Research on Mining and Validation Method of Policy Genome Based on the State Space Reduction

Gang Liu[1, 2, a, *], Tengfei Li[1, b], Wangyang Liu[2, c], Yang Cao[2, d], Yudan Du[1, e]

[1]Harbin Engineering University, Harbin 150001, China

[2]CETC Big Data Research Institute Company Limited, Guiyang 550000, China

[a]liugang@hrbeu.edu.cn, [b]1397159025@qq.com, [c]liuwangyang@cetcbigdata.com, [d]caoyang@cetcbigdata.com, [e]760483155@qq.com

*Corresponding author

Keywords: Domain Policy, State Space, Automatic Summary, Policy Genome

Abstract: The analysis of the policy text are drawing more and more attention in the domain of policy research, becoming an urgent and important problem to be solved in the progress of legal institution in our nation. This thesis puts forward a method to mine and validate the policy genome by the reduction of the state space. Firstly this method preprocesses policy text by the natural language processing technology, and sets up the reasonable expression dimension of policy text by automated-summary-based state space reduction method. Then, based on the state space reduction of policy text, this thesis introduces the concept of policy lineage, and carries out the relevant definition and acquisition of the policy genome by the combination of the nature of biological gene in genetics. Finally, this thesis uses the dominant genes of the policy to calculate of text similarity. When the difference between dominant-gene-based policy text similarity data and traditional policy text similarity data exceeds a certain threshold, then policy latent gene will be mined, and this thesis puts policy dominant gene and latent gene together as a part of policy genome, thus to achieve the goal to calculate the similarity with policy genome instead of policy text.

1. Introduction

The constant derivation and differentiation of the text have resulted in the inextricable link between the texts. This topic introduces the concept of "text consanguinity", it also explains the relationship between the text and the nature of the text. It is an innovative attempt to describe the relationship of text by blood and gene and explain the cause of implied text phenomenon. The related results have not been published in the field of biological genetics. Therefore, the exploration and confirmation of the text genome can effectively analyze the various relationships between texts, and the hidden consanguinity of text can not only serve as an important basis for the analysis of text consistency, but also can be used for reference in other related fields.

2. Relevant work and theoretical basis

This chapter mainly introduces the related work and theoretical basis involved in the process of text consistency research, including the theory of text consanguinity related theory and the related theory of text feature dimension reduction.

2.1 Text consanguinity and text genes

This paper combines the genetics of blood relationship in genetics to put forward the theory of "text consanguinity", which divides the text blood relationship into text explicit blood relationship and text recessive blood relationship, and proposes a complete text genome mining and confirmation method.

Text blood relationship is divided into text explicit blood relationship and text recessive blood relationship, the text dominant blood relationship is the textual blood relationship of the text lineage tree, and the text recessive blood relationship is the blood relationship of the implicit non-lineage representation between the texts.

The constant derived differentiation of text will cause text inconsistent phenomenon and produce text recessive blood relationship, therefore, the mining of text recessive blood relationship is an important problem in text analysis.

In the field of text research, genes correspond to concepts. In the genetic process, the text dominant gene has a certain probability change, which corresponds to the extension of the concept, and the text recessive gene remains unchanged, which mainly corresponds to the connotation of the concept. Therefore, the mining of text recessive genes reflects the "consistency" in the text derivative process through the study of the concept connotation, that is, the hidden blood relationship between the texts. Therefore, in the study of text consistency analysis, the mining of text recessive genes is important.

2.2 Text feature selection and text feature reduction

The process of selecting the best feature subset is called Feature Selection (FS). In this process, the original feature set is F, F' is the selected feature subset, the selected feature subset must be consistent with the original feature set, ie $F' \in F$. Feature selection greatly improves the execution speed of text mining algorithms, and also greatly reduces the memory space occupied by the algorithm.

There are two main methods for text feature dimensionality reduction based on vector space: one is the dimensionality reduction method based on text feature matrix. The other method is feature selection dimension reduction. The research work in this paper is mainly based on text feature selection.

