Data Security and Privacy in Data Mining: Research Issues & Preparation

Dileep Kumar Singh[#1], Vishnu Swaroop[*2]
[#]IT Resource Centre
Madan Mohan Malaviya Engineering College,
Gorakhpur, India
[*]Dept. of Computer Science & Engineering,
Madan Mohan Malaviya Engineering College
Gorakhpur, India

Abstract— **Database mining can be defined as the process of mining for implicit, formerly unidentified, and potentially essential information from awfully huge databases by efficient knowledge discovery techniques. The privacy and security of user information have become significant public policy anxieties and these anxieties are receiving increased interest by the both public and government lawmaker and controller, privacy advocates, and the media. In this paper we focuses on key online privacy and security issues and concerns, the role of self-regulation and the user on privacy and security protections, data protection laws, regulatory trends, and the outlook for privacy and security legislation. Naturally such a process may open up new assumption dimensions, detect new invasion patterns, and raises new data security problems. Recent developments in information technology have enabled collection and processing of enormous amount of personal data, such as criminal records, online shopping habits, online banking, credit and medical history, and driving records and almost importantly the government concerned data**

Index Terms—**Database mining, Database security, Data Privacy, Inferences, Intrusion Detection, Law.**

I. Introduction

Security and Privacy protection have been a public policy concern for decades. However, rapid technological changes, the rapid growth of the internet and electronic commerce, and the development of more sophisticated methods of collecting, analyzing, and using personal information have made privacy a major public and government issues. The field of data mining is gaining significance recognition to the availability of large amounts of data, easily collected and stored via computer systems. Recently, the large amount of data, gathered from various channels, contains much personal information. When personal and sensitive data are published and/or analyzed, one important question to take into account is whether the analysis violates the privacy of individuals whose data is referred to. The importance of 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 privacy is growing constantly. For this reason, many research works have focused on privacy-preserving data mining, proposing novel techniques that allow extracting knowledge while trying to protect the privacy of users. Some of these approaches aim at individual privacy while others aim at corporate privacy.

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Knowledge discovery is needed to make sense and use of data. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. [1,2,3]

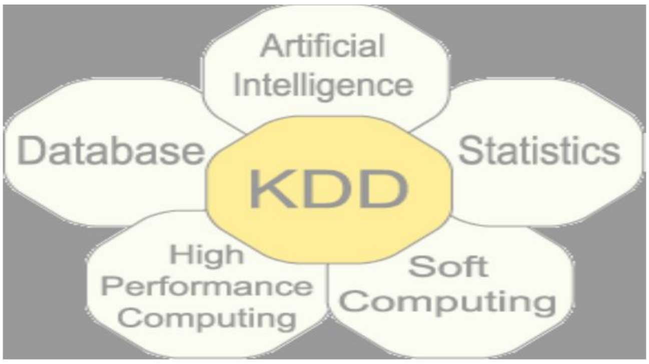

Figure -1

Usually, data mining e.g. data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information 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.[4] Although data mining is a comparatively new term but the technology is not. Companies have used powerful computers to filter through volumes of superstore scanner data and analyze market research reports for many years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.[5] Data mining, the discovery of new and interesting patterns in large datasets, is an exploding field. One aspect is the use of data mining to improve security, e.g., for intrusion detection. A second aspect is the potential security hazards posed when an adversary has data mining capabilities. Privacy issues have attracted the attention of the media, politicians, government agencies, businesses, and privacy advocates.

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of more than collecting, organizing and managing data; it also includes analysis and prediction. Data mining can be performed on data represented in quantitative, textual, graphical, image or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association sequence or path analysis, classification, clustering, and forecasting. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. The databases and data warehouses become more and more popular and imply huge amount of data which need to be efficiently analyzed. Knowledge Discovery in Databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases.[6,7]

The data mining database may be a logical rather than a physical subset of your data warehouse, provided that the data warehouse DBMS can support the additional resource demands of data mining. If it cannot, then you will be better off with a separate data mining database. [8]

II. What is Data Mining?

Data mining is an iterative and interactive process of discovering something innovative. The same as Novel- something we are not aware, Valid- generalize the future, Useful- some reaction is possible, Understandable- leading to insight, many step and process. Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." There are other definitions:

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner".[9]

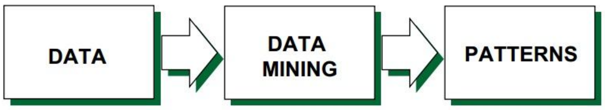

Figure -2

Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases".[10]

Evolution of database technology, data collection, database creation, IMS and Network DBMS, relational data model, Relational DBMS, advance database Models object oriented database, data collection centre, warehousing, multimedia database and recent web database needs to process the approach of data mining.

