was carried out on the classification of mobile Internet user behavior [12]. In this study, the mobile user behaviours, including data use, movability patterns and application use, were classified from three characteristics. Furthermore, traffic heavy users and the pattern of mobility were observed almost related to users' access behaviour. Apps can be grouped by their software usage actions and user experience should distinguish device styles. However, fairness was less and congestion in the network was not controlled.A participatory approach to cultural analysis focused on shared interaction information in locational social networks [13]. The participatory perceived consumer activity knowledge was initially obtained from the social networks based on location. Then a pragmatic classification approach for sorting inadmissible users was introduced because only local users could be cultured mapping. After that, the cultural clustering method to discover cultural clusters has been extracted from day to day activity, mobility, and linguistic perspectives correspondingly. Eventually, on the world map the cultural groups were shown. Nonetheless, because of the compilation from a particular social network, consumer activity information was less.

III. PROPOSED METHODOLOGY

In this section, the proposed IMC-TEM-ESUCM model is briefly explained. At first, the Internet traffic is split into sessions and dialogs by the hierarchical clustering with thresholding heuristics scheme. When the client opens the mobile App and closes it, a session is started. During this session, the Internet traffic generated is called the dialog. Most dialogs are single-type use, such as sharing text and location, while the other dialogs are mixed uses. To classify these segmented dialogs into single-type usages or outliers, a service usage predictor is used. The classification of service usages is then done using a K-means classification, while DBSCAN detects and removes the outliers i.e., mixed forms of uses.

3.1 Segmentation of Internet Traffic

Initially, traffic network data is collected for the various applications in mobile Apps. Once these sample data is gathered, two-stage segmentation is done, i.e. from traffic flow to session and from session to dialog, to segment these traffic flows from coarse-grained level, i.e. session to fine-grained level, i.e. dialog. Next, using the thresholding process, the measured traffic-flow is segmented into multiple sessions. Here, there may be several consecutive dialogs in each session. The idea of hierarchical clustering with a bottom-up segmentation algorithm is implemented for segment dialogs. The time period of the session is separated into several limited intervals. The intervals with null traffic are then processed while the intervals of non-zero are interpreted as a series of initial dialog.

Three merge functions are also implemented to preserve neighboring intervals in a single dialog if they correspond to one application category or mixed types and are: merging heavy-heavy dialog pairs, light-heavy dialog pairs and maximum dialog pairs. It avoids a situation where one application of network traffic is split into two dialogs of the same types of use.

3.2 Extraction of Traffic Features

The dialog encodes two types of information as a sequence of packet length and a sequence of time delays after segmenting the traffic-flow. After that, the discriminative features are extracted according to the packet size and time delay. The size of the packet is determined by the packet length in terms of bytes to be sent. Time delay refers to the time required to enter the destination for the packet.

- Features related to the packet size: Standard variation, median min, max, skewness, kurtosis and mean error are measured as features for a specified packet length. It also defines the minimum and maximum IP packet size values. The range from minimal to peak is then separated into sub-ranges of K of equal size. To achieve a K-size feature vector for each sub-range, the amount of packets is determined, each of which represents the percentage of packets of duration in the k^{th} sub-range. However, the amount of packets whose lengths exceed the lengths of their successive packets are determined with a significant margin as a feature to describe the sequence variance. The longest monotone including both that and declining subsequences of a dialog is also analyzed and the lengths of these subsequences are used to describe the network traffic pattern and skew. The array of packet sizes is also classified and separated into sub-ranges of K of equal size. For each element in the packet length, when its value is in the k^{th} sub-range, this element is substituted by k. On this basis, this sequence is mapped into a new sequence of strings. By using this string series, all continuous subsequences will be classified with size varying from 3 to 20 and their related number of occurrences will also be calculated in this string sequence. Ultimately, as properties, the top-N number of instances is used.
- Features related to the time delay: The time interval is extracted for each two consecutive packets, resulting in a sequence of time delays.

3.3 Classification of Service Usage Types using K-means Clustering Classifier

After the feature extraction is completed, the extracted features are applied to the K-means classification to classify the service usage types. Initially, k values are randomly selected as cluster centroids and all training features are inserted to the nearest cluster. After all features have been inserted to the clusters, the centroids i.e., representing the mean of the features of each cluster are recomputed and with these recomputed centroids, the new centroids of their respective clusters are obtained. As a result, all cluster membership is updated and all features of the training set are inserted to their nearest, maybe re-centered cluster. This iterative process is continued until there is no alter to the centroids or their membership and the clusters are considered settled. Convergence is achieved once the recomputed centroids equal to the past iteration's centroids. The measure of distance is Euclidean in k-means which given two features in the form of (i, j) can be denoted as:

$$d = \sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2} \qquad (1)$$

The algorithm for classifying the traffic features using k-means clustering-based classifier is given below.

