Analysis of Trajectory Data in Support of Traffic Management: A Data Mining Approach

Ahmed Elragal[1] and Hisham Raslan[2]

[1] Department of Business Informatics & Operations, German University in Cairo (GUC), Cairo, Egypt
ahmed.elragal@guc.edu.eg
[2] Teradata Egypt,21 Giza st, Giza, Egypt
hisham.raslan@teradata.com

Abstract. Huge amount of location and tracking data is gathered by location and tracking technologies, such as global positioning system (GPS) and global system for mobile communication (GSM) devices; leading to the collection of large spatiotemporal datasets and to the opportunity of discovering usable knowledge about movement behavior. Movement behavior can be extremely useful in many ways when applied, for example, in the domain of traffic management, planning metropolitan areas, mobile marketing, tourism, etc. In this research, we move towards this direction and propose a framework for finding trajectory patterns of frequent behaviors using GSM data. The research question is "how to use trajectory data analysis in support of solving traffic management problems utilizing data mining techniques?" Our framework is illustrated to explain how GSM data can provide accurate information about population movement behavior, and hence support traffic decisions.

Keywords: trajectory, data mining, traffic management.

1 Introduction

The movement of people or vehicles within a given area can be observed from the digital traces left behind by the personal or vehicular mobile devices, and collected by the wireless network infrastructures. For instance, mobile phones leave positioning logs, which specify their localization, or cell, at each moment they are connected to the GSM network. The increasing use of these technologies will make available large amounts of data pertaining to individual trajectories; therefore, there is an opportunity to discover, from these trajectories, spatiotemporal patterns that convey useful knowledge. Spatiotemporal patterns that show the cumulative behavior of a population of moving objects are useful abstractions to understand population movement behavior. In particular, a form of pattern, which represents an aggregated abstraction of many individual trajectories of moving objects within an observed population, would be extremely useful in the domain of sustainable mobility and traffic management in metropolitan areas, where the discovery of traffic flows among sequences of different places in a town can help decision makers take well informed decisions in different areas such as traffic management and urban planning.

P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 174–188, 2014.

In many application domains, useful information can be extracted from moving object data if the meaning as well as the background information are considered. The knowledge of moving patterns between different places in the geographic space may help the user to answer queries about moving objects or movement behavior. In order to capture and model such pattern relationships, data mining techniques play an essential role.

For the purpose of this research, traffic management is the direction, control, and supervision of all functions related to road-related passenger transportation services. Traffic management is a key towards the creation of a modern and sustainable city. Traffic management becomes a rather critical job in an over-populous city with limited resources. In a city over populated like Cairo with millions of vehicles and very limited roadways, traffic management is impossible without the aid of technology.

The goal of this research is to develop framework for finding patterns in trajectories utilizing data mining techniques. The framework can be used to find hidden patterns in trajectory data to support spatiotemporal semantic decisions in the area of traffic management and analysis. The research relies on a case study based on GSM data from Cairo to support answering the research question "how to use trajectory data analysis in support of solving traffic management problems utilizing data mining techniques?"

Our approach is directed towards building a framework to support the management and analysis of traffic data. Fig. 1. Initial framework explains the suggested framework which we are going to build and test.

Fig. 1. Initial framework

Sample data was used in the experimental case study. It's a real data taken from a GSM operator in Egypt. The data is mapped into known regions and stored in a central data warehouse where knowledge discovery tasks take place to extract useful knowledge in support of traffic management and analysis.

The remaining of this paper organized as follows: section 2, related work; section 3, traffic management; section 4 GSM data in support of traffic management; section 5, proposed framework; section 6, conclusion; and references at the end.

2 Related Work

The literature of trajectory data mining and mining moving objects has witnessed different waves focusing on various themes ranging from those focused on frameworks to those focusing on visual interaction between solution and user.

Many research focused on location information, sequence, and regions of interests e.g., [1], [2], [3], [4], [5], and [6]. [1] propose a spatial context model, which deals with the location prediction of mobile users. The model is used for the classification of the users' trajectories through Machine Learning (ML) algorithms. Predicting spatial context is treated through supervised learning. [2] Developed an extension of the sequential pattern mining paradigm that analyzes the trajectories of moving objects. They introduced trajectory patterns as a descriptions of frequent behaviors, in terms of both space (i.e., the regions of space visited during movements) and time (i.e., the duration of movements). [3] Provided a case study of movement data analysis where statistical means and pattern mining were merged together with visualization techniques to improve understanding of data. [4] Tried to cope with the complexity of trajectory semantics in terms of three evolving steps: Trajectory Modeling which considers spatiotemporal features (like trajectory modeling data type moving point) and semantic trajectory units (like stops and moves); Trajectory Computing Propose corresponding bottom up computational solutions for targeting semantic trajectories; Trajectory Pattern Discovery Investigate the mining and learning algorithms for the computed semantic trajectories. [5] Proposed a model for trajectory patterns and a measure to represent the expected occurrences of a pattern in a set of imprecise trajectories. The concept of pattern groups is introduced to present the trajectory patterns and a new min-max property was identified. A TrajPattern algorithm was devised based on the newly discovered property and the algorithm was applied on a wide range of real and synthetic data sets to demonstrate the usefulness, efficiency, and scalability of this approach. [6] Mined interesting locations and travel sequence in a given Geo-spatial region using GPS trajectories. In their paper they improved location based services by integrating Social Networking into Mobile Web.

