International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.271-284
http://dx.doi.org/10.14257/ijdta.2015.8.6.25

Data Faultage in Data Resource

Daniel Hsu[1], Shardrom Johnson[1,2] and Miao Hui[1]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200444 P. R. China
[2]Information Centre, Shanghai Municipal Education Commission, Shanghai 200003 P. R. China
zoen@shu.edu.cn, jshardrom@shmec.gov.cn, miaohui@shu.edu.cn

Abstract

Faultage is a specialized term in geology, but it can be used to describe some characteristics vividly in data resource. In this paper, we set up the preliminary theoretical system of data faultage to lay the foundation of later research and make contribution to the structure standardization of data resource. More concretely, data faultage in six areas has been enumerated firstly. Then, the conception of data faultage is presented on the theory of geological faultage, and the details of data faultage are discussed on the microscopic view. Finally, we make a verification case based on data faultage, some information from Shanghai Media Group are used to analyze the distribution of its listeners, and the theoretical system of data faultage are verified.

Keywords: *Data resource, Inhomogeneity, Data faultage*

1. Introduction

Nowadays, the scale of data resource has already presented the explosive growth tendency, which makes it hard to obtain useful information in huge data [1]. Generally, making good use of data resource will greatly improve work efficiency, and many techniques have been developed to manage data resource so far, such as data warehouse [2], a collection of data which is subject-oriented, integrated, time-varying, non-volatile, primarily used in organizational decision making [3]. Such techniques are factually effective in a way, but no matter what kind of techniques, one common and meaningful phenomenon in data resource should be paid great attention, and we call it data faultage.

Where is data, where is data faultage. For example, a road tolling system that provides large amounts of information about the movement of vehicles through toll gates [4]. Normally, people would do offline analysis of historical traffic data, but the analysis of real-time data is more important, as it can provide a warning period that enables managers to adjust the traffic management system to prevent congestion occurring. Such problems have become increasingly important and challenging to both academic researchers and industry practitioners [5]. In our life, data faultage appears frequently, such as the following.

1.1. Enterprises Informatization

In the process of enterprise informatization, some problems may cause data faultage, such as lack of historical data, full of repeated work, or inefficient management [6]. These have badly restricted the development of the enterprise.

ISSN: 2005-4270 IJDTA

1.2. Futures and Securities

China uses American futures, and data faultage appears when different trade agreements change or regulation is slack [7]. The institutional reform of Chinese security market also brings about data faultage in transaction data.

1.3. Financial Management

Whether for enterprise or government, the data of financial condition cannot be shared in different levels and information management are difficult [8]. This kind of data faultage makes it hard to get the overall situation of finance.

1.4. Census and Statistics

Data faultage in the field of statistical investigation partly results in time discontinuous, and data cannot accurately reflect the real situation. Other reason is anthropogenic factors, lacking of rigorous and meticulous work attitude [9].

1.5. Medicine Domain

One of the main tasks in traditional Chinese medicine is discovering novel paired or grouped drugs from the Chinese Medical Formula database [10]. While the database is getting bulkier, including structured information, unstructured text and images [11], then data faultage is formed that makes it difficult to get useful information.

1.6. Network Service

When visiting network, the users launch a request to transmit large amounts of data, but the network resources distribution system is unable to meet the customers' requirements [12] then data faultage is formed, and off network.

In addition to the six areas, many other situations also have data faultage, such as servers lose power, information sharing overlap in E-Government [13] and so on. The reason why data faultage happens frequently is complicated and changeable, missing or wrong data, inconsistent values, redundant data or others, all of them will lead to data faultage. No matter what kind of data faultage, their essential character is that data resource can’t meet the users’ demand, and hardly to obtain the key information.

It is inspiration that data resource is similar to geological reservoir from the view of characteristic and representation, which generally reminds us of studying geological knowledge to find out their similarity and differentia. It turns out that the faultage of geological reservoir has a good reference value to data faultage in data resource. Enlightened by the thought, some appropriate ideas and methods are introduced to the theoretical system of data faultage.

