Towards Human Mobility Detection Scheme for Location-Based Social Network

Munairah Al-Jeri
System and software development department
Kuwait Institute for Scientific Research
P.O. Box 24885, Safat 13109, Kuwait.
mujeri@kisr.edu.kw

***Abstract*—Location-based Social networks (LBSN) are a promising source for human mobility mining that attracted a lot of attention in the research community. The aim of this research is to discover the patterns and mobility of LBSN such as Twitter by analyzing the spatiotemporal features of user's tweets. Twitter Streaming Application Program Interface (API) was used to capture Geo-tagged tweets. The analysis shows that geotagged tweets can obtain valuable information on human mobility, such as the relation between movement flow and the time of the day. It provides evidence that Twitter data can be useful for tracking human movement and understand human behavior pattern.**

Keywords—Location-based Social networks, Human Mobility, Twitter, Spatiotemporal Dynamics, Geo-tagged.

I. INTRODUCTION

Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing, and collaboration [1]. Social media lets users post the details of their daily activity, location and share information with the world. It has been observed to grow fast in the past few years. With its rapid growth of applications, it makes the collection of large geo-social data possible from a variety of sources such as Twitter, Instagram, and Foursquare. Retrieved data from the social media produce a point-based geo-social network. For example, more than 300 million tweets are generated every day across the world through Twitter's application. Approximately 15 millions of these tweets have geographic coordinates, and about 20 million tweets include a geographic reference (e.g., place name) in the message content. Such data naturally form massive networks that contain users and links. Links can be directly built by each user's connections (e.g., friends, followers, pins) or they can be indirectly derived from interactions between users such as shared text (e.g., retweets, hashtags), videos, images, and web pages [2]. These tweets are valuable for understanding the mobility, interest of users, and to build new applications that are temporal and location-aware.

Analyzing human mobility is one of the most important fundamentals of many applications such as urban planning [3], traffic and population predicting [4], location inference, and recommender systems [5,6]. Recent researchers have used data pulled from Location-Based Social Network (LBSN) such as georeferenced tweets to analyze mobility pattern instead on just relying on tracking technologies such as mobile phones datasets [7], Wi-Fi, and RFID device [8]. Such techniques have delivered deep understandings of human mobility dynamics, but their ongoing use for monitoring human mobility involves privacy concerns, data access restrictions and high expenditure.

Data shared on the LBSN have made a wide-ranging collection of information on human behaviors in space and time available for researchers [9]. Mainly, the spatiotemporal pattern is essential for human behavior analysis, which attracts insights on human mobility [9–11]. Geo-tagged tweets have proven in geo-social research, which aids in new intelligent geo-social systems. These researches are believed to have significant discovery in the upcoming years with the rise of interests in spatial computing [12].

This paper presents a mechanism for human mobility characterization that illustrates the benefits of the spatiotemporal data retrieved from social media applications. The work is based on data obtained from Twitter. Our primary motivation is to study the geo-located Twitter data for human mobility and activity patterns in the state of Kuwait. The study focuses on understanding the movement of the population of Kuwait, where they flow, at what time and its intensity to efficiently manage essential governorate applications such as traffic, population prediction, and transportation resources. Limited work has been done to explore human mobility at a country scale while taking into consideration the moment of the day as well as the noisy nature of the data. The contribution of this paper is to identify the categories of the most intense places of user's movement through the day by introducing activity category as a new dimension to the mobility pattern analysis. The work approaches mobility mining based on Twitter data that fully consider the following challenges:

- Extract general mobility information related to a particular country while distinguishing the time and day in which the data is retrieved.
- Study the relationship between the moment of the day and its associated spatial place category.

- Consider the fuzzy and noisy nature of data generated by users. Thus, making the obtained result as precise as they could be.
- Take into account the activity level of users within each detected area. Thus, allowing the establishment of a relationship between the different level of users and their movement across the study area.

To fully understand the obtained results and analysis, it's essential to understand the geographic distribution of the urban areas in the state of Kuwait. Kuwait is considered as a desert country, where most of its lands are covered by the Arabian sand and the urbanized areas are mostly concentrated along the coastline. The center city of Kuwait is found in the central district in the capital governorate, where the majority of the commercial and business areas are located.

