Mobile Location Prediction: Pervasive study with NoSQL Database

Mira H Gohil
Assistant Professor
Thakur Institute of Management Studies, Career Development and Research (TIMSCDR)
Mumbai, Maharastra, India
miragohil260178@gmail.com

S. V. Patel
Professor
Sarvajanik College of Engineering & Technology, Surat, Gujarat, India.
patelsv@gmail.com

***Abstract*— Mobile phones have inherent mechanisms that facilitate to record mobility of users. If accurate mobility data is available, than it can be helpful to understand user mobility patterns and can help in user location prediction. Location-awareness is therefore a major component in context-awareness, which in turn, enables our devices to become our hidden supporters. The location prediction is attracting researchers due to heavy utilization of cell phones having GPS trackers.**

Numerous algorithms have been used in location prediction study however, in this paper we present anytime location prediction model (ALPM) to find mobility patterns with the use of graph based data model and taking into account the user past history of mobility and the current timestamp and location.

Keywords- Mobile location prediction, NoSQL, Neo4j, Location prophecy, Next location prediction, Graph database.

I. INTRODUCTION

Mobile communication is growing at an unprecedented growth; almost half of earth's population now uses mobile communications. According to the International Telecommunications Union – Globally, mobile cellular subscriptions correspond to a penetration rate of 97% and mobile broadband penetration will reach 47% in 2015. [1]

When we get the user's mobility data the biggest problem is the use of database structure. Which database structure we use to get the work done easily with better results?

Although relational database have evolved significantly, using them in this use case may not be beneficial. Firstly, relational database do not store relationship values which are of utmost importance in a connected data. Secondly, as relational database models are not flexible, they should be avoided from updating in later stage of process. Thirdly as Query structure in relational database often involves complicated joins to get the output which are considerably less costly in Graph databases.

Researchers in [7][8][9][10][11] mention that mobile location prediction is the problem of sequence pattern mining. Pattern mining is useful in different field like for customer behavior prediction, for mobile commerce, for web pattern mining. In mobile location prediction, user movement data is again a sequence of transaction from one cell to another cell. Every day in the life, user performs some similar pattern in the transaction e.g. working person spend most of the time in the office place and follow similar route in travelling. In holidays, people follow some similar patterns most of the time. So pattern is always there in the user movement transaction. Many algorithms proposed are based on mining the mobility patterns of users, forming mobility rules from the patterns, and finally predicting a mobile user's next movements by using the mobility rules. Finding mobility pattern from mobility data is a problem that involves processing of very large data. Here, we propose the use of graph database as it offers many benefits as compared to other databases.

This paper uses community edition of Neo4J which is a graph database application to find users' home location, working location and given time location from the places visited historically.

The rest of the paper is organized as follows: The Section II explains the rational for using graph database instead of RDBMS. Section III briefly states the methodology Section IV presents the data collection and cleaning Section V and VI describe data visualization and analysis respectively. Section VII presents the result analysis based on the analyzed data and Section VIII draws concluding remarks and further work.

II. DATABASE SELECTION

Relational database store facts and are well suited for structured and tabular data. However when we need to find relationships between these facts, relational databases involves several joins in query and may result in join-bombs explained below. This significantly affects performance.

Relational databases do not store information about relationships hence they are not productive for connected datasets.

Graph databases stand out in such scenario where we have connected dataset. They not only store data elements and their relationship but can also store degree, weight or significance of that relationship.

A. Join Bomb Pronblem

Relational databases are designed to handle tabular data and can handle large datasets. But due to several requirements like Normalization in order to avoid data redundancy and maintaining unique data in tables. Data islands are created.
Hence in every case, in order to get results we need joins between different tables. The issues with joins are that we are not aware of the intermediate latency or memory required to process each Join. And with multiple Join queries in large data sets ultimately deteriorates performance. Hence larger dataset leads to increased processing time. [2]

B. Performance Issues

Below table exhibits the execution time as experimentally shown by Partner and Vukotic. [3] They used both relational database and Neo4J to conduct an experiment for social network containing 1,000,000 people each having about 50 friends. Experiment was to find extended friends like friends-of-friends at degree 2 and so forth. Until relationship depth 2 both database types perform marginally close to each other. However after depth three, the difference in execution time considerably increases. At depth four and five, relational databases become too slow.
In contrast, Neo4j's execution time increases by fraction of seconds. Thus Neo4j outperforms relational databases in use case of connected datasets.