2. Domain Knowledge

2.1. Geological Reference

In geology, the reservoir is affected by sedimentary environment, digenesis and tectonic in the process of forming, in spatial distribution and internal various attributes, there also have some uneven changes, that is called the reservoir inhomogeneity. Faultage is obvious displacement structure caused by fracture of rock formation or rock mass, and it happens because of the reservoir inhomogeneity [14]. Faultage is extensively developed in nature, and it is one of the most important geological structure types in the lithosphere. Big faultage comprises a regional geologic trellis, and it not only controls the regional geological structure and evolution, but also controls and influences region mineralization [15]. Some medium and small-scale faultage often directly determine the shape of some deposits and ore bodies. Active faultage directly affects hydraulic structure and even causes earthquakes. Therefore, the research of faultage has theoretical importance and practical significance.

How faultage is formed? To explain it, firstly start with crack. Crack is an interface which has loss of adhesion. If the materials on two sides of crack are displaced, it is called faultage. So the formation principle of faultage is in close attach with the forming reason of crack. Cracks are produced under the function of tectonic stress, and its block diagram is given in Figure 1.

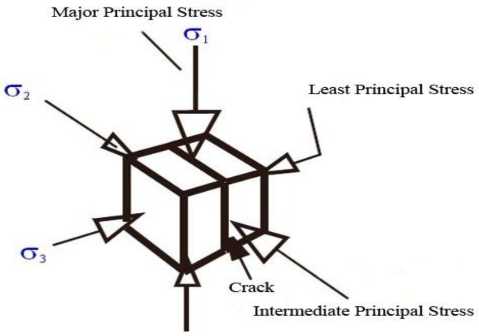


Figure 1. Block Diagram of Crack

From this diagram we can see, σ1 is major principal stress, σ2 is intermediate stress, and σ3 is the least principal stress. Because of the various stress from three directions, and the effect from the characteristics of object itself at the same time, when the difference of various stress is large enough, object cannot bear any more, then crack emerges. If continue to apply force, crack will grow into faultage. Geological faultage in reality can’t keep a well-organized state, they are usually strange rocks with various styles and shapes, and that is exactly the difficulties and the theories meaning of this dissertation.

2.2. Data Resource

From the point of resource, data resource is the reacquainting and highly generalization of data and its own state [16]. Any effective methods of data resource must be based on actual analysis and comprehensive understanding. Different kinds of data resource in reality are benefit from the popularity of computer technology, database technology, and database management system [17]. Data resource itself contains a large amount of data collection objects, and each collection has some data objects with various themes or structure.

As we have known that the properties of data are variable with time, and there must be faultage between the historical data and instant data, thus the analysts are always lacking of the most critical data. Especially under the circumstance of the pace of business is speeding up, this time lag could result in missing good business opportunities. Time behavior makes data inconformity, which will lead to the inhomogeneity in data resource. Inhomogeneity is an inevitable attribute of data resource that will result of data faultage. It is essential to know the phenomenon before handling them, so we should figure out the basic characteristics of data resource and data faultage at first.

Data resource is similar to geological reservoir, as they both have inhomogeneity and faultage structure. There are many kinds of classification on faultage in geologic terms, for example, according to the relationship of relative displacement between hanging wall and footwall, faultage structure can be divided into three types, namely normal faultage, thrust faultage and strike-slip faultage [18]. While in data resource, faultage could be classified with several types by the relationship of data, but the data relationship is so complex that making the classification confused. In practice, ambiguous things are unlikely to apply because of their low accuracy. Therefore, the system of data faultage in this paper is a good way to describe the structural feature of data resource. Before analyzing, it is necessary to make a list of comparison between data resource and geology reservoir, trying to figure out where they are alike and what's the difference.

Table 1. Comparison of Data Resource and Reservoir

Comparison	Reservoir	Data Resource
Constituent Element	*Different types of rocks*	*Different themes of data*
Element Form	*Rocks are in continuous state*	*Data are discrete individuals*
Property Variation	*Unpredictable and stable*	*Changeable and unstable*
Inhomogeneity Definition	*Properties vary with its position*	*Properties vary with time*
Inhomogeneity Consequence	*Cracks and faultage are formed in rocks*	*Data faultage is generated in data resource*
Faultage Application	*oil exploration, disasters forecasting, etc.*	*data mining, data resource structural optimization, etc.*

Table 1 shows several main factors that compared between reservoir and data resource. The basic element of reservoir is rock while data resource is composed of data. Rock is continuous as a whole, and data are discrete values. There absolutely exists some difference in many aspects like property variation and others, but common grounds are so valuable that we should pay much attention. One of the important resemblances is that reservoir and data resource both have inhomogeneity, which may cause faultage in rock and data. Faultage is materialization in geology, and we can study it according to its appearance. While faultage in data resource is suppositional, that we should describe it with the support from various methods like mathematical theory or physical techniques.

