The Design and Application of Statement Neutral Model Based on PCA

Yejun Zhu[1, a]

[1]School of Mechatronics Engineering, Harbin Institute of Technology, Harbin, 150001, China

[a]email: zhuyejun12@163.com

Keywords: Neutral Statement; PCA; Feature Extraction; Descriptive Variable Design

Abstract. In order to solve the management's difficult problem caused by the huge number and numerous styles in the existing statement, the paper puts forward a statement management method of feature extraction neutral statement model based on principal component analysis (PCA). The method designs the descriptive variable of standardization statement's structure and implements fully digitization. By dimension reduction process, the paper obtains statement's digital feature. And under the premise of the loss as little as possible, the paper makes the digital feature be reverted to neutral statement, and then forms neutral statement model.

Introduction

In fact, the enterprises implement decentralized management system for data and statement to save the data in the database, but save the statement model in the statement library for the application. In this management method, enterprises usually need to consume a large amount of storage space to meet storage conditions of all kinds of statement files, but a large number of statement files leads to the difficult in the management and design of statement. For solving the above practical problem, the paper proposes the management method of statement neutral model. Through extraction neutral of statement files in statement library, the paper achieves neutral statement model. With a small amount of data including most of statement's features in statement library, the paper provides some degree of support for the design and management of statement.

In statement neutral model the key link is feature extraction, whose purpose is to reduce the dimension of statement data. Extracting the most striking feature from the original data is the base of forming statement neutral file. At present feature extraction methods mainly are the extraction of statistical feature, geometric feature, motion feature, and frequency domain feature. The extraction method of geometric feature is using the structure feature of prior knowledge and statement to position the salient feature point of statement style, such as cell, merge, data block. But these feature points are easily influenced by the complex semantic to show the great geometric limitation. The extraction method of statistical feature mainly is principal component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA). PCA is a classical statistics algorithm. Its idea is under the premise of retaining information in the greatest degree to extract the feature component of low dimensional data from high dimensional space, and is an extraction method based on the global feature [1]. The main idea of LDA is through the transformation to achieve the minimization and maximization of inter-object distance, and to obtain the optimal projection direction for producing the best classification results, but there will be a small sample problem. ICA is making the data separated into the independent and non Gauss component linear combination, which is an extraction method based on local feature [2]. DCT and wavelet technology are the analysis method based on time-frequency transformation.

Preprocessing of Statement Data

The wide use of statement makes statement's application system is very complex. The needs for statement in all walks of life have many differences. Before the feature extraction of statement, we need to use a standardized structure description system to achieve the unstructured information description, whose purpose is to ensure the integrity of unstructured data translated into structured data. The paper first uses the ontic knowledge representation method to analyze the statement's

structure, whose purpose is to systematically describe the concept and relation of statement's structure. The process of ontology construction usually has the following 4 steps [3]:

(1) Delimiting the scope of ontology, this mainly refers to domain knowledge.

(2) Using natural language to define and express domain knowledge. Terms express the intention definition. Intention definition refers to use a limited number of attribute having an unbreakable connection with terms to define domain knowledge.

(3) Adopting a kind of formal language to make these definitions formalization [4]. First we define the concepts of terms in the design of domain knowledge to constitute the concept set of ontology system. According to the relationship between ontology concepts, we organize these ontology concepts to separate the levels, and then establish the classification system of ontology system.

(4) Assessing, verifying and forming a formal ontology system.

On this basis, the paper establishes the structuration description system of statement, which is as follow:

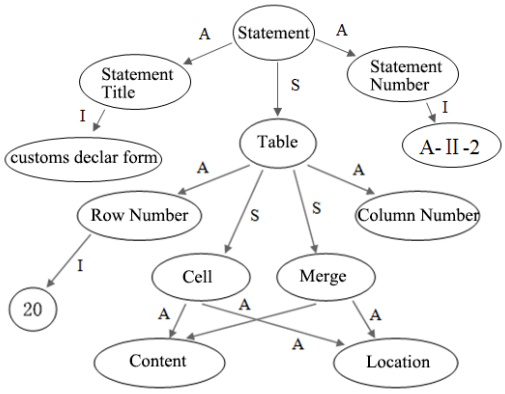

Fig.1. The schematic diagram of statement structure's ontology description

The structuration description of statement is the base of statement's descriptive variable design. Through the introduction, the geometric features of the structuration description statement cannot directly be used in the analysis method of statistical features, such as PCA.

