Acta Technica 62 (2017), No. 5A, 855–866

Big data quality analysis based on data mining

Hai-fei Qin[1]

Abstract. Data mining is the process of acquiring knowledge from data. The emergence of big data is very complicated in the complex data, and the data quality is drastically reduced. Based on the process of the big data, the paper divide big data into four parts,asfollows:the web and social media, the Internet of things(IOT) data, the machine data and the transactional data.And as to the four part data from different angel ananlyze,obtain data source, data type,userrequirement,data characteristics and the degree of concerns are very important in the data quality analysis. Then, as to the telecommunication data, the logging data and the micro-blogging data analyze from a professional perspective further. This paper adopts an analytical method from the outside to the inside step by step. It is a foundation for big data further analysis.

Key words. Big data, data quality,datamining,iot, machine data , transaction data.

1. Introduction

Data mining is the process of extracting implicit and potentially useful information and knowledge from massive, incomplete, noisy, fuzzy, random actual application data. The core of data mining is data, the emergence of big data, which makes traditional relational database, data warehouse and data mining technology losing its inherent advantages. Known as china's four new inventions, "online shopping", "high-speed rail", "mobile payment" and "shared bike", have improved the process of big data development. Big data is famous for its volume, speed, variety, veracity and value. The main difficulty of big data is not volume, because the expansion of computer system can meet the volume need, the real challenge of big data comes from data type variety (variety), demand for timely response (velocity) and uncertainty of data (veracity). Data is the carrier of information, when mining valuable information or applied to a field, data quality should be guaranteed as a basic demand, however, the data often has some quality problems, such as incomplete, inconsistent, redundancy and conflict, error etc., these problems not only influence the judgment of information to the people, and even directly caused economic losses [1].Data qual-

[1] Information department , Chuxiong Normal university, 675000,china

ity is an essential characteristic of data, that determines the reliability of data for making decisions in any organization or business. In practically, every business instance, project failures and cost overruns are due to fundamental misunderstanding about the data quality that is essential to the initiative[1]. The quality of data is recognized as a relevant performance issue of operating processes of decision-making activities and of inter-organizational cooperation requirements[2]. So, the analysis of big data is focused on the quality analysis of big data, and the basis of data mining is also data quality analysis. This paper will put forward its own opinions on big data quality analysis.

2. Related Work

'Data quality is the processes and technologies involved in ensuring the conformance of data values to business requirements and acceptance criteria.' 'Complete, standards based, consistent, accurate and time stamped by Glossary of data quality termspublished.' , Ballou and Pazer identified and discussed four dimensions of data quality: accuracy, completeness, concsistency, and timeliness[3]. Wang and Strong recently analyzed the various attributes of data quality from the perspective of those who use the data[4].Intrinsic, contextual, representational, and accessibility.Data quality standards the data get satisfaction in consistency, correctness, completeness and minimality[5]. GUO Zhimao,ZHOU Aoying think data quality is: data consistency, completeness, correctness, completeness and the minimality, collecting time, collecting source, the credibility of data source, timeliness etc.,A Rula , A Maurino , R Pietrobon , J Lehmann think data quality may depend on various factors such as accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, availability, and verifiability[6].

D Luebbers,UGrimmer,MJarke think most data quality measures are developed on an ad hoc basis to solve specific problems, and fundamental principles necessary for developing usable metrics in practice are lacking Companies must deal with both the subjective perceptions of the individuals involved with the data, and the objective measurements based on the data set in question. Subjective data quality assessments reflect the needs and experiences of stakeholders: the collectors, custodians, and consumers of data products[7].Giri Kumar Tayi and Donald P. Ballou, Guest Editors think "data quality" can best be defined as "fitness for use"[8],Hong Chen, David Hailey, Ning Wang and Ping Yu define the three dimensions of data quality as data, data use and data collection process. Data quality has different definitions from different perspectives. These include: "fit for use in the context of data users" [9], "timely and reliable data essential for public health core functions at all levels of government" [9], and "accurate, reliable, valid, and trusted data in integrated public health informatics networks"[9]. Data quality is commonly conceived as fitness for use, for a certain application or use case. Even datasets with quality problems might be useful for certain applications, as long as the quality is in the required range[3].Recent studies have shown that poor quality data is prevalent in large databases and on the Web. With the variety of data, often from a diversity of sources, data quality rules cannot be specified a priori.Discovering quality issues from the data itself and trading-off accuracy vs efficiency, and identifies a range of open problems for the community[10].By rapidly acquiring and analyzing big data from various sources and with various uses, researchers and decision-makers have gradually realized that this massive amount of information has benefits for understanding customer needs, improving service quality, and predicting and preventing risks[11].

