International Journal of Engineering & Technology, 7 (4) (2018) 3538-3542

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.17981
Research paper



Human trajectory data and internet traffic mining using improved multi-context trajectory embedding service usage classification model

Suryakumar B [1] *, Dr. Ramadevi E [2]

[1] *Research Scholar, Department of Computer Science, NGM College, Polach, Tamilnadu, India*
[2] *Associate Professor, Department of Computer Science, NGM College, Polach, Tamilnadu, India*
**Corresponding author E-mail: suryakumarphd2017@gmail.com*

Abstract

Due to the rapid growth of mobile messaging Apps, the classification of Internet traffic into different types of service usages has become a vital process to handle the location-based social networks. In previous researches, Improved Multi-Context Trajectory Embedding Model (IMC-TEM) was proposed to analyze and mine the human trajectory data using multiple context information of trajectory data. However, this model does not consider Internet traffic classification that investigates how to use encrypted Internet traffic for classifying service usages. Therefore in this paper, IMC-TEM is incorporated with CUMMA model to classify the service usage using both Internet traffic data and contextual information of trajectory data generated by messaging Apps. In this model, four major processes are performed to predict the service usages and end-user behaviors efficiently. Initially, traffic segmentation process is performed based on the hierarchical clustering with threshold heuristics that segments the Internet traffic into sessions and dialogs. After that, features are extracted from the segmented traffic based on the packet length and time delay. Then, Random Forest (RF) classifier is applied to classify the service usage types. Moreover, clustering-Hidden Markov Model (HMM) is introduced to detect mixed dialogs from outliers and decompose those into sub-dialogs of single-type usage. Finally, the performance effectiveness of the proposed model is evaluated through the experimental results using different real-world datasets.

Keywords: *Clustering-HMM; Encrypted Internet Traffic; IMC-TEM; Mobile Messaging Apps; RF Classifier; Service Usage Classification*.

1. Introduction

In a digitized world, the 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 (Yuan, N. J., et al. 2013). 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 like paying bills, booking tickets and shopping.
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 a variety of 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 reconstructing protocol signatures in its payload (Sen, S., et al. 2004; Haffner, P., et al. 2005). 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 research, IMC-TEM was proposed to analyze the human trajectory data (Suryakumar, B., & Ramadevi, E. 2017). In this model, a CNN was used to learn the parameters. Moreover, a frog-leaping optimization algorithm was used for tuning the parameters which are needed to improve the accuracy of the contextual model and social link prediction. However, it considers only the characterization of types of contexts for different Apps. It requires Internet traffic classification to jointly analyze service usage behavior as well to enhance the location recommendation and social-link prediction performances.
Hence in this article, service usage classification with IMC-TEM is proposed based on the encrypted Internet traffic analysis. In this model, network traffic characteristics are also considered with the contextual features. To classify the Internet traffic data generated by messaging Apps, CUMMA model is integrated with IMC-TEM which consists of four phases namely traffic segmentation, traffic feature extraction, usage type prediction and outlier detection. Initially, network traffic is segmented as sessions and dialogs based on the hierarchical clustering with thresholding heuristics. Then, the features are extracted from the segmented traffic according to the packet length and time delay. After that, the service usage type is predicted by using an RF classifier. Finally, the mixed dialogs from outliers are detected and decomposed into sub-dialogs of single-type usage by using the clustering-HMM method. Based on this proposed model, the complexity of contex-

Copyright © 2018 Suryakumar B, Dr. Ramadevi E. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

tual information model is reduced and the scalability of Internet traffic classification is improved.
The rest of the article is structured as follows: Section II presents the works which are related to the Internet traffic and service usage classification methods. Section III explains the proposed methodology. Section IV illustrates the performance evaluation of the proposed model. Section V concludes the research work.