Through the analysis of semantic quantification we extract three kinds of statement's basic attributes from ontology structuration description: statement size (the definition of row number and column number), the position and size of form (the position of cell is determined by two-dimensional array) and content attribute (header, reporting and data). The position and size data of form are merged into a basic property because of the special relationship between them: merge's position data contains size data, but size data does not contain position data. Containing three kinds of statement attributes the basic variable can be used as the vector dimension and data of PCA to participate in the neutral model algorithm. Considering the constant feature of vector dimension and the statistical feature of PCA algorithm, we draw the following statement description variables system:

First the paper makes statement in the test statement library be transformed into the matrix of m* n, which is a p-dimensional (m* n) vector. The statement's basic unit is a cell, and each cell's position is uniquely identified at the same time. Cell contains content and sideline. The content can express a basic attribute, and a sideline contains four lines of up, down, left, right. The existence of sideline also represents the existence of form line. The combination of four sidelines marks the position of merger. Because of their values are not the same, we first use the data of key features which is the data of sideline (form's position data) as our experimental data. The schematic diagram of statement description is as follows:

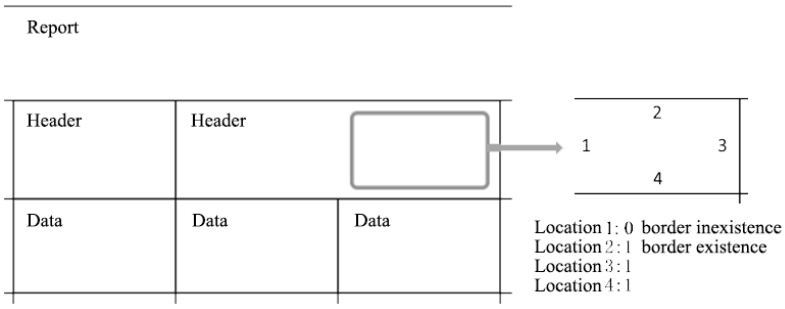

Fig.2. The schematic diagram of statement description variable

The Design of Statement Neutral Model

After determining the test statement library and statement description variables, we choose a certain number of statement samples in the test statement library as the training data set. In this paper the statement neutral model is designed based on the existing statement category of a highway project. The distinction of this statement category bases on the function and purpose of statement, and does not completely base on the style. This research can better manage and design the original statement in the style. Now the paper chooses the A-I type of statement as a sample library to experiment. The following is some sample statements in the library:

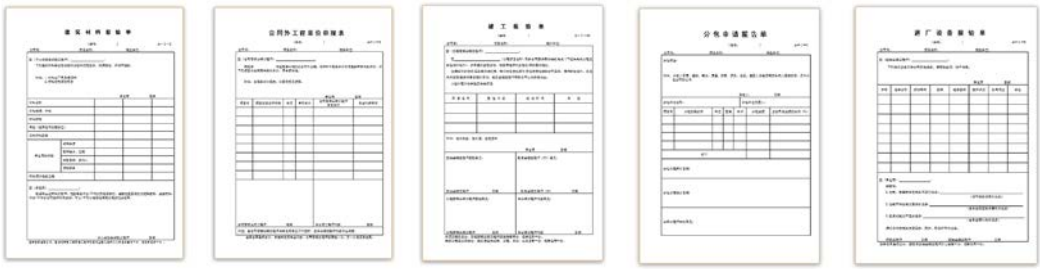

Fig.3. The sample figure of sample statement

The paper assumes the test statement library totals M statements. The data matrix of statement style is $A \in R^{mXn}$. The data matrix of each statement style is arranged with the rearrangement of m*n column vector to form a statement sample training set φ_i.

The first step of feature extraction algorithm is calculating the average of training samples chosen from test statement library. The formula is:

$$m_\varphi = \frac{1}{M}\sum_{i=1}^{M}\varphi_i \tag{1}$$

Restored in statement data, the average expresses the average table of sample training set, which are the common features of M statements. In order to well reflect the difference features between these training samples that is the intra-class features. We need to eliminate the common features of training samples. The general approach is to subtract the same features of the total statement data, such as we want to achieve the sample data of M statements' vector averaged:

$$x_i = \varphi_i - m_\varphi \tag{2}$$

The covariance matrix of matrix is calculated:

$$C_x = \frac{1}{M}XX^T \tag{3}$$

At this point we obtain the matrix. It is a P-order matrix and p = m * n. Then the characteristic value and the characteristic vector will be solved by us. Generally speaking, P is greater than M. From the following singular value decomposition theorem SVD, we can obtain:

The paper makes that A is a dimensional matrix and its rank is R. Then there are two orthogonal matrices and a diagonal matrix which make:

$$A = [a_1, a_2, \cdots, a_r] = U\Lambda V^T \quad (4)$$

We can transform it into the characteristic value and the characteristic v_p vector of another M-order matrix S, so the amount of calculation is greatly reduced. The matrix S is:

$$S_{ij} = X_i^T X_j \ (i, j = 1, \cdots, M) \quad (5)$$

After calculating the characteristic vector v_p of S matrix, based on the singular value decomposition theorem we can obtain:

$$A = U\Lambda^{1/2}V^T = \sum_{p=1}^{r} \lambda_p^{1/2} \mu_{pi} v_{pi}^T \quad (6)$$

The paper can calculate the characteristic vector μ_p.

