ALGORITHMS FOR DATA CLEANING IN KNOWLEDGE BASES

ADEEL ASHRAF¹, SARAH ILYAS¹, KHAWAJA UBAID UR REHMAN¹ AND SHAKEEL AHMAD²

¹Department of Computer Science, University of Management and Technology, Punjab, Pakistan

²Department of Computer Sciences, Faculty of Computing and Information Technology in rabigh, King Abdulaziz University Jeddah

Email: adeel.ashraf@umt.edu.pk

ABSTRACT. *Data cleaning is an action which includes a process of correcting and identifying the inconsistencies and errors in data warehouse. Different terms are used in these papers like data cleaning also called data scrubbing. Using data scrubbing to get high quality data and this is one the data ETL (extraction transformation and loading tools). Now a day there is a need of authentic information for better decision-making. So we conduct a review paper in which six papers are reviewed related to data cleaning. Relating papers discussed different algorithms, methods, problems, their solutions and approaches etc. Each paper has their own methods to solve a problem in efficient way, but all the paper have common problem of data cleaning and inconsistencies. In these papers data inconsistencies, identification of the errors, conflicting, duplicates records etc problems are discussed in detail and also provided the solutions. These algorithms increase the quality of data. At ETL process stage, there are almost thirty five different sources and causes of poor quality constraints.*

Keywords: Data inconsistencies, Extract-Transform-Load, Data Scrubbing, Data Quality.

1. Introduction. Information that is unfinished, wrong, reproduced or inappropriately formatted are eradicated or set through various methods in a database which is called data cleaning. Data cleaning algorithms are used to observe errors by using algorithms, examine for tables and set of rules. Duplicate records and errors are detected by data cleaning tool which contains set of programs. Use of algorithms can reduce cost and time rather than find out errors manually.

Data cleaning is used by database managers, developers and data warehouse specialists [1]. The objective of data cleaning is to reduced cost, searching time and enhances the customer services. Developing of data and clutch its full benefits is an organizational assets. In a business transactional database representation of business unit in which one record can be presented, this is guaranteed by deduplication. Accurate record properties preserver for a specific are guaranteed by validation. Cleaning data which is stored in a reporting database can be used by application users and user can take benefits using this data. The cleaning is used when duplication of data is reported by system. Service enhancement and quality of data is getting through an analysis on data. Cleaning of data, organization and aggregation are included by enrichment services. This service also guarantees the accurate, correct and complete data in a product files and catalogue files etc. [2].

An algorithm for data cleaning is a process to identify incomplete, inaccurate and unreasonable data and then improve the quality through correctness of identified errors. Algorithm process may use various checks like format, correctness, limit and reasonable. Resulting of these processes used another's checks like flag, document and subsequent. Validation checks also involve checking for agreement and may vary from contracts and standards [3].

2. Methodology.

Data Cleaning Approaches

There are five approaches which are discussed in this paper (Data analysis, Definition of transformation workflow and mapping rules, Verification, Transformation and Backflow of cleaned data).

1. Data analysis

A detailed analysis is required when inconsistencies and errors are identified to be destroyed.

2. Definition of transformation workflow and mapping rules

On the basis of different data sources, dirtiness and heterogeneity of data, then there is needed to be executed different steps of cleaning and transformation.

3. Verification

The effectiveness of definitions and workflow of transformation should be evaluated and tested.

4. Transformation

On the basis of different multiple data sources, workflow of ETL executed transformation steps for different operation like refreshing and loading a data warehouse and queries answering.

5. Backflow of cleaned data

Dirty data also replaced by cleaned data when errors are removed in the original sources and on the data staging area, avoid the redoing.

Figure 1:- Data Warehouse

In Figure-1 simply shows that all data is extracted and cleaning and transform. The steps shows in the above figure.

3. Related Work. In 2003 Kim (Kim et.al. 2003) planned a classification of dirty data. Defining and understanding is the first step of data cleaning it is a very important issue that will catch the attention of the researchers [17]. Data cleaning is applied when different databases are combined then duplicate records are identified and the problem is to identify and remove these duplications this problem is called merge/purge problem (Hernandez and Stolfo, 1998) [18]. With the passage of time different researchers referred this problem as record linkage, semantic integration, instance identification or the object identify problem. Many researchers give different solutions for this problem. Knowledge bases (Lee et.al. 2000) [19] regular expression matches and user defined constraints proposed by (Cadot and di Martion, 2003), and filtering (Sung et.al. 2002) [20]

In this paper we reviewed the data cleaning algorithms and their comparison and suggest which algorithm is best in different scenarios to the other data cleaning algorithms.

