"AN ANALYSIS OF TECHNIQUES FOR IMPROVING DATA QUALITY FOR KNOWLEDGE DISCOVERY IN DATABASES"

Sarika Sharma *, Rizwan Shaikh**, Dr. Vipin Tyagi***

ABSTRACT:
The companies in today’s fast changing competitive world are compiling databases on individual customers so that they can understand their customers better. The strategy today is to maintain, monitor and analyze customer records not just for adding values but also to have a competitive edge. These records are maintained in the form of a customer database also known as a data warehouse. Organizational database are pervaded with the poor quality of data as most of the times the data is not collected for the purpose of data mining in the mind, and the sources, which seems to be quite reliable can provide deflecting information. The data mining tools when applied on this effective data will result in poor quality model following the garbage in garbage out for the data. So business organizations are spending a great amount of their budget on improving the data quality and pre-processing of data for the model. The paper presents the various methods to improve the data quality and analyze the research done in the area.

INTRODUCTION:
Companies are maintaining the data warehouse, which is a repository of multiple heterogeneous data sources, and organized for applications like Knowledge Discovery in Databases and Data Mining for storing records of customers comprising of their detailed personal information and every transaction made by them. Data is stored in the large data warehouses and is further used for mining to identify the hidden patterns. This is achieved by the data mining process, which according to Prebon (1999) is considered to be a blend of statistics, Artificial Intelligence, and database research. Sometimes Data Mining and Knowledge Discovery are used as synonyms too [22].

Data mining refers to extracting or mining knowledge from large amount of data and applying intelligent methods in order to extract hidden data patterns. The pool of database so created helps to access information through these special techniques for sustainable customer satisfaction. Data Mining has become an essential tool to help business executives take strategic decisions and to sharpen the competitive edge of their business. It is a technique used for CRM, which helps companies to provide more customized service to their customers.

A basic task of the process is to extract knowledge from data such as resulting knowledge is useful in given application. Recently there has been a vast amount of interest in the discovery of different types of association rules from large databases [28]. Data Warehouses so maintained by the companies are architecture for the knowledge discovery so they play important role in process.

* Sarika Sharma is Assistant Professor at JS College of Engineering, Pune
** Rizwan Shaikh is HOD, MCA Department at JS College of Engineering, Pune
*** Dr. Vipin Tyagi is Assistant Professor at Jaypee Institute of Technology, Guna, M.P
As it is containing information related to real people and real world, database design should be done with common sense business reasoning and learning so that the meaningful information is stored in the database, which can be used to design various strategies [26].

**PROCESS OF KNOWLEDGE DISCOVERY IN DATABASES:**

The basic steps of the process are:

1. **Data Cleaning:** To remove noise and inconsistent data.
2. **Data Integration:** Multiple sources may be combined.
3. **Data Selection:** Data relevant to the analysis task are retrieved from the database.
4. **Data Transformation:** Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. **Data Mining:** Intelligent methods are applied in order to extract data patterns.
6. **Pattern Evaluation:** Identify truly interesting patterns representing knowledge based on some interestingness measure.
7. **Knowledge Presentation:** Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

The first three stages are pre-processing of data i.e. preparing the data for applying Data Mining techniques. Data preprocessing can be organized into following categories; data cleaning, data integration & transformation, and data reduction. Preprocessing plays important role in data mining, as data quality is crucial for many applications.

Data quality can be defined as “consistently meeting customer’s expectations. Improving data quality is a burning issue in many areas outside KDD [12] [6].

Ideally data mining research should focus on what happens traditionally both before and after applying the data mining technique [24]. Data Quality Management is implemented mainly in Data Warehouses. Data Warehouse technology includes data cleaning, data integration and analyzing techniques with functionalities such as summarization, consolidation and aggregation [5].

**SURVEY OF THE LITERATURE:**

Most mining techniques assume that the data is in fairly clean and complete form, which is generally not true for data to be entered in a data warehouse [25]. The information resources that are generally quite reliable may supply conflicting information at times [20] and most of the times data is not collected for the purpose of data mining in mind.