2. Literature survey

Identification of the diverse usage behaviors of smartphone apps (Xu., Q., et al. 2011) was proposed based on the network measurements from a national level tier-1 cellular network provider in the United States (US). In addition, the similarities across the apps were identified in terms of geographic coverage, diurnal usage patterns, etc. Moreover, the diurnal patterns of different genres and mobility of apps were discovered to classify the user behavior. However, some traffic may miss since it does not use standard platform URL API. Also, inaccuracy of identification may happen due to the time difference between user-agent fields.
An empirical analysis (Ghose, A., & Han, S. P. 2011) was proposed for user content and their behavior on the mobile internet. The major aim was analyzing whether there was a positive or negative interdependence between the two activities. Then, a unique panel dataset was used that consists of individual-level mobile internet usage data that encompass individual multimedia content generation and usage behavior. After that, this knowledge was combined with data on user calling patterns for constructing their social network and computing their geographical mobility. Moreover, an individual-level simultaneous equation panel data model was constructed to control the different sources of endogeneity of the social network. However, the performance was not effective due to lack of data like it does not have information about the certain type of content uploaded or downloaded and the destination websites.
Network traffic classification (Zhang, J., et al. 2013) was proposed by using correlation information. In this technique, a novel non-parametric approach was proposed for traffic classification by incorporating the correlation of traffic flows. Based on this approach, the classification performance was improved and evaluated by using some training samples. Also, a detailed analysis was presented for both theoretical and empirical characteristics. However, this approach requires some prior knowledge to classify the traffic flows and also the overall accuracy was less.
An effective network traffic classification (Zhang, J., et al. 2013) was proposed with unknown flow detection. The main objective of this method was avoiding the issue of unknown flows in a semi-supervised network. In this method, flow correlation was incorporated into the semi-supervised model which has the detection ability of unknown flows. Also, flow label propagation was proposed for automatically labeling relevant flows from a large unlabelled dataset to observe the issue of a small supervised training set. Moreover, the compound classification was proposed for mutually identifying the correlated flows that improve the classification accuracy. However, this approach was not suited for traffic classification across the network.
Characterization of user behavior in the mobile internet (Yang, J., et al. 2015) was studied. In this study, the mobile user behaviors were classified from three characteristics such as data usage, mobility patterns and application usage. Also, the traffic heavy users and the mobility pattern were observed as nearly associated with the application access behavior of the users. Users may be clustered via their application usage behavior and application types may be recognized through an interaction of the users. However, fairness was less and network congestion was not controlled.
A participatory cultural mapping approach (Yang, D., et al. 2016) based on the collective behavior data in location-based social networks. Initially, the participatory sensed user behavioral data were collected from the location-based social networks. Then, a progressive home location identification method was proposed for filtering ineligible users since only local users were eligible for cultural mapping. After that, three primary cultural features were extracted from daily activity, mobility and linguistic perspectives correspondingly to introduce the cultural clustering method for discovering cultural clusters. At last, the cultural clusters were visualized on the world map. However, the user behavioral data was less since it was gathered from a particular social network.

3. Proposed methodology

In this section, the proposed IMC-TEM with CUMMA model is explained in brief. The contextual information of trajectory data is characterized by IMC-TEM by maximizing the average log probability for each location. Here, the parameters needed to set IMC-TEM such as vector size and context window length are optimally tuned based on the frog-leaping optimization algorithm. Moreover, the encrypted Internet traffic is classified to recognize the service usage types by using CUMMA model. Initially, Internet traffic from traffic flows is segmented into sessions with a number of dialogs using hierarchical clustering where the traffic flow denotes the encrypted network traffic and session and dialog represents the segments of traffic flow in various granularities. A session is initiated when the user opens the App and terminated while the user closes it. The generated Internet traffic during this session is called as the dialog. Most of the dialogs are single type usage such as text and location sharing whereas the other dialogs are mixed usages. A service usage predictor is used for classifying these segmented dialogs into single-type usages or outliers. Then, an RF classifier is used for classifying the service usage whereas clustering-HMM is used for disaggregating mixed usage types.

3.1. Traffic segmentation

Initially, network traffic data of different usages in mobile messaging Apps is collected. Once these benchmark data is collected, two-stage segmentation i.e., from traffic-flow to session and from session to dialog is performed to segment these traffic-flows from coarse-grained level i.e., session to fine-grained level i.e., dialog. First, the collected traffic-flow is segmented into multiple sessions using the thresholding method. Here, each session may contain multiple consecutive dialogs. To segment dialogs, the concept of hierarchical clustering is adopted with a bottom-up based segmentation algorithm. In this algorithm, the time duration of the session is split into multiple small intervals. Then, the zero-traffic intervals are filtered whereas the non-zero intervals are represented as a sequence of initial dialogs.
Additionally, three merging operations such as merging heavy-heavy dialog pairs, light-heavy dialog pairs and peak dialog pairs are performed to maintain adjacent intervals in one dialog if they correlate to one usage type or mixed usage types. Due to these operations, a condition is avoided where the network traffic of one usage type is split into two dialogs of the same usage types.

3.2. Traffic feature extraction

Once the traffic-flow is segmented, the dialog encodes two types of information such as the sequence of packet lengths and the sequence of time delays. Then, the discriminative features are extracted from two major perceptions such are packet length and time delay. The packet length is computed according to the size of the packet and measured in terms of bytes to be transmitted. Time delay refers to the time taken for the packet to arrive at the destination.

