An Enhancement on Mobile Social Network using Social Link Prediction with Improved Human Trajectory Internet Data Mining

B. Suryakumar[1]
Ph.D. Research Scholar, Department of Computer Science
NGM College, Pollachi
Tamilnadu, India

Dr. E. Ramadevi[2]
Associate Professor, Department of Computer Science
NGM College, Pollachi
Tamilnadu, India

***Abstract*—Generally, the mobile social network has missing and unauthentic links. The prediction of those links is one of the major problems to understand the relationship between two nodes and recommends the potential links to the users derived from the history of user-link interactions and their contextual information. The recommendation problem can be modeled as prediction of the future links between users. Many research works have been developed to understand the relationship between the nodes and construct the models for missing or suspicious links prediction. Among those, Improved Multi-Context Trajectory Embedding Model with Service Usage Classification Model (IMC-TEM-SUCM) has better enhancement on human trajectory data mining by classifying the internet traffic. However, this method requires the prediction of the relationship between the nodes and social links. Hence in this article, the IMC-TEM-SUCM is proposed with the Social Link Prediction (SLP) mechanism for identifying the relationship between two nodes and predicting the stable links. In this technique, a number of nodal features are considered and their influence on the link prediction problem of Foursquare and Gowalla are examined. This extended network is used for computing two features such as optimism and reputation that depict node's characteristics in a signed network. After that, meta-path-based features are considered and their influence of the route length on the problem of link prediction is examined. Moreover, a link prediction process is performed by using the machine learning classification algorithms that use the extracted node-based and meta-path-based features. Also, Cosine coefficient and Jaccard coefficient similarity-based techniques are used for computing the similarity index between any two nodes. A higher similarity indicates a higher chance of forming links between them. Finally, the performance effectiveness of the proposed model is evaluated through the experimental results using different real-world datasets.**

Keywords*—Mobile social network; improved multi-context trajectory embedding model with service usage classification model; social link prediction; machine learning; cosine coefficient; Jaccard coefficient

I. INTRODUCTION

In modern years, location-based social networks have increased due to the emerging of site-enabled mobile devices. Generally, location-based social networks are a digital mirror to human mobility in physical world since it offers a chance to completely understand the spatial and temporal activities/behaviors of people's lifestyles [1]. As a result, the popularity of mobile social Apps can support people to communicate with each other, share photos, information and connect with commercial activities. Different mobile industries monetize their services in messaging Apps. Thus, the service usage analytics in messaging Apps or location-based social network becomes essential for commerce since it can support recognize in-App behaviors of end users and so several applications are enabled. Though it provides in-depth analysis into end users and App performances, a primary process of in-App usage analytics are classifying Internet traffic of messaging Apps into different usage types such as services, locations, etc., and outlier or unknown combination of usage.

Many traffic classification methods have been developed by analyzing TCP/UDP port numbers of an IP packet or recreating protocol signatures in its payload [2-3]. But, the challenges were addressed for examining IP packet content since messaging Apps use unpredictable port numbers. Additionally, several mobile Apps use the Secure Sockets Layer (SSL) and its successor Transport Layer Security (TLS) as a building block for encrypted transmissions. Such challenges were tackled by developing data mining solutions to classify the encrypted Internet traffic data generated by messaging Apps into different service usage types. In previous researches, MC-TEM was proposed to analyze the human trajectory data [4]. In this model, CNN was used to learn the parameters. Moreover, IMC-TEM was proposed that uses a frog-leaping optimization algorithm for tuning the parameters which are needed to improve the accuracy of the contextual model and social link prediction [5]. On the other hand, it considers only the characterization of types of contexts for different Apps. It requires Internet traffic classification to jointly analyze service usage behaviour as well to enhance the location recommendation and social-link prediction performances. As a result, IMC-TEM-SUCM was proposed [6] by classifying internet traffic using Random Forest (RF) classifier. However, this method requires the prediction of the relationship between the nodes and social links to find any link failure between two nodes.

