Context-Aware Data Mining Framework for Wireless Medical Application

Pravin Vajirkar, Sachin Singh, and Yugyung Lee

School of Computing and Engineering,
University of Missouri–Kansas City, Kansas City, MO 64110 USA.
{ppv22e, sbs7vc, leeyu}@umkc.edu

Abstract. Data mining, which aims at extracting interesting information from large collections of data, has been widely used as an effective decision making tool. Mining the datasets in the presence of context factors may improve performance and efficacy of data mining by identifying the unknown factors, which are not easily detectable in the process of generating an expected outcome. This paper proposes a Context-aware data mining framework, by which contexts will be automatically captured to maximize the adaptive capacity of data mining. Context could consist of any circumstantial factors of the user and domain that may affect the data mining process. The factors that may affect the mining behavior are delineated and how each factor affects the behavior is discussed. It is also observed that a medical application of the model in wireless devices offers the advantages of Context-aware data mining. A Context-aware data mining framework is quantified through a partial implementation that would be used to test the behavior of the mining system under varied context factors. The results obtained from the implementation process are elucidated on how the prediction output or the behavior of the system changes from the similar set of inputs in view of different context factors.

1 Introduction

Context-awareness computing work has been carried out by many researchers [4, 2, 1]. Many of them have been worked on defining Context-awareness and some of them are mainly focusing on building Context-aware applications. However, little has been done in building a framework which supports data mining based on Context awareness leading to useful and accurate information extraction.

Context is a powerful, long-standing concept. In computer-human interaction it can be mostly captured via explicit models of communication (e.g., user query inputs). However, implicit context factors (e.g., physical environmental conditions, location, time etc.) are normally ignored due to absence of knowledge base or appropriate model. We believe implicit context-aware factors could be used to interpret and enhance explicit user inputs and thereby affecting data mining results to deliver accurate and precise prediction results.

Nowadays, huge volume of data is available. This data-rich environment does not guaranty for information-rich environment. Data mining is a process that discovers patterns in data that may be used for valid predictions [6]. We focus on building a context-aware data mining framework that can filter useful and interesting context factors, produce accurate and precise prediction using those factors.

This framework was designed keeping the medical applications in mind, even though it is generic in nature and can be used in any domain. Medical professionals are always on-the-go. The use of decision support PDA supported by data mining facility can be of great asset to the medical professionals while working on an emergency or while rushing to attend an emergency. Even in other domains, getting the data mining based services on a mobile device can be of substantial utility, like a stock broker getting the predicted value of stocks on his PDA. Presenting the information on a PDA deals with a set of constraints due to its restricted resources like memory, screen and CPU power. Providing Data Mining services on a PDA is a case of a very *thin* client communicating with a very *fat* server, required a huge amount of processing. Despite the technical challenges that one faces due to the limited computing power of mobile devices, it provides a great deals of portability, mobility and practical usage to the proposed context-aware data mining framework.

2 What is Context-Awareness?

The concept of *Context* have often been interwoven and used in many different applications. Depending upon the application an information may be interpreted as a *meaningful* but in another context, as a meaningless. Thus, when the information has to be conveyed from one element to another, the interpretation of information can be changed particularly depending on the context of the information.

Dey and Abowd [4] defined *context* as a piece of information that can be used to characterize the situation of participants in an interaction. Similarly, [2, 1] defined context as information on the location, environment, identity of people in a certain situation. By sensing context information, context enabled applications can present context information to users, or modify their behavior according to changes in the environment [8]. Chen and Kotz [3] defined *context* as the set of environmental states and settings that either determines what an application behavior is or where the event occurs. Schilit and Theimer [10] emphasized the importance of applications which adapt themselves to context.

There are some definitions which were too broad to apply to any application while some are too specific to certain domains. In real-world datasets, the context-aware factors that constitute context-awareness change rapidly and therefore the factors tend to become subjective and very domain specific. Our challenge in this paper is to make the concept *hybrid* (combining generic and domain specific). From our point of view on context, lack of context-awareness leads to missing a lot of critical and useful information that would affect the data mining process and thereby, affecting the data mining results. Our definition of the context is the information regarding objects (including actors) which supports the entire process from the user query to the mining. More importantly, the context will provide the system the ability to adapt to a changing environment during the data mining process and thereby providing the users with a time sensitive data accurately, efficiently and in a precise manner.

