International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 08 | Aug 2018 www.irjet.net p-ISSN: 2395-0072

A STUDY ON DATA MINING IN SOFTWARE

Sreenivasulu Tholuchuri

MCA, Hyderabad, India

Abstract - *Data mining for software engineering is a process of discovering software engineering data in databases. In simple words, it's a series of actions to extract knowledge from useful patterns and relationships in huge volumes of databases and use that knowledge to improve the software engineering process. It uses tools from artificial intelligence and statics with database management to analyze large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, and retail), science research (astronomy, medicine) and government defense departments.*

Key Words: Data Mining, Software Engineering, Error Detection, Clustering, Association

1. INTRODUCTION

Data Mining also called as knowledge discovery in databases or KDD for making productive use of mined knowledge in operable way.

1.1 Early Ages

During 1980's data storage capacities in computers increased a lot and many big companies started began to store transactional data which resulted collections of huge volume records, often called as data ware houses. This data warehouses were too large to be analyzed with traditional statistical approaches.

With the aim of knowledge discovery several computer science workshops and conferences were held for adapting the techniques from the field of Artificial Intelligence (AI) --- such as neural networks, genetic algorithms, machine learning etc. This led to the First International Conference on Knowledge Discovery and Data Mining (FICKDD), held in Montreal, and the launch in 1997 of the journal Data Mining and Knowledge Discovery which was also the period when many early data-mining companies were formed and products were introduced.

Now a day's vast amounts of data are collected daily. Figuring out such data is an important need.

"We are living in the information age" is a popular saying; however, we are actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools. Businesses worldwide generate gigantic data sets, including sales transactions, stock trading records, product descriptions, sales promotions, company profiles and performance, and customer feedback.

For example, large stores, such as Wal-Mart, handle hundreds of millions of transactions per week at thousands of branches around the world. Scientific and engineering practices generate high orders of thousands of terabytes of data in a continuous manner, from remote sensing, process measuring, scientific experiments, system performance, engineering observations, and environment surveillance. The medical and health industry generates tremendous amounts of data from medical records, patient monitoring, and medical imaging. Billions of Web searches supported by search engines process thousands of terabytes of data daily. Communities and social media have become increasingly important data sources, producing digital pictures and videos, blogs, Web communities, and various kinds of social networks. The list of sources that generate huge amounts of data is endless. This hazardous growing, universally available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This essentiality has led to the birth of data mining. The field is young, dynamic, and promising. Data mining has and will continue to make great strides in our journey from the data age toward the coming information age.

2. DATA MINING

Data Mining is more appropriately named "knowledge mining from data," which seems somewhat long. However, the shorter term, knowledge mining may not reflect the attention on mining from large amounts of data. Though, Mining is expressive term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

In addition, other terms have a same meaning to data mining—for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Many people think data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery.

The knowledge discovery process is shown in Figure below as an iterative sequence of the following steps:

1) Data cleaning (to remove inconsistent data and noise)
2) Data integration (phase where multiple data sources may be combined)
3) Data selection (phase where data relevant to the analysis task are retrieved from the database)
4) Data transformation (phase where data are transformed and consolidated into forms fitting for mining by performing summary or aggregation operations)
5) Data mining (an crucial process where intelligent methods are applied to extract data patterns)
6) Pattern evaluation (phase to identify the truly interesting patterns representing knowledge based on interestingness measures)
7) Knowledge presentation (phase where visualization and knowledge representation techniques are used to show mined knowledge to users)

Steps 1 to 4 are different forms of data preprocessing, where data are prepared for mining. The data mining process may interact with the user or a knowledge base. The interesting patterns are showed to the user and may be stored as new knowledge in the knowledge base. The previous view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term data mining is generally used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data). Therefore, we adopt a broad view of data mining functionality: **Data mining** is the process of discovering interesting patterns and knowledge from huge amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that flow into the system dynamically.