3. Definition of Data Faultage

3.1. Basic Concepts

Regardless of the existence of the various information security safeguards [19], many enterprises also remain confused to data management, especially when data are of low efficiency. So it is necessary to make sense of the characteristics of data resources, and some concepts about data faultage would be introduced in the following.

Definition 1(Inhomogeneity): Inhomogeneity is an attribute that the different properties of data resource vary with time. Inhomogeneity exists between each data set is called macroscopic inhomogeneity. Each data set contains a large collection of data objects, inhomogeneity exists among data objects is called microscopic inhomogeneity.

In reality, inhomogeneity among some systems makes it hard to share information, and this caused some vulnerability management [20]. Therefore, inhomogeneity is a significant feature of data resource, and it may lead to data faultage.

Definition 2(Data Faultage): Data resource has some certain limits because of their attributes, themes, structure or others. As the amount of data is rising constantly, these limitations lead to obvious differences emerge among data objects. Data can no longer been viewed as a unified whole, and that is called data faultage.

Definition 3(Macroscopic Data Faultage): Macroscopic data faultage exists between data collections, and affected by subject, structure and other factors. Macroscopic data faultage is extensive, and we will not discuss it in this paper.

Definition 4(Microscopic Data Faultage): Microscopic data faultage exists inside a data collection, and affected by structure, element, data relationship, *etc.* The main work of this paper is to describe the microscopic data faultage.

Definition 5(Dominant Data Faultage): In data collections, some data are valueless, such as empty data, repetitive data and so on. These data are disturbing to data processing, and cause dominant data faultage in data resource.

Definition 6(Tacit Data Faultage): Tacit data faultage is caused by irrelevant or illogical data, which cannot be found by visual inspection, but detailed Analysis.

We can understand these concepts through the block diagram of data faultage in Figure 2.

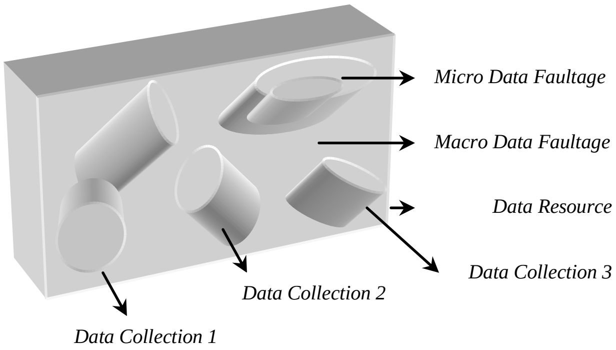


Figure 2. Block Diagram of Data Faultage

Each of these data collections has its local temporal dataset along with spatial data and the geographical coordinates of a given object or target [21]. They have different capacity, shapes, location and so on, which result in the inhomogeneity of data resource. Some collections may have part of the same data, and they may be integrated into one collection, so macroscopic data faultage is changed into microscopic data faultage.

3.2. The Properties of Data Faultage

There is a specialized geological term called faultage effect [22], which means mistakes in seeing of rock strata caused by the activities of faultage. We can learn from faultage effect that the study of faultage should be in numerous ways, and data faultage has different characteristics when considered from various angles. The following three points are summarized:

One is inevitability. Data faultage is inevitable as data resource is inhomogeneous.

Two is variability. The status of data faultage varies with data environment.

Three is inhomogeneity. Data faultage is unevenly distributed in data resource.

The properties of data faultage determine its configuration feature, and that is one of the important parts should be considered. In addition, we need further research and investigation on its characteristics.

3.3. The Influences of Data Faultage

Each kind of phenomenon certainly will be affected by different factors, so is data faultage. This paper presents three influencing factors about data faultage:

Subject oriented. Data are collected according to their subjects, and different subjects will generate different features and data faultage.

Time behavior. Data are updated every now and then, so data faultage is changing with time behavior.

Data capacity. The quantity of data also affects the probability of data faultage. In general, the scale of data faultage is correlated positively with data capacity.

The actual influences are more than that, and general factors should not be considered in different conditions.

4. Microscopic Data Faultage

Microscopic data faultage exists inside data collections. In Figure 3, the processing steps are introduced to make good sense of data faultage.

