Data Mining and Data Warehousing Techniques

RESHMA S. PANCHAL
Ph. D. Research Scholar
Department of Library and Information Science
Dr. Babasaheb Ambedkar Open University,
Ahmedabad

DR. PRIYANKI R. VYAS
Professor,
Department of Library and Information Science
Dr. Babasaheb Ambedkar Open University,
Ahmedabad

Absract:
The purpose of this paper is to aware LIS professionals and students about a Data mining and Data warehousing techniques This paper will discuss about the importance of Data mining process for research work and data storage. Different techniqueshave been adopted by large number of Libraries and Museums at inside and outside of India. This paper will highlight the brief description of whole process towards data mining, data warehousing and its process. And also focus about an applications and experiments using in data mining and warehousing process.

Keywords: *Data Mining, Data Warehousing, Harvesting, Data Processing*

1. Introduction

In Library Management study various research methodologies have been considered important for improving library functioning and services. In order to carry out analysis of user surveys various analytical techniques are used. In such analysis emphasis is given on library use study or user's study. Along with traditional method, statistical or computerised methods are also being used these days. In other fields especially industrial or commercial fields study on customer's behavior is given a great importance. In these fields use of little higher techniques like data mining and data warehousing are found more useful. Due to progress in computer technology such applications have now become easy. Application of data mining and data warehousing is growing in other fields also. Even this technique is also explored in library management to study library use and in user's behavior considering importance of analytical techniques in various fields including library and information science we are going to discuss Data Mining and Data Warehousing in this article. Before we look into the specific applications of data mining and data warehouse, it is worthwhile to know, what data mining and data warehousing technique is?

2. Review of Literature

A related literature has been referend to attempt towards preparing this conceptual paper and identified an important information from different source to re-write the concept of Data mining and Data warehousing process and techniques to make an easy understandability.

2.1 Data Mining and Data Ware Housing: Concept and Scope

If we look at the term "Data Mining or Data Warehousing," it reveals that in this analytical technique, data is used or database might be handled with some special technique. We know that between database and software a bridge of database management system exists, so we may guess that there may be some relation with DBMS or similar treatment to data. If we concentrate the word "Mining", we feel that the technique must be similar to process of exploration like done in mines. As in mining study of rock patterns or layers are studied, similarly in data mining also data patterns might be studied. With this understanding let us look into the actual concept of data mining and data warehousing. Since the data mining technique is largely used in industrial or commercial fields, let us first know the concept as used in these fields.

2.2 Data Mining

Data mining means by inter connecting different databases attempt are made to find out hidden patterns in them, which were never noticed before. In order to search data samples, based on data pattern from the databases which are stored in warehousing are used. In data mining exploration statistical techniques like Regression, Classification and Clustering methods are used. Now first let us try to know about "data mining" and then about "data warehousing".

2.3 Data Warehouse

Data warehouse means when the activities with involvement of database and software for performing specific transactions are carried out and after transactions are completed, data is removed from the system and stored into data warehouse elsewhere. In a normal way after completion of various computerised processes finally data is stored back into the same system database. However, when the data is no longer required in the system it should be removed and kept in data warehouse as all the processes/ transactions are completed in all the respects. The data thus stored in data warehouse is used for data mining for the purpose of finding patterns or data samples out of transactions or processes occurred in the past. Out of the stored back data bases some useful data bases are selected and specific data sample located and used for data mining analysis.

The above explanation reveals that for data mining analysis database from warehouse are used. Now after understanding about data ware house let us try to know about data warehousing.

2.4 Data Warehousing

Data warehousing means from the database stored in data warehouse selection of specific database suitable for data mining is done. In this process after selecting useful databases, their cluster or group is made and some process is carried out to make it usable for data mining. All these processes are collectively called data warehousing.

In data mining analysis it is expected to trace the hidden data patterns or data layers which were so far never been explored. Based on the existence of these hidden data patterns future prospective of forecasting could be predicted. These forecasting helps to take decisions. Data mining analytical techniques are prominently used in dustry and business fields. Data mining technique can reveal data patterns of the customer's behavior which were never understood or noticed so far.

These data layers or patterns are used to predict the future trend of business growth and accordingly growth of any industry or business is planned to achieve the growth targets.

Through the above paras we have touched upon concepts of data mining, data ware house data warehousing and in short about data mining analytical process. In reality these processes are complex in nature but with the help of computing facility the expected results are achieved without much difficulties.

3. Steps of data mining process

The major steps in data mining analysis are as follows:

1.To prepare data pattern/sample:

First the sample data pattern, based on which the hidden data layer or pattern is to be traced. Then pattern is to be prepared considering a specific problem. This is the first step in the process.

