INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

User Profiling of Automobile Driver and Outlier Detection

Bharat S. Dangra
PG Scholar, Department of Computer Engineering
MAEERS - Maharashtra Institute of Technology, Pune, India
Dr. Mangesh V. Bedekar
Department of Computer Engineering
MAEERS - Maharashtra Institute of Technology, Pune, India
Suja S. Panicker
Department of Computer Engineering
MAEERS - Maharashtra Institute of Technology, Pune, India

Abstract:
The central idea of the paper is to profile the driver of vehicle. Profiling is the collection of data associated with specific user. It's the representation of person's identity. So we will profile and learn the driver behavior from his driving habits like the way he change gears, the way he accelerates, the way he takes turn and many more parameters such that any unusual change in driving behavior can be traced and should be reported to the authentic user in real time. All the data will be recorded in the vehicle itself using in vehicle data recorder (IVDR). There are certain diagnostic codes made available by auto manufacturer at the OBD-II port, Controller Area Network (CAN) data recorder of the vehicle. All this learning process is carried out implicitly; no explicit involvement of driver is required other than driving the vehicle. Our focus is on household vehicle generally being driven by various members of family and that's make the vehicle as "Multiuser Single System".

Keywords*: User profiling, Outliers, Part profiles, VAN, OBD-II, CAN*

1. Introduction

Today car security is one of the challenging issues in our society. Despite the various technologies that have been introduced in recent years to deter car thefts and tracking it, it was reported that as many as cars were stolen yearly in the world. According to National Crime Information Center, in 2006, 1,192,809 motor vehicles were reported stolen, the losses were 7.9$ billion. There were an estimated 721,053 motor vehicle thefts in the US in 2012. According to FBI, a motor vehicle is stolen in the US every 44 seconds [12].

In India in 2010 automobile thefts increased by 5.6% compared with thefts in 2009. Total numbers of more than 40,000 cars, costing about 160 crore, are stolen every year in India. Every day, more than 40 vehicles get stolen in Delhi. The city accounts for 9.7% of motor vehicle thefts in the country, second only to the much larger states of Uttar Pradesh (14.1%) and Maharashtra (12.7%), says the latest National Crime Records Bureau report. At 87.6 per one lakh population, Delhi also has a much higher rate as compared to the national rate of 12.5. Only 20% of these stolen vehicles are recovered. In 2011, 14,668 motor vehicles were stolen in Delhi, which is a marginal decrease from 2010's figure of 14,966. While 9,203 of these vehicles were two-wheelers, 5,050 were cars. Only about 2,957 of these vehicles were recovered, show Delhi Police statistics. Motor vehicle thefts comprise of 27% of the total crimes in Delhi. The latest National Crime Records Bureau report says 'auto theft' in the country accounted for 44.4% (1,51,200 cases) of the total theft cases, which accounted for an increase of 2.5% in 2011 as compared to 2010 (1,47,475 cases) [13]. All this information makes us think what if the vehicle itself detect theft, what if it can profile the user and trace unknown behavior as an outlier and that too in real time.