- Packet Length Related Features: For a given sequence of packet lengths, standard deviation, median, minimum, maximum, skewness, kurtosis and standard error of packet length are calculated as features. Here, the variance in packet length is given by,

$$\mathrm{var} = \frac{\sum_{x \in X}(x-\bar{x})^2}{n-1} \quad (1)$$

In equation (1), X refers to the packet length of a given sequence and $n = |X|$. Also, the minimum and maximum values of IP packet lengths are identified. Then, the range from minimum to maximum is split into K equal-sized sub-ranges. For each sub-range, the number of packets is computed to obtain a K-size feature vector, each of which denotes the percentage of packets with length in the k^{th} sub-range.

In addition, the number of packets whose lengths are greater than the lengths of their consecutive packets is calculated with a significant margin as a feature to characterize the variation of the sequence. Also, the longest monotone including both increasing and decreasing subsequences of a dialog is examined and the lengths of these subsequences are used to define the tendency and skewness of the network traffic.

Likewise, the range of packet lengths is identified and split into K equal-sized sub-ranges. For each element in the sequence of packet lengths, this element is replaced with k when its value is in the k^{th} sub-range. This sequence is mapped into a new string sequence based on this manner. By using this string sequence, all the continuous subsequences are identified with size ranging from 3 to 20 and their corresponding number of occurrences in this string sequence is also determined. Finally, the top-N number of occurrences is used as features.

- Time Delay Related Features: The time interval for every two consecutive packets is extracted and thus a sequence of time delays is obtained.

3.3. Usage type prediction

Once all the features are extracted, the service usage types are predicted by using an RF classifier whereas temporal dependencies between consecutive dialogs are neglected to simplify the classification process. As well, these segmented dialogs are treated as a set of independent training instances. Consider **B** as the number of trees for a given set of labeled dialogs with usage types. A random sample is selected continuously with the replacement of the labeled dialogs and **B** number of decision trees is constructed to the **B** samples. Once the training process is completed, predictions for unknown dialog $\mathbf{d}'$ can be obtained by averaging the predictions from all the individual decision trees on the dialog $\mathbf{d}'$. Particularly for each class **k**, RF gives the probability estimation as follows:

$$\mathbf{P}(\mathbf{k}|\mathbf{d}') = \frac{\mathbf{1}}{\mathbf{B}} \sum_{\mathbf{b=1}}^{\mathbf{B}} \mathbf{P_b}(\mathbf{k}|\mathbf{d}') \quad (2)$$

In equation (2), $\mathbf{P_b}(\mathbf{k}|\mathbf{d}')$ refers to the probability estimation of usage type **k** given by $\mathbf{b^{th}}$ tree. It is estimated by calculating the fraction that usage type **k** gets votes from the leaves of $\mathbf{b^{th}}$ tree. The overall decision function of RF is defined as follows:

$$\mathbf{u}' = \mathbf{arg\,max_k}\, \mathbf{P}(\mathbf{k}|\mathbf{d}') \quad (3)$$

Finally, each dialog will be labeled as a single usage type or the unknown mixture of usages.

3.4. Outlier detection and handling

The predicted unknown dialogs are handled by classifying those as outliers or the unknown mixture of usages. For a given dialog, the most probable hidden usage mixture is identified. In this model, a clustering-HMM method is applied for disaggregating a usage mixture consisting of multiple single-type sub-dialogs. Initially, a K-means clustering method is used and then K number of centers is built with mean packet lengths as prior knowledge. After that, each mixed usage dialog is segmented into multiple traffic segmentations where each denotes a single usage sub-dialog. Moreover, each mixed dialog is segmented into multiple sub-dialogs by removing the anomalous sub-dialogs that have very few packets or greatly short in time.

After that, a feature extraction is performed for each sub-dialog to represent those as a feature vector. Here, temporal dependencies between consecutive sub-dialogs are considered to enhance the classification accuracy. As a result, HMM is used for classifying sub-dialogs in a mixed dialog. Especially, for each service usage type **k**, corresponding sequences of packet lengths are given to the HMM model. Thus, **k** number of trained HMM models is obtained. The given sub-dialog with unknown usage type is given into the obtained HMM models to compute the likelihood of this sub-dialog for each usage type **k** and obtain **k** likelihoods. Then, this sub-dialog is classified as the service usage type of the largest likelihood and the previous classified dialogs are used to train the HMM model. Finally, this trained HMM model is utilized for predicting the usage types of sub-dialogs. Based on this model, a sequence of semantically-rich usage activities $\langle \mathbf{u}, \mathbf{l_n}, \mathbf{c_n}, \mathbf{s_n}, \mathbf{b_n}, \mathbf{e_n}, \mathbf{ut_n} \rangle_{\mathbf{n=1}}^{\mathbf{N}}$ is revised, each of which contains user, location, category label, total time, start time, end time and usage type.