The remainder of this paper is organized as follows. Firstly, an overview of the background is in section II. Section III describes the collection and preparation of the data. Next, section IV is devoted for the clustering and classification of the data. Then, the method of the mobility pattern detection is defined in section V. And section VI shows the results and analysis obtained. Finally, the main conclusion and future direction is summed up in section VII.

II. RELATED WORK

Recently, many studies have been done in analyzing LBSN data that provides valuable information on understanding human movement and activity [13,14]. The following are some studies that considered social media data in their analysis of human movement pattern.

Noulas et al. [15] used Foursquare's geo-located information in their approach to model human activity pattern in two metropolitan cities by applying a spectral clustering algorithm. Then with the use of the predefined Foursquare categories that indicates the type of the location and with the results obtained from the clustering algorithm, they identified user communities that share similar categories of places as well as did a comparison of urban neighborhood.

Focusing on social media data, Hassan et al. [16] presented fundamental findings related to the activity categories and individual mobility pattern in a city using Twitter data. They discovered the relationship between popularity of a place and the possibility of selecting this place as a destination. They also found out that the behavior of the spatio-temporal activities in the three major U.S cities have distinct patterns.

Finally, [17] composes the spatial-temporal trajectories of a set of Twitter users by identifying relevant patterns, flows and anomalies in urban area. Such trajectories are mined by building simple Origin-Destination matrices. The proposed trajectory mining approach was applied and validated on a large set of twitter data gathered in Barcelona during the mobile world congress (MWC2012).

Even though there are some similarities between the above-mentioned approaches and the presented work, some main difference can be stated. Such solutions do not take the advantage of the benefits of Fuzzy c-mean clustering algorithm (FCM) [18]. Moreover, the predefined categories of Foursquare are not validated to understand the accuracy results. Furthermore, none of the above-mentioned considers the data based on the moment of the day when the information was generated in their analysis. On the contrary, the presented work analyzes the mobility and activity pattern throughout the day, which allows knowing and understanding of the movements in every moment, category and location. An overview of the proposed work flow is shown in Fig.1.

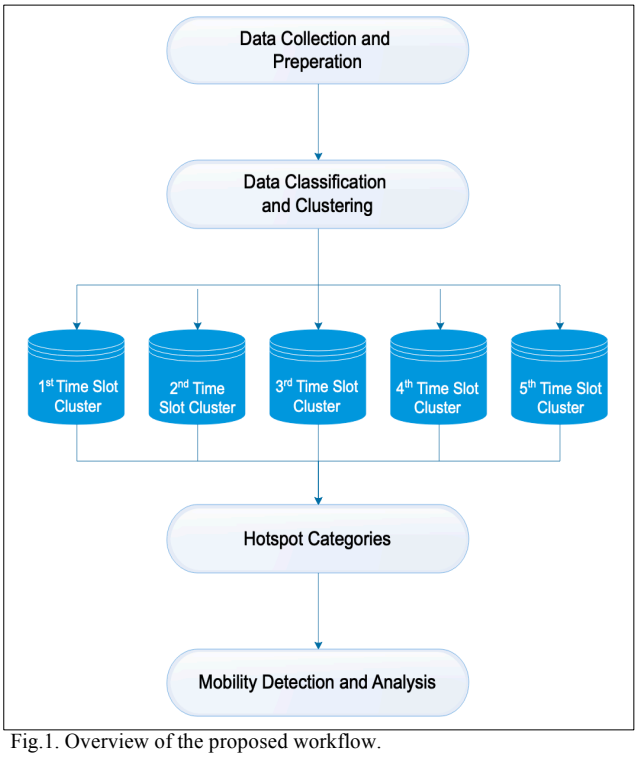

Fig.1. Overview of the proposed workflow.

III. DATA COLLECTION AND PREPERATION

The section describes the method used to extract the data, the characterization of the collected data, and lastly, the process of preparing the final data for analysis by data cleaning and data aggregation.

3.1. Data acquisition

Acquiring data is the first important aspect of any analysis. In this research, the Twitter public Streaming API is used to extract Twitter data using a python script. The script can be easily extended with any other social-network API. Data retrieved is restricted by the geo-location coordinates of a rectangular bounding box for the state of Kuwait defined by two longitudes and two latitudes. To make it easier to handle the massive volume of data, the JavaScript Object

Notation (JSON) format of the retrieved data is transformed into data frame format.