There are empty data, invalid values or some data that are related to user subject but valueless, and we call them pores. The ratio of pore number to the total data which have something to do with subject is called porosity. The pores will reduce the utilization rate of data resource, so prior to the next procedure, the first thing is to make sure whether the porosity is zero or not. A simple way is to see if there are empty data or invalid values in database. If at least one is found, then go on data compression, which mainly deal with dominant data faultage, otherwise, jump out of the processing. The next step is faultage detection, to find out the position and status of tacit data faultage. As in geology, faultage has its own two sides, we can make use of it on oil exploitation, but it also may lead to earthquakes. Similarly, data faultage has advantages and disadvantages. Some data faultage can be used to obtain special and important information, while some are disturbing to analysts. The latter should be processed, so the next step is to do pressure soluble on data faultage. The three main processing steps consist of the whole analysis procedure of microscopic data faultage. In the following part, the three steps will be introduced in detail.

4.1. Data Compaction

Definition 7(Data Compaction): Data compaction is to process data on the basis of databases, in order to have no effect on final evaluation.

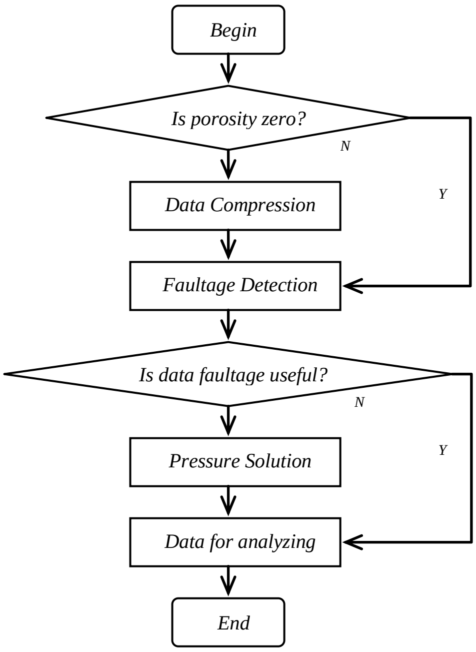


Figure 3. Flow Chart of Microscopic Data Faultage

The operations mainly include: deal with empty data, transform the data with incongruous format, and remove the invalidated data and so on. Data compaction is prepared for faultage detection. The purpose of it is to handle dominant data faultage, minimize disturbance to the results of the faultage data judgment. In order to understand data compaction better, the model of data compaction is given below.

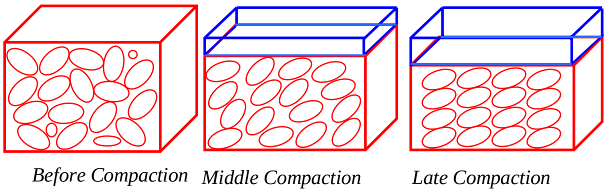


Figure 4. The Model of Data Compaction

It can be seen from Figure 4, before compaction, a bewildering welter of data filled in database. In the process of compaction, data become in better order, inconsistent values are eliminated. At the end of compaction, database arrives in perfect physical conditions, small space usage, and high compactness.

4.2. Faultage Detection

Definition 8(Faultage Detection): Faultage detection is to detect the distribution of tacit data faultage, and determine its status.

After the dominant data faultage is eliminated, tacit data faultage become the biggest problem. In order to do detection better, several concepts should be introduced firstly.

In databases, data are stored in different ways according to their attributes. If there is large amount of data, they are divided into some data collections. Work out the following values of each collection, and then do the judgment.

Definition 9(Interval Data Density): If one data collection in database is expressed as $\{a_1,a_2,\cdots,a_n\}(n>0)$, within a range of $[a_s,a_t],(s<t)$, and $a_{\max},a_{\min}$ are maximums and minimums, and $a_{\max} \neq a_{\min}$.

$$d=\frac{\sum_{i=s}^{t}a_i}{a_{\max}-a_{\min}} \tag{1}$$

Then d is called Interval Data Density in range of $[a_s,a_t]$. Data density is defined to describe the data distribution. If d is bigger, data in this interval are more unevenly distributed. The number of intervals is determined by users, the more intervals there are, the more precise the result will be.