Thus it may contain incomplete relations; and the information stored in the databases is often subject to uncertainty and imprecision [15] [17] [16]. Most Data Mining algorithms assume the input data is static or have been designed for arbitrary inserts and deletions of data records [9], which is not true for the real life data gathered.
There is a clear need seen for improving the quality of data in the data warehouses. The most time consuming part of the data mining process is preparing data for data mining [10]. Thus there is a need to develop data cleaning techniques that can be used before mining on it is done, as quality of the data has strong impact on decision making and help management deliver quality data products [27]. Many authorities have noted that establishing a data warehouse and cleansing the data often account for most of the effort (and cost) of a Data Mining project [19].

The presence of missing or incomplete data is commonplace in large real world databases. Missing values occur for a variety of reasons, e.g. omissions in the data entry process, confusing questions in data gathering process etc. Learning missing data is critical for the data warehousing from the summary information. At the stage of capturing, organizing & retrieving information, the involvement of human being is essential so that the resulting knowledge from the data is useful in given applications [29] [1] [14]

Problems with data quality can result in tangible and intangible damage ranging from loss of customer to loss of vision. Data quality supports the effective decision-making and that data gets to the right person at right time. DQM emphasizes on the improvement of data quality to ensure that a) user of data are involved in the improvement b) pre-determined requirements for excellence are defined in terms of measurable data characteristics c) data conforms to the requirements [6]. Decision makers often confront the issue of whether to utilize information based on incomplete but consistent data or instead rely on complete or less consistent data [2].

There is need to study the link between poor data quality and procedures to detect and eliminate problems. In a typical data mining session, most of the time is spent on extracting and manipulating data, not really doing data mining and exploration. Stored data seems to double every 9 months and this makes demand for data mining and reduction tools increase exponentially. Only 10% of collected data is ever used and forty % of all collected data have errors in them. (Little progress is seen in data cleaning) [8]. Also it is estimated that 70% of the time used in DM project is dedicated to data preparation. Data preprocessing is a lengthy stage in mining, and so limiting the iterations of preparation is a particularly attractive goal [23].

Researchers are trying to develop efficient algorithms to cope with large volumes of data [16] and data preparation is required for speeding of Data Mining process. Another question is “how much data is enough?”. Even if we have infinite data available it may be case that we do not need all of it to obtain the best data mining results [7]. Given a large data set and a classification-learning algorithm, progressive Sampling (PS) uses increasingly larger random samples to learn until model accuracy no longer improves. It is shown that the technique is remarkably efficient compared to using the entire data. How to set a starting sample size is still a problem [11].

Today data management challenges are sparse data problem, reuse of data, identifying missing values and integration and correction of this missing information with observable
data [4]. Few detection and recovery techniques for data corruption have been suggested but this needs further research. Research is going on the issues of conceptual reconstruction for mining massively incomplete data sets and further refinements are open issues [3].

Technical challenges involve availability and reliability of data. There is need to digitize large quantities of information, including in some states, test results, difficulties cleaning up data from multiple sources so that they are compatible and non-redundant; and various data entry & accuracy errors that can lead to incorrect conclusions [18].

Data Quality Mining (DQM) can be defined as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases. Improving data quality can be seen as a goal of its own and DQM opens many new and promising application areas for data mining techniques outside of KDD [13]

**Basically there are three important aspects:**

(a) Employment of data mining methods to measure and explain data quality deficiencies: Research and practical experience are needed in order to understand what methods fit with which data.

(c) Although current process models are aware of data quality aspects, process models for KDD should have an explicit data quality phase. Measuring and improving data quality is a central aspect of the initial phases of a KDD project.

(d) Specialized process models for DQM need to be developed which reflect the change of scope from pure data analysis to data quality measurement and improvement.

Technical challenges also involve availability and reliability of data. There is need to digitize large quantities of information, including in some states, test results; difficulties cleaning up data from multiple sources so that they are compatible & non-redundant and various entry & accuracy errors that can lead to incorrect conclusions [21].

