Method of Chinese data cleaning based on field matching algorithm and its application to freight quality management systems

Wenyang Tan[1,a] Xiushan Jiang[2,b] Shiyi Li [*,c]

[1]School of Traffic and Transportation, Beijing Jiaotong University, China

[2]School of Traffic and Transportation, Beijing Jiaotong University, China

*College of Computing, Georgia Institute of Technology, USA

[a]16120773@bjtu.edu.cn, [b]xshjiang@bjtu.edu.cn, [c]sli606@gatech.edu

Key words: Data cleaning, data quality management

Abstract: Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality. Currently the problem is that people hold large amounts of data but do not get any useful knowledge, being described as "data rich but information poor". This paper will introduce data cleaning and study how to clean the Chinese data in a system based on the character matching algorithm and SNM method.

Introduction

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of it. Data quality problems are present in single data collections, such as files and databases, due to misspellings during data entry, missing information or other invalid data [1]. The accuracy of the business decisions and the cost of inputs depend on the quality of a large real world data set.

Currently the problem is that people hold large amounts of data but do not get any useful knowledge, in the situation of "data rich but information poor"[2]. As time goes on, caused by dirty data loss or potential threat the loss will increase further. Therefore, the study of the data cleaning is imminent.

Foreign studies on data cleaning started earlier and the technology became relatively mature after 10 years' effort [3]. While only in early21st century, studies in China started to research deeply [4]. Because of the difference between English and Chinese, research on Chinese data cleaning problem needs to be further developed. Thus, this paper will study the Chinese data cleaning method and apply the theory in reality.

Methodology of Chinese Data Cleaning

Methodology for Chinese attribute data cleaning

The kinds of abnormal data at attribute level mainly include noisy values, missing values and conflicting values [5,6,7]. The detection and processing methods of these abnormal data are shown in Tab.1.

[*] Corresponding author

Mail address: College of Computing, Georgia Institute of Technology, GA 30332 USA

Tab 1. Methods of attribute data cleaning

Attribute data cleaning	
Detecting	Manual; Statistical method; Clustering method; Method based on association rules
Processing	Ignoring tuples; Fill in vacancy by manual; Fill in vacancy by using global variable; Fill in vacancy by using probability statistic function; Binning; Combining computer and artificial

The methods mentioned are applicable to the abnormal Chinese attribute data, but the algorithm need to be improved. As Chinese is double byte encoding without obvious delimiters, and there exist a large number of homonyms, it is difficult to clean the data at the attribute level, which is also one of the key points in the study of Chinese data cleaning.

Methodology for Chinese Duplicate data cleaning

The cleaning of duplicate records is the most important problem in the process of data cleaning. The core of the problem is matching and merging [8]. Generally, field matching problem is to determine whether the two field values are syntactic substitutions that represent the same semantic entity [9]. If the two fields are semantically equivalent, that is, if they all refer to the same semantic entity, the two fields are equivalent. The most commonly used methods are:

(1) String matching method. The string matching method is able to distinguish the words existing in the dictionary, its speed is quick and accuracy is high. If A and B is the same string or one is the abbreviation of the other, the matching degree is 1. The formula can be written as:

$$\mathrm{match}(A, B) = \frac{1}{|A|}\sum_{i=1}^{|B|} \max_{j=1}^{|B|} \mathrm{match}(A_i, B_j) \quad (1)$$

(2) “Pinyin” matching method. Chinese often shows the phenomenon of homophone. In order to increase the accuracy, “Pinyin” matching method can be used as an auxiliary method to improve the matching accuracy;

(3) Similarity matching method of fields. Through using the similarity calculation formula, we can judge the similarity of fields. The edit distance method is mainly used.

Field matching is the basis for record matching. After achieving the matching results of fields, we would start matching the records. According to the weight of the field, the weighted average is calculated, and the similarity of records can be worked out. Then, we can sort the records in the database, and check whether the records are repeated by comparing the similarity of adjacent records.

Case Study

Background

In the process of building data warehouse for the Railway Bureau and analyzing its data source, we found that there are a large number of data quality problems. Accordingly, data cleaning is needed. After analyzing the data in the data source, we find the following problems:

(1) Structural problems. As the information of the owner in the source database is mainly from the Excel document, there are a large number of data redundancies in the data structure when the database is imported directly.

(2) Spelling error. Error occurred during the entry process or transfer process.

(3) Duplicate records. The same entity corresponds to multiple records. Due to the difference in format and spelling, the database management system cannot correctly identify the data.

