Article

Processing on Structural Data Faultage in Data Fusion

Fan Chen 1, Ruoqi Hu 2, Jiaoxiong Xia 1,2,3,* and Jie Tao 2

1 School of Computer Engineering and Science, Shanghai University, Shangda Road 99, Shanghai 200444, China
2 XianDa College of Economics and Humanities, Shanghai International Studies University, East Tiyuhui Road 390, Shanghai 200083, China
3 Information Centre, Shanghai Municipal Education Commission, Dagu Road 100, Shanghai 200003, China
* Correspondence: jshardrom@shec.edu.cn

Received: 30 December 2019; Accepted: 3 March 2020; Published: 6 March 2020

Abstract: With the rapid development of information technology, the development of information management system leads to the generation of heterogeneous data. The process of data fusion will inevitably lead to such problems as missing data, data conflict, data inconsistency and so on. We provide a new perspective that combines the theory in geology to conclude such kind of data errors as structural data faultage. Structural data faultages after data integration often lead to inconsistent data resources and inaccurate data information. In order to solve such problems, this article starts from the attributes of data. We come up with a new solution to process structural data faultages based on attribute similarity. We use the relation of similarity to define three new operations: Attribute cementation, Attribute addition, and Isomorphous homonuclear. Isomorphous homonuclear uses digraph to combine attributes. These three operations are mainly used to handle multiple data errors caused by data faultages, so that the redundancy of data can be reduced, and the consistency of data after integration can be ensured. Finally, it can eliminate the structural data faultage in data fusion. The experiment uses the data of doctoral dissertation in Shanghai University. Three types of dissertation data tables are fused. In addition, the structural data faultages after fusion are processed by the new method proposed by us. Through the statistical analysis of the experiment results and compare with the existing algorithm, we verify the validity and accuracy of this method to process structural data faultages.

Keywords: data fusion; structural data faultage; isomorphous homonuclear; information entropy; data structure integrity

1. Introduction

In recent years, as information technology and distributed systems develop, heterogeneous data has become more popular. At the same time, groups of heterogeneous data remain independent from each other. Take the data of science and technology resource management integration as an example. The data of science and technology resource management includes EXCEL, XML and other text data, database data. The amounts of data are large. There is a big difference among various types of data in their structures and ways of storage. However, changes to the business require data to be quickly integrated, adapted to the business, and easily accessible [1]. Many organizations face the problem of integrating data from multiple sources. Because of the features of heterogeneous data, the process of data integration is filled with problems like data redundancy, invalidity, repetition, missing, inconsistency, value error, format error and so on [2]. Those problems cause data faultage, then affect the quality of data, and finally bring inconvenience to users. Therefore, how to ensure the
heterogeneous data’s integrity and consistency after data integration has already become an issue that is urgently needed to be studied and solved [3].

The main characteristics of big data include volume, velocity, variety, and veracity [4]. In the process of big data, the veracity of data is very important. Now, data integration has appeared widely in many fields. Additionally, data quality after data integration has become one of the most concerned problems. The data after integrating has different data quality, which lead to the spread of uncertain or inaccurate data in social media. Data quality is one of the major challenges in data integration.

Data integration is the process of retrieving data from different sources and combining them into meaningful and valuable information. Traditional data integration is divided into three steps: schema mapping, record linkage and data fusion [4]. In the last step of data fusion, it is easy to cause inaccurate data, so it is crucial to obtain high-quality data.

Now there is repetitive information among many data tables. It not only takes up storage space, but also brings much inconvenience to the storage of data. For example, the table of master degree in university is divided into three data tables according to the sorts of master degree (academic degree, professional degree, and part-time degree). However, most information in the three data tables is repetitive, which all included basic things like names, genders, schools, etc. The fusion of the three data tables can enlarge the space for data resource. That said, the data after fusion will definitely lead to the data error in data’s structure because of differences among the structure of the tables. Structural errors caused by data fusion bring great inconvenience to the use of data, make data lose "authenticity", and bring great inconvenience to the subsequent process of data.

Guided by the theory of geological faultage, this paper applied the data error caused by heterogeneous data in data fusion to the study process about the structural data faultage theory. Aiming at structural data faultage presenting in the process of data fusion, the paper use the similarity between the name and content of attributes to eliminate structural data faultage, protect data structure’s integrity, and increase resource space.

The rest of the paper is organized as follows: Section 2 introduces the background of data fusion and structural data faultage. The details of structural data faultage processing algorithm based on attribute similarity are given in Section 3. Section 4 takes three doctoral dissertation awarding information tables of Shanghai university as data sets to conduct experiments and verify the usefulness and reliability of the proposed algorithm. Experiment’s result and analysis are provided in section five. Section 6 draws the conclusion.

2. Background

2.1. Data Fusion

Depending on the growth of information technology, more convenient ways to store and share data resource are provided with. Nowadays, as the idea of Internet sharing generally becomes popular, Internet users’ requirements for information no longer concentrate on single data resource, but tend to ask for multiple ones. Data patterns give unified view and format to describe the data in data resource [5]. Therefore, similar data are fused with each other, and unified data can greatly reduce data storage space and facilitate users to browse and operate data.

Data integration is the foundation and key of information system integration. Good data integration systems ensure that users can use heterogeneous data with low cost and high efficiency. Some difficult problems in data integration must be solved to achieve this goal [6]. The autonomy of the data sources can make a difference in quality between different data sources. The heterogeneity of data sources can cause conflicts between data. Because of these two points, the problem of inconsistent data often occurs during data integration [7].

Now the methods of data fusion can be divided as: data fusion, feature fusion and decision fusion [5]. Data fusion mainly eliminates the noise from the input data. Feature fusion and decision fusion emphasis on obtaining valuable information that relates to practical application [8]. Different methods of data fusion have different effects on data. For example, using feature fusion can reduce
the heterogeneous obstacles of data well [9], and the information fusion that close to the source has high accuracy [10]. However, with the accumulation of data, data fusion of different structures will lead to mutual exclusion of data. Some unforeseeable problems such as inconsistency, incompleteness, illegality and repeatability of data begin to show their influence. They will make data resources have the characteristics of data faultage [11]. After data fusion, the heterogeneity of data will lead to structural data faultage, which will cause bad influence on data fusion. Therefore, the process of structural data faultage is very important.

2.2. Structural Data Faultage

Reference [8] indicates that there are four main reasons which lead to data fault: first, when data structure is inconsistent or changes, data fault turns up; second, data created in different time is possible to be different in its properties; third, when the quantity of data continues to climb, the risk of finding data fault may also increase in a way; fourth, the changes of users’ needs will make the chosen data different.

Different data resource has differences that will bring trouble to data sharing in defining data patterns and describing data attribute, and then lead to the structural data faultage. According to the external features, structural data faultage refers to the data attributes that there comes trouble like repetition, missing, inconsistency in data resource. It makes data resource appearing the phenomenon of cracks, separations or fractures from a holistic perspective [12].

