Webology (ISSN: 1735-188X)
Volume 16, Number 2, 2019

Data Mining: A Process Of Extracting Patterns

Abhay Bhatia[1], Parag Jain[2], Praveen Verma[3], Gaurav Gupta[4], Deepak Arya[5], Barkha Chaudhary[6]

[1]Dept. of CSE, Dreams College of Polytechnic, Saharanpur

[2,3,4,5] Dept. of CSE, Roorkee Institute of Technology, Roorkee

[6]Student, Dept. of CSE, Roorkee Institute of Technology, Roorkee

Abstract— The technique of extracting patterns from data is known as data mining. In a nutshell, data mining is the process of analysing observational datasets to discover unexpected relationships and summarise data in new and valuable ways for data owners. It is increasingly vital in modern company for translating data into business intelligence and providing an information edge. Data mining's automated, prospective analysis extends beyond the examination of previous events provided by decision support systems' retroactive capabilities. Business problems that formerly took too long to address can now be answered using data mining methods. In both the business and public sectors, data mining is becoming increasingly popular. Data mining is commonly used to cut costs, improve research, and increase sales in industries such as banking, insurance, healthcare, and retail. Data mining is a huge step forward in terms of the types of analytical tools currently available, but it has limitations. One restriction is that data mining aids in the discovery of patterns and linkages, but does not express to the user the value or significance of such patterns. The user must make these kinds of selections. The second flaw is that data mining can uncover patterns of behaviour and correlations between variables, but not necessarily causality. Professional skills and analytical experts who can arrange the analysis and explain the results are required for successful data mining.

Keywords— Discovering Knowledge, OLAP, Extraction, Mining, Data warehouse

I. INTRODUCTION

The extraction of hidden forecast information from massive databases is known as data mining. It enables you to locate the needle in the haystack of data. It's a trendy new technology that has a lot of promise for helping businesses focusing on the important data that is most appropriate in their data warehouse. This technique predicts future behavioural trends, which allows businesses to make proactive with data-driven decisions. Moreover it allows for prospective analysis with automation that goes beyond the examination of previous events provided by decision support systems' retroactive tools. Business problems that formerly took too long to address can now be answered using data mining methods. They comb through the information for hidden patterns and analyses that experts might overlook because they are out of the ordinary. In both the corporate and public sectors, it is becoming increasingly crucial. For instance, data mining technologies are used in the insurance and banking industries to detect fraud and aid risk assessment. Data mining can be used to anticipate outcomes in a range of situations. It's commonly referred to as Database Knowledge Discovery (KDD). Data mining and knowledge discovery are common database synonyms; however it is actually an element that are associated to the discovery process for knowledge.

The Discovering of Knowledge in any usable Databases process made up of the following phases that lead from unprocessed collection of data for some form of valid and dignified data knowledge:

1. Data Cleansing: In such phase, irrelevant data with its noise data are removed from the complete collection.
2. Assimilating of Data: This phase consist of multiple data sources that are often mixed with or may be combined into some common origin.
3. Process of Data assorts: Here, retrieved data from collection is done and the data relevant to the analysis is decided.
4. Conversion of data: In this phase, selected data firstly transformed into forms and then that is appropriate for the mining procedure.
5. Data mining: It is the vital step in which knowledgeable techniques are applied to extract patterns potentially useful.
6. Evaluation: In this step, firmly extraction of patterns that represents knowledge is identified that are based on some given measures.
7. Knowledge visualizer: It is the last but not the least phase where the discovered knowledge is represented visually to the user and this essential step uses visualization techniques to help users understand the data mining results.

Usually some of the above steps are combined.

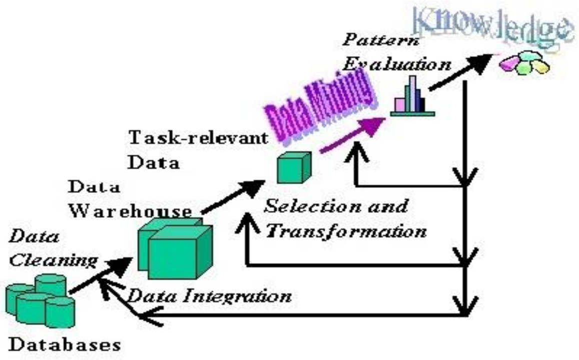

Figure 1

Data cleansing and data assimilation, for example, might be used as a pre-processing phase to create a data warehouse. Data assortment and data transformation can also be integrated when the consolidation of the data is the result of the piece, or as for the case of data warehousing, the selection is done on changed data. This is an iterative procedure. Once the user has access to the newly acquired knowledge, the evaluation metrics can be improved, the mining modified, fresh data picked or changed, or new data sources integrated to produce different, more appropriate findings.

