DATA MINING TOOLS AND TECHNIQUES

Understanding the advantages of using different data mining tools and techniques

Vishal Sr. Programmer
CyberQ Consulting Pvt. Ltd.
New Delhi India-110025

ABSTRACT- Data mining is one of the most applicable areas of research in computer applications among the various types of data mining. This paper is going to focus on web mining. This is the review paper which shows deep and intense study of various techniques available for web mining Tools and Techniques. Web mining - i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage - is the collection of technologies to fulfill this potential. Above definition of web mining is explored in this paper.

INTRODUCTION-

Most internal auditors, especially those working in customer-focused industries, are aware of data mining and what it can do for an organization — reduce the cost of acquiring new customers and improve the sales rate of new products and services. However, whether you are a beginner internal auditor or a seasoned veteran looking for a refresher, gaining a clear understanding of what data mining does and the different data mining tools and techniques available for use can improve audit activities and business operations across the board.

WHAT IS DATA MINING?

In its simplest form, **Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends**. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

Data mining is not particularly new — statisticians have used similar manual approaches to review data and provide business projections for many years. Changes in data mining techniques, however, have enabled organizations to collect, analyze, and access data in new ways. The first change occurred in the area of basic data collection. Before companies made the transition from ledgers and other paper-based records to computer-based systems, managers had to wait for staff to put the pieces together to know how well the business was performing or how current performance periods compared with previous periods. As companies started collecting and saving basic data in computers, they were able to start answering detailed questions quicker and with more ease.

Changes in data access — where there has been greater empowerment and integration, particularly over the past 30 years — also have impacted data mining techniques. The introduction of microcomputers and networks, and the evolution of middleware, protocols, and other methodologies that enable data to be moved seamlessly among programs and other machines, allowed companies to link certain data questions together. The development of data warehousing and decision support systems, for instance, has enabled companies to extend queries from "What was the total number of sales in New South Wales last April?" to "What is likely to happen to sales in Sydney next month, and why?"

However, the major difference between previous and current data mining efforts is that organizations now have more information at their disposal. Given the vast amounts of information that companies collect, it is not uncommon for them to use data mining programs that investigate data trends and process large volumes of data quickly. Users can determine the outcome of the data analysis by the parameters they chose, thus providing additional value to business strategies and initiatives. It is important to note that without these parameters, the data mining program will generate all permutations or combinations irrespective of their relevance.

Internal auditors need to pay attention to this last point: Because data mining programs lack the human intuition to recognize the difference between a relevant and an irrelevant data correlation, users need to review the results of mining exercises to ensure results provide needed information. For example, knowing that people who default on loans usually give a false address might be relevant, whereas knowing they have blue eyes might be irrelevant. Auditors, therefore, should monitor whether sensible and rational decisions are made on the basis of data mining exercises, especially where the results of such exercises are used as input for other processes or systems.

Auditors also need to consider the different security aspects of data mining programs and processes. A data mining exercise might reveal important customer information that could be exploited by an outsider who hacks into the rival organization's computer system and uses a data mining tool on captured information.

DATA MINING TOOLS Organizations that wish to use data mining tools can purchase mining programs designed for existing software and hardware platforms, which can be integrated into new products and systems as they are brought online, or they can build their own custom mining solution. For instance, feeding the output of a data mining exercise into another computer system, such as a neural network, is quite common and can give the mined data more value. This is because the data mining tool gathers the data, while the second program (e.g., the neural network) makes decisions based on the data collected.

Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs. This should be considered as early as possible in the project's lifecycle, perhaps even in the feasibility study.

Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools. Below is a description of each.

- ***Traditional Data Mining Tools.*** Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.
- ***Dashboards.*** Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.
- ***Text-mining Tools.*** The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

Besides these tools, other applications and programs may be used for data mining purposes. For instance, audit interrogation tools can be used to highlight fraud, data anomalies, and patterns. An example of this has been published by the India's Treasury office in the *2002–2003 Fraud Report: Anti-fraud Advice and Guidance*, which discusses how to discover fraud using an audit interrogation tool. Additional examples of using audit interrogation tools to identify fraud are found in Ambrish Jha's 2012 book, *Fraud Detection*.

