International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-6, August 2013

Activity Based Data Management in Mobile Environment Using CART and ID3 Data Mining Techniques

Er. Satwant Kaur, Er.Varinderjit Kaur, Er.Gurpreet Singh

Abstract: ***Mobile clients feature increasingly sophisticated wirelessnetworking support that enables real-time information exchange with remote databases. Location-dependent queries, such as determining the proximity of stationary objects (e.g., restaurants and gas stations) are an important class of inquiries. We present a novel approach to support nearestneighbor queries from mobile hosts by leveraging the sharing capabilities of wireless ad-hoc networks. We illustrate how previous query results cached in the local storage of neighboring mobile peers can be leveraged to either fully or partially compute and verify spatial queries at a local host. The feasibility and appeal of our technique is illustrated through extensive simulation results that indicate a considerable reduction of the query load on the remote database. Furthermore, the scalability of our approach is excellent because a higher density of mobile hosts increases its effectiveness. Most users in a mobile environment are moving and accessing wireless services for the activities they are currently engaged in. We propose the idea of complex activity for characterizing the continuously [1] changing complex behavior patterns of mobile users. For the purpose of data management, a complex activity is modeled as a sequence of location movement, service requests, the co-occurrence of location and service, or the interleaving of all above. An activity may be composed of subactivities. Different activities may exhibit dependencies that affect user behaviors. We argue that the complex activity concept provides a more precise, rich, and detail description of user behavioral patterns which are invaluable for data management in mobile environments. Proper exploration of user activities has the potential of providing much higher quality and personalized services to individual user at the right place on the right time.***

Keywords: mobile environments, CART, ID3, proactive data management, prefetching, pushing

I. INTRODUCTION

Location-based queries are of interest in a growing number of applications and an important sub-class of such queries are nearest neighbor (NN) searches. Increasingly such queries are issued from mobile clients and there exist several algorithms that allow the efficient execution of NN queries on centralized databases. In this study we propose an approach that leverages short-range, ad-hoc networks to share information in a peer-to-peer (P2P) manner among mobile clients to answer location-based nearest neighbor queries. Such a P2P approach can be very valuable for applications where access to the server is not always guaranteed and may be spurious at times. For example, during a natural disaster such as an earthquake or a hurricane, the communication from rescue crews to stationary databases may be intermittent.

Manuscript published on 30 August 2013.
***** Correspondence Author (s)
Er.Satwant Kaur, RIET Phagwar, India
Er.VarinderjitKaur, Assistant Prof. RIET Phagwar, India
Er.Gurpreet Singh, Assistant Prof.& Head CSE STSSIET, Jalandhar, India.

In such a scenario, P2P data sharing can providea robust alternative where fault-resilience is naturallybuilt into the design. Majority [2]of users in a mobile environment donot travel at random. They navigate from place to placewith specific purposes in mind. In many cases, the patternsof location movement and service invocation of mobileusers targeting similar purposes exhibit strong resemblance.The common patterns of location movement may due togeographic relationships between locations or servicedistribution. The regularity in service invocation may comefrom the dependencies between services or the proximity ofservice providers. It is potentially beneficial to discoversuch mobility and service patterns to facilitate network anddata management. Many mobility learning schemes andmotion prediction algorithms have been proposed toexplore the benefit of mobility patterns. Datamining techniques are also used in the discovery of user behavior patterns. However, these techniques aredesigned to discover either mobility patterns or requestpatterns. We argue in this paper that mobile users exhibitcomplex behavior patterns that cannot be easily describedby simple associations between sequences of frequentlytraveling locations or series of repeatedly accessed data. Wepropose the idea of complex activity for characterizing thecontinuously changing complex behavior patterns of mobileusers. A complex activity is a sequence of location movement,service requests, the cooccurrence of location andservice, or the interleaving of all above. An activity may becomposed of subactivities. Different activities may exhibitdependencies that affect user behaviors.