2.2.1. Concepts

Figure 1 shows three geological faultages. The geological faultage on the left side belongs to the field of positive faults. That is, the crust caused by the faultage slides upward, and then the distribution of sedimentary rocks in the crust will meet longitudinal displacement. But on the faultage surface of earth’s crust, there is still contact connection [13]; In Figure 1, the middle geological faultage belongs to the field of lateral fault. Although the faultage has caused lateral displacement in the earth’s crust, there is no marked difference between the two sides of the crust on the faultage surface in sedimentary rocks’ types and distribution, but the dislocation that only shows in both sides of the crust from horizontal dimension; refer to Figure 1, although the geological faultage on the right side has caused the faultage surface, as the time goes by, the distribution of the sedimentary rocks on both sides of the faultage surface does not show any changes. Sedimentary rocks only have differences in the width distribution [14].

**Definition 1.** Structural data faultage. Two subject-related data sets have structural differences on faultage surface. This kind of data faultage is called structural data faultage.

![Figure 1](image)

**Figure 1.** The differences of sedimentary rocks in faultage-surface caused by geological faultage.

The data fault in the data set is similar to the geological fault in the geological layer. Although data and geology belong to two different disciplines, they have certain similarities in terms of properties and other aspects [15].

Figure 2 is part of two personal information tables. As shown in the figure, the two data sets have obvious differences in structure. In the left diagram structure, field xb (gender), gbm (country code), gb (country), nzm (nation code), mz(nation), zzmmm (political status code), zzmm (political status), csrq (date of birth), zjlxm (certificate type code), zjlx (certificate type), zjhm (ID number) all appear in the right one. There is dislocation among them, but it has little influence on the resolution
towards the following faultage. However, the left data structure has attribute hkszssm (Registered city code) and hkszss (Registered city), which cannot be found in the right data structure. Correspondingly, in the right data structure, there are attribute ZSZYDM (own major code) and ZSZYMC (own major name), however the left one does not have relative attributes. If the two databases are fused with each other, the tables definitely have partial missing of the attribute value after fusion. That leads to data structure's incompleteness and inconsistency. This is so-called structural data faultage.

Definition 2. Plastid. Plastid is the word or character that is included by any attribute in data resource and has its independent meaning.

Definition 3. Reconfiguration Cost. The sum of operating cost for all the plastids in one attribute name gradually transforming into another attribute name is called Reconfiguration Cost, regarded as $T_c$. The formula is shown as:

$$T_c = \sum_{i=1}^{n} C_i$$  \hspace{1cm} (1)

In the formula, $C_i$ is a basic meta-operation, including replacement, shift, add, remove and other operations.

Definition 4. Isomorphous Attribute. In a data resource, the set of attributes $e$ in the set of attributes $E$ has an extremely similarity, where $e$ is a subset of $E$.

2.2.2. Effect

With the arrival of information age, companies and departments all strengthen the development of informatization, and gradually build their own information management platform. However, as the informational platform continues to build, related departments’ data all store in dispersive information management system. It causes kinds of “information island” coming into being [16].

“Information island” embodies specifically on [16]:

(a) The stage of the development of informatization

No matter it is enterprise informatization or informatization of government affairs, they all start from primary stage to middle stage, and finally reach advanced stage. At the primary stage of computer application, people can easily use computer from word processing and report printing, and then work on operations, and develop or bring in application systems. These applications that are scattered to develop and bring in generally will not consider the issue of data standard or information sharing. They pursue the goal as “the faster, the better”, and then lead to “information island” continuing to happen. And these data shapes without standard and sharing base provide structural data faultage with hotbed to exist and grow.

(b) Misunderstanding of informatization construction

Figure 2. A schematic of the heterogeneity between property names of two subject-related datasets.
For a long time, as informatization education does not reach enough depth and scope, misunderstanding of “valuing hardware, network and underestimating software and data” commonly exist among enterprises and departments. It makes users willing to spend money and energy to choose the type of appliances and build the network. Some even become “new apps chasers”. It makes network appliances continue to changes and causes a lot of waste. Enterprises and departments have not diligently developed and use the information resources, so the problem of “information island” is overlooked and is not solved for a long period of time. These cognitional misunderstandings have set intangible barriers and make it harder for solve the problem of structural data faultage fundamentally.

As the business develops, enterprises or departments need to comprehensively consider the actual condition among all the platforms or business, to decide the best design. Therefore, the problems of “information island” need to be tackled. And the data from every platform of inner enterprises need to be integrated and optimized [17]. For platforms or departments’ information management systems have differences because of the differences in hardware platform, operation systems and levels of technique. Those will make “information island” data’s integration lead to structural data faultage [18].

Therefore, the existence of structural data faultage have close connection with information management system’s difference. At the same time, for the need of decisions and other demands, the integration towards heterogeneous data is required. Then it makes the solution to structural data faultage become one of the most important research part in the field of data integration nowadays.

2.3. Related Research

At present, data fusion mostly appears in sensor, pedestrian recognition and other fields, and the most basic data table fusion is few. Data tables of the same class are fused together to increase the data storage space. But due to the different structures, different data fusion is bound to generate data errors. For example, the data’s incomplete, conflict, inconsistency and so on in the structure. How to solve the data error and ensure the integrity and consistency of the data structure after data fusion is very important. Data fusion is a combination of technologies designed to resolve conflicts of sources and find truths that reflect the real world [4].

Fendel et al. [19] use information extraction to fuse heterogeneous data in B2B e-commerce. They use a variety of trainable and self-learning AI techniques to improve. However, there are still two problems in the fusion. That is, the incompleteness of the extracted information and the inaccuracy caused by the error of the extracted information. Diego et al. [20] declaratively specify several types of coordination association ideas between data from different data sources for data integration in a data warehouse. Dong et al. [21] say data fusion plays an important role in data integration systems: it can detect and remove dirty data and improve the correctness of integrated data. Several data fusion models are mentioned. But no details are given on how to resolve data conflicts in data fusion. Tong et al. [22] outline 52 data fusion models combined with machine learning. Although many models ensure the quality of data after data fusion and prevent data faultage to some extent. Some models do not use real data for simulation, which makes the reliability of model evaluation weak.

The data fault theory started later than the traditional data preprocessing theory. Although the system has been preliminarily formed after nearly ten years of development and the concepts involved in the data faultage theory are gradually clear, there are still many contents to be improved and developed. The faultage formed in the structure is a structural faultage, which is defined as a structural data faultage by extension to the data.

Now few studies have applied data to the fault field for analysis. The reference [23] pointed out that the distributed data platform produced a large amount of distributed data, and there were faultages between these data formats. In order to provide high-quality data for follow-up work, data faultage needed to be processed. In this paper, a new type of Intermediate Storage System (ISS) is proposed to meet the requirements of data availability and interference minimization. And its effectiveness is verified by applying ISS to data faultage processing. The reference [24] pointed out that faultage is an important concept in geology, and the use of faultage can describe and explain
some data features in data resources. This paper introduces the faultages in the data resources from the micro level, and finally illustrates the dominant and recessive characteristics of the faultages in the data resources by taking the media transmission data as an example. The reference [25] used matrix detection of structural difference and faultage reconstruction to eliminate the original data faultage in dissertation data. But the execution efficiency of the algorithm is low and the time complexity is high.