We now define the types of context factors:

User Context: *User Identity Context* describes the information of user responsible for the query including his/her field of expertise, authorization of tasks or datasets, his/her team members and their expertise fields. *User History Context* describes the history (i.e., user-profiling) built up for each user when he/she queries for a particular information. This helps when the user frequently queries with a similar query or uses the same piece of information.

Application Context describes the evolving context with the application specific and wireless features. *Joint Conference Context*: A specific query on patient can perform jointly to combine related information, such as a conference with neuro-surgeon and diabetics specialist. The context-ware data mining enables all related areas of work related to the patient to be covered and ensure for the patient's well-being as developed at Future Computing Environment [5]. *Time Context*: Time is an important aspect to certain domains such as stock market, war planning or medicine. Considering an example of diagnosis of diabetics, the sugar level tends to drop at night than in day time. It is also higher an hour after meals than before meal. We can have different datasets on the bases of at what period of time the blood sample was collected to get better prediction outputs.

Data Context *Domain Context* describes domain specific context. In our medical application, it is *Patient Context* which captures the personal and medical history of the patient. It also records the immediate family members and their medical history. *Location Context:* The importance of this context lies in the fact that it can be used to cluster datasets. For example, assume that we have collected the dataset from different zone/states/countries etc. If the patient under consideration falls in any one of these parameters, then the system should extract a set of records, which correspond to the particular zone and then use this subset for mining. The datasets primary formed from the population living near a certain location as living area is related to health issues. For example, people living in coastal regions have less probability of getting goiter. Similar, people living in the country side have less tendency to get high blood pressure as compared to suburban folks. It would be a good idea to pick the datasets depending on the location context of the patient. This may improve the accuracy of the system, wherein we just concentrate on the data more relevant to him/her than the generic data. If the patient doesn't fall in any one of the zones, then the system can approximate his datasets to the zone closest to him/her or use the default datasets or combination of two zones etc.

Data Mining Context describes attributes related to the data-mining query: *Data Context* As the data is distributed in a typical domain, it is important to determine which dataset to be mined for a given context. *Attribute Context* The classification of Data mining [11, 7, 6] builds a model (e.g., a decision tree in C4.5) as a prediction model and querying this tree based on the user inputs leads to predicting the value of desired information. Since these models are built with the attributes whose prediction is to be made, it is important to identify which attributes should be chosen for a query in a given context. *Performance Context* describes the time taken for data mining process. This addresses the classic question of speed vs. accuracy issue in data mining. Certain applications would require faster responses for multiple attributes even if the accuracy is not the best. For example, stock applications which predict stock prices of numerous companies pretty quickly with reduced accuracy. Certain applications, however, would require highest accuracy even if it takes time. Medical applications where the prediction values would be critical for a future course of treatment to be given to a patient has to be accurate even if it takes more time. Thus, depending on the context the framework may choose an appropriate algorithm for the mining process.

3 Context-Aware Data Mining Framework

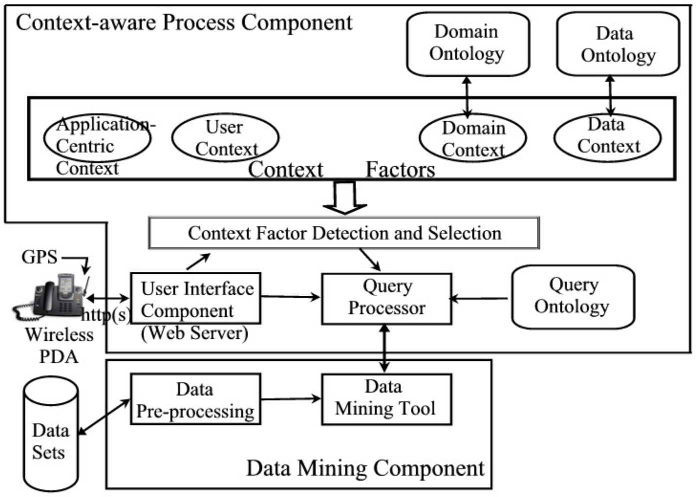


Fig. 1. System Architecture

We propose a context-aware framework which considers context factors during data mining process. As a proof of concept, we have implemented a decision support system for wireless medical applications. Figure 1 shows the system architecture of the Context-aware data mining framework.