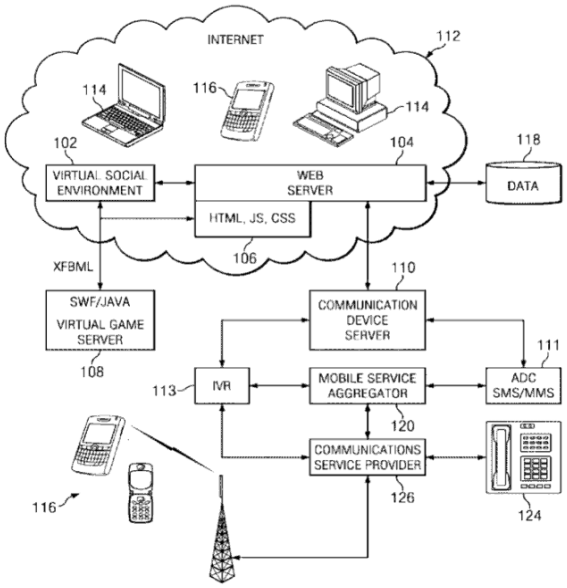


Fig: 1.1: Activity based Mobile environment

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

II. DATA MINING

data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into[3[useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

KNN: In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until [4]classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor.

III. ID3

The ID3 algorithm is considered as a very simple decision tree algorithm (Quinlan, 1986). ID3 uses information gain as splitting criteria. The growing stops when all instances belong to a single value of target feature or when best information gain is not greater than zero. ID3 does not apply any pruning procedures nor does it handle numeric attributes or missing values.

For the decision tree algorithm, ID3 was selected as it creates simple and efficient tree with the smallest
depth. Unlike a binary tree (such as used for Huffman Encoding where there is just a left and right node), the ID3 decision tree can have multiple children and siblings.

The ID3 decision makes use of two concepts when creating a tree from top-down:

1. Entropy
2. Information Gain (as referred to as just gain) Using these two algorithms, the nodes to be created and the attributes to split on can be determined.

The best way to [5]explain these concepts and how they apply to creating a decision tree using ID3 is to show an example.

Entropy

Entropy is the measurement of uncertainty where the higher the entropy, then the higher the uncertainty.

With a decision tree, this measurement is used to determine how informative a node is.

CART for Classification And Regression Trees, is an exploratory method used to study the relationship between a dependent variable and a series of predictor variables. CART modeling is also called C&RT in some programs or statistical texts. CART modeling selects a set of predictors and their interactions that optimally predict the dependent measure. The developed model is a classification tree (or data partitioning tree) that shows how major "types" formed from the independent (predictor or splitter) variables differentially predict a criterion or dependent variable.

CART diagrams should be thought of as a "tree trunk" with progressive splits into smaller and smaller "branches." The initial "tree trunk" is all of the participants in the study. A series of "predictor" variables are assessed to see if splitting the sample based on these predictors leads to better discrimination in the dependent measure. For instance, if our dependent measure is whether the patient has gotten medical case management services, we would first assess whether there are different levels of receiving this service for two groups formed on the basis of one of the predictor variables. The "most significant" of these predictions would define the first split of the sample, or the first branching of the tree. Then, for each of the new groups formed, we would ask if the subgroup could be further significantly split by another of the predictor variables. And so on. After each split, we ask if the new subgroup can be further split on another variable so that there are significant differences in the dependent variable. The result at the end of the tree building process is that we have a series of groups that are maximally different from one another on the dependent variable. At each step, the optimal binary split is made. Different orientations of the same tree are sometimes useful to highlight different portions of the results.

Advantages: The CART method has certain advantages as a way of looking for patterns in complicated datasets. [6]First, the level of measurement for the dependent variable and predictor variables can be nominal or ordinal (categorical) or interval (a "scale"). Second, the level of measurement for the predictor variables can be nominal or ordinal or interval. Third, not all predictor variables need be measured at the same level (nominal, ordinal, interval). Fourth, missing values in predictor variables can be estimated from other predictor ("surrogate") variables so that partial data can be used whenever possible within the tree. Fifth, if an appropriately conservative set of statistical criteria are used to prune the tree after is grown, the resulting models will primarily emphasize strong results without over-capitalizing on chance because the relationships between many variables are being considered at once. On the other hand, it must always be remembered that CART modeling is essentially a "stepwise" statistical method and that there is always a potential for too much to be seen in the data even when very conservative statistical criteria are used. Nonetheless, in those cases in which there is not a strong[7] theory in an area that would clearly indicate which variables are, and are not, probably predictors of some dependent measure, CART will be very useful in identifying major data trends

