ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Study on Data Quality Mining

S. Brintha Rajakumari

Assistant Professor, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, India

ABSTRACT: An enormous amount of data stored in databases and data warehouses, it is increasingly important to develop powerful tools for analysis of such data and mining interesting knowledge from it. Data mining is a process of inferring knowledge from such huge data. Dirty data is a serious problem leading to incorrect decision making, inefficient daily operations, and eventually wasting both time and money. Data quality refers to the accuracy and completeness of the data. To perk up data quality, it is now and then necessary to dirt free the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points, removing unnecessary data fields, identifying anomalous data, and standardizing data formats. Data mining can be carried out on data represented in quantitative, textual, or multimedia forms. It can use a variety of parameters to examine the data which include association, sequence analysis, classification, clustering and forecasting. This paper provides an approach of the data quality in data mining and also discusses about the Data Quality as a versatile issue that represents one of the biggest challenges for data mining.

KEYWORDS: Data Quality, Data Mining.

I. INTRODUCTION

Data quality has serious consequences for the efficiency and effectiveness of organizations and businesses. Data quality has many dimensions like Accuracy, Completeness, Consistency, and Timeliness. Accuracy of data is the degree to which data correctly replicate the real world object. Completeness of data is the amount to which the expected attributes of data are provided. Consistency of data means that data across the venture should be in synchronized with each other or the absence of data conflicts. The timeliness of data is extremely important which depends on user probability.

The concept of data quality has emerged only during past ten years due to the exchange of data among the business organizations, government etc. In particular the concern on data quality has been increased due to the growth of internet. The review of data quality is done here apart from the context of DBMS like data integrity and data security.

Data Quality is classified into four categories, Intrinsic DQ, Accessibility DQ, Contextual DQ and Representational DQ. Each category has many dimensions like Accuracy, Completeness, Consistency, Timeliness, etc. from literature survey in Table1. The scope of the study in this paper includes only the intrinsic and representational data quality categories.

TABLE 1 DQ CATEGORIES AND DIMENSIONS

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Timeliness/Currency.
Accessibility DQ	Accessibility, Access Security
Contextual DQ	Relevancy,Value-added, Completeness
Representational DQ	Content coverage/Amount of data, Consistent Representation/Writing Style, Interactivity, Layout, Multimedia Presentation, Navigation Quality,Organization, Achieves/Documentation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

II. RELATED WORKS

Data Quality Mining can be defined as the premeditated application of data mining techniques for the purpose of data quality measurement and improvement [5]. The goal of Data Quality Mining is to detect, quantify, explain, and correct data quality deficiencies in very large databases.

Poor data quality is always a problem in practical applications of Knowledge Discovery in Database (KDD). Normally the data does not originate from systems that were set up with the main goal of mining this data. It is often has to contract with operative systems that produce data only as a byproduct. Even if during design and implementation of such systems data quality was considered an important aspect, experience shows that such principles get worn out in the long run. The reason is that data quality pays off only from a considered point of view. So after several years in business and a huge number of operative decisions to adapt the system to its changing environment, data quality typically suffers. Typically the owner of the data is not fully responsive of data quality deficiencies. The system might have been doing a good job for years and the owner most likely has its initial status in mind. Doubts concerning data quality may raise surprise or even disaffection. Data mining techniques are used to patch up solution to measure, explain, and improve data quality. In fact, this practical behavior usually was surprisingly successful. A systematized approach is expected in order to support analysts in this crucial situation[1].

There are many opening points to employ common data mining methods for the purposes of Data Quality Mining. Methods for deviation and outlier detection, clustering approaches and dependency analysis are straight forward to be employed for data quality purposes. It is possible that neural networks are trained to recognize data deficiencies and classify them with the supply of training data prepared by a human.

Data quality is a new research area which includes Data mining, Statistics, Knowledge representation, Management information systems and Data integration. Data mining is an analytic process designed to explore large sets of data in search of consistent patterns and/or systematic relationships between attributes or variables. Exploratory data mining is defined as the preliminary process of discovering structure in a set of data using statistical summaries, visualization, and other means.

