Research on data cleaning technology based on instance level

Chuan Li¹, Yunqi Hou², Zhang Yu³

¹Department of Computer Engineering and Science, Shanghai University, Shanghai 200072, China
²Computer Science, University of Guelph, N1G 2W1, Canada
³Accountancy, Donghua University, Shanghai 200051, China
fyszlc@shu.edu.cn

Abstract. In the era of current data explosion, data cleaning becomes an important part of data analysis, and it is also one of the important means to improve data quality. In this paper, the concept, principle, process, detection method and related cleaning algorithm of structural data cleaning are introduced in detail through the data cleaning technology based on instance level. In view of the outstanding data quality problems based on instance level, relevant experiment is designed, the operation and verification process of structural data cleaning will explain concretely through visual programming technology and machine learning algorithm. Finally, the research of data cleaning technology in the future is prospected.

1. Introduction

With the development of Internet of things, AI and data storage technology, the data of various industries is growing exponentially. However, when users use these data according to their own needs, there may exist a large amount of dirty data in the user data, or when users are extracting and integrating different source data, it may also generate some new dirty data which ultimately causes data quality problems, affecting the accuracy of user decisions and the cost of inputs. All these cases highlight the necessity and urgency of using data cleaning methods to effectively detect and repair the user data.

The so-called dirty data mainly refers to inconsistent, inaccurate, abnormal and man-made error data [1]. Data quality refers to the source degree that trusted information can be used as a prescribed application, which is to provide the correct set of data to the right people who make decisions, execute business and achieve organizational goals at the right time and in the right place [2]. Dirty data directly affects the data quality. At present, the main technologies to improve data quality include data cleaning and data integration technology. Data integration technology is used to detect and repair dirty data in mode level, such as naming conflicts, structural conflicts, etc. Data cleaning technology aims at the dirty data of the instance level; data cleaning technology is an important technical means to improve data quality.

2. Data Cleaning Technology

Data cleaning refers to the data processing process that improves data quality by detecting and eliminating dirty data from the instance level [3]. Data cleaning is essentially a means of discovering and solving problems, which is to find and locate, and analyze and correct the error data. The main application technologies include data detection technology, data analysis technology and data correction technology.

Data cleaning can be divided into "domain specific data cleaning" and "domain independent data cleaning". "Domain specific data cleaning" requires cleaning personnel to master relevant knowledge
in specific fields to understand the application scenarios of data [4]. "Domain independent data cleaning" mainly aims at general database users, especially aiming at users of relational databases. It does not need to master specific knowledge in specific fields. And it is also easy to integrate with database management system and suitable for different business fields [5]. In comparison, domain independent data cleaning technology has a wider range of application and lower requirements for participants. Therefore, this paper mainly discusses the structure-oriented data cleaning technology that based on instance level and its implementation process.

3. Realization Principle
The structural data quality problems based on instance level mainly include incomplete records, approximately duplicated records, logical errors, noise data, format errors, etc. First of all, we should analyze and understand the source data which exists different data quality problems, taking a holistic look at the data as a whole, then applying different data processing algorithms and rules to deal with dirty data, finally obtaining high-quality data that meets the needs of users, as shown in Fig.1.

4. Data Cleaning Process
The main process of data cleaning includes data pre-processing, abnormal detection, repair and data verification. The technology used in each link of data cleaning is different. In the detection link, machine learning, data mining and other algorithms can be used to detect anomalous data.

4.1. Data pre-processing
The process of data preprocessing is as follows:
(1) Collecting and organizing the user data set to be processed, carefully analyzing and understanding the data set, forming an intuitive impression in the observation process, directly deleting the data that is obviously not qualified, finally forming a result data set of preliminary screening.
(2) Importing the result data set of preliminary screening into the database to facilitate the subsequent data cleaning process.

4.2. Data anomaly detection and repair
There are many structural data quality problems. The following methods are mainly introduced in detail, including noise data, missing values, redundant data and similar duplicate records.
(1) Missing value detection and repair
The detection method of missing value is relatively simple, it only needs to detect whether the data value is empty, N/A or NULL, that is invalid value. Data missing is a common data quality problem.
We can ignore missing data which accounts for a small and insignificant proportion. It is necessary to communicate with the relevant personnel to some very important data, and the data must be retrieved. In order to solve the problem of missing values, the method of filling missing values is mainly used to supplement missing data.

(2) Noise data detection and repair
The main idea of noise data detection is to find some local outliers or global outliers, and then remove the abnormal outliers. At present, k-means clustering algorithm is commonly used for outlier selection.

(3) Redundant data detection and repair
Data redundancy refers to the characteristic that the same data appears repeatedly in the database or has no influence on the result of the data set. This paper mainly discusses the problem of data redundant characteristics. We often encounter a large amount of data being cleaned, but some characteristics will not have a direct impact on the data application results, if useless characteristics participate in the calculation, it will greatly increase the computer overhead or reduce the use value of the data itself. PCA (Principal Component Analysis) dimensionality reduction algorithm is mainly used to redundancy characteristics. The main purpose of dimensionality reduction is to transform the samples of high-dimensional space into low-dimensional space by mapping or transformation, then deleting redundant or irrelevant characteristics by characteristic selection.