Simulator implementation for collection of mobile environment data set:

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

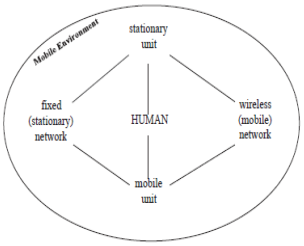


Figure 1: Data set collection based on mobility

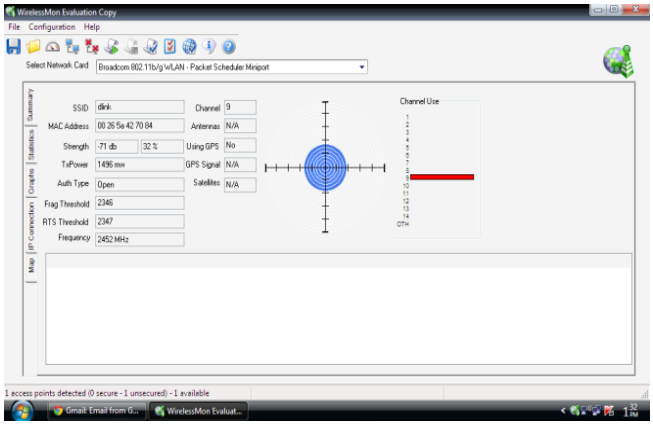


Figure 2: Strength, Auth. Type, Threshold and Frequency of the mobile network

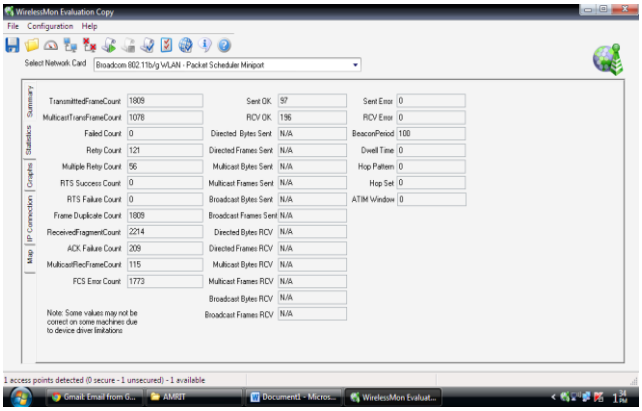


Figure 3: Transmitted Frame, Multicast Frame, Retry Count, RTS Success Count Statistics.

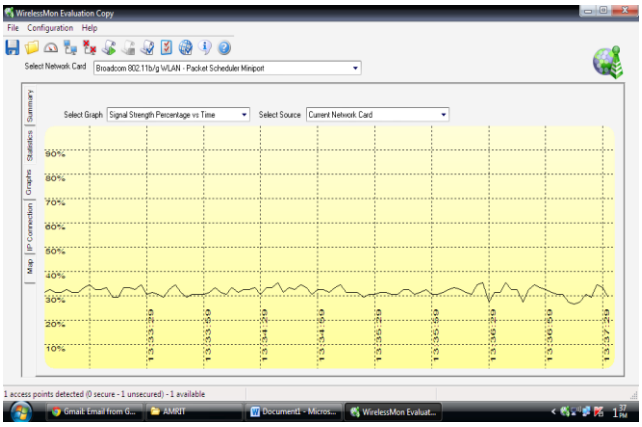
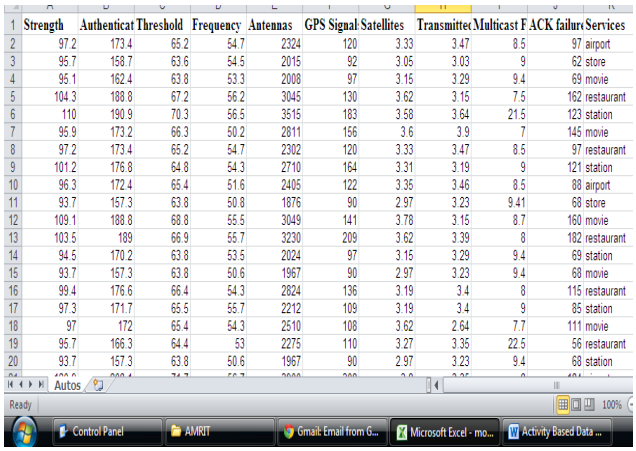


Figure 4: Signal Strength during mobile activity.