In this context, achieving good data quality is an intrinsic objective of any data mining activity otherwise the process of discovering patterns, relationships and structures is seriously deteriorated. From another perspective, data mining techniques may be used in a wide spectrum of activities for improving the quality of data.

III. DATA QUALITY METHODOLOGY

The following five steps process helps to comprehensibly improve and maintain the data quality as in Figure1.

1.Outline: The first step for creating a successful data quality program is to understand what data quality means in the context of a particular organization. Quality data is broadly defined as fit for use. It can be trusted and it is suitable for its intended purpose.

2. Improve: Extensive set of rules help to improve data fit for the business, reduce costs, find more opportunities and develop operational efficiencies. From simple information to complicated information, cross-checking of information can quickly and reliably change the data.

3. Integrate: Data integration is the process of extracting and integration data from multiple heterogeneous sources to be loaded into an integrated information resource [2].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

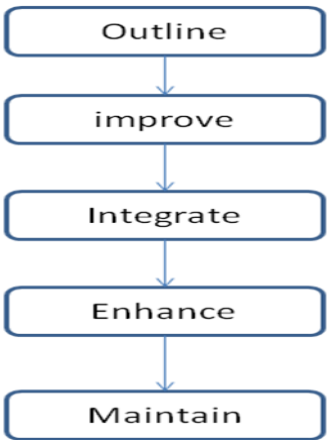

Figure 1. Steps in Data Quality Methodology

4. Enhance: The term data enhance refers to a set of procedures for adding detail to the records in an existing database. It is necessary to understand and analyze data in the database. Most companies fall short on capturing data such as demographic, lifestyle and firm graphics.
5. Maintain: Measuring data quality enables to maintain high quality data. Dashboard software allows the management to create reports and quickly identify problem areas. Measuring data is the only method for creating consistent data quality[3][4].

IV. DATA QUALITY MINING

Data Quality Mining can be defined as the premeditated application of data mining techniques for the purpose of data quality measurement and improvement [5]. The goal of Data Quality Mining is to detect, quantify, explain, and correct data quality deficiencies in very large databases.

Poor data quality is always a problem in practical applications of Knowledge Discovery in Database(KDD). Normally the data does not originate from systems that were set up with the main goal of mining this data. It is often has to contract with operative systems that produce data only as a byproduct. Even if during design and implementation of such systems data quality was considered an important aspect, experience shows that such principles get worn out in the long run[. The reason is that data quality pays off only from a considered point of view[6]-[9]. So after several years in business and a huge number of operative decisions to adapt the system to its changing environment, data quality typically suffers. Typically the owner of the data is not fully responsive of data quality deficiencies. The system might have been doing a good job for years and the owner most likely has its initial status in mind. Doubts concerning data quality may raise surprise or even disaffection. Data mining techniques are used to patch up solution to measure, explain, and improve data quality. In fact, this practical behavior usually was surprisingly successful. A systematized approach is expected in order to support analysts in this crucial situation[10][11].

There are many opening points to employ common data mining methods for the purposes of Data Quality Mining. Methods for deviation and outlier detection, clustering approaches and dependency analysis are straight forward to be employed for data quality purposes. It is possible that neural networks are trained to recognize data deficiencies and classify them with the supply of training data prepared by a human[12][13].

Data quality is a new research area which includes Data mining, Statistics, Knowledge representation, Management information systems and Data integration. Data mining is an analytic process designed to explore large sets of data in search of consistent patterns and/or systematic relationships between attributes or variables. Exploratory data mining is defined as the preliminary process of discovering structure in a set of data using statistical summaries, visualization, and other means.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

In this context, achieving good data quality is an intrinsic objective of any data mining activity otherwise the process of discovering patterns, relationships and structures is seriously deteriorated. From another perspective, data mining techniques may be used in a wide spectrum of activities for improving the quality of data[14-19].