The results of the study of data quality and big data quality were studied by many of the above researchers. The third part is based on the process of the big data, and the data is divided into four parts, the web and social media, the Internet of things data, the machine data, and the transactional data, and the basic analysis of the four pieces of data. The fourth part mainly from the professional point of view, the three parts of the data comprehensive analysis.

3. Big data basic analysis

Data is a complex entity, but it is not a new individual, it is produced in the process of world change. Based on the process of data generation, the paper divides the data into four parts: web and social media data, Internet of things data, machine data and transaction data. For this part from data source, data content structure, storage mode, service object, concern object, long-term concern object, data security level, purpose of data owner, purpose of data acquisition, data representation, data characteristics and typical application.

Table 1 is the basic analysis of big data, from data source, data type, storage mode, service object, concerned object, Long-term concern object, the purpose of concern, data security level, the purpose of owner , data capture lever , data representative, data characteristics and typical application.

Data Source: Web and social media data from the major web pages, microblog, WeChat, Facebook and other social media data, generated by the public, such as: public websites, social networks, etc..The IOT data from sensor, GPS, logistics information, shared bike, LBS (location-based services) etc.. Machine data comes from machine performance indicators for various exploration and inspection, mainly provided by machines, such as medical health, geological, satellite, etc.. Transaction data: generally refers to the various relational database data, which conforms to the ACID characteristics of data, such as: oralce,SQL server and so on.

Data Type: Web and social media, 90% are unstructured and semi-structured, and about 10% are structured. IOT, unstructured, semi-structured and structured are coexist,the semi-structured is dominant. Machine data: unstructured, semi-structured and structured are coexist,the structured is dominant. Transaction data ,the data is all structured.

Table 1. basic analysis of big data

data	Web and social media	IOT(Internet of things)data	Machine data
angles			
Data Source	Web and micro-blog, WeChat, Facebook etc.	sensor, logistics, location based services etc.	resource explore, medical, health data etc.
Data Type	90% are unstructured and semi-structured data, 10% structured	unstructured ,semi-structured data, and structured are coexist	unstructured ,semi-structured data, and structured are coexist
Storage Mode	document-oriented,column-oriented,Graphbased,key-value	graph based and relation database	graph based and relation database
Service Object	mass group and enterprise	enterprise	enterprise
Concerned Object	data owner, expert, scholar	enterprises, competitor	measure, analyst,competitor
Long-term Concern Object	data owner, expert, scholar	enterprises, competitor	analysts, decision makers, competitors
The purpose of concern	data support demand, technical research	data support, demand,development demand, technical research	resource search demand, health care services demand
Data security level	common	high	more higher
The purpose of owner	data support	data support	analyze, forecast and support decision making
Data capture lever	common	hard	harder
Data rep	web and micro-blog	logistics information, sharing cycle, location services	resources, healthcare, health
Data characteristics	subjective data volume, speed, variety, veracity and value	objective data, small noise, may allow error	objective data, low noise, can allow machine error
Typical applications	analysis of time and space	location services, logistics management, sharing bicycles	exploration,healthcare

Storage Mode: Web and social media, storage are document-oriented, column-oriented, graph based, key-value based big data file systems. IOT, storage are column-oriented,key-value, graph based and relational database. Machine data, graph oriented, key-value and relational database. Transaction data is oriented to relational database.

Service Object, Concerned Object, Long-term Concern Object: Web and social media, are services for the mass group and enterprise. Concerned Object are the data owner and the researchers. Long-term Concern Object are the data owner and the researchers. IOT, is services for the enterprise. Concerned Object and Long-term Concern Object are the enterprises and the competitor.Machine data, are services for enterprises. Concerned Object are the measure, the analyst,the competitor. The Long-term Concern Object is the analysts, the decision makers and the competitors. Transaction data, is services for the enterprise. Concerned Object are enterprises,competitor. Long-term Concern Object are analysts, enterprise.