4. Result and discussion

In this section, the performance effectiveness of the proposed ICM-TEM-SUCM is evaluated and compared in terms of recall by considering context factors only, service usage types only and both factors. The experiments are conducted in MATLAB 2018a based on two applications called location recommendation and social link prediction for both proposed and existing models. Here, three open geo-social networking datasets namely Foursquare_S, Foursquare_L and Gowalla are used. Among these datasets, Foursquare_L and Gowalla are utilized for location recommendation whereas Foursquare_S and Gowalla are utilized for link prediction. In this analysis, the locations are recommended based on the service usages i.e., number of users, check-ins, number of items, etc., in a particular location. The examples of services are coffee shop, Apple store, Sandwich shop wherein users get different services. Such services may be recommended at different locations like Airport, Train station, Stadium, etc. The following 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_S	4163	483814	32512	121142
Foursquare_L	266909	33278683	-	3680126
Gowalla	216734	12846151	736778	1421262

The examples of recommendation results based on the service usages, locations and time-aware are shown in the following Figure 1-3.

```
enter Serivice Category Food

Serching_Item =

    'Food'


recomended_location =

    'Coffee Shop'
    'Sandwich Shop'
    'Burgers'
    'Pizza'
```

Fig. 1: Traffic-Based Service Usage-Based Location Recommendation.

```
enter Location Mall

Serching_Item =

    'Mall'


recomended_location =

    'Mall'
    'Toys&games'
    'Pet Store'
    'Starbucks'
    'vietnamese'
```

Fig. 2: Traffic-Based General Location Recommendation.

```
enter Time 7

recomended_location =

    'American'
    'Other-Architecture'
    'Stadium'
    'Theater'
    'Camping&Outdoors'
```

Fig. 3: Traffic-Based Time-Aware Location Recommendation

4.1. Analysis on location recommendation

The effectiveness of location recommendation along both general location recommendation and time-aware location recommendation is evaluated by considering home-city recommendation settings. Here, 20% trajectories with only home-city locations are selected as a testing dataset and the remaining trajectories are chosen as training data. Consider n is the rank of the target location within this list to form a top-k recommendation list by collecting the top k ranked locations from the list.
Also, hit_k is defined for a single test case as either the value is 1 if the target location is identified in the top k results, otherwise, the value is 0. The overall Recall_k refers to the ratio of hits in all the test check-in records.

$$\text{Recall}_k = \frac{\text{Number of hit}_k}{\text{Number of all cases}} \quad (4)$$

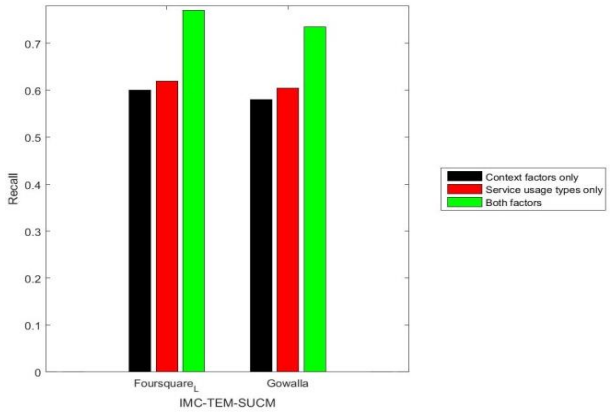


Fig. 4: Comparison on Traffic-Based General Location Recommendation.

Figure 4 and Figure 5 shows that the impact of contextual factors with service usage for general and time-aware location recommendation on Foursquare_L and Gowalla dataset. From the analysis, it is observed that all the considered contexts including usage type are useful for both recommendation tasks. Therefore, it is concluded that the proposed IMC-TEM-SUCM achieves higher recall value while considering all the contextual and network factors whereas, only contextual information are utilized in IMC-TEM.

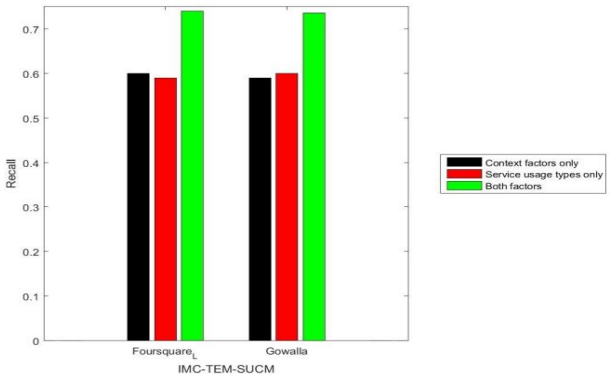


Fig. 5: Comparison on Traffic-Based Time-Aware Location Recommendation.