2. Finding Patterns

There are certain characteristics which every driver exhibits while driving and they can be divided into two types one is static characteristics and another is dynamic characteristics. There are various characteristics like how driver accelerates, how he applies brake, how he changes gears and many more. From above characteristics if they are taken into consideration individually like we are concentrating on how driver accelerates then it is termed as static characteristics, whereas if more than one characteristic is taken into consideration simultaneously then it is termed as dynamic characteristics like when driver takes turn we have to consider how he accelerates, how he applies brake, how he change gears, steering wheel angle and its velocity, whether he gives indicator or not and likewise. There are unique pattern in which every driver exhibits these characteristics and from these characteristic we will generate a profile which will uniquely represent the driver's driving behavior. All these characteristics, whether static or dynamic will constitute part profiles which in turn together form the whole profile. In this paper, we are only considering static characteristics. All these characteristics we are going to record in vehicle itself using In-vehicle data recorders (IVDR)[1]. There are certain diagnostic codes made available by auto manufacturer. In a VAN (Vehicle Area Network), there are diagnostic codes made available by namely at these locations OBD-II (On Board Diagnostic) Port and CAN (Controller Area Network) Data Recorder. There are various signals that can be gathered at OBD-II recorder like the vehicle speed, Engine speed, Mass air flow rate, Coolant temperature, Throttle percentage, Fuel use, Temperatures (of oil, coolant, air intake, etc.), Engine load, Electrical voltages, Fault conditions and the signals that can be gathered at CAN (Controller Area Network) Data Recorder are Acceleration, Engine Speed, Wheel Speeds, Brake Pressure, Throttle Percentage, Steering Wheel Angle, Steering Wheel Velocity, Vehicle Speed. The characteristic we have discussed earlier are totally different from the diagnostic codes which have been made available by the auto manufacturers at various ports. Driver profile is based on these characteristic no doubt and we can derive these characteristics out of the codes available at various ports. Say for example we can derive that whether the turn taken by the driver is cautious move or the aggressive one. It depends on the speed of vehicle, the velocity with which he moves steering wheel and the angle of steering wheel while taking turn and likewise we can derive all other mentioned characteristics. From the signals which we are gathering at OBD-II and CAN there are some signals from which we can derive drivers behavior whereas there are some signals like Engine speed, Mass air flow rate, Coolant temperature which describes the internal state of vehicle. We will also collect this data and will call it as vehicle data. So in driving session while logging driver’s behavior we will also record vehicle data generated in response from the action taken by driver like what is Engine speed, Coolant temperature when driver is going with certain acceleration. Both these logs i.e. driver characteristics log and vehicle data log are therefore linked with each other with time factor.

3. Architecture

Our system comprises of two components namely Vehicle and Remote Server.

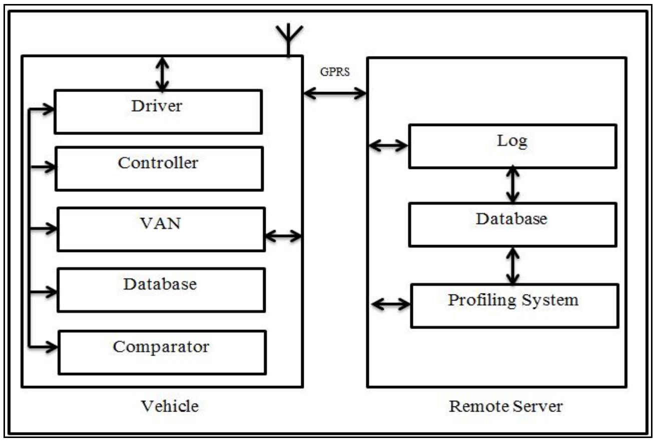

Figure 1: System Architecture

Vehicle consists of driver, VAN (Vehicle Area Network), Controllers (OBD-II, CAN), comparator and database for storing logs and profiles and Remote Server consists of database and profiling system. Both these components can connect to each other as and when required through GPRS. So when driver will drive the vehicle his behavior in terms of signals can be gathered at Controllers from where VAN will collect these signals and store them in log files at database. VAN is a vehicle area network which is a local area network in and around a moving vehicle. It enables devices in and around the vehicle to communicate either directly or through wireless protocols over internet. Say after the session (session comprise of the time driver start his vehicle to the time he turn off the ignition system) or in some idle time VAN will forward the stored log on to remote server. After receiving log from VAN it will be stored on database and then followed by preprocessing. Preprocessing simply reformats the log file. After this Profiling system will take the data collected from number of sessions and will process the data and generate profiles out of it. These profiles will be forwarded on to vehicle for real time comparison for finding outliers. Profiling system will also make sure to update the profiles with the changing behavior of drivers.

4. Learning

The description of what information is of interest to a user is commonly referred to as a user profile. In this scenario the interested information is various characteristics exhibited by automobile driver while driving. Learning is required for deriving pattern out of these characteristics. Learning will be carried out implicitly. No extra input is demanded from the driver. The only required input is that the driver has to drive the vehicle the way he drives. Learning will take time. Driver profile cannot be generated overnight. As described earlier that learning will be carried from the characteristics executed by the driver and its effect on vehicle state. There will be a blind profiling stage in which no feedback will be provided to driver. Driver profile will be generated from the data collected in blind profiling stage over the period of time. Initially we will be concentrating on single driver for a single vehicle with static characteristics but the system can be extended to Multiuser single device along with dynamic characteristics. Every driver has unique pattern of driving which distinguish him from other drivers. As described earlier in system architecture that we will log the record for every driving session and forward that data on to remote server; remote server will acquire and store the data. When profiling system will process the data it will find the repetitions if there is any and if these pattern repeats beyond certain threshold system will raise the confidence and consider that pattern of that characteristic as part profile. Part profiles let us know which characteristic is performed and in what manner it is performed. So the collection of part profile helps to build the profile of driver [7].