II. DATA MINING NEEDS

- A vast amount of data is now being gathered. Every 12 months, the amount of data collected is predicted to roughly double. One of the most sought-after aspects of data mining is the ability to find knowledge from enormous amounts of data. The data can grow in two ways: in terms of size, for example, in the case of image data, or in terms of dimension, for example, in the case of gene expression data.
- There is frequently a significant gap between the data that is saved and the knowledge [12] that may be drawn from it. This type of migration does not happen by itself. It's how the process of data mining works. Although some initial knowledge of the data is known in exploratory data analysis, data mining might be useful for a more in-depth comprehension of the data.
- Analyzing of data manually data analysis is long time process, but it becomes as bottleneck when any large amount of data is made to be analyzed.
- Computer science and engineering technologies and processes are rapidly developing, resulting in new requirements. Data mining techniques are currently used in a variety of data-rich sectors. Genetic data analysis and image mining.

III. BASIC WORKING OF DATA MINING

Based on open-ended user questions, data mining software examines stored operational data associations and trends. Statistics, machine learning, and neural networks are just a few examples of logic software. In general, one of four types of relationships can be found:

• Class: The data is organised into groups using the stored data. A restaurant chain, for example, can look into a customer's purchase history to see when they come in and what they normally order. This data can be utilised to boost traffic by offering daily promotions.

• Cluster: Data is organised into groups based on logical correlations or consumer preferences.
You can use data mining to determine market categories and consumer preferences, for example.

• Association: You can mine data to identify the association. The beer diaper example is an example of associative mining.

• Sequential patterns: To forecast behavioural patterns and trends, data is analysed. Outdoor equipment companies, for example, can forecast whether a customer will buy a backpack based on their previous purchases of sleeping bags or hiking boots. There are five main components to data mining:

- Extract transaction data, transform it, and load it into your data warehouse system.
- Data storage and management in multi-dimensional database systems.
- Providing access of data to the IT analysts and business professionals.
- Analysis of data with the help of desired application software's.
- Presenting a valid format of the usable data in the form of figure or a table.

Analyzation at different levels:-

• Neural Network [Artificial intelligence]: A nonlinear based models that are predictive in nature and learns through specified training and is structurally similar to a biological neural network.

• Genetic algorithm: Natural evolution is as an optimization strategy that employs mechanisms including natural selection with gene combination and mutation.

• Decision Tree: A tree-like structure that represents a series of decisions. These decisions generate rules for classifying documents. Specific decision tree methods include classification and regression tree (CART) and Chi-Square automatic interaction detection (CHAID). CART and CHAID is decision tree techniques used to classify datasets. They provide a set of rules that can be applied to new (unclassified) records to predict which records will give a particular result. CART segments the dataset by creating a two-way split, and CHAID uses a chi-square test to generate the segment to create a multi-directional break. CART usually requires less data preparation than CHAID.

• Nearest Neighbour Method: The closest neighbour approach is a methodology for classifying each record in a dataset based on a combination of classes of k records that are most similar in the historical dataset (k 1).

• Rule guidance: Extract applicable and "if-then" rules from the database itself on a statistical significance for extraction.

• Data visualization: Visually interpret the complex relationships between the multidimensional data. On the other hand graphical based tools are used for explaining such relationships between the data.

IV. ARCHITECTURE FOR DATA MINING

These advanced methodologies must be properly integrated with data warehouses and flexible, interactive business analytics tools in order to achieve their full potential. Many data mining tools today operate outside of the warehouse, necessitating additional steps to extract, import, and analyze data. Integration to the warehousing also gets simplifies for the application of data mining outcomes and when new insights require operational implementation. The resulting analytics data warehouse can be used to improve business processes throughout your organization, including ad campaign management, fraud detection, and upcoming product usage. The architectural view of advanced analytics in a huge data warehouse is shown in Figure 2.

