Research and Application of Network User Behavior Data Mining under the Background of Big Data

Qian Cheng (corresponding author*), Zeng Xiaowen, Zhou Yanying

Water Resources and Electric Engineering
Guangdong Technical College
Guangzhou, Guangdong, China

Abstract — **In this paper, the author studies the application of network user behavior data mining under the background of big data. In this research work we focused mainly on the precise mobility profile building thorough trajectory and behavioral pattern mining using the GSM CGI Cell-ID, where all the concerned issues like precise spatial extraction, stay points detection and mobility profiling are addressed properly thorough the proposed framework. The proposed framework utilized both spatial term and semantic information for mobility profile building which is not addressed in any of previous related work, so makes it suitable for any LBS due to of its novel and generic nature.**

Keywords - application; network user behabior; data mining; big data; cloudsim.

I. Introduction

The smart devices which include smart phones, ipad, ipods, PDAs became the vital part of today's human life and this trend of adoption is much faster than any other in history. While the growth of mobile usage is quite rapid as there are 4.5 Million of mobile users by 2010 which is as much as double of PC users worldwide [1-2]. So almost every of the company Google, Nokia, Apple, Samsung launched their products in the market with OS X, Android, Windows and Symbian operating system with an open API's for development. On the other hand these smart phones also opened a door of available sensors throughout the network spread over geographical area which can be widely used for location based services [3, 4, 5, 6, 7].

This location data retrieved from smart phones can play a vital role in determining the user trend. As mobile phone is connected with the base station where each base station represents the cell where user is residing in at a particular time it record the spatiotemporal trend of user automatically even without disturbing the user routine. So we can apply the data mining techniques on such kind of data to extract meaningful information. As this location information can provide the list of significant locations for user during mobility this can be used for Location based services (LBS) [8].

The location information and mobility path can be used for the potential applications which include mobile advertisement, city wide route sensing, region pollution, traffic safety management, social networking, potential warning system, soured analysis, route tracking expand and communication. In these applications the low level mobility data is unity recommendations interoperated into high level information in term of stay locations, patterns and finally user profiling.

In our work we are focused on the precise extraction of mobility profile against user mobility so that it can be rendered for any location based service. As described before this mobility profiling is all location based where location is logged in by the user using different methods i.e. Indoor and Outdoor.

There are many ways to record the user mobility which can be Wi-Fi, Bluetooth, Infrared, GPS and GSM depending on the situation and type of intended application. Most important work regarding the location extraction based on algorithms is done in their work where they formally defined the term mobility mining to extract patterns through profiling. While the current mobility trends are studied in detail by their work, the mobility is defined as key prediction indicator of human life.

So the tracking of true location is basis of every mobility based application. As there are two kind of technical solutions available for the location recording indoor and outdoor. In case of indoor like Bluetooth, RFID or infrared, these short range and cannot compete the outdoor like GSM and GPS which are categorized broadly in their work. On the other hand Wifi is another solution to location tracking as well were geo location of the user is determined by the terminal it is connected with. As per the feasibility and their wide usage outdoor technologies are widely used for location tracking which includes GPS, Assisted faux GPS and GSM. Where GPS is coordinated based which provides the exact location of the user in term of longitude and latitude. But GPS needs long start-up time on device, high consumption of energy which is discouraging for the user. And most importantly there are many applications where exact location of user does not matter and application can use the relative position of the user for prediction of trends where GSM can serve the purpose well enough.

The GSM location positioning system is known as landmarked-based which means it is not coordinate based. There are different methods available for tracking of location even in GSM which includes Assisted-GPS (A-GPS), Time difference of arrival (TDOA) and Enhanced observed time difference (EOTD) but all of these methods need extra hardware equipment installation in the network to work within the performance. Secondly network operator use this kind of hardware installation or backbone information for its own network management purposes which is generally hidden from the public use.

So the Cell global identity (CGI) of GSM network is only viable and readily piece of information for location tracking. Where no extra hardware installation required and there is no need to bother user as it is tracked automatically as soon as user connects with the network. The CGI information is continuously changed as the sure moves in the network from one cell to another of the network. This CGI information is four header i.e. Mobile country code (MCC) assigned to every country, Mobile network code (MNC) assigned to every operator, Location area code (LAC) created by operator for identification of Cells, Cell ID which is given by the Cell to each user connected. So MCC, MNC, LAC, Cell Id provide the unique set of identification for location extraction of any user at particular time span. As shown in their work cell information data is enough for most of the applications where it can be used for location prediction of mobility quite evidently. On the other hand Google also shown its full support towards the cell based location prediction and tracking, which shows that Cell data is valuable and readily available for all kind of location aware services now a days.