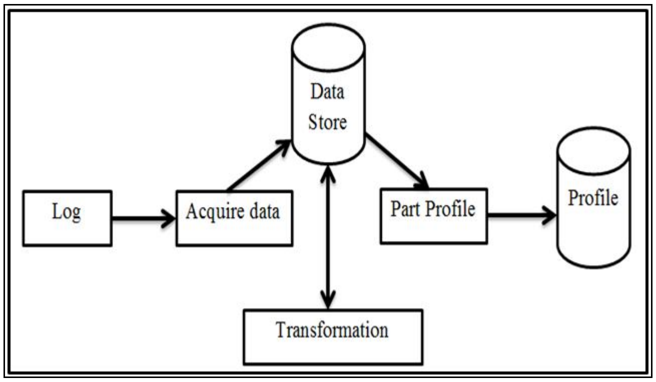

Figure 2: Remote Server

After blind profiling stage, profiles that are generated will be forwarded on to the vehicle. As we know that profile is the collection of part profiles and which is nothing but the characteristics or combination of them. So in every new driving session there will be comparison between characteristics of new session with the characteristics of profiles we have. Characteristics can be compared statically or dynamically. Dynamic characteristic is nothing but the collection of static characteristics. Say for example characteristics like change in gear can be taken into consideration individually whereas when driver is taking turn it can be seen as combination of various characteristics like steering wheel angle, steering wheel velocity, acceleration, breaking, speed so all these characteristics can be club together in single category say "turn". So in every new session there will be comparison between the characteristics of new session with that of characteristics of profiles stored on the vehicle. In comparator we can apply functions like cosine similarity [2] so as to find similarity between current driving sessions with that of the profiles we have so that any unusual behavior can be traced.

5. Structure

The generated profile will have to be structured and there are various ways to structure them.

5.1. Vector Space Model

Representation based on vector space model is one of the popular representations. It represents user profile in the n-dimensional space. Each element of vector is composed of a keyword and its weight. The weight can be taken as Boolean value or real value, which respectively represents for whether user is interested in the keyword and how much user is interested in the keyword. It appears as

$U = \{Key_1{:}Value_1, Weight1, Key_2{:}Value_2{:}Weight_2\ldots Key_n{:}\ Value_n{:}\ Weight_n\}$

The vector can usually be generated by collecting and training relevant data. The centroid-based classification method is used to deal with the relevant characteristics [2].

5.2. Ant Based and LCS

A hybrid method is proposed, which uses the ant-based clustering and LCS classification method to find and predict user’s behavior. In this paper, user profile created based on user pattern. Ant-based clustering approach is used to discover behavior patterns. The picking and dropping operations are biased by the similarity and density of data items within the ants’ local neighborhood, ants are likely to pick up data items that are either isolated or surrounded by dissimilar ones. They tend to drop them in the vicinity of similar ones. In this way, a clustering and sorting of the elements on the grid is obtained. The classification algorithm, Longest Common Subsequence (LCS), approach for discovering user behavior patterns using a graph partitioning model. The longest common subsequence (LCS) problem is to find the longest subsequence common to all sequences in a set of sequences [6].

5.3. K-means

It is a clustering algorithm and learning is carried in unsupervised manner. Aims to partition n observations into k clusters, in which each observation belongs to clusters with the nearest mean. [9].

5.4. ID3

Forms a decision tree which is a flow chart like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. It is a classification technique which is a two-step process. First there is a learning step then comes the classification. In this process whole focus is on finding splitting criterion that best partitions the tuples into individual classes [14].

6. Store and Access

These structures we will have to store somewhere such that it should be easier to store them and manipulate them. For that we have

6.1. Flat Files

A flat file database is a database that stores data in a plain text file [8]. Each line of the text file holds one record, with fields separated by delimiters, such as commas or tabs. While it uses a simple structure, a flat file database cannot contain multiple tables like a relational database can. Fortunately, most database programs such as Microsoft Access and FileMaker Pro can import flat file databases and use them in a larger relational database. Flat file is also a type of computer file system that stores all data in a single directory. There are no folders or paths used organize the data. In this project flat files can be used to store the real time data while driver is driving. Data for particular session will be stored in flat files which then will be sent to remote server.

