



GeoHash tag based mobility detection and prediction for traffic management

Jai Prakash V Verma¹ · Sapan H Mankad¹ · Sanjay Garg¹

Received: 5 November 2019 / Accepted: 3 May 2020 / Published online: 16 July 2020
© Springer Nature Switzerland AG 2020

Abstract

User mobility detection and prediction can help in many ways for planning and monitoring a population distribution and displacement in a specific area or location. It also helps in planning for resource distribution and allocation. In this paper, we present a study on the mobility of the population which can be monitored by sensors, GPS devices through location detection. All the public transports like buses, taxis, etc. have GPS devices posted that can provide movement of a vehicle in city traffic. For experimental analysis, the T-Drive Taxi Trajectories dataset was selected, in which taxi location data are collected with the information of taxi ID, timestamp, latitude, and longitude. GeoHash tags were generated for the location of a taxi based on latitude and longitude. A graph was built based on vertices as GeoHash tag and edges as a direct link between the GeoHash tags. Graph-Based data analysis was applied to identify the importance of GeoHash tag based on the in-degree and PageRank of the vertices. The mobility path and movement of traffic can be predicted that can be used for disaster management and urban development for city planning.

Keywords User mobility · Geoscience dataset · Big data analytics · Graph analytics · GeoHash tags

1 Introduction

In the Asia-Pacific regions, both developing and developed countries facing traffic management issues in urban areas because of the increasing population. It increases the number of vehicles on roads. Over the years, a wide variety of traffic management systems have been developed for urban traffic control [16]. Mobility path information of cell phone users plays an important role in a wide range of cell phone applications, such as context-based search, advertising, early warning systems [1, 17], traffic planning and monitoring [8, 15], route prediction [10, 11], and air pollution exposure estimation [3]. Mobility and demographic based user profile learning is important for these applications [6]. User's Mobility Detection is the approach by which one can find the location of a user using a GPS sensor equipped in a smart device. There is a large number of

ways by which we can get this Geolocation information of a user. For example, using sensors, we can detect the location of a user. With the advent of the Internet of Things, a large number of devices are connected and communicating with the devices carried by humans. Thus, there is a need for a robust system to detect a user's movement to learn how it works in different applications. User mobility includes the study of different areas like the user's point of interest, routes, traffic, individual mobility pattern, etc. Analysis of patterns of tracks and investigating the evolution of patterns over time [23].

An analysis is done at a particular interval of time. There is a time-dependent social graph that provides random and social interactions. There is a systematic way of detecting and tracking human movement by their mobility data generated from different Geo Location sensors. Characteristics of group mobility are group evolution, periodicity

✉ Jai Prakash V Verma, jaiprakash.verma@nirmauni.ac.in; Sapan H Mankad, sapanmankad@nirmauni.ac.in; Sanjay Garg, sgarg@nirmauni.ac.in | ¹Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, India.



and meeting duration [5]. Nguyen et al. [19] presents a novel method for mobility detection based on social events and their relationships as a knowledge graph. The graph based analysis is presented with the data collected from social media.

Characterization of group evolution is done over time by considering first the structure change rates of growth, contraction, birth, and death; second the group meeting periodicity; and third the group meeting duration and correlation with the strength of the bond of group and group's stability [23]. The mobility of users in networks is responsible for dynamic changes in user accesses to base stations. The rapid movement of a dense group of users in a network causes degradation of the quality of service. Dynamic base station switching schemes monitor and model the movements of the group by understanding the utilization of the network. Big data analytics can be used for developing and validating the detection methods to monitor group mobility, validate connected/idle duration of models and simulations based on dynamic switching of users between base stations [20].

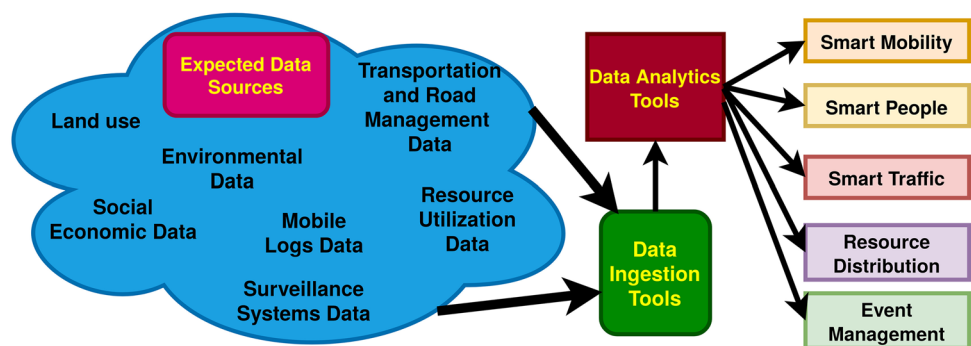
As per [4], mobile movement data were collected for analyzing user mobility. These datasets cover periods: July to August 2007 and November 2007 to June 2008 from 10 mobile operators of Portugal and May to October 2007 from 5 mobile operators of France. Both datasets contain more than a billion calls from 2 million users in Portugal (~ 20% of the total population) and 17 million users in France (~ 30% of the total population) [4]. Another mobile movement-based user mobility detection was analyzed in Haiti for understanding the population disbursement before and after an earthquake. Mobile phone location data was collected for the largest mobile phone operator in Haiti (Digicel) and analyzed the movements of 1.9 million mobile phone users during the period from 42 days before, to 341 days after the devastating [14]. The prior approach is taken into our previous paper considering a small location belongs to organization premises where the mobility of individuals was collected and analyzed using

graph-based analysis. By extending this work we are proposing detection of the trajectory of a taxi movement in the city based on the GPS location data collected. [25]