Through PCA, the paper solves the characteristic value and characteristic vector of the training samples' total population scatter matrix. In general, transformed by the above processes any statements can be expressed as the form of the linear combination, whose corresponding weighting coefficients satisfy the coefficients of PCA [5][6].

The statement neutral model is established on the basis of the classification of sample library, so it contains two characteristics, which are the intra-class characteristic and the inter-class characteristic. The former is reflected in the variance of the total population scatter matrix, and the PCA processing aims at it. The latter is the difference between the different categories of statements. Statement neutral model proposed in this paper is convenient to simplify the statement design and management, and the emphasis point is intra-class characteristic. The characteristic table contains all intra-class characteristics, while the average table contains part of intra-class characteristics and a large number of inter-class characteristic [7]. The characteristic table can independently show the characteristics, but it lacks the common inter-class characteristics in the category. On the basis, the paper puts forward the generative process of report of statement neutral file, which is as follows:

1. Through the normalization of experimental data, the paper obtains a certain number of characteristic tables of the sample library (characteristic matrix u_l, $l = 1,2,\cdots,k$) and an average table (co-average matrix m_φ);

2. Through the following processing of the characteristic table u_l, the paper obtains the characteristic general table u^* containing all intra-class characteristics. For each element u_{ij}^* from u^*, there is:

$$u_{ij}^* = \begin{cases} 1 & \sum_{l=1}^{k} u_{lij} \neq 0; \\ 0 & \sum_{l=1}^{k} u_{lij} = 0; \end{cases} \quad (7)$$

3. In the average table m_φ eliminating the intra-class characteristics which is the part of characteristic in the characteristic general table u^*, the paper can get the basic table b, which is:

$$\mathrm{b}_{ij} = \begin{cases} 1 & m_{\varphi ij} = 0 \vee m_{\varphi ij} - u_{ij}^* = 0; \\ 0 & m_{\varphi ij} \neq 0 \wedge m_{\varphi ij} - u_{ij}^* \neq 0; \end{cases} \quad (8)$$

4. The basic table b completely does not contain the intra-class characteristics. It only shows the inter-class characteristic of statement category, so it is called the base table. Combining the base table b and characteristic table u_l, the paper gets a series of neutral statement table z_l, which is the neutral statement model of this statement category. There is:

$$z_{lij} = \begin{cases} 1 & b_{ij} \neq 0 \vee u_{lij} \neq 0; \\ 0 & b_{ij} = 0 \wedge u_{lij} = 0; \end{cases} \qquad l = 1,2,\cdots,k \quad (9)$$

Application Examples and Analysis

According to the original structure of the sample, the paper transforms it into the cell statement of 20 * 10 be standardized based on the proportion, and does not change the style. According to the statement variable description system, the paper transforms it into the 800-dimension(20 * 10 * 4) vector and the sample totals 9 tables, which is the experimental data is 9 * 800 matrix.

The following figure is the average table and the base table of sample library achieved by the algorithm. The latter only contains the inter-class characteristic of this statement category and is the basis of neutral file.

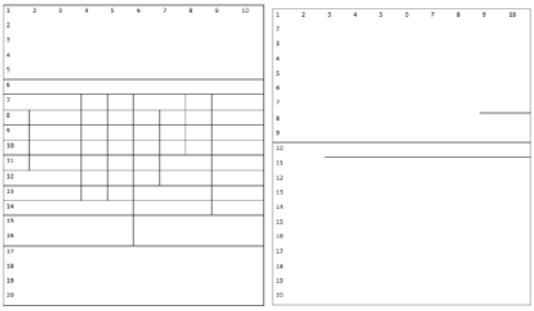

Fig.4. The average table and the base table of sample library

The following table is the characteristic value's information of the total scatter matrix, which totals 8 nonzero characteristic values. Through the normalization of characteristic vectors the paper gets the following 8 characteristic tables:

Tab.1. The characteristic value and contribution rate of statement data

	Eighth pivot element	Seventh pivot element	Sixth pivot element	Fifth pivot element	Fourth pivot element	Third pivot element	Second pivot element	First pivot element
Characteristic value	3.0133	3.7855	7.2036	9.0724	10.6094	14.5599	20.1176	28.3606
Contribution rate	0.0312	0.0391	0.0745	0.0938	0.1097	0.1505	0.2080	0.2932
Accumulative contribution rate	1.0000	0.9688	0.9297	0.8552	0.7614	0.6517	0.5012	0.2932

