A Review of Data Cleaning Algorithms for Data Warehouse Systems

Rajashree Y.Patil,[#], Dr. R.V.Kulkarni [*]

[#] *Vivekanand College, Kolhapur, Maharashtra, India*

[*] *Shahu Institute of Business and Research (SIBER) Kolhapur, Maharashtra, India*

***Abstract*— In today's competitive environment, there is a need for more precise information for a better decision making. Yet the inconsistency in the data submitted makes it difficult to aggregate data and analyze results which may delays or data compromises in the reporting of results. The purpose of this article is to study the different algorithms available to clean the data to meet the growing demand of industry and the need for more standardised data. The data cleaning algorithms can increase the quality of data while at the same time reduce the overall efforts of data collection.**

***Keywords*— ETL, FD, SNM-IN, SNM-OUT, ERACER**

I. INTRODUCTION

Data cleaning is a method of adjusting or eradicating information in a database that is wrong, unfinished, inappropriately formatted or reproduced. A business in a data-intensive profession like banking, insurance, trade, telecommunication, or transportation might use a data cleaning algorithm to methodically inspect data for errors by using a set of laws, algorithms and search for tables. On average data cleaning tool consist of programs that are able to correct a number of specific types of errors or detecting duplicate records. Making use of algorithm can save a database manager a substantial amount of time and can be less expensive than mending errors manually.

Data cleaning is a vital undertaking for data warehouse experts, database managers and developers alike. Deduplication, substantiation and house holding methods can be applied whether you are populating data warehouse components, incorporating recent data into an existing operation system or sustaining real time dedupe efforts within an operational system. The objective is an elevated level of data precision and reliability that transmutes into enhanced customer service, lower expenses and tranquillity. Data is a priceless organisational asset that should be developed and honed to grasp its full benefit. Deduplication guarantees that a single correct record be present for each business unit represented in a business transactional or analytic database. Validation guarantees that every characteristic preserved for a specific record is accurate. Cleaning data prior to it being stored in a reporting database is essential to provide worth to clients of business acumen applications. The cleaning procedure usually consists of processes that put a stop to duplicate records from being reported by the system. Data analysis and data enrichment services can help improve the quality of data. These services include the aggregation, organisation and cleaning of data. These data cleaning and enrichment services can ensure that your database-part and material files, product catalogue files and item information etc. are current, accurate and complete.

Data Cleaning Algorithm is an algorithm, a process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and commissions. The process may include format checks, completeness checks, reasonable checks, and limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g. taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking and correction of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules and conventions.

Data cleaning Approaches - In general data cleaning involves several phases.

a. Data Analysis: In order to detect which kinds of errors and inconsistencies are to be removed, a detailed data analysis is required. In addition to a manual inspection of the data or data samples, analysis programs should be used to gain metadata about the data properties and detect data quality problems.

b. Definition of transformation workflow and mapping rules: Depending on the number of data sources , their degree of heterogeneity and the "dirtiness " of the data, a large number of data transformation and cleaning steps may have to be executed.

c. Verification: The correctness and effectiveness of a transformation workflow and the transformation definitions should be tested and evaluate .e.g. on a sample or copy of the source data, to improve the definitions if necessary.

d. Transformation: Execution of the transformation steps either by running the ETL workflow for loading and refreshing a data warehouse or during answering queries on multiple sources.

e. Backflow of cleaned data: After errors are removed, the cleaned data should also replace the dirty data in the original sources in order to give legacy applications the improved data too and to avoid redoing from the data staging area.

II. REVIEW OF LITERATURE

While a huge body of research deals with schema translation and schema integration, data cleaning has received only little attention in the research community. A number of authors focussed on the problem of data cleaning and they suggest different algorithms to clean dirty data.

Wejje Wei, Mingwei Zhang, Bin Zhang Xiaochun Tang

In this paper, a data cleaning method based on the association rules is proposed. The new method adjusts the basic business rules provided by the experts with association rules mined from multi-data sources and

generates the advanced business rules for every data source. Using this method, time is saved and the accuracy of the data cleaning is improved.

Applied Brain and Vision Science-Data cleaning algorithm

This algorithm is designed for cleaning the EEG resulting from the brain function of "stationary" behaviour such as for an eyes-open or eyes-closed data collection paradigm. This algorithm assumes that the useful information contained in the EEG data are stationary. That is, it assumes that there is very little change in the on-going statistic of the signals of inertest contained in the EEG data. Hence, this algorithm is optimal for removing momentary artifacts in EEG collected while a participant is in eyes-closed or eyes-opened state.

Yiqun Liu,Min Zhang,Liyun Ru,Shaoping Ma

They propose a learning-based algorithm for reducing Web pages which are not likely to be useful for user request. The results obtained show that retrieval target pages can be separated from low quality pages using query-independent features and cleansing algorithms. Their algorithm succeeds in reducing 95% web pages with less than 8% loss in retrieval target pages.

