International Journal of Information Technology and Management

Vol. VIII, Issue No. XI, February-2015, ISSN 2249-4510

REVIEW ARTICLE

DATA MINING USING CLUSTERING: A RESEARCH

International Journal of Information Technology and Management
Vol. VIII, Issue No. XI, February-2015, ISSN 2249-4510

Data Mining Using Clustering: A Research

Jyoti[1] Mr. Kaushal[2]

[1]M.Tech Scholar

[2]Associate Professor

---------------------------♦-----------------------------

INTRODUCTION

The technologies for generating and collecting data have been advancing rapidly. At the current stage, lack of data is no longer a problem; the inability to generate useful information from data is! The explosive growth in data and database results in the need to develop new technologies and tools to process data into useful information and knowledge intelligently and automatically.

Data mining (DM), therefore, has become a research area with increasing importance (Weiss and Indurkhya, 1998; Technology Forecast, 1997; Fayyad et al., 1996; Piatetsky-Shapiro and Frawley, 1991).

DM is the search for valuable information in large volumes of data (Weiss and Indurkhya, 1998). It is the process of nontrivial extraction of implicit, previously unknown and potentially useful information such as knowledge rules, constraints, and regularities from data stored in repositories using pattern recognition technologies as well as statistical and mathematical techniques (Technology Forecast, 1997; Piatetsky-Shapiro and Frawley, 1991). Many companies have recognized DM as an important technique that will have an impact on the performance of the companies.

DM is an active research area and research is ongoing to bring statistical analysis and artificial intelligence (AI) techniques together to address the issues.

Current trends on data mining Just five years ago, only 50 researchers took part in the knowledge discovery and data mining conference workshop. Today, however, knowledge discovery nuggets, the well-known monthly electronic newsletter by Gregory Piatetsky-Shapiro, has more than 4,000 readers. Moreover, data mining continues to attract more and more attention in the business and scientific communities.

In a 1997 report, Stamford, Conneticut-based Gartner Group mentioned: ``Data mining and artificial intelligence are at the top of the five key technology areas that will clearly have a major impact across a wide range of industries within the next three to five years.'' Many companies currently use computers to capture details of business transactions such as banking and credit card records, retail sales, manufacturing warranty, telecommunications, and myriad other transactions. Data mining tools are then used to uncover useful patterns and relationships from the data captured.

Currently, data mining techniques, tools, and researches are being expanded to the various fields. For example, the DM tool, intelligent text-mining system, extracts text fragments relevant to user queries, automatically creates and processes a series of new queries, and assembles a new text. The output enables the user to see the new aspects of a given theme. This tool is a rule based system using complex heuristics.

Data warehousing is one of the most important research areas related to DM. A data warehouse is a read-only database developed for analyzing business situations and supporting decision makers. The data warehouse includes large volumes of subject oriented data, where all levels of an organization can find the information in a timely manner. DM goes together with the data warehousing which is necessary to organize historical information gathered from large-scale client/server-based applications. In other words, DM can add values to the information assets of organizations in different sectors, through effective induction of large corporate data warehouses into a client-server. Therefore, developing an advanced client-server induction system capable of supporting efficient and effective data mining of databases in business environment is one of the active research areas.

REQUIREMENTS AND CHALLENGES OF DM

DM is a relatively new field and there are many challenges to be faced. Extracting useful information from data can be a complicated and sometimes a difficult process. In this section, we look at some of the requirements and challenges of data mining (adapted from Chen et al., 1996).

ABILITY TO HANDLE DIFFERENT TYPES OF DATA

Many database systems have complex data types, such as hypertext, multimedia data, and spatial data. If a DM technique is robust and powerful, it should be able to perform effective DM on various types of data structures. Though ideal, it is impractical to expect a DM technique to handle all kinds of data and to perform different goals of DM effectively. In general, a specific DM system is built for mining knowledge from a specific kind of data.

GRACEFUL DEGENERATION OF DM ALGORITHMS

The DM algorithms should be efficient and scaleable. The performance of the algorithm should degenerate gracefully. In other words, the searching, mining, or analyzing time of a DM algorithm should be predictable and acceptable as the size of the database increases.

VALUABLE DM RESULTS

DM system should be able to handle noise and exceptional data efficiently. The discovered information must precisely depict the contents of the database and be beneficial for certain applications. Also, the quality of the discovered information should be interesting and reliable.

REPRESENTATION OF DM REQUESTS AND RESULTS

DM identifies facts or conclusions based on sifting through the data to discover patterns or anomalies (Technology Forecast, 1997). To be effective, the systems should allow users to discover information from their own perspectives and the information should be presented to the users in forms that are comfortable and easy to understand. Highlevel query languages or graphical user interface is required to express the DM requests and the discovered information.