3. The algorithm

3.1. Relate Definition

A key step in data fusion is record matching. It matches different records from different data sources that point to the same entity. Reference [26] records three main methods of record matching: content-based matching, structure-based matching, mixed matching. We discuss the content-based matching based on data fusion. Some similarity algorithms are used to match the values of one or more patterns recorded. The semantic similarity metric is a function that describes the degree to which two concepts are similar. To determine whether two entities are same, we need to measure the similarity [27]. Similarly, for structural data faultage generated in data fusion, we use similarity to process for data comparison.

The heterogeneity between data is an obstacle to the successful fusion of data. Determining the relationship between two data is an effective way to solve this problem [28]. The feature vectors of a data column describe the attributes of various aspects in a data column that are used to determine the degree of difference between data columns. According to the formal description of data column feature vector, the difference degree of each component is firstly defined and investigated. Then combine them for the difference between the data columns. Each data column has an attribute, so we summarize the difference between the data columns as the difference between the attributes.

From the perspective of attribute similarity, this paper will indicate the solution when there are structural data faultages in data resource—Structural Data Faultage Processing Algorithm based on Attribute Similarity (SDFPAAS). Because the data faultage solved by this algorithm is analyzed from the perspective of structure, which involves the discussion of attributes. First, we introduce several concepts involved in the algorithm in order to describe the algorithm.

**Definition 5. Attribute Similarity.** Similarity between two attributes \( P_1, P_2 \) in one dataset or between an attribute \( P_1 \) in one dataset and an attribute \( P_2 \) in another dataset is called Attribute Similarity \( Sim(P_1, P_2) \).

The sets of data waiting to be processed may have differences in individual attribute definitions, attribute ranges, attribute value types and so on, because of the various of storage system and medium. In such situation, extra analyzation and processing are required. The degree of retention of attribute information in source data set after data resource integration have direct connection with data resource's quality after the integration [29].

Attribute Similarity needs to comprehensively considerate the name of attribute, the description of attribute, the type of values, the distribution of values, value ranges and etc. The accurate calculation towards attribute similarity is the basement to achieve structural data faultage processing, and also the premise to analyze and decide the following data.

(a) Similarity between the name of attribute column \( N(P_1, P_2) \)

Attribute columns with the same name definitely describe the same semantics. Attribute column name similarity analysis is also called string similarity analysis. The method of string similarity comparison is usually used for calculating the similarity between attribute names \( N = (P_1, P_2) \). Now, there are three main ways to calculate string similarity: literal similarity matching, semantic similarity matching and correlation statistical matching [30]. The representative methods in literal similarity matching are calculation method based on edit distance [31] and method based on the same character and words [32]. We calculate the similarity between attribute names using the method based on edit distance.
The reconfiguration cost in definition 3 is edit distance. It represents the minimum numbers of insert, delete, and substitution that required to convert from one string to another. It can be shown as Formula (2), \( \text{Len}(\cdot) \) represents the length of the attribute name:

\[
N(P_1, P_2) = 1 - \frac{T_c}{\max(\text{Len}(P_1), \text{Len}(P_2))}
\] (2)

(b) Similarity between the description of attribute column \( D(P_1, P_2) \)

\( D(P_1, P_2) \) refers to the comparison of data that belongs to different attributes. The similarity between the description of the attribute column is calculated in the same way as the similarity between attribute names. It compares the similarity between each data. The formula is (3).

\[
D(P_1, P_2) = \frac{\sum_i^n N_i(P_{1i}, P_{2i})}{n}
\] (3)

(c) Similarity between value type of attribute column \( F(P_1, P_2) \)

\( F(P_1, P_2) \) is used to indicate that whether attributes’ type of value is matching. Because in the real world, different data resources may have different data types on similarity attributes. In order to considerate different situation generally, this paper will divide the data types into basic data types, such as numeric type, character type, date type. If two attribute types belong to one same data type, then \( F(P_1, P_2) = 1 \), otherwise \( F(P_1, P_2) = 0 \).

(d) Similarity between value distribution of attribute column \( S(P_1, P_2) \)

\( S(P_1, P_2) \) refers to value ranges of attributes. Calculation of the similarity between values can be divided into two ways [33]. The first case is when the value type is numeric type or similar numerical type. KL divergence and JS divergence can be used to calculate the similarity between distributions [34]. KL divergence is an asymmetric measure of the difference between two probability distributions. The asymmetry measure means \( KL(A, B) \neq KL(B, A) \). The formula for KL divergence is as follows.

\[
KL(P_1 \parallel P_2) = \sum_{i=1}^{N} P_1(x_i) \log \frac{P_1(x_i)}{P_2(x_i)}
\] (4)

It is easy to cause deviation and low accuracy when comparing the attribute similarity in our study. JS divergence measures the similarity of two attribute distributions. It solves the problem of asymmetry in KL divergence. And its value range is from 0 to 1. The formula for JS divergence is Formula (5).

\[
JS(P_1 \parallel P_2) = \frac{1}{2} KL(P_1 \parallel \frac{P_1 + P_2}{2}) + \frac{1}{2} KL(P_2 \parallel \frac{P_1 + P_2}{2})
\] (5)

We choose to use JS divergence to calculate the similarity between the distribution of attribute values. The higher the JS divergence, the greater the difference. In order to make the similarity measurement results consistent, the higher the JS divergence, the higher the similarity. We modify the formula of JS divergence. The Formula is (6).

\[
S(P_1, P_2) = 1 - JS(P_1 \parallel P_2)
\] (6)

If the value type is character type, now there are few researches on the similarity between character types. In this paper, the frequency of each character type under this attribute is first counted and then JL divergence is used to calculate the similarity between distributions.

(e) Similarity between value range of attribute column \( R(P_1, P_2) \)

\( R(P_1, P_2) \) are used for making sure whether attributes that value belongs to numeric type is similar in domain or not. Generally, if the value ranges of an attribute match each other, then the attribute must have connections. The range of values can be compared by taking the maximum value and the minimum value directly. If value range of two attributes matches each other, then \( R(P_1, P_2) = 1 \), otherwise \( R(P_1, P_2) = 0 \). For character data, the value range can be defined by using the frequency of plastids between attributes. Take the data with the highest occurrence frequency and the data with the lowest occurrence frequency as the upper and lower limits respectively. Then it can use the method of name similarity comparison to compare.

According to the above definition of attribute similarity, we use Euclidean distance to calculate the joint attribute similarity. The calculation formula of attribute similarity is shown as Formula (7).
\[
\text{Sim}(P_1, P_2) = \sqrt{F(P_1, P_2)^2 + R(P_1, P_2)^2 + N(P_1, P_2)^2 + D(P_1, P_2)^2 + S(P_1, P_2)^2} \quad (7)
\]

**Definition 6.** Attribute Cementation. During the process of data resource fusion, the operation that combines the same or similar attributes in data set which waiting for fusion into a same attribute is called attribute cementation.

