International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:10, 2015

Research of Data Cleaning Methods Based on Dependency Rules

Yang Bao, Shi Wei Deng, Wang Qun Lin

***Abstract*—**This paper introduces the concept and principle of data cleaning, analyzes the types and causes of dirty data, and proposes several key steps of typical cleaning process, puts forward a well scalability and versatility data cleaning framework, in view of data with attribute dependency relation, designs several of violation data discovery algorithms by formal formula, which can obtain inconsistent data to all target columns with condition attribute dependent no matter data is structured (SQL) or unstructured (NoSql), and gives 6 data cleaning methods based on these algorithms.

***Keywords*—**Data cleaning, dependency rules, violation data discovery, data repair.

I. Introduction

WITH the accelerating process of enterprise informatization, the enterprise has accumulated huge amounts of data resources. However, user input error, the merger of enterprises and the change of business environment with the passage of time, will affect the quality of stored data, caused a lot of "dirty data". The purpose of Data Cleaning is to detect and eliminate the data in the presence of errors and inconsistencies, to improve the quality of data. At present the research on data cleaning mainly focus on the following aspects: duplicate objects detection, missing data processing, abnormal data detection, logical error detection and inconsistent data processing etc.

In view of the current situation of the data quality, many domestic and foreign scholars have put forward some data cleaning framework and methods, representative: National University of Singapore proposed intelligent data cleaning framework based on knowledge [1]; a data cleaning tool can be extended is Ajax from France INRIA Research Organization [2]; Berkeley University of California proposed the interactive data cleaning system [3]. Some scholars take the technology of data mining and data cleaning combination, proposed the data cleaning framework and model based on rough-set theory, neural network, expert system, but they generally lack of interoperability, it is difficult for users to gradually analyze and adjust the processes and procedures of data cleaning, therefore it lacks the versatility.

Based on the analysis of current situation of data quality and the existing data cleaning framework, this paper presents the Data Cleaning method based on Dependency Rules, which solves the interactivity and dependency rules judgment difference problem of traditional data cleaning methods in a certain extent.

Y. Bao, S.W. Deng, and W.Q. Lin are with Beijing Institute of System Engineering, Beijing, China (phone: +8613810955822; e-mail: baoyang18@163.com, s.w.deng@163.com, linwangqun2005@163.com).

II. The Principle of Data Cleaning

Data cleaning are widely used in Data Warehouse, Knowledge Discovery in data (KDD) and data quality management (DQM) of these three areas, because it is related to the academic field, so the concept of data cleaning has no unified definition. But the purpose of different application areas is the same that is eliminating the "dirty data", so as to improve the quality of data. In this paper, the definition of data cleaning is to detect and eliminate the "dirty data".

A. Basic Principle of Data Cleaning

The basic principle of data cleaning is: analyzing the reasons of "dirty data" and the type of error data, then according to the data quality problem of each type, propose cleaning rules and the corresponding cleaning algorithm, and implement these rules and algorithm, then transform "dirty data" to "clean data" which meet user or application requirements.

B. Types and Causes of Dirty Data

There are many types of dirty data, reasons of each kind of dirty data are not same, this paper described them including single data source and multiple data sources, classified dirty data as mode layer problem and instance layer problem respectively, Table I lists the types, and reasons of "dirty data".

C. Data Cleaning Process

In general, data-cleaning process includes the following procedure: first, analyze the source data set of dirty data, get the data quality metadata. On this basis, define cleaning rules, select cleaning strategy, determine cleaning algorithm, using the sample data set to test the correctness and efficiency of these cleaning rules and algorithm, show all the parameters which reflect the quality of the data to the user, by the user to decide whether or not to modify or redefine the cleaning rules, improve cleaning algorithm, then store the last confirmed cleaning rules in the rule library, finally use them to clean dirty data [4]. Usually some data error will appear at the times of cleaning, so often need clean "dirty data set" iteratively, will get clean data that user satisfied. Specifically, data cleaning includes the following steps:

Step1. Data analysis, in order to detect the type of "dirty data", need for a detailed data analysis, in addition to manual inspecting data attributes and the sample data set, also should be automatically access the metadata about data quality through analysis program.

Step2. Define the cleaning rules and workflow: according to the metadata data obtained from the first step analysis

about data quality, define the cleaning rules under the considering of the number, the heterogeneity degree between different data source, and the number of "dirty data", then make the execution order of cleaning rules. Each cleaning rule can be stated as follows: cleaning type, condition, action, strategy, manipulated record-set and operation-set}, the order of workflow definition is reasonable arrangements from the selected rules.