Algorithm:

Input: F, k //F: set of features to be classified and k: integer
Output: k clusters
Begin
Initialize k random centroids;
Do again

```
for(all features f in F)
 shortest ← 0;
 membership ← NULL;
 for(all centroid C)
   d ← Distance(C); //d: distance
   if(d < shortest)
     shortest ← d;
     membership ← C;
   end if
 end for
end for
```

Update centroids C;
Until convergence
End

3.4 Detection and Prevention of Outliers

The predicted unknown dialogs were classified as outliers or an unpredictable use combination. The most probably unknown mix is identified for a given dialog. A clustering-based algorithm such as DBSCAN is used for detecting and handling the outliers i.e., mixed usages from the multiple single-type sub-dialogs. This algorithm is based on density-reachability and density-connectivity concepts. These two concepts are depending on input parameters such as epsilon (eps) and the least number of points $(minPts)$. Epsilon defines the distance around a feature that defines its eps-adjacent.

For a given feature f, when the number of features within the eps-adjacent is at least $minPts$, then g is defined as a core feature. All features within its eps-adjacent are referred to be directly density-reasonable from f. Also, a feature e is said to be density-reasonable if it is within the eps-adjacent of a feature that is directly density-reasonable or density-reasonable from f. A cluster is defined as the group of features in a dataset that are density-connected to a specific core feature. Any feature that is not part of a cluster is detected as mixed usage or outlier.

Thus, the mixed usage types are detected as outliers and neglected to recommend the locations.

DBSCAN Algorithm:

- At first, all features in the dataset are considered to be unassigned.
- Then, an arbitrary unassigned feature e is selected from the dataset.
- If it finds e is a core feature, it finds all the density-connected features based on eps and $minPts$.
- It assigns all these features to a new cluster.
- If it finds e is not a core feature, then e is considered to be outlier and it removes from the feature set.
- Once each feature is assigned, the algorithm terminates.

IV. EXPERIMENTAL RESULTS

- In this section, the effectiveness of the IMC-TEM-ESUCM model is evaluated and compared with the existing IMC-TEM-SUCM model in terms of accuracy, precision, recall and f-measure. The experiments are carried out for both proposed and existing models MATLAB 2017b based on two open geo-social networking datasets are used, namely $Foursquare_L$ and $Gowalla$. Effectiveness of location recommendation along both common and time-aware locality suggestion is evaluated by considering home-city recommendation settings. Here, 20% trajectories with only home-city locations are selected as a testing dataset and the residual trajectories are chosen as training data. Table 1 gives the basic statistics of the considered datasets.

.

Table.1 Statistics of Datasets

Dataset	Number of Users	Number of Check-ins	Number of Links	Number of Locations
$Foursquare_L$	266909	33278683	-	3680126
$Gowalla$	216734	12846151	736778	1421262

4.1 Precision

It is computed based on the ratio of predicted locations at True Positive (TP) and False Positive (FP) rates.

$$Precision = \frac{TP}{TP+FP}$$

The performance of general and time-aware location recommendation on different datasets using precision is given in Table 2.

Table.2 Performance of Location Recommendation using Precision

Dataset	Precision			
	General Location Recommendation		Time-aware Location Recommendation	
	IMC-TEM-SUCM	IMC-TEM-ESUCM	IMC-TEM-SUCM	IMC-TEM-ESUCM
$Foursquare_L$	0.932	0.942	0.800	0.812
$Gowalla$	0.940	0.951	0.815	0.826