Attributes of Data set:

1.	Strength
2.	Authentication type
3.	Threshold
4.	Frequency
5.	Antennas
6.	GPS Signals
7.	Satellites
8.	Transmitted Frame
9.	Multicast Frame
10	ACK failure count
11.	Services

Collected Data Set:



Strength	Authenticat	Threshold	Frequency	Antennas	GPS Signal	Satellites	Transmitte	Multicast F	ACK failure	Services
97.2	173.4	65.2	54.7	2324	120	3.33	3.47	8.5	97	airport
95.7	158.7	63.6	54.5	2015	92	3.05	3.03	9	62	store
95.1	162.4	63.8	53.3	2008	97	3.15	3.29	9.4	69	movie
104.3	188.8	67.2	56.2	3045	130	3.62	3.15	7.5	162	restaurant
110	190.9	70.3	56.5	3515	183	3.58	3.64	21.5	123	station
95.9	173.2	66.3	50.2	2811	156	3.6	3.9	7	145	movie
97.2	173.4	65.2	54.7	2302	120	3.33	3.47	8.5	97	restaurant
101.2	176.8	64.8	54.3	2710	164	3.31	3.19	9	121	station
96.3	172.4	65.4	51.6	2405	122	3.35	3.46	8.5	88	airport
93.7	157.3	63.8	50.8	1876	90	2.97	3.23	9.41	68	store
109.1	188.8	68.8	55.5	3049	141	3.78	3.15	8.7	160	movie
103.5	189	66.9	55.7	3230	209	3.62	3.39	8	182	restaurant
94.5	170.2	63.8	53.5	2024	97	3.15	3.29	9.4	69	station
93.7	157.3	63.8	50.6	1967	90	2.97	3.23	9.4	68	movie
99.4	176.6	66.4	54.3	2824	136	3.19	3.4	8	115	restaurant
97.3	171.7	65.5	55.7	2212	109	3.19	3.4	9	85	station
97	172	65.4	54.3	2510	108	3.62	2.64	7.7	111	movie
95.7	166.3	64.4	53	2275	110	3.27	3.35	22.5	56	restaurant
93.7	157.3	63.8	50.6	1967	90	2.97	3.23	9.4	68	station