6.2. Relational Database

A Relational database stores data in tables. The data stored in a table is organized into rows and columns. Each row in a table represents an individual record and each column represents a field. A record is an individual entry in the database. Field is a piece of information in a record.

There are various advantages of relational database and they are Data can be easily accessed. Data can be shared. Data storage and redundancy can be reduced. Data inconsistency can be avoided. Data Integrity can be maintained. Standards can be enforced. Security restrictions can be applied. Independence between physical storage and logical data design can be maintained. High-level data manipulation language (SQL) can be used to access and manipulate data [10].

6.3. XML

XML (Extensible Markup Language) is a flexible way to create common information formats and share both the format and the data on the World Wide Web, intranets, and elsewhere. XML uses human, not computer, language. XML is readable and understandable, even by novices, and no more difficult to code than HTML. XML is completely compatible with Java™ and 100% portable. Any application that can process XML can use your information, regardless of platform. It is extendable. The XML tag names are readable and convey the meaning of the data. The information structure is easily discerned by both humans and computers as each XML tag immediately precedes the associated data. The data structure follows a noticeable and useful pattern, making it easy to manipulate and exchange the data. [11].

6.4. CSV

The principle advantage is that CSV format can be read by any spreadsheet program. As it is a plain text file, it can also be read by word processor or simple notepad programs. Its advantages are it is human readable and easy to edit manually. It is simple to implement. It can be processed by almost all existing applications. It provides a straightforward information schema and is faster to handle. It is smaller in size and easy to generate [11].

7. Weights

Weight is part and parcel of user profiling. Every profile will have some weights. It's the weights which decides whether the user is authentic or an outlier. Weights will be assigned to each characteristics and the collection of characteristics. We are not deciding weight explicitly. Weights will be decided by taking driver behavior into consideration which characteristics he exhibits and how much he exhibits and this will make sure that each profile will have different weights though characteristics are same. So this makes every profile unique. In every session some characteristics may vary and thereby weight will vary but what matters is the overall weight of profile. If the weight matches with that of any stored profile then driver is authentic else he is an outlier. Driver's behavior is not permanent and can drift with time. So there might be some change in the behavior of authentic user. So for authentic user if particular characteristic varies over a period of time then using various feedback algorithms we can update the profile. This is where frequency factor will play the role.

8. Updating User Profile

8.1. Frequency Factor

With the passage of time the behavior of driver will change so our profile should be updated with the current behavior of driver. For that we will be take frequency factor into consideration. Frequency comprises of two parameters and they are frequency and regency. In this we will be focusing on the characteristic which are frequent and are recent too. So the characteristic which is more frequent and is recent will have more weight and the one which is not recent and is not frequent, the weight of that characteristic will be decreased. This is how we will be increasing or decreasing the weights of characteristics and thereby our profile will be updated with the current pattern of driver. There are various feedback algorithms which ensure that at any point in time driver profiles are consistent with the current behavior of driver. [3] talks about Rocchio algorithm which is the development of traditional vector space model theory and probability model theory. This algorithm updates user profile by taking positive feedback and negative feedback. To map this algorithm with driver's behavior we can divide driver's behavior into three categories cautious, moderate and aggressive on basis of some threshold where cautious and moderate comes under positive feedback and aggressive behavior if observed will be taken as negative feedback. So this is how through Rocchio[2] any profile can be continuously adjusted through feedback to become updated profile. [5] presents a methodology of consistency in which they also make sure that the profiles should be updated with current behavior by considering negative and positive response.[4] describes about the long term and short term behavior. Driver's behavior is not permanent and can drift with time. In this paper they describe various approaches to deal with behavior drifting and one such is forgetting mechanism. The main idea behind it is that natural forgetting is gradual process. The behavior which is observed often the characteristics related to that behavior their relevance should be increased whereas the behavior which is not frequent anymore its relevance should be decreased gradually [3] describes that profiles evolves over a period of time. The characteristics which are of importance earlier might not be of any or little interest and relevance of some characteristics increases over period of time. So the profile should remain updated by taking into consideration all changes in behavior. For this they have considered forgetting factor mechanism which makes sure that profiles describes the current behavior to its fullest.