Helena Galhardas, Daniela Florescu, Dennis Shasha

This paper presents a language, an execution model and algorithms that enable users to express data cleaning specifications declaratively and perform the cleaning efficiently. They use as an example a set of bibliographic references used to construct the Citesser Web Site. They propose a model to clean textual records so that meaningful queries can be performed.

Timothy Ohanekwu,C.I.Ezeife

This paper proposes a technique that eliminates the need to rely on match threshold by defining smart tokens that are used for identifying duplicates. This approach also eliminates the need to use the entire long string records with multiple passes, for duplicate identification.

Yu Qian, Kang Zhang

This article focuses on a role of visualization in effective data cleaning. This addresses a challenging issue in the use of visualization for data mining: choosing appropriate parameters for spatial data cleaning methods. On one hand, algorithm performance is improved through visualization. On the other hand, characteristics and properties of methods and features of data are visualised as feedback to the user.

Chris Mayfield, Jennifer Nevialle, Sunil Prabhakar

In "A statistical method for integrated data cleaning and imputation" this paper they focus on exploiting the statistical dependencies among tuples in relational domains such as sensor networks, supply chain systems and fraud detection. They identify potential statistical dependencies among the data values of related tuples and develop algorithms to automatically estimate these dependencies, utilizing them to jointly fill in missing values at the same time as identifying and correcting errors.

Data cleaning based on mathematical morphology-S.Tang

In the field of bioinformatics and medical image understanding, data noise frequently occurs and deteriorates the classification performance; therefore on effective data cleansing mechanism in the training data is often regarded as one of the major steps in the real world inductive learning applications. In this paper the related work on dealing with data noise is firstly revived , and then based on the principle of mathematic morphology , the morphological data cleansing algorithms are proposed and two concrete morphological algorithms.

Kazi Shah Nawaz Ripon, Ashiqur Rahman,G.M. Atiqur Rahaman

In this paper, they propose novel domain independent techniques for better reconciling the similar duplicate records. They also introduced new ideas for making similar-duplicate detection algorithms faster and more efficient. In addition, a significant modification of the transitive rule is also proposed.

Payal Pahwa, Rajiv Arora, Garima Thakur

This paper addresses issues related to detection and correction of duplicate records. Also, it analyses data quality and various factors that degrade it. A brief analysis of existing techniques is discussed, pointing out its major limitations. And a new framework is proposed that is an improvement over the existing technique.

R.Kavitakumar , Dr.RM.Chandrasekaran

In this paper they designed two algorithms using data mining techniques to correct the attribute without external reference. One is Context-dependent attribute correction and another is Context-independent attribute correction.

Jie Gu(

In this paper , they propose a combination of random forest based techniques and sampling methods to identify the potential buyers. This method is mainly composed of two phases : data cleaning and classification, both based on random forest.

Subhi Anand , Rinkle Rani

World Wide Web is a monolithic repository of web pages that provides the internet users with heaps on information. With the growth in number and complexity of websites, the size of web has become massively large. This paper emphasizes on the Web Usage Mining process and makes an exploration in the field of data cleaning.

Li Zhao, Sung Sam Yuan, Sun Peng and Ling Tok Wang

They propose a new comparison method LCSS, based on the longest common subsequence, and show that it possesses the desired properties. They also propose two new detection methods, SNM-IN and SNM-INOUT, which are variances of the popular detection method SNM.

Aye T.T.

This paper mainly focus on data pre-processing stage of the first phase of Web usage mining with activities like field extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from the single line of the log file. This data cleaning algorithm eliminates inconsistent or unnecessary items in the analysed data.

Yan Cai-rong, Sun Gui-ning, Gao Nian-gao

By analysing the limitation of traditional structures of knowledge base, an extended tree-like knowledge base is built by decomposing and recomposing the domain knowledge. Based on the knowledge base ,a data cleaning algorithm is proposed. It extracts atomic knowledge of the selected nodes firstly, then analyses their relations, deletes

the same objects, builds an atomic knowledge sequence based on weights.

Chris Mayfield, Jennifer Neville, Sunil Prabhakar

They present ERACER, an iterative statistical framework for inferring missing information and correcting such errors automatically. Their approach is based on on belief propagation and relational dependency networks, and includes an efficient approximate inference algorithm that is easily implemented in standard DBMSs using SQL and user defining functions.

Mohammad, H.H.

They developed a system uses the extract, transform and load model as the system main process model to serve as a guideline for the implementation of the system. Besides that, parsing technique is also used for identification of dirty data. Here they selected K-nearest Neighbour algorithm for the data cleaning.

Shawn R. Jeffery, Minos Garofalakis, Michel J. Franklin

In this paper they propose SMURF , the first declarative , adaptive smoothing filter for RFID data cleaning. SMURF models the unreliability of RFID readings by viewing RFID streams as a statistical sample of tags in the physical world , and exploits techniques grounded in sampling theory to drive its cleaning processes.

Manuel Castejon Limas, Joaquin B. Ordiieres Mere , Francisco J.

