International Journal of Marketing and Technology
(ISSN: 2249-1058)

CONTENTS

Chief Patron

Patron

Chief Advisors

Editorial Board

Editorial Advisory Board

Title

DATA MINING ISSUES AND KEY TO SUCCESS

Author(s)

Deep Mala Sihint,

B.Sc, M.A (English),

MCA, Lecturer,

Jayawant Institute of Management Studies,

Tathwade, Pune-411033

Abstract:

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge, and can be used for applications ranging from business management, production control and market analysis to engineering design and science exploration. A fruitful direction for future data mining research will be the development of techniques that incorporate privacy concerns. Data Mining is the process of automatic discovery and understandable models and patterns from large amounts of data. Data mining has attracted tremendous interest in the research community as well as commercial market place. The last few years have witnessed a flurry of technical innovations and introduction of commercial products. The next major challenge facing data mining is to make a transition from a niche technology to a main stream technology. In this paper we will present key technical issues that must be addressed for a successful transition.

Keywords: - Data Mining, Data, Information, Process, growth, Storage

Introduction:

The rapid growth in information science and technology in general and the complexity and volume of data in particular have introduced new challenges for the research community. Many sources produce data continuously. [2] The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of data bases are increasing ever faster. The increase in the use of electronic data gathering devices such as point- of-sale or remote sensing devices has contributed to this explosion of available data. Figure1 from the Red Brick company illustrates the data explosion.

Data Storage became easier as the availability of the large amount of computing power at low cost i.e. cost of processing power and storage is falling, made data cheap. Data mining has resulted from the recent convergence of large databases of customer or member information, high speed computer technology and sophisticated analytical techniques. The use of

data mining tools represents a phenomenal opportunity for associations to improve the performance of organizational, marketing and delivery strategy and create long term Sustainable growth. Data mining describes a series of tools that allow associations to comprehensively understand the relationship of members to the association by analyzing their historical behavior with the association. By creating analytical models of member behavior it is possible to make predictions of future member behavior to guide organizational, marketing and delivery strategy. This article represents an introduction to the potential of these tools, and a brief description of the data mining process.[3]