Figure 5: Collected Data Set

Implementation of ID3 and KNN

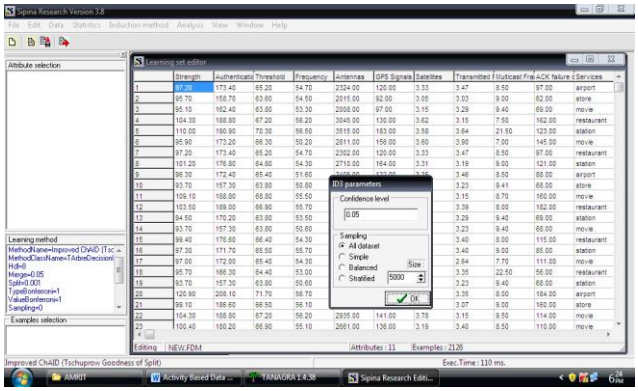


Figure 6: Import Data Set

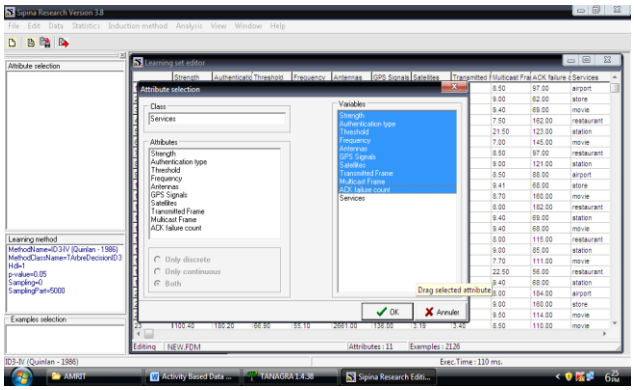


Figure 7: Define Class Variables

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

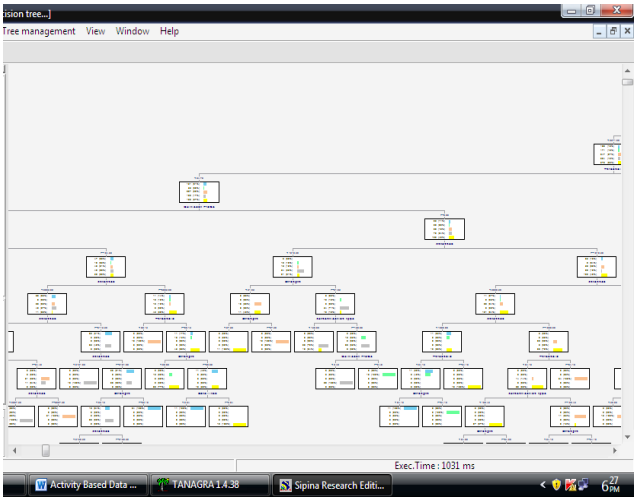


Figure 8: Generating Tree based on class attribute services

Error rate			0.4911						
Values prediction			Confusion matrix						
Value	Recall	1-Precision		airport	store	movie	restaurant	station	Sum
airport	0.2896	0.5640	airport	75	23	69	59	33	
store	0.3115	0.6049	store	0	81	91	36	52	
movie	0.5413	0.5054	movie	54	24	321	92	102	
restaurant	0.5802	0.4931	restaurant	22	53	114	293	23	
station	0.6130	0.4023	station	21	24	54	98	312	
			Sum	172	205	649	578	522	