4. Results and Discussion on Problems, Methods, Solutions and Literature Review Comprehensive Data Cleaning.

The following papers discussed: methods, problems and challenges in data cleaning, current approaches etc. In today's focused surroundings, there is a requirement for more exact data for a superior basic leadership. The reason for this article is to contemplate the diverse calculations accessible to clean the information to take care of the developing demand of industry and the requirement for more institutionalized information [4]. The information cleaning calculations can expand the nature of information while in the meantime decrease the general endeavors of information accumulation. Information cleaning is a technique for modifying or killing

data in a database that isn't right, unfinished, improperly designed or recreated. An efficient managing an account, protection, exchange, telecom, or transportation may utilize an information cleaning calculation to systematically review information for mistakes by utilizing an arrangement of laws and calculations. Making utilization of calculation can spare a significant measure of time and can be less costly than retouching mistakes physically. The target to upgrade client administration and lower costs. Information is a precious authoritative resource and acceptance ensures that each trademark protected for a particular record is exact. In cleaning information examination and information advancement administrations can enhance the nature of information. These administrations incorporate the total, association and cleaning of information. These information cleaning and advancement administrations can guarantee that your database-part and material records, item list documents and thing data and so on are present, exact and finish. Information cleaning calculation procedure may incorporate organization checks, fulfillment checks, sensible checks, and breaking point checks, audit of the information to recognize anomalies or different mistakes, and appraisal of information by branch of knowledge specialists. Information cleaning approaches incorporate Data investigation, Definition of change work process and mapping rules, Verification, Transformation and Backflow of cleaned data [5]. Different creators proposed diverse calculations and methodologies like an information cleaning strategy in light of the affiliation tenets is proposed, a learning-based calculation to separate target pages from low quality pages, an execution model and calculations, a system that kills the need to depend on match edge, A measurable technique for coordinated information cleaning and ascription, novel space autonomous strategies for better accommodating the comparable copy records, two calculations utilizing information mining procedures to amend the characteristic without outer reference: One is Context-subordinate trait revision and another is Context-free property adjustment, an arbitrary backwoods based systems and examining strategies to distinguish the potential purchasers, Web Usage Mining process and makes an investigation in the field of information cleaning for web multifaceted nature, concentrate on information pre-preparing phase of the main period of Web use mining with exercises like field extraction and information cleaning calculations, ERACER an iterative factual structure for gathering missing data and rectifying such mistakes consequently, K-closest Neighbors calculation for the information cleaning and another strategy for anomaly discovery and information cleaning for both typical and no ordinary multivariate information sets is proposed [6].

Now a day, tradition of several organizations for the purpose of meeting the requirements, they used data warehouses to promote decision making and accomplish their goals through user/customer satisfaction. For any work related decision, it enables the top level management timely manner decision to access information. DSS (decision support system) is one of the sources that implement in data mining. Its run and dependency of the better decision on a factor called DQ (data quality). Using data scrubbing to get high quality data and this is one the data ETL (extraction transformation and loading tools). One of the most important and needy in DW (data warehouse) is DS (data scrubbing). There are increasing relation between data quality and data scrubbing to get effective DS and high DQ. DS provides the solution of those constraints in which DQ is weak and burden of financial cost is high. These constraints are: missing values, dirty and noise data, ambiguous, inconsistency, conflicting, similar column, uncertain data and duplicated records. The causes and sources have several constraints, include: merge data which have different sources, representing same information difference, input error etc. At ETL process stage, there are almost thirty five different sources and causes of poor quality constraints [9].

This paper also includes the comparison and analysis study of two tables of data scrubbing algorithms, which methods they are adopting and pros and cons of each algorithm, time complexity and accuracy in details. It also included comparison and analysis of data scrubbing framework and also find out which framework is best.

In paper the issues to be addressed is the data cleaning and solution. Sort the data quality problems that are processed by the data cleaning and main solution approaches. Data cleaning is particularly necessary when the integration of heterogeneous data sources and should be addressed as well as the diagram related data transformations. Purging data or scrubbing, deals with detecting and removing errors and data inconsistencies

to improve data quality. Data quality issues are present in the unique data collections, such as files and databases. This is because the sources often consist of redundant data in several representations. Data warehouses require and provide extended support for data cleaning. They load and continuously update massive amounts of data from various sources so the probability that some of the sources contain high "incorrect data". Because of the wide range of data inconsistencies can and volume simple data, data cleansing is considered one of the biggest problems in data warehouses. Federated database systems and information systems on the steps in front of web processing data similar to data warehouses. A data cleaning approach must meet several requirements. First, it should detect and remove all Main errors and inconsistencies in individual data sources and in integrating multiple sources [10].