When the types of two attributes in data set are different, if one of them is numeric type and the other one is character type, then there is little probability that these two attribute types are similar. Especially for a database with standardized design, there is no need to consider attribute cementation for these fields with little probability to be similar. The judgement of similarity degree of correlation attribute refers to Table 1.

| Plan                        | Similarity |
|-----------------------------|------------|
| Data type                   |            |
| Same                        | 1          |
| Different                   | 0          |
| Field length                |            |
| Same                        | 1          |
| Different                   | 0          |
| Primary key constraint      |            |
| all primary keys            | 1          |
| all non-primary keys        | 1          |
| One is primary key and one is non-primary key | 0 |
| Range of constraints        |            |
| All have constraints        | 1          |
| All don’t have constraints  | 1          |
| One has constraints and one don’t have constraints | 0 |
| Value content               |            |
| The variance is approximate | 1          |
| The minimum is approximate  | 1          |
| The maximum is approximate  | 1          |

For the unfused attributes that without cementation processing, in order to make sure the integrity of structure, these attributes need to be added into attribute results. This involves attribute addition.

**Definition 7.** Attribute Addition. During the process of data resource fusion, in order to make sure information’s integrity, the attribute sets whose similarity with other attributes are lower than threshold will be added into attribute results. This operation is named attribute addition.

After each source data sets are processed in the process of data faultage, data sets merge with each other. After the fusion, the information like the source of each record in data sets need to be record in order to make sure of information’s integrity.

**Definition 8.** Incursion. During the process of data resource fusion, for signing every source of data record after fusion and fail to match the attributes whose similarity is higher than the threshold, the attributes need to be newly added. Such operation is called incursion.

**Definition 9.** Mapping. The operation that matches the entity record in data resource and heterogenous data is called mapping.

The essence of mapping is the operation that equivalently transfers heterogenous information into another pattern. It is the basic step of structural data faultage processing, and also the final goal of it.

**Definition 10.** Isomorphous Homonuclear. There are two or more records in data resource show a same entity object in the real world. After the processing of data faultage, the records are combined into one. This operation is called isomorphous homonuclear.
In the data fusion of different data tables, the main difference is the fusion of attributes. When attribute’s definition is the same, data can be fused straightly based on the attribute value. For instance: in two attributes, one is named by lowercases letter, and the other one is by capital letter; when two attributes are named in a same way and only the units of measurement of values are different, all of them can use attribute cementation processing to treat it. For the data fusion of different attribute definition, attribute addition processing can be used to meet such situation. In order to reduce the data redundancy, isomorphous homonuclear can unify the attributes that use different name to indicate the same content. In the end, the data tables after structural data faultage processing should satisfy data structure’s integrity and consistency.

3.2. The Main Idea

The definition and detection of structural data faultage describe and measure the differences among several data tables from the perspective of structure. Therefore, when processing the data tables with structural data faultage, the primary consideration is the impact of different data tables’ structures on particular data—since most of the time, the entities in fused table are different. Moreover, the situation will result in losing key data without being well treated, then which will bring difficulties to the subsequent data analysis. In this way, structural data faultage should be dealt from the completeness and consistency of data tables comprehensively.

After data fusion, the biggest problems are the missing and repetition of data. This algorithm is mainly used for eliminating structural data faultage, and ensure the integrity of data structure and efficient resource utilization. The three main operations of the algorithm are attribute cementation, attribute addition and isomorphous homonuclear. The three operations were respectively explained in Definition 6, Definition 7 and Definition 10.

The premise of this algorithm is to carry out subsequent process according to attribute similarity. Compared with other semantic similarity computing methods, the semantic computing method combining multiple relationships can improve the accuracy of semantic computing [35]. According to the definition of relevant attributes in 3.1, We calculate the joint similarity between attribute names, attribute descriptions, attribute types, attribute distributions, and attribute value ranges. Thus, the calculation result of attribute similarity is more accurate.

The algorithm firstly worked out the similarity of unfused and fused attributes after data fusion, and then decided what to do with structural data faultage based on attribute similarity. The algorithm can determine whether the data was partially missed or totally missed, as well as the relevance among the attributes. The flowchart of this algorithm is shown in the Figure 3.
Attribute cementation primarily focuses on the partial missing of data, which refers to the partly missing data fields in some fused attributes after data fusion. Some attributes as well as their field’s content all missed after data fusion. This is the complete missing of data. Attribute addition can add the missing attributes as well as the data fields into the data fusion table completely, and then make sure that none of the data fields are missing. After the processing of data structure’s completeness, data fusion table must have some problems like content repetition and data redundancy. Isomorphous homonuclear processing can unify the content of similar attributes, and explain the correlative connection of data in the data dictionary. Those processesions are to make sure the consistency of data’s structure, reduce data redundancy and make it convenient for accessing data.

The steps of algorithm are as follows:

(a) First: Data fusion. The data tables need to be processed are fused according to the actual situation.

(b) Second: After the completion of data fusion, structural data faultage processing began. First of all, extract the data content without being fused in data resource.

(c) Third: Calculate the attribute similarity between the unfused and fused data.

(d) Fourth: Set the threshold T. Attribute similarity is sorted from high to low. Each attribute similarity corresponds to two attributes. At the same time, check whether the two attribute names corresponding to the attribute similarity belong to the same semantics. If inconsistencies occur, the comparison is terminated, and the similarity of the previous attribute of this attribute similarity is taken as the threshold.

(e) Fifth: Set the unfused data attributes that are bigger than or equal to the threshold T as partly missing of data. Set the attributes smaller than the threshold T as complete missing of data.

(f) Sixth: Process the attributes as well as the field content whose data is partly missed through attribute cementation. Based on fused attribute units, map the corresponding data fields that are partly missed to the attribute in the table of data fusion.
Seventh: Process the attributes as well as the field content whose data is completely missed through attribute addition. Make the complete missing part of data incursion to the table of data fusion.

Eighth: Do the operation of Isomorphous Homonuclear. At this point, the data fusion table has the highest content integrity, but there are many redundant data and duplicate data in the data table. Many data express the same semantic meaning with different names which result in a large proportion of space resources and high data redundancy. After the seventh step, we calculate the attribute similarity between the data attribute columns. The attributes with large correlation are selected for Isomorphous Homonuclear. And label the correlative connection of attribute value in the data dictionary.

The Isomorphous Homonuclear is also based on attribute similarity. We set up the attribute connection diagram according to the similarity. The attribute is listed as the node and the similarity represents the edge. It can be found that the span between two of these similarity values is much larger than that between other similarity values by sorting the similarity between each attribute column and other attribute columns in descending order. The attribute column before this span has similar distribution to that column, and the data distribution after that is completely different. We select the critical value of this span as the threshold $\theta$.