Therefore in this paper, the proposed SLP-IMC-TEM-SUCM is proposed to predict the social links for a better understanding of the relationship between two nodes. In this technique, a number of nodal features are considered and their

influence on the link prediction problem of Foursquare and Gowalla are examined. This extended network is used for computing two features such as optimism and reputation that describe nodes characteristics in a network. After that, meta-path-based features are considered and their route length influence on the problem of link prediction is also examined. Moreover, a link prediction process is performed by using the machine learning classification algorithms that use the extracted node-based and meta-path-based features. Also, Cosine coefficient and Jaccard coefficient similarity-based techniques are used for computing the similarity index between any two nodes. A higher similarity indicates a higher possibility of making links between them. Thus, the stable link is predicted to reduce the average delay during packet reception in mobile social networks.

The remaining article is prepared as follows: Section II presents the works related to the social link prediction techniques. Section III explains the proposed methodology. Section IV shows the performance evaluation of the proposed technique. Section V concludes the research work.

II. Literature Survey

Efficient routing in mobile social networks [7] was proposed by exploiting friendship relations. In this technique, a new metric was introduced to detect the value of friendships between nodes accurately. According to this metric, the neighbourhood of each node was defined as the set of nodes having close friendly relations with that specific node either directly or indirectly. Moreover, a friendship-based routing was presented to periodically differentiate the friendship relations used in the broadcasting of messages. However, the complexity and maintenance cost of this model was high.

The problem of search with local information was addressed [8] in joint social and communication networks. In this model, the end-to-end delay distribution and success probability were derived to differentiate the delay and success probability on different social links. Moreover, greedy routing algorithms were used to enhance the delay and chain completion success. Next, this analysis was extended to the joint social and communication networks. However, the complexity of this analysis was high.

A link prediction algorithm [9] was proposed on the social network. In this study, two improved algorithms were proposed such as CNGF algorithm based on neighbourhood information and KatzGF algorithm based on the overall information of the network. Moreover, the link prediction algorithm was proposed based on the information about nodes multiple attributes which defects the inactive social network. However, it does not consider the service usage types to improve the LBSN recommendation.

User's location prediction method [10] was proposed in LBSN. In this study, a model was proposed that influences the global temporal preferences and spatial correlation for predictions. Here, global temporal preferences were used to exploiting the historical check-ins of other users that models the temporal reputation of locations. The spatial correlation was used to estimate the distance that a user was willing to stay a site based on his/her current site. However, the effectiveness was less and required to further improvement based on the user's location.

Characterization of user behavior in the mobile internet [11] was studied. In this study, the mobile user behaviors were classified from three characteristics such as data use, mobility patterns and application use. Also, the traffic heavy users and the mobility pattern were observed as nearly associated with the application access characteristic of the users. Users may be clustered via their application use characteristic and application types may be recognized through an interaction of the users. However, fairness was less and network congestion was not controlled.

A link prediction method in LBSN [12] was proposed by developing social and mobility patterns. In this study, three new methods were proposed by merging social and mobility patterns such are: interior and exterior of commonplaces, familiar neighbours of places and total and partial overlapping of places. However, the location prediction was required to suggest locations that users could stay.

III. Proposed Methodology

In this section, the proposed IMC-TEM-SUCM with SLP algorithm is explained in brief. In this technique, node and meta-paths-based features are considered for predicting the social network links. The link prediction problem of considered datasets is solved based on the following process:

The link prediction problem is modeled as the classification process where there are two classes in which one class is used for the existence of future links and the other class is used for the non-existence of future links. For this case, a number of features namely node-based and meta-path-based features should be defined for utilizing in the classifiers.

- Node-based Features: The node-based features are applied on the Foursquare (G_F) and Gowalla (G_G) layer wherein the network infrastructure is modified via giving appropriate signs to the links. Every link in G_F or G_G is given a positive or negative sign by considering two nodes n_1 and n_2 in G_F or G_G. If they pursue each other, then there are two directed links between them and a positive sign is assigned on both directions. If one node follows the other node, then a positive sign is assigned in the actual direction and a negative sign is assigned in the opposed direction. According to this, G_F or G_G becomes a network with both positive and negative signs.