III. Architecture of Data Mining

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses…etc. This knowledge contributes a lot of benefits to business strategies, scientific, medical research, governments and individual.

The architecture contains modules for secure safe-thread communication, database connectivity, organized data management and efficient data analysis for generating global mining model.[11]

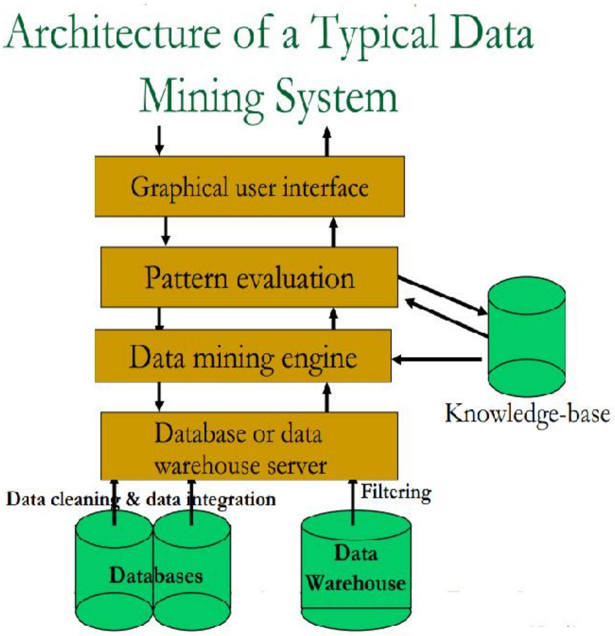
Architecture of a Typical Data Mining System
Graphical user interface
Pattern evaluation
Data mining engine
Knowledge-base
Database or data warehouse server
Data cleaning & data integration
Filtering
Databases
Data Warehouse

Figure -3

IV. Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. In some cases, users may have no idea of which kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel.

- Concept description – characterization and discrimination
- Association – correlation and causality
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Trend and Evolution Analysis
- Other Pattern – direct or statistical analysis

In above first two functionalities involves first generalize, summarize and contrast data characteristics second association, multi-dimensional vs single-dimensional association. Next two functionalities that is classification and prediction finding models that describe and distinguish classes or concepts for future prediction i.e. classify countries based on climate or classify cars based on gas mileage, presentation means decision-tree, classification rule, neural network and cluster analysis like class label is unknown – group data to form new classes, clustering based on the principle i.e. maximizing the intra-class similarity and minimizing the interclass similarity. Last three functionalities one is outlier analysis i.e. a data object that does not comply with the general behavior of the data, It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis second is trend and evolution analysis i.e. trend and deviation by regression analysis, sequential pattern mining, periodicity analysis and similarity based analysis and last includes all other type of pattern-directed or statistical analysis.[12]

V. Data Security Issues

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive.[13]

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.[14,15,16]

VI. Major Issues in Data Mining

A. Mining Methodology and User Interaction

- Mining different kinds of knowledge in database
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data Mining query language and ad-hoc data mining
- Expression and visualization of data mining results
- Handling noise and incomplete data
- Pattern evaluation

B. Performance and Scalability

- Efficiency and scalability of data mining algorithms
- Parallel, distributed and incremental mining methods

C. Issues Relating to the diversity of Data Type

- Handling relational and complex types of data
- Mining information from heterogeneous databases and global information systems like web database.

D. Issues Related to Applications and Social Impacts

- Application of discovered knowledge, domain specific data mining tools, intelligent query answering, decision making

E. Mining methodology and user interaction issues

- Mining different kinds of knowledge in databases
  - Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis .
  - These tasks may use the same database in different ways and require the development of numerous data mining techniques.