V. CONCLUSION

Data quality is a multidisciplinary area, because data in a variety of formats and with a variety of media are used in every real life, business activity and manipulate the quality of processes that use data. With the advent of networks and the internet data are created and exchanged with much more confused processes and need more sophisticated management. The data quality methodology was defined as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. Improving data quality can be seen as a goal of its own and data quality methodology opens many new and promising application areas for data mining techniques. This paper presented new data quality methodology for data quality mining.

REFERENCES

1. R. Agrawal, R. Srikant, ``Mining Sequential Patterns", Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.
2. Udayakumar R., Khanaa V., Kaliyamurthie K.P., "High data rate for coherent optical wired communication using DSP", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) 4772-4776.
3. Angeles, P. and MacKinnon, L., 2004. Detection and Resolution of Data Inconsistencies, and Data Integration using Data Quality Criteria. QUATIC'2004.
4. Hariharan V.S., Nandlal B., Srilatha K.T., "Efficacy of various root canal irrigants on removal of smear layer in the primary root canals after hand instrumentation: A scanning electron microscopy study", Journal of Indian Society of Pedodontics and Preventive Dentistry, ISSN : 0970-4388, 28(4) (2010) pp.271-277.
5. Data Quality Concepts, Methodologies and Technique, Batini C,Scannapieco M.
6. Udayakumar R., Khanaa V., Kaliyamurthie K.P., "Optical ring architecture performance evaluation using ordinary receiver", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp. 4742-4747.
7. M. Angélica Caro,Data Quality in web applications:A state of the art, IADIS International Conference on WWW/Internet 2005.
8. Kanniga E., Selvaramarathnam K., Sundararajan M., "Embedded control using mems sensor with voice command and CCTV camera", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp.4794-4796.
9. jochen hipp, udo grimmer, Data Quality Mining.
10. Udayakumar R., Khanaa V., Kaliyamurthie K.P., "Performance analysis of resilient ftth architecture with protection mechanism", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp. 4737-4741
11. M. Angelica Caro,Coral Calero, Ismael Caballero, Mario Piattini., Data Quality In Web Applications: A State Of The Art ,IADIS International Conference on WWW/Internet 2005, pp 364-368.
12. S.Brintha Rajakumari and C.Nalini ,"An efficient Data Mining data Set preparation using aggregation in relational database" , Indian Journal of Science and Technology,Vol 7(S5),Pp 44-46, June 2014.
13. S.Brintha Rajakumari and C.Nalini ,"An efficient cost Model for data storage with horizontal layout in the cloud" , Indian Journal of Science and Technology,Vol 7(S3),Pp 45-46, March 2014.
14. S.Christy, S.Brintha Rajakumari, Dr.M.Suryakala "Quality Data Representation in Web Portal-A Case Study" in Second International Conference on Trendz in Information Science and Computing (TISC–2011) held at Sathyabama University, Chennai on December 17th -19th, 2010.
15. B.Vamsi Krishna, Significance of TSC on Reactive power Compensation, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN (Online): 2278 – 8875,pp 7067-7078, Vol. 3, Issue 2, Febuary 2014
16. B.Vamsi Krishna, Realization of AC-AC Converter Using Matrix Converter, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN (Online): 2278 – 8875,pp 6505-6512, Vol. 3, Issue 1, January 2014
17. D.Sridhar raja, Comparison of UWB Band pass filter and EBG embedded UWB Band pass filter, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN 2278 – 8875,pp 253-257 ,Vol. 1, Issue 4, October 2012
18. D.Sridhar raja, Performances of Asymmetric Electromagnetic Band Gap Structure in UWB Band pass notch filter, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN (Online): 2278 – 8875,pp 5492-5496, Vol. 2, Issue 11, November 2013
19. Dr.S.Senthil kumar, Geothermal Power Plant Design using PLC and SCADA, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN 2278 – 8875,pp 30-34, Vol. 1, Issue 1, July 2012