In addition, internal auditors can use spreadsheets to undertake simple data mining exercises or to produce summary tables. Some of the desktop, notebook, and server computers that run operating systems such as Windows, Linux, and Macintosh can be imported directly into Microsoft Excel. Using pivotal tables in the spreadsheet, auditors can review complex data in a simplified format and drill down where necessary to find the underlining assumptions or information.

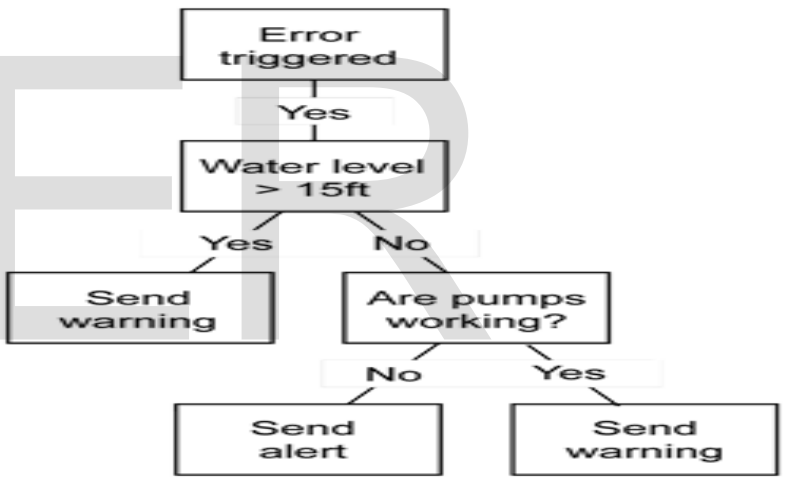
When evaluating data mining strategies, companies may decide to acquire several tools for specific purposes, rather than purchasing one tool that meets all needs. Although acquiring several tools is not a mainstream approach, a company may choose to do so if, for example, it installs a dashboard to keep managers informed on business matters, a full data-mining suite to capture and build data for its marketing and sales arms, and an interrogation tool so auditors can identify fraud activity.

DATA MINING TECHNIQUES AND THEIR APPLICATION

In addition to using a particular data mining tool, internal auditors can choose from a variety of data mining techniques. The most commonly used techniques include artificial neural networks, decision trees, and the nearest-neighbor method. Each of these techniques analyzes data in different ways:

- ***Artificial neural networks*** are non-linear, predictive models that learn through training. Although they are powerful predictive modeling techniques, some of the power comes at the expense of ease of use and deployment. One area where auditors can easily use them is when reviewing records to identify fraud and fraud-like actions. Because of their complexity, they are better employed in situations where they can be used and reused, such as reviewing credit card transactions every month to check for anomalies.
- ***Decision trees*** are tree-shaped structures that represent decision sets. These decisions generate rules, which then are used to classify data. Decision trees are the favored technique for building understandable models. Auditors can use them to assess, for example, whether the organization is using an appropriate cost-effective marketing strategy that is based on the assigned value of the customer, such as profit.

Figure 1. Decision tree



- ***The nearest-neighbor method*** classifies dataset records based on similar data in a historical dataset. Auditors can use this approach to define a document that is interesting to them and ask the system to search for similar items.

Each of these approaches brings different advantages and disadvantages that need to be considered prior to their use. Neural networks, which are difficult to implement, require all input and resultant output to be expressed numerically, thus needing some sort of interpretation depending on the nature of the data-mining exercise. The decision tree technique is the most commonly used methodology, because it is simple and straightforward to implement. Finally, the nearest-neighbor method relies more on linking similar items and, therefore, works better for extrapolation rather than predictive enquiries.