TABLE I
THE TYPES AND REASONS OF "DIRTY DATA"

Source	Layer	Type	Reason
Single data Source	*mode*	field constraints	Not in the constraints
		Attribute dependency conflict	Two dimensions of value is not consistent
		the only conflict (such as primary key)	primary dimension exactly the same, but exists two different record attached
		referential integrity conflict	beyond the set range, no corresponding object data
	instance	single attribute embedded value	a single attribute value contains too much information
		spelling mistakes	error in the input data, data transmission error
		blank value	the user is not willing to reveal information, a lot of optional input form design
		noise data	real data, deliberate error, data acquisition equipment error, data transmission error
		data duplication	refers to the same entity in reality with a plurality of not identical records in a data set to show
Source	Layer	Type	Reason
multiple data source	mode	naming conflicts	the name with a different entity
		structural conflict	the unit of measurement is not consistent, inconsistent use of different code, one code has inconsistent meaning, the same meaning code have different format
	instance	time inconsistency	different time level data are compared and calculated at the same level
		size inconsistency	different levels of data were compared and calculated at the same level
		data duplication	the same data appears two times or more above the combined database
		…	

Step3. Perform the data cleaning rules: according to the execution process determined on the last step, by the rule execution engine to execute these rules in the source data set, and the data cleaning results are displayed to the user.

Step4. Verification: To evaluate the efficiency and accuracy of the workflow execution of cleaning rules, inspect the cleaning efficiency and the correctness, find it whether satisfy user or application requirements. According to the results of data cleaning, changing the rules, modify the workflow, the real cleaning procedure need data analysis, design, verification iteratively, then get satisfied data cleaning rules and workflow.

Step5. Execute the step 2-3 repeatedly, until finishing all the data quality problems.

Step6. Execute the step1-4 repeatedly, until meeting the requirements of user or application on the cleaned data.

III. DATA CLEANING METHODS BASED ON DEPENDENCY RULES

At present, the market has been the emergence of many data cleaning methods and commercial tools, such as IBM, DataStage, Oracle, OWB, SQLServer DTS etc. They have provided some data cleaning function, but also have significant limitations: lack of generality, lack of scalability and ease of use, limited cleaning function. Basing on the advantages of current data cleaning framework, this paper propose an interactive data cleaning framework, and give some methods based on dependency rules, it will make "dirty data" and domain knowledge be represented as rules, has good interactivity, expandability and generality.

A. Data Cleaning Frame

Fig. 1 is the cleaning frame [5] based on dependency rules.

1) Heterogeneous data source: As the general frame, the ability of heterogeneous data sources support is very important, not only to support the large relation data base system (such as Oracle, SQLServer, DB2, Sybase etc.) and all kinds of common data access interface (such as ODBC, JDBC), but also should be as much as possible to support NoSql data sources, the data in a variety of file formats (such as Excel forms, text file data).
2) Target data source: Store the "clean" data after data cleaning, usually data warehouse, for data integration, data mining and business online analytical processing (OLAP) to provide data supporting.
3) Common data access interface: The interface can access the data from different platform and data source of different position in the network, support connecting and access each other among many types of data sources, effectively shielding the underlying heterogeneity of the data source.
4) Rule execution engine: Analyze the rules configuration file is received, get which rules need to perform, and set reasonable arrangements for the order of rules execution.
5) Rules Library: It is meta-database used to store the data cleaning rules.
6) User-defined rules module: It is interactive interface between user and data cleaning system, through this interface, legal users can define new cleaning rules and domain rules, the coverage and assess the coverage and accuracy rate by sample data.
7) Automatic rule generation module: The rules store can be generated automatically. First, the user provided data set is randomly partitioned into training and test datasets, in which, the training data set is used to generate the rules, and the test data set used to evaluate and validation for generated rules. Rules learning device select machine learning algorithms from algorithm library, and apply them into the training data set, the generated rules into the temporary rules repository. Rules extractor extracts from it, and inspects them with the test data; compares the test

parameters and the presets parameters threshold to decide whether be stored in the rule base ultimately.

8) Algorithms library: It is used to store the data cleaning algorithms, such as the record matching algorithm, the edit distance algorithm etc.