4.2. Analysis on social link prediction

This analysis is used for predicting whether a social link between a pair of users exists or not only based on their trajectory data. Consider P_T is the set of all user pairs with real friend links and P_R is the number of all user pairs identified by a candidate as friends. Then, the effectiveness of social link prediction is measured based on the value of recall which is computed as follows:

$$\text{Recall} = \frac{|P_T \cap P_R|}{|P_T|} \quad (5)$$

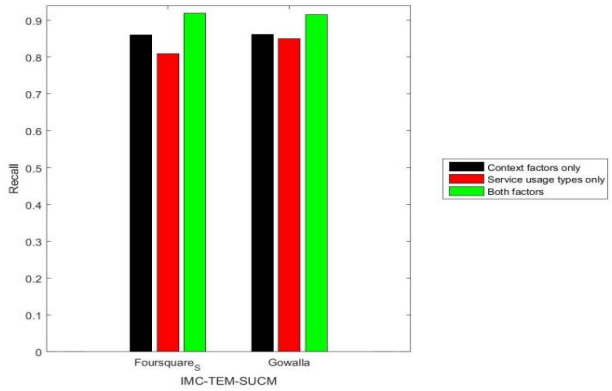


Fig. 6: Comparison on Social Link Prediction.

Figure 6 shows that the impacts of contextual features for social link prediction on Foursquare_S and Gowalla dataset. From the analysis, it is observed that all the considered contexts and usage types are useful for predicting the social links. Thus, it is concluded that the proposed IMC-TEM-SUCM achieves higher recall value while considering all the contextual and network factors whereas, only contextual information are utilized in IMC-TEM.

5. Conclusion

In this paper, an Improved Multi-Context Trajectory Embedding Model with Service Usage Classification Model (IMC-TEM-SUCM) is proposed to enhance the social link prediction based on the encrypted Internet traffic in mobile messaging Apps. In this model, human trajectory data mining is considered by jointly modeling contextual information of trajectory data, network characteristics and temporal dependencies. Initially, the traffic-flows of service usages are collected 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 based on the packet length and time delay for the training process. Additionally, the service usage classifier i.e., an RF classifier is learned to classify the segmented dialogs. Furthermore, the anomalous dialogs with mixed usages are detected and segmented into multiple sub-dialogs of single-type usage using the clustering-HMM method. Finally, the experimental results

demonstrate the performances of the proposed model for service usage analytics with human trajectory data mining. In the future, this model could be enhanced by the routing algorithm to reduce the high traffic-flows and increase the end-to-end success probability.

References

[1] Yuan NJ, Zhang F, Lian D, Zheng K, Yu S, & Xie X (2013), "We know how you live: exploring the spectrum of urban lifestyles", *Proceedings of the first ACM conference on Online social networks*, pp. 3-14. https://doi.org/10.1145/2512938.2512945.

[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. https://doi.org/10.1177/003335491012500414.

[3] Sen S, Spatscheck O, & Wang D (2004), "Accurate, scalable in-network identification of p2p traffic using application signatures", *ACM Proceedings of the 13th international conference on World Wide Web*, pp. 512-521.

[4] Haffner P, Sen S, Spatscheck O, & Wang D (2005), "ACAS: automated construction of application signatures", *ACM Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, pp. 197-202. https://doi.org/10.1145/1080173.1080183.

[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] Xu Q, Erman J, Gerber A, Mao Z, Pang J, & Venkataraman S (2011), "Identifying diverse usage behaviors of smartphone apps", *ACM Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pp. 329-344. https://doi.org/10.1145/2068816.2068847.

[7] Ghose A, & Han SP (2011), "An empirical analysis of user content generation and usage behavior on the mobile Internet", *Management Science*, *57*(9), 1671-1691. https://doi.org/10.1287/mnsc.1110.1350.

[8] 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. https://doi.org/10.1109/TPDS.2012.98.

[9] Zhang J, Chen C, Xiang Y, Zhou W, & Vasilakos AV (2013), "An effective network traffic classification method with unknown flow detection", *IEEE Transactions on Network and Service Management*, *10*(2), 133-147. https://doi.org/10.1109/TNSM.2013.022713.120250.

[10] 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. https://doi.org/10.1109/TETC.2014.2381512.

[11] 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*, *7*(3), 30-53. https://doi.org/10.1145/2814575.