Policy Texts with Topic Detection and Information Entropy Evolution Analysis

Li-Xia Chen and Guo-He Feng

September 27, 2021

Policy texts with topic detection and information entropy evolution analysis

Li-Xia Chen[1[0000-0002-5781- 5249]] and Guo-He Feng [2[0000-0002-0774-1544]]

[1] School of Economics and Management, South China Normal University,
Guangzhou 510006, Guangdong, China
[2] Scientific Laboratory of Economic Behaviors, South China Normal University,
Guangzhou 510006, Guangdong, China
ghfeng@163com

Abstract. Talent policy has always been an important tool for countries to seize the talent highland and seek innovative development. However, talent policy text with complex themes, uneven distribution and unclear structure have caused great trouble for scholars to perform Talent Management. We constructed a large-scale unannotated corpus related to talent policy from Sougou Engine and collected 287 talent policies from the local government in Guangdong Province, China, which has been fuelled by rapid market growth and an unprecedented population surge resulted in a labor force of record size. We proposed a novel clustering model called LDA2Vec, which merges LDA and Word2Vec, and performed topic evolution analysis regarding topic similarity and topic entropy. The talent policies mainly included five topics: (i) talent introduction; (ii) talent training; (iii) talent guarantee; (iv) talent incentive, (v) talent evaluation. Talent policy in Guangdong province has gone through an evolutionary process from monism to pluralism. The introduction of China's innovation-driven strategy in 2013 as the dividing line directly impacts topic content and the intensity of five topics particularly.

Keywords: Talent Management; Economics of Human Resources; Science and Technology Innovation Policy; Topic Evolution; LDA2Vec; Semantic Similarity; Topic Entropy.

1 Introduction

Since an international management consulting firm created the term the War of Talent in 1997 [3], the topic of Talent Management attracts the attention of a large number of scholars. Talent Management was defined as activities and processes that involve identifying talents who differentially contribute to the organization’s sustainable competitive advantage [1]. Major corporations are struggling to find and retain the talent needed for growth and development. Yet, even in the wealthiest countries of East Asia, where innovation and education are seemingly fundamental principles, people still face barriers that prevent them from achieving professional success and personal fulfilment. It is clear that bridging these two issues is the remedy for economic stability and the long-term prosperity of countries. In this regard, the policy has always been an essential tool for governments to seize the talent highland and seek innovative development. Based on this, studying talent policy in content innovation and finding new competitive growth points beyond policy content have real significance for the economics of human resources development. However, previous researches on policy texts were more descriptive but weak in presenting a comprehensive overview that could depict the relevance and evolution of policy topics [2], which is helpful for the increased presence of talented people in the labour force and the long-term prosperity of corporations. Therefore, based on the LDA2Vec topic modeling method, our study deals with science and technology talent policy in southern China, specifically in the Province of Guangdong, which has been fuelled by rapid market growth and an unprecedented population surge resulted in a labour force of record size. LDA2Vecis a new topic model algorithm proposed by Moody [4]. It is a hybrid algorithm that mixes Word2Vec and LDA. LDA2Vec optimizes the result of text topic classification by adding a Word vector into the word layer of LDA. For example, Zhong et al. [5] used LDA2Vec to extract features from texts.

2 Research Design

2.1 Data Sources

In the paper, two hundred and eighty-seven talent policies issued by Guangdong Province from 2001 to 2020 were crawled down from various local governments websites as data set. The research on topic evolution requires dividing the acquired data into several contiguous sub-periods. Therefore, we choose three years as a time slice and divides the collected data into 7. The output of word2vec is a set of vectors representing each word existing in the corpus. In the paper, a total of 3273626 items related to science and technology talent was retrieved from Sogou Engine to construct a large-scale unannotated corpus. Each item contained a title and the main body. Then, all items were split into sentences, and the Skip-gram algorithm was applied to learn word vector representation.

2.2 Data Pre-processing

Building a policy ontology library. We constructed the government knowledge base of science and technology talents based on ontology to reduce data redundancy by distinguishing subordinate and cognate words and fusing knowledge with the same meaning. First, we collected documents from websites such as Legal Star, the Chinese government official website, and the Ministry of Science and Technology of the People's Republic of China. Next, we selected the terms to relate to talent policy. Then a topdown approach is adopted to construct the domain knowledge system. Finally, an information resource knowledge ontology is categorized with evident general characteristics, including attributes and relationships. Focusing on the research results of the knowledge topic classification of the Comprehensive E-Government Subject Headings and drawing on the division of some government policies, we established a talent policy knowledge ontology framework system as shown in Fig. 1, based on three processes of cultivating talent, attracting talent and using skills.