3. State space reduction method based on automatic summary

The technical route and workflow of this paper are shown in Fig. 1, which mainly includes five stages. In the first stage, the traditional vector-based text similarity calculation algorithm is used to calculate similarity data between texts. In the second stage, the dominant gene is extracted from the text. The gene has strong ability to express the text and can fully and accurately express the subject of the text. In the third stage, the similarity between the genomes is calculated using the extracted text dominant genes. In the fourth stage, the traditional vector space-based text similarity calculation data and the dominant gene-based text similarity calculation data are compared and analyzed. In the fifth stage, we excavate the recessive genetic genes of the text, take the recessive genes as part of the genome, and calculate the text similarity iteratively. Finally, we can replace the traditional text similarity calculation with the genome similarity calculation, so as to realize the next step of the text analysis research.

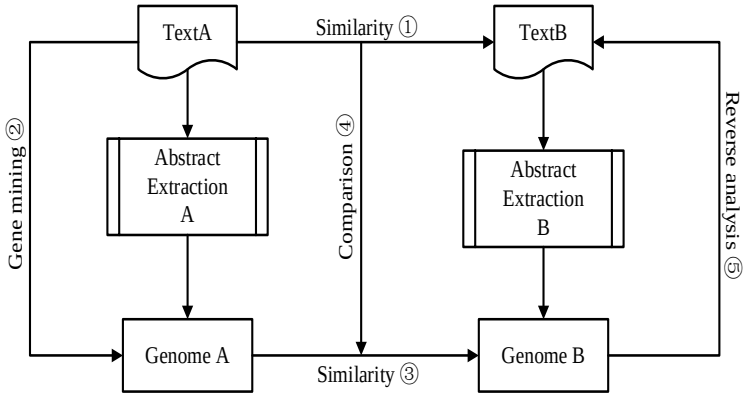

Figure 1. Technology roadmap

3.1 Text preprocessing

The commonly used text representation models include: Vector Space Model (VSM) and Boolean Model (BM). This paper uses the VSM model.

The preprocessing stage processes the text into a text feature matrix as a data training set for text analysis. The specific steps are shown in Fig. 2.

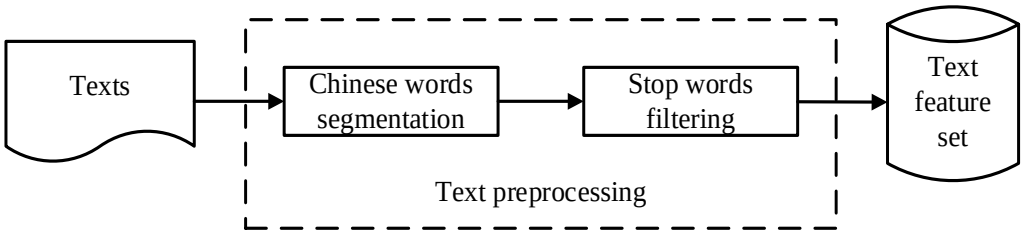

Figure 2. The general process of text preprocessing

3.2 Weight calculation of text feature words

In this paper, we use the method of calculating the weight of feature words based on text summary. The most important feature of this method is that the processing of feature words is easy to operate and calculate. The main basic process of text feature word weight is shown in Fig. 3.

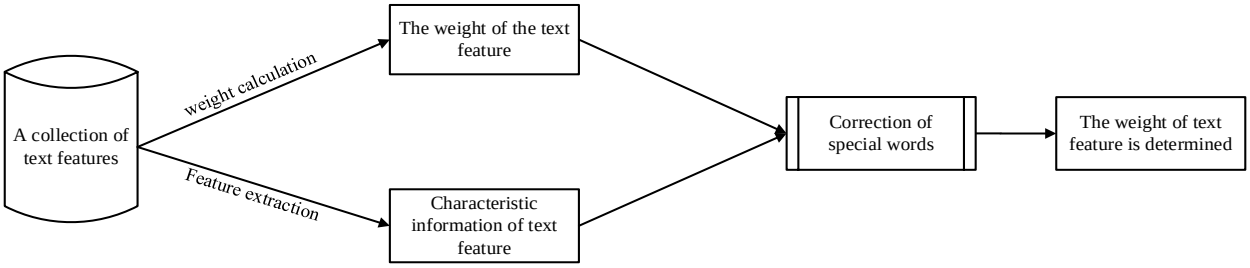

Figure 3. Basic process of weight calculation of text feature words

(1) TF normalization algorithm based on word frequency

The formula for calculating TF-IDF value of feature words is shown in Formula 1.