Frame structure

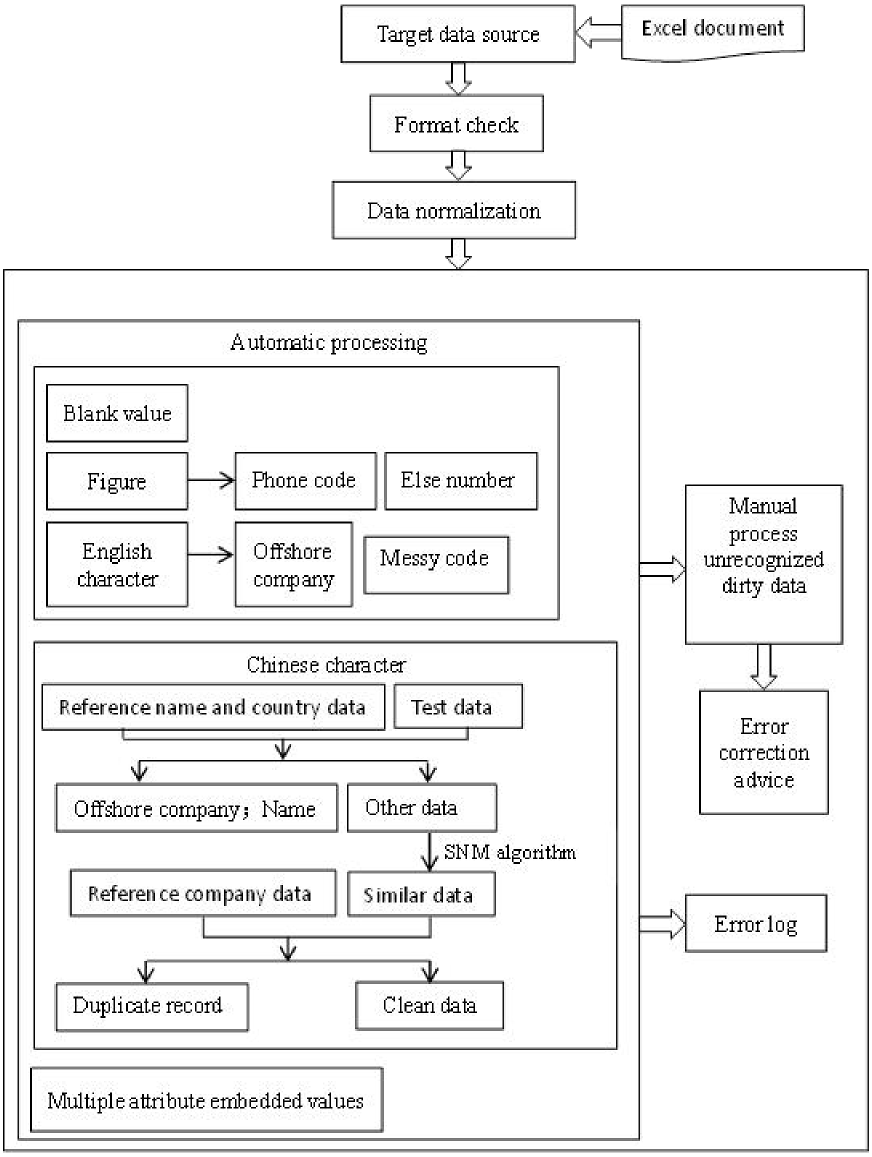

Fig.1.Data cleaning frame structure

(1) For the blank value, we replace it with “other”.

(2) For the numeric record, we need to judge whether it is a phone number.

(3) For the English record, the record within 10 characters will be regarded as messy code and changed to “other”, and the record above 30 characters will be changed to “offshore company”. For example, the record “GHHH” is a messy code, and record “HENG YANG STEEL TUBE GROUP INTL TRADING” is considered as offshore company.

(4) For the Chinese record, we need to build word segmentation dictionaries of Chinese surname and the name of common country. If the first few characters match with a record of the dictionary, we replace the record with “individual” or “overseas country”.

(5) After clean the above problems, we need to clean the duplicate records. The most reliable and simplest way to detect similar duplicate records, as shown in fig.1,is Sorted-Neighborhood Method (SNM).

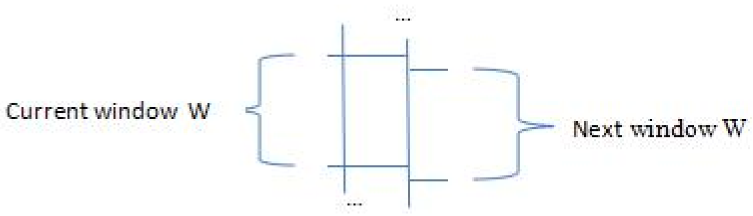

Fig.2. Schematic diagram of SNM method

In order to improve the efficiency and reduce the time complexity of the algorithm, we select an improved SNM to solve the problem. We set the size of each window Wi as:

$$W_i = \mathrm{Int}(W_i + \frac{M_{atch-num}}{W_{i-1}}(w_2 - w_1) \tag{2}$$

Where w_1 is the minimum size of window, and w_2 is the maximum size of window. $M_{atch-num}$ is the number of similar records in the window. The speed of each window movement vi can be written as:

$$v_i = \mathrm{Int}(w_{i-1} - 1 + \frac{W_{i-1}-2}{1-w_{i-1}} \times M_{atch_{num}}) \tag{3}$$

Thus, the larger the number of similar records in the window, the larger the window size, and the smaller the moving speed. If all records are similar, the movement speed is 1. We adopt the revised database data last year to establish the standard corporate name reference database.