3.2. Data Characterization

Data retrieved provided real geographic coordinates of the user's spatial distribution in Kuwait from December 2017 to February 2018. During the 62 days of collection, more than 21,000 thousands of unique users were observed, and 323,859 thousands of data records were obtained. Data record represents spatial points on the area of interest that defines the user location for a specific date and time. For the spatiotemporal analysis, we are only considering three parameters that represent a record: $R_n = (u, p, t)$, where u represents a user, t is the timestamp of the record, and p is the position defined by latitude and longitude coordinates.

3.3. Data Cleaning

The Fetched data may be incomplete or contains noise. The need for data cleaning will arise from errors happening in the analysis. Data cleaning is the process of preventing and correcting these errors. It included removing tweets without geolocation information, tweets that have geolocation coordinates out of Kuwait's boundary, removal of users with less than 25 posts, and the elimination of spam accounts that do not represent real users. Spam users were identified based on the geographic location of their tweets. Speed and distance between two consecutive tweets posted by the same user in the same day were considered as the main key in detecting spammers. Since there is no realistic transportation to help a user to move more than 1.65 kilometers in less than two minutes.

3.4. Data Aggregation

Next, tweets that are posted by the same user in the same day and in similar time and location are being aggregated. The objective of this step is to avoid the disturbance when analyzing the movements of users since it does not represent a real movement in a space-time dimension. When tweets are close in time and space, they will be aggregated in one tweet representing the actual situation of the user. The criteria used to determine when two data records Ri and Rj are overlapped is:

- $u_i \neq u_j$;
- the geographic distance between pi and pj is less than 250m;
- ti and tj are less than 1-hour time difference;
- Both data records Ri and Rj are posted on the same day.

The distance threshold is set based on most cellular provider's adaptation of the Assisted-GPS technologies that provide location information such as: GPS (with an average accuracy of 10 m), Wi-Fi (70–80 m), and cellular position (100–300 m) to avoid mistakes caused by location inaccuracy within a short distance [19].

IV. DATA CLASSIFICATION AND CLUSTERING

The data should be undergone thorough primary processes before the movement pattern detection discussed in section V. These processes consists of separating the data based on its timestamp, then applying a clustering algorithm to group similar data, next classifying users into different levels based on their activity, and ending the process by identifying social activity areas.

4.1. Time Slot Separation

The mobility analysis depends on the assumption that cities activity is not the same during the full day, but it changes over time. The 24-hour period of a day is divided into five different time slots: early morning, morning, evening, late evening and night. After the data has been cleaned and aggregated, it is split into five different databases based on the divided time slots. Details of each dataset time range are shown in Table I.

TABLE I. TIME RANGE OF THE DEVIDED DATASET

Time Slot	Time Range
Early morning	00:00 – 08:00
Morning	08:00 – 12:00
Evening	12:00 – 16:00
Late evening	16:00 – 21:00
Night	21:00 – 00:00

4.2. Clustering algorithms

The next step in the analysis is to apply the Fuzzy c-means clustering algorithm to each of the five datasets. Clustering is defined as grouping a set of objects in one group, where objects in the same group are more similar to each other and dissimilar to other groups. The primary objective of this step is to cluster the data and to help identifying the areas of social activity. Spatial coordinate data of each tweet were only considered as inputs for this algorithm. *Dunn* index and *DB* index, validity indices were performed to find the suitable value for the number of clusters to be generated for each dataset. The *compactness* and the *separation* between clusters were also considered as part of the validity.

Clustering will be functional if the clusters are maximum separated from each other and the objects within clusters should be more and more close (compactness) to the centroid. *Dunn* index is defined as the ratio of the *separation* to *compactness* which indicates that if the value of *Dunn* index is large, clusters are well separated. *DB* index is defined as the ratio of *compactness* to *separation*, and a small value indicates that they are more compact [20].

The unsupervised clustering algorithm is based on the Euclidean distance between tweets data points. The clusters

are formed in each dataset according to the distance between points and the cluster centers for each cluster. Every data point is related to every cluster in its dataset with a degree of belonging to each cluster. Data points that lie far away from the center of a cluster will have a low degree of belonging to that cluster, while nearby data points will have a high degree of belonging to the cluster. The result is a membership matrix between all tweets and generated clusters.

4.3. User Classification

Another essential step for mobility analysis is to identify user's activities and to classify them into levels. The activity level of each cluster is measured to enrich the information about users and discover the kind of users in each cluster in each time slot.