Context-aware Process Component handles context factors for Data mining and Query processor components. In this component, Context factors such as *Application Context*, *User Context*, *Domain Context*, *Data Context* are defined as described in Section 2. In order to identify context factors from the implicit information of the user's query, we perform the processes including checking the attributes of the user query; learning about the entities of the query context.

For instance, Heather has an allergic symptom and is living in Huntsville, AL. The model can analyze Huntsville's pollen grain content at that period of the year and season and invokes subsequent processes such as selecting the desired datasets for the data mining component; discovering missing entities: Comparing with the datasets, missing entities that might help in mining the data accurately can be identified; classifying the discovered entities based on the Context-Aware factors; resolving any missing, ambiguous and conflicting information with minimum user's involvement.

Ontology Component defines a high level meta-data information about the existing datasets, users, and mining tools. For instance, the domain ontology stores knowledge about the existing datasets like relationships between datasets at structural level and also at the semantic level. For example, if one gives queries for diabetics datasets with a missing parameters for say, blood pressure. The missing value can be predicted from the blood pressure dataset. Interestingly, we may mine more datasets to fill in the missing attributes.

Data-mining Component consists of two subcomponents. The first one is a preprocessing component which converts the existing datasets format into the format acceptable by the data mining tool. The second subcomponent is the actual data mining tool (Weka3[1]) which requires ARFF format as input. The tool mines the datasets, given the input parameters, like the datasets name, query elements and values, etc. This subcomponent accepts the preprocessed datasets and the query parameters and returns the queried result.

User Interface Component is a Web Server component which interacts with the client subsystems on the PDA over HTTP(S). Once it receives a request from the user, it forwards it to the Query Processor. At the same time, it also passes the query to the Context Factor component that determines which context factors are applicable for the given query. After determining which ones are applicable it extracts the relevant information from the query and passes it to the Query Processor.

Query Processor has all the information explicitly given by the user in the form of the query and all implicit information from the Context component. The Query Processor may use the output of one query as an input to other query and so on till it gets all the inputs required for the original query that the user had requested. In case of still-missing attributes, user input will be requested. Thus the Query Processor can be considered as a Meta task manager, managing multiple atomic tasks which are part of the same high level task. After the Meta task is executed the final result is returned to the user over HTTP(S) to the PDA as the prediction or query results. Using appropriate domain-specific query ontology it will construct an appropriate query/set of queries.

4 Medical Scenarios

The usage of Context-aware factors tends to change the data mining behavior by the pruning the user query by attaching its context factors. These factors

[1] http://www.cs.waikato.ac.nz/ ml/weka/index.html

could affect the change by picking different and the correct datasets, changing its query values itself, attaching missing query values to its attributes, and/or affecting the change using different data mining algorithms, etc.

Case 1: Diabetes Treatment Dr. Smith is a female diabetics specialist. He has a patient named Martha whose report show a high level of sugar in blood. Dr. Smith can always rely on his experience and prudence, though he would like to explore the advantages of data mining techniques which can derive some useful information from historical data. For this scenario, we have selected a diabetes dataset[2]. The attributes from 1^{st} - 8^{th} are the patients inputs while 9^{th} (Class) is the output results which supports his treatment decision for Martha. The dataset attributes for the Diabetics are follows: [min, max, mean, SD].

1. Number of times pregnant [0, 17, 3.84, 3.36]
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test [0, 199, 120.89, 31.97]
3. Diastolic blood pressure (mm Hg) [0, 122, 69.10, 19.35]
4. Triceps skin fold thickness (mm) [0, 99, 20.53, 15.95]
5. 2-Hour serum insulin (mu U/ml) [0, 846, 79.79, 115.24]
6. Body mass index (weight in kg/(height in $m)^2$) [0, 67.1, 31.99, 7.88]
7. Diabetes pedigree function [0.078, 2.42, 0.471, 0.331]
8. Age (years) [21, 81, 33.24, 11.76]
9. Class variable (0 or 1) [Discrete][500-negative, 268-positive]

Table 1 shows the interpretation of the given diabetes dataset from the perspective of our context-aware model.

