A Review of Data Cleansing Concepts – Achievable Goals and Limitations

Kofi Adu-Manu Sarpong
Institute of Computer Science
Valley View University, Accra-Ghana
P.O. Box VV 44, Oyibi-Accra

John Kingsley Arthur
Institute of Computer Science
Valley View University, Accra-Ghana
P.O. Box VV 44, Oyibi-Accra

ABSTRACT

Data Cleansing is an activity involving a process of detecting and correcting the errors and inconsistencies in data warehouse. It deals with identification of corrupt and duplicate data inherent in the data sets of a data warehouse to enhance the quality of data**.** The study looked into investigating some research works conducted in the area of data cleansing. A thorough review into these existing works was studied to determine the achievable goals and the limitations that arose based on the approaches conducted by the researchers. They identification of errors by most of these researchers has led into the development of several frameworks and systems to be implemented in the area of data warehousing.

Generally, these findings will contribute to the emerging empirical evidence of the strategic role data cleansing play in the growth of organizations, institutions and other government agencies in terms of data quality and reporting purposes and also to gain competitive advantage since they will overcome the mere existence of dirty data.

General Terms

Data warehousing, data cleansing, quality data, dirty data

Keywords:

Data inconsistency, identification of errors, organization growth

1. INTRODUCTION

Data warehousing is a technology that seeks to create a repository for all sources of data needed by the organization so that, information could be accessed and used from one point. Companies over the last couple of decades have done more data logging and capturing with the advent of computers with database capabilities. Many have found that these data are quite useful to augment or focus market groups if only the information was available for statistical analyses.

Cleansing data of errors in structure and content is important for data warehousing and integration. Current solutions for data cleaning involve many iterations of data "auditing" to find errors, and long-running transformations to fix them. There are various approaches used in cleaning data in manufacturing industries, schools/colleges/universities, organizations and many more. Users need to endure long waits, and often write complex transformation scripts.

In this paper, we discuss and review five different papers in the area of data cleansing. The following papers are discussed: Problems, Methods, and Challenges in Comprehensive Data Cleansing, Data Cleaning: Problems and Current Approaches, Matching Algorithms with a Duplicate Detection System, Open User Involvement in Data Cleaning for Data Warehouse Quality and the Three Tier level Data Warehouse Architecture for Ghanaian Petroleum Industry. This paper looks at different dimensions of the data cleansing process and in some cases a proposed framework and/or algorithm are discussed.

2. REVIEWED PAPERS ON DATA CLEANING

There has been quite a number of works and researches done on data cleaning with its application to enterprise data warehouse as well as data marts for operational systems for many organizations. This research reviews some of them with special reference to the methods used and the frameworks proposed to clean dirty data in operational databases from different sources.

2.1 Problems, Methods, and Challenges in Comprehensive Data Cleansing

In this paper, [1] classified data quality problems into syntactical anomalies which concern data formats and values for data representation (e.g. lexical errors, domain format errors and irregularities). Lexical errors name discrepancies between the structure of data items and the specified format. Domain errors specify where the given value for an attribute A does not conform to the anticipated domain format G (*dom*(A)). Irregularities deal with non-uniformed use of values and other abbreviations which normally are noticeable when different currency format is used to specify employee salary.

The authors also discussed the Semantic anomaly which hinders data collection from being comprehensive and non-redundant representation of the world such as integrity constraints, contradictions, duplicates and invalid tuples. Finally, Coverage anomaly is discussed by the authors in regard to data quality problems. Here the amount of entities and their properties are represented as missing values and missing tuples from real world data collections. In the work of Müller and Freytag no appropriate framework was designed to support the cleansing process [1].

2.2.1 What the research achieved

The research was able to identify several problems and challenges such as the inability of researchers to state the details of the implementation of cleaning in comprehensive data cleaning. This will help many researchers to be able look into the identified problems and come up with solutions appropriately.

2.2.2 Limitations

i. Maintenance is not considered in the framework for cleaning of data. Hence after cleaning of data, the question is what next? Data needs to be maintained.

Although several cleaning tools were compared in terms of the file formats that they can clean, however, how effective they do the cleaning was swept under the carpet. The effectiveness of these tools should be clearly known.