Another wave focused on developing framework on trajectories analysis and pattern detection e.g., [7], and [8]. [7] Proposed a reverse engineering framework for mining and modeling semantic trajectory patterns. They applied data mining to extract general trajectory patterns, and through a new kind of relationships, they model these patterns in the geographic database schema. They used a case study able to shows the power of the framework for modeling semantic trajectory patterns in the geographic space. [8] developed MoveMine which is able to integrate data mining functions including moving object pattern mining and trajectory mining based on novel methods. MoveMine is able to perform trajectory clustering, classification and outlier detection. The output of the system could be written in Google maps and Google earth. Their system consists of three layers: 1- data collection and cleaning, 2- Mining, 3- Visualization interface.

Another research defined trajectory data warehouse (TDW) that is loaded by spatiotemporal observations and studied how standard data warehousing tools can be used to store trajectories and to compute OLAP operations over them e.g., [9], [10], [11]. [9] Presented an approach for storing and aggregating spatio-temporal patterns by using a Trajectory Data Warehouse (TDW). [10] Investigated the extension of Data Warehousing and data mining technology so as to be applicable on mobility data. In his work, he presented the developed framework for analyzing mobility data

and some preliminary results. [11] Aims was to make trajectories as a first class concept in the trajectory data conceptual model and to design a TDW, in which data resulting from mobile information collectors' trajectory are gathered. These data will be analyzed, according to trajectory characteristics, for decision making purposes, such as new products commercialization, new commerce implementation, etc.

Further research focused on the techniques and algorithms needed to find patterns and develop solutions e.g., [12], [13], and [2]. [12] Proposed a data preprocessing model to add semantic information to trajectories in order to facilitate trajectory data analysis in different application domains. [13] Described the analysis, pre-processing, modeling, and storage techniques for trajectory data that constitute a Moving Object Database (MOD). MOD is the backbone of the 'PATH-FINDER' system, which specifically focuses on extracting further information about the movement of vehicles in the Athens municipal area.

Few research studies were directed towards visual interaction and security e.g., [14], [15], and [16]. [16] Proposed a visual-interactive monitoring and control framework extending the basic Self-Organizing Map (SOM) algorithm. The framework implements the general Visual Analytics idea to combine automatic data analysis with human expert supervision. It provides facilities for visually monitoring and interactively controlling the trajectory clustering process. They Applied the framework on a trajectory clustering problem, to demonstrate its potential in combining both unsupervised (machine) and supervised (human expert) processing, to produce appropriate cluster results. [14] Proposed a general framework to solve the conflict between the data mining methods, which want as precise data as possible, and the users who want to protect their privacy by not disclosing their exact movements. The framework allows user location data to be anonymized, thus preserving privacy, while still allowing interesting patterns to be discovered. The framework allows users to specify individual desired levels of privacy that the data collection and mining system will then meet. [15] Studied the privacy threats in trajectory data publishing and show that traditional anonymization methods are not applicable for trajectory data due to its challenging properties: high-dimensional, sparse, and sequential. Their primary contributions are (1) to propose a new privacy model called LKC-privacy that overcomes these challenges, and (2) to develop an efficient anonymization algorithm to achieve LKC-privacy while preserving the information utility for trajectory pattern mining.

3 Traffic Management

In recent years, urban traffic congestion has become a huge problem in many cities across many countries. In order to reduce congestion, governments usually invest in improving city infrastructures. However, infrastructure improvements are very costly to undertake; hence, existing infrastructure and vehicles have to be used more efficiently. To reduce traffic congestion, it is necessary to conduct further research on the various characteristics of traffic flow patterns. In general, road traffic system consists of many autonomous, such as vehicle users, public transportation systems, traffic lights and traffic management center, which distribute over a large area and interact with one another to achieve an individual goal. Traffic management objective

is to increase the efficient passages of every vehicle, while at the same time reduce the number of vehicles on the street. Therefore, research on traffic information control and traffic guidance strategies are particularly necessary and important [17].