For a given network structure, a priori information on the link existence between a source and the destination node, the difficulty is to understand the sign of this link. Node-based feature involves the reputation that refers to the reputation of a node in the mobile social network. A higher reputation for a node indicates the node becomes more tolerable by other users and is expected to be followed. The reputation value for node n_1 is calculated by considering the positive and negative incoming links to the node. Consider, the number of positive incoming links and the number of negative incoming links to n_1 are $d_{in}^{+}(n_1)$ and $d_{in}^{-}(n_1)$. The Normalized Reputation (NR) of n_1 is computed as follows:

$$NR = \frac{d_{in}^{+}(n_1) - d_{in}^{-}(n_1)}{d_{in}^{+}(n_1) + d_{in}^{-}(n_1)} \quad (1)$$

Another node feature is optimism that can be calculated related to the reputation. Nodes with higher optimism values indicate that they are expected to follow others. The number of positive outgoing links and the number of negative outgoing links from n_1 are denoted as $d_{out}^{+}(n_1)$ and $d_{out}^{-}(n_1)$. The Normalized Optimism (NO) of n_1 is computed as follows:

$$NO = \frac{d_{out}^{+}(n_1) - d_{out}^{-}(n_1)}{d_{out}^{+}(n_1) + d_{out}^{-}(n_1)} \quad (2)$$

Here, the link between n_1 and n_2 are predicted by using common neighbors of these nodes $\{CN(n_1, n_2)\}$ that refers to the number of nodes with links to both n_1 and n_2. Thus the reputation and optimism values of n_1 and n_2 such as $NR(n_1), NR(n_2), NO(n_1)$ and $NO(n_2)$ are utilized as node-based features.

- Meta-Path-based Features: This feature is used for capturing the interlayer connectivity information. A route between two users in a mobile social network has important data about their link such as relationship. Initially, meta-paths between two target nodes are mined up to a fixed path length according to the traversing the target network method like breadth-first search. By using cluster-based meta-path, a meta-path with length 2 from n_1 to n_2 in cluster C is defined as n_1 follows n_x $(x \neq 2)$ which is friends with n_2

In this proposed technique, cluster-based meta-paths with length 2, 3 and 4 are employed as features. For each pair of nodes n_1 and n_2, all cluster-based meta-paths of different lengths that pass via every cluster in G_F or G_G are computed with having one of the following conditions:

- A route from n_1 to C has not more than a length i and located in G_F or G_G. Additionally, there exists a route from C with length 1 which is located on G_F or G_G.
- A route from n_1 to C has not more than a length 1 and located in G_F or G_G. As well, there exists a route from C with not more than length i and is located on G_F or G_G.

Three different lengths are considered for the cluster-based meta-paths along with the above conditions. As a result, a set of six features is obtained for every pair of nodes. Once all features are extracted, Support Vector Machine (SVM), Naive Bayes and K-Nearest Neighbor (KNN) classifiers are used to predict the stable social links. Among these classifiers, SVM creates a set of hyperplanes in a high dimensional space and splits different classes by optimizing the space to the functional margins. Also, Gaussian kernel function maps the actual finite dimensional space into a higher dimensional space. Similarly, Naive Bayes generates the number of algorithms in which the particular feature value is considered as an independent of other features. To train the parameters of this model, maximum-likelihood estimation is used. Another classification method KNN has the K nearest training samples in the feature space and locally approximates the function. A weighted average of the functions is computed and the value of K is optimized to train this classifier.

Finally, the performance of these classifiers is compared with a number of baseline links prediction methods such as cosine coefficient and Jaccard coefficient similarity methods. These methods compute a similarity index between any two nodes and a higher similarity indicates a higher possibility of making links between them. Thus, the stable link is predicted to reduce the average delay during packet reception in mobile social networks.

IV. Experimental Results

This section presents the experimental results of the proposed SLP-ICM-TEM-SUCM by considering Cosine coefficient and Jaccard coefficient similarity techniques using MATLAB 2018a and compared in terms of accuracy, Mean Squared Error (MSE) and Mean Absolute Error (MAE). Here, two open geo-social networking datasets namely $Foursquare_S$ [13] and $Gowalla$ [14] are used for link prediction. Table I gives the basic statistics of the considered datasets.