Data security level, The purpose of owner, Data capture lever, Data represent: Web and social media security level is common, this data has already appeared,the purpose of owner is to provide data support better and to serve the public better. Obtain data is relatively easily. Data representative: web, micro-blog. IOT, Security level is high. The purpose of owner is to provide data support,and it support analysis, obtain data is difficult. The data represents: logistics information, sharing bike, LBS(Location Based Service), meteorological data. Machine data, security level is higher, the purpose of owner is to provide analysis, prediction, support decision-making, obtain data is difficult, data represents: resources, medical, health. Transaction data, security level is very high, the purpose of owner is to provide enterprise internal data support, support enterprise decision-making, provide personalized service,obtain data is extremely difficult.The data representation is ORACLE, SQL SERVER.

Data characteristics: Besides the characteristic of big data volume, speed, variety, veracity and value, all kinds of data have the following characteristics, Web data and social media data is subject data, has noise and some error. IOT is objective data, has noise, can allow partial error. Machine data, objective data, low noise, can allow machine error. Transactional data: both subjective and objective data are available, but preliminary cleaning has been done, almost no mistakes,consistent with transaction ACID rules,.

As can be seen from Table 1 data security at the lowest level is the most easy to access web and social media, but also the most complex web and social media data, fast growth, arbitrariness, subjectivity, objectivity, weak noise, high value and low density, it is almost all the characteristics of big data by all the people.

In summary, data characteristics and application requirements are strongly correlated with data quality. The correlation between data sources and data quality is also high, and the attention paid to groups has a great impact on data quality, such as mobile payment, transactional data and so on.Because it is relevant to people's vital interests, the higher the attention, the higher the requirement of data quality, which is not allowed to have any errors. IOT data can be both a machine to produce and need to be carried in the network, so the machine transmission error and noise are inevitable, but this part of the data is objective, as long as does not affect the overall situation, it is allowed. Web and social data credibility is lower, it saved some people's subjective data, with many colors of the individual, this quality of data is difficult to guarantee, it must be based on user requirements and concerns.

4. Big data quality Professional analysis

Data mining is based on data, and data acquisition is a troublesome and controversial process. For example, enterprise data, is the foundation of enterprise operation, is the core competitiveness of an enterprise.It is the secret in the enterprise and competitors, who want to acquire it must be loyal to the enterprise and service to the enterprises. The enterprise belongs to the management character, it wants to be fast and profit. But the development of science and technology requires a prudent learning environment and painstaking research, innovation and development ability, the two are partly contrary. At present, the vast majority of data exist in enterprises, the science and technology exist in research institutes and university, which requires the university-enterprise cooperation. But it's a rare thing to be able to work together, and there's not a lot of people in the school business that can touch the core competitiveness data. But as the experts and scholars of universities and research institutes in china, they have to find the data for analysis to ensure the leadership of science and technology and the practicability of cultivating talents for the future.So the data capture is become more and more important. Data quality research is become more and more popular.

According to big data basic analysis,we known the data source, concerned object,and user requirement become imporant in the data quality.Therefore, this paper will use three different field data to analyze the data quality.

4.1. Telecom data analysis

Telecommunication, as the leading enterprise of information enterprises, has a large number of data and advanced technologies, which is also an important component of big data analysis,and it is also representative of transactional dataThe following will analyze the telecom data.