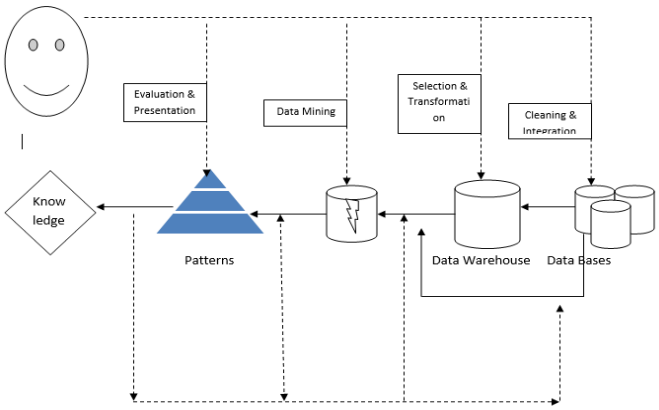

Fig -1: Data Mining Steps

3. GOALS OF SOFTWARE ENGINEERING

Requirement Analysis: In this phase of SE task gathering of software requirements from client, analyze and documenting data are done. It's a functional or non-functional need to be implemented in the system. Client's acceptance is mandatory to proceed for further process.

Whatever document prepared in this phase is called Software Requirement Specifications (SRS)

System Design: System design is a process of defining user interfaces, modules, architecture and the data for a system to satisfy client requirements. Here we will implement overall product design as per client requirement different types of SDK will be used.

Development/Programming: The source code of the program is written in different programming languages as per client requirement. It is called programming process in software development. Coding reserved for actual writing of source code. It is a main part in the software development. Software development organization requires good programmers to define the standard style of code called Coding standards. It gives a good appearance to the code written by different software programmers. It should be understandable, reusable which follows good programming practices. Naming conventions, limitation of data types and using of variables, constants are main coding standards.

Error Detection/Bug Fixes: Error detection or bug fix is a essential process for effective and proper software project planning. Some data related software bugs are kept in bug repositories. It contains information related to bugs. A bug fix contains data and code related. There are many types of programming bugs, design bugs, data bugs that create errors in system implementation may require fixes that are successfully resolved by development team.

Testing: Software testing is not a cost effective. It is the important phase in software development. There are different stages in testing to validate or verify software. Verification and validation processes are concerned with checking that software being developed meets its specification and delivers the functionality expected by the people paying for the software.

During testing, errors can mask (hide) other errors. When an error leads to unexpected outputs, you can never be sure if later output anomalies are due to a new error or are side effects of the original error. Because analysis is a static process, you don't have to be concerned with interactions between errors. Consequently, a single analysis session can discover many errors in a system.

Maintenance: Good software should deliver the required functionality and performance to the user and should be maintainable, dependable, and usable. Software will be written in such a way so that it can be develop to meet the changing needs of customers. This is a demanding attribute because software change is an impending requirement of a changing business environment.

Agile procedures, used in the maintenance process itself, are likely to be effective, whether or not an agile way has been used for system development. Incremental delivery, design for change and maintaining simplicity all make sense when software is being modified. In fact, you can think of an agile development process as a process of software expansion.

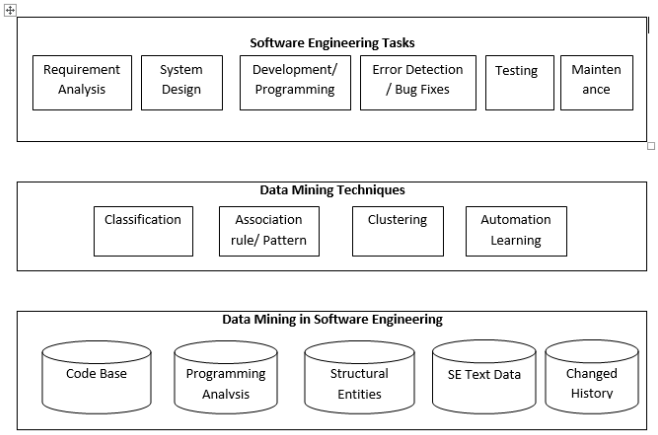

Fig -2: Software Engineering Tasks, Data Mining Techniques & Data Mining in Software Engineering

4. TECHNIQUES IN DATA MINING

4.1 Association rule

Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining. The problem of mining association rules can be reduced to that of mining frequent item sets.