Figure 1 shows the potential fields where the outcome of the proposed work can be exploited. One of the outcomes helps smart city users to enhance the usage of the available transport facility by sharing the location information. Smart Mobility means the recommendations for sharing transport services to enhance cost-effectiveness and time saving based on smart city-data. The generated recommendation helps the people of the smart city to enhance their living standards by using recommendations about the resources and events. It can lead to a smart living environment by using these available resources optimally. This also helps people to enhance their living standards using available resources and facilities. Knowledge and pattern regarding the traffic plan help the people of the smart city to use the resources effectively and save time for transportation. The smart traffic will also be helpful for the governmental agencies for planning different events in the city smartly. The resource like land-use, environment, socio-economical, energy consumption and transportations distribution for a city should be transparent and equal. This will enable the distribution and establishment of these types of resources. Many environmental alerts before and after any disaster help the people of the smart city to make a better decision about their movement and displacement. It also enhances the accountability of the governmental agencies for any events.

The rest of the paper is organized as follows. Section 2 presents the related research work in the area of mobility detection and prediction. Section 3 presents the proposed research work of this paper. This paper is mainly based on the project related to the smart city for user mobility detection and prediction. Section 4 describes the methodology, tools, and concepts applied to achieve the proposed outcome. Section 5 presents an experimental analysis done for justification of proposed research work. In Sect. 6 result and analysis are discussed.

Fig. 1 Data source and expected outcomes for proposed work



2 Related work

Much research has been conducted in the field of user mobility detection and prediction about many application areas. Table 1 shows a comparative analysis of different approaches applied in the area of user mobility detection and prediction.

Yavas et al. (2005) [28] proposed a new algorithm for mobility prediction for mobile users. Mobile holders always reveal information about their location to the internet services provider. The algorithm is mainly predicting the next inter-cell movement of a mobile user in a Personal Communication Systems network. Kim et al. [9] proposed a mobility model with an emphasis on user movement on specific popular places, referred to as hotspot regions. They considered mobility characteristics including pause time, speed and directions of movements. Cho et al. (2011) [2] proposed a social network based user mobility prediction model. Data are collected from mobile user mobility and a social network. It predicts a daily routine with the geographic location of the user. User day to day movement pattern is identified based on social network data collected from his mobile movement.

Liu et al. (2013) [12] proposed a mathematical model for analyzing a user's moving pattern based on mobile movement data. The mobile location data are captured based on the voice call from mobile in particular locations for one year. The machine learning based model predicts user location with good accuracy. Faye et al. (2017) [7] proposed a data analysis model for predicting user mobility using data captured from smart devices often carried by users. These devices include a smartphone, smartwatches, etc. Different sensors present in most commercial smart devices can be used to deliver mobility information and patterns. It also provides a mobility assistant mechanism based on the combination of mobile wi-fi activity data.

Watanabe et al. (2017) [27] developed a novel proof-of-concept framework called RouteDetector to identify a route of the train based on readings of smart sensors devices attached. A machine learning based analysis predicted a potential path for a train in a schedule for specific locations.

Deciding a suitable data source for user mobility analysis is an important concern. A brief analysis of various data sources for user mobility in the multi-user context modeling environment is presented in [18]. They have described factors for choosing an ideal dataset for such tasks along with their desired characteristics.

Senaratne et al. in 2018 [22] introduced a visual analytics based approach for comparing mobile usage patterns and detecting anomalies in daily routines across regions and user groups. A GSM user internet usage database of

358 users collected over a period of seven months from Santiago de Chile is used to explore the Spatio-temporal patterns derived from the user movement traces. They further demonstrate their contribution in terms of similarity of user movements, classifying home and work area of users, region partitioning based on origin and destination, and temporal change detection. The outcomes of their work can be helpful for smart city urban planning and transportation management.

A recent study by Liu et al. [13] proposed a big data approach to examine geographic patterns of time-space aggregate human activity and its impact on land use characteristics. A practical approach for city policymakers and planners to understand the patterns of land use and human activity with new and emerging location-based big data is presented in their study. The work in [21] proposed a framework for big mobile data, based on real data traffic collected from second-, third- and fourth-generation networks from almost 7 million users and in densely populated areas. Their findings are helpful in the context of urban planning, traffic control, and mobile network resource optimization, etc.

3 Proposed work

The proposed research work presented in this paper, is based on the research project focused on the design and development of integrated data structure for large scale data captured for a project based on user mobility (please refer Fig. 2). Due to the large size, data generation speed, and diverse nature of smart city data, this problem is considered under the Big Data problem. A prototype has been designed and developed to demonstrate the effectiveness of the analytics service for Big Data Analytics. The prototype has been implemented using open-source solutions available for Big Data Analytics, and its results are evaluated concerning the parameters such as efficiency and effectiveness. The experiment analyzes and visualizes the data which contains the GPS trajectories of 10,357 taxis from Feb. 2 to Feb. 8, 2008, within Beijing [29, 30]. Information and Communication Technologies (ICT) and the Internet of Things (IoT) play the key roles in Smart City projects. It is a very challenging task to handle a large amount of data generated in different processes and connected devices of projects related to land-use, environment, social network and economy, energy consumption and transportations.