We classify the attribute column $P'$ that satisfy $\text{Sim}(P, P') \geq \theta$ as approximate matching attribute of $P$. And connect $P'$ with $P$ to form the attribute connection diagram. If there is an approximate matching attribute $P'$ in the attribute connection diagram of attribute column $P$, and there is an attribute $P$ in the attribute connection diagram of attribute column $P'$. It indicates that attribute column $P$ and attribute column $P'$ match each other. The attribute $P$ and $P'$ express the same semantic meaning and can be combined with each other. If no other attribute column matches the attribute column, there is no attribute that can be combined with it. The attribute connection diagram is illustrated in Figure 4. Finally, the attributes that match each other are represented by one attribute according to the attribute connection diagram.

![Figure 4. Attribute connection diagram.](image)

4. Experiment

In order to make logical choice towards master’s thesis used for sampling, and ensure the quality of postgraduate degree, Shanghai academic degree committee obtains information of doctor’s thesis in Shanghai from China Academic Degrees & Graduate Education Development Center regularly, in order to ensure data sources are authoritative and unique.

According to the data from China Academic Degrees & Graduate Education Development Center, there are three basic types of PhD degree’s information sheet. They are the table of Academic degree, the table of Professional doctor degree and the table of Equivalent degree: doctor degree. The fusion of three data tables can greatly improve the quality and efficiency of paper sampling. However, due to the data structures of the three degree-granting tables are not consistent, there is a lack of uniform definition standards for the data of degree types and the data of degree granting institution. And the data on hardware, operation system and database system cannot reach the same level. The data table after fusion has various differences in data definition and storage structure. It
can make it hard to do highly efficient information sharing and data interaction, and will lead to the problems like the high repetition rate of data, low information utilization and complex data retrieval [36]. It is not enough to support the work of sampling master’s thesis. Therefore, it is necessary to analyze and process the macroscopic data faultage towards the three tables of degree.

4.1. Experiment Design

The laboratory uses python as development language. Data fusion and the process of structural data faultage were carried out on Pycharm development platform.

The laboratory chooses the table of normal doctor degree, the table of Equivalent degree: doctor degree and the table of professional doctor degree in Shanghai university in 2005 as experimental input data. Partial data information of the three tables are shown in Figure 5, Figure 6 and Figure 7 respectively (all in English).
The table of Academic degree records and conserves the degree condition in every year from every field. It mainly includes the relative information of degree-conferees (names, ID numbers, name of school, tutors’ name, student number, major and etc.), information about thesis (title, keywords and etc.). The characters of the three charts are different, resulting in the attributes they include and names of attributes used for naming the same content are different. The data volume size of the three data tables can refer to Table 2. The table of doctor Academic degree refers to data 1. The table of equivalent degree: doctor degree refers to data 2. The table of Professional doctor degree refers to data 3.

Table 2. The size of the three data tables.

| data1 | data2 | data3 |
|-------|-------|-------|
| The number of attributes | 67 | 57 | 61 |
| The amount of data | 12446 | 209 | 785 |

The three data tables have differences in many areas like the storage structure and data’s definition because of the variety of the real application conditions. In order to ensure the unity and integrity of the paper sampling data’s structure, and to deal with the data problems encountered in data fusion, the analysis and processing are required towards the structural data faultage during data fusion.

4.2. Process of the Experiment

4.2.1. Data Fusion

The differences among data tables mainly reveal on the differences of attribute value in data resource. The fusion of data table is the fusion of data resource’s attributes. According to the mapping relationship, the data resources with the same attributes are integrated together to complete the fusion.

According to the algorithm mentioned before, as the distinction of the three data tables referred to the difference of attribute value they contained, and the accuracy of feature fusion is higher than other ways of fusion from the reference [37], the three data tables need to do feature fusion based on attributes’ feature. After fusing data 1, data 2 and data 3, there would be a new table data. Based on the statistics, after this data fusion, the attribute values in the new table data were the common attributes owned by the three data tables. However, some parts of the content of the new one had
lost. It meant there were some attributes that had not been fused. And for the fused parts, there were also some repetition and missing. From the external structure, they are all structural data faultage existing after data fusion.

Table 3 lists all the attribute values of the three data tables. The attributes with shadow in the table are the attributes that cannot be fused to each other, and the attributes with no shadow are the attributes that have been fused to each other.

| Attribute Name | Attribute Constraint | Attribute Name | Attribute Constraint | Attribute Name | Attribute Constraint | Attribute Name | Attribute Constraint |
|----------------|----------------------|----------------|----------------------|----------------|----------------------|----------------|----------------------|
| SSDM           | int(20)              | XFSM           | int(20)              | XWSYDWM        | int(20)              | QXM            | int(20)              |
| SSNC           | varchar(255)         | XFS             | varchar(255)         | XWSYDWM        | varchar(255)         | QX             | varchar(255)         |
| XM             | varchar(255)         | DSM             | varchar(255)         | ZXXM           | varchar(255)         | GZDWXZM        | int(20)              |
| XPY            | varchar(255)         | BYNY            | int(20)              | ZXXM           | varchar(255)         | GZDWXZ         | varchar(255)         |
| XBM            | int(20)              | HXWRQ           | varchar(255)         | YJXKSY         | varchar(255)         | GZDWSSM        | int(20)              |
| XB             | varchar(255)         | XWZSBH          | varchar(255)         | XWLB           | varchar(255)         | ZYXWLYM        | int(20)              |
| GB             | int(20)              | LWTM            | varchar(255)         | WLB            | varchar(255)         | ZWJB           | varchar(255)         |
| GB             | varchar(255)         | LWGC            | varchar(255)         | ZYDM           | int(20)              | ZWJB           | int(20)              |
| MZM            | int(20)              | LWLX            | int(20)              | YJKDM          | varchar(255)         | ZCJB           | varchar(255)         |
| MZ             | varchar(255)         | LWLX            | varchar(255)         | YJKMC          | varchar(255)         | ZCJB           | int(20)              |
| ZZMM            | varchar(255)         | LWXLYL          | int(20)              | ZYSZM          | varchar(255)         | GZDW           | varchar(255)         |
| ZZMM           | varchar(255)         | LWXLY            | varchar(255)         | ZSZYM          | varchar(255)         | SQXWNY         | int(20)              |
| CSRO            | int(20)              | QZWXM            | int(20)              | ZSZYM          | varchar(255)         | SQXWNY         | int(20)              |
| ZJLM            | varchar(255)         | QZXW             | varchar(255)         | KSH            | varchar(255)         | XWSYDWM        | int(20)              |
| ZJLX            | varchar(255)         | QZXL              | int(20)              | KSFSM          | int(20)              | XM             | int(20)              |
| ZJHM           | varchar(255)         | HQZXWNY          | int(20)              | KSFS           | varchar(255)         | XMPY           | varchar(255)         |
| HKSZSSM         | int(20)              | QZXWDM            | int(20)              | RXNY           | int(20)              | SQH            | varchar(255)         |
| HKSZSS          | varchar(255)         | QZXWDM           | varchar(255)         | XH             | int(20)              | ZYXWLY         | int(20)              |
| ID             | int(20)              | XWSYDWM          | int(20)              | ZSZYM          | varchar(255)         | ZPSTATE        | varchar(255)         |
| YJXKSY         | varchar(255)         | XWSYDWM          | varchar(255)         | QZXWKKM        | int(20)              | ZP             | varchar(255)         |
| GDLX             | int(20)              | GDLX             | varchar(255)         | QZXWXK          | int(20)              | GZKZ            | int(20)              |
| GPSTATE         | varchar(255)         | ZP               | varchar(255)         | GZWX            | int(20)              | GZDWSS          | varchar(255)         |

4.2.2. Structural Data Faultage Processing

According to the data table after data fusion, the attributes that are not been integrated by the three tables are ZJHM (ID Number), XWSYDWM (Degree-conferring unit code), SQXWNY (Date of application), SQH (Application number), GZDW (Work units), ZCJB (Professional title code), ZCJB (Professional title), ZWJB (Position code), ZWJB (Position), XM (Name), XMPY (Spell-Name), ZYXWLYM (Professional degree field code), ZYXWLY (Professional degree field). Calculate the similarity between these attributes and integrated attributes based on Definition 5.