So in our work we will use the CGI information for the determination of location against user mobility which is rather in expensive way to track the user anytime during data recording. And this makes our work feasible for any kind of location aware application that just use location based profile without worry of expensive location tracking device installations during mobility data recording.

II. OVERVIEW OF GSM NETWORK ARCHITECTURE

GSM Topology varied from operator to operator which is kept hidden from the user and it’s generally unavailable for public use. This topology information is used by the operator for the arrangement of different cells in the GSM network.

However basic network architecture of GSM network is same for every network operator as shown in Figure 1. In GSM network Base Transceiver Station (BTS) is basic unit which defines its own cell defined by Location area code where multiple user Cell-Ids are allotted to users connected to network within this cell.

As shown in Figure 1 Mobile Station (MS) is linked with the Base Station Controller (BSC) and Mobile Services Switching Centre (MSC). The MSC acts as a communication bridging between multiple MSCs and BSCs to serve as a channel between two different cells belongs to different MSCs. Radio cells in the network are determined by the number of BTS where different BTS connected to a MSC have same Location Area (LA). While more than one MSC together serve as a communication mean and make geographic region identified by unique code called Location Area Code (LAC) as shown in Figure 2.Where conceptually each cell has a hexagonal shape.

Each operator has different setup topology for BSCs connected with MSC; result in different number of Cell IDs in same LAC for each operator. Intact user density determines the LAC size in urban areas LAC has few Cell IDs while in case of normal areas its much higher. If the number of users connected with MSC goes beyond its capacity this will make a hand shake with new MSC and handover upcoming users. So any GSM cell is uniquely identified by Cell ID in combination with LAC and Country Code. However LAC code is most important for the identification of geo-location in GSM network.

The one of main character of any GSM network is that cells in it are not arranged as polygonal shape in fact they are in overlapped and have bubble shape because of coverage and tradeoffs issues. This overlapping is quite high in urban areas where mobile towers are in vicinities due to of high number of users in area Whip in case of rural areas its very low where one cell can even cover 35Km which limitation in GSM network due to of lesser number of users. This overlapping is the basis of one important phenomenon which is called cell oscillation where user is assigned multiple Cell-Ids even at a stationary state, so while in analysis it appears that user is moving and it can lead to inappropriate mobility profiling.

This L AC is organized by the operator in such a way that it can serve a purpose of every network service like messaging, GPRS, Value added services so it depends on number of user's density in particular area. It is obvious that this LAC overlapping is quite often in urban area where numbers of users are very high especially in center of cities, shopping malls, schools or business centers for load balancing and quality of services, while on there this overlap other hand in case of urban areas where users are not dense is less likely and one single LAC can serve up to 35Km.

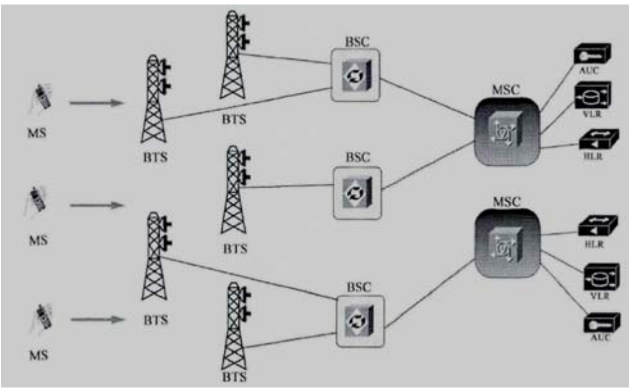

Figure 1. The model and architecture.

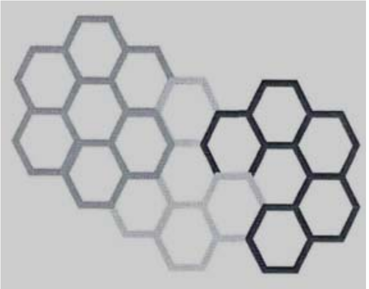

Figure 2. The network topology.

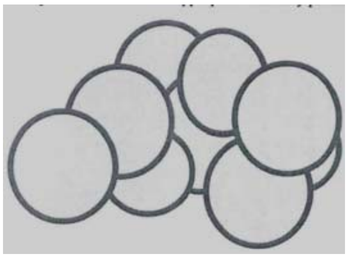

Figure 3. The network topology: Actual bubble shape.

III. DATA MINING ALGORITHM

The main effect of data mining algorithm is to extract the required information from a large amount of data, including the structural data, semi-structured, and unstructured data sources, such as audio, video, data, for data algorithm. This algorithm must have model, and first search algorithm. Currently the common data mining algorithms are mainly the decision tree method, bionic global optimization of genetic algorithm and neural network, statistical analysis and row exclusive counterexample method, etc. In order to improve the effect of data mining process effectively, a detailed research should be applied to the cloud computing method. In this way. more effective implicit knowledge in the mass data information can be discovered in order to improve the effect of application of information data.

