(IJRSSH) 2014, Vol. No. 4, Issue No. IV, Oct-Dec **ISSN: 2249-4642**

REVIEW OF KNOWLEDGE DISCOVERY PROCESS (KDDS)

Mr. Ranbir

Assistant Professor

ABSTRACT

In this paper we will discuss and work with the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, of not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Data bases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

INTRODUCTION

The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transaction and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the “essence” of information stored, and the discovery of patterns in raw data. We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for amass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became over shelling.

WHAT KIND OF INFORMATION ARE WE COLLECTING?

We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files.

- **BUSINESS TRANSACTIONS**

Large department stores, for example, thanks to the widespread use of bar codes, store millions of transactions daily representing often terabytes of data. Storage space is not the major problem, as the price of hard disks is continuously dropping, but the effective use of the data in a reasonable time frame for competitive decision making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world (M.S. Chem, et al.) [9]. Every transaction in the business industry is often "memorized" for perpetuity. Such transaction are usually time related and can be inter-business deals such as purchases, exchanges,banking, stock, etc., or intra-business operations such as management of in-house wares and assets.

- **SCIENTIFIC DATA**

Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated. Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about oceanic activity, on in an American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed.

- **MEDICAL AND PERSONAL DATA**

From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies andorganizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele. Regardless of the security issues this type of data often reveals, this information is collected, used and even shared. When correlated with other data this information can light on customer behavior and the like.

- **SATELLITE SENSING**

There are a countless number of satellites around the globe: Some are geo-stationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. NASA, which controls a large number of satellites, received more data every second than what all NASA researchers and engineers can cope with. Many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.

- **GAMES**

Our society is collecting a tremendous amount of data and statistics about games, players and athletes. From hockey scores, basketball passes and car-racing lapses, to swimming times,boxer's pushes and chess positions, and all the data are stored. Commentators and journalists are using this information for reporting, but trainers and athletes would want to exploit this data to improve performance and better understand opponents.

- **DIGITAL MEDIA**

Many radio stations, television channels and film studio are digitizing their audio and video collections to improve the management of their multimedia assets. Data mining such as the NHL and the NBA have already started converting their huge game collection into digital forms. In addition the proliferation of cheap scanners, desktop video cameras and digital cameras is one of the causes of the explosion in digital media repositories.

- **COMPUTER ANALYSIS DESIGN (CAD) AND SOFTWARE ENGINEERING DATA**

There are multitudes of Computer Analysis Design (CAD) system for architects to designbuildings or engineers to conceive system components or circuits (J Han et al) [10], These systems are generating a tremendous amount of data. Moreover, software engineering is a source of considerable similar with code, function libraries, objects, etc., which need powerful tools for management and maintenance.

- **VIRTUAL WORLDS**

There is a remarkable amount of virtual reality object and space repositories available. Management of these repositories as well as content-based search and retrieval from these repositories are still research issues, while the size of the collection continues to grow. There are many applications making use of three dimensional virtual spaces. Ideally, these virtual spaces are described in such a way that they can shares objects and places.

- **TEXT REPORT AND MEMOS (E-MAIL MASSAGE)**

Most of the communications within and between companies or research organization or even private people, are based on reports and memos in textual forms often exchanged by e-mail. These messages are regularly stored in digital forms for future use and reference creating formidable digital libraries.

- **THE WORLD WIDE WEB REPOSITORIES**

Many believe that the World Wide Web will become the compilation of human knowledge. Since the inception of the World Wide Web in 1993, documents of all sorts of formats, contents and description have been collected and inter-connected with hyperlinks making it the largest repository of data ever built. Despite its dynamic and unstructured nature, its heterogeneous characteristic, and its very often redundancy and inconsistency, the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers.

WHAT ARE DATA MINING AND KNOWLEDGE DISCOVERY?

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

DATA CLEANING: also known as data cleansing, it is a phase in which noise data andirrelevant data are removed from the collection.

DATA INTEGRATION: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

KNOWLEDGE REPRESENTATION: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a preprocessing phase to generate a data warehouse. Date selection and data transformation can also be combined where the consolidation of the date is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data. The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measure can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results. Data mining derives its name from the similarities between searching for valuable information in a large database and mining. Both imply either sifting through a large amount of material the material to exactly pinpoint where the values reside. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

CONCLUSION

In this research paper we defined that how we get or acquire data and where from we can get the data. After that I described that how we manage that data and how we can store in a structured manner in our data warehouse.

REFERENCES

1. Berry MJA and Linoff GS. 2000. Mastering Data mining: The art and science of customer relationship management, Canada Wiley.
2. New W. 2004. Pentagon failed to study privacy issues in data mining
3. Verykios, VS; Bertino, E; Fovino, IN; Provenza, LP; Saygin, and Theodoridis, Y. 2004. State-of –the-art in Privacy Preserving Data Mining. SIGMOD Record. Volume 33, Issue 1:50-57.
4. Wang J. 2003. Data Mining Challenges and Opportunities. London, IRM Press.
5. Agrawal, R.; and Srikant, R. 2000. Privacy-Preserving Data Mining. In Proceedings of the ACM SIGMOD International Conference of Data, 439-450.
6. Lindell, Y.: and Pikas, B. 2000. Privacy Preservation Data Mining. In Advances in Cryptology-CRYPTO 2000.
7. Goldreich, O.; Micali, S; and Wigdeerson, A 1987. How to Play any mental Game. In Proceedings of the Nineteenth annual ACM Symposium on the theory of computing, 218-299.
8. Evfimievski, S. 20002. Randomization Techniques for Privacy Preservation Association rule mining. SIGKDD Explorations 4(2); 43-48.
9. 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.
10. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
11. G. Piatetsky-Shapiro, U.M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT press, 1996.