- Interactive mining of knowledge at multiple levels of abstraction
  - Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
  - User can interact with the data mining system to view data and discovered patterns at multiple granularities and from different angles.
- Incorporation of background knowledge:
  - Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.
- Data mining query languages and ad hoc data mining:
  - Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval.
  - High-level data mining query languages need to be developed to allow users to describe ad hoc data mining tasks by facilitating the specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered pattern.
- Presentation and visualization of data mining results:
  - Discovered knowledge should be expressed in high-level languages, visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.
- Handling noisy or incomplete data:
  - Data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases.
- Pattern evaluation—the interestingness problem:
  - Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns, particularly with regard to subjective measures that estimate the value of patterns with respect to a given user class, based on user beliefs or expectations.

VII. GOAL OF DATA MINING

Prediction example that given sales-purchase system recording of the goods from previous years one can predict what amount of goods that need to have in stock for the forthcoming season? Verification that one can check how a disease likes any viral is related to environmental situation? Exception detection that, is it possible to identify credit/debit transactions that are in fact frauds?

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment. Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together. Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor.

- Prediction – To foresee the possible future situation on the basis of previous events.
- Description – What is the reason that some events occur?
- Verification – We think that some relationship between entities occurs.
- Exception Detection – There may be situations (records) in the databases that correspond to something unusual.

VIII. USER INFLUENCE ON THE PROTECTION OF PERSONAL PRIVACY AND SECURITY

Consumers are increasingly aware that ubiquitous, more powerful computers and widespread access to the Internet make it easier for legitimate and shady businesses as well as government agencies to collect, access, and use personal information. Consequently, consumers have become more assertive in demanding that their personal information be protected and that they be given greater control over the collection and use of such information. Such activism has caused businesses and governments to change their procedures or modify their products. The Internet will continue to shift market power toward consumers, who can decide how much they want to pay for what they want to buy, and let sellers compete for their business. Electronic commerce enables companies to customize their products and services to suit the individual consumer. To meet the specific preferences of individuals, companies will have to tailor their marketing based on consumers' personal information about their shopping habits, likes and dislikes, as well as demographic and other characteristics. Such an exchange of information raises potential privacy and security concerns

IX. PREPERATION IN DATA MINING

As data mining initiatives continue to evolve, there are several issues Congress may decide to consider related to implementation and oversight. These issues include, but are not limited to, data quality, interoperability, mission creep, and privacy, [17] As with other aspects of data mining, while technological capabilities are important, other factors also influence the success of a project's outcome. We generate an enormous amount of data as a by-product of our everyday transactions (purchasing goods, enrolling for courses, etc.), visits to Web sites and interactions with government (taxes, census, car registration, voter registration, etc.). Not only is the number of records we generate increasing, but the amount of data gathered for each type of record is increasing.

A. Data Quality

Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data. To improve data quality, it is sometimes necessary to "clean" the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database.

B. Data Mining Application Areas

There are many areas of data mining application in most popular are Science (astronomy, bioinformatics, drug discovery), Business (advertising, customer relationship management, investment, manufacturing, entertainment, telecom, e-commerce, banking, marketing, health), web (serach engines, bots), government (law enforcement, proofing tax chater, anti-terror).

C. Interoperability

Related to data quality, is the issue of interoperability of different databases and data mining software. Interoperability refers to the ability of a computer system and/or data to work with other systems or data using common standards or processes. Interoperability is a critical part of the larger efforts to improve interagency collaboration and information sharing through e-government and homeland security initiatives. For data mining, interoperability of databases and software is important to enable the search and analysis of multiple databases simultaneously, and to help ensure the compatibility of data mining activities of different agencies. Data mining projects that are trying to take advantage of existing legacy databases or that are initiating first-time collaborative efforts with other agencies or levels of government may experience interoperability problems. Similarly, as agencies move forward with the creation of new databases and information sharing efforts, they will need to address interoperability issues during their planning stages to better ensure the effectiveness of their data mining projects.