The proposed system architecture used for this work is shown in Fig. 3, which is divided into three segments. The functionality of each segment contributes to meet out the objective of the project. The lowest segment consists of different sensors, heterogeneous repositories. It is

Table 1 Comparative analysis of different approaches applied in the area of user mobility detection and prediction

Approach	Year	Objective	Methodology	Pros	Cons
Data mining based mobility prediction [28]	2005	Predicting next inter-cell movement of the mobile user in the personal network.	Comparison of prediction methods with transition matrix and ignorant prediction.	Proposed method gives more precision compared to both the methods.	Spatial correlation between cell towers was ignored
User mobility model from real-world wireless user traces [9]	2006	Estimation of the physical location of users from a large trace of mobile devices associating with access points in a wireless network.	Extraction of user path and hotspot regions using approaches like centroid and Kalman filter.	A software model generating realistic user tracks using the mobility characteristics.	Metrics for validating the synthetic tracks in the initial stage.
Location-based social networks Analysis [2]	2011	It combines the periodic day to day movement patterns with the social movement of the friendship network.	Relation between human geographic movement, its temporal dynamics, and the ties of the social network.	Analyze the human geography and daily routine patterns as well as the effect of social ties.	Still accuracy is just 36.1%
Analyzing mobile phone location data using machine learning [12]	2013	The user activities are being predicted whether the user is in a steady-state or moving position.	Multi-class support vector machine, Multi-nomial logistic regression, decision tree and random forest	Predict human movement based on the high spatial and temporal regularities information with the location of individuals visited multiple times.	Potential causes for miss-classifications are not discussed
Ensemble based approach for user profile learning [6]	2015	Learning a user profile based on mobility and demographic analysis	A data collection source from three social resource sites is built.	A novel dataset has been contributed.	Demographic profiling can be strengthened further by adding more attributes.
User mobility using mobile sensing [7]	2017	Analysing mobility behavior for 13 participants data collected with the SWIPE open-source system.	It uses the network and the activity required from the smartphones and smartwatches.	User movement through the mobile sensors and characterizes the mobility of the user.	Implementation part is missing
Tracking human mobility using mobile device sensors. [27]	2017	Identifies a route for a train trip by simply reading smart device sensors: an accelerometer, magnetometer, and gyroscope.	Machine-learning technique for experiments demonstrated that the Route Detector successfully identified routes used for a trip by train	Extracts departure/arrival times of vehicles from the sequence of the detected human activities.	The analysis was very specific, a generalized framework can be proposed.

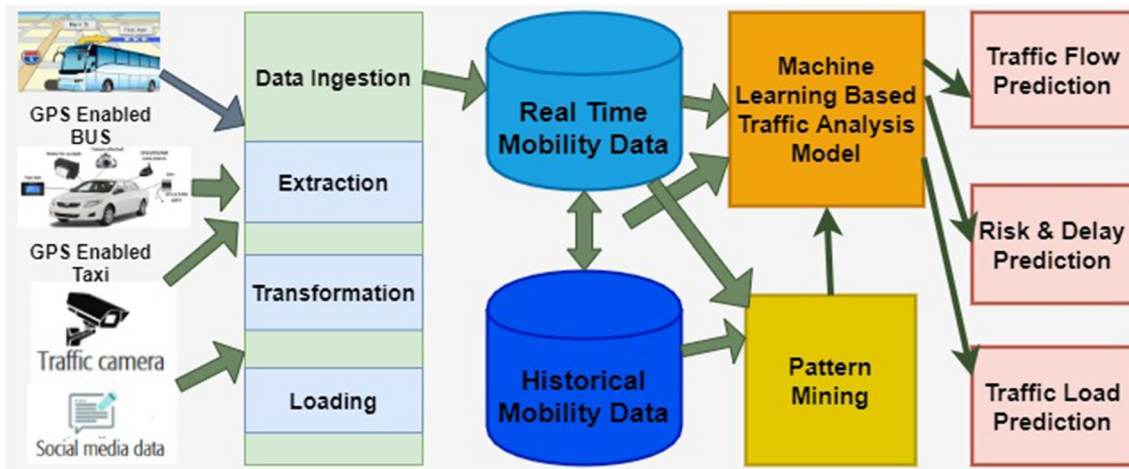


Fig. 2 Traffic detection and prediction system for smart city projects

responsible for data acquisition, data cleansing and data classification with state-of-the-art approaches. The middle segment supports the new scenario to develop links that were not possible in the lowest layer. Moreover, once the data are collected from the heterogeneous sources then the mapping between resources has been established. Then, data are made semantically relevant and browse the table, which helps the end-users to select parameters for the analysis. Traditional metadata formats such as DBLP and open library are used to describe and store it. Then, mapping of this data is done with the usage of resource description semantics, which contains all the links of various resources. An analytic engine is a topmost segment that processes application-specific data. Further, it utilizes the data available with the data segment and also helps the user in query submission, algorithm processing, and

workflow to get information from repositories. To handle the aforementioned issues, big data mining has emerged as a new technique, which is used to identify large data sets because of complexity, cardinality, and continually. These are being used in various applications such as network traffic, businesses, etc. Moreover, these are useful to generate non-obvious relations and associations from a huge data set of smart cities. Since the main focus of this research work is User Mobility Detection and Prediction System, we will mainly focus on the mobility of the user and explain in detail. To achieve it, various statistical modeling, machine learning, and data mining techniques can be applied.