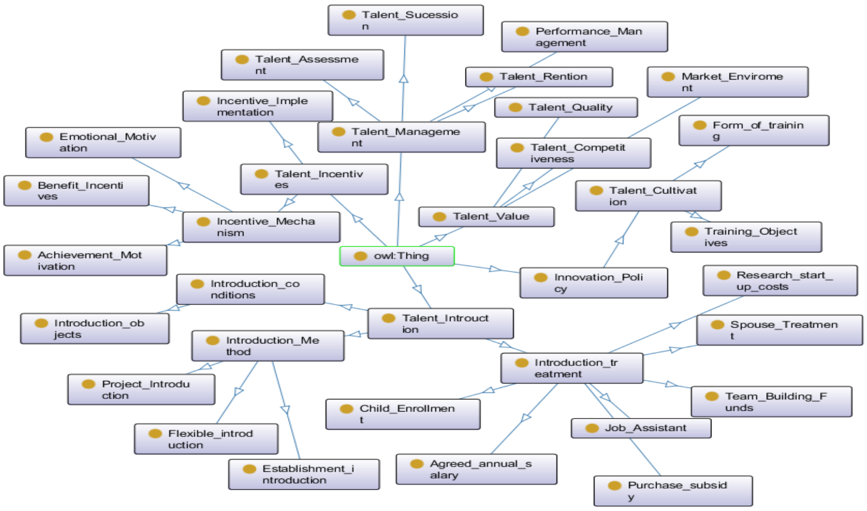

Fig. 1. Talent policy knowledge ontology framework system

Using the ontology editing tool Protégé 5.5, we first established the 'innovation policy class'. Then, we divided five first-level categories, including talent development, talent motivation, talent introduction, talent management, and talent value according to the theme taxonomy. Next, we constructed the sub-categories according to the implementation mode, implementation object, and implementation theory, respectively. Under each class, synonyms were attributed to the ontology library according to the synonym class and intergeneric relationship. Thus, excluding generic keywords as well as synonymous keywords, as shown in Fig. 2.

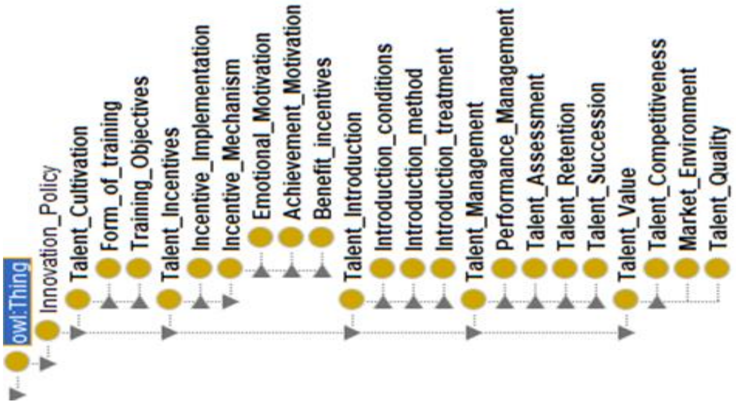

Fig. 2. Class hierarchy of knowledge ontologies

A hybrid domain concept extraction method. The paper proposes a hybrid domain concept extraction method for science and technology policies to exclude generic and low-frequency words. Based on matching rules, statistical methods are used to filter candidate concept words. First, the simple structural information is extracted from the data set. Second, the forward maximum matching method is used to scan the dictionary and select the word that matches the longest word in the dictionary as the target sub-word for the subsequent matching. After that, HTP is used to mark the lexicality, and verbs such as "accept", "issue", "implement", adverbs such as "basically", "exactly", "above" and adjectives are filtered. Next is to identify the syntax and structure of the string, eliminate the ambiguity between different combinations of words and then obtain the candidate words. Finally, the domain concepts at the semantic level are obtained through manual validation and ontology libraries.

2.3 Topic extraction with LDA2Vec

To meet the input requirements of lda2vec, there are two steps we need to do: Step 1: One part of the input to the lda2vec model comes from the output of the LDA model, i.e., keywords, the topic-word distribution matrix and the document weights. To determine the optimal number of topics, we first combine the consistency and perplexity methods for preliminary judgment, followed by a fast manual verification using the pyLDAvis visualization method. For document topics with unknown distribution, the smaller the value of complexity, the better the model, and the greater the score for consistency. Step 2: Another part of the input to the lda2vec model comes from the output of the word2vec model. After the above process, we chose word vectors of the selecting keywords in the vocabulary as one part of the input to the lda2vec model.

2.4 Topic evolution analysis

Topic evolution analysis is mainly divided into two types, the evolution of the topic content and topic intensity. At present, the traditional topic evolution methods are based on term frequency, burst terms, co-word segmentation, citation analysis, etc. In the paper, we innovatively propose the topic similarity for topic content and the information entropy for topic intensity.