A new method of outlier detection and data cleaning for both normal and no normal multivariate data sets is proposed. It is based on an iterated local fit without a priori metric assumptions. They propose a new approach supported by finite mixture clustering which provides good results with large data sets.

Kollayut Kaewbuadee,Yaowadee Temtanapat

In this research they developed a cleaning engine by combining an FD discovery technique with data cleaning technique and use the feature in query optimization called "Selective Value" to decrease the number of discovered FDs.

Data cleaning is a very young field of research. This represents the current research and practices in data cleaning. One missing aspect in the research is the definition of a solid theoretical foundation that would support many of the existing approaches used in industrial setting. The number of data cleaning methods is used to identify a particular type of error in data. Unfortunately, little basic research within the information systems and computer science communities has been conducted that directly relates to error detection and data cleaning. In-depth comparisons of data cleaning techniques and methods have not been yet published. Typically, much of the real data cleaning work is done in a customized, in-house, manner. This behind-the scene process often results in the use of undocumented and ad hoc methods. Data cleaning is still viewed by many as a "black art" being done "in the basement". Some concerned effort by the database and information systems groups is needed to address this problem. Future research directions include investigation and integration of various methods to address error detection. Combination of knowledge-based techniques with more general approaches should be pursued. The ultimate goal of data cleaning research is to devise s set of general operators and theory that can be combined in well-formed statements to address data cleaning problems. This formal basis is necessary to design and construct high quality and useful software tools to support the data cleaning process.

REFERENCES

[1] "A Data Cleaning Method Based on Association Rules" by Weijie Wei, Mingwei Zhang, Bin Zhang ,www.atlantis-press.com

[2] Applied Brain and Vision Science-Data cleaning algorithm

[3] "Data Cleansing for Web Information Retrieval using Query Indepandent Features" by Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma- www.thuir.cn

[4] "An Extensive Framework for Data Cleaning " by Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon

[5] "A Token-Based Data Cleaning Technique for Data Warehouse" by Timothy E. Ohanekwu International Journal of Data Wrehousing and Mining Volume 1

[6] "The role of visualisation in effective data cleaning" by Yu Qian, Kang Zhang – Proceddins of 2005 ACM symposium on applied computing

[7] "A Statistical Method for Integrating Data Cleaning and Imputation" by Chris Mayfield, Jennifer Neville, Sunil Prabahakar-Purdue University(Computer Science report-2009)

[8] "Data cleansing based on mathematical morphology" by Sheng Tang published in ICBBE 2008 The second International Conference-2008

[9] "A Domain Independent Data Cleaning Algorithm for detecting similar-duplicates" by Kazi Shah Nawaz Ripon, Ashquir Rahman and G.M. Atiqur Rahaman – Journal of Computer Vol 5, No. 12,2010

[10] P.Pehwa "An Efficient Algorithm for Data Cleaning" www.igi-global.com -2011.

[11] "Attribute Correction-Data cleaning using Association Rule and Clustering Methods" by R.KavithaKumar, Dr. RM. Chandrasekaran, IJDKP,Vol.1,No.2 March-2011.

[12] Random Forest Based Imbalanced Data Cleaning and Classification – Jie Gu –Lamda.nju.edu.cn

[13] Data Cleansing Based on Mathematical Morphology S.Tang-2008 – ieeeexplore.ieee.org. Bioinfornatics and Biomedical Engineering , 2008 ICBBE 2008. The 2nd International conference.

[14] "An efficient Algorithm for Data Cleaning of Log File using File Extension" International journal of Computer Appliactions 48(8):13-18, June-2012 Surabhi Anand , Rinkle Rani Aggarwal.

[15] A New Efficient Data Cleasing Method – Li Zhao, Sung Sam Yuan, Sun Peng and Ling Tok Wang – ftp10.us.freebsd.org

[16] Computer Research and DEvleopment (ICCRD),2011,3rd International Conference ."web log cleaning for mining of web usage patterns" –T.T.Aye.

[17] "Mass Data Cleaning Algorithm based on extended tree-like knowledge base" – Yan Cai-rong,SUN Gui-ning , GAO Nian-gao Computer Enginerring and application -2010

[18] ERACER-A database approach for statistical inference and data cleaning- Chris Mayfield, Jennifer Neville, Sunil Prabhakar w3.cs.jmu.edu

[19] "Adaptive cleaning for RFID Data Streams" by Shawn R. Jeffery, Minos Garofalakis, Michael J. Franklin blg.itu.dk

[20] "Outlier Detection and Data Cleaning in Multivariate Non-Normal Samples: The PAELLA Algorithm" by Manuel Castejon Limas, Joaquin B. Ordieres Mere, Francisco J. Martinezn de Pison , Ascacibar and Eliseo P. Vergara Gonzaalez

[21] Informatics and Computational Intelligence (ICI) 2011 , Mohamed H.H . IEEE Xplore Digital Library. " E-Clean : A Data Cleaning Framework for Patient Data"