Location Context	The record history used in diabetes dataset is explicitly from the area near Phoenix, Arizona. The user has a choice to pick location specific if he believes that the problem has no regard to location change. Some results are more location sensitive than the other. Here in this case, there is no direct relation to location.
Application Context	Dr. Smith may want to discuss the discovered result with gynecologist (if Martha is pregnant), or with a blood pressure expert to jointly come to an effective diagnostic treatment.
User Context	According to user's specified area of interest, a user profile can be developed. For an advanced user, his various query records and statistics allow him know about the risk factors of the results. In this case, he may query diabetes dataset, jointly with other datasets (e.g. blood pressure, pregnancy, etc). The selecting these datasets can be done by analyzing Dr. Smith's profile and a flavored output can be generated.
Data Mining Context	Five of input details about Martha are known and rest are unknown or uncertain. The application would try to fill in the void data by asking relative queries to the model and using the supporting dataset. The input attribute 7 from the diabetes which asks the user about the pedigree value for Martha. The application would very closely speculate the value based on family history, genetic history, and nursing assessments.

Table 1. Case1: Interpretation of Context-Aware Factors

[2] http://kdd.ics.uci.edu/

Case 2: Heart Attack Risk In this scenario Dr. Smith is interested in determining whether the major blood vessel is <50% or >50% narrowed as a measure of a probable heart attack risk. This could form the basis of further action that he may take in treating the patient. The effect of context factors can be explained in the light of a carefully selected scenario. The datasets we have chosen is the heart datasets[3] as our proof-of-concept. This real-life datasets was customized to make use of other related datasets and to prove the validity of our context factors. Let us consider the datasets attributes first.

1. (age) Age in years
2. (sex) Sex (1: Male; 0: Female)
3. (chest_pain) chest pain type (1: Typical angina; 2: Atypical angina; 3: Non-anginal pain, Value 4: Asymptomatic)
4. (trestbps) resting blood pressure (in mm Hg on admission to the hospital)
5. (chol) Serum cholestoral in mg/dl
6. (fbs) (Fasting blood sugar >120 mg/dl) (1: True; 0: False)
7. (restecg) resting electrocardiographic results (0: Normal; 1: ST-T wave abnormality; 2: Definite left ventricular hypertrophy)
8. (thalach) Maximum heart rate achieved
9. (exang) Exercise induced angina (1 = yes; 0 = no)
10. (oldpeak) ST depression induced by exercise relative to rest
11. (slope) The slope of the peak exercise ST segment (1: upsloping; 2: flat; 3: downsloping)
12. (ca) Number of major vessels (0-3) colored by flourosopy
13. (thal) the heart status (3: Normal; 6: Fixed defect; 7: Reversible defect)
14. (Family-Hist) - History of any heart disease within immediate family (1: True; 0: False)
15. (Smoke-Disease) - Symptoms of smoke disease
16. (Location) - Location of the person where he lives.
17. (num) Diagnosis of heart disease (angiographic disease status) (0: <50% Diameter narrowing; 1: >50% Diameter narrowing)

We have selected attribute 17 as the pivot element to form the prediction value in the classification tree. Apparently, we need a few more factors to determine the value of our pivot element. Most of the elements from attribute 4 through attribute 13 are standard clinical tests and should be available to doctor. Table 2 shows the interpretation of the given heart attack risk datasets from the perspective of our context-aware model.

It is worthwhile to note that though this scenario brings up the possible uses of different context factors, it also highlights the different behavior of the system. First, the system picks up a data value from a record to plug in to another datasets. The idea is the system to *know* where to pick and plug the data. Second, the system *trims* or *customizes* the data. For instance, the system can filter relevant datasets attributes for mining. Third, the system picks up related auxiliary datasets, mines it, and grabs a predicted value to use this output for querying the main datasets classification tree. Finally, we also think of

[3] http://kdd.ics.uci.edu/

Data Context	This parameter refers to any health affects caused by say, smoking. Determining whether a person has smoking ill effects is in itself a sub-problem. Here we refer to dataset which has information like (1) Smoking since (2) Cigarettes per day (3) Quitting period. Based on these input parameters from the user, the system picks up another dataset referring to smoking, say *Smoking Effects*. We perform mining with this dataset to build the classification tree and thereby to predict the person's smoking problems. Using the input parameter for the given patient, the system will query the above tree and predict whether the patient has smoking problems. The predicted output is the input to the original user query and is used to plug in the smoking attribute to the original classification tree for the heart disease. It is so because one of the inputs to the original query is output after the system selects related datasets. In this case, the system will select the auxiliary datasets only if the patient ever smoked. Thus based on the context different datasets are picked up to query and used in plug-in the missing attribute value(s).
Patient Context	The system knows the patient's immediate family members. For each of these members it can access the *Historical patient repository*. We are assuming that hospitals maintain some medical record of each patient in history about what disease they had in past, etc. If any one of these members had any such a medical condition, it can be picked up as input.
Location Context	Location is less likely affect to the data mining behavior because Location here is not an input to the system. But there are cases that a context factor can affect the output given the same input parameters.