In addition, data cleaning should not be performed in isolation, but with the diagram related data transformations from complete metadata. Mapping data cleaning functions and other data transformations should be specified declaratively and reusable for other data sources as well as regards the processing of applications [11]. Major data quality problems are solved by data cleaning and data transformation. As we shall see, these issues are closely linked and should be treated consistently. Data transformations are needed to support any changes in the structure, representation or content of data the quality of the data from a source depends largely on the degree to which it is governed by pattern and integrity constraints controlling the authorized data values. For sources without schema, such as files, there are few restrictions on the data that can be entered and stored, resulting in a high probability of errors and inconsistencies [12, 13].

Since the data source cleaning is expensive, preventing incorrect data to be displayed is obviously an important step to reduce the cleaning problem. This requires a proper design of database schema and integrity constraints both from data entry applications [16].

Sources multiple problems the problems present in single sources are compounded when multiple sources must be integrated. Each source can contain incorrect data and data sources can be represented in different ways, overlap or contradict. This is because the sources are usually developed, deployed and maintained independently to serve specific needs. At the schema level, model design differences and the data schema must be addressed by the scheme of translation steps and schema integration, respectively. Scheme level conflict included many conflicts appear only at the instance (data conflicts). Reflected in the metadata schemes are generally insufficient to assess the quality of data from one source, especially if only a few integrity constraints are applied. It is therefore important to analyze the actual instances to get the real (revised) metadata about the characteristics of data value or unusual patterns [14].

Data warehouse of an enterprise joins the information from various wellsprings of the association/undertaking to bolster endeavor wide basic leadership, reporting, investigating and arranging. A distribution center is a subject situated, coordinated, time-variation and nonvolatile gathering of information in backing of administration's basic leadership process. Data cleaning is a crucial errand in information warehousing process keeping in mind the end goal to get right and subjective information into the Data Warehouse [15]. The procedures on information stockroom delicate and rely on upon the exactness and consistency of information. Information in information distribution center from outside sources ordinarily contains blunders like spelling errors, conflicting traditions crosswise over information sources, and/or missing fields. Enormous measures of time and cash are spent on information cleaning, recognizing and amending mistakes in information. Late writing proposes a few cutting edge answers for coordinating and combining information sources. This article means to build familiarity with the effects of poor information quality on an undertaking. These effects incorporate client disappointment, expanded operational expense and less powerful basic leadership.

Information cleaning decides and recognizes the undesirable, degenerate, conflicting and defective information to upgrade the nature of information. Information cleaning procedure may incorporate configuration checks, culmination checks, sensibility checks, limit checks, survey of the information to distinguish exceptions different blunders. Wellsprings of mistake in information are information section blunders, estimation blunders, refining blunders and information coordination blunders. Information blunders can worm in at each progression of the procedure from starting information securing to record stockpiling. Information extraction is a standout amongst the most tedious undertakings of Data Warehouse improvement. The proposed system fundamentally engaged for describing information quality. Information cleaning procedure is partitioned into four sections: information access interface, information quality examination, and information change and results evaluation. The information quality investigation ought to be done just once and ought not to be a redundancy work. The procedure ought to be consecutive and not iterative. Information quality is the extent to which information meet the particular needs of particular clients. In quality information

fulfillment, legitimacy, exactness, accuracy, non-duplication, determination honesty, openness and opportuneness. Information quality issues incorporate Issues connected with information sees, qualities, presentation and security and privacy. Imperfect information leads to breakdowns in the inventory network, poor business choices, and second rate client relationship administration. Error rate = number of failed fields/number of aggregate fields. Poor information quality affects the common endeavor from various perspectives. At the operational level, poor information drives straightforwardly to client disappointment, expanded cost, and brought down representative employment fulfillment. It likewise increments operational expense since time and different assets are spent distinguishing and amending blunders. With the wrong information the associations resolve can be lost.

In this paper comprehensive data cleaning challenges, problems and methods are discussed Many challenges and problems was identified i-e researchers had not ability to implement the detail of cleaning in complete range of data cleaning [15]. Here is one limitation that in the framework, maintenance is not assumed or considered for data cleaning. Two problems Single source and multiple source problems are discussed in this paper and current approaches. Existing algorithm for identification of duplicate records through introducing a framework which is called integrated framework. With very least iterations, measurement of the developed algorithms beside the standard algorithm to see the effect. This limitation had also inability to determine the duplication of the record.