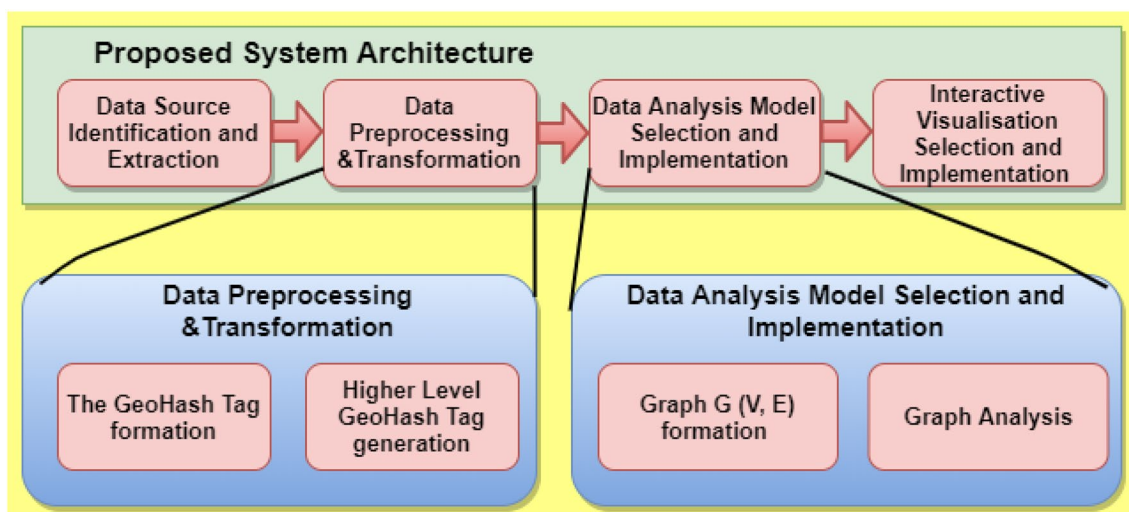










Fig. 3 Proposed data analytics approach

Table 2 List of sensors for mobility detection and prediction systems

Sr No.	Type of sensor	Example	Image	Working
1	Inductive loop	12v/24v/110v/220v AC/DC Ground Sensors Traffic Inductive Loop Vehicle Detector signal control		Built into the roads and detects the presence of conductive metal by measuring the change in magnetic field.
2	Magnetic	3Pcs MAG3110 3-Axis Digital Earth Magnetic Field Geomagnetic Sensor Module I ² C Interface For Arduino		Detects the presence of ferrous metal object through the magnetic anomaly which they cause in the Earth's magnetic field.
3	Video image processor	Wireless Video Sensor Platform		Analyzes video images of roadways surveillance cameras and provides traffic flow data across several lanes.
4	Microwave radar	MR-WBBM8030XF AC 110/250V Microwave Radar Sensor Switch Dial Code Mobile Radar Sensor Suitable for Aisle Garage Courtyard etc		Transmits electromagnetic signals and receives echoes from objects of interest.
5	Active infrared	Beams active infrared sensor		Illuminates the detection zones with low-power infrared energy transmitted by laser diodes and then uses the reflected energy to detect vehicles
6	Passive infrared	HC-SR501 PIR (Passive Infrared Sensor) Motion Sensor for Arduino		These sensors detect the energy emitted by vehicles, or by the atmosphere and reflected by vehicles.
7	Laser radar	Whistler 5050EX High Performance Laser Radar Detector: 360 Degree Protection and Bilingual Voice Alerts		These active sensors transmit scanning infrared beams in the near infrared spectrum over one or more lanes.
8	Audio	Anti spy RF Detector Wireless Bug Detector Signal for Hidden Camera Laser Lens GSM Listening Device Finder Radar Radio Scanner Wireless Signal Alarm		These are passive sensors and use different audio signals.

4 Methodology

This section describes the concepts and methodology for Big Data analysis, user mobility and graph-based data analytics with their process flow and execution. Different types of sensors for user mobility detection are discussed.

4.1 Big data analysis

A huge amount of data are generated every day from different resources at every time. The handling of these overflowing data is called big data analytics. It is the study of data, categorizing it in terms of uses, its application, and method of how the data is obtained, the size of data & many more. We can represent or categorize them in any way as per our requirement & for the best possible outcome. For example, it can store in tabular form, binary tree form, graphical form, etc. and analyze it using Excel, SQL, Hadoop, etc. In general, big data is a lot of data that cannot be processed or analyzed normally. We need specialized tools to solve these like efficient software, better hard drive, faster processor, etc. So, for example, we need to analyze 5GB of data, so if one processor takes 1 hour to complete 1 task N processor can do it $1/N$ hour. This is the approach used to analyze the overflowing data. User mobility data is generated in huge volumes with a very high speed (velocity). Also, these types of data are generated from different types of devices in different formats and variations. User mobility detection and prediction problem consider under Big Data Analytics because its data fulfills all the three dimensions (Volume, Velocity, and Verity) of Big Data. For this type of big data, we are proposing a big data environment for storage and computation in a distributed manner with Hadoop and Spark.

4.2 User mobility detection sensors

Sensors that can be used for mobility detection: The accelerometer, magnetometer, and gyroscope can be chosen as they can detect fine-grained detail about motion. The light sensor can be chosen as changes in the level of light detected may be evident between different locations. Table 2 depicts a detailed description of customarily used

sensors in various mobile manufacturing companies. In the initial time of ATMS and ITS systems, the data was being captured using different types of sensors present at fixed positions. These sensors were able to detect the nearby vehicles passing through it. Earlier, inductive loop detectors were most popular to detect the vehicles but nowadays there are a variety of fixed position sensors available as listed in Table 2.

4.3 Graph based data analytics

A social information graph can be build where GeoHash tag of an individual taxi represents the vertex. For edges between these vertices, the link is identified based on the GeoHash tags collected with a different timestamp. The difference between these timestamps for an individual taxi shows the relationship between the vertices of the graph. A graph $G(V, E)$ is built with a set of vertices V and a set of edges E . Here a set of vertex V is represented by GeoHash tags which are generated by a combination of latitude and longitude and set of edges E represents the link between these vertices. The dataset collected from the T-Drive Taxi Trajectories dataset provides the values of attribute taxi-id, timestamp, latitude, and longitude. Here a set of edges shows a link between these GeoHash tags that have timestamp differences more than by given threshold value. As per algorithm 1, steps 1 to 3 are showing the vertex generation using latitude and longitude from the given dataset D . Step 4 to 6 shows the process to generate edges from these vertices. Step 7 is for building a graph using the set of vertices V and a set of edges E . Step 8 to 13 describe how in-degree is generated for each vertex and printed. Step 14 to 18 shows the execution of the page rank algorithm. Table 3 provides notations used for page rank algorithm.