$$ntf_{k,D} = \alpha \times (a + (1-a) \times \frac{tf_{k,D}}{tf_{\max}(D)}) \tag{1}$$

In formula (1), $tf_{k,d}$ represents the word frequency of the feature word k in the text D, $tf_{max}(D)$ indicates the word frequency of the most frequently occurring word. Among them, both α and a are regulators of the formula. The study shows that the effect of taking α as 7 and a as 0.4 is better.

(2) The weight algorithm based on the location of feature words

The calculation of feature word weights should consider the following aspects: text title(T), the subtitle(S), the first sentence(pf) and the end sentence(pl). Among them, the key words in T, S, pf and pl have different important coefficients of C_t, C_s, C_{pf} and C_{pl} respectively. And $C_t > C_s > C_{pf} > C_{pl}$. It reflects the contribution of feature words to text topic content in different positions of text.

1) Sentence Title: A title vocabulary is a vocabulary in a text such as a title, and a subtitle. Formula (2) is used to calculate the weight of valid title words.

$$W_t(S_i) = \begin{cases} 1 & S_i \text{ Contains valid title words} \\ 0 & \text{others} \end{cases} \tag{2}$$

2) Position of the sentence: In the Chinese text, the first sentence and the last sentence of the paragraph play a more important role. Therefore, for the sentence S_i, according to the position of the

sentence where the vocabulary is located and the position of the paragraph where the sentence is located, you can get the weight calculation formula (3) of vocabulary.

$$W_p(S_i) = \begin{cases} 1 & S_i \text{ is the first sentence of the paragraph} \\ 0.5 & S_i \text{ is the paragraph ending sentence or the second sentence} \\ 0 & \text{others} \end{cases} \tag{3}$$

By analyzing the different positions of sentences and the correlation between words and text titles, we can finally measure the feature information of sentences and words in the text comprehensively and accurately, and combine the weight of feature information to calculate the weight of feature words. Therefore, based on the characteristic information of the sentence itself, we can get the formula (4) for calculating the weight of the vocabulary.

$$FS(S_i) = \delta * (\beta * W_t(S_i) + \chi * W_p(S_i)) \tag{4}$$

In formula (4), δ, β and χ are the regulatory factors of the formula, among which: $\beta + \chi = 1$.

(3) Linear weight weighting algorithm

This paper is based on word frequency of TF normalization algorithm and weighting algorithm based on the key position, this paper proposes a comprehensive measure of text feature weight linear weight weighted algorithm (LWA), calculating formula for formula (5).

$$W_k(S_i) = ntf_{k,D} + FS(S_i) \tag{5}$$

Where $W_k(S_i)$ represents the comprehensive weight value of k in sentence S_i, and $k \in S_i$, $S_i \in D$. In formula (1) and formula (4), the setting of α and δ regulatory factors is to ensure that $ntf_{k,d}$ and $FS(S_i)$ are on the same order of magnitude, and more closely related to the artificial definition of the weight of feature words by experts in the field.

Compared with the traditional method of using word frequency to measure the weight of key, in this paper, the linear weighted algorithm (LWA) will be the absolute frequency value of normalized transformation, combined with the key position information, comprehensive relatively accurate to measure key weight, avoid text content in the training set of variable length of text, and cause of key weight calculation of interference by the length of the text. Therefore, the method of this paper is more advantageous, and it can be seen from the analysis that the time complexity of LWA algorithm is $O(n)$, that it is feasible for large-scale text.

3.3 Standardization of text feature words

Definition 1 (correlation degree) the number of synonym of the feature word w in the text.

Definition 2 (influence degree) the influence of feature words on text is defined by formula (6).

$$Influence(W) = weight \times correlation \tag{6}$$

In formula (6), *weight* is the weight value of the feature word w in the text, *correlation* is the relevance degree of the feature word w in the text.

