Survey On A Novel Cost-Based Model For Data Repairing

S. Banupriya[1], S. M. Jagatheesan[2]
Research Scholar, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, India.
Associate professor of Computer Science, Gobi Arts & Science College, Gobichettipalayam, India.

Abstract- *Data warehousing and Data mining has become one of the important factors in the business field. This needs to have strategies to manage large volumes of structured, unstructured and semi-structured data. It's challenging to analyze such large scale of data to extract data meaning and handling uncertain outcomes. The consistent data in an inconsistent database is usually characterized as the data that persists across all the database instances that are consistent and minimally differ from the inconsistent instance. Those are the so-called repairs of the database. Almost all data sets are dirty, i.e. the set may contain inaccuracies, missing data, miscoding and other issues. One of the biggest challenges in data analytics is to discover and repair dirty data; failure to do this can lead to inaccurate analytics and unpredictable conclusions. Data cleaning is an essential part of managing and analyzing data. In this survey paper, data quality troubles which may occur in data processing to understand clearly why an organization requires data cleaning are examined, followed by data quality.*

Keywords- Integrity Constraints, Semantics, Data repairing, Data cleaning, Dirty data, Violation.

I. INTRODUCTION

When overlapping or redundant information from multiple sources is integrated, inconsistencies or conflicts in the data may appear as violations of integrity constraints on the integrated data. Conflicts in this data may be introduced for many reasons, including misspellings or differing conventions used during data entry, different processes and time-scales for performing updates and so on.

Data quality is essential to all businesses, which demands dependable data cleaning solutions. Data cleaning, which is to detect and repair data errors, has played an important part in the history of data management, because high-quality business decisions must be made based on high-quality data. Data cleaning is not magic; it cannot guess something from nothing. What it does is to create decisions from evidence. Certain data patterns of semantically related values can give evidence to precisely capture and rectify data errors.

The data cleaning process gets more complex when data comes from heterogeneous sources. Here, data quality problem has to be solved by data cleaning and data transformation. Despite of the various viewpoints on the effect of data quality, in the end, all have the probability to produce in economic expenses for groups. By some evaluations it is known that the in organizations and companies issue of dirty data already reached to epidemic amounts. The issue is equally prevalent and hypothetically equal beyond frightening in health care and other organization [7].

First and foremost, Experian approximates average 12% loses in business due to wrong records causing productivity reduction, resources wastage, and significantly, misused chances for marketing of cross-channel. The Experian investigation also focuses that approximately one-third of responders think that they waste almost 10% or more budget in marketing because of outcome obtained from inaccurate data.

II. LITERATURE REVIEW

In this work et.al (2) G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma has proposed two central criteria for data quality are consistency and accuracy. Inconsistencies and errors in a database often emerge as violations of integrity constraints. This paper studies effective methods for improving both data consistency and accuracy. Employment a class of conditional functional dependencies (CFDs) to specify the consistency of the data, which are able to capture inconsistencies and errors beyond what their traditional counterparts can catch. To improve the consistency of the data, two algorithms are suggested: one for automatically computing a repair DB that satisfies a given set of CFDs, and the other for incrementally finding a repair in response to updates to a clean database.

In this work et.al (1) M. Arenas, L. E. Bertossi, and J has proposed the problem of the logical characterization of the notion of consistent answer in a relational database that may violate given integrity constraints. This notion is captured in terms of the possible repaired versions of the database. Integrity constraints capture an important normative aspect of every database application. In this paper provides a logical

characterization of consistent query answers in relational databases that may be inconsistent with the given integrity constraints. Intuitively, an answer to a query posed to a database that violates the integrity constraints will be consistent in a precise sense: It should be the same as the answer obtained from any minimally repaired version of the original database.

In this work et.al (9) X. Chu, P. Papotti, and I. Ilyas tackle the problem in a novel, unified framework. It is well recognized that business and scientific data are increasing exponentially and that they have become a first-class asset for any institution. However, the quality of such data is compromised by sources of noise that are hard to remove in the data life-cycle: imprecision of extractors in computer-assisted data acquisition may lead to missing values, heterogeneity in formats in data integration from several sources may introduce duplicate records, and human errors in data entry can violate declared integrity constraints. These issues compromise querying and analysis responsibilities, with possible damage in billions of dollars. Given the use of clean data for any operation, the capability to improve their quality is a key constraint for useful data management.