Topic similarity. It’s widely used for NLP processing tasks to calculate word similarity, mainly exploring relevance in words [6]. By comparing the topic similarity with a predefined similarity threshold, we can identify the highly relevant topic. In the paper, cosine similarity is used to measure the relevance of words on adjacent time slices. If the similarity value is greater than or equal to the predefined threshold, it is determined that there is an evolutionary relationship between the two words. Presumed there exist two vectors, A and B, their cosine similarity θ is given by the dot product and the vector length, as follows:

$$\cos\theta = \frac{A \cdot B}{|A| \cdot |B|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} \cdot \sqrt{\sum_{i=1}^{n} {B_i}^2}} \tag{1}$$

Topic entropy. The topic intensity reflects the evolution trend and the change of the focus. The paper introduces Shannon's concept of information entropy into the topic intensity analysis, which contributes to measure the degree of category division. The process of information entropy to calculate topic intensity is as follows:

(1) LDA2Vec divides the topic policy text Q into T topic clusters, and each topic cluster contains the extracted ranked topN keywords. In order to calculate the topic intensity in different time slice, the paper needs to rearrange the keywords according to the time series. For each keyword, the paper calculates its word frequency in the document collection of each time slice and ranks them in order from highest to lowest. If the word frequency of the keyword in that time slice is the highest value, then the keyword is classified into this time slice. At this time, the policy text Q is divided into i time slices {group_1, group_2, ..., group_i}, and the keyword set under each time slice is W {W_1, W_2, ...W_j}, so that:

$$group_i = \left\{W_{(i,1)}, W_{(i,2)}, \dots \dots, W_{(i,j)}\right\} \tag{2}$$

(2) The term frequency of keywords in certain time series is given by:

$$\mathrm{P}\left(\mathrm{W}_{\mathrm{i,j}}\right) = \frac{\mathrm{count}（\mathrm{W}_{\mathrm{i,j}})}{\mathrm{size}（\mathrm{i}）} \tag{3}$$

Among them, count（$W_{i,j}$）is the number of occurrences per keyword in the document set, and size(i) is the total number of words in a time series.

(3) After obtaining the term frequency, we calculate the information entropy of the entire document topic. The topic intensity is obtained by a linear transformation of topic entropy in equation (5) to correlate the final result positively. H(groupi) means the entropy of one topic cluster on a specific time series, and WQ represents the topic intensity.

$$\mathrm{H}(\mathrm{group_i}) = \sum_{\mathrm{i=1}}^{\mathrm{j}} -\mathrm{P}\left(\mathrm{W_{i,j}}\right) \log_2 \mathrm{P}(\mathrm{W_{i,j}}) \tag{4}$$

$$\mathrm{W_Q} = 1 - \frac{\mathrm{H(group_i)} - \min\{\mathrm{H(group_1)},\dots,\mathrm{H(group_i)}\}}{\max\{\mathrm{H(group_1)},\dots,\mathrm{H(group_i)}\} - \min\{\mathrm{H(group_1)},\dots,\mathrm{H(group_i)}\}} \tag{5}$$

According to the concept of information entropy, the larger the entropy value of the keyword set, the greater the intensity value of the topic cluster, and vice versa.

2.5 Visualization of topic evolution

A visualization method is used to show the overall landscape and local details of the evolution of the topic content and topic intensity over time. In this paper, topic content evolution can be visualized by the Sankey diagram. Sankey Diagram, namely the Sankey energy diversion diagram, originated from the "Energy Efficiency Diagram of Steam Engine" drawn by Sankey in 1898. It is mainly composed of edge, flow and nodes. The edge represents the flow data, the flow represents the route of topic evolution, and the nodes represent different categories. The width of the edge is proportional to the size of the data volume. The change of topic intensity is shown by folded line chart.

3 Results

3.1 Topic modeling

As shown in Fig. 3, the perplexity reaches the lowest point when the number of topics is 5, while the highest performance of the intra-topic consistency score is found when the number of topics is set to 4. Therefore, in this paper, the number of topics is considered in the interval 4 - 7.

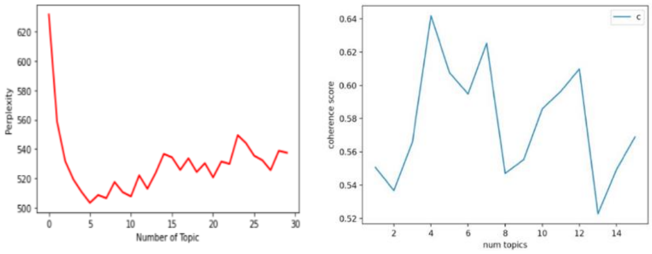

Fig. 3. The perplexity and coherence score in the data sets

PYLDAvis contributes to obtain the performance of different topic counts in Fig. 4, where λ is the parameter that regulates which of the two attributes is important, set between 0 and 1.

