

Research on the Construction of a New Degree Quality Evaluation Model Based on Data Fusion and Rule Sampling

Shardrom Johnson^{1,2} and Miao Hui¹

¹*School of Computer Engineering and Science, Shanghai University, Shanghai 200444 P. R. China*

²*Xianda College of Economics & Humanities Shanghai International Studies University, Shanghai 200083 P.R China*
jshardrom@shmec.gov.cn, miaohui@shu.edu.cn

Abstract

With the rapid development of higher education, how to safeguard and promote the quality of degree training has increasingly become the focus of all sectors of society and training units. Strengthening evaluation is an important process to ensure the quality of the degree-granting. To weaken the human factor and reduce the complexity of human intervention in the evaluation process, this paper presents a new degree evaluation model. This model consists of a command management unit, data unit, sampling rules unit, index system unit, evaluation system unit and information feedback unit. In this model, data cleaning and data integration are used to deal with multi-source heterogeneous degree data, and the rule sampling method is applied to achieve the complex and diverse sampling requirements. To prove the scientific and effective nature of this evaluation model, we applied this model to a sampling of master's dissertations from Shanghai in 2014. The result of using this evaluation model on this sampling met the requirement of the Municipal Degree Committee.

Keywords: *Quality assurance, system model, data integration, rule sampling*

1. Introduction

To ensure postgraduate quality and degree-granting quality, the postgraduate training unit evaluates dissertations through a third party. This has become one of the important components of the current graduate education teaching reform [1]. Dissertation sampling was first seen in 1997 when the Shanghai Academic Degree Committee Office invested over 30 million yuan in dissertation sampling and “double-blind” review for doctoral and master’s dissertations before degrees were awarded. Then, other provinces and cities emulated this. The purpose of this measure is to enhance the quality consciousness of postgraduate training units and to ensure the postgraduate quality [2]. Since the dissertation sampling work was trial implemented in Shanghai and other provinces, it has made some achievements and been widely used in major cities throughout the country. In recent years, dissertation sampling has basically become one of the important means to ensure master’s dissertation quality and to improve postgraduate education level [3].

Some scholars have described the functions, procedures and responsibilities of the dissertation sampling system, considering that the dissertation sampling system’s protection should continue from concept to program and main responsibility [4]. Some other scholars introduced the basic situation of the dissertation double-blind review system at Southeast University, and they further illustrated the significance of quality management, supervision and evaluation [5]. In summary, most researchers were still focused around the dissertation double-blind review system, and they did not focus on the fact that it was necessary to sample doctorate and master’s dissertations awarded last year. This paper proposes a degree-quality evaluation model based on doctorate dissertation

and master's dissertation samples, and this model has great potential to help realize the educational information evaluation system.

2. Degree Quality Evaluation Model

The quality evaluation model consists of a command management unit, data unit, sampling rules unit, index system unit, evaluation system unit and information feedback unit. The command management unit is the core of the whole model and is responsible for the organization and coordination of the evaluation work. The data unit is responsible for collecting data required for evaluation, data cleaning and integration. The sampling rules unit determines the rules based on sampling requirements. The index system unit formulates the corresponding evaluation index unit based on the data unit results. The evaluation system unit is the specific implementation process of evaluation where dissertations will be judged based on evaluation indicators made by the evaluation system unit. The information feedback unit is responsible for giving the command management unit and other correlative units feedback on the evaluation results.

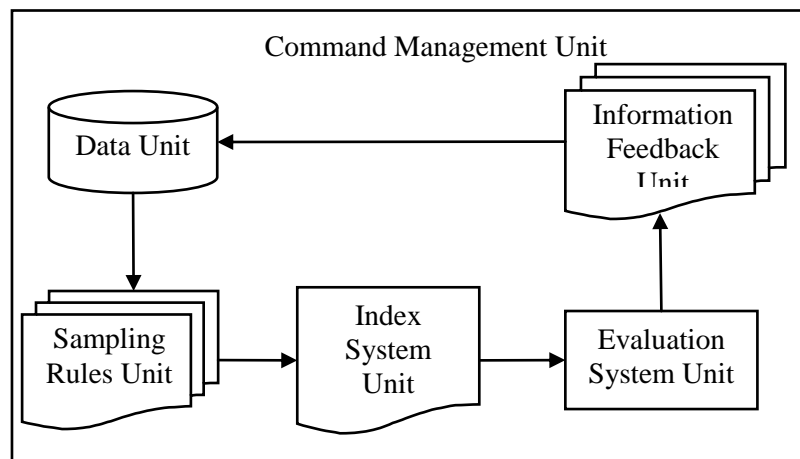


Figure 1. Degree Quality Evaluation Framework

Degree evaluation process:

Step 1: Collect degree-granting data, teachers' and students' data from each degree-granting unit, then execute data cleaning and data fusion on these data.

Step 2: Determine sampling rules based on sampling requirements, and then screen the sampling lists.

Step 3: Determine the index system and evaluation methods based on the evaluation objects.