D. Privacy

As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy. Concerns about privacy focus both on actual projects proposed, as well as concerns about the potential for data mining applications to be expanded beyond their original purposes. For example, some experts suggest that anti-terrorism data mining applications might also be useful for combating other types of crime as well.[18] So far there has been little consensus about how data mining should be carried out, with several competing points of view being debated. Some observers contend that tradeoffs may need to be made regarding privacy to ensure security. Other observers suggest that existing laws and regulations regarding privacy protections are adequate, and that these initiatives do not pose any threats to privacy. Still other observers argue that not enough is known about how data mining projects will be carried out, and that greater oversight is needed. There is also some disagreement over how privacy concerns should be addressed. Some observers suggest that technical solutions are adequate initiatives.[19,20]

Data mining has attracted significant interest especially in the past decade with its vast domain of applications. From the security perspective, data mining has been shown to be beneficial in confronting various types of attacks to computer systems. However, the same technology can be used to create potential security hazards. In addition to that, data collection and analysis efforts by government agencies and businesses raised fears about privacy, which motivated the privacy preserving data mining research. One aspect of privacy preserving data mining is that, we should be able to apply data mining algorithms without observing the confidential data values. [21,22] This challenging task is still being investigated. Another aspect is that, using data mining technology an adversary could access confidential information that could not be reached through querying tools jeopardizing the privacy of individuals. Some initial research results in privacy preserving data mining have been published. However, there are still many issues that need further investigation in the context of data mining from both privacy and security perspectives. This workshop aims to provide a meeting place for academicians to identify problems related to all aspects of privacy and security issues in data mining together with possible solutions. Researchers and practitioners working in data mining, databases, data security, and statistics are invited to submit their experience, and/or research results.

E. Laws and Regulations

The legal and policy foundation for data mining is based on the some specified protocols, which established penalization for data security and privacy Government Act, which requires consequence to provide a level of security for data mining, that is adequate with the level of security provided for data.

F. Interesting Challenges

- Threats imposed by data mining techniques to privacy/security and possible remedies.

- Statistical approaches to ensure privacy in data mining.
- Statistical disclosure control applied to privacy preserving data mining.
- New methodologies for privacy preserving data mining.
- Security leaks in existing privacy preserving data mining techniques.
- Privacy preserving data mining for specific applications particularly e-commerce.
- Effect of distributed data sources to privacy and security.
- Data quality, privacy, and security measures.

There has been much interest recently on using data mining for counter-terrorism applications. For example, data mining can be used to detect unusual patterns, terrorist activities and fraudulent behavior. While all of these applications of data mining can benefit humans and save lives, there is also a negative side to this technology, since it could be a threat to the privacy of individuals. This is because data mining tools are available on the web or otherwise and even naïve users can apply these tools to extract information from the data stored in various databases and files and consequently violate the privacy of the individuals. Recently we have heard a lot about national security vs. privacy in newspapers, magazines and television talk shows. This is mainly due to the fact that people are now realizing that to handle terrorism; the government may need to collect information about individuals. This is causing a major concern with various civil liberties unions.

We are beginning to realize that many of the techniques that were developed for the past two decades or soon the inference problem can now be used to handle privacy. One of the challenges to securing databases is the inference problem. Inference is the process of users posing queries and deducing unauthorized information from the legitimate responses that they receive. This problem has been discussed quite a lot over the past two decades. However, data mining makes this problem worse. Users now have sophisticated tools that they can use to get data and deduce patterns that could be sensitive. Without these data mining tools, users would have to be fairly sophisticated in their reasoning to be able to deduce information from posing queries to the databases. That is, data mining tools make the inference problem quite dangerous. While the inference problem mainly deals with secrecy and confidentiality we are beginning to see many parallels between the inference problem and what we now call the privacy problem.