Compared with the traditional method of representing the text as the vector matrix corresponding to the number of feature words and the number of texts, the state space reduction method based on automatic abstraction reduces the feature word vector dimension of a text by 1-P. Efficiency is a big improvement. After repeated experiments, the value of P in this paper is: 0.8.

4. The excavation and confirmation of the text genome

4.1 Text genome

Definition 4 (text dominant gene) determining the feature words of the text topic.

Definition 5 (text recessive gene) feature words that have a certain effect on text expression.

Definition 6 (text genome) the text dominant gene and the text recessive gene.

Formula (7) is used to verify whether the feature word w is a recessive gene.

$$Evaluation(w) = \frac{c(w)}{f(w)} \tag{7}$$

Among them, $c(w)$ is the contribution of the feature words W to the text topic expression. In the follow-up study, the influence of the word on the similarity of text is used as the measurement

standard. $f(w)$ stands for the frequency of characteristic word W in the text. The higher the contribution of W, the more important it is to the text, and the recessive gene is to be used as the text. On the contrary, it means that the feature is not important to the text, so it is abandoned.

4.2 The acquisition of the text genome

On the basis of the definition of the text genome, the principle of selection of text dominant genes is: firstly, language units that include more semantic information and can better represent text should be selected. Secondly, dominant genes need to satisfy the statistical regularity of text distribution on them. Thirdly, the selection process of dominant genes should be relatively easy to achieve, and its space and time complexity are relatively small. Combining with the principle of selecting dominant genes in text, this paper argues that the feature set processed by the state space reduction method based on automatic summary filters out a lot of "noise data" and can express the theme of text comprehensively and accurately. Therefore, this part of the feature set is used as the dominant gene of the text to calculate the text similarity for the first time.

Then, the text similarity data based on text dominant genes are analyzed. If the value fits well with the traditional text similarity value, the text dominant gene is considered to be complete and can replace the text content for consistency analysis; If the value is much lower than the traditional text similarity value, the recessive gene mining is carried out and been taken as part of the text genome to participate in the text similarity calculation, the text genome is used to replace the text content for consistency analysis.

Through the previous research, this paper establishes vector coordinates for the genomes involved in text similarity calculation, in which each text is represented by a vector, and each feature word in the text is equivalent to one dimension in the vector space, each The coordinate value of the feature word in the corresponding dimension is the weight value of the feature word, that is, each text can be represented as $D(w_1, w_2, ..., w_n)$, This paper uses the cosine of the angle between the text vectors to measure the similarity between texts. The similarity calculation formula of two texts is formula (8).

$$sim(D_1, D_2) = \cos\theta = \frac{\sum_{i=1}^{n}(w_{1i} \times w_{2i})}{\sqrt{(\sum_{i=1}^{n} w_{1i}^2)(\sum_{i=1}^{n} w_{2i}^2)}} \tag{8}$$

The process of text similarity calculation based on text dominant genes is described as follows:

(1) On the basis of state space reduction, the vector representation of text genome is constructed, the coordinate value of genome in the corresponding dimension is the importance of gene.

(2) Traversing text D_1 and D_2, if there is a common gene in the text, then multiply the important degree of both, as the molecular part of formula (8).

(3) Calculate the mold of text D_1 and D_2 separately, and use it as the denominator of (8).

(4) Calculate the similarity $sim(D_1, D_2)$ of two texts according to formula (8).

This paper proposes a method to extract dominant genes of texts based on automatic abstraction-based state space reduction method, reduce the number of feature words involved in text similarity calculation, and only perform recessive gene mining for texts with poor fitting degree. Filter out the impact of "noise data" on similarity calculations.

4.3 Text recessive gene mining

In this paper, the reverse bloodline tracking method based on similarity data (RBTA) is used to mine political text recessive genes. The main process is described as follows:

(1) If the similarity data sim_1 is lower than the traditional vector space-based text similarity calculation data sim_2 0.05, text recessive gene mining is performed.

(2) To obtain the feature words of the 1-P part filtered by the state space reduction, and sort according to the importance degree.

(3) Take one feature word at a time, using the formula (9) computing text similarity data sim_2, take text-based dominant gene data sim_1 text similarity, difference of the two as $c(w)$ values.