A good way to apply advanced data mining techniques is to have a flexible and interactive data mining tool that is fully integrated with a database or data warehouse. Using a tool that operates outside of the database or data warehouse is not as efficient. Using such a tool will involve extra steps to extract, import, and analyze the data. When a data mining tool is integrated with the data warehouse, it simplifies the application and implementation of mining results. Furthermore, as the warehouse grows with new decisions and results, the

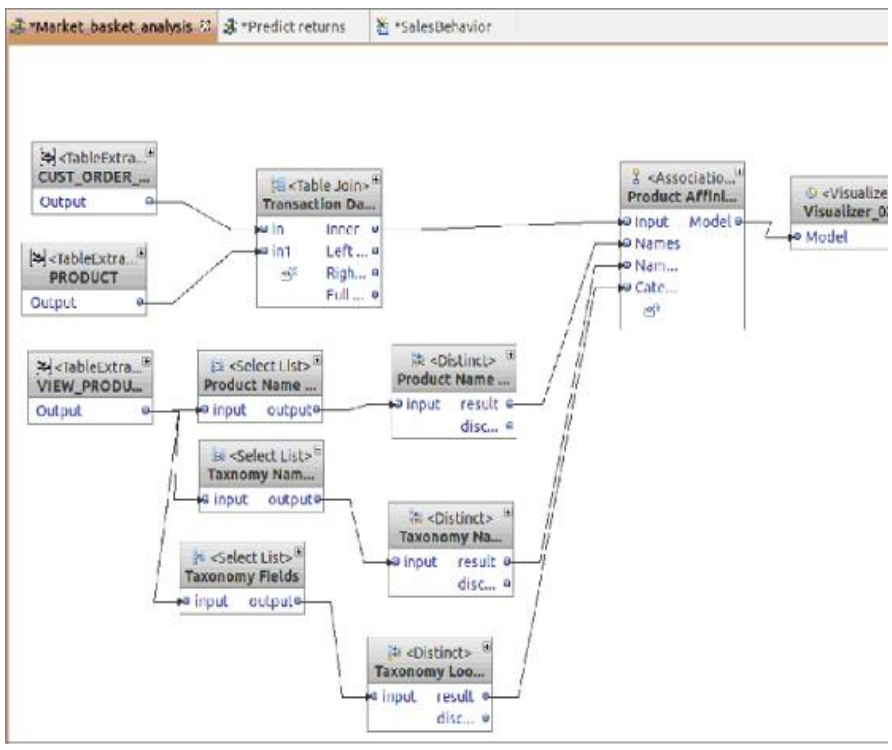
organization can mine best practices continually and apply them to future decisions.

Regardless of the technique used, the real value behind data mining is modeling — the process of building a model based on user-specified criteria from already captured data. Once a model is built, it can be used in similar situations where an answer is not known. For example, an organization looking to acquire new customers can create a model of its ideal customer that is based on existing data captured from people who previously purchased the product. The model then is used to query data on prospective customers to see if they match the profile. Modeling also can be used in audit departments to predict the number of auditors required to undertake an audit plan based on previous attempts and similar work.

- **Association-**Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, you make a simple correlation between two or more items, often of the same type to identify patterns. For example, when tracking people's buying habits, you might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.

Building association or relation-based data mining tools can be achieved simply with different tools. For example, within Info Sphere Warehouse a wizard provides configurations of an information flow that is used in association by examining your database input source, decision basis, and output information. Figure 2 shows an example from the sample database.