Depth	RDBMS execution time (s)	Neo4j execution time (s)	Records returned
2	0.016	0.01	~2500
3	30.267	0.168	~110,000
4	1543.505	1.359	~600,000
5	Unfinished	2.132	~800,000

Fig 1: Performance comparison of relational and graph database [3]

III. METHODOLOGY

The history data of the user through which we can predict the user location is very essential to collect but at the same time privacy and security are also equally important. So the data can be collected either by developing an android based system or with the tie up of data center. Data was provided in agreement by Nokia Research Center (NRC) by means of signed Memorandum of understanding (MoU) and agreeing to the terms and conditions. Data obtained for this research was collected by Nokia Research Center Lausanne together with its Swiss research partners (Idiap and EPFL). In this dataset user historical timeline data is available.

The visualization and prediction of the data is most important to understand and implement system. For visualization we used the graph tool community edition of Neo4J (NoSQL graph database) and prediction of the location is done using the cypher query language available with Neo4j package.

IV. DATA COLLECTION & CLEANING

In January 2009, Nokia Research Center Lausanne and its Swiss academic partners Idiap and EPFL started an initiative to create large-scale mobile data research resources. This included the design and implementation of the Lausanne Data Collection Campaign (LDCC), an effort to collect a longitudinal smartphone data set from nearly 200 volunteers in the Lake Geneva region. This data set is a subset of the Lausanne Data Collection Campaign which is collected from 71 volunteer users in Switzerland between October 2009 and February 2011. The average duration of a participant's data is 13.5 months since not all users were active for the entire duration. The population is composed of university students and professionals. Users carried their smartphone as their actual mobile phone, and were asked to charge the phone at least once a day. In summary, the data contains roughly 10 million location points, 1.4 million application usage events (such as opening, closing, minimizing, etc. without including system applications) and 8.8 million non-empty Bluetooth scans [4].

For our task, we deliberately studied visit_10min.csv and visit_20min.csv file patterns with fields as shown in the figure 2 among the massive date set MDC obtained from Nokia research center.

Files utilized from MDC dataset are 1) Visits_10min.csv – 5.82 MB, 2) Visits_20min.csv – 5.05MB 3) Places.csv - 9.36 MB. Dataset contain 80 users data gather over the period of 3 years.

In data cleaning, Timestamp data divided into separate parts viz. Years, Month, Day, and Hours. Each hour was further divided into minutes of 3 groups with gap of 20 minutes and with hour-id (as values1, 2 and 3).

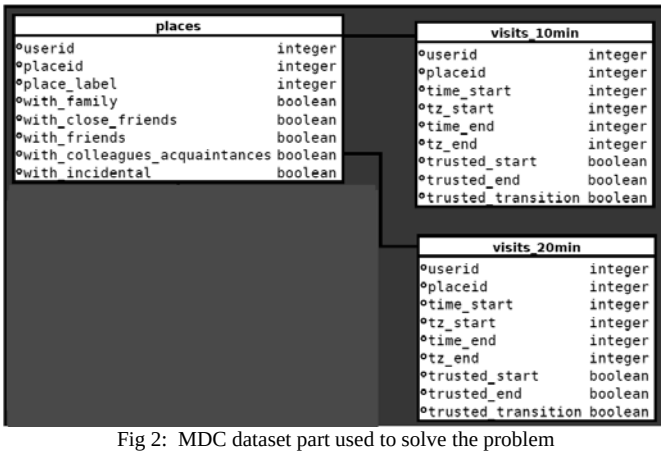

Fig 2: MDC dataset part used to solve the problem

UserId and PlaceId has been converted into Nomenclature values. Values has been assigned for Monday to Sunday (as values 1-7) and referred Holiday or Workday as numerical values 0 or 1 respectively. Regional public holidays have also been considered. [5] and have been marked as holiday (workday=0).

V. DATA VISUALIZATION

Neo4j is a highly scalable, fully transactional ACID, NoSQL (Not only SQL) graph database that stores data structured as graphs, it allows developers to achieve

excellent performance in queries over large, complex graph datasets and at the same time, it is very simple and intuitive to use. It runs over all the common platforms. It can be used and embed inside applications as well [6]. Graph database Neo4j 3.0.3 version has been used and applied the concept of TimeLine Tree Model for visualization as in figure 3.