(4) Similar duplicate records detection and repair.
Similar duplicate records are physical descriptions of two or more instances of representing the same entity[6]. In order to detect duplicate records, the general method is to calculate the similarity of each attribute of the instance. If the similarity exceeds a certain threshold, the two records are considered to be the same entity, and the similar record will be deleted. If the similarity is slightly below the threshold, the record will not be deleted until the user confirms it. Other data similarity detection methods include Fuzzy Match/merger algorithm, ordered adjacency point algorithm, etc.

4.3. Data validation
After the user data set is processed by each step of data cleaning, whether the result set can meet the user's business environment needs, whether it can express the attribute characteristics and correctness of the original data set, it requires the data cleaning personnel to verify the validity of the data result set. If the result set does not meet the requirements, we need further position and analyze to the result set, after repeating iterations, until a higher quality user data is obtained.

5. Experimental results and analysis
In order to verify the usefulness and applicability of the method, we use a data set of communication behaviour of wireline subscribers to analyze and verify it. The experimental data uses 7 user features, a total of 30,000 records, the data types are 64-bit floating-point type, some of the data as shown in table 1.

| Time | Package Price | Flow /Month | Phone Bill /Month | Call Duration /Month | Amount of arrears | Month of arrears |
|------|---------------|-------------|-------------------|----------------------|------------------|-----------------|
| 27.0 | 389.0         | 140.2       | 390.0             | 14.3                 | 0.00             | 0.0             |
| 29.0 | 159.0         | 0.00        | 5.00              | 0.00                 | 5.00             | 1.0             |
| 28.0 | 389.0         | 0.00        | 0.00              | 0.00                 | 0.00             | 0.0             |
| 20.0 | 389.0         | 0.00        | 0.00              | 57.09                | 0.00             | 0.0             |
| ......|               |             |                   |                      |                  |                 |
| 59.0 | 159.0         | 0.00        | 4.00              | 0.00                 | 0.00             | 0.0             |
| 23.0 | 159.0         | 60.22       | 10.78             | 16.52                | 0.00             | 0.0             |
| 25.0 | 389.0         | 226.43      | 390.00            | 619.55               | 0.00             | 0.0             |
| 23.0 | 89.0          | 165.65      | 114.86            | 16.40                | 81.00            | 1.0             |
The experiment uses Python programming language; we need build Python3.6.4 development runtime environment. We write and execute Python code through Anaconda3 (64-bit) editing tool. At the same time, Numpy and Pandas scientific computing packages, Matplotlib drawing toolkit and scikit-learner machine learning packages are imported to assist program development and verification.

1) Redundant data cleaning.
PCA (Principal Component Analysis) was used to remove redundant characteristics from telecom user data, so as to obtain the minimum data set that maximizes to express the characteristics of original data. Improving data execution speed and reducing the resource cost of computer. The data visualization diagram after dimension reduction is shown in Fig.2(a).

2) Noise data and abnormal point cleaning.
K-means clustering algorithm is used to filter the clustering data and delete the noise and abnormal points in each cluster to eliminate the influence of outliers on the data set.

(a) Dimension reduction (b) Dimension reduction clustering (c) Screening outliers and clustering Fig.2 Data cleaning process

It is seen from the results shown in Fig.2(b) that the user data set is divided into 8 categories after deleting redundant characteristics. There are many discrete points in the original data set. The data set basically tends to be regular after removing abnormal points. The data visualization diagram after screening outliers is shown in Fig.2(c).

6. Conclusion and Prospect
In this paper, the basic concept, implementation principle and cleaning process of structural data cleaning are introduced in detail in terms of data cleaning based on instance level, and the cleaning method is described in detail in combination with experiments, including redundancy detection and outlier detection, etc. In the era of big data, user data is not only growing fast, but also becoming more and more diverse. Data quality problems emerge in endlessly, so data cleaning technology plays a crucial role in the era of big data. Compared with foreign countries, data cleaning technology started late relatively in China, and various cleaning tools are relatively immature. Facing the era of big data, data cleaning technology has a long way to go.

References
[1] Fan Wenfei. Extending dependencies with conditions for data cleaning[C] 8th IEEE International Conference on Computer and Information Technology,2008:185-190.
[2] McGilvray D. Data Quality Engineering Practice[M].Diao Xingchun, Cao Jianjun, Zhang Jianmei. Beijing: Electronics Industry Press,2010:245-246.
[3] Newcombe H B, Kennedy J M, Axford S J,et al. Automatic Linkage of Vital Records[J]. Science,1959,130,(3381):954-959.
[4] Chaudhuri S, Ganjarn K, Ganti V, et al. Data Clearning in Microsoft SQL Server 2005[C]. Proceedings of the ACM SIGMOD Conference. Baltimore,MD,2005:918-920.
[5] Kalashnikov D V,Mehrotra S. Domain-Independent Data Clearning via Analysis of Entity-Relationship Graph[J]. ACM Transactions on Database Systems,2006,31(2):716-767.
[6] Garcia S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining[J]. Knowledge-Based Systems,2016,98:1-29.