Step 4: Automatically select experts for initial evaluation based on the subject type or the professional category of each dissertation. If the result of the initial evaluation is qualified, end the evaluation. If the result of the initial evaluation is unqualified, re-select experts to evaluate this dissertation. If the result of re-evaluation is qualified, this dissertation is qualified. If the result of re-evaluation is unqualified, this dissertation is unqualified.

Step 5: Analyze the result of the evaluation, and give the command management unit and other correlative units feedback.

2.1. Command Management Unit

The command management unit consists of the heads of universities, leaders of research institutes, people in charge of the degree management department and some relevant professionals. This unit is mainly responsible for making evaluation policy, grasping the direction of evaluation, coordinating the efforts of each evaluation unit, summarizing and analyzing evaluation results, determining trends from data and constructing better evaluation mechanisms.

2.2. Data Unit

Data collection is the basis of the degree evaluation. The reliability of evaluation data is directly related to the effectiveness of the evaluation results. To collect as much data as possible and ensure the scientific and comprehensiveness of data collection, a comprehensive information collection system must be established. And the way of data collection should also be diversified. There are many available methods such as questionnaires, seminar discussions, and degree databases.

The collected data may have many problems because of the diversity of data sources and acquisition methods. The inconsistent concept representation of the information may result in data inconsistencies. The degree-conferring data of each degree program lacks unified specification, which leads to inconsistent data formats. The degree-related data is duplicate-entered in different systems, leading to the data repeat problem. These above data problems need to be addressed with the techniques of data fusion and data cleaning. Data cleaning is the use of mathematical statistics and data mining techniques to detect and eliminate erroneous data, incomplete data and duplicate data from data sources. Data fusion is used to integrate the heterogeneous data sources into a uniform data collection. This achieves the purpose of improving the efficiency of data sharing by providing users with a unified access interface [6].

To implement data cleaning and data fusion, the structural features and characteristic differences of the data tables are described below.

2.2.1. Structural Features

The structural features of a data table consist of the names of data columns, the types of unit values, the distribution of unit values, the distribution of the length of unit values and the distribution of unit value symbols [7]. To facilitate computer processing and analysis, one column of a data table is represented by $\xi = (p, q, f, g, h)$, wherein, p represents the name of this data column, and q indicates the type of this unit value, including number type, string type, date type, *etc.* If the type of a unit value is not clear, the type of that unit value will be replaced by the string type. Then, f represents the distribution function of unit values. It is a discrete function. The domain of this function is a collection of all unit values appearing in this data column, and the range of this function is $[0,1]$. The length of the distribution function of the unit values is represented by g . The domain of g is the collection of all lengths of unit values appearing in this data column. The range of g is $[0,1]$, which indicates the proportion of a specific length value in its data column. In addition, h represents the symbol distribution function of unit values. The type of symbol can be divided into numbers, letters, double byte characters, separators and other characters. The domain of h is five symbol types. The range of h is $[0,1]$, which indicates the proportion of a specific symbol in its data column.

2.2.2. Characteristic Differences

To examine the difference in degree between the various properties, it is necessary to define the investigation method in advance and to quantify the degree of difference [8].

Based on the formal description of the column feature vectors, the differences in degree among various components is defined below.

a) Name difference

The names of two columns are P_{a_1} and P_{b_1} . The difference between P_{a_1} and P_{b_1} can be calculated by the following formula:

$$d(P_{a_1}, P_{b_1}) = \begin{cases} \frac{|length(P_{a_1}) - length(P_{b_1})|}{\max\{length(p_{a_1}), length(p_{b_1})\}} & P_{a_1} \subseteq P_{b_1} \text{ OR } P_{b_1} \subseteq P_{a_1} \\ 1 & \text{Other} \end{cases} \quad (1)$$

Wherein, $length(p_{a_1})$ and $length(p_{b_1})$ represent the length of P_{a_1} and P_{b_1} respectively.

b) Value type difference

Two data unit value types are q_{a_1} and q_{b_1} , and the value type difference between q_{a_1} and q_{b_1} is defined as $d(q_{a_1}, q_{b_1})$. The value type differences between various unit values are shown in the following table.

Table 1. Value Type Differences between Common Unit Value Types

Type	Text	Number	Date	Boolean
Text	0	2	3	3
Number	2	0	1	1
Date	3	1	0	4
Boolean	3	1	4	0

c) Value distribution difference

Two data unit value distribution functions are f_{a_1} and f_{b_1} , and the difference between f_{a_1} and f_{b_1} can be defined as

$$d(f_{a_1}, f_{b_1}) = \sum_{x \in D_f} |\Delta f(x)|. \quad (2)$$

Wherein, $D_{f_{a_1}}$ and $D_{f_{b_1}}$ are the domain of f_{a_1} and f_{b_1} respectively, $D_f = D_{f_{a_1}} \cup D_{f_{b_1}}$, $\Delta f = f_{a_1} - f_{b_1}$.