2.2 Data Cleaning: Problems and Current Approaches

According to [2], the classification of data quality problems can be divided into two main categories: single-source and multiple-source problems. At the single-source, Rahm and Do divide these into schema level and instance level related problems without considering the occurrence in a single relation.

The single-source problems deal with attribute, record, record type and source whereas the multiple-source problems deal with naming conflicts, schema-level conflicts and the identification of overlapping data which refers to same real-world entity.

2.2.1 What the research achieved

i. The research clearly provided a classification of data quality problems in data sources differentiating between single- and multi-source and between schema- and instance-level problems.

ii. The research also, further outlines the major steps for data transformation and data cleaning and emphasized the need to cover schema- and instance-related data transformations in an integrated way.

iii. The research provided an overview of commercial data cleaning tools. It was identified that, while the state-of-the-art in these tools is quite advanced, they do typically cover only part of the problem and still require substantial manual effort or self-programming.

Limitations

i. The research work is insufficient of the design and implementation details of the best language approach for supporting both schema and data transformations. For instance, operators such as Match, Merge or Mapping Composition have either been studied at the instance (data) or schema (metadata) level but may be built on similar implementation techniques.

ii. Data cleaning is not only needed for data warehousing but also for query processing on heterogeneous data sources, e.g., in web-based information systems. This environment poses much more restrictive performance constraints for data cleaning that need to be considered in the design of suitable approaches; however, they were untouched in this work.

2.3 Matching Algorithms with a Duplicate Detection System

According to [3], detecting database records that are approximate duplicates, but not exact duplicates, is an important task. These duplicates concern the same real world entity due to data entry errors, abbreviations that are not standardized or differences in schema level records from multiple databases. In this paper, the approximate duplicate detection problem was explored and the following contributions were made:

- The author proposed how to compute the transitive closure of "is duplicate of" relationships incrementally by using the union-find data structure (R_i and R_j) operation where R_j is compared to R_i using the matching algorithm.
- The other contribution made by the author was the implementation of heuristic method for minimizing the number of expensive records while comparing individual records.

2.3.1 What the research achieved

The research contributed to the existing algorithms for detecting duplicates of records by introducing an integrated framework, which draws from the existing algorithms.

Limitations

The research was able to measure the effectiveness of the developed algorithm against the standard algorithm, however, with very minimal iterations. It is also identified that minimal iterations could result in inability to identify duplicate records. Therefore, there is doubt on the outcome of this research.

2.4 Open User Involvement in Data Cleaning for Data Warehouse Quality

The issue of data cleaning caught the attention of researchers not so long ago and the data quality management in data warehouse has to overcome several inherent problems. Data quality is vital to organizations since it enables them make informed and accurate decisions [8, 9]. According to [4] data cleansing is a relatively new research field. Although after Maletic and Marcus in the year 2000 who proposed a framework which other authors/researchers have looked into and proposed slight different architectures, the outcome still needs further investigation. The process is computationally expensive on very large data sets and thus it was almost impossible to do with old technology [5, 8].

The new faster computers allow performing the data cleansing process in acceptable time on large amounts of data. There are many issues in the data cleansing area that researchers are attempting to tackle. They consist of dealing with missing data, determining record usability, erroneous data, etc. Different approaches address different issues [5, 8].

The following problems were identified in the paper:

- Data warehouse projects mostly integrate the data quality phase into the extract-transform-load (ETL) process but do not allocate enough time for efficient data validation with respect to data cleansing.
- When data has been extracted from operational data sources, feedback of data cleaning results is not taken into

account even though it's important to avoid redoing the same data cleaning tasks.

- It can be realized that some erroneous data require manual intervention and others require combining the manual and automatic data cleansing approaches as cited by [3].

Some of the issues Bradji., *et al*., discussed in their work was to extend the data cleaning process for users at the operational sources to validate the data, and they also considered the adaptation of ETL to support their data cleaning process.

The purpose of this research is to establish that involvement of the user in the data cleaning process for the purposes of validating the data at the operational sources before propagating the cleaned data.