To achieve this task, users are classified according to their activity level in the target social network area. Based on the average post per day, users are classified into three activity level: inactive, active and highly active users. The level of activity of each cluster is determined by the activity level of each user's tweets. The result is the percentage of users for each of the level of activity associated with each cluster.

4.4. Hotspots Identification and Categories

Hotspot analysis is used to identify locations of statistically significant point-of-interest (POI) in a large volume of data by converging points that are in proximity to one another on a calculated distance. The number of tweets per square meter of each cluster in each time slot was created and plotted on the map, known as the density heat map. The produced heat map is used to identify major hotspots of tweets by using a color gradient to show the location of dense point data across all of Kuwait and to get a more precise visualization of social activity locations and categories of user's POI.

With the use of Foursquare API [21], we retrieved the most-likely Foursquare category that can be associated with the geographic position of each identified POI. Given a GPS location, the Foursquare API offers an ordered list of places that could be related to it. To validate the category of the POI, we manually checked each location in google maps with the use of Fusion Table [22]. The identified POI falls into five categories: work and study area, shopping malls, eat outs and resorts, residential, and others.

V. MOBILITY DETECTION SCHEME

The final task aims to extract the flow movement of users through the results achieved from previous actions in section IV. The task starts with identifying the most representative cluster for each user in each of the five-time slots. This was accomplished by obtaining the cluster with the highest membership degree generated by the fuzzy algorithm to be the representative for the user in that timeslot. At the end, each user is represented by five variables that indicate the general movement of the user during the day across the time slots. Table II shows an example of these variables, where it shows that user1 is closer to cluster A_1 in the early morning slot and moves to cluster A_2 in the next slot, morning slot. If there are no records for a user in a specific time slot, then a *Null* value will be representing that time slot, similarly with user4.

With the use of the resulted individual mobility flow, the flow movement from one time slot to another is obtained (e.g., from evening slot to late evening slot or from night slot to early morning slot). Firstly, the sum of occurrences is calculated for each cluster. Then, the count of movement between clusters in different slots is computed. Lastly, the percentage of users moving from one slot to another is obtained.

Specific user mobility can be extracted from the mobility flow table which will aid in the future prediction of real time spatiotemporal data. The result information, combined with the activity level for each cluster and the categories of POI, offers a global and precise visualization of user's behavior studies in the total area of Kuwait.

TABLE II. EXAMPLES OF INDIVIDUAL MOBILITY FLOW

Users	Time Slots				
	Early morning	***Morning***	***Evening***	***Late evening***	***Night***
user1	A_1	A_2	C_3	D_4	C_5
user2	D_1	B_2	B_3	A_4	L_5
user3	A_1	E_2	L_3	A_4	C_5
user4	-	A_2	L_3	A_4	B_5

VI. RESULTS AND ANALYSIS

In this final section, the results and analysis of the flow movement of Twitter user's in Kuwait are discussed. All of the analyses for each of the time slots have been carried out. The flow movement from the late evening slot to the night slot is taking as the reference scenario in the below discussed analysis, to give a better understanding and visualization of the mobility movement.

6.1. Results

To analyze the mobility of Twitter users, the Twitter Streaming API was used to target the country of Kuwait. Some details of the generated dataset are seen in Table III.

All tweets were classified according to time and location. Each tweet was classified either as "weekend" tweet or "weekday" tweet based on the day it was posted. The processed dataset is composed of 91,601 weekend tweets and 226,884 weekday tweets. Moreover, each tweet is assigned to it's district and governorate that it was posted from based on its location indicated by the longitude and latitude fields.

The result of this classification shows that both weekends and weekdays tweets are mainly generated in the most central areas of the capital governorate (Sharq and Jibla) between the hours of 5pm and 6pm.

TABLE III. DATASET DETAILS

Features	Kuwait Database
Time Period	24/12/2017 - 14/02/2018 (62 days)
Covered Area	17,818 km²
Raw # tweets / users	323,859 / 21,616
Process # tweets / users	318,485 / 16742

According to the timestamp of the tweets, the data was divided into five datasets. The time range of each slot was previously defined in Table I. After the time slot separation of the data, the fuzzy clustering algorithm was applied to each dataset. Details of each slot clusters are shown in Table IV. Expectedly, the late evening slot is the most active slot with the most generated number of clusters, where users tend to go out and tweet more about their activities.