The similarity of attributes missing from part of the content must be high, because such attributes already exist in the fusion table. While the similarity of attributes missing completely from the content must be low because such attributes do not exist in the fusion table. Firstly, the threshold value should be determined. When the incomplete fusion attributes and the fused attributes are completely different under a certain threshold value, it indicates that the threshold value has reached a critical value. In addition, the previous threshold value of the threshold value should be taken. The process of determining the threshold value is shown in Table 4. When the threshold value is 1.12448335244, the attribute ZJHM and the QZXWXKM of the result table appear, which are two completely different attributes. This indicates that partial and complete deletion of attribute content has been completed before this threshold value.
Finally, the threshold value of attribute similarity was determined to be 1.1313708499. The attribute value whose attribute similarity was bigger than or equal to threshold $T$ are ZJHM and ZJHM in result tables, XWSYDWM and XWSYDWM in result tables, XM and XM in result tables, XMPY and XMPY in result tables. Vividly, those unsuccessfully fused attributes all had appeared in the new database. However, the attributes in the new database had not been completely fused and were partially missed. Therefore, the attribute cementation processing is adopted to recombine these several attribute units.

After comparing with all the attribute similarity in the new table, the attributes that are smaller than threshold $T$ were SQXWNY, SQH, GZDW, ZCJBM, CJB, ZWJBM, ZWJB, ZYXWLYM, ZYXWLY. We found that there were no these attributes in new table after data fusion. It meant there were complete missing among these attribute value in the new table. Then we take the operation of attribute addition. Make those attribute values intrude into the new table again. It meant to add the attribute units into the new table. Table 5 shows the attributes for which attribute cementation and attribute addition operations are required. At that time, the result table of data fusion is the most complete one on data structure, without missing and the situation that the same attributes were not fused.

Table 5. Failed to merge properties and the subsequent actions.

| Attributes in Data Fusion that Failed to Fuse | Operation |
|---------------------------------------------|-----------|
| Isomorphous Attribute                        | Attribute Cementation |
| ZJHM; XWSYDWM; XM; XMPY                      |           |
| Non-isomorphous Attribute                    | Attribute Addition |
| SQXWNY; SQH; GZDW; ZCJBM; CJB; ZWJBM; ZWJB; ZYX |            |

After the first step of structural data faultage processing, which were attribute cementation and attribute addition. The data structure’s integrity of this data result table could be ensured, and the attribute value of the three data tables were all fused into the new table.

After finishing the integrity processing in structural data faultage, the semantic analysis of each attribute unit in the new data table calculated by the algorithm found GZDWXZM (Work unit property code) and GZDWXZ (Work unit property) referred to a same entity object. And there are many similar attributes in data table. Define this kind of attribute as isomorphous attribute. For example, XB (gender) and XBM (gender code), gender code 1 refers to male, and gender code 2 refers to female. At present, the storage space for data is limited, so using two or more attributes to show one same object means a large occupation of storage space. It will make the content repetitive, waste resources, and bring inconvenience for the stock check of the thesis. Therefore, we used isomorphous
homonuclear processing to build the single representation of the set of isomorphic attribute, and uniform standard definition of some units to ensure the consistency of data. In this experiment, the character type and numeric type attributes that associated with each other are replaced by the numeric type, and all the attributes that have associative relationships are shown in Table 6. All attributes in Table 5 can be replaced by isomorphic homonuclear operations which ultimately reduces the volume of the data table and increases the storage space of the data table, making the paper sampling more accurate and easier to obtain information.

### Table 6. Set of properties for the operation of isomorphic homonuclear.

| Field Name | Corresponding Relation |
|------------|------------------------|
| XBM        | XBM--XB                |
| GBM        | GBM--GB                |
| MZM        | MZM--MZ                |
| ZZMMM      | ZZMMM--ZZMM            |
| ZJLXJM     | ZJLXJM--ZJLX           |
| HKSSZSM    | HKSSZSM--HKSSZSS       |
| XWSYDWM    | XWSYDWM--XWSYDWM       |
| XWLBM      | XWLBM--XWLBM           |
| GDLXM      | GDLXM--GDLX            |
| LWLXM      | LWLXM--LWLX            |
| LWXTLYM    | LWXTLYM--LWXTYL        |
| QZXWM      | QZXWM--QZXW            |
| QZXWXKM    | QZXWXKM--QZXWXK        |
| QZXWDWM    | QZXWDWM--QZXWDW        |
| GZDWXZM    | GZDWXZM--GZDWXZ        |
| GZDWSSM    | GZDWSSM--GZDWSS        |
| GZXZM      | GZXZM--GZXZ            |
| ZWJB       | ZWJB--ZWJB             |
| ZYXWLYM    | ZYXWLYM--ZYXWLY        |
| QXM        | QXM--QX                |
| ZCJBM      | ZCJBM--ZCJB            |
| ZYXWLYM    | ZYXWLYM--ZYXWLY        |
| XXFSM      | XXFSM--XXFS            |

### 5. Experiment’s Results Analysis

After the process of structural data faultage, the data table should include all the data fields that need to be fused. Figure 8 showed the comparison of the data structure’s completeness after the processing of faultage. It can be straightly seen that the structure’s integrity of the data table after data fusion is guaranteed by structural data faultage processing. Make a comparison of the integrity of data before and after the attribute cementation and attribute addition: the integrity of data structure that was just fused account for 83%, while the one that had finished the complete faultage processing reached 100%.
Figure 8. The comparison of data structure’s integrity.

Figure 9 shows the change of table size in the disk during structural data faultage processing. Since the missing part of attribute value is different by using different data tables as the bottom table to fuse, three data tables are respectively used as bottom table for fusion. And the size changes during the processing of three structural data faultages are counted. According to its size change, the data table after structural data faultage processing occupies less space in the whole storage resource, increasing the additional data resource storage space.