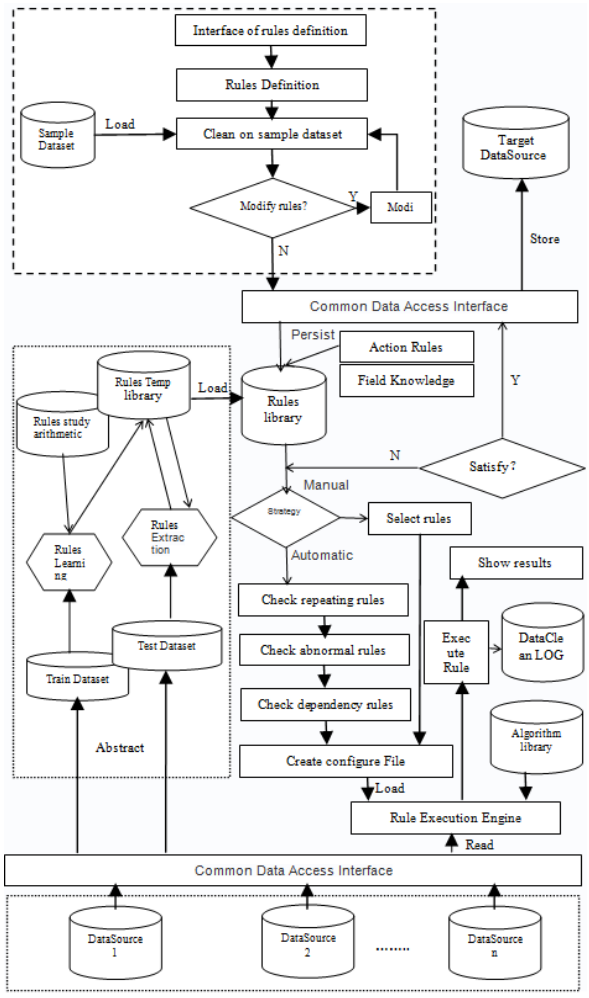

Fig. 1 Data cleaning frame

B. Violation Data Discovery Algorithms

The most important and tedious work of data cleaning is discover dirty data, good judgment logic and scientific statistics can be as accurate as possible positioning errors occur location and coverage rate [6]. This paper will put forward a set of data discovery algorithm based on the attribute dependence rules, using some statistics formalization formula to express.

1) Element and Structure Matching

The matching algorithm only consider properties of individual columns and utilize the relationship between columns by their column name or domain similarities (interpreted) or similar data distributions (uninterpreted).

$$H(X) = -\sum_{x \in T} p(x) \log\ p(x) \tag{1}$$

X be an attribute with alphabet T, x is each value, p(x) is probability distribution of X, (1) describes the uncertainty of values in an attribute. If two joint alphabet T1 and T2,(2) is the correlation between the two attributes probability distributions, it consider joint probability distribution p(x,y) and marginal probability distributions p(x) and p(y) over two attributes, and can measure the amount of information captured in one attribute about the other.

$$MI(X,Y) = \sum_{x \in T_1} \sum_{y \in T_2} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

Other, the value of conditional dependency attributes is (3), it measures the uncertainty of attribute X given knowledge of attribute Y.

$$H(X \mid Y) = -\sum_{x \in T_1} \sum_{y \in T_2} p(x,y) \log p(x,y)$$
$$MI(X,Y) = MI(Y,X) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \tag{3}$$

Based on the above, can transform input tables into dependency graph A and B, and based on the graph, calculate the distance metric between T1 and T2 (two equal size dependency graphs), according to (4), the minimum distance got from the metric is always the distance of a single best matching node pair, that is the important step of matching strategies, further data cleaning or improvement is possible by exploiting inter-attribute correlations.

$$D_M^U(A,B) = \sqrt{\sum_{i,j} (a_{ij} - b_{m(i)m(j)})^2} \tag{4}$$

2) Functional Dependencies

Functional Dependencies (FD) states that the value of an attribute is uniquely determined by the values of some other attributes, but how to evaluate and calculate the FD is difficult before cleaning the violation data, as follows, will give three algorithms, which is the basis of identifying the more inconsistencies.