Association rules found a relationship between things and other transactions or interdependence. Assuming that I={i1,i2,…im} is a collection and the related data task D is a collection of database transactions, in which each transaction T is a collection, and making $T \subseteq I$. Every transaction has an identifier TD . Assuming that A is a set of items, $A \subseteq T$. Association rules are the containing type of $A \Rightarrow B$, among them, $A \subset I$, $B \subset I$ and $A \cap B = \Phi$. The rules $A \Rightarrow B$ is in the transaction which sets up with support s , s is the percentage for the transaction contains $A \cup B$ in the D.

The LIPI algorithm through scanning data sets of frequency, then finding relevant data and finishing dig, its principle as follows:

Assuming that I={i1,i2,…in} is a collection, which composed of different characteristics, the characteristics set is not an empty set, but is a subset of the set of I, which can be expressed as(x1x2…xm) , every xk is a term.

The sample variance and sample proportion of variance have established the following relationships:

$$S^{*2} = \frac{S^2}{\mu^2} \quad (1)$$

Proof: by definition 4

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\alpha_i - \bar{\mu}}{\bar{\mu}^2}\right)^2$$

$$= \frac{1}{(n-1)\bar{\mu}^2}\sum_{i=1}^{n}(\alpha_i - \bar{\mu})^2 \quad (2)$$

$$= \frac{1}{\bar{\mu}^2}S^2$$

The algorithm is to propose the original data processing for many times, then use the effective information contained in the original data.

The study of the data mining algorithm has great significance in improving the effect of data processing. User data information needs to be extracted in the huge amounts of data, in order to promote the development of various fields. Cloud computing is a relatively new computing mode, its application in data mining also needs to do some further researches on the existing basis, continuously improving its application efficiency and improving data information of the application efficiency.

The data mainly displays in some aspects, such as by the department of statistics data format, which does not have a uniform requirements. When using VF or SQL or some using TEXT or a variety of other formats, each system has general check, data format, and compatibility.

Complicated statistical data have originated from the enterprises and institutions' direct submitting, or the result of the whole system of internal between different departments. For the lack of effective statistical data storage and management of professional means, it leads to the deep processing of statistical data.

It is known that the statistics business involves all aspects of society including index the great amount of data. Although there are abundant data resources, yet the lack of professional analysis of statistical data at a deeper level and lack of refining and mining tools, leads to the disunion between a large number of accurate data and resource usage and people's growing demand of statistical information.

The decision of government and enterprises and the current statistical work has been simply filled in form, and then submitted, but it is lack of effective means for the subsequent development. The data mining of data warehouse technology can solve the above mentioned problems effectively mainly because it has the following obvious advantages:

(1) Based on data warehouse algorithm, the data time-consuming preconditioning in data mining can be solved. Through the establishment of data warehouse, it can avoid

data extraction, cleaning, conversion and loading process every time.

(2) Another feature of the data warehouse is to store data through subject organization, which provides convenience for data mining to choose the appropriate data source. According to different areas, data warehouse is divided into the national economic statistics, social statistics analysis, and enterprise survey, etc. The national economy includes consumption statistics, labor statistics, people's living standard, and statistics, etc.

(3)The data is collected by statistics departments at all levels, all existing in different types of database such as EXCEL, Fox Pro, etc. Because historical data cannot be stored in the database, many knowledge can't be excavated in the mining database, such as forecasting and application; on the other hand, data warehouse storage management can get the data from the PLTP system and the history of offline business data and the external data sources of heterogeneous distributed, thus it's good for the heterogeneous data and source data to summarize in order to finish the more efficient usage. Except for the requirements of data mining and data warehouse environment, data mining needs to be based on data cube environment and data warehouse technology and hence meet the demands of data mining technology. Therefore, it is necessary for data mining and data warehouse to work together. On the one hand, data mining technology has become a very important application of data warehouse and a relatively independent tool; On the other hand, the data mining technology is an important step of catering the process of data mining, and it improves the efficiency of data mining and ability, and ensures that data mining have the extensiveness and completeness of data source.

TABLE I. S EQUENCE DATABASE

Sid	Sequence
S1	ACAABC
S2	ACABCB
S3	ACABC
S4	ABCAB

TABLE II. CA SEQUENCE DATABASE

Sid	Sequence
S1	ABC
S2	BCB
S3	BC
S4	B

IV. RESULTS AND DISCUSSION

We have implemented the proposed methodology on the dataset which is spatial outlier free. Our methodology is based on overlapping area so we plotted the cell appearance over different locations which show the bubble effect in GSM network as shown in Figure 4. And also shown in Figure 5 where one semantic location is identified by multiple cells and one cell can appear at multiple semantic locations.