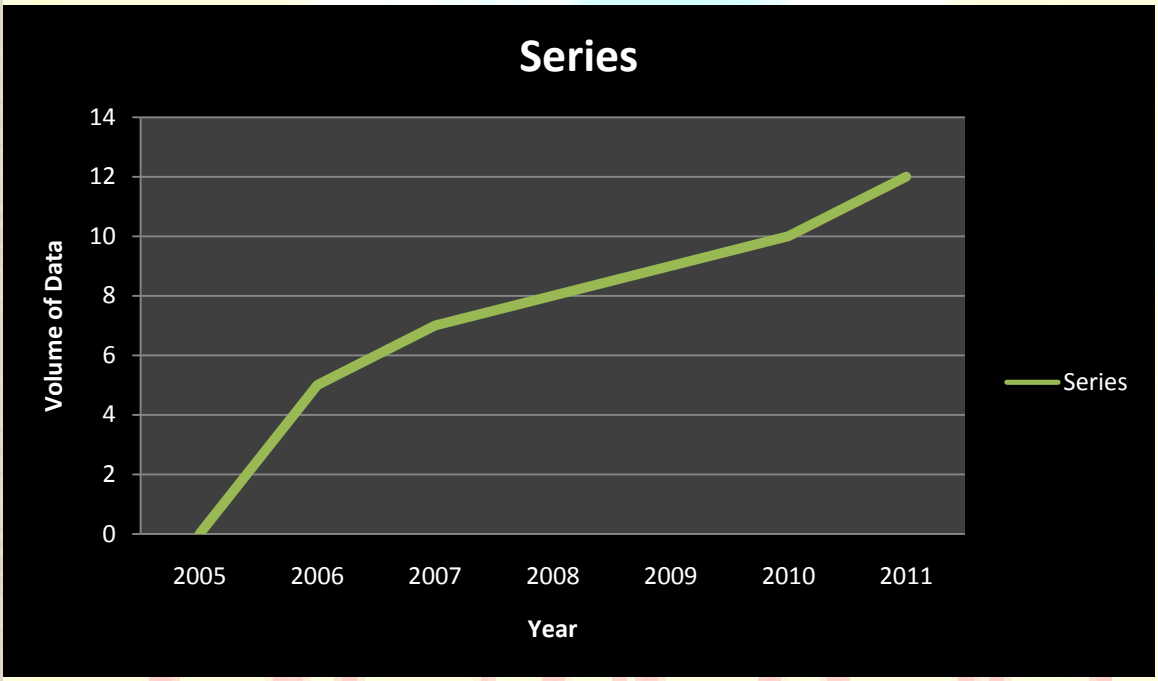

Figure1: - The Growing Base of Data

Definition:

The term data mining has been stretched beyond the limits to apply to any form of data analysis. Some of the numerous definitions of data mining are:[4]

Data mining or knowledge Discovery in databases (KDD) as it is also known is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization,

learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies.

-William J Frawley , Gregory Piatetsky- Shapiro and Christopher J Matheus

Data mining is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represents valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database.

-Marcel Holshemier & Arno Siebes(1994)

Analogy:

The analogy with the data mining process is described as:

Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands off low value as no direct use can be make of it; it is the hidden information in the data that is useful"

-Clementine User Guide, a data mining Toolkit

Basically data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. Data mining tends to work from the data up and the best techniques are those developed with an orientation towards large volume of data, making use of as much of the collected data as possible to arrive at reliable conclusions and decisions. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which the time knowledge is required. Once knowledge has been acquired this can be extended to large sets of data working on the assumption that the large data set has a structure similar to the sample data.[5]

Stages in Data Mining:

The phase depicted starts with the raw data and finish with the extracted knowledge which was acquired as a result of the following stages:[5]

1) **Selection: -** Selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined.
2) **Preprocessing:-** This the data processing cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries for example unnecessary to note the sex of a patient when studying Pregnancy. Also the data is reconfigured to ensure a consistent format as there is a possibility of inconsistent formats because the data is drawn from several sources e.g. sex may be recorded as 'F' or 'M' and also 1 or 0.
3) **Transformation:-** The data is not merely transferred across but transformed in that overlays may added such as the demographic overlays commonly used in market research. The data is made useable and navigable.
4) **Data Mining: -** This stage is concerned with the extraction of patterns from the data. A pattern can be defined as given a set of facts (data) F, a language L, and some measure of certainty C a pattern is S and L that describes relationships among a subset Fs of F with a certainty c such that S is simpler in some sense than the enumeration of all the facts in Fs.