(4) Obtaining the word frequency $f(w)$ of the characteristic word, calculate the evaluation value of the feature words by formula (7), repeat the process (3), until calculate all the evaluation value.

(5) Sort by the evaluation value of the feature word, and obtain the characteristic words with the evaluation value greater than K as the text recessive gene, and reject the unqualified feature words.

(6) The text dominant gene and the extracted text recessive gene are jointly involved in the text similarity calculation to obtain the similarity value. If the value fits well with the traditional text similarity value, the text recessive gene is found and the process ends.

In the process of text recessive gene mining, considering the accuracy and efficiency of text similarity calculation, this paper believes that if the text similarity data based on dominant genes is 0.05 less than the traditional text similarity data, then the text recessive gene mining. In the process of calculating the evaluation value of the feature word, if the evaluation value of the feature word is greater than 0, the feature word is considered to be a recessive gene, otherwise, the feature word is discarded.

5. Experimental results verification and its application

This chapter mainly carries out relevant experimental verification from the reduction of policy text state space and the mining and establishment of policy text genomes.

5.1 Comparative analysis of results

In order to evaluate the effectiveness of the linear weighted weighting method proposed in the paper, 12 samples were randomly selected from the policy text samples for detailed analysis. First, multiple domain experts are required to manually evaluate the random sample. For the obtained evaluation data, the lowest and highest two data are removed, and then averaged, and the average value is used as the final weight value of the domain expert evaluation. Then, the weights of the random samples are obtained by using the linear weighted weighting method and the classical word frequency method proposed in this paper. Finally, the effectiveness and accuracy of the improved algorithm proposed in this paper are verified by comparing the results of the above three methods. The partial calculation results of the weight of feature words are shown in table 1.

Table.1. Part of the calculation result table of the weight of policy text feature words

	Call for bids	Bid	Activity	Project	The State Council	According to law	Purchase	Government	Cargo	Serve	Formulate
methods	10	8.88	7.1	6.82	5.7	5.42	10	6.62	4.93	4.84	4.57
Classical method [6]	15	11	10	9	5	4	47	26	7	6	3
Artificial assessment	10	9	8	7	6	5	10	8	5	5	5

5.2 State space reduction and dominant gene establishment

Taking the first chapter of Tender and Bidding Law and Government Procurement Law as examples, Table 2 illustrates in detail the similarity values of policy texts and the number of genes under different reduction thresholds P. From Table 2, it can be seen that if the reduction threshold is too small ($P \leq 0.6$), although the number of genes involved in similarity calculation can be greatly reduced and the efficiency of similarity calculation of policy texts can be improved, the accuracy of similarity is greatly affected; when the threshold $P = 0.8$, with the increase of threshold P, the similarity value of policy texts increases very slowly. Therefore, in order to compromise the efficiency and accuracy of policy text similarity, the threshold of state space reduction is P=0.8.

Table.2. The similarity and gene number of Bid 1 and Purchase 1 in different thresholds

Numble	*P*=1	*P*=0.9	*P*=0.8	*P*=0.7	*P*=0.6	*P*=0.5
Similarity	0.499	0.450	0.454	0.421	0.004	0.003
Number of genes in Bid 1	149	128	111	96	82	72
Number of genes in Purchase 1	298	254	226	194	165	141
Total number of genes	447	382	337	290	247	213

5.3 Policy text similarity analysis and recessive gene mining

Experimental data analysis at this stage includes the similarity calculation data of the policy text based on the dominant gene, the comparative analysis of the policy text similarity data under the two methods, and the data mining of the recessive gene of the policy text.

(1) Comparison and analysis of similarity data

This paper argues that if the error between the two data is within 0.05, the dominant genes extracted from policy text are considered to be complete; if the error between the two data is large, it is necessary to mine the recessive gene of the policy text. Comparative analysis of dominant gene-based policy text similarity calculation data and traditional vector space-based policy text similarity calculation data as shown in Fig.4, the basis for selecting data in the table is that the similarity value is greater than 0.1, and “Bid X Purchasing Y” means Chapter X in the Tendering and Bidding Law and Chapter Y in the Government Procurement Law.