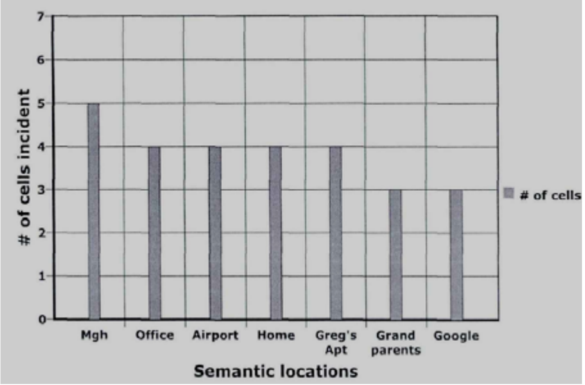

Figure 4. The data mining result on network topology.

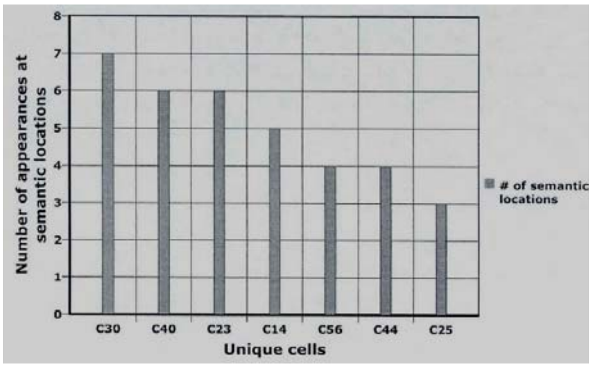

Figure 5. The users' behavior analysis data.

V. CONCLUSIONS

In this paper, the author studies the application of network user behavior data mining under the background of big data. In this research work we focused mainly on the precise mobility profile building thorough trajectory and behavioral pattern mining using the GSM CGI Cell-ID, where all the concerned issues like precise spatial extraction, stay points detection and mobility profiling are addressed properly thorough the proposed framework. The CGI information is continuously changed as the sure moves in the network from one cell to another of the network. This CGI information is four header i.e. Mobile country code (MCC) assigned to every country, Mobile network code (MNC) assigned to every operator, Location area code (LAC) created by operator for identification of Cells, Cell ID which is given by the Cell to each user connected. So MCC, MNC, LAC, Cell Id provide the unique set of identification for location extraction of any user at particular time span.

As shown in their work cell information data is enough for most of the applications where it can be used for location prediction of mobility quite evidently. On the other hand Google also shown its full support towards the cell based location prediction and tracking, which shows that Cell data is valuable and readily available for all kind of location aware services now a days. The proposed framework utilized both spatial term and semantic information for mobility profile building which is not addressed in any of previous related work, so makes it suitable for any LBS due to of its novel and generic nature.

ACKNOWLEDGEMENTS

This work is supported by the Key Project of College "innovation and strong school project" independent innovation ability to enhance the class of natural science" (No. CXZ10).

REFERENCES

[1] D. McDonagha, A. Brusebergb, C. Haslamc. Visual product evaluation: exploring users' emotional relationships with products. Applied Ergonomics, 2002, 33, p.p.231-240.

[2] Kaikai Chi, Yi-hua Zhu, Xiaohong Jiang, Xianzhong Tian. Practical throughput analysis for two-hop wireless network coding. Computer Networks, 2013, pp. 233-256.

[3] Luiz Filipe M. Vieira, Mario Gerla, Archan Misra. Fundamental limits on end-to-end throughput of network coding in multi-rate and multicast wireless networks. Computer Networks, 2013, pp. 5717-5727.

[4] Yang Xu, Xiao yao Xie, Huan guo Zhang. Modeling and Analysis of Electronic Commerce Protocols Using Colored Petri Nets. Journal of Software, 2011, pp. 67-78.

[5] D.E. Leidner. Virtual partnerships in support of electronic commerce: the case of TCIS. Journal of Strategic Information Systems, 1999, pp. 81-93.

[6] Jiali Yun, Liping Jing, Jian Yu et al. A multi-layer text classification framework based on two-level representation model. Expert Systems with Application, 2012, 39(2), pp. 2035-2046.

[7] Changxing Shang, Min Li, Shengzhong Feng, Qingshan Jiang, Jianping Fan. Feature selection via maximizing global information gain for text classification. Knowledge-Based Systems, 2013, pp. 54-68.

[1] D. McDonagha, A. Brusebergb, C. Haslamc. Visual product evaluation: exploring users' emotional relationships with products. Applied Ergonomics, 2002, 33, p.p.231-240.