Urban traffic analysis and control is a complex problem that is difficult to analyze with traditional analytical methods. The degree of complexity of vehicle movement in urban centers is such that modeling and simulation techniques have been gaining popularity as analysis tool. Simulation entitles the study of particular problems, allowing providing solutions based on experimentation. [18] Presented the results of a project to build modeling and simulation tools with this purpose. The first stage of this project was devoted to define and validate a high level specification language representing city sections. This language, called ATLAS (Advanced Traffic LAnguage Specifications) focuses on the detailed specification of traffic behavior. ATLAS is a specification language defined to outline city sections as cell spaces. A static view of the city section to be analyzed can be defined and a modeler is able to define complex traffic models in a simple fashion. [19] Used the term 'urban data-mining' which they described as a methodological approach that discovers logical or mathematical and partly complex descriptions of urban patterns and regularities inside the data.

Trying to understand, manage and predict the traffic phenomenon in a city is both interesting and useful. For instance, city authorities, by studying the traffic flow, would be able to improve traffic conditions, arrange the construction of new roads, the extension of existing ones, and the placement of traffic lights. This target can be served by analyzing traffic data to monitor the traffic flow and thus to discover traffic related patterns. These patterns can be expressed through relationships among the road segments of the city network. We aim to discover, by using aggregated mobility data, how the traffic flows in this network, the road segments that contribute to the flow and how this happens. We believe that, the application of new information technologies such as trajectory mining technology to urban traffic information control can make it possible to create and deploy more intelligent systems for traffic control and management to support road managers in traffic management tasks.

4 GSM Data in Support of Traffic Management

Traffic management is analyzed and studied either based on original data collected by means of cameras, GPS, or other tracking systems. In the absence of this, traffic management in studied using synthetic data. However, synthetic data does not really represent the real problem and is hard to generalize. Meanwhile, collecting traffic data needs certain setup, which is not always available. Therefore, in our framework we used GSM data to understand the traffic patterns and behavior. In the following we will describe how GSM data can help for that purpose.

GSM mobile network is a radio network distributed over land areas called cells, each served by at least one fixed-location transceiver, known as a cell site or base station. When joined together these cells provide radio coverage over a wide geographic area. This enables a large number of portable transceivers (e.g., mobile phones, pagers, etc.) to communicate with each other and with fixed transceivers and telephones anywhere in the network, via base stations. The most common example of

a cellular network is a mobile phone (cell phone) network. A mobile phone is a portable telephone which receives or makes calls through a cell site (base station), or transmitting tower. As the phone user moves from one cell area to another cell whilst a call is in progress, the mobile station will search for a new channel to attach to in order not to drop the call. Once a new channel is found, the network will command the mobile unit to switch to the new channel and at the same time switch the call onto the new channel.

Every time a subscriber makes or receives a call the network generates a Call Details Record CDR to store the details of the call (caller ID, called ID, time of the call, call duration, etc.) one of the CDR parameters is the cell id. It is possible to follow the path (trajectory) of a subscriber by linking the CDRs generated by the subscriber in a period of time. To study the traffic in certain areas over specific roads, we need to identify the cell ID's used by the mobile operator to cover those areas then map their locations to geographical areas (semantics) as shown in Fig. 2. Cells mapped on traffic roads. Using the CDRs generated with the identified cells over a period of time we can identify moving pattern of the operator subscribers in these areas.

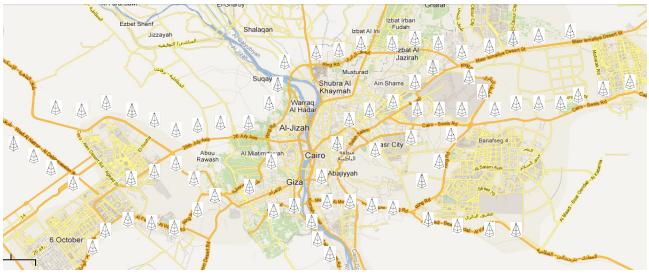

Fig. 2. Cells mapped on traffic roads

5 The Proposed Framework

The proposed framework aim is to establish an enhanced approach towards building analysis engine that can detect movement pattern that if known to the traffic department, is expected to influence their decisions and hence enhance traffic quality and save time, fuel consumption, and thereafter help Cairo becomes an environment friendly city. Based on the conducted business analysis, the traffic department is interested in analyzing data in specific areas such as traffic volume and Traffic pattern on specific roads.