TABLE IV. TIME SLOTS DETAILS

Time Slot	# of Cluster	# of Tweets
Early morning	15	41,728
Morning	11	45,477
Evening	14	78,698
Late evening	18	136,656
Night	15	20,800

Users were classified based on their average tweet posts. 30% of users were classified as inactive users and only 6% were classified as highly active users. Noticing that late evening slot contains the highest percentage of inactive and active users than the rest of the other time slots. While highly active users are more active in the evening slot.

6.2. Analysis

The density heat map in Fig.2 and Fig.3 is the result of the tweets concentration in the late evening and night time slots. In both figures, red color areas indicate higher density of tweets, while areas of lower densities are represented by a yellow color. The green points symbolize the midpoint of each cluster in each time slots. Noting that the displayed clusters represents 80% of the generated tweets in both time slots. The rest of the clusters were discarded as no significant POI could be identified in those areas, thus no additional contribution is added to the analysis of user's movement. Expectedly, tweets concentration is higher in the center of the city in the late evening slot (cluster A in Fig.2),

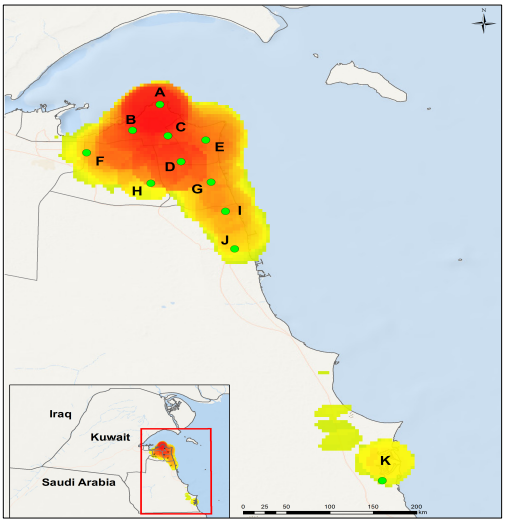

Fig. 2. Heat map of tweets and generated clusters in late evening slot.

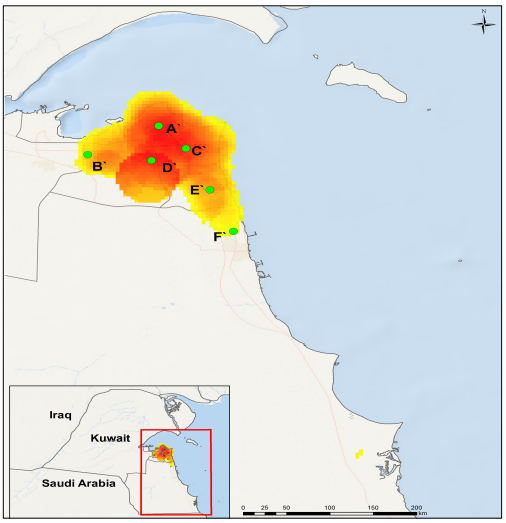

Fig. 3. Heat map of tweets and generated clusters in night slot.

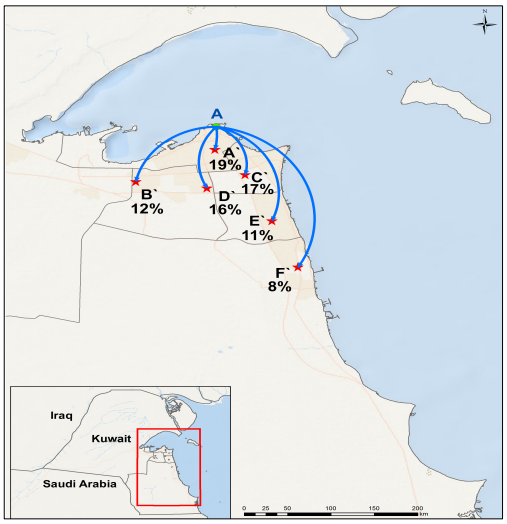

Fig. 4. Movement flow of users from late evening slot to night slot

while it is spreading to the rest of the residential areas in the night slot. Noticing that the busiest areas in Kuwait always represent a high number of POI. The areas showing this behavior are mainly located at the city center. As seen in Fig.2, the densest areas in the most active clusters in the late evening slot are more situated in the center of the city with more POI identified (cluster A, cluster C) than in the rest of the clusters in that time slot.

