International Journal of Computing and Applications | **ISSN: 0973-5704**
Volume 13, Number 1, (January-June 2018), pp 51-54

MINING KNOWLEDGE FROM A DATABASE: AN APPROACH USING DATA CLEANSING DATA INTEGRATION AND TEXT JOINS

Manu Sehgal[1] and Deepshikha Bhargava[2]

[1]*Assistant Professor, PG Department, of Information Technology, GGDSD College, Chandigarh, India. Email: manuleo12@gmail.com*
[2]*Professor, Deputy Director & Head of Institution, Amity, Institute of Information Technology, Amity University, Rajasthan, Jaipur, India. Email: deepshikhabhargava@gmail.com*

Abstract: Data mining is a methodology for having a deep study of various applications in a database that helps to recognize for various unknown patterns in data that can be used to predict future behavior of various statistics in a database. Basically data mining is a tool not to change the appearance but to notice unknown interrelation between the data. We may apply techniques which can be used for data integration, data cleansing and improve the quality of data saved in the database.

Keywords: Data mining, data cleansing, text joins, data integration.

Nomenclature

(i) *Data Mining:* Data mining is one step at the core of the knowledge discovery process, dealing with the extraction of patterns and relationships from large amounts of data.

(ii) *Data Cleansing:* Data scrubbing, also called data cleansing, is the process of amending or removing data in a database that is incorrect, incomplete, improperly formatted, or duplicated.

(iii) *Text Joins:* Text-to-Join is a great way for you to collect email addresses from contacts through a text message.

(iv) *Data Integration:* Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information.

1. INTRODUCTION

Many people have a myth that database is just a collection of data that helps in data storage and retrieval. Basically a database rich and interrelated data and information, which can form set of very large, interconnected, mixed information networks. Various data can be extracted from these information networks.. In this paper, we have a brief overview of database which is a collection of many information networks and we also discuss the ways how information networks can be used to improve data quality and consistency, makes the process easy for data integration, and we are able to create some interesting knowledge. This discussion involves literature review, database issues related to networks as well as data mining tasks related to the network and how we can use networks for data cleaning how to recognize or exploit data from a range of knowledge which is avail able on various data networks.

General tasks related to a database are basically indexing, recovery of data, and updations etc. we have observed that various objects in databases are not independent tuples but they are, inter-related data that can explored. Various researches done earlier has not given attention towards the linked data or interconnected data.

2. EVOLUTION OF DATA MINING

Even today we are searching the database which is not updated for years and years. We had seen a rapid growth of electronic data management techniques for searching the desired data. Every year extra operations

are being automated on the data found from various data bases All these information grasp useful data and use various methods, that can be helpful to get better business decisions making also can improve the chances of success.

(I) Firstly to give a outline of an already accessible methods which plays a important role for searching practical data from databases.

(II) Secondly it also provides a feature of categorization of data that recognizes significant features to have a deep insight of information detection and data mining software tools.

(III) Thirdly to explorer the beforehand accessible information discovery and data mining software tools .Therefore These tools may prove to be either gainful packages available for buying.

(IV) Fourthly it identifies various unique features that discovery software should have so that they can attract various users for searching the data in the databases.

3. LITERATURE REVIEW

This paper addressed problems of similarity-based operations in data integration. Data integration has been a topic of research for more than twenty years and has gained a growing interest over recent years because of the continuously increasing availability of data from sources in local or global scopes. Some of the work done on this topic is

- "Text Joins for Data Cleansing and Integration in an RDBMS "Luis Gravano Panagiotis G. Ipeirotis Nick Koudas Divesh Srivastava 2001" In this paper it has been discussed that the data provided in an organization is always noisy due to sometimes incomplete or incorrect information
- "Approximate String Joins in a Database (Almost) for Free "Luis Gravano Panagiotis G. Ipeirotis H.V. Jagadish Columbia University Columbia University University of Michigan 2003" This paper lay stress on management of sting data which is becoming a hot topic these days.
- "Duplicate Record Detection: A Survey " Ahmed K. Elmagarmid, Senior Member, IEEE, Panagiotis G. Ipeirotis, Member, IEEE Computer Society, and Vassilios S. Verykios, Member, IEEE Computer Society 2007" This paper is used to discuss about the compromises which are made to complete a database and about the quality of the data which leads to data entry errors
- "Problems, Methods, and Challenges in Comprehensive Data Cleansing "Heiko Müller, Johann-Christoph Freytag Humboldt-Universität zu Berlin zu Berlin 2014"his paper lays stress on the representation of data which is consistent, consolidate as well as free from duplication

4. MOTIVATION OF THE RESEARCH

Data mining is a step at the base level for knowledge discovery that deals with extraction of patterns and relationships from loads and loads. There is a drastic increase in demand for better decision support In this paper, we are going to give an summary of common information detection tasks, various approaches which can be used to solve these tasks, and accessible software techniques for employing these approaches.