Reviewing the work done in this paper, the following were realized:

i. Data extracted from different operational sources could not be easily sent for re-validation for the fact that some of the errors could be at the schema or instance level and could not be easily taken care of.

ii. The proposed framework only work for small amount of data being propagated and was not tested for large data sets.

2.4.1 What the research achieved

i. The research extended the data cleaning and extract-transform-load (ETL) processes to better support the user involvement in data quality management.

ii. Systems user interface is graphical and hence easy to be operated by end-users.

2.4.2 Limitations

The data warehouse that was used for the testing of results was small. Huge data warehouse will bring out the exact outcome of the research.

i. There is an extensive involvement of users. For example, after detection of duplicates (errors), the end user is asked to validate. Continues human validation may result in other human errors in the data warehouse.

2.5 Three Tier level Data Warehouse Architecture for Ghanaian Petroleum Industry

According to [6], data warehouse (DW) is a modern proven technique of handling and managing diversity in data sources, format and structure and he added that recent advances in database technologies are leading to the proliferation of different kinds of information design with independent supporting hardware and software.

This technology was developed therefore to integrate heterogeneous information sources for analysis purposes. In this paper they realized that, Information sources are progressively becoming autonomous and they change their content rapidly due to perpetual transactions (data changes) and may change their structure due to continual users' requirements evolving (schema changes).

However, handling correctly all type of these changes was the real challenge the research is supposed to comprehend. Strategic decision making is an ongoing process and part of human life. Businesses all over the world are faced with challenging decision making which is the very livelihood to their very existence.

There has been intensive research in the field of databases including the grandfather of data warehousing Inmon which they offered tremendous insight into how the power of computing can be harnessed in strategic decision making. The achievement in the petroleum industry has been highlighted and written about by Nimmagadda and his fellow professionals. However, how this can be implemented in Africa remains to be solved.

In this paper, [6] investigated the various data capturing systems which are currently in use at the Ghana National Petroleum Cooperation (GNPC) – a state agency responsible for the overall management and supervision of the hydrocarbon endowment of the sub-region and made recommendation that a tree tier level petroleum data warehouse architecture for cutting-edge decision making on petroleum resources should be adapted as the architecture for a data warehouse.

2.5.1 What the research achieved

It provides a comprehensive and effective framework for managing information of GNPC.

2.5.2 Limitations

The case for this study is narrowed and hence, the findings of this research to the other companies in the process of commencing exploratory work in their respective fields due to issues of underlying geological rock differences and the contents of such fields (that is gas or oil or both).

3. CONCLUSION

Data cleansing has become a major activity performed by most organizations that have data warehouses. Every organization needs quality data to improve on its services it renders to its customers. In view of this a thorough review of approaches and papers in that regard are discussed and their limitations also stated. This is to help future development and research directions in the area of data cleansing. The papers reviewed in this report looked at critical aspects of data cleansing and the various types of data that could be cleansed. Several algorithms have been proposed in the various works discussed.

4. REFERENCES

[1] Heiko Muller, Johann-Christoph Freytag. (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing, pp. 21.

[2] Rahm, E., Do, H.H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bull. Vol 23 No. 4, pp. 3-13

[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] Jonathan I. Maletic, Andrian Marcus. (2000). Data Cleansing: Beyond Integrity Analysis, pp. 8.

[5] Louardi BRADJI, Mahmoud BOUFAIDA. (2011). Open User Involvement in Data Cleaning for Data Warehouse Quality. International Journal of Digital Information and Wireless Communications (IJDIWC) 1(2), pp. 573.

[6] Deku JerryYao,Mohammad Sarrab 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

[7] Chapman, A. (2005). Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, Version 1.0. Report for the Global Biodiversity Information Facility,Copenhagen,pp7

[8] Bradji, L., Boufaida, M. (2011). Knowledge based data cleaning for data warehouse quality.In: Proc. 2011 International Conference on Digital Information Processing and Communications, ICDIPC2011,LNCS,Part II, CCIS no 189,pp.373 -384

[9] Vassiliads, P.(2009). A Survey of Extract-Transform-Load Technology. In InternationalJournal of Data Warehousing & Mining,vol.5 ,no. 3, pp. 1-27