A. Accuracy

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

$$Accuracy = \frac{TP+TN}{TP+TN+False\ Positive\ (FP)+FalseNegative\ (FN)} \quad (3)$$

Fig. 1 shows the comparison of SLP-IMC-TEM-SUCM for different similarity-based techniques such as cosine coefficient and Jaccard coefficient techniques in terms of accuracy (%). When considering $Gowalla$ dataset, the accuracy of the Jaccard coefficient technique is 12.21% higher than Cosine coefficient technique. Similarly, the Jaccard coefficient technique achieves 11.91% higher accuracy than the Cosine coefficient technique. From the analysis, it is observed that the SLP-IMC-TEM-SUCM using Jaccard coefficient-based similarity technique has high accuracy than the Cosine coefficient-based similarity technique for both datasets.

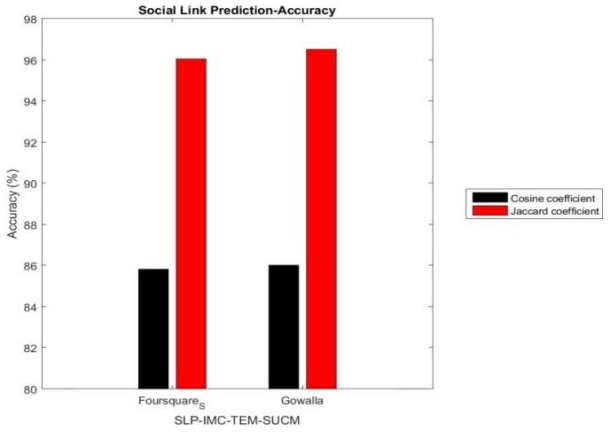

Fig. 1. Comparison of Accuracy.

TABLE I. STATISTICS OF DATASETS

Dataset	No. of Users	No. of Check-ins	No. of Links	No. of Locations
$Foursquare_S$	4163	483814	32512	121142
Gowalla	216734	12846151	736778	1421262

B. Mean Squared Error (MSE)

It is the expected variation of the errors between the real and predicted values.

Fig. 2 shows the comparison of MSE for SLP-IMC-TEM-SUCM using different similarity-based techniques such as cosine coefficient and Jaccard coefficient techniques. When considering *Gowalla* dataset, the MSE of the Jaccard coefficient technique is 47.58% less than Cosine coefficient technique. Similarly, the Jaccard coefficient technique achieves 46.22% less MSE than the Cosine coefficient technique. From the analysis, it is observed that the SLP-IMC-TEM-SUCM using Jaccard coefficient-based similarity technique has less MSE than the Cosine coefficient-based similarity technique for both datasets.

C. Mean Absolute Error (MAE)

It defines the mean absolute variation between the real and predicted values.

Fig. 3 shows the comparison of MAE for SLP-IMC-TEM-SUCM using different similarity-based techniques such as cosine coefficient and Jaccard coefficient techniques. When considering *Gowalla* dataset, the MAE of the Jaccard coefficient technique is 47.37% less than Cosine coefficient technique. Similarly, the Jaccard coefficient technique achieves 46.55% less MAE than the Cosine coefficient technique. From the analysis, it is observed that the SLP-IMC-TEM-SUCM using Jaccard coefficient-based similarity technique has less MAE than the Cosine coefficient-based similarity technique for both datasets.

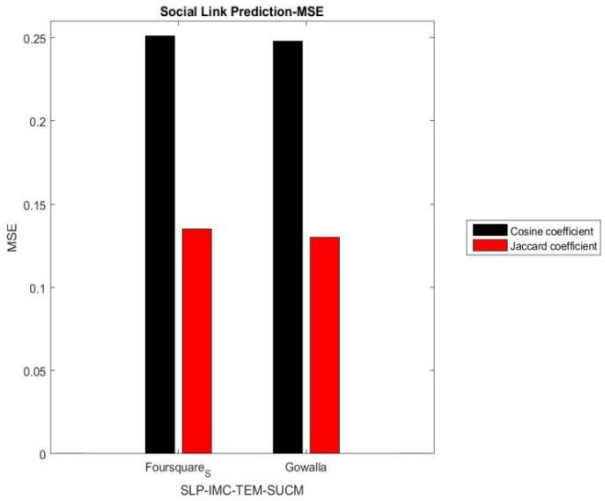

Fig. 2. Comparison of MSE.