In this work et.al (10) X. Chu, M. Ouzzani, J. Morcos, I. F. Ilyas, P. Papotti, N. Tang, an Y. Ye. KATARA has present Katara, a novel end-to-end data cleaning system powered by knowledge bases and crowdsourcing. Table understanding, including identifying column types and the relationship between columns, has been addressed by several techniques. KATARA differs from them in two main aspects. First, this paper focus on finding coherent table patterns for the purpose of data cleaning. In other words, instead of explaining table semantics at the schema level, they need to and table patterns that can align in- formation at the instance level. Second, existing techniques do not explicitly assume the presence of dirty data.

In this work et.al(4) L. Berti-Equille, T. Dasu, and D. Srivastava has proposed that as data types and data formats change to keep up with evolving technologies and applications, data quality problems too have evolved and become more complex. Data streams, Web logs, Wikipedia, biomedical applications, a social networking websites generate a mind boggling variety of data types with attendant data quality problems. In real world data, different types of glitches co-occur in complex patterns such as the simultaneous occurrence of outliers and missing values in the same record. In the last two decades, a large body of work in the DB community has focused on declarative data cleaning and more recently, on constraint based data repair.

In this work et.al (5) L. E. Bertossi, S. Kolahi, and L. V. S. Lakshmanan has proposed matching dependencies were recently introduced as declarative rules for data cleaning and entity resolution. Enforcing a matching dependency on a database instance identifies the values of some attributes for two tuples, provided that the values of some other attributes are succinctly similar. Matching functions naturally introduce a lattice structure on attribute domains, and a partial order of semantic domination between instances. Using the latter, define the semantics of clean query answering in terms of certain/possible answers as the greatest lower bound/least upper bound of all possible answers obtained from the clean instances. They show that clean query answering is intractable in some cases. Then they study queries that behave monotonically w.r.t. semantic domination order, and show that they can provide an under/over approximation for clean answers to monotone queries. Moreover, non-monotone positive queries can be relaxed into monotone queries.

In this work et.al (3) F. Chiang and R. J. Miller has proposed Integrity constraints are the primary means for preserving data integrity. Constraints represent domain specific rules and relationships that hold over any database instance that accurately reflects the domain. Typically, constraints are defined at design time when a data architect with domain knowledge precisely defines the semantics of the application data. If every constraint is enforced within a database, then the data, as it evolves, will continue to conform to the constraints. In reality however, constraints may not be enforced. Relaxed enforcement policies may allow the data to become inconsistent with respect to the constraints. To manage this, several approaches have proposed techniques to repair the data, by finding minimal or lowest cost changes to the data that make it consistent with the constraints.

III. DATABASE INTEGRATION

Often many different databases are integrated together to provide a single unified view for the users. Database integration is difficult since it requires the resolution of many different kinds of discrepancies of the integrated database. One possible discrepancy is due to different sets of integrity constraints. Moreover, even if every integrated database locally satisfies the same integrity constraint, the constraint may be globally violated. For example, different databases may assign different addresses to the same student. Such conflicts may fail to be resolved al: all and inconsistent data cannot be “cleaned” because of the autonomy of different databases. Therefore, it is important to be able to find out, given a set of local integrity constraints, which query answers returned from the integrated database are consistent with the constraints and which are not.

TYPES OF DATA

Database refers to the collection of data that surround us. Each type of data differs in the way it is created, stored and analyzed. Based on nature and characteristics, data is categorized under these three types.

Structured Data:

- Structured data refers to the type of data that is stored in databases in a systematic manner.
- It constitutes 20% of the total data and is used in most of the programming and computer-related activities.
- Examples :
  - Meta-data (Time and date of creation, File size, Author etc.)
  - Library Catalogues (date, author, place, subject, etc)
  - Census records (birth, income, employment, place etc.)
  - Economic data (GDP, PPI, ASX etc.)

Unstructured Data:

- Unstructured data, as the name suggests, is not systematic at all; it refers to any kind of data that carries unknown form or structure.
- It is also understood better for the challenge it represents in terms of processing it.
- It cannot be stored, obviously, in the way structured data can be stored in spreadsheets.
- Examples :
  - Media (MP3, digital photos, audio and video files)
  - Text files (Word processing, spreadsheets, presentations etc.)
  - Social Media (Data from FaceBook, Twitter, LinkedIn)

Semi-structured Data:

- Semi-structured data is not the kind of data that conforms to the formal structure of data models associated with relational databases or other forms of data tables.
- However, it does contain tags or other markers to segregate semantic elements and enforce hierarchies of records and fields within the data.
- Therefore, it is also known as a self-describing structure.
- Examples :
  - Personal data stored in a XML file.