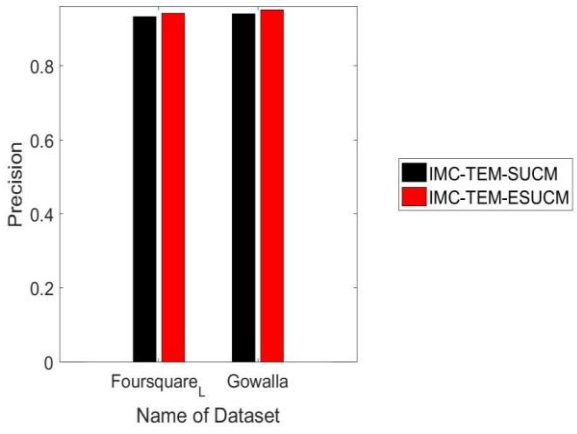

Figure.1 Comparison of Precision for General Location Recommendation

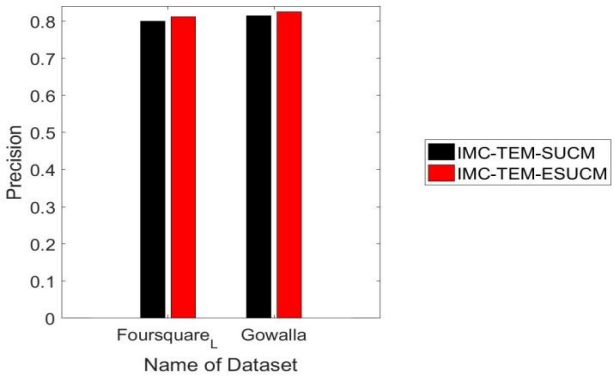

Figure.2 Comparison of Precision for Time-aware Location Recommendation

Figure 1 & 2 shows the precision of IMC-TEM-SUCM and IMC-TEM-ESUCM for general and time-aware location recommendation on different datasets, respectively. From this analysis, the IMC-TEM-ESUCM model has a better precision for both general and time-aware location recommendation system compared to the IMC-TEM-SUCM.

4.2 Recall

It is computed based on the ratio of predicted locations at TP and False Negative (FN) rates.

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

The performance of general and time-aware location recommendation on different datasets using recall is given in Table 3.

Table.3 Performance of Location Recommendation using Recall

Dataset	Recall			
	General Location Recommendation		Time-aware Location Recommendation	
	IMC-TEM-SUCM	IMC-TEM-ESUCM	IMC-TEM-SUCM	IMC-TEM-ESUCM
$Foursquare_L$	0.944	0.952	0.781	0.793
Gowalla	0.950	0.960	0.790	0.802

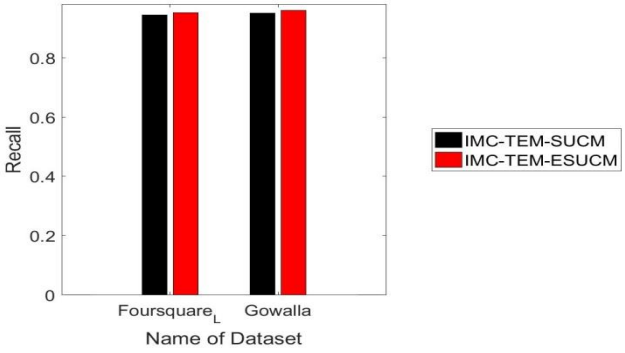

Figure.3 Comparison of Recall for General Location Recommendation

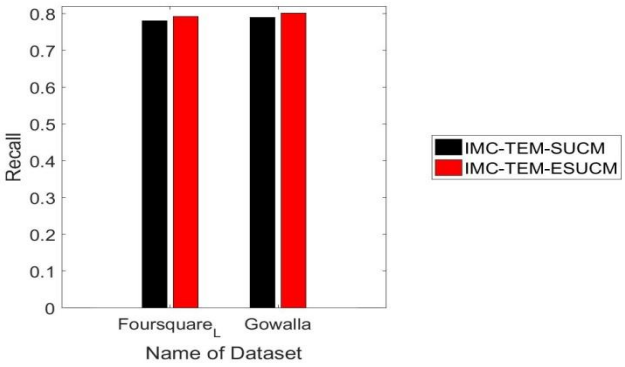

Figure.4 Comparison of Recall for Time-aware Location Recommendation

Figure 3 & 4 shows the recall of IMC-TEM-SUCM and IMC-TEM-ESUCM for general and time-aware location recommendation on different datasets, respectively. From this analysis, the IMC-TEM-ESUCM model has a better recall for both general and time-aware location recommendation system compared to the IMC-TEM-SUCM.

4.3 F-measure

It is computed by using both precision and recall as follows:

$$F - measure = 2 \cdot \left(\frac{Precision \cdot Recall}{Precision + Recall}\right)$$

The performance of general and time-aware location recommendation on different datasets using f-measure is given in Table 4.