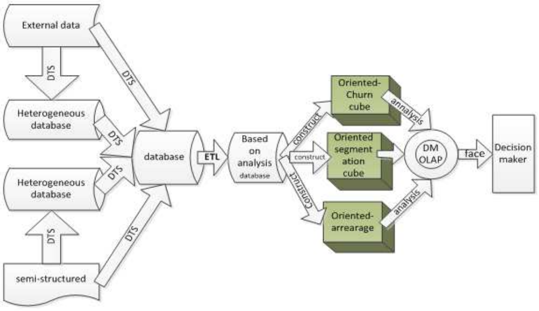

Fig. 1. Telecom data analysis system structure

Figure 1 telecom data analysis of the data mainly comes from "financial center", "quality service center system 10000 ", "big data service center", "network-centric" homogeneous or heterogeneous database and external unstructured or semi-structured data. Through DTS (data transformation service) converts the data from different data sources to the same homogeneous database for further cleansing. In the database, the data is extracted, cleaned, transformed, loaded, updated and purified, and then loaded into the data warehouse. According to the analysis requirement, the fact table and dimension table oriented-topic are constructed, and the data is aggregated. In the process of aggregation, it is necessary to verify the correctness of data step by step, build data cube, check the data and aggregation data in the data cube. The process of data from database to data cube is a process of gradual refinement and repeated verification of data quality. Through for the OLAP data cube from different angles roll-up, drill, determine the analysis of the theme of the process. Further mining of the analysis topic may require further clustering, classification, missing value completion, data normalization and other cleaning to ensure the data quality. The whole process of telecom data analysis is also the whole process of the data quality analysis.

4.2. Log data analysis

Log data is the geological exploration data, is the typical representative of machine data. The analysis of the logging data is the analysis of the machine data, and we're going to analyze the logging data.

The analysis of log data mainly includes the following six steps,as follows Fig 2 :

Step 1, Obtain the log curve from the exploration log.

Step 2, Select the logging curve according to the user needs and oriented-topic read the log data. The appropriate characteristic curve can be selected according to the known well and the data to be obtained.

Step 3, According to the visual analysis, the feature data of subject-oriented is extracted, which is based on the visual analysis of the data read. Through the probability distribution histogram of the values, it can be seen whether the data conforms to the normal distribution, the skew and the dispersion of the data. Through the box figure can quickly provide variable distribution of some of the key attributes, such as the central trend of variables, divergence of variables, the outliers and missing value.

Stept 4, The comprehensive test of the characteristic data can be used to analyze the missing values, outliers, center, center of gravity and mean value of each variable. The method of missing value processing is: the missing value is eliminated, the missing value is filled according to the correlation between the variables, the missing value is filled according to the similarity between the cases, and the tool can handle the missing value data and so on.

Step 5, According to the comprehensive test of the characteristic data, compare the situation of the center, center of gravity and mean of a variable, to see whether there is an impact of dimensionalization among the variables, and if so, the data should be standardized.

Step6, To see if there's a negative impact on the data before and after the standardized processing of the boxes, and if there's no exception to the data, it can be analyzed by clustering or classification of the data.

4.3. Microblogs data analysis

Micro-blog data is a typical representative of big data spatio-temporal data. All the characteristics of big data are reflected in micro-blog data, and the analysis of micro-blog data is the analysis of big data web and social data. The microblog data analysis process is shown in figure 3.

Micro-blog data analysis can be used before sohu data sets or data sets compiled by fudan university, or the microblog data can be obtained from the web, and the micro-blog data can be divided into text data and image data. For the data of text, firstly, the text of repetition and retweeting is eliminated, and obtained the text that is not repetition but has noise, named short text,by using the lexical library to cut the short text,obtain word set. In this case, the word set is noisy, and eliminated stop word,obtain the word set is small noise, the word set is sorted (the word weighted), the feature word are extracted, obtain the feature set, classify the feature set and the classified data set of segmentation is obtained. Form the image collection of the microblogs, the first images to be catalogued,labeled and digitized to get digitized pictures, to eliminate irrelevant information, to enhance the relevant information, to get the key words of the different images, association analysis of feature words set and feature image of segmentation,and obtain the words and images in the same theme.

The analysis of big data quality has different definitions in all walks of life, and the method of analysis is different, but the goal is to achieve the same goal, which is to meet the user's requirements. In telecommunication, the basic requirement of data quality is accurate, and the data basic requirement in logging data is the norm, standards based. In the microblog, the basic requirement of data quality is consistent,but all the data must meet people's needs,meet the customer's analysis needs and conform to the theme requirements.