It can be seen from the data in Fig.4 that the method proposed in this paper has a high degree of fit with the traditional method for calculating the similarity of policy texts. To some extent, the method of this paper is generally superior to the traditional policy text similarity method based on vector space in two aspects of similarity accuracy and computational efficiency.

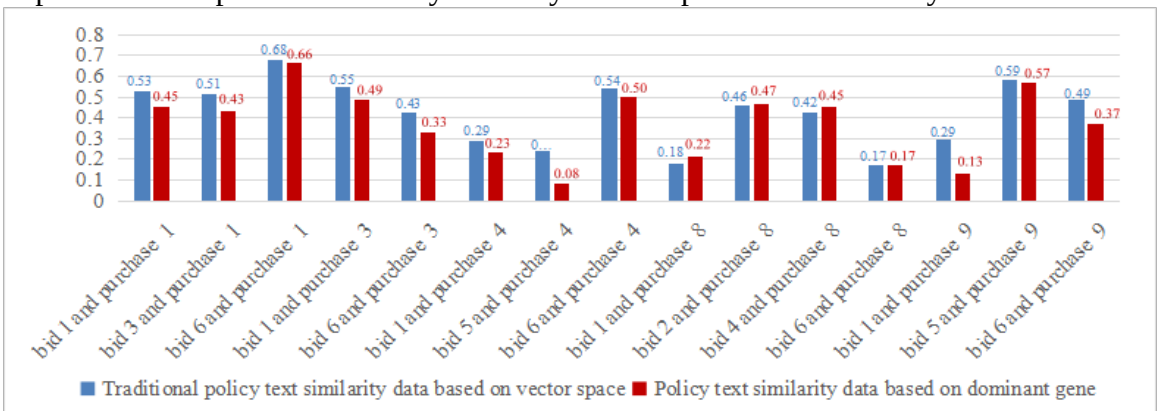

Figure 4. Comparison of two policy text similarity calculation

It can be concluded from Fig. 4 that compared with the conventional method, the degree of similarity data fitting of "bid 1 purchase 1", "bid 3 purchase 1", "bid 1 purchase 3", "bid 6 purchase 3", "bid 1 purchase 4","5 purchase 4”, “bid 1 purchase 9” and “bid 6 purchase 9” is not ideal, so it is necessary to carry out recessive gene mining.

(2) Recessive gene mining

According to the analysis of policy text similarity data based on dominant genes and traditional policy text similarity data, it is necessary to mine some policy text recessive genes. Following is a detailed analysis of the process of recessive gene mining in the policy text of "Bid 1 Purchase 1". According to the previous analysis, the number of complete feature words of "Bid 1 Purchase 1" is 351. The number of dominant genes selected by the state space reduction algorithm based on automatic summary is 256, and the number of reduced feature words is 95.Therefore, this part mainly excavates 95 feature words, evaluates them according to formula (7), and finally identifies 41 recessive genes of policy texts. The recessive genes and dominant genes of policy texts are taken as part of the genome of policy texts, formula (8) is used to calculate the similarity of the policy text of "Bid 1 Purchase 1" again. The result is increased from 0.454453943432082 to 0.499036869258002.

The comparison of the number of feature words based on vector space and genome-based similarity calculation is shown in Fig.5.

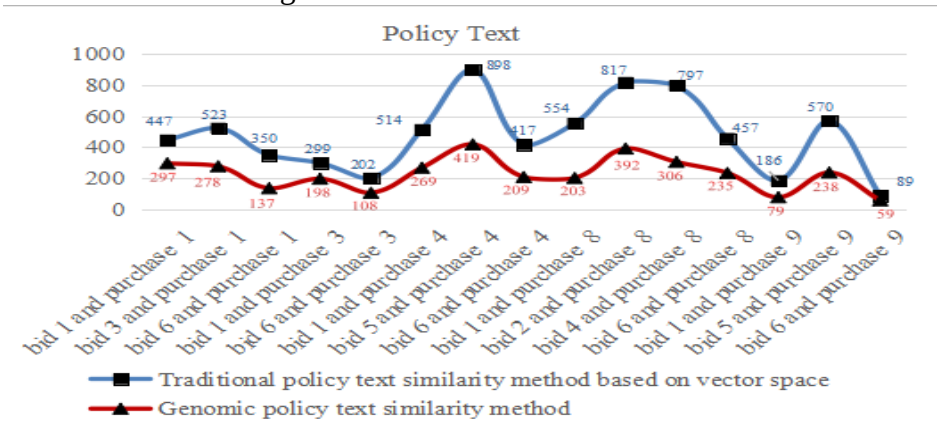

Figure 5. Comparison of the number of Characteristic words involved in the calculation of the two methods