d) Value length diversity

Two data unit value length distribution functions are g_{a_1} and g_{b_1} respectively, and the difference between g_{a_1} and g_{b_1} can be defined as

$$d(g_{a_1}, g_{b_1}) = \sum_{x \in D_g} |\Delta g(x)|. \quad (3)$$

Wherein, $D_{g_{a_1}}$ and $D_{g_{b_1}}$ are the domain of g_{a_1} and g_{b_1} respectively, $D_g = D_{g_{a_1}} \cup D_{g_{b_1}}$, and $\Delta g = g_{a_1} - g_{b_1}$.

e) Symbol distribution diversity

Two data unit value symbol distribution functions are h_{a_1} and h_{b_1} respectively, and the difference between h_{a_1} and h_{b_1} can be defined as

$$d(h_{a_1}, h_{b_1}) = \sum_{x \in D_h} |\Delta h(x)|. \quad (4)$$

Wherein, $D_{h_{a_1}}$ and $D_{h_{b_1}}$ are the domain of h_{a_1} and h_{b_1} respectively, $D_h = D_{h_{a_1}} \cup D_{h_{b_1}}$, and $\Delta h = h_{a_1} - h_{b_1}$.

2.2.3. Characteristic Difference Vector and Feature Vector Distance

Based on the various components' differences as described above, the characteristic difference vector of two data columns can be defined as

$$\Delta\xi = (d(p_{a_1}, p_{b_1}), d(q_{a_1}, q_{b_1}), d(f_{a_1}, f_{b_1}), d(g_{a_1}, g_{b_1}), d(h_{a_1}, h_{b_1})). \quad (5)$$

The feature vector distance is defined as

$$d(\xi_{a_1}, \xi_{b_1}) = \sqrt{d^2(p_{a_1}, p_{b_1}) + d^2(q_{a_1}, q_{b_1}) + d^2(f_{a_1}, f_{b_1}) + d^2(g_{a_1}, g_{b_1}) + d^2(h_{a_1}, h_{b_1})} \quad (6)$$

2.2.4. Vector Distance Matrix and Structural Difference Matrix

Assuming data table A has n columns and data table B has m columns, there is a vector distance matrix D of size $n \times m$. The first row of matrix D represents the vector distance between the first column of data table B and each column of data table A . In the same way, the last row of matrix D represents the vector distance between the column m of data table B and each column of data table A .

$$D = \begin{bmatrix} d(\xi_{a_1}, \xi_{b_1}) & \cdots & d(\xi_{a_n}, \xi_{b_1}) \\ \vdots & \ddots & \vdots \\ d(\xi_{a_1}, \xi_{b_m}) & \cdots & d(\xi_{a_n}, \xi_{b_m}) \end{bmatrix}. \quad (7)$$

The structural difference matrix D_T can be obtained by the following steps. Set threshold value T of vector distance. If an element $d(\xi_{a_i}, \xi_{b_j})$ in vector distance matrix D is greater than the threshold value T , set the corresponding position element in the structural difference matrix D_T as a value of 0. If an element $d(\xi_{a_i}, \xi_{b_j})$ in vector distance matrix D is less than the threshold value T , set the corresponding position element in the structural difference matrix D_T as a value of 1.

2.2.5. Algorithm

The data reconstruction and fusion algorithm are proposed based on the structural difference matrix.

Algorithm 1. Data reconstruction and fusion algorithm

Step 1: Compare the number of rows and columns of structural difference matrix D_T . If the number of columns is greater than the number of rows, map properties of the data table corresponding to the columns of the differences' matrix structure D_T to the new reconstructed data table. If the number of rows is greater than the number of columns, map properties of the data table corresponding to the rows of the differences' matrix structure D_T to the new reconstructed data table.

Step 2: If the number of columns of matrix D_T is greater than the number of rows, extract those rows whose sum is equal to zero. If the number of rows of matrix D_T is greater than the number of columns, extract those columns whose sum is equal to zero.

Step 3: Add column names and their constraints corresponding to columns or rows extracted in the previous step to the new reconstructed data table.

Step 4: Import data from the original table to the new table based on the structure of the newly reconstructed table.

2.3. Sampling Rules Unit

Sampling is picking a set of the elements in a collection based on established rules. Sampling must satisfy the principle of coverage, randomness and equality. In general, sampling rules should be regarded as a compound rule, which can be broken down into several simple rules. Those simple rules, when executed in a certain sequence, should have the same effect as those simple rules executed as a compound rule.

2.3.1. Rule Definition

The formal definition of a rule can be expressed as a tuple composed of a subset and a selected method, and it can be defined as $t: (A, h)$. Its scope is the set A . And the set A is a subset of the complete set. The set A is called as the target collection or the target subset of the rule t . Then h represents the selection method of the rule t . Namely, under the action of the rule t , the method h can select elements from the subset A as the result of rule execution based on certain algorithms. When the system receives a rule, the rule will operate under the selection methods of this rule control. This process is a rule execution.