Table 2. Case 2: Interpretation of Context-Aware Factors

hybrid behavior, where the system picks values as combination of all approaches discussed above and then computes the values by using some domain-specific formulae. Example, for height/weight ratio parameter, we can grab height and weight as two different values and then compute its ratio.

5 Implementation and Experimental Results

The proposed framework has been implemented in a wireless J2ME[4] enabled, CLDC[5] (Connected-Limited Device Configuration) complaint PDA (Personal Digital Assistant). The communication from PDA to the server is made through HTTPS connection, so make the connection secure. The connection is established at the physical layer and the user is given the desired output on his/her PDA screen. The application contains palm-sized PDA that would communicate with the computer, enabling *Context-aware data mining framework*. The wireless component is assumed to have following capabilities and features.

Now we will show the experimental results regarding the heart attack risk case. The dataset selected is the heart dataset[6] that stores the history of patients with heart diseases. The dataset consists of following attributes:

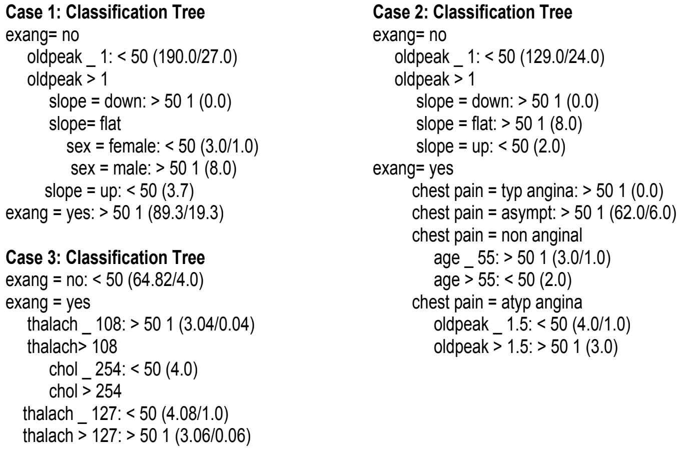
[4] http://java.sun.com/j2me/

[5] http://0511704376-0001.bei.t-online.de/seiten/indexeh.htm

[6] http://www.cs.waikato.ac.nz/ ml/weka/arff.html

– Personal details: *age* in years and *sex* (1 = male; 0 = female)
– Cardiac Details: *painloc* : chest pain location (1: substernal; 0: otherwise)
– The prediction output for this dataset is *num*: the diameter of the artery (angiographic disease status) (0: < 50% diameter narrowing; 1: > 50% diameter narrowing)

The user requests a query, which consists of inputs. He/she expects an outcome prediction value *num*. The system performs data mining using all attributes and entire datasets to construct the model. Using this model, it will apply the query variables to get the prediction result. We applied the J48 algorithm [11] and build a decision tree using C4.5 [7] on the heart dataset. Figure 2 (Case 1) shows the classification tree generated using the decision tree.



Case 1: Classification Tree

```
exang= no
    oldpeak _ 1: < 50 (190.0/27.0)
    oldpeak > 1
        slope = down: > 50 1 (0.0)
        slope= flat
            sex = female: < 50 (3.0/1.0)
             sex = male: > 50 1 (8.0)
       slope = up: < 50 (3.7)
exang = yes: > 50 1 (89.3/19.3)
```

Case 3: Classification Tree

```
exang = no: < 50 (64.82/4.0)
exang = yes
    thalach _ 108: > 50 1 (3.04/0.04)
    thalach> 108
        chol _ 254: < 50 (4.0)
        chol > 254
   thalach _ 127: < 50 (4.08/1.0)
   thalach > 127: > 50 1 (3.06/0.06)
```