If we are able to predict the place of the user, it will provide an unprecedented opportunity to develop novel application to provide the user specific and place appropriate content which is also described as context-aware mobile recommendation systems.

As each user travel patterns are different based on user's mobility history which can be termed as Places of Interest (PoI). Although the mobility pattern varies with different population, however it is restricted to limited geographical locations and period to great extent. The main objective of this research is to develop user specific model which learn from user's history data and then if we implement the model in current context, it will predict the user's future locations.

After preparing the architecture of the user 5947, visualization of the user 5947 when all nodes and relationships are matched and returned in graphical manner are as shown in the figure 4.

Analysis using NoSQL Neo4j's Cypher Query Language

- Working location and home location of the user.
- Holiday home location of the user
- Users most significant pathway used for the day to day travelling.
- Possible path between the two places used by the user
- Also we can see the user sequential travelling data with weights. So we can come out with the conclusion of the important places in user mobility pattern

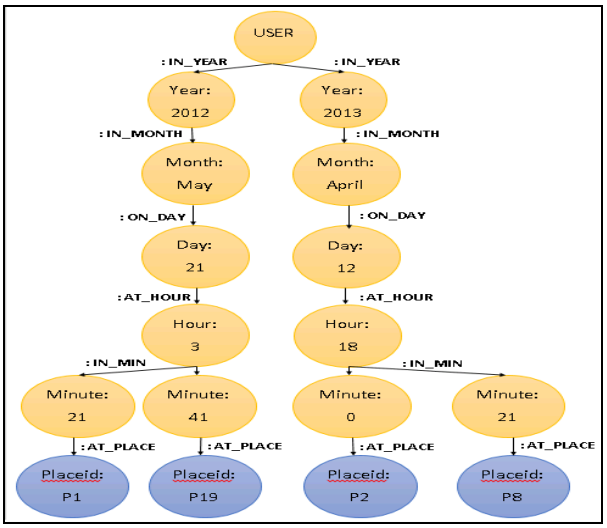

Fig 3: Architecture of Time line data of user places for Neo4J

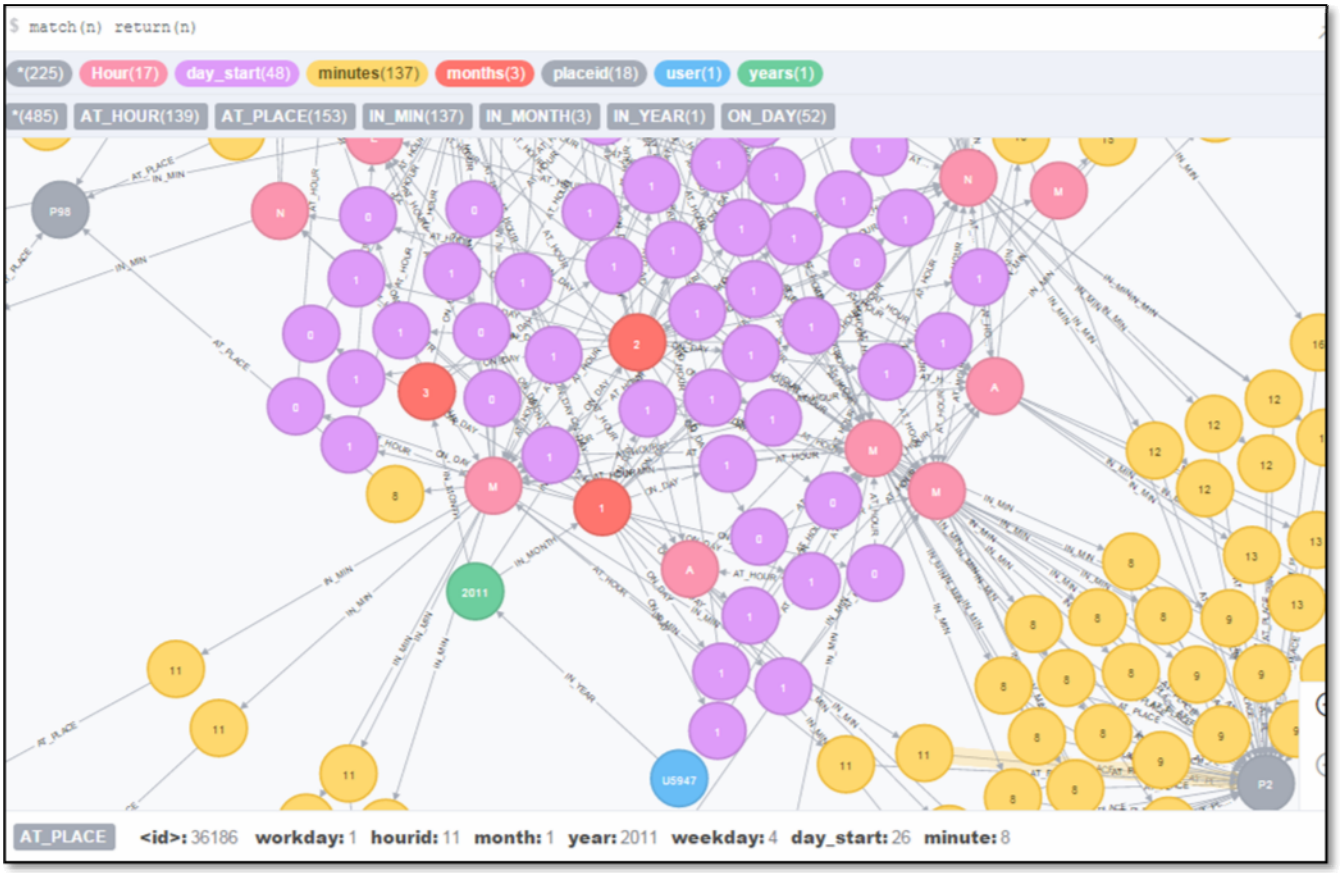

Fig 4: Architectural visualization of user 5947 as sample in Neo4j

We can conclude that the First-best location is a home location. Similarly, the same condition has been used for place visits on days of weekend; this count confirms the home location as users are most likely at home during the weekends for the data collected over the entire year.

In this way we are able to calculate the probability from total frequency of Home location discovered with the total visits of all places gives us the probability for that place.

In the same way, the work location has been found during weekdays by assuming the hours (between 9-12 and between 13-16 hours)