The laboratory uses information entropy to measure the structural data faultage processing’s result after data fusion. In physics, entropy is the parameter used for describing the randomness of things. In the field of mathematics, entropy is an abstract concept [38]. In information system, information entropy can be used to show the nondeterminacy of the information. The more certain an information system is, the lower the information entropy will be. On the contrary, the more the information entropy is, the less certain the system will be. Therefore, information entropy can be used as the measurement towards a system’s nondeterminacy, as well as the measurement of a system’s complexity [39]. The higher the entropy is, the more disordered the data is [40].

Figure 10 shows the information entropy of the three data tables, compared with the amount of data contained in the three data tables in table 2. The information entropy of the data table and the amount of data contained in the data table correspond to each other.
Figure 10. Information entropy of each data table.

Calculate the information entropy respectively of the data table after data fusion, attribute cementation and isomorphous homonuclear processing. During the data fusion, if there are different tables that are used as bottom table, the changes of information entropy will also be different after data fusion. Figure 11 respectively use data1, data2 and data3 as the bottom table for fusion. Since the chosen bottom table is different, the attributes that can be fused are also different. That leads the information entropy is different after fusion. It can be straightly observed the changes of information entropy of data tables after the operation of data fusion, attribute cementation, attribute addition and isomorphous homonuclear during the structural data faultage processing. The result shows that after data fusion, the entropy of data information decreases by structural data faultage processing. It shows that the data table system is definite and orderly after structural data faultage processing.

Leave alone the way used for data fusion, the result revealed that the information entropy of the table of data fusion was the smallest after structural data faultage. It can be seen from the change trend of information entropy that the information entropy became the biggest after the operation of attribute cementation and attribute addition. At that time, the data table is the most complete, and contains the most amounts of data. The redundancy rate and the repetition rate of data were quite high, so that the value of entropy was the biggest. After the operation of isomorphous homonuclear, the redundancy rate and the volume of data in the table reduced. The information entropy of the data
fusion table is reduced to the minimum. It meant that the final system of the data table is ensured and the simplest.

The structural data faultage processing algorithm based on attribute similarity (SDFPAAS) is compared with the matrix detection of structural difference and faultage reconstruction algorithm (SDMDFS) proposed in reference [25]. SDMDFS first constructs the structural difference matrix based on the data column difference, then detect faultages according to the matrix. Finally, the faultage reconstruction is carried out according to the detection results. As the algorithm proposed in reference [25] requires detection first and then faultage reconstruction, the algorithm is inefficient. The method we come up with detects and processes the fused data at the same time. We compare three indicators: efficiency, structural integrity and data uncertainty. The efficiency is represented by the processing time. The information entropy is used to describe the uncertainty of the data. The results are shown in Table 7. The algorithm proposed in this paper is more efficient, which proves the effectiveness of the algorithm.

| Algorithm       | SDMDFS | SDFPAAS |
|-----------------|--------|---------|
| Elapsed time    | 6 min  | 30 min  |
| Structural integrity | 100%   | 100%    |
| Data uncertainty | 0.0466 | 0.0533  |

Table 7. The comparison of two algorithms’ running time.

The method proposed in this paper solves the structural data fault through experimental analysis. The simulation and method comparison are carried out in the actual scene to make the result evaluation more reliable. After structural data faultage processing, the completeness and the consistency of the data fusion table are assured. The result data table structure is ordered and determined. At the same time, the storage space for data was enlarged, which makes the spot check towards thesis, as well as the data storage become easier. The efficiency of the algorithm is also verified by comparison.

6. Discussion

Nowadays, data tables are used in many fields to store the data. The fusion among data tables is good for enlarging the storing space. In this paper, data errors caused by data fusion are associated with geological faultage, and data errors on structure are classified as structural data faults. In order to eliminate the structural data faultage, a structural data faultage processing algorithm based on attribute similarity is proposed. Structural data faultages are processed according to attribute similarity. Finally, apply the structural data faultage theory in the sampling data of Shanghai Doctoral Dissertation, and make practical use of the theory. It can be found that this theory can effectively increase the storage space of data, and at the same time, can make sure of the completeness and effectiveness of data table’s structure. This theory set the foundation for providing with high quality of data set for the spot check of academic dissertation. Through practice, we further demonstrated the usefulness and reliability of structural data faultage algorithm, and rise the quality of sample data.

The theory of data faultage was started late, and has been generally in the face of advances in recent years. Therefore, there is little research in this field. The new method put forward in this passage can deal with structural data faultage during data tables fusion efficiently, and expand the space for data storage. In the future, we are supposed to do further research on content data faultage. This paper processed the error that data had in the structure. If the faultage is in the content of data resource, then this kind of data faultage belongs to content data faultage. There are some examples: data updating [41]; different data fusion and integration requirements [42]; the conflicts of definition in data attributes [43] and so on. Further research on content data faultage in data fusion also need in the future.

Author Contributions: Conceptualization, F.C; data curation, F.C. and J.T.; formal analysis, R.H.; funding acquisition, J.X.; investigation, R.H.; resources, J. X.; supervision, J.X. and J.T.; validation, J.T.; visualization, F.C.
and R.H.; writing – original draft, F.C. and J.X.; writing – review & editing, F.C. and R.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** The theory of this research work is supported by the National Natural Science Fund Project (40976108; 61303097), the Shanghai Science Fund Project (17ZR1428400), the Shanghai Construction of Key Disciplines Fund Project (J50103), the Second (2016) Shanghai Research Project for Private University (2016-SHNGE-08ZD), and the Shanghai University Graduate Innovation Fund Project (SHUCX070037; SHUCX120105) the support of this paper work.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Fengguang, X.; Xie, H.; Liquin, K. Research and implementation of heterogeneous data integration based on XML. Electronic Measurement & Instruments, 2009. In Proceedings of the ICEMI '09. 9th International Conference, Beijing, China, 16–19 August 2009.

2. Gal, A. The health problems of data integration. *Indian J. Med Paediatr. Oncol.* **2008**, 29, 65–65.

3. Li, Q.; Li, Y.; Gao, J.; Zhao, B.; Fan, W.; Han, J.; Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In Proceedings of the Acm Sigmod International Conference on Management of Data, Snowbird, UT, USA, 22–27 June 2014.

4. Dong, X.L.; Divesh, S. Big data integration. In Proceedings of the 2013 IEEE 29th international conference on data engineering (ICDE), Brisbane, Australia, 8–12 April 2013.

5. Yi, S.; Xue, Z.; Lin, T.; Jing, H.; Wang, Z.; Wang, X.; Gao, X. Data Integration Technology Research and Application in Integrated Distribution Network Planning Platform. *Power Syst. Technol.* **2016**, 7, 2119–2205.

6. Hasselbring, W. *Information System Integration*; ACM: New York, NY, USA, 2000.

7. Hull, R.; Zhou, G.; A framework for supporting data integration using the materialized and virtual approaches. *Acm Sigmod Rec.* **1996**, 25, 481–492.

8. Snively, E.; Russell, A.P.; Powell, G.L.; Theodor, J.M.; Ryan, M.J. The role of the neck in the feeding behaviour of the Tyrannosauridae: Inference based on kinematics and muscle function of extant avians. *J. Zool.* **2014**, 292, 290–303.