In the most active time slot, late evening slot, tweets are mostly generated in shopping malls, and in eat outs and resorts areas while it is less in the residential areas. In the next time slot, a reverse movement gradually takes place and activity decreases in shopping areas and increases in residential areas. The flow percentage of users from cluster A in the late evening slot to the night slot is seen in Fig.4, where the red stars represent the midpoint of the night slot clusters and the green point represent the most active cluster from the previous time slot. Observing that a high number of POI are categorized as eat outs and resorts in cluster A` and C`, while most POI in clusters B', E` and F' are identified as residential houses. Noticing, that the most tweeted POI in cluster D` are identified as Others such as: airport and garages, with more tweets generated by new identified users.

It was possible to extract user's movement from one-time slot to another based on the categories of the identified POI. For instance, Fig.5 shows user's interest in shopping malls decreases by 35% and increases in residential areas by 20% indicating the end of the day when most are back home, while the interest in eats outs and resorts remains the same in both time slots.

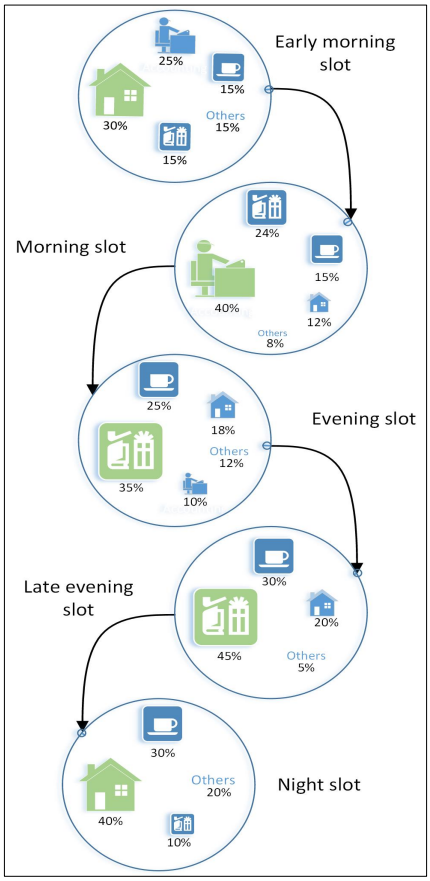

Fig. 5. Categories of users movement throughout the full day.

Finally, regarding the activity level of the users in each cluster, the result obtained shows that the most tweeted areas have a high number of users with low activity level. While less crowded areas have more users with higher activity level, as shown in cluster F` and cluster A' in Fig.3. Even though cluster F` is considered as the least tweeted cluster and the least cluster with identified POI, the majority of users activity level in the identified POI are classified as high active users. While noticing in cluster A', that the percentage of highly active users is lower than inactive users in most of the identified POI even though cluster A' is represented as the most tweeted cluster in that time slot. Fig.6 shows the result of user activity level in each cluster in the late evening slot.

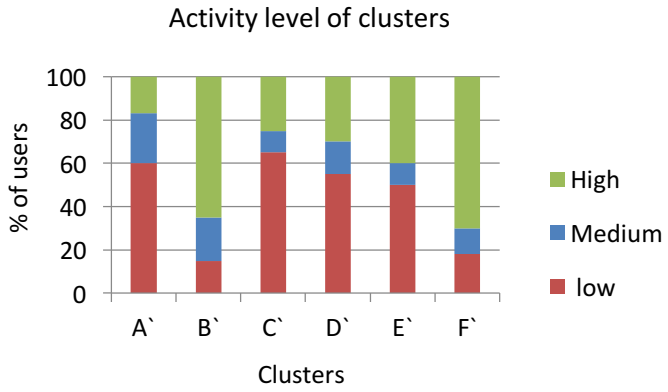

Fig. 6. Activity level of users in each cluster in the late evening slot

VII. CONCLUSION

In this paper, we have presented a workflow for the acquisition, clustering and classification, and detection of mobility pattern of Twitter data. Tweets with geotagged information were processed and analyzed using FCM algorithm in different time slots to generate spatial clusters. Social hotspots were identified in each time slot. Finally, the mobility of users in the five time slots has been carried out.