The adjacent implementation of rules can be expressed as $t_1: (A_1, h_1) \rightarrow t_2: (A_2, h_2)$. This indicates that the rule t_1 is executed on subset A_1 , then executed on subset t_2 . The rule can be abbreviated as $t_1 \rightarrow t_2$. And it can also be expressed as a compound rule $t_{12}: (A_1 \cup A_2, h_{12})$.

Definition 1 Probabilistic Rule A probabilistic rule is used to specify the probability of elements being pumped into a subset. Then, individuals in this subset are randomly selected as results of sampling based on this probability, while the number of individuals drawn cannot exceed a certain upper limit. This can be formally expressed as $t_p: (A, h_p(A, Q, r))$. Wherein, the upper limit is determined by $[|A| \times r]$.

Definition 2 Equilibrium Probability Rule An equilibrium probability rule is used to specify a target subset A , an element attribute p and a sampling probability r . The subset A will be divided based on the value of the element attribute p , then a probabilistic rule will be executed on each subset. This rule can be formally expressed as $t_B: (A, h_B(A, q, r, p))$, and

$$t_B: (A, h_B(A, q, r, p)) \equiv t_{p_1}: (A_1, h_p(A_1, q, r)) \rightarrow t_{p_2}: (A_2, h_p(A_2, q, r)) \rightarrow \dots \rightarrow t_{p_n}: (A_n, h_p(A_n, q, r)). \quad (8)$$

Definition 3 Coverage Rule A coverage rule is used to specify how at least one of the elements in a target subset should be drawn. It can be formally expressed as $t_C: (A, h_C(A, q))$. Based on whether there are elements in the sampling status $q: (T, R, S_0)$ to be drawn, the sampling method of this rule can be described as

$$h_C: (A, q) = \begin{cases} \emptyset & (\exists x)(x \in A \wedge x \in R) \\ \{Rand(A \cap S_0, 1)\} & (\forall x)(x \in A \rightarrow x \in R) \end{cases} \quad (9)$$

Definition 4 Attribute Value Coverage Rule An attribute value coverage rule is used to divide the target subset by a certain attribute value, then execute the coverage rule on each division. It can be formally expressed as $t_C: (A, h_C(A, q, p))$, in which, when p is an attribute of elements in subset A , the elements in A can be divided into n disjoint subsets in light of different values of attribute p , namely $A_1, A_2, \dots, A_n, n \in N^+$.

Definition 5 Required Rule A required rule is a special kind of rule. It is used in the situation when all elements in a target subset should be drawn. It can be formally expressed as $t_N: (A, h_N(A))$, wherein, $h_N(A) \equiv A$.

2.3.2. Rule Conflicts

Rules might cause different sampling behaviors based on the status of sampling. Taking $t_1 \rightarrow t_2$ for example, the status of sampling before the execution of the rule t_2 has been changed by the rule t_1 , but in $h_1(A_1) \cup h_2(A_2)$, the sampling method $h_2(A_2)$ does not depend on the status of sampling after the execution of $h_1(A_1)$. This leads to the sampling results of two rules executed in order not necessarily being equivalent to the two rules respectively executed. Namely, if $h_{12}(A_1 \cup A_2)$ is not necessarily equivalent to $h_1(A_1) \cup h_2(A_2)$, then there will be a rule conflict.

One must define rule priorities to handle rule conflicts. When there are rule conflicts, it is more reasonable to decide which rule to execute first or to give priority to a sampling condition based on the level of the rules' priorities. When a lower priority rule conflicts with a higher priority rule, selection parameters will indicate that the lower priority rule will be adjusted. If this adjustment doesn't make sense, the lower priority rule will be discarded.

2.4. Index System Unit

The index system unit is the core of the evaluation system and is directly related to the objectivity and effectiveness of the evaluation result. The index system unit needs to establish a scientific, reasonable and feasible evaluation index based on the data unit. This evaluation index should not only reflect the requirements of the State Council Academic Degree Committee on Graduate Education but also take the characteristics and advantages of various degree programs into account. The earlier method of indicator system establishment can be divided into an analysis method, Delphi method, synthesis method and indicator properties grouping method [9]. The early-established index system is usually able to achieve the comprehensive principle, but it is less independent between each indicator. And there is a phenomenon of index connotation overlapping [10]. Thus, it is necessary to filter early-established index systems and ultimately to determine the weight of each indicator.

2.5. Evaluation System Unit

The evaluation method is divided into two types, a percentile system and a hierarchical system [11]. The comparison between the percentile system and the hierarchical system is as follows.

Table 2. The Comparison between The Grading And Hundred-Mark Systems

	Hierarchical system	Percentile system
Division level	Vague	Exact
Discrimination	Low	High
Data analysis	Hard	Easy
Scope of application	Non-cognitive areas such as capacity, emotional <i>etc.</i>	Admission and selection examinations

2.6. Information Feedback System Unit

Data itself does not have any meaning, so it is necessary to transform data into knowledge by data analysis. How to make better use of data and implement value evaluation has become the focus of research in the evaluation field [12]. Degree evaluation results are just a kind of data, and they can generate value only through systematic data analysis and mining. Furthermore, this data can provide support to decision making in education and research work in each degree program.