(6) If the record contains English characters and Chinese characters, we need to process the data manually.

Results

As the freight quality management systems in 2016 contain nearly 3 million records, we choose the first 1000 data in the systems as the test data.

(1) Data dictionary

Construct a data dictionary to hold the entity's attribute information in the database. The following is an important data table structure related to data cleaning.

Tab 2 The table of customer information

Field type	Date type	Primary key	Field meaning	Remarks
ID	Int(255)	Yes	Customer number	Auto-incrementing
FHR	Varchar(255)	No	Shipper name	Important fields for data cleaning
FHDZ	Varchar(255)	No	Shipping address	
SHR	Varchar(255)	No	Consignee name	Important fields for data cleaning
SHDZ	Varchar(255)	No	Consignee address	
DATE	Int(255)	No	Date	
PLDM	Int(255)	No	Category code	
PL	Varchar(255)	No	Category	
ZL	Decimal(10)	No	Weight	

(2) Data preprocessing and correction

In this part ,we process the number, messy code, and identify the offshore company data and name data. The results are shown as Tab.3.

Tab 3. Partly dirty data example of customer information at data cleaning

Field type	Dirty data	Clean data
FHR	liu yu feng	individual
FHR	li hai peng 13221119700	individual
FHR	Heng shui lao bai gan ying xiao you xian gong si ;	Heng shui lao bai gan ying xiao you xian gong si
SHR	99 xu zhou tie lu jing ying ji tuan you xian gong si	xu zhou tie lu jing ying ji tuan you xian gong si
SHR	e luo si you xian gong si	overseas country
SHR	HENG YANG STEEL TUBE GROUP INTL TRAD	overseas country
SHR	GHHH	other

(3) Duplicate records correction

Calculate and judge the similarity of data based on the condition set. The data files are stored in the database and the similar duplicate records will be displayed on the page. The similar duplicate records will be replaced by the correct data in the case of user confirmation. Partial duplicate records are shown as Tab.4.

Table 4. Partly duplicate data example of customer information at data cleaning

Test data	Reference data	Similarity	mark
Qing hai qing hua mei hua you xian ze ren gong si	qing hai rui cheng yan ye you xian ze ren gong si	0.50	0
qing hai rui cheng yan ye you xian gong si		0.83	1
qing hai rui cheng yan ye you xian ze ren gong si		0.92	1
qing hai sheng rui cheng yan ye you xian ze ren gong si		0.92	1
qing hai rui cheng yan ye you xian ze ren gong si		1.00	

Finally, the experimental results of 1000 data are tested. The results show that 174 dirty data entries are modified, and 535 similar duplicate data entries are detected.

Conclusion

This paper studied the application of cleaning the data in the freight quality management systems. We firstly develop a cleaning rule by analyzing the data, and then propose different approaches for different issues. The character matching algorithm and an improved SNM method are used in the process. The case analysis shows that the method is feasible and effective.

References

[1] Rahm E, Do H. H. Data Cleaning: Problem and Current Approaches. IEEE Data Engineering Bulletin, 2000, 23(4): 3-13.

[2] Wang Yuefen, Zhang Chengzhi, et al. A Survey of Data Cleaning. New Technology of Library and Information Service, 2007, 12:50-56.

[3] Galhardas H, Florescu D. An Extensible Framework for Data Cleaning. In: Proceedings of the 16th IEEE International Conference on Data Engineering. 2000:312-312.

[4] Ye Ou, Zhang Jing, Li Junhuai. Survey of Chinese data cleaning, Computer Engineering and Application. 2012, 48(14):121-129.

[5] Guo Zhimao, Yu Ronghua, et al. An Extensible System for Data Cleaning. Computer Engineering. 2005, 29(3):181-183.
[6] Lee, M.L. Low, W.L. IntelliClean: A Knowledge-Based Intelligent Data Cleaner. In: Proceeding of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000, 290-294.
[7] Masek W, Paterson M A. Faster algorithm computing string edit distance. Journal of Computer System Science. 1980, 20(1):18-31.
[8] Chen Ting, Guo Ying, Liu Yunchao. Chinese Field Matching Algorithm. Computer Engineering, 2003,29(13):118-124.
[9] Guo Zhimao, Yu Ronghua, et al. An Extensible System for Data Cleaning. Computer Engineering. 2005, 29(3):181-183.