The framework includes process, technologies, data, and decisions as the main aspects towards knowledge extraction. The process outlines the detailed tasks to take place at each phase and highlights the main layers which data passes through to be transformed into knowledge. The technologies used in each layer or by the tasks are also presented. The phases of the proposed framework are presented in Fig. 3. The proposed framework. The following are the tasks performed to reach the results:

1- Build Logical Data Model
2- Data loading
   - Load CDRs
   - Load Cell information
   - Create and load lookups
3- Data preparation & Semantic annotation
   - Data Quality assessment
   - Define POI to add semantics to the data
   - Identify Commuters (users)
4- Data Preprocessing
   - Trajectory Extraction
   - Build ADS
5- Perform Analytics.

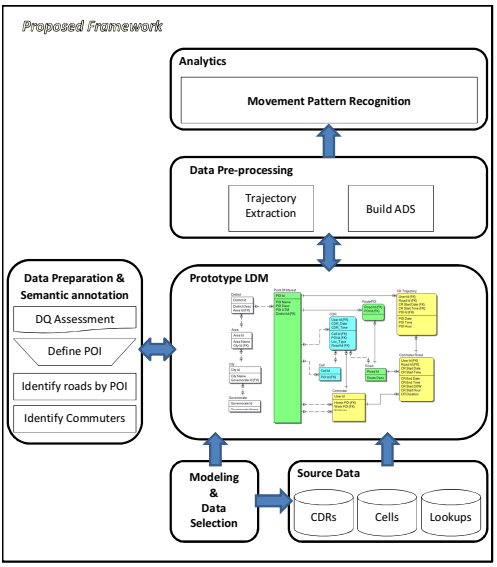

Fig. 3. The proposed framework

While using GSM data to study traffic pattern, we will not be violating people's privacy i.e., we will be using disguised data, which does not reveal real identify of people.

In the following we will provide description of each task.

5.1 Build the Logical Data Model

The foundation for the analysis engine is a well-designed Logical Data Model. The model will be used to physically realize the engine and will help the analytical users plan and develop queries and analytics. Fig. 4. LDM shows the proposed LDM that will be used to build analysis engine.

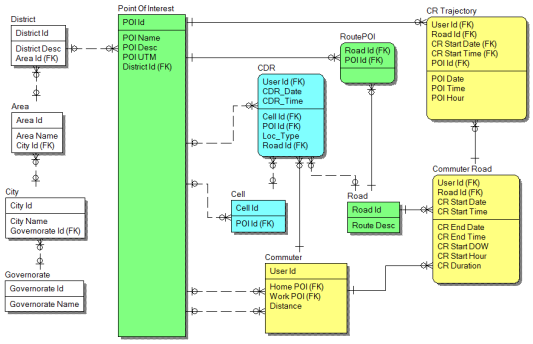

Fig. 4. LDM

The model is comprised of entities, attributes and relations; the model entities are colored to differentiate the type of information stored in each entity; blue entities are data loaded from the source (CDRs and Cells), white (uncolored) entities have semantics information (lookups: governorate, cities, areas, districts), green represents entities identified for analysis (POIs and roads), and yellow entities contain information extracted using analysis algorithms (trajectories).

5.2 Data Loading

Data loading includes the following tasks:
1. Load CDRs and Cell information
2. Create and load lookups

Load CDRs and Cell Information
Call Detail Record tables for voice and GPRS are merged into the table "CDR". The CDR table has the following structure: (User_Id, CDR_Date, CDR_Time, Cell_Id)

- Four Months of voice and data CDR's are merged into the table "CDR". Total number of CDRs is 10,314,009,634 ≈ 1.5B
- Greater Cairo cells loaded to "Cell" table in the cell id column; the POI column will be mapped later. Total number of cells is 14,175

After loading the CDR table, cell ids are loaded to the "Cell" table in the cell id column; the POI column will be mapped later. POI's were manually identified using Google earth by locating all cells on the map then grouping the cells in a specific area to a POI as we will explain later. The cell table has the following structure: (Cell_Id, POI_Id)

Create and Load Lookups
Lookups tables Governorate, City, Area, and District are created and loaded with the information required to add semantics to the analysis results.

5.3 Data Preparation and Semantic Annotation

Data Quality Profiling
Data quality is the suitability of data to meet business requirements. Because different organizations and applications have different uses and requirements for the data, data quality requirements will also differ. So data doesn't have to be perfect, but it needs to meet business requirements. We will be assessing the data for Consistency (Format and Content), Completeness, Uniqueness, and Integrity

We use data profiling tool to perform the required data quality assessments. The selected tool is part of Teradata Warehouse Miner TWM. The following tables, Table 1. Value analysis for the CDRs table, and Table 2. Value analysis for the cells table, show value analysis for the CDR and Cell table respectively.