Algorithm 1: Pseudo code of the proposed approach

```

Input: Set of Vertex  $V$  and edge  $E$ 
 $D$  set of data tuple with attribute set (Taxi-id, Latitude, Longitude, and Timestamp)
Output:  $PR_i$  : Page Rank Score of the vertex  $i$ 
1 for each  $d_i$  in  $D$  do
2    $v_i = \text{GeoHashTag}(d_i(\text{Latitude}, \text{Longitude}))$ 
3   Add  $v$  to set of vertices  $V$ .
4 for each  $v_i$  in  $V$  do
5    $e(v_i, v_j) = \text{TimeStamp}[v_j] - \text{TimeStamp}[v_i]$  // if difference is more then by given
   threshold value.
6   Add  $e$  to set of edges  $E$ 
7 Graph  $\leftarrow (V, E)$ 
8 for each  $v_i$  in  $V$  do
9    $\text{deg}(v_i)^- \leftarrow$  The number of links coming in to a vertex  $P_i$ .
10   $\text{deg}(v_i)^+ \leftarrow$  The number of links going out to a vertex  $P_i$ .
11   $\text{deg}(v_i) = \text{deg}(v_i)^- + \text{deg}(v_i)^+$ 
12 #Query: Get in-degree of each vertex.
13 Print In-degree of  $V$ .
14 for each  $v_i$  in  $V$  do
15   $PR_i \leftarrow$  Page Rank Score of the vertex  $i$ .
16   $\text{result}[i] \leftarrow PR_i$ 
17 #Run PageRank algorithm, and show results.
18 Print results # PageRank Score of each vertex
    
```

5 Experimental analysis

This section describes the execution plan to achieve the proposed objectives. For experiment analysis, a Big Data environment is set up with Hadoop for distributed storage [24] and Spark for distributed computing [26]. Spark's GraphX component provides the API for graph analytics like in-degree, PageRank, etc. Here we applied spark's PySpark which provides an interface with Resilient Distributed Datasets in apache spark and python. GraphFrame is used to access all the API of the GraphX component of the spark which is implemented in the scala programming language.

5.1 Dataset selection

For experimental analysis, the T-drive Taxi Trajectories from Nokia MDC datasets are selected. A total 10,357 taxi

covered around 9 million kilometers to generate this dataset. Around 15 million points covered for GPS trajectories, the dataset was generated in the period of 2 to 8 February 2008. Figure 4 shows the data distribution with time and distance intervals [29, 30]. The format of data contains in *taxi id, timestamp, longitude, latitude*.

5.2 Data preprocessing

Data transformation is the process of converting/transforming data from one form or structure into another form or structure. Data transformation is critical to operations such as data integration and data management. Data transformation can include a range of activities such as: converting data types, data cleaning by removing null values or duplicate data, enrich the data, or perform aggregations, depending on the requirements of the project. Data Discovery performs knowledge gain of user mobility with

Table 3 Page rank notations

Symbol	Meaning
P_i	A vertex i , here vertex is representing a mobile tracer device
dmp	Damping factor: the probability that the user opens a new vertex (the mobile tracer device) to begin a new random walk.
$PR(P_i)$	Page Rank Score of the vertex i
$\text{deg}(P_i)^-$	The number of links coming in to a vertex P_i (In-Degree of P_i)
$\text{deg}(P_i)^+$	The number of links going out from a vertex P_i (Out-Degree of P_i)