End users should be able to specify task commands for the DM system and the results from the DM system should be understandable and usable.

Mining at different abstraction levels It is very difficult to specify exactly what to look for in a database or how to extract useful information from a database. Besides, the value of a piece of information is in the eyes of the beholder ± one person’s ``gold mine’’ could easily be another person’s garbage. To facilitate the mining process, the systems should allow the users to mine at different abstraction levels. For example, a high-level query might disclose an interesting trace that warrants further exploration. Thus, it is important for DM tools to support mining at different levels of granularity.

Mining information from different sources of data In the ages of the Internet, Intranets, Extranets, and data warehouses, many different sources of data in different formats are available. Mining information from heterogeneous database and new data formats can be challenges in DM. The DM algorithms should be flexible enough to handle data from different sources.

Protection of privacy and data security DM is a threat to privacy and data security because when data can be viewed from many different angles at different abstraction levels, it threatens the goal of keeping data secured and guarding against the intrusion on privacy. For example, it is relatively easy to compose a profile of an individual (e.g. personality, interests, spending habits, etc.) with data from various sources.

DATA MINING STEPS

In general, there are three main steps in DM: preparing the data, reducing the data and, finally, looking for valuable information. The specific approaches, however, differ from companies to companies and researchers to researchers. For example, IBM (reported in Technology Forecast, 1997) defined four major operations for DM:

1. Predictive modeling: using inductive reasoning techniques such as neural networks and inductive reasoning algorithms to create predictive models.

2. Database segmentation: using statistical clustering techniques to partition data into clusters.

3. Link analysis: identifying useful associations between data.

4. Deviation detection: detecting and Fayyad et al. (1996),

A REVIEW OF DATA MINING TECHNIQUES

1. Industrial Management & Data Systems 101/1 [2001] 41±46 1 Retrieving the data from a large database.

2. Selecting the relevant subset to work with.

3. Deciding on the appropriate sampling system, cleaning the data and dealing with missing fields and records.

4. Applying the appropriate transformations, dimensionality reduction, and projections.

5. Fitting models to the preprocessed data.

Classifying DM techniques

Many DM techniques and systems have been developed and designed. These techniques can be classified based on the database, the knowledge to be discovered, and the techniques to be utilized. In this section, we review one of the classification schemes proposed by Chen et al. (1996).

Based on the database

There are many database systems that are used in organizations, such as relational database, transaction database, object-oriented database, spatial database, multimedia database, legacy database, and Web database. A DM system can be classified based on the type of database it is designed for. For example, it is a relational DM system if the system discovers knowledge from relational database and it is an object-oriented DM system if the system finds knowledge from object-oriented database.

Based on the knowledge

DM systems can discover various types of knowledge, including association rules, characteristic rules, classification rules, clustering, evolution, and deviation analysis. DM systems can also be classified according to the abstraction level of the discovered knowledge. The knowledge may be classified into general knowledge, primitive-level knowledge, and multiple level knowledge.

Based on the techniques

DM systems can also be categorized by DM techniques. For example, a DM system can be categorized according to the driven method, such as autonomous knowledge mining, data driven mining, query-driven mining, and interactive DM techniques. Alternatively, it can be classified according to its underlying mining approach, such as generalization based mining, pattern-based mining, statistical- or mathematical-based mining and integrated approaches.

Major DM techniques

In this section, we review and discuss the major DM techniques.

STATISTICS

Statistics is an indispensable component in data selection, sampling, DM, and extracted knowledge evaluation. It is used to evaluate the results of DM to separate the good from the bad. In data cleaning process, statistics offer the techniques to detect ``outliers'', to smooth data when necessary, and to estimate noise. Statistics can also deal with missing data using estimation techniques.

Techniques in clustering and designing of experiments come into play for exploratory data analysis. Work in statistics, however, has emphasized generally on theoretical aspects of techniques and models. As a result, search, which is crucial in DM, has received little attention. In addition, interface to database, techniques to deal with massive data sets, and techniques for efficient data management are very important issues in DM. These issues, however, have only begun to receive attention in statistics (Kettenring and Pregibon, 1996).