5. Conclusion

Big data analysis becomes very difficult to deal with big data, such as volume, variety, velocity, high value and low density (value). The nature of the big data variety, velocity, veracity has made big data quality analysis very difficult. In order to solve this problem, the author from the perspective of fundamental analysis in big data,put forward the data quality is relate to the data datasource,data type, user requirement,data characteristics ,obtain the degreeof concern. Next,the author from the perspective of professional from three field in telecom data,logging data micro-blog data, comprehensive analysis. This work lays a foundation for big data further analysis.

References

[1] C. Batini, C. Cappiello, C. Francalanci, A. Maurino: *Methodologies for Data Quality Assessment and Improvement.* Acm Computing Surveys *41* (2009), No. 3, 1–

22.

[2] S. Thota: *Subash Thota: Big Data Quality.* DOI *10* (1985), No. 2, 257–262.

[3] D. P. Ballou, H. L. Pazer: *Modeling data and process quality in multiinput,multi-output information systems.* Management Science *31* (1985), No. 2, 171–180.

[4] R. Y. Wang, D. M. Strong: *What data quality means to data consumers.* Tamkang Journal of Science and Engineering *12* (1996), No. 4, 41–52.

[5] R. Y. Wang, H. B. Kon, S. E. Madnick: *Data quality requirements analysis and modeling, Proceedings of the 9th International Conference on Data Engineering. Vienna.* IEEE Computer Society (1993) 797–804.

[6] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann: *Quality assessment for linked open data: A survey.* Semantic Web Journal *7* (2016), No. 1, 521–528.

[7] D. Luebbers, U. Grimmer, M. Jarke: *Systematic Development of data mining-based data quality tools.* Proceedings of the 29th VLDB Conference (2003).

[8] G. K. Tayi, D. P. Ballou: *Examining data quality.* Communications of the ACM *41* (1998), No. 2.

[9] H. Chen, D. Hailey, N. Wang, P. Yu: *A Review of Data Quality Assessment Methods for Public Health Information Systems.* Int. J. Environ. Res. Public Health *11* (2014), No. 4, 587–606.

[10] B. Saha ,D. Srivastava: *Data quality:The other face of Big Data.* IEE International Conference on Data Engineering *2*, (2014), No. 9, 264–275.

[11] L. Cai, Y. Zhu: *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.* Data Science Journa *14* (2015), No. 1, 589–597.