Table 1. Value analysis for the CDRs table

Column Name	Count	null	Unique	Zero	Positive	Negative
User_ID	10314009634	0	15424023	0	10314009634	0
CDR_Date	10314009634	0	122			
CDR_Time	10314009634	0	86400	138790	10313870844	0
Cell_ID	10314009634	0	15077	0	10314009634	0

Table 2. Value analysis for the cells table

Column Name	Count	null	Unique	Zero	Positive	Negative
Cell_ID	14175	0	14175	0	14175	0
Cell_Status	14175	151	4			
POI_Id	14175	0	522	0	14175	0

Data Issues

1. Number of CDRs with missing cells: 214026321 (2% of total CDRs)
2. Number of missing cells: 1127 (8% of total Cells)

The ratios of discovered data issues (2%) will not affect the accuracy of the analysis results and can be ignored

Define Points of Interest POI

This step is comprised of the following:

1. Group cell sites to define Points of Interest POI
2. Update POI column in the cell and CDR tables

Group Cell Sites to Define Points of Interest POI

Points of Interest POI under investigation are defined by nearest cell sites. We also use the cell sites on the roads to monitor traffic events; such points are also considered POI. To locate cell sites, all cells were added to Google earth map and the cell sites near by the roads were selected (manually) and defined as POI. To add the cells to the map we use the site coordinates. The coordinates are in Universal Transverse Mercator notation (UTM).

Cell sites usually have more than one cell; also the number of cells is very large to handle! (See Fig. 5. Cell sites located on greater Cairo map).

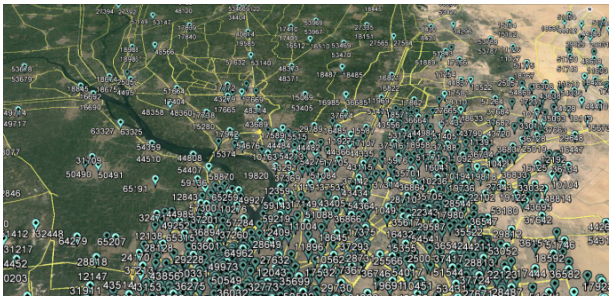

Fig. 5. Cell sites located on greater Cairo map

To reduce the number of cells we grouped the cells in one district to one or more points on the map in the center of the grouped cells. These points are used in our research as Points of Interest POI (see Fig. 6. POIs located on greater Cairo map).

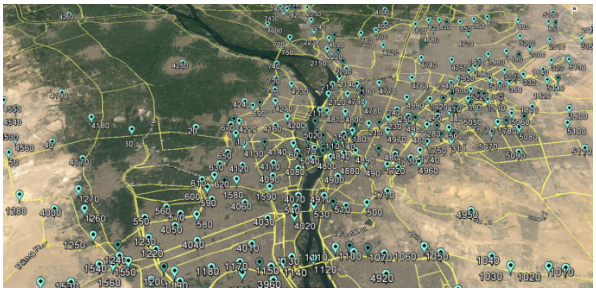

Fig. 6. POIs located on greater Cairo map

The process of grouping cells to a centralized POI is a manual task, and also judgmental. We tried to choose the POIs to be close to a real POI and of course identified all possible POIs close to the main roads to be able to track movements on the roads. Using this process we reduced the number of points on the map from 14K to about 500 point which is more manageable and easier to locate on the map. The identified POIs and the POI demographics are then loaded in the POI table.

Update POI Column in the Cell and CDR Tables
As a result of grouping cells to POI a mapping table is created. The mapping table is used to populate the POI column in the cell table and by joining the cells table and the CDR's table on the cell id column we can update POI column in the CDR table.

Identify Roads Using POI
Roads are part of Regions of Interest that are defined by the business users. For the purpose of the study some roads are selected to capture traffic volume and study movement patterns. The selected roads are stored in the road table. Each road is identified by POIs close to the road and a mapping table that relates a road to a group of POIs is used to populate the RoutePOI table.

Identify Commuters
We mean by commuters, the users of the mobile devices who generate voice and data CDRs. To maintain the privacy of the subscribers we only have anonymous identification number (user id) for the subscribers in the CDR table. We can use user id to select the CDRs made by each user (subscriber) and use it for the purpose of this research. To populate the commuter table we select the distinct users from the CDR table and load them in the commuter table. The number of commuters (users) identified is 13,887,256 which is a considerable number of commuters, relative to greater Cairo population, that can fairly represent greater Cairo movement behavior.

5.4 Data Preprocessing

The following are the algorithms used to extract user trajectories and the Analytical Data Set ADS that will be used for the analytics part.

Trajectory Extraction
In this task we will extract the movement trajectories on specific roads from the CDRs. In this task the movement of the commuters (the required trajectories) are inferred from the CDR's generated on the specified roads. The results are stored, as shown in the logical model, in two tables, "Commuter Road" and "CR Trajectory". The "Commuter Road" stores all the trajectory a specific user has done on any road. While, the details or the stops on the road are stored in the "CR Trajectory" table.