Table.4 Performance of Location Recommendation using F-measure

Dataset	F-measure			
	General Location Recommendation		Time-aware Location Recommendation	
	IMC-TEM-SUCM	IMC-TEM-ESUCM	IMC-TEM-SUCM	IMC-TEM-ESUCM

$Foursquare_L$	0.786	0.795	0.542	0.553
Gowalla	0.781	0.792	0.547	0.558

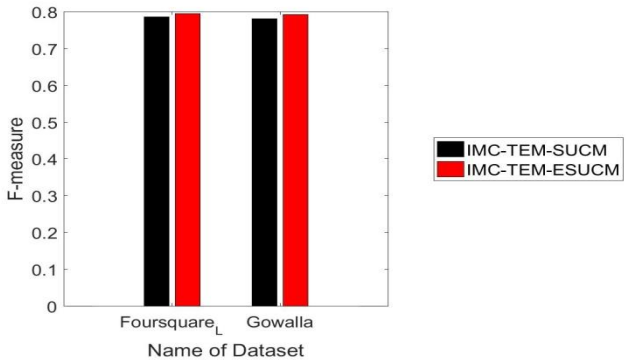

Figure.5 Comparison of F-measure for General Location Recommendation

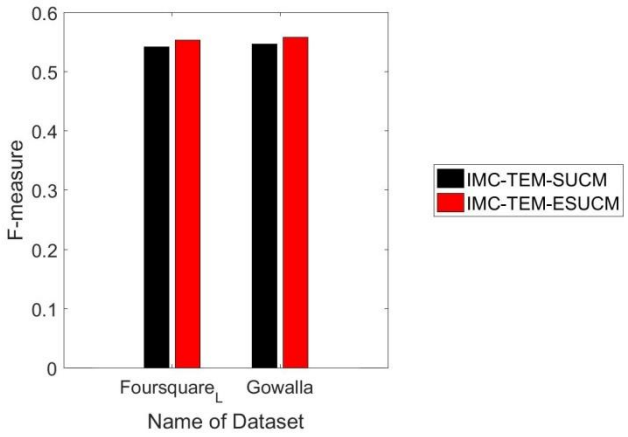

Figure.6 Comparison of F-measure for Time-aware Location Recommendation

Figure 5 & 6 shows the f-measure of IMC-TEM-SUCM and IMC-TEM-ESUCM for general and time-aware location recommendation on different datasets, respectively. From this analysis, the IMC-TEM-ESUCM model has a better f-measure for both general and time-aware location recommendation system compared to the IMC-TEM-SUCM.

4.4 Accuracy

It is computed based on the TP and True Negative (TN) among total number of locations predicted.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

The performance of general and time-aware location recommendation on different datasets using accuracy is given in Table 5.

Dataset	Accuracy (%)			
	General Location Recommendation		Time-aware Location Recommendation	
	IMC-TEM-SUCM	IMC-TEM-ESUCM	IMC-TEM-SUCM	IMC-TEM-ESUCM
$Foursquare_L$	92	92.58	81.67	81.81
Gowalla	91.5	91.94	82.38	82.46

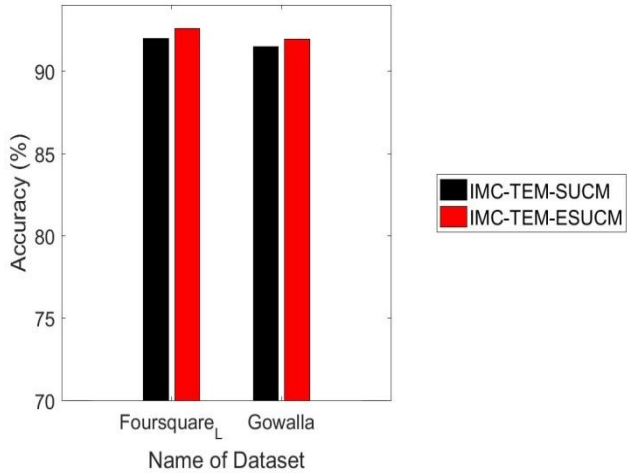

Figure.7 Comparison of Accuracy for General Location Recommendation

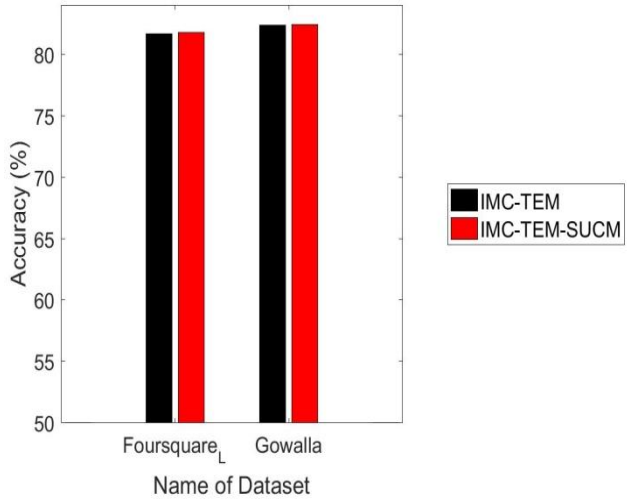

Figure.8 Comparison of Accuracy for Time-aware Location Recommendation

Figure 7 & 8 shows the accuracy of IMC-TEM-SUCM and IMC-TEM-ESUCM for general and time-aware location recommendation on different datasets, respectively. From this analysis, the IMC-TEM-ESUCM model has a better accuracy for both general and time-aware location recommendation system compared to the IMC-TEM-SUCM.