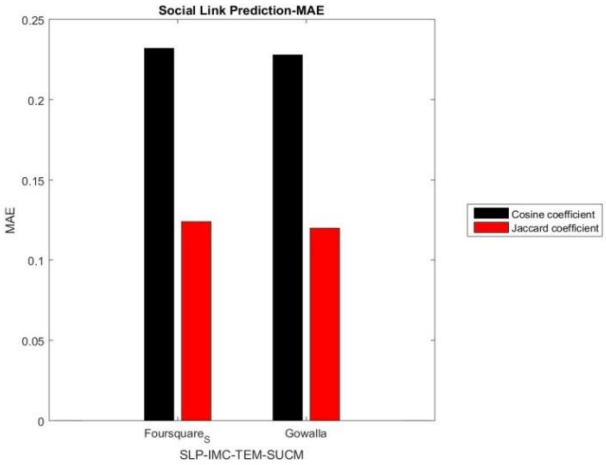

Fig. 3. Comparison of MAE.

V. CONCLUSION

In this article, SLP-IMC-TEM-SUCM is proposed to predict the missing or suspicious links. In this technique, the problem of link prediction in social networks is modeled as the classification process using two classes; one is for the existence of forthcoming links and other one is for the non-existence of forthcoming links. To achieve efficient classification, node and meta-path-based features are considered and extracted. After that, different classifiers are applied to classify these features. Further, the outcomes of each classifier are analyzed based on the similarity measure such as Cosine and Jaccard coefficient. A higher similarity indicates the higher probability of creating links between two users. Thus, the stable link is predicted to reduce the average delay during packet reception in mobile social networks. Finally, the experimental results demonstrate that the proposed SLP-IMC-TEM-SUCM using Jaccard coefficient technique achieves better performance than the Cosine coefficient-based similarity technique for predicting the social links.

REFERENCES

[1] N.J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu and X. Xie, "We know how you live: exploring the spectrum of urban lifestyles," in Proc. 1st ACM Conf. Online Soc. Netw., pp. 3-14, 2013.

[2] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in ACM Proc. 13th Int. Conf. World Wide Web, pp. 512-521, 2004.

[3] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: automated construction of application signatures," in Proc. ACM SIGCOMM Workshop Min. Netw. Data, pp. 197-202, 2005.

[4] B. Suryakumar and E. Ramadevi, "A multi context embedding model based on convolutional neural network for trajectory data mining," Int. J. Comput. Sci. Mob. Appl., vol. 5, no. 9, pp. 1-9, 2017.

[5] B. Suryakumar and E. Ramadevi, "An improved multi-context trajectory embedding model using parameter tuning optimization for human trajectory data analysis," Int. J. Appl. Eng. Res., vol. 13, no. 22, pp. 1-9, 2018.

[6] B. Suryakumar and E. Ramadevi, "Human trajectory data and internet traffic mining using improved multi-context trajectory embedding service usage classification model," Int. J. Eng. & Technol., vol. 7, no. 4, pp. 1-9, 2018.

[7] E. Bulut and B. K. Szymanski, "Exploiting friendship relations for efficient routing in mobile social networks," IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 12, pp. 2254-2265, 2012.

[8] K. Neema, Y. E. Sagduyu, and Y. Shi, "Search delay and success in combined social and communication networks," in IEEE Glob. Commun. Conf., pp. 3077-3082, 2013.

[9] L. Dong, Y. Li, H. Yin, H. Le, and M. Rui, "The algorithm of link prediction on social network," Math. Probl. Eng., 2013.

[10] J. Manotumruksa, "Users location prediction in location-based social networks," in Proc. 6th Symp. Future Dir. Inf. Access, pp. 44-47. BCS Learning & Development Ltd., 2015.

[11] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," IEEE Trans. Emerg. Top. Comput., vol. 3, no. 1, pp. 95-106, 2015.

[12] J. Valverde-Rebaza, M. Roche, P. Poncelet, and A. de Andrade Lopes, "Exploiting social and mobility patterns for friendship prediction in location-based social networks," in 23rd IEEE Int. Conf. Pattern Recognit., pp. 2526-2531, 2016.

[13] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen, "LCARS: a location-content-aware recommender system," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 221-229, 2013.

[14] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "A random walk around the city: New venue recommendation in location-based social networks," in IEEE Int. Conf. Priv. Secur. Risk Trust Soc. Comput., pp. 144-153, 2012.