A state diagram that explains the methodology used to detect trajectories by following the usage of a specific user in a specific period is shown in Fig. 7. State diagram for detecting movements from CDRs, followed by the algorithm used to populate the two tables. The CDR table is sorted by user_id and time stamp as the algorithm is built to detect the trajectories user by user.

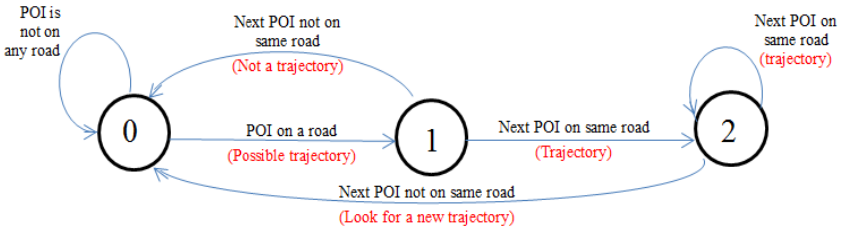

Fig. 7. State diagram for detecting movements from CDRs

```
Start
Position pointer to first row in CDR table
Loop1: Do while not EOF
  SET Luser_id = user_id;
  SET first_row_of_the_user = T;
  Loop2: Do while user_id = Luser_id
    IF First_row_of_the_user = T THEN
      SET LRoute_Id = Route_Id, store 1st row values;
      SET First_row_of_the_user = F;
      SET State = 1;
    Else
      IF LRoute_Id = Route_Id THEN
        IF state = 1
          INSERT new row in CR_Traj using 1st row values
          SET State = 2;
        End IF
        INSERT new row in CR_Traj using this row values
        Store this row attribute values as last row
      Else
        IF State = 2 THEN
          INSERT a new row in Com_Rd using first and last
        END IF;
        SET LRoute_Id = Route_Id;
        SET State = 1;
      END IF;
    END IF;
    Skip to next row
  END Loop2;
END Loop1;
```

Build ADS

To study the pattern of movement, we aggregate the trajectories per route per day of week per hour of the day. The result is the Analytical Data Set ADS used for performing the analytics in this area. {Route_id, CR_Start_DOW, CR_Start_Hour, CR_Rt_DOW_Hr_Cnt}.

5.5 Performing Analytics and Sample of the Case Study Results

Traffic Pattern

Identifying traffic patterns could be the most valuable information for traffic management; we understand that traffic pattern varies (day and night, hour of day, day of week, summer and winter, vacation times, etc.) The analysis engine should be able to support analysis over different time dimensions. For example, the day pattern is of importance to be known as in the following business questions:

BQ#1 - The night pattern (after 8PM, to 6AM); per road, per day
BQ#2 - The daytime pattern, per road, per day

The ADS was used to explore the different movement behaviors on a specific road and on all roads. Microstrategy was used to analyze the data and tables and the analysis results and graphs shown in shown in Fig. 8. Traffic pattern - all selected roads, Table 3. Daytime movement (6am – 8pm), and Table 4. Night movement (8pm – 6am). Again, busiest road is highlighted in yellow in the road row; while busiest day is highlighted in yellow in the totals row.

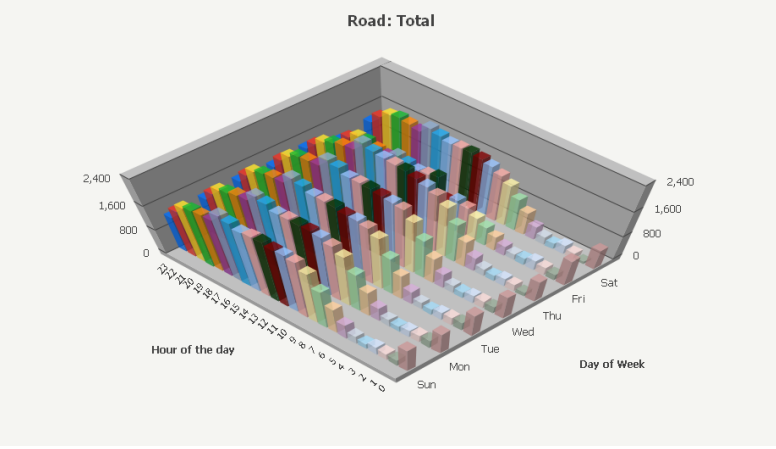

Fig. 8. Traffic pattern - all selected roads

Table 3. Daytime movement (6am – 8pm)