Figure 9: Calculate Error Rate

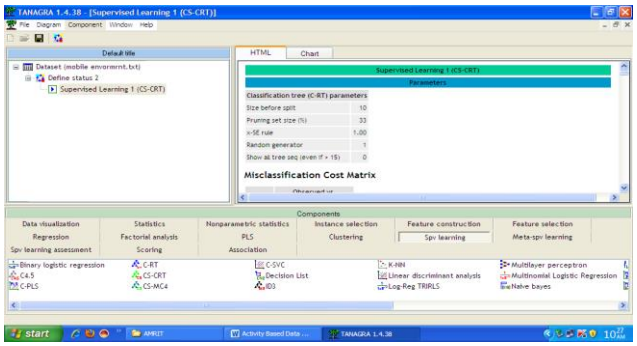


Figure 10: Parameters of CART Algorithm

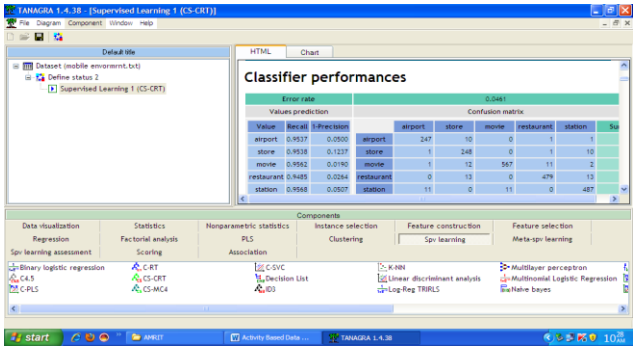


Figure 11: Calculate error rate

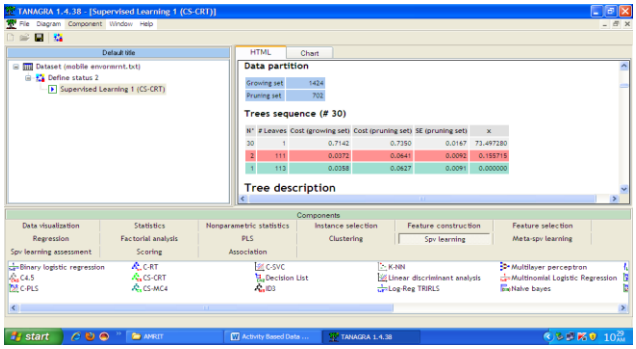


Figure 12: Calculate no of nodes

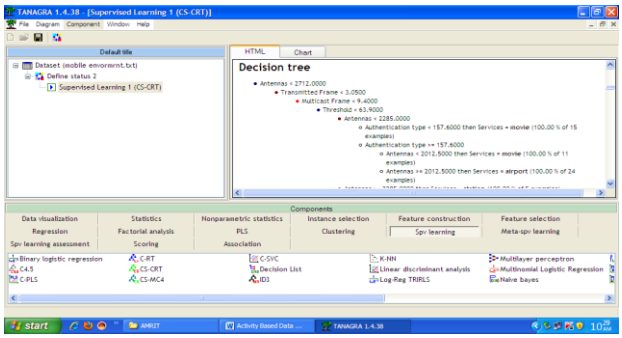


Figure 13: create decision tree

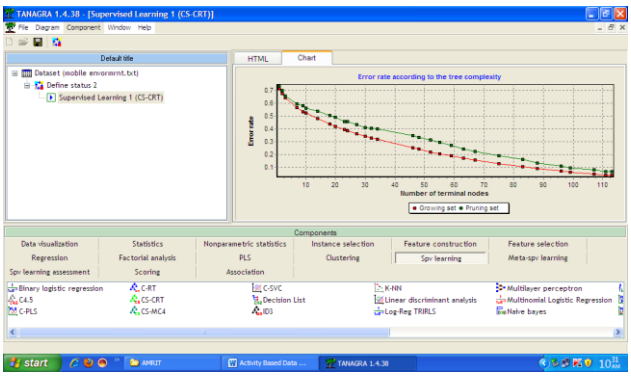


Figure 14: show growing set and pruning set

Result:

	ID3	CART
ERROR RATE	0.40	0.04
No. of Nodes	61	217
No. of Leaves	31	109
Computation Time	94 ms	125 ms

IV.CONCLUSION

In this Research paper, we wanted to highlight the approaches for creating a decision tree. They are mainly available into academic tools from the machine learning community. We note that they are an alternative quite credible to decision trees and predictive association rules, both in terms of accuracy than in terms of processing time. After analysis Order ID3 AND CART algorithm is more suitable to find accurate and consuming less access time to mine data with minimum error rate 0.04. cart is a best algorithm for mining a data on mobile environment data set.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

REFERENCES

1. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc. Int’l Conf. Very Large Databases (VLDB), pp. 487-499, 1994.
2. W.-C.P. Jiun-Long Huang and M.-S. Chen, “Exploring Group Mobility for Replica Data Allocation in a Mobile Environment,” Proc. 12th Int’l Conf. Information and Knowledge Management, pp. 161-168, 2003.
3. J.-L. Huang and M.-S. Chen, “On the Effect of Group Mobility to Data Replication in Ad-Hoc Networks,” IEEE Trans. Mobile Computing, vol. 5, no. 5, pp. 492-507, May 2006.
4. W. Ma, Y. Fang, and P. Lin, “Mobility Management Strategy Based on User Mobility Patterns in Wireless Networks,” IEEE Trans. Vehicular Technology, vol. 56, no. 1, pp. 322-330, Jan. 2007.
5. W.-C. Peng and M.S. Chen, “Allocation of Shared Data Based on Mobile User Movement,” Proc. Third Int’l Conf. Mobile Data Management, pp. 105-112, 2002.
6. M. Sricharan, V. Vaidehi, and P. Arun, “An Activity Based Mobility Prediction Strategy for Next Generation Wireless Networks
7. V.S. Tseng and K.W. Lin, “Mining Sequential Mobile Access Patterns Efficiently in Mobile Web Systems,” Proc. 19th Int’l Conf. Advanced Information Networking and Applications, vol. 2, pp. 762-767, Mar. 2005.
8. H. Cao, N. Mamoulis, and D. Cheung, “Mining Frequent Spatio-Temporal Sequential Patterns,” Proc. Fifth IEEE Int’l Conf. Data Mining, 2005.
9. V.S. Tseng, H.-C.Lu, and C.-H. Huang, “Mining Temporal Mobile Sequential Patterns in Location-Based Service Environments,” Proc. Int’l Conf. Parallel and Distributed Systems, vol. 1, pp. 1-8, 2007.
10. W.-C. Peng and M.-S. Chen, “Shared Data Allocation in a Mobile Computing System-Exploring Local and Global Optimization,” IEEE Trans. Parallel and Distributed Systems, vol. 16, no. 4, pp. 374- 384, Apr. 2005.