V. CONCLUSION

In this article, an IMC-TEM-ESUCM model is proposed to increase the location recommendation by classifying the Internet traffic with the human trajectory data. Initially, the traffic-flows of service usages are gathered with the corresponding usage types by users. Then, the traffic from traffic-flows is hierarchically segmented into sessions and to dialogs. After that, the traffic features are extracted and given to the K-means clustering-based classifier to classify the service usage types. Besides, the anomalous dialogs with mixed usages are detected and neglected from the feature set using the DBSCAN based clustering method. Finally, the experimental results demonstrated that the performance of the proposed IMC-TEM-ESUCM model achieves better performance for recommending the locations than the existing IMC-TEM-SUCM model.

REFERENCES

1. Yuan, N. J., Zhang, F., Lian, D., Zheng, K., Yu, S., & Xie, X. (2013). We know how you live: exploring the spectrum of urban lifestyles. In Proceedings of the first ACM conference on Online social networks (pp. 3-14).
2. Rice, E. (2010). The positive role of social networks and social networking technology in the condom-using behaviors of homeless young people. Public health reports, 125(4), 588-595.
3. Sen, S., Spatscheck, O., & Wang, D. (2004). Accurate, scalable in-network identification of p2p traffic using application signatures. In Proceedings of the 13th international conference on World Wide Web (pp. 512-521). ACM.
4. Haffner, P., Sen, S., Spatscheck, O., & Wang, D. (2005). ACAS: automated construction of application signatures. In Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data (pp. 197-202). ACM.
5. Suryakumar, B., & Ramadevi, E. (2017). A multi context embedding model based on convolutional neural network for trajectory data mining. International Journal of Computer Science and Mobile Applications, 5(9), 1-9.
6. Suryakumar, B., & Ramadevi, E. (2018). An improved multi-context trajectory embedding model using parameter tuning optimization for human trajectory data analysis. International Journal of Applied Engineering Research, 13(22), 15633-15637.
7. Suryakumar, B., & Ramadevi, E. (2018). Human trajectory data and internet traffic mining using improved multi-context trajectory embedding service usage classification model. International Journal of Engineering & Technology, 7(4), 3538-3542.
8. Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J., & Venkataraman, S. (2011). Identifying diverse usage behaviours of smartphone apps. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (pp. 329-344). ACM.
9. Ghose, A., & Han, S. P. (2011). An empirical analysis of user content generation and usage behavior on the mobile Internet. Management Science, 57(9), 1671-1691.
10. Zhang, J., Xiang, Y., Wang, Y., Zhou, W., Xiang, Y., & Guan, Y. (2013). Network traffic classification using correlation information. IEEE Transactions on Parallel and Distributed Systems, 24(1), 104-117.
11. Zhang, J., Chen, C., Xiang, Y., Zhou, W., & Vasilakos, A. V. (2013). An effective network traffic classification method with unknown flow detection. IEEE Transactions on Network and Service Management, 10(2), 133-147.
12. Yang, J., Qiao, Y., Zhang, X., He, H., Liu, F., & Cheng, G. (2015). Characterizing user behavior in mobile internet. IEEE transactions on emerging topics in computing, 3(1), 95-106.
13. Yang, D., Zhang, D., & Qu, B. (2016). Participatory cultural mapping based on collective behavior data in location-based social networks. ACM Transactions on Intelligent Systems and Technology (TIST), 7(3), 30.

AUTHORS PROFILE

B. Suryakumar received M.Sc and MCA degrees in Computer Science from Annamalai University, Tamilnadu. Currently, he is doing Ph.D., in Computer Science at Bharathiar University, Coimbatore. His research interest **lies in the area of Data Mining, Networking and Data security.**

Dr. E.RamaDevi received PhD degree in Computer Science from Mother Teresa Womens University, Kodaikanal. Currently she is an Associate Professor in Computer science at NGM College, Pollachi, India. She has got 15 years of research experience and has more than 23 years of teaching experience. Her research interest includes areas like Big Data, Data Mining, Knowledge base System, Intelligent and Control System and Fuzzy Logic. She has presented various papers in national and International Conferences and published 20 research papers on refereed journals.