1. *Integration of different techniques:* Various tools which are available these days use a one technique or various techniques. We have no problem with the types of techniques but our major concern is that which technique is best suitable to solve the problem The best possible tool is to provide a variety of different methods for solving various different kinds of problems.
2. *Extensibility:* One of the major drawback of this method is that different techniques outperform each other for different problems. With the increasing number of projected methods and reported applications, it becomes more clear that any fixed arsenal of algorithms will

never be able to cover all arising problems and tasks. Now we say that It is significant to provide an design that allows for easy usage of new methods, and revision of existing methods.

3. *Flawless incorporation with databases:* With the present data analysis we found that s products can be categorized in two forms. Firstly the one is drill-down analysis and reporting, which is provided by vendors of RDBMS's, These are systems which is going to provide a tight connection with the original database and deploy the processing power. They are also not allowed for testing customer provided hypotheses, moreover, automatically recognizing various patterns and models. Secondly we have a category which consists of (stand-alone) pattern detection tools, which are used to recognize various patterns in the data or information. These tools are used to access the database offline which means that; data is extracted from the database and then transferred into the detection engine. In addition, these tools are often too little prepared with data handing ability, which leaves the data preprocessing only to the user. Which can further lead repeating export-import processes.
4. *Helping provided to both analysis experts and novice users:* We usually have three stages at use age of KDD technology in an organization:
    1. Firstly potential of KDD is discovered. We perform naive studies which are performed, often by external consultants (who are basically data mining specialists).
    2. Secondly when we have proved profitability of KDD, then it is useful for solve business problems. We have various Users teams of analysis experts which have their expertise in KDD technology and also we have domain experts who have deep insight knowledge of the application domain.
    3. Third step basically leads to Full utilization of KDD knowledge inside the association. Various End users are free to carry out their analysis based on their individual choice and requirements without any restrictions.

5. RESEARCH METHODOLOGY

The methodology which we use for this approach is as follows:

Step I: We need to select a database (it can be an existing database or we can create a new database by adding data in fields according to our requirement.)

Step II: We need to find out the drawbacks or weakness in the database (that can be duplication of a record, null data, insignificant entry etc)

Step III: We apply various techniques like data cleansing, data integration, text joins and data mining

Step IV: We use SQL queries with joins to show few similarity based operations

Step V: Compare the two database one before operation and after some queries applied. Find the difference based on various parameters like execution time and storage space and processing time.

Step VI: Show the results of the compassion of the two databases.

6. CONCLUSION AND FUTURE SCOPE

This study has given a brief view of data mining methodology, database issues and data mining tasks. We have discussed various techniques for data mining and we can use one of the techniques for future for developing an algorithm or modify the previous existing algorithm for a new problem area.

REFERENCES

[1] A.N. Pathak, Manu Sehgal, Divya Christopher, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011. A Study on Fraud Detection Based on Data Mining Using Decision Tree.

[2] Ramesh C ponia, Shiv Kumar Gupta, Ritu Vijay "A compuratural Study of mobile data network" Oriental Journal of Computer Science and Technology (OJCST) ISSN 0974-6471, December 2011 Volume 4, No. 2. ublished Oreintal scientic publishing company India pp 387-392.

[3] Bhargava D., Sinha M., "Performance analysis of agent based IPSM" International Computer Science and Software engineering ICSSE 2012 joint conference on pp. 253-258, May 30 June 1, 2012. DOI 10-11-09/ ICSSE 2012.6261961.

[4] Eike Schallehn and Schallehn, Eike. "Efficient similarity-based operations for data integration", ULB Halle, Germany, HALCoRe, 1971.

[5] www.sigkdd.org

[6] Text Joins for Data Cleansing and Integration in an RDBMS "Luis Gravano Panagiotis G. Ipeirotis Nick Koudas Divesh Srivastava.

[7] Efficient Similarity-based Operations for Data Integration Prof. Dr. Gunter Saake, Prof. Dr. Kai-Uwe Sattler, Dr. Ralf- Detlef Kutsche.