Figure 2. Information flow that is used in association

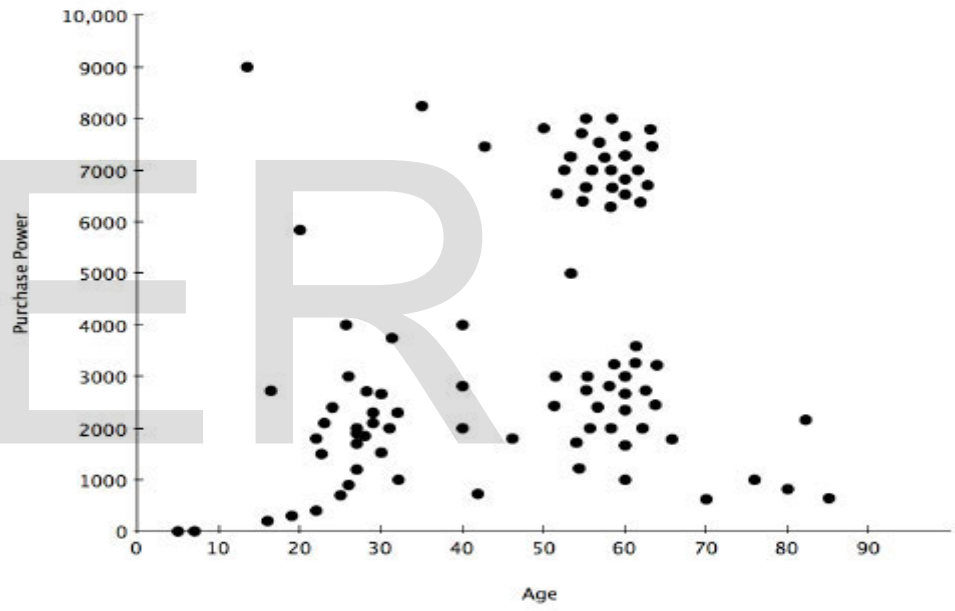


- **Classification** You can use classification to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, you can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, you might apply it into a particular class by comparing the attributes with our known definition. You can apply the same principles to customers, for example by classifying them by age and social group.

Additionally, you can use classification as a feeder to, or the result of, other techniques. For example, you can use decision trees to determine a classification. Clustering allows you to use common attributes in different classifications to identify clusters.

- **Clustering** By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree.
- Clustering can work both ways. You can assume that there is a cluster at a certain point and then use our identification criteria to see if you are correct. The graph in Figure 3 shows a good example. In this example, a sample of sales data compares the age of the customer to the size of the sale. It is not unreasonable to expect that people in their twenties (before marriage and kids), fifties, and sixties (when the children have left home), have more disposable income.

Figure 3. Clustering



In the example, we can identify two clusters, one around the Rs 2,000/20-30 age group, and another at the Rs 7,000-8,000/50-65 age group. In this case, we've both hypothesized and proved our hypothesis with a simple graph that we can create using any suitable graphing software for a quick manual view. More complex determinations require a full analytical package, especially if you want to automatically base decisions on *nearest neighbor* information.

Plotting clustering in this way is a simplified example of so called *nearest neighbor* identity. You can identify individual customers by their literal proximity to each other on the graph. It's highly likely that customers in the same cluster also share other attributes and you can use that expectation to help drive, classify, and otherwise analyze other people from your data set. You can also apply clustering from the opposite perspective; given certain input attributes, you can identify different artifacts. For example, a recent study of 4-digit PIN numbers found clusters between the digits in ranges 1-12 and 1-31 for the first and second pairs. By plotting these pairs, you can identify and determine clusters to relate to dates (birthdays, anniversaries).

- **Prediction** is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other data mining techniques, prediction involves analyzing trends, classification, pattern matching, and relation. By analyzing past events or instances, you can make a prediction about an event. Using the credit card authorization, for example, you might combine decision tree analysis of individual past transactions with classification and historical pattern matches to identify whether a transaction is fraudulent. Making a match between the purchase of flights to the US and transactions in the US, it is likely that the transaction is valid.

- **Sequential patterns** often used over longer-term data, sequential patterns are a useful method for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps drive the structure of the decision tree and the output.