Road Name	Sun	Mon	Tue	Wed	Thu	Fri	Sat
26th of July axis	2,029	2,053	2,052	2,059	2,034	1,310	1,689
6th of October bridge	2,241	2,386	2,368	2,439	2,293	1,133	1,740
Salah Salem	2,199	2,237	2,225	2,241	2,161	1,168	1,656
Nasr road	3,525	3,358	3,396	3,374	3,349	1,931	2,868
Ring road – north	3,609	3,523	3,493	3,604	3,695	2,599	3,099
Ring road – south	5,170	5,115	5,128	5,114	5,125	3,052	4,197
Suez road	818	823	831	837	818	474	647
Ismailia road	1,770	1,880	1,822	1,855	1,776	1,140	1,514
Cairo-Alex Agg rd	2,506	2,556	2,522	2,556	2,640	1,558	1,999
Cairo-Alex Desert rd	1,151	1,169	1,178	1,211	1,265	657	830
	25,018	25,100	25,015	25,290	25,156	15,022	20,239

Table 4. Night movement (8pm – 6am)

Road Name	Sun	Mon	Tue	Wed	Thu	Fri	Sat
26th of July axis	614	614	660	630	617	599	619
6th of October bridge	616	657	675	674	635	566	601
Salah Salem	760	764	778	773	754	694	703
Nasr road	1,305	1,303	1,353	1,296	1,299	1,158	1,184
Ring road – north	1,476	1,483	1,522	1,518	1,458	1,279	1,333
Ring road – south	1,850	1,852	1,908	1,860	1,805	1,645	1,683
Suez road	239	244	240	232	219	233	217
Ismailia road	538	538	540	527	509	455	462
Cairo-Alex Agg rd	888	872	907	912	905	747	799
Cairo-Alex Desert rd	288	281	300	300	316	265	264
	8,574	8,608	8,883	8,722	8,517	7,641	7,865

Ranking Roads Using Clustering

Further analysis was done to cluster the movement pattern on the road using TWM data mining tool, the result of clustering the ADS to 3 clusters is listed in Table 5. Traffic pattern – K-mean clustering. The number of clusters was chosen heuristically as per common perception to traffic it is low, medium, or high. Indeed different other techniques could also be used. Additionally, different interpretations could be generated. The graph in Fig. 9. Traffic pattern cluster mean, shows the cluster means.

Clustering results: (Clustering column: Road DOW Hour average)

Table 5. Traffic pattern – K-mean clustering

Cluster ID	Weight	Mean	Variance
1	0.49	45.56	793.79
2	0.37	164.32	1299.34
3	0.15	326.72	3895.40

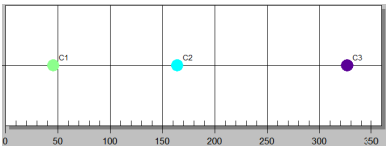

Fig. 9. Traffic pattern cluster mean

Road Rank

Using the clustering model to score road average traffic volume per day of week and hour of day we get the results for each road. By aggregating road scored hours to the cluster level we get the results as shown in Table 6. Roads average volume of traffic per day of week and hour of the day in each cluster, which can be used to rank the road based on the number of hours in each cluster. The road rank is illustrated using the colors green, yellow, and red.

Table 6. Roads average volume of traffic per day of week and hour of the day in each cluster

Road Name	Number of hours per cluster		
	1	2	3
26th of July axis	71	97	
6th of October bridge	71	97	
Salah Salem	68	100	
Nasr road	49	50	69
Ring road – north	43	57	68
Ring road – south	43	19	106
Suez road	168		
Ismailia road	93	75	
Cairo-Alex Agg rd	61	104	3
Cairo-Alex Desert rd	151	17	

6 Conclusion

In this research, we have proposed a framework, which would enable the analysis of trajectory data using data mining techniques. While mainstream literature focus on finding this knowledge based on GPS data, we have been able to show how this could be achieved based on GSM data. Traffic management decision makers could use our framework to make related decision e.g., traffic volume on specific roads, and traffic pattern. Our analyses confirm that long-term GSM activity data is well suited to identify typical movement patterns done by communities especially when GPS data is not available. In addition, we believe our methods explained how that estimation of movement quantities from GSM activity data is possible.

Using real data in the case study was definitely for the benefit of the research; however, the data size was a major obstacle that caused the research to halt several times before Teradata granted the use of one of its servers to the research. This is very important to mention as GSM CDRs are always huge in volume and for deployment the required volume of data will be much more as data from all operators should be integrated to provide complete view for whole population.