2.Data warehousing:

Once target data pattern corresponding to problem is prepared the suitable databases are selected from the data ware house in which it is likely to get data similar to data sample and cluster of them is prepared. Once useful databases are selected, all the personal or confidential details are removed from them in order to make them anonymous to prepare improved database structure.

4. To select Algorithm

Now considering the target problem which was used to trace the data pattern a suitable algorithm is chosen. The main algorithm in this case are Regression and Classification. In regression method numerical data is handled. In this method data is handled in terms of quantity for example weight, speed or age, where as non numerical data is handled by in classification type of algorithm. In this case descriptive or qualitative things are handled such as colour, name, gender etc.

5. To select suitable software for analysis

Once the above three steps are completed, then comes the last step of data mining analysis. This is to choose the available analytical package, such as SPSS, SAS or S-Plus to analyse the data and present the conclusions/results. Finally, the results are used to provide solution to the problem which was selected for locating the data patterns in the first step in data mining analysis.

So far through the above paras we have understood the principal base of data mining and data warehousing concepts. Now let us try to understand how this technique is applied in the practical life. We are going to have a non business example e.g. Library management.

6. Application of Data Mining and Data Warehousing Technique in Library Management

Discussion about the use of data mining and data warehousing in the library and information science was first initiated by Nicolson and others. While doing this the first use of term "Bibliomining" was also done by them. The reason for exploring this new concept was felt because when one tries to use technique of data mining in library and information field its result was heading towards the software and data bases. Actually, search should direct towards behavior of library users. In order to have solution to this problem Nicolson &Staton decided to use slightly modified data mining technique known as bibliomining concept. While doing this they also found similarity with the concept of bibliometrics, which is used for exploring research on finding streams of scientific communication.

Up till now in studying relating to user group, emphasis was on using frequency evaluation and aggregate measure type of evaluation. However, in such evaluation useful several datasets/ streams or patterns were not found as they remained hidden. It is challenging to dig out such patterns. However, is possible by using data mining and data warehousing techniques. This makes it possible to understand specific needs of user groups and by which services they would be satisfied. It would be also possible to forecast about the future services and to take necessary decisions by using this new technique. Now let us try to know about how the concept of data mining and data warehousing is applied for study in library and information field.

In this process by applying statistical and pattern recognition tools on comprehensive data available in library systems it would be possible to predict the decisions about which library services need to be introduced. There are the following six steps in data mining process such as:

1. To identify area of the problem (in vision) and prepare data sample relevant to it.
2. To select databases relevant to targeted problem/data sample from the data warehouse
3. Create data warehousing
4. Select statistical data analysing tools
5. Choose data mining analytical tools
6. To prepare analysis report using traditional tools to present conclusion and prepare implementation plan.

Now let us discuss the above steps to understand them clearly.

7. Steps of Data Mining and Data Ware Housing

7.1 To fix Target problem Area

The first step in data mining is to fix the problem and its scope. The problem area could be specific or general. The information about the target problem can be located and relevant data sample is prepared.

While searching the information, it is decided first whether data mining can be done direct or indirect. Suppose library management wishes to decide to reduce the manpower in the library as library's budget has been cut. Naturally manpower on the circulation counter has to be reduced first which is going to impact on the facilities provided by the circulation counter and there would be cut in the services also. In this situation it becomes necessary to study how readers are going to face difficulties and inconvenience. This type of data mining could be of direct data mining. In case of indirect data mining example could be as, previous to this situation readers used to get messages as part of services directly to users in their department as such they need not have to come to library personally and their visits to library would tend to reduce. One can find solution to this problem via indirect data mining way.
While pursuing regional problem through data mining there would be need to use various tools. Searching sample of data pattern pertaining to a specific problem is not an easy task. It is necessary to use internal as well as external data together and one also needs to select comprehensive group of data bases as it is very difficult to get frequency or reoccurrence of a particular data pattern. It becomes therefore necessary to first prepare competent structure of data warehousing

7.2 Selecting Databases Corresponding to Problem

Once the type of data mining has been decided and data sample to relevant to problem is identified, next step is to select databases corresponding to data sample from warehousing. One needs all types of data for example transactional, not consolidated data as well as normal data. In many problems one needs system based local data as well as the external information. In most of the systems once transactions are completed data gets deleted. In order to understand this, let us see the following examples.

Example: Book issued by readers has been returned to library and it is recorded in the system. With this, issue and return transaction(cycle) is completed and the data about it automatically gets deleted. Second example is the reader reserves the book which he/she needs but it is currently issued to someone else, hence reservation data is created in the system. After some time, previous reader returns the book to library which is also recorded in the system. The next reader immediately gets a message about it as his claim gets activated. He comes to the library and gets the books issued. With this reservation, claim activation and claimant getting the book the entire transaction(cycle) gets completed. Above are the examples of defective library software. In standard library software it does not happen i.e. transaction data is never lost even after completing it. In western countries for security reasons such transaction data is purposely removed from the main system and kept in data warehouse. Under any circumstances it is better to remove such data and keep it in data warehouse which helps in cleaning database and secondly removed data is stored in warehouse for further use. In data mining point of view automatically deleting such data from system or removing it purposely both are harmful which makes it difficult to get data for data mining analysis. Best way is to store it in data warehouse by removing personal or confidential details from the database or to make it anonymous. Such data is always useful for data mining.