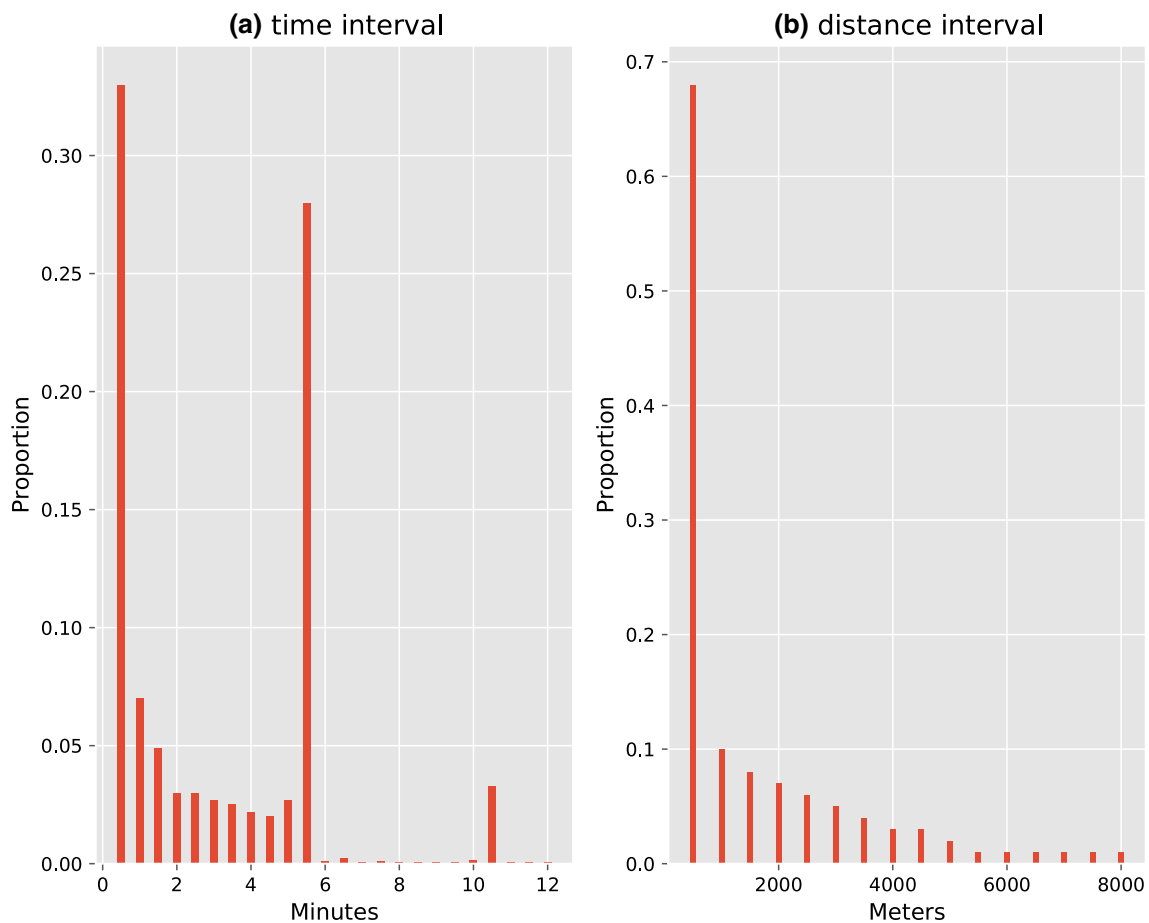


Fig. 4 Histograms of time interval and distance between two consecutive points [29, 30]

the correlation between locations. Data Structure containing records of booking of a taxi with timestamp and latitude and longitude of the pickup location. Extracting Data combine these multiple data files into one single file containing all the record of *taxi-id*, *latitude*, *longitude* and *timestamp*. Here transformation of latitude and longitude are concatenated together to form a unique area called GeoHash tag. The Geohash is the method for encoding regions with specific precision of latitude and longitude. Geohashes offer properties like arbitrary precision and the possibility of gradually removing characters from the end of the code to reduce its size. After applying the geo hash method we get the data in the form of the shorter hash. From the GeoHash file and the simulator file, we can get the group of id containing the same location/ area (see Fig. 5). For example:

2-4,	URTFD,	-18375.0
2-4,	URTFD,	-12837.0
...		

5.3 Graph based user mobility detection and prediction

Figures 6 and 7 show the Python code used to build a Graph-based onset of vertices V and set of edges E generated as shown in Fig. 5c. Geohash represents the GPS location of an individual taxi. Edges between these Geohash tags show the link between the vertexes. Figure 6 shows the steps for building graph $G(V, E)$. The graph shows the connections between the GeoHash tags. A Big Data environment is set up using Hadoop to achieve distributed storage and Spark for achieving distributed computing. The GraphX component of the Spark framework provides API to execute graph algorithms and Google’s PageRank algorithm. GraphX is implemented in the Scala programming language and can be only used by the Scala program. To overcome this issue, the Graphframe, a python implementation to execute GraphX API is used. In-degree and PageRank algorithms are executed as per Fig. 7.