9. Analysis of Outlier

Outlier has to be found in real time. So analysis of it should be carried out from the time when the driver starts the vehicle. Analysis will be carried out in the vehicle itself at comparator; no involvement of remote server is required. Analysis will be carried out in two steps and both these step will takes place simultaneously. These steps are Derivative and Composite

9.1. Derivative Approach

In this process every characteristic is taken into consideration individually. Every characteristic is compared with the characteristic earlier defined in profile and on that weights will be assigned.

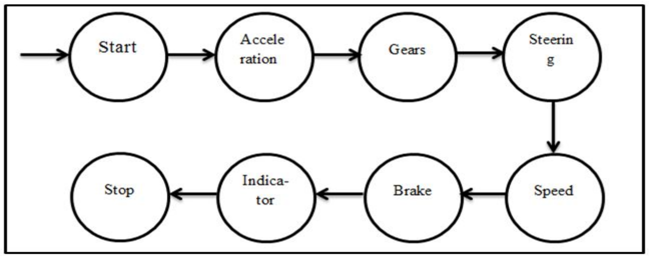

Figure 3: State Diagram

This approach can be explained with the help of an example. Say driver wants to go from one place to another and therefore will cover certain distance and on the way will have to take a turn to reach to destination. So driver starts his journey by starting his vehicle (every static characteristic is considered as and when encountered and real time comparison of these characteristics is carried out with the profiles we have stored on the vehicle). Say after starting the vehicle he accelerates then change gears and then moves steering wheel likewise he followed every state as described in (fig 3) state diagram. After starting vehicle he accelerates and his acceleration pattern varies from the profiles we have on vehicle then system will be on alert and will look for an outlier. Say gear shifting pattern also varies then system will raise the confidence for an outlier and now speed and steering pattern also varies then system will raise the confidence for outlier and if confidence found beyond threshold then system will confirmed driver as an outlier. This is how driver will shift from one state to another state and system will look for outlier and gradually raise the confidence in favor of outlier or against it.

9.2. Composite Approach

In this more than one characteristic is taken into consideration. This is the approach when we are considering dynamic characteristic and which is explained in learning module of this paper.

10. Reporting an Outlier

Say an outlier is found then what actions will the system take. As soon as the outlier is found a message will be forwarded to the authentic user and further action depends upon the type of response obtained from authentic user. In case of negative response no action will take place but in case of positive response there will be series of actions which system will take and they are

- There will be speed lock of the vehicle beyond that certain speed, speed of vehicle will not be increased.
- There will be auto decrement of gears. As soon as driver shift from higher gear to lower gear he cannot again shift into the higher gear.
- After certain amount of time the ignition system of vehicle will be disabled.

But there is question that instead of taking two steps and then disabling the ignition system in third place why can't we disable it in first place? Doing this involves a risk factor of accident. Say a driver is driving and is found as an outlier and confirmed by the authentic user after this if we disable the ignition system all of a sudden then vehicle will stop in the middle of the road and there are chances of having an accident if there is a vehicle following this vehicle. So we will follow above three steps and that too sequentially. Another thing we can do is that after the driver is confirmed as an outlier we will alert that driver that you are found as an outlier so we will be taking certain steps and in last we will be disabling the ignition system so before all this happens please take the vehicle on the side of road and park it otherwise the ignition system of the vehicle will be disabled anyhow.

11. Hypothesis

There is certain hypothesis that should be taken into consideration such that these typical behaviors should not be treated as an outlier. The hypothesis like

- The change in driving pattern of the same authentic user is possible so he should not be treated as an outlier. In weekdays authentic driver used to follow certain path from home to office and now he is planning to go for trip somewhere in weekends. So there is possibility that the behavior of authentic driver might differ from the regular behavior he used to show on weekdays like the today's driving speed is more than the normal speed, today he has covered more distance than the normal and likewise some of the characteristics might vary but on that basis he should not be treated as an outlier. His other characteristics should be taken into consideration like the way he change gears, his breaking pattern and many more which makes him the authentic driver.
- Say if someone known to the authentic driver but unknown to the system wants to drive the vehicle. But when he will drive he will be treated as an outlier. So in this scenario authentic driver will tell the system explicitly that not to look for outlier in this particular session by sending a message to the system.
- Another scenario could be that in case of emergency in the house; a member of the family needs to be hospitalized as soon as possible so in this case also there is huge possibility that the behavior of authentic driver might be different. So if there is change in behavior outlier will be detected and reported to the authentic driver. In this scenario what authentic driver can do is that he can neglect the outlier by sending negative response to the system or another thing he can do is that he will tell the system explicitly that not to look for outlier for this session.