9. Banerjee, T.P.; Das, S. Multi-sensor data fusion using support vector machine for motor fault detection. *Inf. Sci.* **2012**, 217, 96–107.

10. Xia, F.; Zhang, H.; Peng, D.; Li, H.; Xu, L.; Yang, L. In Proceedings of the Research of the Reliability Coefficient in Information Fusion, Kuala Lumpur, Malaysia, 3–5 April 2009.

11. Xia, J.; Wang, J.; Yan, C.; Xu, J. Research on Data Faultage Phenomena. *J. Comput. Appl. Softw.* **2013**, 77, 9–13, doi:10.3969/j.issn.1000-386x.2013.08.003.

12. Wang, J. Research on Data Faulting in Data Resources. Master’s Thesis of Shanghai University, Shanghai, China, 2013.

13. Ward, S.N. On the consistency of earthquake moment rates, geological fault data, and space geodetic strain: The United States. *Geophys. J. Int.* **2010**, 134, 172–186.

14. Jingguang, S.; Tianpeng, G. The Research of Expression Method on Geological Fault Modeling. *J. Geo-Inf. Sci.* **2016**, 18, 1331.

15. Xu, J.; Xia, J.; Zhou, S. Application of data fault analysis in data processing of broadcasting station. *Comput. Appl. Softw.* **2016**, 33, 158.

16. Buitelaar, P.; Cimiano, P.; Frank, A.; Hartung, M.; Racioppa, S. Ontology-based information extraction and integration from heterogeneous data sources. *Int. J. Hum. Comput. Stud.* **2008**, 66, 759–788.

17. Qiu, D.; Liu, J.; Zhao, G.; Design and application of data integration platform based on web services and XML. In Proceedings of the 2016 6th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 17–19 June 2016.

18. Johnson, S.; Zhou, S.; Miao, H.; Hsu, D. A Theory of Data Faultage. In Proceedings of the 2015 3rd International Conference on Computer and Computing Science (COMCOMS), Hanoi, Vietnam, 22–24 October 2015.

19. Fensel, D.; Ding, Y.; Omelayenko, B.; Schulten, E.; Botquin, G.; Brown, M.; Flett, A.S. Product data integration in B2B e-commerce. *Ieee Intell. Syst.* **2001**, 16, 54–59.

20. Calvanese, D.; De Giacomo, G.; Lenzerini, M.; Nardi, D.; Rosati, R. Data Integration in Data warehousing. *Int. J. Coop. Inf. Syst.* **2001**, 10, 237–271.
21. Dong, X.; Naumann, F. Data Fusion—Resolving Data Conflicts for Integration. Proceedings of the VLDB Endowment 2009, 2, 1654–1655.
22. Meng, T.; Jing, X.; Yan, Z. et al. A survey on machine learning for data fusion. Inf. Fusion 2020, 57, 115–129.
23. Steven, Y.K., Imranul, H.; Brian, C.; Indranil, G. In Making Cloud Intermediate Data Fault-Tolerant; Joseph, M., Hellerstein, S.C., Mendel, R., Eds.; In Proceedings of the 1st ACM Symposium on Cloud Computing, Indianapolis, IA, USA, 10–11 June 2010; ACM Press: Indianapolis, IA, USA, 2010; pp. 181–192.
24. Daniel, H.; Shardjom, J.; Miao, H. Data Faultage in Data Resource. Int. J. Database Theory Appl. 2015, 8, 271–284.
25. Zhou, S. Application of Data Fault in Degree Data Preprocessing. Master’s Thesis of Shanghai University, Shanghai, China, 2016.
26. Dorneles, C.F.; Gon, A.R.; Mello, R.D.S. Approximate data instance matching: A survey. Knowl. Inf. Syst. 2011, 27, 1–21.
27. Koudas, N.; Sunita, S.; Divesh, S. Record linkage: Similarity measures and algorithms. In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, Chicago, IL, USA, 27–29 June 2006.
28. Lu, J.; Xue, X.; Lin, G.; Huang, Y. A New Ontology Meta-Matching Technique with a Hybrid Semantic Similarity Measure; Advances in Intelligent Information Hiding and Multimedia Signal Processing; Springer: Singapore, 2020; pp. 37–45.
29. Madnick, S.E.; Wang, R.Y.; Lee, Y.W.; Zhu, H. Overview and Framework for Data and Information Quality Research. J. Data Inf. Qual. 2009, 1, 1–22.
30. Chang, A.X.; Spitkovsky, V.I.; Manning, C.D.; Agirre, E. Evaluating the word-expert approach for Named-Entity Disambiguation. arXiv 2016, arXiv:1603.04767.
31. Niu, Y.; Qiao, C.; Li, H.; Huang, M. Word Embedding based Edit Distance. arXiv 2018, arXiv:1810.10752.
32. Kang, L.; Liheng, X.; Jun, Z. Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model. IEEE Trans. Knowl. Data Eng. 2015, 27, 636–650.
33. Teng, S.; Li, J.; Li, R.; Zhang, Z. The calculation of similarity and its application in data mining. Comput. Knowl. Technol. 2016, 75, 807–813.
34. Barz, B.; Rodner, E.; Garcia, Y.G.; Denzler, J. Detecting Regions of Maximal Divergence for Spatio-Temporal Anomaly Detection. IEEE Trans. Pattern Anal. Mach. Intell. 2019, 41, 1088–1101.
35. Duan, J.; Wu, Y.; Wu, M.; Wang, H. Measuring Semantic Similarity between Words Based on Multiple Relational Information. Ietie Trans. Inf. Syst. 2020, 103, 163–169.
36. Jung, T.; Li, X.Y.; Wan, Z.; Wan, M. Control Cloud Data Access Privilege and Anonymity with Fully Anonymous Attribute-Based Encryption. Ietie Trans. Inf. Forensics Secur. 2015, 10, 190–199.
37. Raja Khurram Shahzad. Android Malware Detection Using Feature Fusion and Artificial Data. In Proceedings of the 2018 IEEE 16th Intl Conference on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Athens, Greece, 12–15 August 2018.
38. Wilson, A.G. The Use of the Concept of Entropy in System Modelling. J. Oper. Res. Soc. 1970, 21, 247–265.
39. Wellmann, J.F.; Regenauer-Lieb, K. Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models. Tectonophysics 2012, 526–529, 207–216.
40. Jeremy, D.S.; Charles, S.Z. The compression–error trade-off for large gridded data sets. Geosci. Model Dev. 2017, 10, 413–423.
41. Xu, W. Research on spatial data change discovery and update method based on spontaneous geographic information. Build. Mater. Decor. 2015, 401, 232–233.
42. Lu, Y.; Chen, J. Research on Deep Web database integration technology. J. Shanghai Norm. Univ. Nat. Sci. 2016, 45, 422–427.
43. Li, G. Research on Multi-Project Management Method and Application of Networked Collaborative Design Under Generalized Resource Constraint; Chongqing University: Chongqing, China, 2011.