Definition 10(Average Data Density): If one data database has data collection $\{a_1,a_2,\cdots,a_n\}(n>0)$, $a_{\max}{}',a_{\min}{}'$ are maximums and minimums on interval that is specified, and $a_{\max}{}' \neq a_{\min}{}'$.

$$d_e=\frac{\sum_{i=1}^{n}a_i}{a_{\max}{}'-a_{\min}{}'} \tag{2}$$

d_e is the average data density of this data collection. The total data density is $c\times d_e$, and c is the number of intervals.

Definition 11(Density Probability): Density Probability is the ratio of interval data density to total data density.

$$p=\frac{d}{c\times d_e}\times 100\%(0<p<1) \tag{3}$$

Shannon entropy and Fisher information functional are known to quantify certain information-theoretic properties of continuous probability distributions of various origins [23]. Here we introduce information entropy to describe the quantitative measure of information.

Definition 12(Information Entropy): It is the average probability of measuring whether the data faultage is formed or not. It is determined by data density and threshold specified by users.

The variable of interval is expressed as X. Density probability distribution of intervals is represented as follows:

$$\begin{bmatrix} X \\ p(x) \end{bmatrix} = \begin{bmatrix} x_1, x_2, \cdots, x_n \\ p_1, p_2, \cdots, p_n \end{bmatrix} (0 < p_i < 1, \sum_{i=1}^{n} p_i = 1) \tag{4}$$

The information entropy is expressed as:

$$Hn = -\sum_{i=1}^{n} p_i \ln p_i \tag{5}$$

If $p_1 = p_2 = \cdots = p_n$, namely all the intervals have the same data density, then the information entropy gets the maximum value $H_n = \ln n$.

Supposing that data density of arbitrarily siding-to-siding block equals to average data density, and then calculates the information entropy of this dimension, that is called theoretical entropy, expressed as H_t. As the inhomogeneity of data resource, invalided values, and the different density on various ranges are not able to be eradicated, making the actual entropy is different from theoretical entropy, so we call it entropy anomaly, expressed as H_e. If actual entropy is higher than theoretical entropy, we call it positive entropy anomaly, and in the opposite case, it is named as negative entropy anomaly. Take theoretical entropy H_t as zero point, H_e >0 is positive entropy anomaly, H_e <0 is negative entropy anomaly.

After we are acquainted with these concepts, three steps are designed to detect data faultage.

1. Figure out the data density and Information Entropy of data according to the formula given above.

2. Confirm the threshold condition. Here the threshold is specified by users, and it is established by some similar data collections, and average information entropy which is from the calculation of average data density.

3. H_e is valid entropy anomaly. If H_e is within the threshold condition, data are high-quality that meets the demand. Otherwise, tacit data faultage is found in the data collection.

Other data collections should be handled in the same way. Wind-down the quantity of collections and loop operation, the tacit data faultage will be found.

This method makes use of mathematical formula and information theory. The critical process is to calculate the actual entropy and threshold correctly, then compare the two values. We can work out the status of faultage data preliminarily by this method.

4.3. Data Pressure Solution

Definition 13(Data Pressure Solution): The process of treating faultage data with pressure solution algorithm and obtaining useful information is called data pressure solution.

The faultage data are neither good nor bad. According to the needs of users, faultage data can be eliminated, at the same time, they can be used to discover some special information. Here we introduce a method to eliminate faultage data, namely data pressure solution.

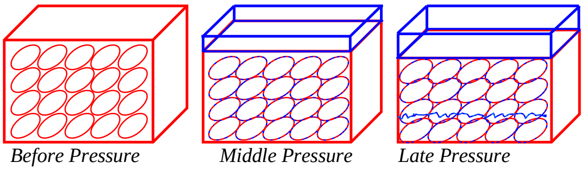


Figure 5. The Model of Data Pressure Solution

As can be seen from Figure 5, boundary data that have gone through first step of pressure solution are partly merged, the storing amount of the data is narrowed down. When coming to pressure solution on a grand scale, the faultage data disappear, and data with same subject flock together, meanwhile, the data space utilization reaches a maximum.

At present, the specific steps of data pressure solution are explored only in a preliminary way. This step still has many problems to deal with, for example, the model is monotonously to some extent, what kind of algorithm is suitable for processing data, and some useful data may be disposed in this process, *etc.*

5. A Verifying Case

Broadcasting station is one of the important forms of public services, and it plays an indispensable role in our daily life. In this case, we take a famous radio station in Shanghai as an example. There are amounts of people listen to the radio station through various means, and one of the ways is an application on mobile terminals such as iphone or ipad, and mass data about IP addresses are collected. To analyze the IP addresses will help us know the listeners’ distribution well, and make appropriate adjustments according to the distribution.