A. Sequential path with weight

To find sequential path and to discover the weight of the path between place, Cypher query has been written to visit all the places visited by the user and assign a weight value (default =0) and increment this value as those places are visited sequentially by the user. As from Figure 5 we were able to find the most visited path between two places on the basis of highest weight and the neo4j code is also there in the figure itself. Below figure is visualization of all the paths visited with their weights.

B. All possible paths between two places.

By using the shortest path function with up to degree of 5, we can find all the possible paths between any two known place nodes from the data of user visited places. This gives us the snapshot of the possible paths the user must have travelled between two places. For e.g. - all places visited between home and work location can be discovered. If these visited locations are given a semantic label, then either same or similar and different places can be proposed in location based advertising systems.

VI. RESULTS ANLYSIS

Using graph we can get the details regarding two important location of the user: Home and Work location. We can find the prominent path between two locations easily. We can also predict the user next location easily if we have the few pervious data pattern and consider each location separately.

Using holiday location pattern data we can predict the users next holiday time and travel companies can offer better plan accordingly. By analyzing user shopping history we can find out next shopping cycle of the user.

Using graph it is very easy to find the similar daily pattern of the users and we can find two user similar route patterns and it is very helpful in pool car which eventually reduce traffic and pollution.

Using 2 years of data, we can find out the profile of the user using user specific model like half day or full day worker, party person, postman or field job person, regular office going person or student.