Social network offers a vast amount of rich spatiotemporal location data. Understanding human movement can be obtained with the use of such data that will aid in the development of social sustainability. As this study shows, analyzing Twitter data can be useful source of knowledge to extract mobility pattern. The collected data can be used for several mining purposes and it allows to classify all the parameters involved, i.e., user's activity, POI and geo-located data points. Future studies will involve developing efficient approaches for finding multi-dimension sequential pattern of each user's trajectories, the prediction movement using the enhanced mobility detection scheme with sentiment enrichment of the collected data.

ACKNOWLEDGMENT

This work was supported by Kuwait Foundation for the Advancement of Science (KFAS).

REFERENCE:

[1] Wise, E.K. and Shorter, J.D., 2014. Social networking and the exchange of information. Issues in Information Systems, 15(2).

[2] Dredze, M., García-Herranz, M., Rutherford, A. and Mann, G., 2016. Twitter as a source of global mobility patterns for social good. arXiv preprint arXiv:1606.06343.

[3] Liu, Y., Wang, F., Xiao, Y. and Gao, S., 2012. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, 106(1), pp.73-87.

[4] Peng, C., Jin, X., Wong, K.C., Shi, M. and Liò, P., 2012. Collective human mobility pattern from taxi trips in urban area. PloS one, 7(4), p.e34487.

[5] Zheng, V.W., Zheng, Y., Xie, X. and Yang, Q., 2012. Towards mobile intelligence: Learning from GPS history data for collaborative recommendation. Artificial Intelligence, 184, pp.17-37.

[6] Cheng, Z., Caverlee, J., Lee, K. and Sui, D.Z., 2011. Exploring millions of footprints in location sharing services. ICWSM, 2011, pp.81-88.

[7] Palchykov, V., Mitrović, M., Jo, H.H., Saramäki, J. and Pan, R.K., 2014. Inferring human mobility using communication patterns. Scientific reports, 4, p.6174.

[8] Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.F. and Vespignani, A., 2010. Dynamics of person-to-person interactions from distributed RFID sensor networks. PloS one, 5(7), p.e11596.

[9] Roick, O. and Heuser, S., 2013. Location Based Social Networks–Definition, Current State of the Art and Research Agenda. Transactions in GIS, 17(5), pp.763-784.

[10] Gonzalez, M.C., Hidalgo, C.A. and Barabasi, A.L., 2008. Understanding individual human mobility patterns. nature, 453(7196), p.779.

[11] Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. and Ratti, C., 2014. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3), pp.260-271.

[12] Caverlee, J., Cheng, Z., Sui, D.Z. and Kamath, K.Y., 2013. Towards Geo-Social Intelligence: Mining, Analyzing, and Leveraging Geospatial Footprints in Social Media. IEEE Data Eng. Bull., 36(3), pp.33-41.

[13] Sun, Y. and Li, M., 2015. Investigation of travel and activity patterns using location-based social network data: A case study of active mobile social media users. ISPRS International Journal of Geo-Information, 4(3), pp.1512-1529.

[14] Wu, L., Zhi, Y., Sui, Z. and Liu, Y., 2014. Intra-urban human mobility and activity transition: Evidence from social media check-in data. PloS one, 9(5), p.e97010.

[15] Noulas, A., Scellato, S., Mascolo, C. and Pontil, M., 2011. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. The social mobile web, 11(2).

[16] Hasan, S., Zhan, X. and Ukkusuri, S.V., 2013, August. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing (p. 6). ACM.

[17] Gabrielli, L., Rinzivillo, S., Ronzano, F. and Villatoro, D., 2014. From tweets to semantic trajectories: mining anomalous urban mobility patterns. In Citizen in Sensor Networks (pp. 26-35). Springer, Cham.

[18] Bezdek, J.C., 1981. Objective function clustering. In Pattern recognition with fuzzy objective function algorithms (pp. 43-93). Springer, Boston, MA.

[19] Zandbergen, P.A., 2009. Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. Transactions in GIS, 13, pp.5-25.

[20] Kovács, F., Legány, C. and Babos, A., 2005, November. Cluster validity measurement techniques. In 6th International symposium of hungarian researchers on computational intelligence.

[21] Foursquare documentation page. https://developer.foursquare.com/docs (accessed on January 12,2018).

[22] Fusion table help page. https://support.google.com/fusiontables/answer/2571232?hl=en (accessed on January 15, 2018).