3. Evaluation Model Application

To verify the scientific nature and effectiveness of the evaluation model, we applied the evaluation model to a sampling of master's dissertations from Shanghai in 2014, using the Shanghai Academic Degree and Graduate Education Information Platform as a carrier [13]. Taking data fusion and the establishment of sampling rules for example, the application of the evaluation model for the master's dissertation sampling of Shanghai in 2014 is introduced below.

3.1. Data Collection

In the year of 2014, 39,768 master's degree were awarded in Shanghai, including 19,188 academic degrees, 19,958 professional degrees and 622 equivalent degrees. Thirty-five programs were involved and 139 professional categories were covered. The format of degree-granting data collected by the data unit was as shown below.

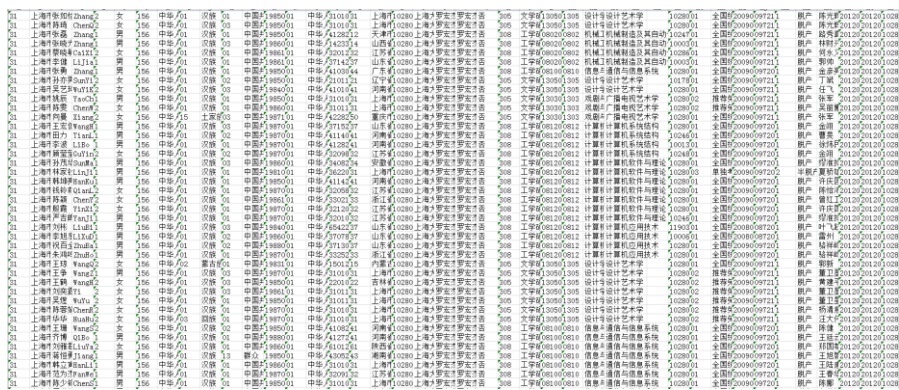


Figure 2. Academic Degree-Granting Data Table Partial Screenshot

The academic degree-granting data table contains 64 fields such as name, ID number, date of degree awarded, degree type, degree-granting unit, mentor's name, and major. The professional degree-granting data table contains 59 fields such as name, ID number, degree-granting data, degree type, degree-granting unit, mentor's name, and major. The equivalent degree-granting data table contains 58 fields such as name, ID number, date of degree awarded, etc. In the dissertation sampling work, there is another data table and a dissertation blind evaluation data table, which includes name, ID number, concealed evaluation result, mentor's name, the major and other information.

3.2. Data Fusion

First, we calculated the name difference, value type difference, value distribution difference, value length difference and symbol distribution difference between the columns of the academic degree-granting information table and the professional degree-granting data table. Then, we calculated the vector distance matrix and set the threshold as the median minimum value of each line, $T=0.326519$. Next, we calculated the structural difference matrix.

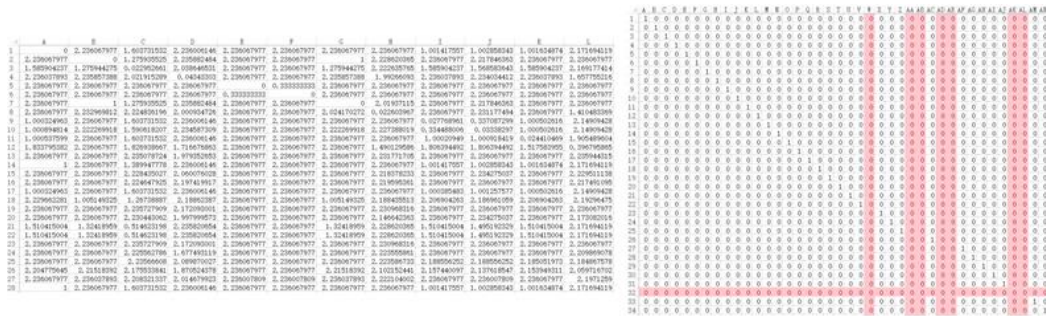


Figure 3. The Partial Screenshot of Experimental Result

Based on the structural difference matrix, data reconstruction and fusion were executed using algorithm 1. The thirty-second line of matrix D_T corresponds to the column named ZHHGBH in the professional degree-granting data table. Then, we mapped the data column named ZHHGBH and its constraints in the professional degree-granting data table to the new table. Because this field was not in the academic degree-granting data table, we filled the corresponding data column of this field with empty values in the new table. In addition, the value of the twenty-third, twenty-seventh, twenty-eighth, thirtieth, thirty-first, thirty-seventh and thirty-eighth columns indicated that the columns named YJXKSY, YJXKDM, YJXKM, ZSZYDM, ZSZYMC, XXFSM and XXFS in the academic degree-granting data table should also be processed. However, once the whole structure of the academic degree-granting data table had been mapped to the new table in the reconstruction process, it was not necessary to deal with them again.