Association Rule mining approach is applied to the records in order to discover the patterns that are possibly to cause high severity defects.

Max association rule mining algorithms employ a support–confidence framework. Even though minimum support and confidence thresholds help weed out or exclude the exploration of a good number of uninteresting rules, most of the rules generated are still not interesting to the users. Regrettably, this is especially true when mining at low support thresholds or mining for long patterns. This has been a big obstacle for successful application of association rule mining

4.2 Classification:

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The models are derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

A neural network, when used for categorization, is typically a collection of neuron-like processing units with weighted connections between the units. There are various other methods for constructing classification models, such as Bayesian classification, support vector machines, and k-nearest-neighbor classification

Regression analysis is a statistical technique/approach that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data.

The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

4.3 Clustering:

Clustering plays a central role in customer relationship management, which groups customers based on their similarities. Using relevance mining techniques, we can better understand features of each customer group and develop customized customer reward programs.

Clustering techniques consider data tuples as objects. They partition the objects into groups, or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how "close" the objects are in space, based on a distance function. The "quality" of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster.

The specified set of entities to cluster needs to be identified, before applying clustering to a software system. The next phase is attribute selection. Max software clustering methods at first transform a fact base to a data table, where each row describes one entity to be clustered. Each column contains the value for a specific attribute. After accomplishment of all preparation steps the clustering algorithm can start to execute. Clustering algorithms used in software engineering are: graph-theoretical algorithms, construction algorithms, optimization algorithms, hierarchical algorithms. For high dimensional data, many of the existing methods fail due to the curse of dimensionality, which contribute particular distance functions problematic in high-dimensional spaces which led to new era of

clustering algorithms for high-dimensional data that focus on subspace clustering and correlation clustering that also looks for arbitrary rotated subspace clusters that can be modeled by giving a correlation of their attributes.

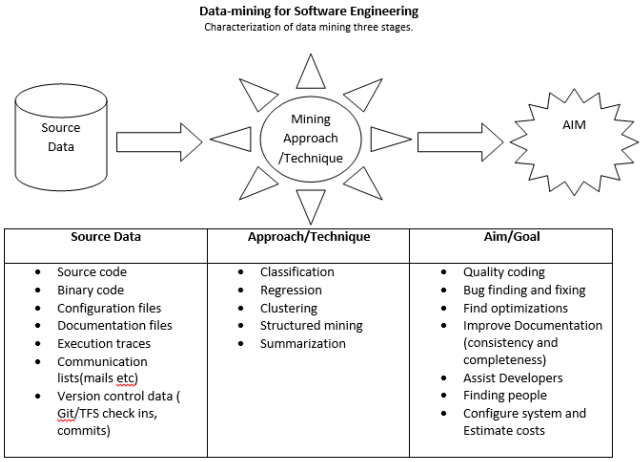

Fig -3: Characterization of Data Mining

5. CONCLUSIONS

In this paper, I have tried to provide an analysis of Data Mining and its origin. Why data mining is essential in this era of computer world. An analyzed information about Software Engineering and it various phases. How Data Mining in Software Engineering is classified and Data Mining Techniques used in process for fruitful knowledge discovery.

REFERENCES

[1] Tao Xie,Jian Pei,Ahmed E. Hassan, "Mining Software Engineering Data"

[2] Lovedeep, Varinder Kaur Atri, "Applications of Data Mining Techniques in Software Engineering", IJEECS

[3] Jiawei Han,Micheline Kamber & Jian Pei, "Data Mining Concepts and Techniques", Third Edition

[4] Ian Sommerville, "SOFTWARE ENGINEERING" Ninth Edition