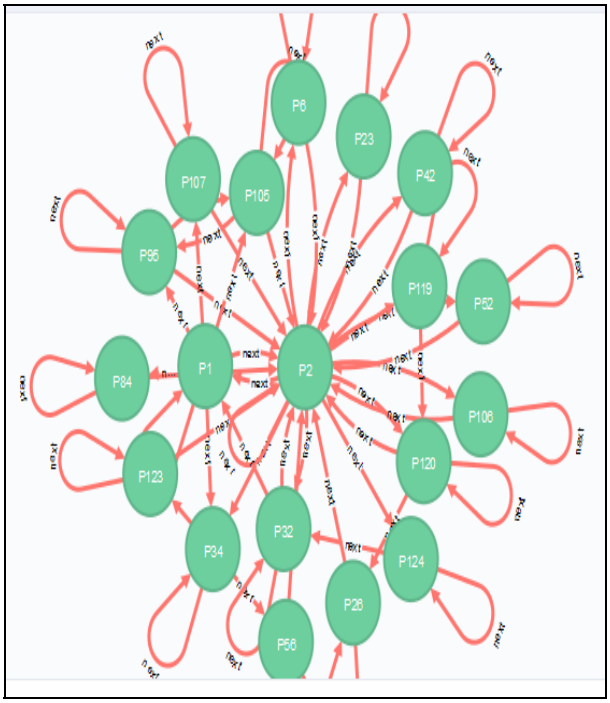

Fig 5: Shows all paths between sequentially visited places

```
match (pi1:placevalue)-[r:next]->(pi2:placevalue)
where pi1.placeid<>pi2.placeid
with  max(r.weight)as m
match (pi1:placevalue)-[r:next]->(pi2:placevalue)
where r.weight =m
return pi1,pi2,r.weight
```

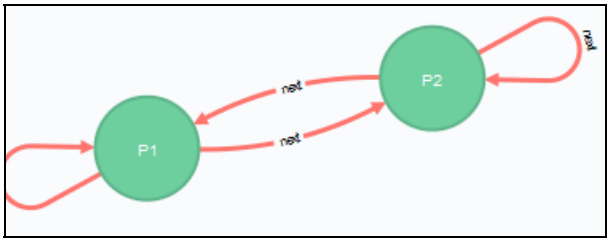

Fig 6: Shows all possible paths between placeid P2 and P1

Analysis of the users mobility prediction:

U5947 – User 5947 is a predictable case. P1 clearly shows to be this person's home location with 81% probability

percentage of being at home during non-working hours and 82% during weekends. This person visits place P2 on weekdays during working hours 9-12 and 13-16 and has a probability percentage of 55% on weekdays. User 5947 seems to be Regular Office-goer due to limited places of visit.

U5964 – This user has also predictable mobility his home location is P1 on weekdays and has 71% probability he will be at home on those hours. And He is at P6 location on Weekends with probability of 57%.During Weekdays, he is most likely to be in P3 during the working hours on weekdays and has probability of 53%.Possibility that this person stays in different home on weekdays and another home on weekends. Most common path travelled is from P3-P4.

U6177 – User 6177 has fixed home P2 with accuracy percentage of 71% and 73% respectively during whole week. And P2 also seems to be the place during weekdays.
This person is not mobile. This resembles a student or a person working from home.

U5927 - User 5927 is an unusual case where the user exhibits P1 place as home location during weekdays and P5 during weekends. During weekdays he seems to be travelling a lot between two places P1 and P2. User is Unpredictable.

U5967* – this user seems predictable as the home location seems to be P1 appears to be his home location on weekdays and weekend And P2 seems to be his work location during weekdays. This user seems to be a student who mostly spends time at home location P1 throughout the week and visits P2 during weekdays. This could either a Part-time worker or student.

U5953 – This User 5953 seems like a regular user, with P3 as his home location on weekdays and visits P1 during working hours of weekdays. However his highest weighted graph is from P1-> P5, which is unusual, hence this user's data needs to be analyzed closely for each week.

U5973 - Also predictable, his home location on weekdays and weekend is P2 with probability of 97% and 91% respectively. So we can say that P2 is definitely his home location.
On weekdays, we come to know, that he is at place P1 during working hours, with prediction probability of 64%

U5928 – **U5966**- Data shows that this user visits place P1 during weekdays and weekends his probability being 95% and 85% respectively. However during the weekdays from hours 9-12, 13-16, we understand that he visits two places P2 and P1 with probability percentage 48% and 34%. This could be a student who visits home and college like place on weekdays and only home on weekends.

U5966* – This user shows similar pattern to U5967, who shows strongly that P1 is this person's home location and on weekdays his visits show his home and another place P2. This could either a Part-time worker or student.

U5925 - This user seems to be working person, he has P1 has his home location and shows good percentage he is at P1 during night on weekdays and weekends. During working hours, he visits P11 a lot. This could be his potential workplace. This User set is highly predictable. This seems like a regular office going person.