After the reconstruction and fusion, it was necessary to conduct a reconstruction and fusion between the new table and the equivalent degree-granting data table. The final data table is shown in Table 3, in which the last data fields highlighted in bold were added after data fusion.

Table 3. Field Names and Constraints after Data Fusion

Field Name	Constraints	Field Name	Constraints	Field Name	Constraints
SSDM	int(20)	XXFSM	int(20)	HKSZSSM	int(20)
SSMC	varchar(255)	XXFS	varchar(255)	HKSZSS	varchar(255)
XM	varchar(255)	DSXM	varchar(255)	XWSYDWM	int(20)
XMPY	varchar(255)	BYNY	int(20)	XWSYDW	varchar(255)
XBM	int(20)	HXRWRQ	varchar(255)	XZXM	varchar(255)
XB	varchar(255)	XWZSBH	varchar(255)	ZXXM	varchar(255)
GBM	int(20)	LWTM	varchar(255)	YJXKSY	varchar(255)
GB	varchar(255)	LWGJC	varchar(255)	XWLBM	varchar(255)
MZM	int(20)	LWLXM	int(20)	XWLB	varchar(255)
MZ	varchar(255)	LWLX	varchar(255)	ZYDM	int(20)
ZZMMM	int(20)	LWXTLYM	int(20)	YJXKDM	varchar(255)
ZZMM	varchar(255)	LWXTLY	varchar(255)	YJXKMC	varchar(255)
CSRQ	int(20)	QZXWM	int(20)	ZYMC	varchar(255)
ZJLXM	varchar(255)	QZXW	varchar(255)	ZSZYDM	varchar(255)
ZJLX	varchar(255)	QZXLM	int(20)	ZSZYMC	varchar(255)
ZJHM	varchar(255)	HQZXWNY	int(20)	KSH	varchar(255)
KSFSM	int(20)	GZDWXZM	int(20)	BZ	varchar(255)
KSFS	varchar(255)	GZDWXZ	varchar(255)	ZHHGBH	varchar(255)
RXNY	int(20)	GZDWSSM	int(20)	SQXWNY	int(20)
XH	int(20)	GZDWSS	varchar(255)	SQXWXS LB	varchar(255)
QZXWDWM	int(20)	GZXZM	int(20)	ZCJBM	int(20)
QZXWDW	varchar(255)	GZXZ	varchar(255)	ZCJB	varchar(255)
QXM	int(20)	ZP	varchar(255)	ZWJBM	int(20)

QX	varchar(255)	ZPSTATE	varchar(255)	ZWJB	varchar(255)
----	--------------	---------	--------------	------	--------------

3.3. Data Cleaning

By comparing sampling requirements, those columns which were not associated with the sampling theme were deleted. Finally, all data columns that met sampling requirements and their meanings were shown in the following table.

Table 4. Field Names and Field Meanings after Data Cleaning

Field Name	Meaning	Field Name	Meaning
XM	Name	LWLX	Dissertation type
XB	Gender	LWXTLY	Source of topic selection
GB	Country	QZXW	Pre-degree
MZ	Nation	HQZXWNY	Pre-degree awarded date
ZZMM	Politics status	QZXWDW	Pre-degree unit
CSRQ	Birthday	XZXM	Principal name
ZJLX	Type of the certificate	XWLB	Degree type
ZJHM	ID number	XWSYDW	Degree-granting unit
HKSZSS	Registered permanent residence	XWSYDWM	Degree-granting unit code
XXFS	Learning style	ZYDM	Major code
DSXM	Mentor's name	YJXKDM	First level discipline code
BYNY	Graduation date	YJXKMC	First level discipline name
HXWRQ	Degree awarded date	ZYMC	Major name
XWZSBH	Degree certificate number	ZSZYDM	Own major code
LWTM	Dissertation topic	ZSZYMC	Own major name
LWGJC	Dissertation keywords	YJXKSY	First level discipline granting
RXNY	Enrollment date	KSFS	Examination method

a). Missing value processing

Because the account information of equivalent education masters was not recorded, the column named HKSZSS was filled with an empty value. And the column named XXFS of the equivalent degree was filled with an “equivalent to applying for master's degree” value. The column named RXNY of the equivalent degree was filled with the value of the column named SQXWNY in the original table. The column named BYNY of the equivalent degree was filled with the value of the column named HXWRQ in the original table. The columns named YJXKSY, ZSZYDM and ZSZYMC of the professional degree remain empty. The column named YJXKDM was filled with the corresponding top four of the discipline code. The column named YJXKMC was filled with the corresponding value of the column named YJXKDM in the subject code data table.

b). Repeated value processing

We detected the repetitiveness of the data table after data reconstruction. At first, we calculated the cosine similarity between each two records. Those records for which the cosine similarity was equal to one were repeated records. If all field values of two records were directly the same, we deleted one of them. If some field values in two records were the same, such as dissertation topic, dissertation keywords, name and ID number, those records would be sent as feedback to the command management unit. How to deal with those records should be decided by human judgment.