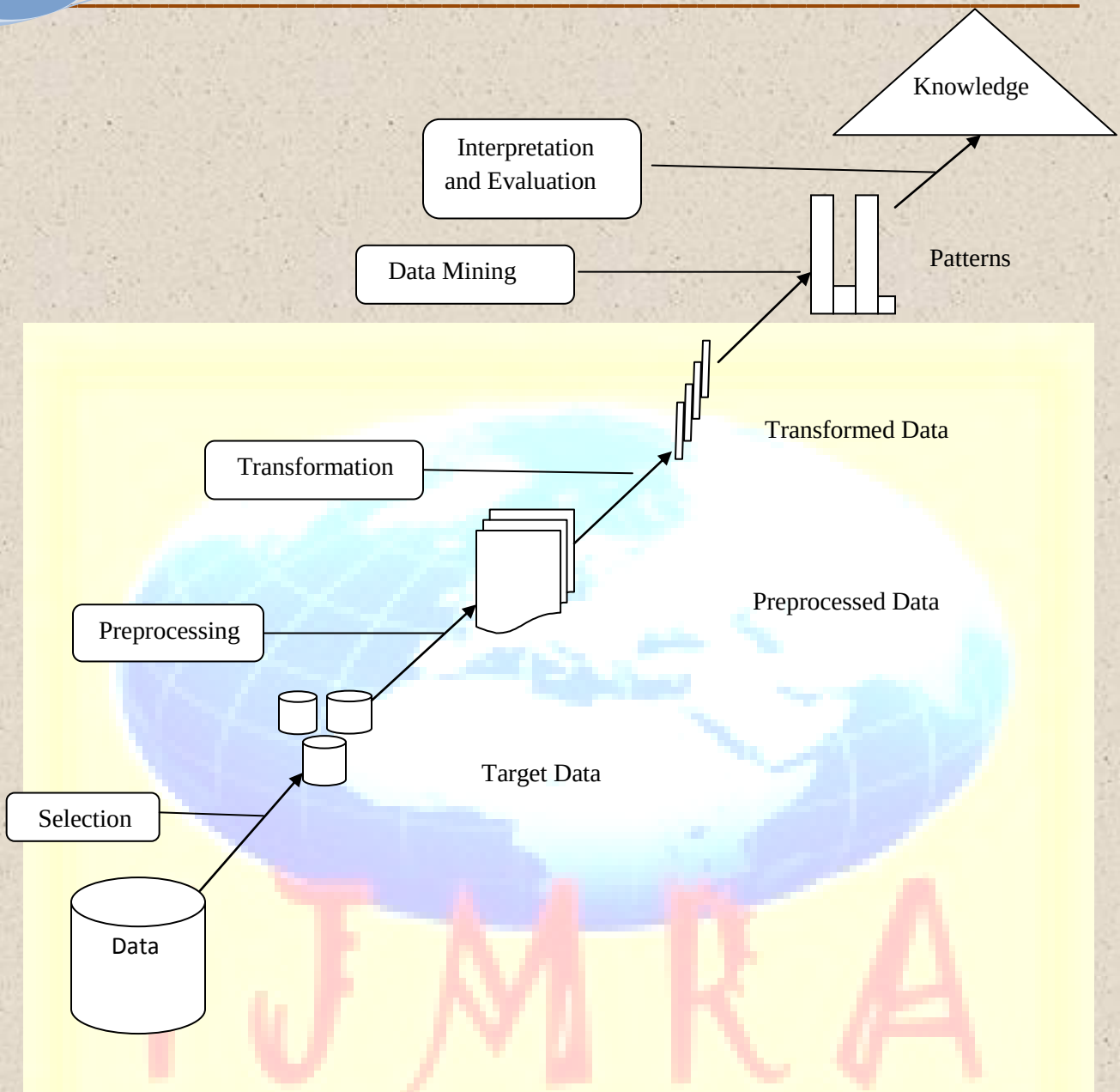

Figure2:- Stages/ Processes identified in data mining by Usama Fayyad & Evangelous Simoudis

5) **Interpretations and Evaluation:-**The patterns identified by the system are interpreted into knowledge which can then be used to support human decision- making e.g. prediction and classification tasks, summarizing the contents of a database or explaining observed phenomena.

Data Mining problems /Issues:

Data mining systems rely on databases to supply the raw data for input and this raises problems in that databases tend to be dynamic, incomplete, noisy and large. Other problems arise as a result of the adequacy and relevance of the information stored.[6]

1) **Limited Information:** - A database is often designed for purposes different from data mining and sometimes the properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Inconclusive data causes problems because if some attributes essential to knowledge about the application domain are not present in the data it may be impossible to discover significant knowledge about a given domain. For example cannot diagnose malaria from a patient database if that database does not contain the patient's red blood cell count.

2) **Noise and Missing Values:** - Databases are usually contaminated by errors so it cannot be assumed that the data they contain is entirely correct. Attributes which rely on subjective or measurement judgments can give rise to errors such that some examples may even be mis-classified. Errors in either the values of attributes or class information are known as noise. Obviously where possible it is desirable to eliminate noise from the classification information as this affects the overall accuracy of the generated rules.

   Missing data can be treated by discovery systems in number of ways such as:

   - Simply disregard missing values
   - Omit the corresponding records
   - Infer missing values from known values
   - Treat missing data as a special value to be included additionally in the attribute domain.
   - Average over the missing values using Bayesian Techniques

3) **Uncertainty: -** Uncertainty refers to the severity of the errors and the degree of noise in the data. Data precision is an important consideration in a discovery system.