The 1th: Soft Functional Dependencies, that is defined as the strength based on a rule $\boldsymbol{X} \rightarrow$(two dependence attributes) over a relation $\boldsymbol{T}$, $|\boldsymbol{dom}(\boldsymbol{X})|_T$ denotes the number of distinct values in the attribute $\boldsymbol{X}$ of table $\boldsymbol{T}$, $|\boldsymbol{dom}(\boldsymbol{X},\mathrm{Y})|_T$ denotes the number of distinct values in the concatenation of attributes $\boldsymbol{X}$ and $\boldsymbol{Y}$ in table $\boldsymbol{T}$.

$$S(X \rightarrow Y, T) = \frac{|dom(X)|_T}{|dom(X,Y)|_T} \tag{5}$$

The 2th: Approximate Functional Dependencies, that is to evaluate the exact proportion of elements with violations, (6) gives the error measure $g(\boldsymbol{X} \rightarrow \boldsymbol{Y},\mathrm{T})$,it is the ratio of the minimum number of elements that need to be removed from $\boldsymbol{T}$ to make $\boldsymbol{X} \rightarrow \boldsymbol{Y}$ hold on $\boldsymbol{T}$, where $\boldsymbol{S}$ is a subset of objects in $\boldsymbol{T}$ that do not violate $\boldsymbol{X} \rightarrow \boldsymbol{Y}$,|| is the statistical number, the smaller g value is the more likely to be a FD [7].

$$g(X \to Y, T) = \frac{|T| - \max\{|S| \mid S \subseteq T, S \notin X \to Y\}}{|T|} \quad (6)$$

The 3th: Probabilistic Functional Dependencies, to compute the probability of $\boldsymbol{X}\to\boldsymbol{Y}$, need get the average of probabilities of $\boldsymbol{X}\to\boldsymbol{Y}$ holding for each distinct value of $\boldsymbol{X}$.

TABLE II
ALGORITHMS COMPARISON

Algorithms rules	Statistical Measures	Features
Soft FDs	S(X→Y,T)	Easy to compute by domain
Approximate FDs	G(X→Y,T)	More accurately tell the violation elements proportion
Probabilistic FDs	Pr(X→Y,T)	Normalization with domain instead of elements

$$\Pr(X \to Y, V_X) = \frac{|V_Y, V_X|}{|V_X|}$$
$$\Pr(X \to Y, T) = \frac{\sum_{V_X \in D_X} \Pr(X \to Y, V_X)}{|D_X|} \quad (7)$$

In (7), for each distinct value V_X of $\boldsymbol{X}$, find the $\boldsymbol{Y}$-value V_Y that occurs in the maximum number of elements with value V_X for $\boldsymbol{X}$, $|V_Y, V_X|$ be the number of elements with values V_X for $\boldsymbol{X}$ and V_Y for $\boldsymbol{Y}$, $|V_X|$ be the number of elements with values V_X for $\boldsymbol{X}$, D_X is all distinct values of $\boldsymbol{X}$ in T.

All three algorithms are essentially to analyze and clean the "dirty data", but with different statistical measures during the process of "dirty data" discovery. Table II is the three algorithms feature comparison.

C. Data Cleaning Methods Based on the Discovery Algorithms

For data cleaning methods, some research institutions put forward data preprocessing, sorted neighborhood method, multiple traversal cleaning method, using domain knowledge for cleaning, using integrated data cleaning based on database. Based on the violation data statistics, this paper mainly puts forward some cleaning methods to the data sets with attribute dependency rule characteristic, as follows.

1) Statistical Method

Attribute can be viewed as a random variable; the number of its values is the same to the amount of recorded value in the field. Considering the confidence interval for this attribute, if the value of the property is not in the confidence interval, the property is wrong.

Advantage: Can randomly select sample data to analyze, accelerate the speed of detection.

Disadvantage: When the parameter model is complex, approaching the optimal parameter values need many iterations, learning is costly, low accuracy.

2) Relationship Mode Recognition

Find abnormal field not conform to existing mode based on data mining and machine learning algorithm.

Advantage: Can find the mode applied to most attribute value, if combine related data mining technology such as classification, clustering.

Disadvantage: The process is complex of the finding action, and the quality of is related to selected method.

3) Clustering Method

Data sets will be grouped into classes or clusters, with high similarity between data objects in the same cluster, while the difference of objects in different clusters is relatively large, scattered in the outside, the data that cannot merge to any class of data called "isolated point" or "singular point".

Advantage: Can detect outliers, is unsupervised mode detection, effective for many types of data, and time for comparing the data objects and clusters is short.