3.4. Sampling Rules

3.4.1. Rules Description

To the target subsets described clearly, the descriptions of elements' attributes were introduced. For an element α with an attribute p of value x , the element attribute can be described as $\alpha.p = x$.

Table 5. Element Attributes Required by the Master's Dissertation Sampling in Shanghai

Attributes	Meaning	Attributes	Meaning
p_1	Mentor's name	p_2	Degree-granting unit
p_3	Confidentiality period	p_4	Country
p_5	Enrollment date	p_6	Degree awarded data
p_7	Registered permanent residence	p_8	Major name
p_9	Municipal blind review score	p_{10}	Nation
p_{11}	Degree type		

Based on the requirements for master's dissertation sampling published by the Municipal Degree Committee [14], eleven sampling rules were sorted out, including three required rules, one must-not-be rule, four probabilistic rules, two equilibrium probability rules and three attribute value coverage rules.

The first required rule was $t_{N_1}: (A_1), A_1 = \{(\alpha \in S) \wedge (\alpha.P_1, \alpha.P_2, \alpha.P_8) \in B_1\}$, and it represented and picked those master's dissertations instructed by mentors who have not passed a doctoral dissertation in the past three years. Wherein, B_1 was a collection of mentors who have not passed a doctoral dissertation in the past three years. The second required rule was $t_{N_2}: (A_2), A_2 = \{\alpha | (\alpha \in S) \wedge (\alpha.p_3 \neq \epsilon) \wedge (\alpha.p_3 < d)\}$, and it represented and picked those master's dissertations that had applied for confidentiality within one-year after decryption. The third required rule was $t_{N_3}: (A_3), A_3 = \{\alpha | (\alpha \in S) \wedge (0 \leq \alpha.p_9 < 60)\}$, and it represented and picked those master's dissertations with a score of less than 60 in dissertation blind evaluation. Those three required rules can be combined into

$$t_N: (A_N), A_N = \{(\alpha \in S) \wedge (((\alpha.P_1, \alpha.P_2, \alpha.P_8) \in B_1) \vee ((\alpha.p_3 \neq \epsilon) \wedge (\alpha.p_3 < d)) \vee (0 \leq \alpha.p_9 < 60))\} \quad (10)$$

The must-not rule was $t_M: (A_4), A_4 = \{\alpha | (\alpha \in S) \wedge (\alpha.p_9 \geq 60)\}$, and it represented that those master's dissertations with a score greater than 60 in dissertation concealed evaluation were not picked.

The first probabilistic rule was $t_{P_1}: (A_5, h_P(A, q, 0.1)), A_5 = \{\alpha | (\alpha \in S) \wedge (\alpha.p_4 \neq \text{China})\}$, and it represented that ten percent of overseas master's dissertations were picked. The second probabilistic rule was $t_{P_2}: (A_6, h_P(A, q, 0.1)), A_6 = \{\alpha | (\alpha \in S) \wedge (\alpha.p_6 - \alpha.p_5 > 1460)\}$, and it represented that ten percent of those master's dissertations corresponding to master's degrees delayed more than one year were picked. The third probabilistic rule was $t_{P_3}: (A_7, h_P(A, q, 0.1)), A_7 = \{\alpha | (\alpha \in S) \wedge (\alpha.p_1 \in B_2)\}$, and it represented that ten percent of those dissertations instructed by those mentors who guided more than three postgraduates awarded master's degrees in the same year were picked.

The fourth probabilistic rule was as follows:

$$t_{P_4}: (A_8, h_P(A, q, 0.1)), A_8 = \{\alpha | (\alpha \in S) \wedge (\alpha.p_{11} = \text{part-time master})\} \quad (11)$$

It represented that ten percent of part-time master's dissertations were picked.

The first equilibrium probability rule was $t_{B_1}(S, \alpha, p_8, h_B)$, and it represented that at least one dissertation from each major was picked. The second equilibrium probability rule was $t_{B_2}(S, \alpha, p_2, h_B)$, and it represented that at least one dissertation from each unit was picked.

The first attribute value coverage rule was $t_{C_1}(S, \alpha, p_4, h_C)$, and it represented that at least one dissertation from each arisen country was picked. The second attribute value coverage rule was $t_{C_2}(S, \alpha, p_{10}, h_C)$, and it represented that at least one dissertation from each arisen nation was picked. The third attribute value coverage rule was $t_{C_3}(S, \alpha, p_7, h_C)$, and it represented that at least one dissertation from each arisen registered permanent residence was picked.

3.4.2. Sampling Rules Priority and Execution Order

In the dissertation sampling, fairness and unpredictability are very important. Rule execution should follow the "without replacement" principle, namely, already picked dissertations can no longer be canceled. In this principle, rules are executed from higher priority to lower priority.