As referred above for data mining one needs operational or transactional data from local databases so also one needs data from external type to support it. In library data about readers is stored in software database or it is stored in servers of the institute. For example, reader's data, their respective transactional data, data about readers browsing from e-resources or from web sites on internet. The data even if it is of internal type sometimes it becomes difficult to retrieve external data about readers if it is available outside the system. Such data could be with departmental system or with the central system of parent institute.

Readers could also be concerned about their existence in social data society or in geographic location databases or in population or census data or it could be in pin code database etc.

So far, we have understood that we can find out the answer to some questions by using the software data. However, with the help of data which is invisible or so for remained hidden in the system can be traced from combination of system data and other external databases which could be used to predict the future possibilities and by analysing it, solution to the problems could be found and accordingly the decisions to provide values added services to the user is taken if analysis conclusion suggest it. In order to locate such hidden data, data mining technique is used and such data can be used for studying the problems relating to reader's difficulties. As the problem is simple or difficult corresponding data mining process needs to be chosen. In the next para we are going to look in to the creation of data ware housing and about handling databases for it.

7.3 Creation of Data Warehousing

Once it is ensured that data/ database for data mining are available then next task is to select useful databases from the data ware house. Preparing cluster of databases it is a challenging task. Databases selected for data warehousing need to be reformatted reformat in which steps like cleaning and anonymising are involved. The main purpose behind this is to remove individual and confidential details from the databases. Next step is to create improvised database cluster and convert old data in to the new cluster. Now let us see how is actually done through an example. Librarian prepares a query pertaining to the current problem, collects the databases from data warehouse useful for data mining and they are inter-connected through common fields. This makes a cluster of databases. From this consolidated data, cluster of information of individual or confidential type are removed to make it anonymous. The database thus finally formed is flat type database, which is going to be used for data mining analysis. The database structure is now ready to receive the data in it. The data mining process suggest which data from the system the librarian should decides to preserve and transfer to data warehouse and which is to be omitted or destroyed as such data may not be necessary to keep in the system. Preserved data would be useful for data mining. Analysis requires removing individual detailed from it by filtering the same. Maintaining data with individual and confidential data in the software is risky as it could be misused.

7.4 Selecting Analytical Tools

Once data ware housing process is completed means preparations needed for data mining analysis are done. Next task is to look for suitable data mining analytical tools. Data pattern needed for solving the problem is fixed and in the point of view of the problem the entire data is handled. While doing this expected data pattern are verified. At the end of analysis conclusion report is prepared. Sometimes system-built provisions are also used for this purpose. These days library management systems (LMS) provide tool to retrieve data samples by writing a small script and after running script data samples can be searched needed for data warehousing process. There are tools available like Online Analytical Processing (OLAP) tools which also could be used.

7.5 Statistical Packages available for Data Mining

The main objective of data mining process is to search data patterns corresponding to an identified problem which is suitable for data mining for exploring the solution. For this several tools like statistical technique to artificial intelligence are used. As we have seen in the definition of data mining it is about gathering data samples which were hidden till now and which are matching with the target problem. In order to search such sample automatic analytical technique tools are used. In data mining process two broad categories of processes are involved-1.to describe and 2.to predict. In the first category it is confirmed and fixed the availability of expected data patterns, where as in the second category based on the possibility data which could be possible but not visible is predicted Nowadays readymade software packages suitable for data mining are available. Data obtained from data warehousing process is imported from the respective databases, analysed and for finding conclusions from analysis suitable tools are used. An algorithm suitable for the solution predicted is selected and used with analytical tools to arrive at a conclusion. The list of analytical tools is as follows:

1.SAS - Statistical Analysis software

2.SPSS - Statistical Package for Social Sciences
3.S-Plus
4.Oracle Data Mining Suite (Darwins)
5.Microsoft SQL Server
6.Free walk data mining suite - Open-source type.

7.6 Implementing the Conclusion

Once report of the data mining process (static report or decision report) is ready the conclusion it is once again tested with the original data base and credibility of conclusion is verified. It i ascertained whether conclusion or solution to the problem is acceptable. If it is not acceptable then the data mining process is revised by refixing the problem and matching sample and data mining process is carried out once again. However, if the decision or conclusion is acceptable then further task structure for decision implementation is prepared and it is implemented.