Open user involvements are better supported by ETL (extract-transform-load) and data cleaning. System user interfaces are graphical and end-user can easily to operate. Limitation for open user involvement is that selection of a data warehouse may be small while large data warehouse may provide exact results. Three tier data warehouse for Ghana petroleum industry to manage information through effective and comprehensive framework. Limitation for the study is narrowed because this included the restricted area of Ghana not geographically studies.

Conclusion. Data cleaning has become very important in organizations, colleges, schools and everywhere in the world. Six papers are reviewed and all the paper are adopt different approaches to solve the data inconsistencies and errors. These algorithms guarantees data cleaning, so theses algorithms increase the quality of data. In these papers data inconsistencies, identification of the errors, conflicting, duplicates records etc problems are discussed in detail and also provided the solutions. At ETL process stage, there are almost thirty five different sources and causes of poor quality constraints. These reviewed research also help and develop the direction in the future.

REFERENCES

- [1] Rahm, E., Do, H.H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bull. Vol 23 No. 4, pp. 3-13
- [2] Informatics and Computational Intelligence (ICI) 2011, Mohamed H.H . IEEE Xplore Digital Library. "E-Clean : A Data Cleaning Framework for Patient Data"
- [3] Monge, A. E. (2000). Matching Algorithms within a Duplicate Detection System. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, pp. 18-19.
- [4] Louardi BRADJI, Mahmoud BOUFAIDA. (2011). Open User Involvement in Data Cleaning for Data Warehouse Quality. International Journal of Digital Information and Wireless Communications (IJDWC) 1(2), pp. 573.
- [5] Deku JerryYao,Mohammad Sarrah and Hamza Aldabbas (2012).Three Tier level Data Warehouse Architecture for Ghanaian Petroleum Industry. International Journal of Database Management Systems (IJDMS) Vol.4, No.5, pp 1
- [6] Vassiliads, P.(2009). A Survey of Extract-Transform-Load Technology. In InternationalJournal of Data Warehousing & Mining,vol.5 ,no. 3, pp. 1-27
- [7] Monge, A. E. (2000). Matching algorithms within a duplicate detection system. IEEE Data Eng. Bull., 23(4), 14-20.
- [8] Salim, N., & Ibrahim, R. (2011, December). Towards data quality into the data warehouse development. In Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on (pp. 1199-1206). IEEE.

- [9] Paul, A., Ganesan, V., Challa, J. S., & Sharma, Y. (2012, March). HADCLEAN: A hybrid approach to data cleaning in data warehouses. In *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on* (pp. 136-142). IEEE.
- [10] Yan, H., Diao, X. C., & Li, K. Q. (2008, November). Research on information quality driven data cleaning framework. In *Future Information Technology and Management Engineering, 2008. FITME'08. International Seminar on* (pp. 537-539). IEEE.
- [11] Ahmed, I., & Aziz, A. (2010). Dynamic approach for data scrubbing process. *International Journal on Computer Science and Engineering*, 2(02), 416-423.
- [12] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4), 3-13.
- [13] Housien, H. I., Zuping, Z., & Abdulhadi, Z. Q. (2013). A Comparison Study of Data Scrubbing Algorithms and Frameworks in Data Warehousing. *International Journal of Computer Applications*, 68(25).
- [14] Kulkarni, Prerna S., and J. W. Bakal. "Survey on Data Cleaning." structure 3.4 (2014).
- [15] Patil, R. Y., & Kulkarni, R. V. (2012). A review of data cleaning algorithms for data warehouse systems. *International Journal of Computer Science and Information Technologies*, 3(5), 5212-5214.
- [16] Akmal, H., Fox, R., Iqbal, S., & Khan, Y. D. An Educational Data Science Approach Towards Prediction Of Student Performance. *IACB, ICE & ICTE*.
- [17] Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A taxonomy of dirty data. *Data mining and knowledge discovery*, 7(1), 81-99.
- [18] Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2(1), 9-37.
- [19] Noh, J. B., Lee, K. C., Kim, J. K., Lee, J. K., & Kim, S. H. (2000). A case-based reasoning approach to cognitive map-driven tacitknowledge management. *Expert systems with applications*, 19(4), 249-259.
- [20] Maletic, J. I., & Marcus, A. (2009). Data cleansing: A prelude to knowledge discovery. In *Data Mining and Knowledge Discovery Handbook* (pp. 19-32). Springer, Boston, MA.