The overall sampling rate for the master's dissertation sampling of Shanghai in 2014 was 5%. And rules execution should reflect the requirements of the dissertation sampling. In the list of rules, there was only one must-not-be rule, and it indicated that those dissertations that had passed the master's dissertation blind evaluation were not required to be evaluated again. Therefore, at the beginning of the sampling process, those dissertations should be excluded. That is, the priority of the must-not-be rule was the highest. And the purpose of the required rules focused on some dissertations that should be all picked. So, the required rules should be executed after the execution of any must-not-be rules.

In the rest of the rules, the probabilistic rules aimed at dissertation collections prone to having problems based on previous experience. So, the sampling rate for the probabilistic rules was higher than the overall rate. The sampling rate for the equilibrium probability rules was the same as the overall requirement, and the purpose of the equilibrium probability rules was to make up for the lack of sampling rate in a random way. So, the priority of the equilibrium probability rules was lower than that of the probabilistic rules. And the purpose of the attribute value coverage rules was to inspect dissertations distributed with different attribute values, thus acquiring an overall understanding of each dissertation subset divided by different attribute values. The larger the scope of the target subset involved before the attribute value coverage rules' execution, the more dissertations in the target subset of the attribute value coverage rules had been already picked. The attribute value coverage rules were executed last so that fewer additional operations were involved. Namely, the attribute value coverage rules had the lowest priority.

The sampling rules execution order was as follows:

$$t_M \rightarrow t_N \rightarrow t_{p_3} \rightarrow t_{p_2} \rightarrow t_{p_4} \rightarrow t_{p_1} \rightarrow t_{B_1} \rightarrow t_{B_2} \rightarrow t_{C_1} \rightarrow t_{C_2} \rightarrow t_{C_3}. \quad (12)$$

After the execution of the dissertation sampling, 2,010 dissertations had been picked. Of these, 35 degree-granting units and 117 majors were covered. The sampling result met the requirements of the Municipal Degree Committee.

4. Conclusion

This paper presents a new degree evaluation model that consists of a command management unit, data unit, sampling rules unit, index system unit, evaluation system unit and information feedback unit. To prove the scientific and effective nature of this evaluation system, this evaluation model was applied to a sampling of master's degree

dissertations from Shanghai in 2014. And taking data fusion and the establishment of sampling rules for example, the practical application of this evaluation system was described. This evaluation model can ensure the orderly progress of sampling work and can reduce the complexity of manual intervention. In addition, this evaluation model is significant to the construction of educational information platforms.

References

- [1] K. Xiangpei, "Practice and Thinking of Postgraduate Dissertation Sampling Review", Chinese High Education Evaluation, vol. 4, (2002), pp. 48-50.
- [2] M. Jie, "Analysis of Optimize Shanghai Postgraduate Dissertation Random Sampling System", Shanghai Education Evaluation Research, vol. 4, (2013), pp. 56-59.
- [3] H. Baoyin, X. Weiqing, Z. Yan and H. Tongliang, "Speed up the Establishment of a Sound Quality Guarantee and Supervision System of Academic Degrees and Graduate Education in China", pp. 3, (2014), pp. 1-9.
- [4] G. Lilan, "Research on Dissertation Sampling System in China", Xiangtan University, Hunan, (2011).
- [5] Y. Zhibiao, "Postgraduate Dissertation Quality Internal Management and External Supervision Practices-Take Southeast University Dissertation Double-blind Review and Dissertation Sampling in Jiangsu Province as Example", vol. 5, (2011), pp. 31-37.
- [6] C. Yueguo and W. Jingchun, "Survey of Data Integration", Computer Science, vol. 31, no. 5, (2004), pp. 48-51.
- [7] H. Xinrong, "The Semantics, Characteristics and Essence of Big Data", Journal of Changsha University of Science & Technology, vol. 6, (2015), pp. 5-11.
- [8] S. Zhou, "Probability Theory and Mathematical Statistics (Fourth Edition)", Higher Education Press, (2008).
- [9] L. Ren, "Research on the Establishment Method of Assessment Index System", Electronic Design Engineering, vol. 21, no. 1, (2013), pp. 34-36.
- [10] C. Yantai, "Classification and Research Progress on Evaluation Methods", Journal of Management Science, vol. 7, no. 2, (2004), pp. 69-79.
- [11] D. Wei, T. Zilong and X. Hua, "Comparative Analysis and Application Thinking of Percentile System and Hierarchical System", China Electric Power Education, vol. 32, (2013), pp. 63-64.
- [12] X. Jiaoxiong, X. Jun and W. Difeng, "Study on pretreatment of data clustering and its application", Shanghai University, Shanghai, (2007).
- [13] X. Jiaoxiong, Y. Xue and S. Jinlong, "The Conception of Provincial Academic Degree and Graduate Education Information Platform Construction", Academic Degree and Graduation Education, vol. 11, (2014), pp. 33-39.
- [14] "Shanghai Municipal Academic Degrees Committee", Shanghai Master's Dissertation Sampling Approaches, The 9th Degree in Shanghai, (2014).

Authors



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.