TECHNIQUES FOR MINING TRANSACTIONAL/ RELATIONAL DATABASE

Mining association rules in transactional or relational database has been the most attractive topics in database field (Agrawal et al., 1993; Han and Fu, 1995;Mannila et al., 1994; Savasere et al., 1995; Srikant and Agrawal, 1995). The task is to derive a set of strong association rules in the form of ``A1 ^ . . . ^ Amˆ)B1 ^ . . . ^ Bn,'' where Ai …for i 2 f1; . . . ;mg† and Bj…for j 2 f1; . . . ; ng) are attribute-value sets, from the associated data sets in a database. For example, one may find the association rule: if a customer buys one brand of beer, he/she usually buys another brand of chips in the same transaction. Because mining association rules might require scanning through a massive transaction database repeatedly, the required processing power could be enormous. Artificial intelligence (AI) techniques AI techniques are widely used in DM.

Techniques such as pattern recognition, machine learning, and neural networks have received much attention. Other techniques in AI such as knowledge acquisition, knowledge representation, and search, are relevant to the various process steps in DM.

CONCLUSION

Having the right information at the right time is crucial for making the right decision. The problem of collecting data, which used to be a major concern for most organizations, is almost resolved. In the millennium, organizations will be competing in generating information from data and not in collecting data. Industry surveys indicated that over 80 percent of Fortune 500 companies believe that data mining would be a critical factor for business success by the year 2000 (Baker and Baker, 1998). Obviously, DM will be one of the main competitive focuses of organizations. Although progresses are continuously been made in the DM field, many issues remain to be resolved and much research has to be done.

PROPOSED WORK

Clustering Process: In proposed algorithm, the input remains in the same order in which data items are entered. The whole process is divided into two phases.

Phase-I: In phase-I, the cluster size is fixed and the output of the first phase forms initial clusters. Here, the input array of elements is scanned and split up into sub-arrays, which represent the initial clusters.

Phase-II: In phase-II, the cluster sizes vary and the output of this phase are the finalized clusters. Initial clusters are inputs for this phase. The centroids of these initial clusters are computed first, on the basis of which distance from other data elements are calculated. Furthermore the data elements having less or equal distance remains in the same cluster otherwise they are moved to appropriate clusters. The entire process continues until no changes in the clusters are detected.

REFERENCES

Agrawal, R., Imielinski, T. and Swami, A. (1993) . Mining Association Rules between Sets of Items in Large Databases, Paper presented at the ACM SIGMOD, May.

Baker, S. and Baker, K. (1998), ``Mine over matter'',Journal of Business Strategy, Vol. 19 No. 4, pp. 22-7.

Chen,M.S., Han, J. and Yu, P. (1996) , ``Data mining: an overview from a database perspective'', IEEE Transactions on Knowledge and Data Engineering, Vol. 8 No. 6, pp. 866-83.

Etzioni, O. (1996), ``The World-Wide Web: quagmire or gold mine?'', Communication of the ACM, Vol. 39 No. 11, pp. 65-8.

Fayyad, U., Djorgovski, S.G. and Weir, N. (1996), ``Automating the analysis and cataloging of sky surveys'', in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, pp. 471-94.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) , ``From data mining to knowledge discovery: an overview'', in Fayyad, U., Piatestsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds), Advances in Knowledge Kettenring, J. and Pregibon, D. (1996) , Committee on Applied and Theoretical Statistics: Workshop on Massive Data Sets, Paper presented at the National Research Council, Washington, D.C.

Lu, H., Setiono, R. and Liu, H. (1996) , ``Effective Data Mining Using Neural Networks'', IEEE Transactions on Knowledge and Data Engineering, Vol. 8 No. 6, pp. 957-61.

Mannila, H., Toivonen, H. and Verkamo, A.I. (1994), Effective Algorithms for Discovering Association Rules, paper presented at the AAAI Workshop, Knowledge Discovering in Databases, July.

Piatetsky-Shapiro, G. and Frawley, W.J. (1991), Knowledge Discovery in Database, AAAI/MIT Press.

Savasere, A., Omiecinski, E. and Navathe, S. (1995), An Effective Algorithm for Mining Association Rules in Large Databases, paper presented at the 21st International Conference, Very Large Data Bases, September.

Srikant, R. and Agrawal, R. (1995), Mining Generalized Association Rules. Paper presented at the 21st International Conference, Very Large Data Bases, September. Technology Forecast: 1997 (1997), Price Waterhouse World Technology Center, Menlo Park, CA Tufte, E.R. (1983), The Visual Display of Quantitative Information, Graphics Press, Cheshire, CN.

Tufte, E.R. (1990), Envisioning Information, Graphics Press, Cheshire, CN. Weiss, S.H. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, San Francisco, CA. Discovery and Data Mining, MIT Press, Cambridge, MA. Han, J. and Fu, Y. (1995), Discovery of Multiple-Level Association Rules form Large Databases, Paper presented at the 21st Int'l Conf. Very