From the comparative analysis in Fig.5, it can be seen that compared with the method of iteratively calculating the similarity of policy texts, the genome-based policy text similarity calculation method proposed in this paper reduces the feature words involved in the similarity calculation in the calculation process. The number, in turn, increases the efficiency of the similarity calculation.

6. Conclusion

By introducing the concept of policy consanguinity, this paper proposes a method of mining and validating policy genome based on state space reduction, and calculates the similarity of policy gene instead of text content, which makes large-scale policy text analysis possible. On the basis of policy text preprocessing, the weights of text feature words are measured by using linear weight weighting algorithm (LWA), which solves the problems of unstable weights and low efficiency caused by expert evaluation in the field. This paper proposes a vocabulary replacement algorithm based on influence degree (IWRA), which solves the synonymous relationship between feature words existing in policy texts, IWEA redefines the normative words and resolves the problem of inaccurate calculation of the similarity of the policy text caused by different replacement relations in the process of lexical standardization. Solving the problem of "high dimensional sparsity" in vector space model by text state space reduction.

The method proposed in this paper still needs to be improved. Firstly, we need to maintain the stop-use vocabulary and synonym forest in the policy field; secondly, we need to define the policy gene more perfectly and formally, and form the genome of the policy field into understandable language to facilitate manual analysis and confirmation; finally, we need to study the inheritance relationship between policy recessive genes and policy consanguinity, describing and explaining specific policy phenomena with policy consanguinity, confirming the rationality of policy consanguinity theory.

Acknowledgments

This work is sponsored by the Humanities and Social Sciences Research Planning Fund Project in Ministry of Education under grant number 19YJAZH053 and the Special Fund Project of Basic Scientific Research Business fee in Central Colleges and Universities under grant number 3072019CF0601.

References

[1] Xu L H, Sun S T, Wang Q. Text similarity algorithm based on semantic vector space model [C] //2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE, (2016):1-4.

[2] Kartsaklis D, Pilehvar M T, Collier N. Mapping Text to Knowledge Graph Entities using Multi-Sense LSTMs, J. (2018).

[3] He L, Xianghong Z, Dayou L, Fang J. A Feature Selection Method Based on Maximal Marginal Relevance, J. Journal of Computer Research and Development. (2012).

[4] Zhang W, Hu H, Hu H, Fang J. Semantic distance between vague concepts in a framework of modeling with words, J.Sci. Soft Computing, (2019), 23(10):3347-3364.

[5] Zhang Y, Ma Y, Meng X. Efficient Spatio-textual Similarity Join Using MapReduce [C] // 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). ACM, (2014).

[6] Mwebaze J, Mcfarland J, Booxhorn D, Valentijn E. A data lineage model for distributed sub-image processing [C] // Proceedings of the 2010 Annual Conference of the South African Institute of Computer Scientists and Information Technologists, SAICSIT Conf. 2010, Bela Bela, South Africa, October 11-13, 2010. DBLP, (2010).

[7] Kwakkel J H, Walker W E, Marchau V A W J. Classifying and communicating uncertainties in model-based policy analysis, J.Ei. International Journal of Technology, Policy and Management, (2010), 10(4):299.

[8] Broeren M.L.M, Saygin D, Patel M.K. Forecasting global developments in the basic chemical industry for environmental policy analysis, J.Sci. Energy Policy. (2014).1, 64:273-287.

[9] Wang N, Zeng J, Ye M, Chen M. Text mining and sustainable clusters from unstructured data in cloud computing, J.Ei. Cluster Computing, (2018), 21(1):779-788.