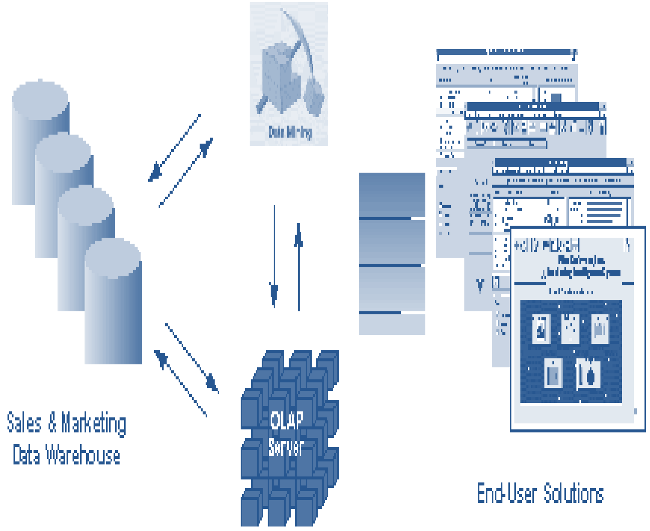

Figure: - 2

A data warehouse, which contains a combination of internal and external data about all customer contacts and competitors' activities, is an excellent place to start. Information about potential customers' backgrounds is also useful for acquiring new customers. This warehousing approach can be implemented using relational database systems like Iris, CoCo, and Rice, and must be tuned for flexible and quick data access. The OLAP (On-Line Analytical Processing) server enables you to browse your data warehouse using more powerful end-user business models. The multidimensional structure will enable users to analyze data the way they want to see their business. They are summarized by product lines, regions, and other key business perspectives. To combine ROI-focused business analytics directly into this infrastructure, data mining servers must be integrated with data warehouses and OLAP servers. Advanced process-centric metadata templates define data mining goals for specific business topics, such as campaign management, customer acquisition, and ad optimization. Thanks to the process of integration of the data warehousing services with the operational decisions that can be implemented directly and can easily be tracked. As the warehousing growing with its new and implementable findings as well as the results, organizations can continually kept on identifying the best practices and try them to be applied for the decisions ahead. Such design is a significant departure from traditional decision-making aids. Advanced Analysis Server generally applies the user's business model directly towards the warehouses, which returns a proactive analysis to the most relevant information needed, in addition to giving data to end users via query and reporting tools. These findings improve OLAP server metadata by adding a dynamic metadata layer that describes the data's extracted view. After that, you may use reports, infographics, and other analytical tools to plan future activities and assess their impact.

V. TECHNIQUES OF MINING

Neural Networks/Pattern Recognition - In a black-box approach, neural networks are deployed. One prepares a test data set, and then allows the neural network to learn patterns based on known results before unleashing it on massive amounts of data. A credit card business, for example, has 3,000 records, 100 of which are known fraud records. The data set is updated to ensure that the neural network recognizes the difference between fraudulent and authorized records. The grid learns the fraud records' patterns. The network is then applied to the company's million-record data set, and it produces records that are identical or similar to the fraud records. Neural networks are notorious for not teaching analysts anything about the data and instead focusing on discovering patterns that fit. The Post Office has employed neural networks for optical character recognition to help automate the delivery process without relying on humans to read addresses.

Memory - Based Reasoning - Rather than a pattern, MBR looks for "neighbor" data. We would build up a set of claims we want adjudicating and let the approach locate comparable claims if we were looking for the insurance claims and you want to know which one of the adjudicators must look on it and which ones they can just pass through the system without aligning it.

Detection based in cluster/ Analysis of Market - This is where the traditional beer or the diapers purchased together works for analyzing originated from. It looks for clusters. Essentially, this method identifies correlations between products or consumers, or wherever else we wish to identify data associations.

Link Analysis - This is yet another method for linking records that are similar. Although not frequently utilized, some tools are specifically designed for this purpose. As the name implies, the approach looks for and displays relationships in customers, transactions, and other data.