- **Combinations** In practice, it's very rare that you would use one of these exclusively. Classification and clustering are similar techniques. By using clustering to identify nearest neighbors, you can further refine your classifications. Often, we use decision trees to help build and identify classifications that we can track for a longer period to identify sequences and patterns.

- **Long-term (memory) processing** Within all of the core methods, there is often reason to record and learn from the information. In some techniques, it is entirely obvious. For example, with sequential patterns and predictive learning you look back at data from multiple sources and instances of information to build a pattern.

In others, the process might be more explicit. Decision trees are rarely built one time and are never forgotten. As new information, events, and data points are identified, it might be necessary to build more branches, or even entirely new trees, to cope with the additional information.

You can automate some of this process. For example, building a predictive model for identifying credit card fraud is about building probabilities that you can use for the current transaction, and then updating that model with the new (approved) transaction. This information is then recorded so that the decision can be made quickly the next time.

MOVING FORWARD

Using data mining to understand and extrapolate data and information can reduce the chances of fraud, improve audit reactions to potential business changes, and ensure that risks are managed in a more timely and proactive fashion. Auditors also can use data mining tools to model "what-if" situations and demonstrate real and probable effects to management, such as combining real-world and business information to show the effects of a security breach and the impact of losing a key customer. If data mining can be used by one part of the organization to influence business direction for profit, why can't internal auditors use the same tools and techniques to reduce risks and increase audit benefits?

EFFECTIVE WAY TO OPTIMIZE DATA MINING

With enterprises operating out of multiple geographic locations, multi-database mining is becoming important for effective and informed decision making. The following mining techniques will help you optimize your data mining efforts.

Step 1: Handling of incomplete data

Incomplete data affects classification accuracy and hinders effective data mining. The following techniques are effective for working with incomplete data.

1. The ISOM-DH model handles incomplete data using independent component analysis (ICA) and self-organizing maps (SOM). It uses existing data to estimate the missing data and visualize the handled high-dimensional data.
2. Another data mining technique is based on the evolution of strategies built using parametric and non-parametric imputation methods. Genetic algorithms and multilayer perceptron's have to be applied to develop a framework to construct imputation strategies which address multiple incomplete attributes.
3. Network approaches based on multi-task learning (MTL): the learning of a problem/instance in relation to others) for pattern classification, with missing inputs, can be compared with representative procedures used for handling incomplete data on two well-known data sets.

Step 2: Ensure efficiency and scalability of data mining algorithms

A great deal of expertise and effort is currently required for the implementation, maintenance, and performance-tuning of a parallel data mining application. These data mining techniques can help:

1. Ensure parallel and scalable execution of data mining algorithms.
2. Grid-enable data mining applications without any intervention on the application side.
3. Opt for scalable data mining instead of mere associations when mining market basket data.
4. Remove barriers to the widespread adoption of support vector machines.

Step 3: Mining of large databases

It's a good data mining technique to combine set architectural alternatives for coupling mining with database systems. Such data mining techniques could include:

1. Encapsulation of the data mining algorithm in a stored procedure.

2. Caching the data to a file system on the fly, then mining.
3. Tight-coupling, primarily with user-defined functions.
4. SQL implementations for processing in the DBMS.

Step 4: Handling of relational and complex data types

It's critical to develop a system to support the interactive mining of multiple-level knowledge in large relational databases and data warehouses. This requires tight integration of online analytical processing (OLAP) with a wide spectrum of data mining functions including characterization, association, classification, prediction, and clustering. The system should facilitate query-based, interactive mining of multidimensional databases by implementing a set of advanced data mining techniques including:

- OLAP-based induction
- Multidimensional statistical analysis
- Progressive deepening for data mining refined knowledge
- Meta-rule guided mining, and data and knowledge visualization
- Assessing data mining results via swap randomization
- Analyzing graph databases by aggregate queries
- Image classification using sub-graph histogram representation
- A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems

Step 5: Data mining techniques for heterogeneous databases

Heterogeneous database systems play a vital role in the information industry in 2011. Data warehouses must support data extraction from multiple databases to keep up with the trend.