In analytical process, there are also data faultage existing in data. IP addresses distribution is diverse every day, especially fluctuate in holidays, and this kind of data faultage is in macroscopic view. In one day, there are thousands of listeners in one area, while a few listeners in other areas, and their listening time is unpredictable, such kind of phenomenon is inevitable data faultage. In the following part, we will analyze the data in September 16, 2011 according to the method that this paper has proposed. Processing steps are as follows:

Step One: Get IP addresses from load balancing servers.

Remote login in three streaming servers to get log files. Pick off the streaming media access log on September 16, 2011.

Step Two: Delete duplicate and invalid IP addresses.

Firstly, import log data into excel. Make use of macro script to filter data and obtain IP addresses. Finally, delete duplicate and invalid data by the means of excel.

Step Three: Query IP addresses attribution.

Query these IP addresses in particular program that we have developed, and it has been shown in Figure 6. Get their attributions and save the query result.

Step Four: Analyzing and processing data faultage.

Divide the IP addresses into collections by their attributions. Then statistics data, detect data faultage in each collection according to the method mentioned above.

Step Five: Decide the approach of data faultage.

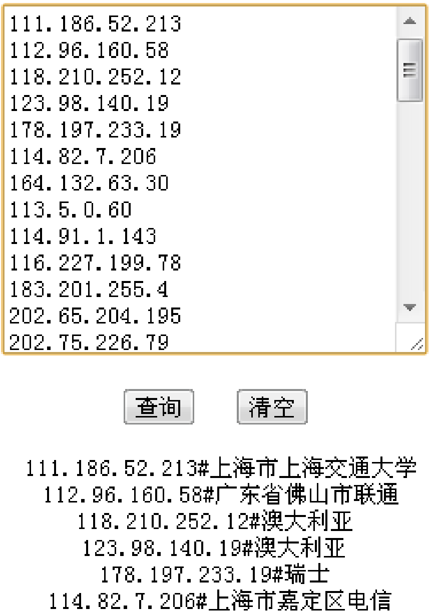
If data faultage should do pressure solution, then go on it in accordance with the actual situation. Otherwise, take advantage of the data faultage result for decision-making directly.

From the five steps, data are intuitive and clear to us. The number of IP addresses that we have query is 3437, with 10 invalid data, and the IP number of Shanghai has reaches

2110. As it is a local radio of Shanghai, so the data faultage between Shanghai and other area is obvious. Therefore, Shanghai can be viewed as a separate area for further analysis. Other data cover two parts from the view of geographical position, the listeners in China and foreign countries. Listeners from foreign countries are widely scattered, so we won't talk about it in this case.

Figure 6 shows the interface of query program. The query result can be saved as excel file for further analysis.

In Figure 7, East area includes Jiangsu and Zhejiang, which are the neighboring cities of Shanghai, so the number of east listeners is higher than other cities. Listeners in West North area are the least, as they are far from Shanghai. From the analysis we know that one influencing factor of data faultage is the distance from Shanghai.



```
111.186.52.213
112.96.160.58
118.210.252.12
123.98.140.19
178.197.233.19
114.82.7.206
164.132.63.30
113.5.0.60
114.91.1.143
116.227.199.78
183.201.255.4
202.65.204.195
202.75.226.79
```

```
111.186.52.213#上海市上海交通大学
112.96.160.58#广东省佛山市联通
118.210.252.12#澳大利亚
123.98.140.19#澳大利亚
178.197.233.19#瑞士
114.82.7.206#上海市嘉定区电信
```