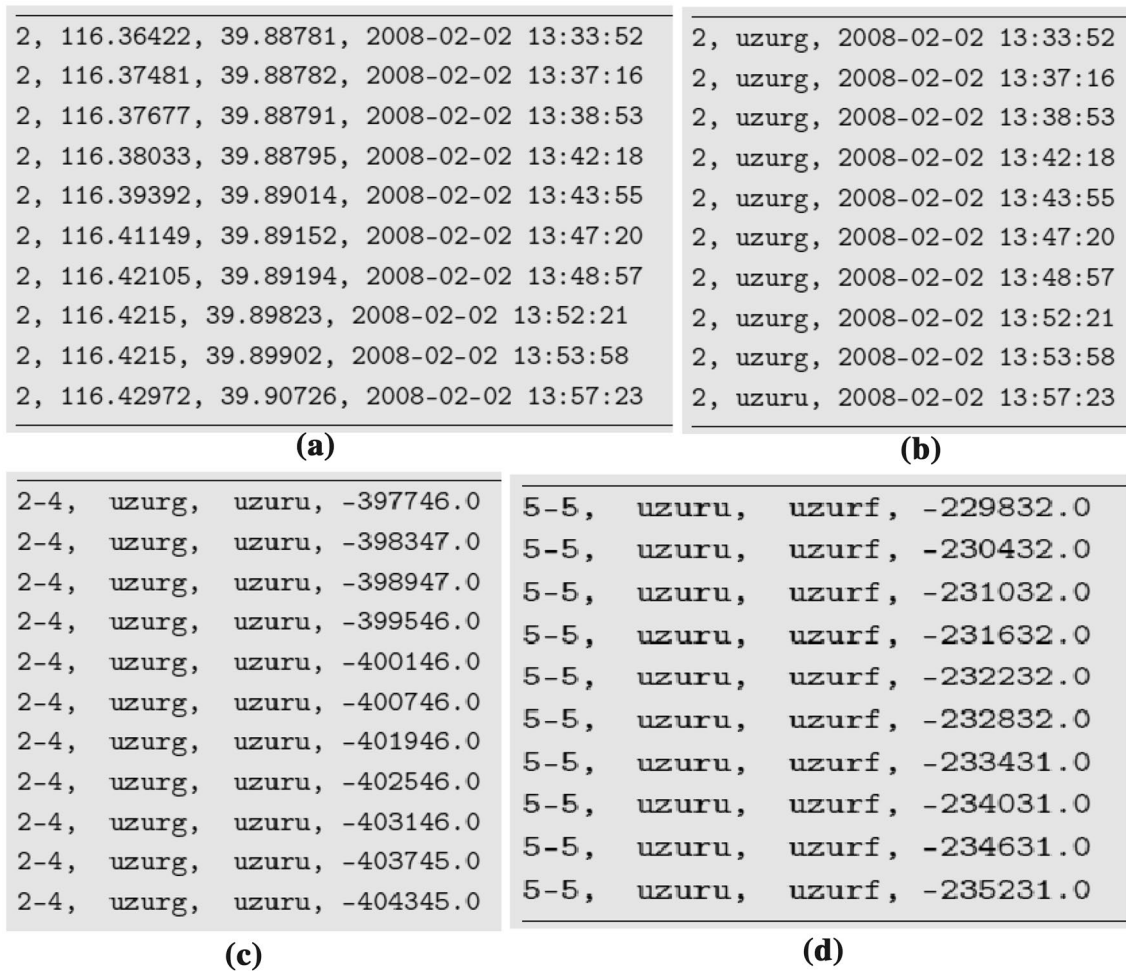


Fig. 5 Data file after applying GeoHash and edges generation

5.4 Results

TableS 4 and 5 show the importance of GeoHash tags based on graphs' In-Degree and Google's PageRank algorithms. The importance of geolocations based on the combination of latitude and longitude on different levels is observed. It is also observed that there are various effects of traffic on different levels of GeoHash tags generated. Table 4 shows top 20 GeoHash tags using the In-degree algorithm and Table 5 shows top 10 Geohash tags using PageRank algorithm. This can be also used to detect the mobility patterns in a city which helps for the planning of traffic arrangements in urban areas. The technique GeoHash which we have used for data transformation gives the dissimilarity with greater precision. But this precision leads to higher computation cost which is the major drawback. So to reduce the computation cost the value of precision needs to be decreased.

Table 4 Top 10 important GeoHash tags based on in-degree of vertices

Sr No	ID	InDegree	Sr No	ID	InDegree
1	uzurg	3661217	6	uzurz	185536
2	uzuru	2862895	7	uzurc	122739
3	uzury	1769219	8	uzurb	30348
4	uzurv	900578	9	uzuxb	16716
5	uzurf	461286	10	uzgzv	4254

6 Discussion

The approach proposed in this work and the analysis thereafter may help in traffic planning at the city level as well as for infrastructure setup. It is more motivated compared to our previous work on user mobility detection and prediction in small premises for a smart city project [25]. As per the result computed Tables 4 and 5 the

```

pyspark --packages graphframes:graphframes:0.1.0-spark1.6
import graphframes
from graphframes import *
from pyspark.sql import SQLContext
from pyspark import SparkContext
sqlCtx = SQLContext(sc)
# For Vertax
lines =
sc.textFile("hdfs://localhost:8020/user/file2.txt")
parts = lines.map(lambda l: l.split(","))
v = parts.map(lambda p: {"id": p[0], "name": p[1]})
v1 = sqlContext.createDataFrame(v, ["id", "name"])
# For Edges
edges =
sc.textFile("hdfs://localhost:8020/user/file3.txt")
parts_ed = edges.map(lambda l: l.split(","))
e = parts_ed.map(lambda p: {"id1": p[1], "id2": p[2],
"Relationship": p[3]})
e1 = sqlContext.createDataFrame(e, ["src", "dst",
"relationship"])
#Create a GraphFrame
g = GraphFrame(v1, e1)

```

Fig. 6 Python code for Graph Generation using GeoHash Tags

```

g.inDegrees.show()
#Run PageRank algorithm, and show results.
results = g.pageRank(resetProbability=0.01, maxIter=20)
results.vertices.select("id", "name").show()
#what are the most visited device
results = g.pageRank(resetProbability=0.15, maxIter=10)
results.vertices.select("id", "name").show()

```

Fig. 7 Python code for execution of In-Degree and Pagerank algorithm

historical trajectory can be used to predict the future trajectory of moving vehicles. The graph shows the connection between different vertices where vertex representing the taxi id. The graph-based analysis will help plan the city traffic as well as infrastructural setup. Here vertex is representing the trajectory of taxi's at an instance of time and the edges are representing the taxi's trajectory for taxi movement. PageRank algorithm of the graph showing the highly influenced vertex in the graph which can be used for traffic management and planning.

However, the method becomes inefficient when the number of data increases. The approach isnt scalable as it needs to train the model separately for every taxi,

Table 5 Top 20 GeoHash tags based on pagerank algorithm

Sr No	ID/Name	Sr No	ID/Name
1	uzury	11	Uzury
2	uzuru	12	Uzuru
3	uzury	13	Uzurg
4	uzury	14	Uzurg
5	uzuru	15	Uzuru
6	Uzuru	16	uzurf
7	Uzuru	17	uzurf
8	Uzuru	18	uzurg
9	Uzuru	19	uzurg
10	Uzuru	20	uzurf

especially span a large area of the road network. Here we are suggesting that the trajectory of a specific taxi is also correlated to the trajectory of taxis in its nearby region. This approach is scalable to even large amounts of data that require big data engines while also being able to predict long-term trajectories in large cities.