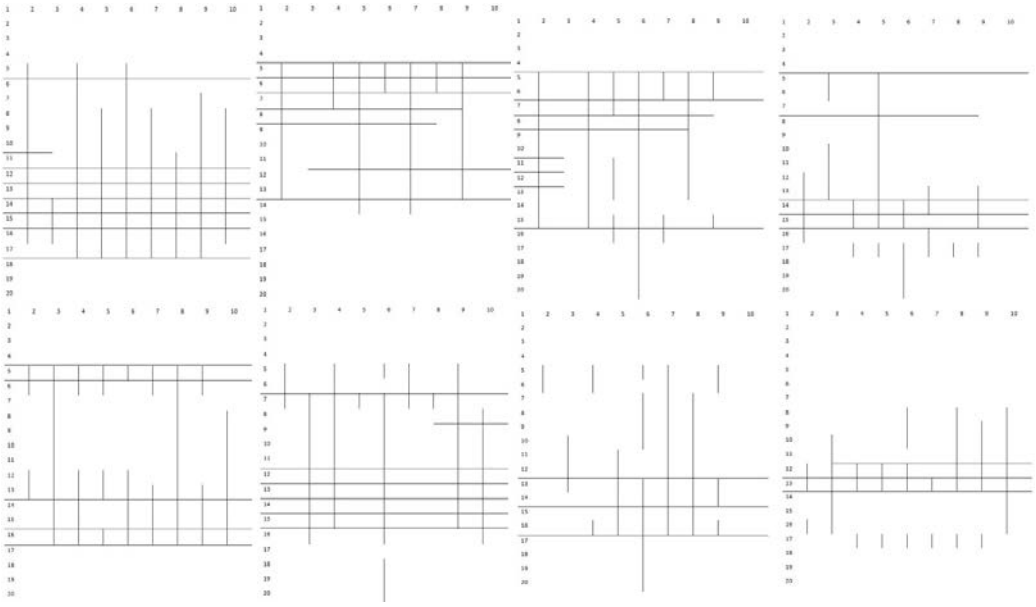

Fig.5. The statement sample library's characteristic table one to eight

This kind of statement neutral model is shown by the following figure, which chooses the number of pivot elements whose contribution rates reach 85%. Formed by the base table and the characteristic table, 5 neutral statement files are as follow:

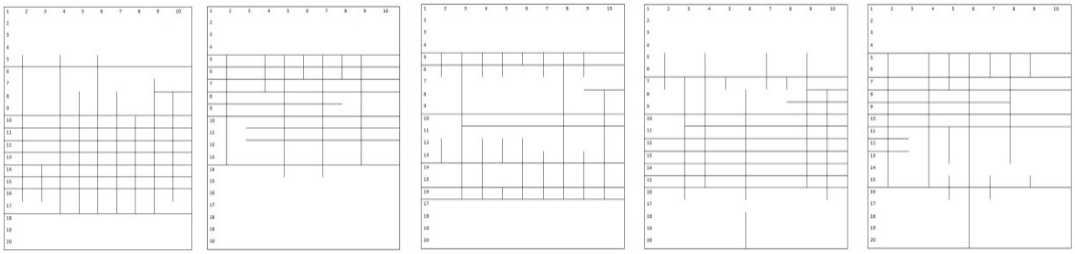

Fig.6. The neutral statement file set of statement sample library

Conclusion

The statement is an information organization system containing the style and significance. And the processing objects of statement neutral model algorithm based on PCA are statement's statistical data. The statement neutral model realizes the feature extraction of statement's style and data. The paper focuses on the design and realization process of statement neutral model. The results can also be used to analyze other statements' features, such as the pattern recognition of statement classification.

Acknowledgement

At the point of finishing this paper, I'd like to express my sincere thanks to all those who have lent me hands in the course of my writing this paper. I'd like to take this opportunity to show my sincere gratitude to my supervisor, Ms.Lin and express my gratitude to my classmates who offered me references and information on time.

References

[1] Carreira-Ferpian M A.A Review of Dimension Redcuction Techniques Technical Report C5-96-09[R]. England, Sheffield: Department of Computer Science University of Sheffield, 1997.

[2] Wu Xiaoting, Yan Deqin. Research and analysis of data dimension reduction method [J]. Computer Application Research

[3] Sun Yu,Yuefei Sui.The Ontology Revision[C].International Joint Conference on Artificial Intelligence (IJCAI2005), 2005, (4):1583-1584.

[4] Lin Zefei. The overview of theoretical research on ontology conceptual model building [J]. Information Research

[5] Yu Chenglong. Feature selection algorithm based on PCA [J]. Computer Technology and Development

[6] Kirby M,Sirovich L.Application of the Karhunen –Loeve Procedure for the Characterization of Human Faces [J].IEEE Trans. Pattern Analysis and Machine Intelligence,1990,12(1):103-108.

[7] Sirovich L,Kirby M.Low -dimensional Procedure for the Characterization of Human Faces[J].Journal of the Optical Society of America,1987,4(3):519-524.