Figure 6. The Program of Query Attribution

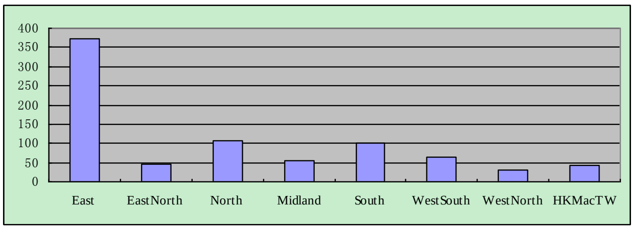


Figure 7. The Number of Listeners In China

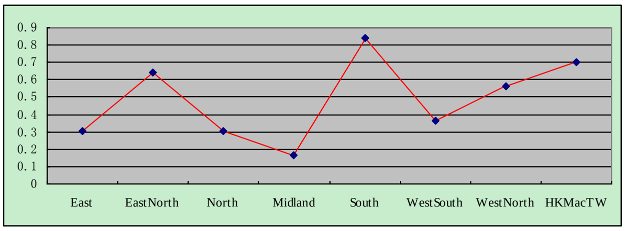


Figure 8. Interval Data Density of Listeners' Distribution

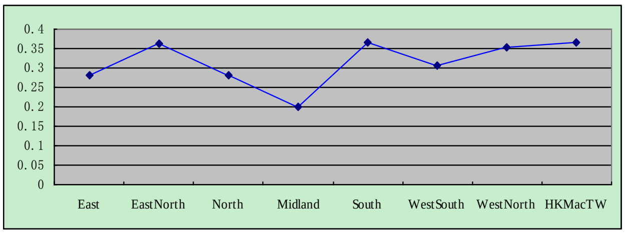


Figure 9. Information Entropy of Each Area

In Figure 8, Interval data density of each area has been shown. If the value is high, indicating that the distribution of listeners in this area vary widely, and data faultage would be generated to a great extent in this area. Then, calculate the density probability of each area according to data density.

In Figure 9, information entropy of each area has been given. The information entropy should be kept in a limited range which is determined by threshold condition, and it is decided by user according to actual demand. In this case, suppose that the threshold range is 0.25~0.35, then four areas do not meet the requirement. Therefore, the residual analysis and processing procedure of data faultage will mainly about these four areas.

This verifying case shows the preliminary faultage characteristics of data resources. Obviously, data faultage not only exists in the above condition, but also in other aspects. For example, the listening period vary with lifestyle regularity of listeners, and there must be data faultage between busy hours and idle hours. Therefore, to find the position of data faultage is helpful to reasonably arrange the sequence and duration of program.

6. Conclusion

As a phenomenon that cannot be neglected in data resource, data faultage has significant influence on data analyzing and resource utilization [24]. It is important to understand the data faultage correctly, to make a theoretical research on data faultage, and to integrate theory with practice. This kind of research is of great value for the development of data resource [25]. Basing on the thought, we propose a series of definitions and methods for data faultage, which are of important implication in the future studies in this regard [26].

In the follow-up research, there are a lot of problems to be resolved [27-28]. For instance, how does the data faultage between microscopic and macroscopic transform into each other, how to describe irregular data faultage, *etc.* Much still remains to be done, but still waters run deep, the exploration of data faultage will never cease.

Acknowledgements

This research was financially supported by the National Educational System Reform Project (01-109-091), by the National Natural Science Foundation of China (Grant NO. 40976108 and 61303097), by Research Funds of Shanghai Leading Academic Discipline Project (J50103), and by the Graduate Innovation Fund Project of Shanghai University (Grant NO. SHUCX070037 and SHUCX120105). Thanks for the help.

References

[1] X. Jiaoxiong, “Clustering Preprocessing of Data Resource”, Shanghai Popular Science Press, Shanghai, **(2011)**.

[2] J. Han and M. Kamber, “Data Mining: Concepts Techniques, Morgan Kaufmann Publishers”, San Francisco, **(2001)**.

[3] W. H. Inmon, “Building the Data Warehouse”, Wiley Publishing, Inc., Indianapolis, **(2005)**.

[4] C. Harrison and I. Abbott Donnelly, "A Theory of Smart Cities", Proceedings of the 55th Annual Meeting of the International Society for the Systems Sciences, University of Hull Business School, UK, July 17-22, **(2011)**.

[5] S. Yilun, "Optimal Attack Strategies in a Dynamic Botnet Defense Model", Applied Mathematics & Information Sciences, vol. 6, no. 1, **(2012)**, pp. 29-33.

[6] X. Huizhen, "The Resources Integration in Enterprise's Informatization", Sci./Tech Information Development & Economy, vol. 21, **(2011)**, pp. 135-137.

[7] W. Lanjun, "The United States Financial Regulatory System and its Reform", Study and Research, vol. 4, **(2011)**, pp. 75-78.

[8] J. Jingbo, "Financial Management Information Problems in Implementation and Measures", Theoretic Observation, vol. 5, **(2007)**, pp. 97-99.

[9] X. Jianying, "The quality of statistical data of ascension", Statistics and Management, vol. 6, no. 15, **(2011)**.