X. Conclusion

Data mining has become one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. In the context of homeland security, data mining can be a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. While data mining represents a significant advance in the type of analytical tools currently available, there are limitations to its capability. One limitation is that although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. A second limitation is that while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. Successful data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output. Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance. However, some of the homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Some efforts that have attracted a higher level of congressional interest include the Terrorism

As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of a project's outcome. One issue is data quality, which refers to the accuracy and completeness of the data being analyzed. A second issue is the interoperability of the data mining software and databases being used by different agencies. A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected. As data miners, our tasks are colliding with these concerns. In analytic customer relationship management (CRM), we often analyze customer data with the specific intent of understanding individual behavior and instituting sales campaigns based on this understanding. Researchers in economics, demographics, medicine and social sciences are trying to understand the relationships between behaviors and outcomes. Both privacy and security are politically popular areas of concern, with growing public awareness and activism in the U.S., Europe, and in many other countries. Therefore, the temptation to legislate and regulate to protect the public may outweigh the consequences of restricting both online and offline commerce. On the other hand, the burden is on business to show where federal legislation is necessary to enhance electronic commerce, with clear benefits and consumer protections. Finally, elected and public officials should be informed of the costs and consequences to consumers, businesses, and the economy of legislative or regulatory proposals to protect privacy and security.

References:

[1] *Introduction to Data Mining and Knowledge Discovery*, Third Edition ISBN: 1-892095-02- 5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Dunham, M. H., Sridhar S., "*Data Mining: Introductory and Advanced Topics*", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "*From Data Mining to Knowledge Discovery in Databases*," AI Magazine, American Association for Artificial Intelligence, 1996.

[4] Larose, D. T., "*Discovering Knowledge in Data: An Introduction to Data Mining*", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[5] L. Getoor, C. P. Diehl. "Link mining: a survey", *ACM SIGKDD Explorations, vol. 7, pp. 3-12, 2005.*

[6] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. "*From Data Mining to KDD : An Overview*", AAAI/MIT Press, 1996.

[7] Han J. et Kamber M., "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, Canada, 2002.

[8] *Introduction to Data Mining and Knowledge Discovery*, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999

[9] David Hand, Heikki Mannila, and Padhraic Smyth," *Principles of Data Mining", MIT Press, Cambridge, MA, 2001.

[10] Peter Cabena, Pablo Hadjinian, Rolf Stadler, JaapVerhees, and Alessandro Zanasi, *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NJ, 1998.

[11] Mafruz Zaman Ashrafi, David Taniar, Kate A. Smith, "*Data Mining Architecture for Clustered Environments" , Proceeding PARA '02 Proceedings of the 6th International Conference on Applied Parallel Computing Advanced Scientific Computing*, Pages 89-98, Springer-Verlag London, UK ©2002

[12] Clifton, C. and D. Marks, "Security and Privacy Implications of Data Mining", *Proceedings of the ACM SIGMOD Conference Workshop on Research Issues in Data Mining and Knowledge Discovery, Montreal,* June 1996.

[13] Z. Ferdousi, A. Maeda, "Unsupervised outlier detection in time series data", *22nd International Conference on Data Engineering Workshops*, pp. 51-56, 2006

[14] Morgenstern, M., "Security and Inference in Multilevel Database and Knowledge Base Systems," *Proceedings of the ACM SIGMOD Conference, San Francisco,* CA, June 1987.

[15] S. A. Demurjian and J. E. Dobson, "Database Security IX Status and Prospects Edited by D. L. Spooner ISBN 0 412 72920 2, 1996, pp. 391-399.

[16] Lin, T. Y., "*Anamoly Detection -- A Soft Computing Approach*", Proceedings in the ACM SIGSAC New Security Paradigm Workshop, Aug 3-5, 1994,44-53.,1994

[17] Scott W. Ambler, "Challenges with legacy data: Knowing your data enemy is the first step in overcoming it", Practice Leader, Agile Development, Rational Methods Group, IBM, 01 Jul 2001.

[18] Agrawal, R, and R. Srikant, "Privacy-preserving Data Mining," *Proceedings of the ACM SIGMOD Conference, Dallas, TX, May* 2000.

[19] Clifton, C., M. Kantarcioglu and J. Vaidya, "Defining Privacy for Data Mining," Purdue University, 2002 (see also Next Generation Data Mining Workshop, Baltimore, MD, November 2002.

[20] Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 2002.

[21] Fung B., Wang K., Yu P. "Top-Down Specialization for Information and Privacy Preservation. ICDE Conference, 2005.

[22] Wang K., Yu P., Chakraborty S., " Bottom-Up Generalization: A Data Mining Solution to Privacy Protection.", ICDM Conference, 2004.