For example, three heterogeneous data mining programs are needed to model the behavior of telecom organizations. First, the client's attribute weight is calculated from original data using the neural network method. Then, based on attribute weight, exceptional client characteristics are identified using the decision tree method. Finally, the distinguishing model is generated adaptively on the basis of clustering. The combination of three algorithms helps effective distinguishing of exceptional client.

CONCLUTIONS

In this paper, we have discussed some web data mining research issues in context of the **A Study on Web Data Mining** project at **SHRI VENKATESHWARA UNIVERSITY, Gajraula, Amroha (UTTAR PRADESH).** We have defined data mining Tool and Techniques. In particular, we discussed web data mining with respect to Tool and Techniques. An important part of our warehousing project is to design for web data mining to generate some useful knowledge from the WWW data. Currently we are exploring the ideas discussed in this paper.

REFERENCES

1. H. Vernon Leighton and J. Srivastava. Precision Among WWW Search Services (Search
Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. http://www.winona.msus.edu/is-f/libraryf/ webind2/webind2.htm, 1997.
2. R. Cooley, B. Mobasher and J. Srivsatava. Web Mining: Information and Pattern Discovery on the Word Wide Web. Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.
3. J. Han, Yue Huang, et al. Intelligent Query Answering by Knowledge Discovery Techniques, IEEE TKDE, 1996.
4. S. K. Madria, M. Mohnia, J. Roddick. Query Processing in Mobile Databases Using Concept Hierarchy and Summary Database. In proceedings of 5th International Conference on Foundation of Data Organization, Japan, Nov. 1998. 16
5. Sourav S. Bhowmick, S. K. Madria, W.-K. Ng, E.-P. Lim, Web Bags : Are They Useful in Web
warehouse? In proceedings for 5th International Conference on Foundation of Data Organization, Japan, Nov. 1998.
6. T. Bray, Measuring the Web. In Proceedings of the 5th Intl. WWW Conference, Paris, France, 1996.
7. Wee-Keong Ng, Ee-Peng Lim, Chee-Thong Huang, Sourav Bhowmick, Fengqiong Qin. Web
Warehousing : An Algebra for Web Information. In Proceedings of the IEEE Advances in Digital
Libraries Conference, Santa Barbara, U.S.A., April 1998.
8. Shian-Hua Lin, Chi-Sheng Shih, Meng Chang Chen, et al. Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach. In Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.
9. Sourav S. Bhowmick, W.-K. Ng, E.-P. Lim. Information Coupling in Web Databases. In Proceedings of the 17th International Conference on Conceptual Modelling(ER'98), Singapore, November 16-19, 1998.
10. D. Backman and J. Rubbin, Web log analysis: Finding a Recipe for Success. http://techweb.comp. com/nc/811/811cn2.html, 1997.
11. M.S. Chen, J. Han and P.S. Yu, Data Mining: An Overview from a Database Perspective. IEEE
Transaction on Knowledge and Data Engineering, 8:866-833, 1996.
12. J. Han and Y. Fu. Discovery of Multi-level Association Rules. In Proceedings of International
Conference on Very Large Databases, pages 420-431, Zurich, Switzerland, Sept. 1995.

Vishal has obtained his **MCA** degree in 2006 from **Uttar Pradesh Technical University**. [U.P] INDIA .His research interest is Information technology. Recently he is doing research from **Shri** Venkateshwara **University** Gajraula, Amroha, (U.P.)-244236, India

Dr. Saurabh Gupta has obtained his Ph.D. Degree from Lucknow University in the year 1999. He has completed his Ph.D. in the area of Computer Engineering. His research interests are, new innovation in e-Governance etc. He has published many of the valuable research papers in various national and international Journals. He is presently working as a State Informatics Officer in National Informatics Centre Shimla (Himachal Pradesh), India.