4) **Size, Updates, and irrelevant Fields: -** Databases tend to large and dynamic in that their contents are ever-changing as information is added, modified or removed. The problem with this form the data mining perspective is how to ensure that the rules are up-to-date and consistent with the most current information. Also the learning system has to be time –sensitive as some data values vary over time and the discovery system is affected by the 'timeliness' of the data. Another issue is the relevance or irrelevance of the fields in the database to the current focus of discovery for example post codes are fundamental to any studies trying to establish a geographical connection to an item of interest such as the sales of a product.

What is data mining good for?

Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning twists thrown in. Like statistics, data mining is not a business solution, it is just a technology. For example, consider a catalog retailer who needs to decide who should receive information about a new product. The information operated on by the data mining process is contained in a historical database of previous interactions with customers and the features associated with the customers, such as age, zip code, and their responses. The data mining software would use this historical information to build a model of customer behavior that could be used to predict which customers would be likely to respond to the new product. By using this information a marketing manager can select only the customers who are most likely to respond. The operational business software can then feed the results of the decision to the appropriate touch point systems (call centers, direct mail, web servers, email systems, etc.) so that the right customers receive the right offers. [1]

How does data mining work?

Data mining uses sophisticated analytic techniques to search large volumes of data. In searching this data, these techniques build models for patterns that accurately predict member and/or Customer behavior. By understanding anticipated future member behavior the association can proactively develop strategies to take advantage of this knowledge. Two keys to successful data mining is a large, integrated store of membership data linked by a common denominator such as member number and an in-depth understanding of the association business processes within which data mining is to be applied such as new member prospecting, membership retention, etc.

While some associations have a well established membership database with extensive information on member behavior, the integration of several different databases such as membership, conference registration and product sales works equally well. The key is the common identifier so that individuals can be matched to data across different databases.

Key factors for successful Data Mining [7][8]

1. Identify the primary business objectives of the association

Like a carpenter's toolbox, the tools of data mining can be used to build different types of models. The first step is to identify which business goals, for example member retention, new product sales or increased conference attendance are most important. Based on this decision, the correct series of tools can be applied to the data to build the corresponding behavioral model.

2. Identify the existence and availability data

The next step is to identify what association behavioral data exists that will be relevant to the identified business goals. If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this stage to guide future member relationship management.

3. Consolidate data into a temporary data warehouse

Association data is sometimes contained in an association management software system such as IMIS, but just as frequently is stored in multiple databases for different association functions or

domains such as conference registration or membership. For analysis, all data needs to be consolidated so that it can be treated consistently.

4. **Clean the data so that missing and invalid values are treated and all known valid values are made consistent**

The purpose of this step is to identify gaps, flaws or other problems with the data and establish protocols to clean or otherwise treat these records so that the final accuracy of the model is maximized.

5. Transform the data for more robust analysis

Data transformation involves altering the distribution of variables to allow for the use of statistical methodologies that require data to be linearly related to an objective variable and to create new variables by combining existing variables to form ratios.

6. Implement data mining algorithms

Based on the data and the desired business outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular business objective to be achieved and the quality of the data to be analyzed.

7. Develop conclusions and make recommendations

Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration by the association.

Conclusion:

Data mining can be beneficial for businesses, governments, society as well as the individual person. However, the major flaw with data mining is that it increases the risk of privacy invasion. Currently, business organizations do not have sufficient security systems to protect the information that they obtained through data mining from unauthorized access, though the use of data mining should be restricted. In the future, when companies are willing to spend money to

develop sufficient security system to protect consumer data, then the use of data mining may be supported.

Bibliography:

- http://www.thearling.com
- http://books.google.co.in
- http://www.associationlaboratory.com/pdf/SuccessfulAssociationData.pdf
- http://www.springerlink.com/content/bgr7687721232681/
- Data Mining Concepts and Techniques by Jiawei Han and Micheline Kamber(Second Edition)
- Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery". International Journal of Information Technology and Decision Making, Volume 7, Issue 4 7: 639–682
- Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.
- Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". http://www.britannica.com/EBchecked/topic/1056150/data-mining.Retrieved 2010-12-09.