Table 1. Query results on top 10 user's data

Queries	Users									
	U5947	U5964	U6177	U5927	U5967	U5953	U5973	U5928	U5966	5925
1) GRAPH PATTERN FOR HOME LOCATION ON WORKING DAY (hours0-4 , 18-23,w=1)	P1	P1	P2	P1	P1	P3	P2	P3	P1	P1
2) DETAIL PLACE DATA ON WORKING DAY (max records,w=1)	P1	P1	P2	P1	P1	P3	P2	P3	P1	P1
3 Probability for home place on working day (hours 0-4, 18-23,w=1)	81%	71%	71%	55%	96%	56%	97%	64%	95%	64%
4) Maximum frequency location on holiday with workday=0 where all holiday included(hours 0-4 , 18-23,w=0)	P1	P6	P2	P5	P1	P3	P2	P7	P1	P1
5) Probability Calculation of Home place on Holiday (hours 0-4,18-23,w=0)	82%	57%	73%	49%	83%	48%	91%	64%	85%	59%
6) Workplace prediction month wise day wise (hours 9-12, 13-16 ,w=1)	P2	P3	P2	P1,P2	P2>P1	P1	P1	P1,P7	P2>P1	P11
7) Place data for working day pattern (hours 9-12, 13-16 ,w=1)	P2	P3	P2	P1,P2	P2>P1	P1	P1	P1,P7	P2>P1	P11
8) Probability for Work place Prediction (hours 9-12, 13-16 ,w=1)	55%	30%	53%	25% , 25%	54% > 31%	53%	64%	25%, 23%	48% > 34%	64%
10) Common path highest weighted graph pattern	P1 -> P2	P3 -> P4	P2 -> P5	P1 -> P5	P1-> P2	P1 = P5	P2- >P1	P3 -> P1	P2 -> P1	P1-> P11

VII. CONCLUSION

From the observations it can be concluded that graph database can efficiently be used on large dataset to perform query operations to find locations like home location, work location; time based location and also discover paths between any two visited places and other analysis.

Future work aims to analyze the correlation between the different users and further it can be extended as social networking analysis.

REFERENCES

[1] Weiser, M.: The Computer for the Twenty-First Century. Scientific American 265, 94-104 (1991).

[2] Neo4j, the World's Leading Graph Database (Neo4j Graph Database) http://www.Neo4j.com

[3] A. Vukotic and N. Watt, Neo4j in action. Shelter Island: Manning, 2014.

[4] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In Proceedings of Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf.. on Pervasive Computing, Newcastle, June 2012.

[5] Region-du-leman.ch, 'Office du Tourisme du Canton de Vaud / Région du Léman (Suisse).' 2015. [Online]. Available: http://www.region-du-leman.ch. [Accessed: 01- Jul- 2015].

[6] Ankur Geol, Neo4J Cookbook, Packt publications, 2015

[7] V. ManoChitra, S. Shanthi, S. Syed Shajahaan, Mining Mobile Sequential Pattern in a Location Aware Environment, IJCSMC, Vol. 2, Issue. 10, October 2013

[8] M. I. Thariq Hussan , Dr. B. Kalaavathi,"Multi-Cluster Based Temporal Mobile Sequential Pattern Mining Using Heuristic Search", WSEAS TRANSACTIONS on COMPUTERS, 2013

[9] Ilkcan Keles, Mert Ozer, I. Hakki Toroslu, Pinar Karagoz, and Salih Ergut, "Location Prediction of Mobile Phone Users using Apriori-based Sequence Mining with Multiple Support Thresholds", AVEA, Istanbul, Turkey, 2014

[10] Vignesh.S 1, Robert.P 2, Vignesh.U 3, Bharathidasan.D4, Rajasekaran," Mining Frequent Patterns and Prediction of User Behavior in Mobile Commerce", International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 3, 2013

[11]Takashi Washio, Hiroshi Motoda, "State of the Art of Graphbased Data Mining", ACM digital library, Volume 5 Issue 1, July 2003

AUTHORS PROFILE

Ms. Mira H Gohil is a research student at the Department of Computing Science, Veer Narmad South Gujarat University (VNSGU). She has 8 years of teaching experience in Computer Networks, Wireless and Mobile Technology. She is currently working as Assistant professor with TIMSCDR (Thakur Institute of Management Studies Career Development and Research), Mumbai, Maharashtra, India. Her current research area is Wireless and Mobile Technology. PH- 8652857218. E-mail: miragohil260178@gmail.com

Dr. S. V Patel is M.E. (Microprocessor System & Applications), Ph.D (Computer Science) and having 35 years of experience in teaching and research. He is Professor associated with Sarvajanik College of Engineering & Technology, Surat, Gujarat, India. His research interests include Software Engineering, wireless sensor networks. Ph-9825740103. Email:patelsv@gmail.com