Visualization - This method aids the user in comprehending their data. Visualization creates a link between text-based and graphical presentations. Users can view data relationships rather than read about them using tools like decision trees, rules, clusters, and pattern visualization. Over the last few years, many of the more powerful data mining tools have made progress in improving their visual content. This is the way data mining and analysis will be in the future. Data volumes have expanded to the point where humans will soon be unable to process it properly using any text-based method. We'll most likely see a visualization-based approach to data mining emerge, similar to Microsoft's Photosynth. The technology is already in place; all that is required is a visionary analyst to sit down and put it all together.

Decision Tree/Rule Induction - Actual data mining algorithms are used in decision trees. Decision trees assist with classification and spit forth descriptive information, allowing consumers to better understand their data. The rules followed in a process will be generated using a decision tree process. When approving a loan, a bank lender, for example, follows a series of rules. The decision tree can define the criteria for the lending institution if we have consideration that is generally based on the data related to loan that a bank has, with the results of the paid or default cases of the loans and the limitations of default level of acceptance. The early decision support (or expert) systems were similar to these decision trees.

Genetic Algorithms - We can generate a data set and then provide the GA the opportunity to conduct various actions to see if a certain path or result is desirable. The GA will evolve in such a way that the end outcome will hopefully improve. The most prevalent applications include optimization of the process, including the scheduling, batching of relevant, re-engineering process and workflow.

OLAP – The term "online analytical processing" refers to a method of quickly responding to multi-dimensional analytical questions. Users can traverse data using OLAP by asking logical questions about it. Drilling down into data, from highly summary views of data to more granular views, is a common feature of OLAP. This is usually accomplished by traversing data hierarchies. If you're looking at population data, for example, Start with the most populous continent, then drill down to the most populous country, go with state level to city level, and ultimately to the neighborhood level from where you can try to find out roots. Drilling up hierarchies (drill up), shooting across several data dimensions or(shot across), and many more data browsing advanced techniques, such as automated time variation while drilling is done upward or downward with time hierarchies, are all part of OLAP. By far the most widely used and deployed approach is OLAP. It is also the most user-friendly and intuitive.

VI. ADVANTAGES

- **Marking/Retailing: -** Direct marketers can benefit from data mining because it can provide them with useful and accurate trends about their clients' purchase habits. Marketers can more precisely direct their marketing attention to their customers based on these tendencies. For example, a software company's marketers may offer upcoming software to customers who have purchased n-number of software previously. Furthermore, data mining also assists marketers for prediction on which products their clients are likely to be emphasized in purchasing. Marketers can use this forecast to surprise their customers and make their buying experience more enjoyable. In a similar way, data mining can benefit retail establishments. For example, as per the trending’s provided by data miner, the managers attached to the store can easily arrange certain stock items, or can provide a genuine and specific discount for attracting the targeted customers.

- **Bank/Credit:-** Financial institutions can benefit from data mining in areas like credit reporting and loan information. A bank, for example, can assess the level of risk associated with each loan by looking at prior customers with comparable characteristics. Furthermore, credit card providers can use data mining to detect possibly fraudulent transactions of credit card users. Although the mining technique cannot guarantee a 100 percent prediction with accuracy of fraudulent transactions, it does assist lowering losses to the issuers of credit card.

- **Law enforcement:-** By evaluation of trends in crime motive and type, located, major habits, and other behavior of pattern in work, mining can also assist law enforcement in identifying and apprehending the criminal suspects.

- **Researchers: -** Researchers can benefit from data mining since it speeds up the data analysis process, giving them ample time to work with other projects too.

VII. LIMITS IN MINING

While data mining software is extremely useful, it is not a stand-alone programme. Data mining demands individuals who are both technically talented and analytically trained, as well as those who can plan and analyse the results.