IV. DATA QUALITY CRITERIA

Data quality has different definition on different field and period. Researcher and expert made different understanding about data quality. According to quality management data quality is appropriate for use or to meet user needs or it is quality of data to meet customer needs [9].Also, another definition for data quality is fitness for use. Indeed, quality of data is critical for improvement process activity as it can be addressed in different field including management, medicine, statistics and computer science. The widespread collection of definition through data quality may give opportunity to better understand the nature of data process.

Data quality is generally described as the capability of data to satisfy stated and implied needs when used under specified conditions [6]. Data accuracy, completeness and consistency are most popular initiatives to address Data quality [10], beside other dimensions like Accessibility, Consistent representation, timeliness, Understandability, Relevancy, etc. Moreover, data quality is combination of data content and form. Where data content must contain accurate information and data form essential be collected and visualized in an approach that creates data functioning.

Content and form are significant consideration to reduce data mistakes, as they illuminate the task of repairing dirty data needs beyond simply providing correct data. Likewise, while developing a scheme to improve data quality it is essential to identify the primary reasons of dirty data. The causes are categories into organized and unintentional errors. The basic sources of producing systematic errors include while programming, wrong definition for data types, rules not defined correctly, data collection’s rules violation, badly defined rules, and trained poorly.

The sources of random errors can be errors due to keying, unreadable script, data transcription complications, hardware failure or corruption, and errors or intentionally misrepresenting declarations on the portion of users specifying major data. Human role on data entry usually result error, this error can be typos, missing types, literal values, Heterogeneous ontologies (i.e. Different nature of data), outdated values or Violations of integrity constraints.

Therefore, the most common dimensions of dirty data including data duplication are:

- Inaccurate data refers to any field contains wrong values. A right value of data will bring accurate and signified arrangement of consistency and unambiguous.
- Incomplete data from missing data is produced by data sets basically missing values. These type of data considered concealed when the amount of values identified in a set, but the values themselves are unidentified, and it is also known to be condensed when there are values in a set that are eliminated.
- Inconsistent data is data redundancy; i.e. same data value is stored in different files which may be in different formats.
- Duplicate data is entries that have been added by a system user same data multiple times.

V. CONCLUSION

In this paper, an overview is initiated to identify the potential of data cleaning in database analytics in the process of gathering, arranging and processing information. It is important to understand data quality criteria of dirty data to able to clean data sets without failure. A comparison of commercialized tools is presented by obtaining comments from different customers. Most of the tools mostly concerns to organize data sets and clean messy data and very methods uses machine learning. But they didn't give much importance to database characteristics, which may lead to big challenge while cleaning data. There are many available data repairing algorithms, still it required human expert to take intelligent decision if the cleaning process is correct or not. Machine learning algorithms will probably replace most jobs in the world, with the fast evolution of big data and accessibility of programming tools like Python and R; machine learning is increasing mainstream existence for data scientists. Machine learning applications are highly automated and self-modifying which continue to improve over time with minimal human intervention as they learn with more data.

REFERENCES

[1] M. Arenas, L. E. Bertossi, and J. Chomicki. Consistent query answers in inconsistent databases. TPLP, 3(4-5), 2003.
[2] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma. Improving data quality: Consistency and accuracy. In VLDB, 2007.
[3] F. Chiang and R. J. Miller. A unified model for data and constraint repair. In ICDE, 2011.
[4] L. Berti-Equille, T. Dasu, and D. Srivastava. Discovery of complex glitch patterns: A novel approach to quantitative data cleaning. In ICDE, pages 733–744, 2011.
[5] L. E. Bertossi, S. Kolahi, and L. V. S. Lakshmanan. Data cleaning and query answering with matching dependencies and matching functions. In ICDT, 2011.
[6] F. P. Sidi, Hassany, S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, A. Mustapha, "Data Quality: A Survey of Data Quality Dimensions". In: 2012 International Conference on Information Retrieval & Knowledge Management (CAMP). pp. 300–304. IEEE, Kuala Lumpur, Malaysia, Malaysia (2012).
[7] VHAInc: The Cost of Dirty Data. (2012).
[8] X. Chu, P. Papotti, and I. Ilyas. Holistic data cleaning: Put violations into context. In ICDE, 2013.
[9] X. Chu, M. Ouzzani, J. Morcos, I. F. Ilyas, P. Papotti, N. Tang, and Y. Ye. KATARA: reliable data cleaning with knowledge bases and crowdsourcing. PVLDB, 8(12), 2015.
[10] I. Taleb, H. T. El Kassabi, M. A. Serhani, R. Dssouli, C. Bouhaddioui, "Big Data Quality: A Quality Dimensions Evaluation". In 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress. pp. 759–765. IEEE (2016).