Case 2: Classification Tree

```
exang= no
    oldpeak _ 1: < 50 (129.0/24.0)
    oldpeak > 1
        slope = down: > 50 1 (0.0)
        slope = flat: > 50 1 (8.0)
        slope = up: < 50 (2.0)
exang= yes
       chest pain = typ angina: > 50 1 (0.0)
       chest pain = asympt: > 50 1 (62.0/6.0)
       chest pain = non anginal
           age _ 55: > 50 1 (3.0/1.0)
           age > 55: < 50 (2.0)
       chest pain = atyp angina
           oldpeak _ 1.5: < 50 (4.0/1.0)
           oldpeak > 1.5: > 50 1 (3.0)
```

Fig. 2. The Experimental Results

Using the meta-level understanding, we know that the personal detailed attributes can be patient-contexts. In our case, we considered *Sex* as a context factor. We will use the same query inputs to analyze the change. The application is now *context-aware* so it knows that part of the query where the *Sex* variable is actually a context input. Male persons can have different factors affecting more predominantly as compared to female ones. If we mine the datasets differently we may get interesting results. So if the user query variable is *Sex=male*, then it may not make sense to mine the entire dataset and then query the model and get the results. We have retrieved only those records that are males, and then mine the dataset excluding the *Sex* column. We performed a vertical and horizontal trimming of dataset. The data model mined out of this could be called as *specialized* dataset. In our experiments we mined the datasets separating on basis of *Sex* as context information. Figure 2 (Case 2) shows the C4.5 tree corresponding to males. As is evident from the figure the new model shows emergence of some new factors in the decision making tree which were not appearing in the non Context-aware factor domains. The Context-aware factors can achieve different results and also give more insight of the trend in the dataset and their interrelations. Figure 2 (Case 3) shows the classification tree when entering a constraint *Sex = female*.

Thus, we can improve the accuracy of the data mining process or simple bring out the ”suppressed” trends within the data by suitably applying context factors to the data mining process.

6 Conclusion

In this paper we have introduced an application model for providing accuracy and precision to data-mining prediction results based on context-aware factors. The model is wrapped on a wireless framework to provide user-friendliness, mobility and practical usage. The model proposed is generic in nature and can be applied to most of the fields. Two scenarios were provided as a proof of concept to our proposed model. Much of the work is focused on combining Context-aware, data-mining and wireless communication and towards feasibility of such a model. The model is tested and implemented with live and customized medical datasets to make the model efficient.

References

1. Brown, P.J.: The Stick-e Document: a Framework for Creating Context-Aware Applications. Electronic Publishing 96 (1996) 259-272
2. Brown, P.J., Bovey, J.D., Chen, X.: Context-Aware Applications: From the Laboratory to the Marketplace. IEEE Personal Communications, **4(5)** (1997) 58-64
3. Chen, G., Kotz, D.: A Survey of Context-Aware Mobile Computing Research. Dartmouth Computer Science Technical Report TR2000-381
4. Dey, A.K., Abowd, G.D.: Towards a better understanding of Context and Context-Awareness. GVU Technical Report GITGVU-99-22, College of Computing, Georgia Institute of Technology. **2** (1999) 2 – 14
5. Dey, A.K., Futakawa, M., Salber, D., Abowd, G.D.: The Conference Assistant: Combining Context-Awareness with Wearable Computing. In Proceedings of the 3rd International Symposium on Wearable Computers (ISWC ’99), 21-28, San Francisco, CA, October 1999. IEEE Computer Society Press.
6. Edelstein, H. A.: Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation, 1999. ISBN: 1-892095-02-5
7. Ragone, A.: Machine Learning C4.5 Decision Tree Generator
8. Salber, D., Dey A.K., Orr, R.J., Abowd, G.D.: Designing For Ubiquitous Computing: A Case Study in Context Sensing, GVU Technical Report GIT-GVU 99-29, (http://www.gvu.gatech.edu/)
9. Schilit, B., Adams, N., Want, R.: Context-Aware computing applications. In Proceedings of IEEE Workshop on Mobile Computing Systems and Applications, 85-90, Santa Cruz, California, December 1994. IEEE Computer Society Press.
10. Schilit, B., Theimer, M.: Disseminating Active Map Information to Mobile Hosts. IEEE Network,**8(5)** (1994) 22-32
11. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.