8. Conclusion

In this article we have discussed the concept of data mining and data warehousing. About how this technique is applied in industry/business fields. Subsequently we explored whether this technique could be used in library management research. We understood the concept of data mining first, then about the steps of use of data mining technique, then about data warehouse, next about data ware housing and finally about analysing data by using data mining and statistical tools. At the end of the article we have understood about library management application by referring two live examples. In short data mining technique could be listed in the following steps:

1. According to data mining problem fix the target area
2. Search the data mining pattern matching with the problem
3. Carryout data warehousing process once databases from data warehouse are selected, cleaned by removing person specific details.
4. Data imported from comprehensive data bases in the relevant data structure
5. Accept an algorithm matching with data pattern and the problem.
6. Analyse data by using data mining and statistical tools.
7. Draw the conclusion, prepare report, verify the result and make decision based on conclusions and implement it.

Reference

1. Ertan, G., Comfort, L., & Martin, O. (2023). Political polarization during extreme events. *Natural Hazards Review, 24*(1), doi:10.1061/(ASCE)NH.1527-6996.0000603
2. Huang, C., Zhang, Q., Guo, D., Zhao, X., & Wang, X. (2023). Discovering association rules with graph patterns in temporal networks. *Tsinghua Science and Technology, 28*(2), 344-359. doi:10.26599/TST.2021.9010090
3. Kaur, R., Ginige, J. A., & Obst, O. (2023). AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Systems with Applications, 213*, doi: 10.1016/j.eswa.2022.118997
4. Keskin, S., & Yazıcı, A. (2023). FSOLAP: A fuzzy logic-based spatial OLAP framework for effective predictive analytics. *Expert Systems with Applications, 213*, doi: 10.1016/j.eswa.2022.118961
5. Khaleghi, N., Rezaii, T. Y., Beheshti, S., & Meshgini, S. (2023). Developing an efficient functional connectivity-based geometric deep network for automatic EEG-based visual decoding. *Biomedical Signal Processing and Control, 80*, doi: 10.1016/j.bspc.2022.104221
6. Li, G., Xu, S., Wang, S., & Yu, P. S. (2023). Forest based on interval transformation (FIT): A time series classifier with adaptive features. *Expert Systems with Applications, 213*, doi: 10.1016/j.eswa.2022.118923

7. Liu, Y., Dai, H., Li, J., Chen, Y., Yang, G., & Wang, J. (2023). BP-model-based convoy mining algorithms for moving objects. *Expert Systems with Applications, 213*, doi:10.1016/j.eswa.2022.118860
8. Moon, J., Posada-Quintero, H. F., & Chon, K. H. (2023). A literature embedding model for cardiovascular disease prediction using risk factors, symptoms, and genotype information. *Expert Systems with Applications, 213*, doi:10.1016/j.eswa.2022.118930
9. Nicholson S. (2003). Avoiding the great data wipe out-Three. *American Libraries, 34*(9), p. 36.
10. Nicholson, S. (2003). The bibliomining process: Data warehousing and data mining for library decision making. *Information technology for Libraries, 22*(4).
11. Pascual Espada, J., SolísMartínez, J., Cid Rico, I., & Emilio Velasco Sánchez, L. (2023). Extracting keywords of educational texts using a novel mechanism based on linguistic approaches and evolutive graphs. *Expert Systems with Applications, 213*, doi: 10.1016/j.eswa.2022.118842
12. Qiu, J., Li, T., Lü, F., Huang, Y., Li, C., Zhang, H., ... He, P. (2023). Molecular behavior and interactions with microbes during anaerobic degradation of bio-derived DOM in waste leachate. *Journal of Environmental Sciences (China), 126*, 174-183. doi: 10.1016/j.jes.2022.04.015
13. Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., ... Cavedon, L. (2023). The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys, 55*(2), doi:10.1145/3490234
14. Shi, B., Dong, B., Xu, Y., Wang, J., Wang, Y., & Zheng, Q. (2023). An edge feature aware heterogeneous graph neural network model to support tax evasion detection. *Expert Systems with Applications, 213*, doi: 10.1016/j.eswa.2022.118903
15. Venkata, P., Pandya, V., & Sant, A. V. (2023). Data mining model based differential microgrid fault classification using SVM considering voltage and current distortions. *Journal of Operation and Automation in Power Engineering, 11*(3), 162-172. doi:10.22098/joape.2023.10185.1722
16. Widyantara, I. M. O., Hartawan, I. P. N., Karyawati, A. A. I. N. E., Er, N. I., &Artana, K. B. (2023). Automatic identification system-based trajectory clustering framework to identify vessel movement pattern. *IAES International Journal of Artificial Intelligence, 12*(1), 1-11. doi:10.11591/ijai.v12.i1.pp1-11
17. Bibliomining. www.bibliomining.com
18. Literature written by Nicholoson, S. on bibliomining http://bibliomining.com/nicholson.