CDRs were used to extract commuter's trajectories. However, CDRs are not generated unless the commuters make network activity (call, SMS, data, etc.). In other words, not all trajectories of the commuters traveling on the roads are extracted, only trajectories for commuters that used their devices while traveling. Moreover, the commuter should make more than one network activity on the road with different cell id to be considered as a traveler in order to distinguish the moving commuter from a resident who lives close to the road. To overcome this problem, other GSM generated records can be used such as the cell registration record that is generated automatically every time a user moves between cells; however, the data size is much more than the usage CDRs. We need to take into consideration that the use of data services is increasing as a trend while voice calls are decreasing and data services generate a lot of CDRs without the user intervention (e.g. check mails and messages, perform software updates, etc.) which makes data CDRs a viable alternative in the near future. In all cases, we believe this area still needs more investigation to enhance the intelligence of the trajectory extraction algorithm.

The complete deployment of the framework by traffic department or concerned government agencies, data from the three mobile networks operating in Egypt should be integrated to get movement pattern of the whole population.

Future work includes the full deployment of the framework and its application in different business domains. Also, comparative study between GSM-based analysis versus GPS-based.

References

[1] Anagnostopoulos, T., Anagnostopoulos, C., Hadjiefthymiades, S., Kyriakakos, M., Kalousis, A.: Predicting the Location of Mobile Users: A Machine Learning Approach. In: ICPS 2009, London (2009)

[2] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F.: Trajectory Pattern Mining. In: KDD 2007, California (2007)
[3] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F.: Trajectory Pattern Analysis for Urban Traffic. In: IWCTS 2009, Seattle (2009)
[4] Yan, Z., Parent, C., Spaccapietra, S., Chakraborty, D.: A Hybrid Model and Computing Platform for Spatio-semantic Trajectories. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 60–75. Springer, Heidelberg (2010)
[5] Yang, J., Hu, M.: TrajPattern: Mining Sequential Patterns from Imprecise Trajectories of Mobile Objects. In: Ioannidis, Y., et al. (eds.) EDBT 2006. LNCS, vol. 3896, pp. 664–681. Springer, Heidelberg (2006)
[6] Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y.: Mining Interesting Locations and Travel Sequences from GPS Trajectories. In: WWW 2009, Madrid (2009)
[7] Alvares, L.O., Bogorny, V., de Macedo, J.A.F., Moelans, B., Spaccapietra, S.: Dynamic Modeling of Trajectory Patterns using Data Mining and Reverse Engineering. In: ER 2007, Aukland (2007)
[8] Li, Z., Ji, M., Lee, J.-G., Tang, L.-A., Yu, Y., Han, J., Kays, R.: MoveMine: Mining Moving Object Databases. In: Proceedings of the ACM SIGMOD Conference, Indianapolis (2010)
[9] Leonardi, L., Orlando, S., Raffaetà, A., Roncato, A., Silvestri, C.: Frequent Spatio-Temporal Patterns in Trajectory Data Warehouses. In: SAC 2009, Honolulu (2009)
[10] Marketos, G.: Mobility Data Warehousing and Mining. In: VLDB 2009, Lyon (2009)
[11] Oueslati, W., Akaichi, J.: Mobile Information Collectors Trajectory Data Warehouse Design. International Journal of Managing Information Technology (IJMIT) 2010 (2010)
[12] Alvares, L.O., Bogorny, V., Kuijpers, B., Fernandes, J.A., Moelans, B., Vaisman, A.: A Model for Enriching Trajectories with Semantic Geographical Information. In: GIS 2007, Seattle (2007)
[13] Brakatsoulas, S., Pfoser, D., Tryfona, N.: Modeling, Storing and Mining Moving Object Databases. In: International Database Engineering and Applications Symposium (IDEAS 2004), Coimbra (2004)
[14] Gidófalvi, G., Huang, X., Pedersen, T.B.: Privacy–Preserving Trajectory Collection. In: ACM GIS 2008, Irvine (2008)
[15] Mohammed, N., Fung, B.C.M., Debbabi, M.: Walking in the Crowd: Anonymizing Trajectory Data for Pattern Analysis. In: CIKM 2009, Hong Kong (2009)
[16] Schreck, T., Bernard, J., Tekusova, T., Kohlhammer, J.: Visual Cluster Analysis of Trajectory Data With Interactive Kohonen Maps. In: VAST 2008, Columbus (2008)
[17] Jin, X., Itmi, M., Abdulrab, H.: A Cooperative Multi-agent System Simulation Model for Urban Traffic Intelligent Control. In: SCSC 2007 (2007)
[18] Tártaro, M.L., Wainer, G.: Defining Models of Urban Traffic Using the TSC Tool. In: Proceedings of the 2001 Winter Simulation Conference (2001)
[19] Behnisch, M., Ultsch, A.: Urban data-mining: spatiotemporal exploration of multidimensional data. Building Research & Information, 520–532 (2009)