As a result, the constraints of data mining are generally data or personnel-related rather than technological. Despite the fact that data mining can help with pattern and relationship discovery, it does not provide the user with information about the value or significance of these patterns. This level of commitment is required of the user. Similarly, the usefulness of the discovered patterns is determined by how they are compared to "real-world" scenarios. Users can test their models with data that contains information on known terrorists, for example, to examine the effectiveness of a data mining programme meant to identify possible terrorist suspects in huge numbers. Although there may be an increase in certain characteristics, this does not always imply that the programme has discovered a suspect who is behaving considerably differently than the initial model. Another disadvantage of data mining is that, while it is possible to find links between behaviours and variables, causal relationships are not always possible to find. For instance, the programme can determine that certain behavioural patterns, such as the proclivity to purchase tickets just before a scheduled departure, are linked to certain events. Income, education level, and internet usage are all factors to consider. It does not imply necessarily that one or more of such type of variables are responsible for purchasing ticket. In truth, a person's behaviour might be influenced by a variety of other factors, such as their employment (the need to travel on short notice), their marital status (relatives with illnesses that require care), or their hobbies.

VIII. ISSUES

The social aspect of this technology is one of the most important issues it raises. It's all about protecting people's privacy. Data mining enables you to examine your daily transactions and acquire vast volumes of data about your personal purchasing habits and preferences.

The data's reliability is the second concern. Of course, data analysis is only as good as the information it is analyzing. Integrating competing or duplicate data from several sources is a major implementation difficulty. A bank, for example, may have credit card accounts in multiple databases. One cardholder's address (or name) may differ from that of another. It’s mandatory for software that it must convert data from system one to system two and recently the most addressed one is chosen as per it has been entered.

The third question concerns whether a relational database structure or a multidimensional database structure is preferable. The data is stored in tables in a relational structure, which enables for ad hoc queries. A multidimensional structure, on the other hand, uses an array of cubes to create a subset for each category. Multidimensional structures make multidimensional data mining easier, although relational structures have outperformed multidimensional structures in client/server situations thus far. And, because to the Internet's tremendous growth, the entire world is becoming one giant client/server ecosystem.

The final issue has been charged. Data mining and data warehousing have a tendency to reinforce each other. The more powerful a data mining query, the more useful the information extracted from the data, the higher the need to gather and manage data, and the larger the need for quicker and more powerful data mining searches. This puts more strain on systems that are more expensive, larger, and faster.
Faster and more powerful data mining queries impose a strain on pricier, larger, and more powerful computers.

IX. FUTURISTIC ASPECTS

In the momentry-term, the profitable outcomes will be there, if mundane, business-related fields. Micro-marketing initiatives will reach unprecedented levels of success. Advertising will be more precise in its targeting of potential buyers.

In the medium term, Data mining could become as popular and straightforward as e-mail. We may use these technologies to get the cheapest flights to Australia, track down a long-lost classmate's phone number, or compare lawn mower pricing..

The long-term the possibilities are extremely amazing. Consider intelligent agents unleashed on data from medical research or data from subatomic particles. Computers may offer new remedies for diseases or new understandings of the universe's nature.

References

1. M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
2. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
3. Nisbet, Robert, John Elder, Gary Miner, 'Handbook of Statistical Analysis & Data Mining Applications, Academic Press/Elsevier.
4. Pang Ning Tan Michael Steinbach and Vipin Kumar, Introduction to Data Mining (2005).
5. Cipolla, Emilt. Data Mining: Techniques to Gain Insight into Your Data.
6. Wang, X.Z. (1999) Data mining and knowledge discovery for process monitoring and control. Springer, London.
7. C. Angeli, S. K. Howard, J. Ma, J. Yang, and P. A. Kirschner, “Data mining in educational technology classroom research: can it make a contribution?” Computers & Education, vol. 113, pp. 226–242, 2017.
8. Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevier.
9. Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.
10. Kantardzic, M., (2011) Data Mining: Concepts, Models, Methods and Algorithms, Wiley-IEEE Press.
11. Wu, W., Lee, Y.T., Tseng, M.L. & Chiang, Y.H. (2010). Data mining for exploring hidden patterns between KM and its performance.Knowledge-Based Systems,23,397-401. doi:10.1016/j.knosys.2010.01.014
12. Rohit, B. Gupta, R. Kumar and A. Kumar, "Towards Information Discovery On Large Scale Data: state-of-the-art," 2018 International Conference on Soft-computing and Network Security (ICSNS), 2018, pp. 1-9, doi: 10.1109/ICSNS.2018.8573666.