**METHODS TO IMPROVE DATA QUALITY FOR DATA MINING:**

The quality of information can be decided only by the organization, which is going to use this information. But there is need to have an agreed framework for the analysis of the quality of data. Some of the guidelines for the data quality are provided by the Department of Defense, which can be applied for most of the applications. Richard Wang (1999) has taken into consideration about fifty research papers written by various people and designed a framework for the data quality analysis. The elements of this framework can be used by the companies for knowing the status of quality of their data, which they are using for building the data-mining model. Various people have analyzed the impact of the data quality for CRM applications and they have provided the implications.
Most of the researchers felt that in a typical data mining session most of the time is spent on the extracting and manipulating the data (pre-processing), not really doing the data mining explorations. Because of the bad data, missing values, etc. the data mining process becomes a time consuming exercise.

Various methods are given for the data cleaning, preparing, improving quality of data by the various researchers working in this field. Some of the methods are summarized in the table 1 given below:

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Authors</th>
<th>Methods/ Solutions</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.</td>
<td>Wu and Barbara (2002)</td>
<td>Learning missing values from summary constraints</td>
<td>Linear Algebra and constraint programming as a basis for the summarized missing values.</td>
</tr>
<tr>
<td>8.</td>
<td>Parthasarathy and Aggarwal (2002)</td>
<td>On the use of conceptual reconstruction for mining massively incomplete data sets</td>
<td>Using correlation structure of data representations on which data mining algorithms can directly be applied.</td>
</tr>
</tbody>
</table>

Many researchers have tried to give the models for improving data quality in the databases on which data mining algorithms are applied to generate a model. It can be derived from the above analysis that a majority of the authors have tried to improve data quality by using mathematics and the programming solutions. So it can be considered as a separate research area not exactly or directly related to the data mining research.

CONCLUSIONS AND FUTURE SCOPE:

On the basis of the above, it can be concluded that further research in the area is solicited, as there are no standard methods for improving the data quality for knowledge discovery...
process. The different techniques of improving data quality at different levels of organizations for an overall improvement in business are required to agree upon some common metrics. Following can be the active areas of the further research:

1. To find technique(s) for Data Reduction and aggregation (rolling up of a customer transactional data into a single, summary non-transactional record).
2. To improve the quality of data in terms of:
   • Fill missing data
   • Remove redundant data
   • Combine related data
   • Remove bad data
3. To find optimal size of data sample which is sufficient to represent total data and analyze dependencies of data variables on each other.

REFERENCES:

1. Ankerst, M. (2001), Human Involvement and Interactivity of Next Generation Data Mining Tools, Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD.
7. Donimgos, P., Hulten, G. (2001), Catching Up with the Data: Research Issues in Mining Data Streams, Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD.


23. Skalak, D.B. (2001), Speeding-up Mining or “Why is data mining iterative?” Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD.


Biographical notes:

1. **Sarika Sharma** is Assistant Professor (IT) at the JS College of Engineering, Pune, India. She has done MCA from the Banasthali Vidyapith and is currently pursuing Ph.D at GGSIP University, Delhi. She has published 15 research papers for various national and international journals and presented papers at conferences of National & International repute. Her research interests are data mining, data warehousing, customer relationship management and software engineering.
   e-mail: sarika4@gmail.com
   Ph: 09970826076

2. **Rizwan Shaikh** is Assistant Professor and HOD (MCA Department) at JS College of Engineering, Hadapsar, Pune. He is ME(Computer), MBA and have rich 13 years experience of Industry and Academics. He has done several software projects and research interests are, Artificial Intelligence, Management, and Robotics.

3. **Dr. Vipin Tyagi** is Assistant Professor at Jaypee Institute of Technology, Guna, MP. He is a keen researcher and author of various research papers in international and national journals. His Research interests are Fuzzy Logic and robotics.
   e-mail: tyagivipindr@rediffmail.com