Received November 16, 2017

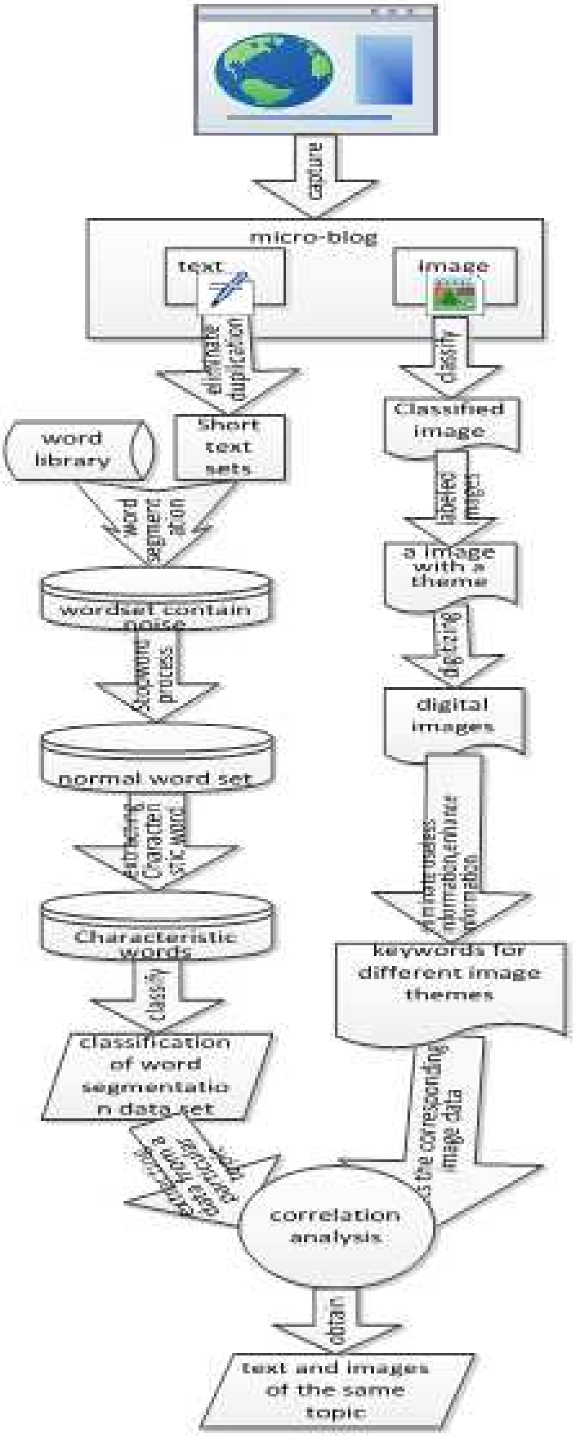

Fig. 2. The process of logging data

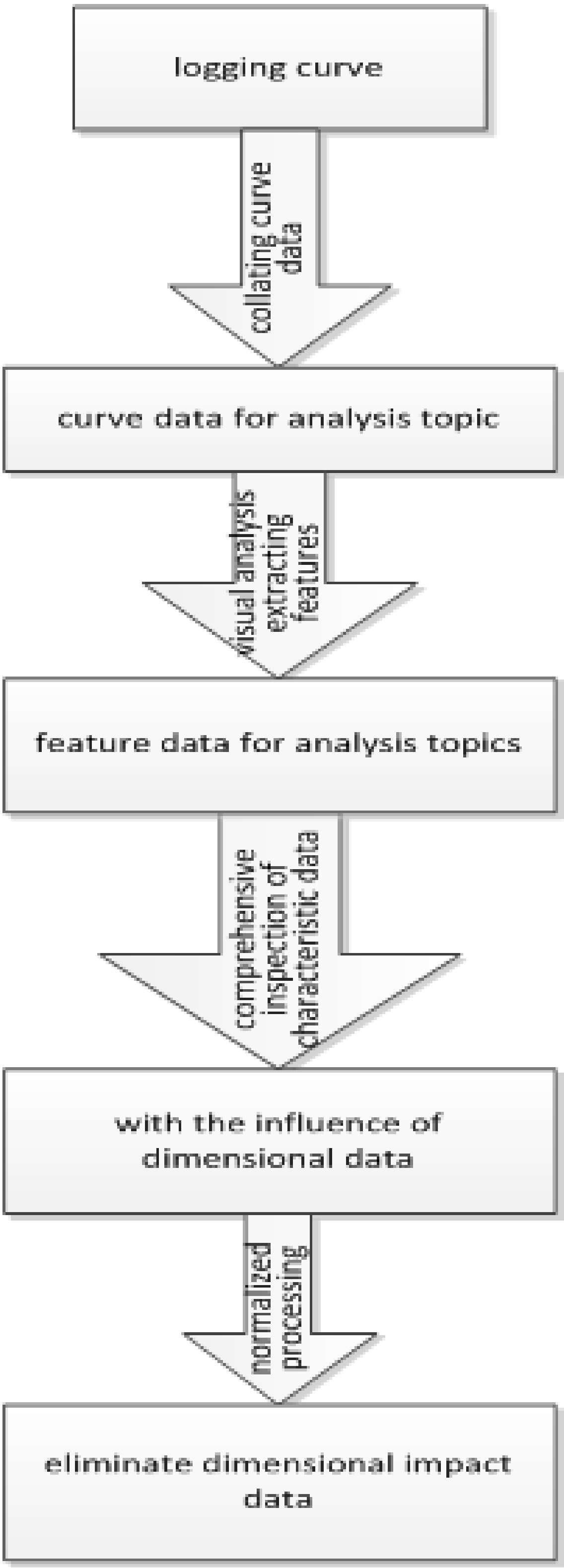
logging curve
collating curve data
curve data for analysis topic
visual analysis extracting features
feature data for analysis topics
comprehensive inspection of characteristic data
with the influence of dimensional data
normalized processing
eliminate dimensional impact data

Fig. 3. The process of micro-blog data