12. Applications

- ***Security***
  No doubt by tracing the unauthentic behavior and conforming that as an outlier and then following certain measures as described earlier we can provide security against automobile theft.
- ***Recommendation***
  If we are having more than one profile then we can check for that profile in which the performance of system is maximum and that profile can be recommended to the other user such that overall performance of the system can be increased. We can also give recommendation to the driver by taking into consideration Speed, Time, Alternative Routes of the Multiuser.
- ***Insurance***
  Insurance premium can be decided by taking into consideration driver behavior like his average speed, what safety parameters driver follows like whether he gives indicator while taking turn, whether he takes sharp turn or smooth turn, the way he applies brake whether it is hard braking or soft braking, how he accelerates and many more. So if the driving skills of driver are smooth he will have to pay fewer premiums whereas the one whose driving skills are found to be risky will have to pay more premiums.
- ***Tracking the vehicle***
  We can track the location from the distance travelled and the initial turns taken. Rather than searching in all the direction we can predict that where there is the possibility of vehicle to be found.
- ***Safety Alert***
  We can calculate risk involvement from drivers driving behavior and can alert him about the measures he should follow for a safe ride.

13. Conclusion and Future Scope

From all this we can conclude that by profiling the automobile driver and learning his driving pattern we can detect the outlier. Detection of outlier will be carried out at run time. This can be useful to check automobile theft to a great extent. Moreover if we share our profile with insurance company it will be useful to decide the premium for driver as the company will come to know whether the driver driving is a smooth driver or the rough one. Other than that we can track our vehicle and also we can get recommendation so as to increase the performance of vehicle.

Our future work includes working with both static and dynamic characteristics. We will also consider more than one part profile of same characteristics and will consider combination like vehicle data with static characteristic, vehicle data with dynamic characteristics, combination of all the three for profiling automobile driver.

14. References

1. Tsippy Lotan. ”An In-Vehicle Data Recorder for Evaluation of Driving Behavior and Safety”, In TRB 2006 Annual Meeting.
2. Xu Qi. “Research on User Profiling Technology for Personalized Demands” In Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on (Volume:3), pp. 198 – 201, 11-12 May 2010.
3. Wei Wang, Dongyan Zhao, Haining Luo, Xin Wang. “Mining User Interests in Web Logs of an Online News Service Based on Memory Model”, Published in: Networking, Architecture and Storage (NAS), 2013 IEEE Eighth International Conference, pp. 151- 155, 17-19 July 2013.
4. Jie Yu, Fangfang Liu. “Mining User Context based on Interactive Computing for Personalized Web Search”, Published in: Computer Engineering and Technology (ICCET), 2010 2nd International Conference on (Volume: 2), pp. 209-214, 16-18 April 2010.
5. Muhammad Ali Zeb, Maria Fasli. “Interest Aware Recommendations based on Adaptive User Profiling”, Published in: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on (Volume:3) pp. 357-360 22-27 Aug. 2011
6. J. Jojo, Akshaya, Sugana. “User profile creation based on navigation pattern for modeling user behaviour with personalised search”, Published in: Current Trends in Engineering and Technology (ICCTET), IEEE 2013 International Conference on pp. 371-374 3-3 July 2013 Coimbatore, India.
7. N. Prasanna Balaji, Chenapaga Ravi, P. Krishna Prasad and V. Chandra Prakash. “Data Mining Techniques and Analysis of Concept Based User Profiles from Search Engine Logs”, International Journal of Computer Science and Telecommunications [Volume 2, Issue 7, October 2011]
8. http://docs.oracle.com/cd/E17984_01/doc.898/e14711/flat_files.htm
9. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu. "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, july 2002.
10. Renjie Zhang, Yuling Zhao. "Research on access of relational database", Computer Engineering and Technology (ICCET), 2010 2nd International Conference on (Volume:7), pp. V7-738 - V7-741, 16-18 April 2010, Chengdu
11. http://www.w3.org
12. http://www.iii.org/issue-update/auto-theft
13. http://timesofindia.indiatimes.com/city/delhi/40-vehicles-stolen-in-Delhi-every-day/articleshow/14619149.cms.
14. Data Mining: Concepts and Techniques 2nd Edition Jiawei Han and Micheline Kamber The University of Illinois at Urbana-Champaign