[10] Z. Zhongmei, "Efficiently Mining Positive Correlation Rules", Applied Mathematics & Information Sciences, vol. 5, no. 2, **(2011)**, pp. 39-44.

[11] M. Xikun and Y. Jingjie, "Discussion on the Integration and Construction of Electronic Medical Records System", Information of Medical Equipment, vol. 27, **(2012)**, pp. 59-60.

[12] G. Mingmin, "Analysis of Network Security and Strategy from the Campus Network", Zhongnan Tribune, vol. 6, **(2011)**, pp. 97-98.

[13] L. Xin and G. Feng, "Privacy Protection Method in E-government Information Resource Sharing", Journal of Computer Applications, vol. 32, **(2012)**, pp. 82-85.

[14] L. Minggao, "Quantitative Reservoir Geology, Geological Publishing House", Beijing, **(2011)**.

[15] S. Chunqing, "Basis of Geology", Higher Education Press, Beijing, **(2005)**.

[16] S. Jiulin, "Scientific data resources and sharing", China Basic Science, vol. 1, **(2006)**, pp. 30-33.

[17] A. Kandel, M. Last and H. Bunke, "Data Mining and Computational Intelligence", Physica-Verlag, Germany, **(2010)**.

[18] W. Shenghe and X. Qihua, "Hydrocarbon Reservoir Geology", Publishing House of Oil Industry, Beijing, **(1998)**.

[19] H. K. Kim, S. K. Kim and S. H. Kim, "Decision Support System for Zero-day Attack Response", Applied Mathematics & Information Sciences, vol. 6, no. 1, **(2012)**, pp. 221-241.

[20] T. Wenya, "Design and Implementation of Web-Based Food Regulatory Information Resources Management Platform", Applied Mathematics & Information Sciences, vol. 5, no. 2, **(2011)**, pp. 105-111.

[21] A. M. Khedr and W. Osamy, "Target Tracking Mechanism for Cluster Based Sensor Networks", Applied Mathematics & Information Sciences, vol. 1, no. 3, **(2007)**, pp. 287-303.

[22] Z. Zuoxun, "Structural Geology (Third Edition)", China University of Geosciences Press, Wuhan, **(2008)**.

[23] P. Garbaczewski, "Information Dynamics in Quantum Theory", Applied Mathematics & Information Sciences, vol. 1, no. 1, **(2007)**, pp. 1-12.

[24] X. Jiaoxiong, L. Mengfang, B. Minjie and X. Jun, "Digitizing Strategy on the Same Ontology in Heterogeneous Data Source", Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications, Westin Resort, Macau, October 24-26, **(2011)**.

[25] S. Johnson, "Optimization Strategy of Energy-Description and Collision-Description Clustering", Journal of Information and Computational Science, vol. 8, no. 8, **(2011)**, pp. 1251-1260.

[26] X. Jiaoxiong, W. J. Y. Chenqiong and X. Jun, "Research on Data Faultage Phenomena", Computer Applications and Software, vol. 30, no. 8, **(2013)**, pp. 9-13,77.

[27] X, Jiaoxiong, X. Jun and W. Gengfeng, "Cluster Pre-processing on Ontic Kernel and Histogram", Journal of Shanghai University (Natural Science), vol. 14, no. 1, **(2008)**, pp. 19-25.

[28] B. Minjie, X. Jiaoxiong and X. Jun, "Database Preprocessing with AHP", Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, August 10-12, **(2010)**.

Authors

Daniel Hsu (Xu Jun), He obtained the master degree in Software Engineering from Shanghai University in 2006. He is currently a doctoral fellow in Shanghai University. His research interests are in the areas of artificial intelligence and data mining.

Shardrom Johnson (Xia Jiaoxiong), He obtained his PhD degree from Shanghai University in 2007. He is employed in the Information Centre of Shanghai Municipal Education Commission and he is also an associate professor in Shanghai University. His research interests are in the areas of data mining, intelligent decision support system and educational informatization. His first monograph is Clustering Preprocessing of Data Resource, which has made a significant contribution to the research of data resource. And, his second monograph is “I and Shanghai Educational Informatization of the 12th Five-year”, which has made a significant contribution on Shanghai educational informatization.

Miao Hui, He is a graduate student in Shanghai University, and he majors in software engineering. His research direction is the structural optimization of data resources, and the current research emphasis is data faultage.