7 Conclusion and future work

In this paper, we have contributed to the design and implementation of a prototype with an objective to demonstrate the effectiveness of the analytics service for Big Data Analytics. User mobility detection based on the location data generated from different Geo location sensors is the main objective of this paper. This can help to track user movement and pattern prediction of the path, thus the population movement and distribution during a specific time period can also be identified and predicted. We have used the T-driveTaxi Trajectories from Nokia MDC datasets for predicting taxi movement. Geohash tag is generated and a Hadoop and Spark based Big Data environment is set up for data analysis. Pattern identification based on the past mobility pattern can also be generated where machine learning and deep learning algorithms can play a vital role.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Ashbrook D, Starner T (2003) Using gps to learn significant locations and predict movement across multiple users. *Pers Ubiquitous Comput* 7(5):275–286
- Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '11*, pp 1082–1090, New York, NY, USA, 2011. ACM
- Demirbas M, Rudra C, Rudra A, Bayir MA (2009) Imap: indirect measurement of air pollution with cellphones. In: *2009 IEEE international conference on pervasive computing and communications*, pp 1–6, March 2009
- Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. *Proc Nat Acad Sci* 111(45):15888–15893
- Eagle N, As P (2006) Reality mining: sensing complex social systems. *Pers Ubiquitous Comput* 10(4):255–268
- Farseev A, Nie L, Akbari M, Chua TS (2015) Harvesting multiple sources for user profile learning: a big data study. In: *Proceedings of the 5th ACM on international conference on multimedia retrieval*, pp 235–242. ACM
- Faye S, Bronzi W, Tahirou I, Engel T (2017) Characterizing user mobility using mobile sensing systems. *Int J Distrib Sens Netw* 13(8):1550147717726310
- Harrington A, Cahill V (2004) Route profiling: putting context to work. In: *Proceedings of the 2004 ACM symposium on Applied computing (SAC '04)*. Association for Computing Machinery, New York, USA, pp 1567–1573. <https://doi.org/10.1145/96790.968214>
- Kim M, Kotz D, Kim S (2006) Extracting a mobility model from real user traces. In: *Proceedings IEEE INFOCOM 2006. 25TH IEEE international conference on computer communications*, pp 1–13, April
- Laasonen K (2005) Clustering and prediction of mobile user routes from cellular data. In: *Proceedings of the 9th European conference on european conference on machine learning and principles and practice of knowledge discovery in databases, ECMLPKDD'05*, pp 569–576, Berlin, Heidelberg, 2005. Springer-Verlag
- Laasonen K (2005) Route prediction from cellular data. *CAPS'05*, pp 147–158
- Liu F, Janssens D, Wets G, Cools M (2013) Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Syst Appl* 40(8):3299–3311
- Liu W, Wenjie W, Thakuria P, Wang J (2020) The geography of human activity and land use: a big data approach. *Cities* 97:102523
- Xin L, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. *Proc Nat Acad Sci* 109(29):11576–11581
- Lv Q, Qiao Y, Ansari N, Liu J, Yang J (2016) Big data driven hidden Markov model based individual mobility prediction at points of interest. *IEEE Trans Veh Technol* 66(6):5204–5216
- Makino H, Tamada K, Sakai K, Kamijo S (2018) Solutions for urban traffic issues by its technologies. *IATSS Res* 42(2):49–60
- Marmasse N, Schmandt C (2002) A user-centered location model. *Pers Ubiquitous Comput* 6(5–6):318–321
- Mehta P, Voisard A (2012) Analysis of user mobility data sources for multi-user context modeling. In: *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, pp 9–14. ACM
- Nguyen HL, Jung JJ (2019) Social event decomposition for constructing knowledge graph. *Future Gener Comput Syst* 100:10–18
- Nunes IO, de Melo POSV, Loureiro AAF (2016) Group mobility: detection, tracking and characterization. In: *2016 IEEE international conference on communications (ICC)*, pp 1–6, May
- Qiao Y, Cheng Y, Yang J, Liu J, Kato N (2017) A mobility analytical framework for big mobile data in densely populated area. *IEEE Trans Veh Technol* 66(2):1443–1455
- Senaratne H, Mueller M, Behrisch M, Lalanne F, Bustos-Jimenez J, Schneidewind J, Keim D, Schreck T (2018) Urban mobility analysis with mobile network data: a visual analytics approach. *IEEE Trans Intell Transp Syst* 19(5):1537–1546
- Suzuki M, Kitahara T, Ano S, Tsuru M (2018) Group mobility detection and user connectivity models for evaluation of mobile network functions. *IEEE Trans Serv Manag* 15(1):127–141
- Verma JP, Mankad SH, Garg S (2018) Big data analytics: performance evaluation for high availability and fault tolerance using mapreduce framework with hdfs. In: *2018 Fifth International conference on parallel, distributed and grid computing (PDGC)*, pp 770–775, Dec
- Verma JP, Mankad SH, Garg S (2019) A graph based analysis of user mobility for a smart city project. In: *Next generation computing technologies on computational intelligence*, pp 140–151, Singapore, 2019. Springer Singapore

26. Verma J, Patel A (2018) Comparison of mapreduce and spark programming frameworks for big data analytics on HDFS. 09 2018
27. Watanabe T, Akiyama M, Mori T (2017) Tracking the human mobility using mobile device sensors. *IEICE Trans Inf Syst* E100.D 8:1680–1690
28. Yava G, Katsaros D, Ulusoy ö, Manolopoulos Y (2005) A data mining approach for location prediction in mobile environments. *Data Knowl Eng* 54(2):121–146
29. Yuan J, Zheng Y, Xie X, Sun G (2011) Driving with knowledge from the physical world. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '11, pp 316–324, New York, NY, USA, 2011. ACM
30. Yuan J, Zheng Y, Zhang C, Xie W, Xie X (2010) Guangzhong Sun, and Yan Huang. Drive: driving directions based on taxi trajectories. In: Proceedings of 18th ACM SIGSPATIAL conference on advances in geographical information systems. ACM SIGSPATIAL GIS 2010, November 2010. Best Paper Award

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.