Disadvantage: Effectiveness is highly dependent on the clustering method, for large data sets, it is a lot of expense.

4) The Method Based on Proximity (Distance)

To quantify the structure and nodes similarity between data objects, based on the algorithms (1)-(3), can calculate the distance metric (two equal size dependency graphs), according to (4), can get the minimum metric distance, which is always the distance of a single best matching node pair, then objects from the other objects are considered abnormal point.

Advantage: The distance measurement among data objects is relatively simple, easy to calculate, other types of data can be converted into numerical data to measure.

Disadvantage: The effectiveness of the method depends on the proximity metrics. If the abnormal points close to each other, the method basically invalid.

5) The Method Based on Classification

Training a classification model that can distinguish between "normal" and "abnormal data" data, by the model, establish a classifier by describing only the normal class.

Advantage: Combined with the preference of data, i.e. the number of normal value in one attribute is much larger than the abnormal value.

Disadvantage: The obtained classification model may be too dependent on the training sample.

6) The Method Based on Dependency Rules

Define data association rules on attributes, to find the rules can give more information, the data not conforming to the rules is considered abnormal data.

Advantage: Can find the relevance of data values, and calculate the support and confidence of rules by the statistical number of data values.

Disadvantage: The strong rules is not necessarily correct rules, calculation amount is large.

Besides the 4th method, other methods can finish according to the Functional Dependencies algorithm, for all columns which have attribute dependency relation, through structured query language(SQL) or NoSql data structure storage and retrieval (such as HASH), obtain all inconsistent data to all target columns with condition attribute dependent, then based on specified rules, select the highest proportion of statistical

numerical value (get by (5)-(7)) as the reference standard to repair the error data.

IV. Conclusion

With the development of information technology, people pay more and more attention to the research and application of data cleaning in single data source, heterogeneous data source and data warehouse, this paper introduces the concept, the key step of data cleaning, cleaning framework, statistic algorithm finding wrong data and main cleaning methods briefly. However, considering the current rapid growth of mass data, the account for the proportion of semi-structured data is increasingly apparent, how to realize dynamic data cleaning, improve cleaning efficiency and guarantee cleaning accuracy is also our worthy of further study.

References

[1] Lee, M. L., Ling, T. W., Low, W. L. IntelliClean: A knowledge-based intelligent data cleaner. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000.290 -294.

[2] Galhardas, H., Florescu, D., Shasha, D., et al. AJAX: an extensible data cleaning tool. In: Chen, W.D., Naughton, J. F., Bernstein, P.A., eds. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Texas: ACM, 2000. 590.

[3] Raman, V., Hellerstein, J. Potter' swheel: an interactive data cleaning system. In: Apers, P., Atzeni, P., Ceri, S., et al, eds. Proceedings of the 27th International Conference on Very Large Data Bases. Roma: Morgan Kaufmann, 2001.381 ～ 390.

[4] Dasu T., Johnson T. Exploratory data mining and data cleaning (M). John Wiley, 2003.

[5] Ye H. Z., Wu D, Chen S. An Open Data Clean ing Framework Based on Semantic Rules for Continuous Auditing (C) In Proceedings of the 2nd International Conference on Computer Engineering and Technology, Chengdu, China. 2010: 158- 162.

[6] S. Song and L. Chen. Differential dependencies: Reasoning and discovery. ACM Trans. Database Syst., 36(3):16, 2011.

[7] D. Z. Wang, X. L. Dong, A. D. Sarma, M. J. Franklin, and A. Y. Halevy. Functional dependency generation and applications in pay-as-you-go data integration systems. In WebDB, 2009.

Yang Bao was born in ShenYang City, LiaoNing Province, China in 1978.He received the bachelor in 2001 and the master in 2004,both in computer application major from Harbin Institute of Technology, HarBin, China. Last year, he was admitted into Tsinghua University, BeiJing, China, studying for a doctor degree in Software Institute, the major is software engineering, the specific is the technology about database and data processing.

Since 2004, he is working in Beijing Institute of System Engineering, BeiJing, China as the research associate, mainly engaging in software engineering research. His papers such as Research on the Data Quality of the Large-scale Software System published on Computer Engineering & Design, 2011, Third-Party Software Testing and Evaluation published on Journal of Computer Research and Development 2009, etc. His research interests include software developing, computer software architecture research, software testing, database technology.