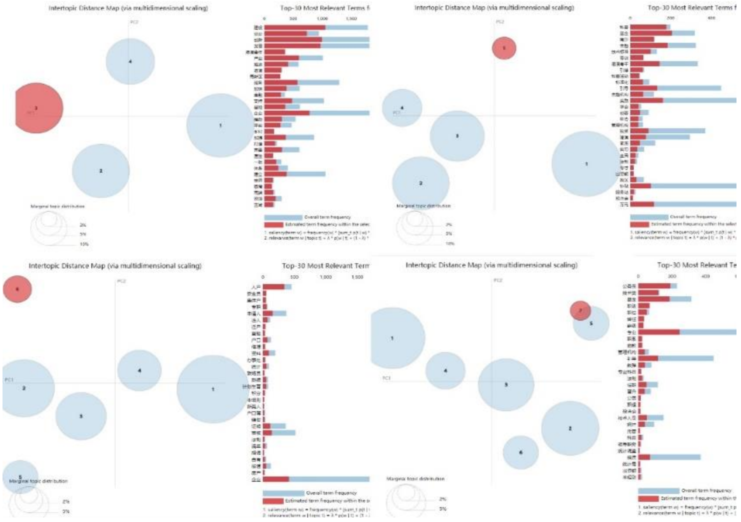
Intertopic Distance Map (via multidimensional scaling)
Top-30 Most Relevant Terms
Marginal topic distribution
Overall term frequency
Estimated term frequency within the selected topic

Fig. 4. Topic distribution map of policy texts

When the number of topics is set to 4, we find that the concepts of talent cultivation and talent value coexist in the 4th bubble; when the number of topics is set to 6, the concepts of "household entry", "collective household", and "applicant" in the 6th bubble are similar to the talent introduction in the 2nd bubble; finally, when the number of topics is set to 7, the bubbles overlap and thus discarded. After manual judgment, when the number of topics is set to 5, we can fully express the concept of each theme independently. As a result, the optimal number of topics is set to 5. After data preprocessing, the obtained word vectors and document vectors are fed into the lda2vec model for fusion training to perform topic-word extraction. For some synonyms, this paper uses the talent policy ontology library to put forward the concept of synonymy and only records the highly relevant keywords in the year of the first occurrence, as shown in Table 1. The five topics obtained from the LDA2vec modeling results are summarized as talent introduction, talent training, talent guarantee, talent incentive, and talent evaluation.

Table 1. Topic identification results

Time Series	Number	Highly-relevant Topic Keyword
2001-2003	9	Overseas students, Technic talents, Graduates, Scientific research institutions, High-tech, Talent Enrollment System, Compulsory Education, Income Distribution System, Professional titles
2004-2006	11	Scientific and technological talents, Outstanding talents, Transformation of scientific and technological achievements, Patents, Intellectual property rights, Science and technology industry bases, Science and technology economy, Special funds, Equity, Performance indicators, Technic qualifications
2007-2009	14	Postdoc, Shenzhen City, Distinguished Talents, Graduate Students, Technical Community, Municipal Government, Colleges and Universities, Training and Employment, Civil Servants, Staffing, Preferential Policies, Special Allowances, Salary, Audit System
2010-2012	9	High-skilled talents, High-tech zone construction, New situations, Informatization, Social security, Backup work, Scientific research funding, Technical achievements, Comprehensive assessment
2013-2015	11	High-level talents, Scientific research projects, Basic research, Employers, Entrepreneurial parks, Innovative companies, Industry-university-research, Human resources, Domain experts, Innovation plans, Expert reviews
2016-2018	12	Pearl River Delta, Fintech, Workstations, Industrial system, Science and technology services, Vocational education, Management system, Information system, Infrastructure, Innovation projects, Independent innovation, Evaluation mechanism
2019-2020	10	Scientific and technological innovation, Key technology, Leading talents, Talent echelon, Gathering innovation platform, Equity incentive, Scientific research reward system, Continuous investment system, Supervision and evaluation mechanism, Scientific research integrity

3.2 Visualization of topic evolution

The visualization of topic content evolution is shown in Fig. 5.

After obtaining the highly relevant keywords from the above topic, the cosine similarity equation (1) calculates the similarity among the topic keywords in adjacent time slices. According to existing research, the paper sets 0.3 as the similarity threshold. If the similarity value is greater than or equal to 0.3, we can determine an evolutionary relationship between the two words, representing topic content evolution.

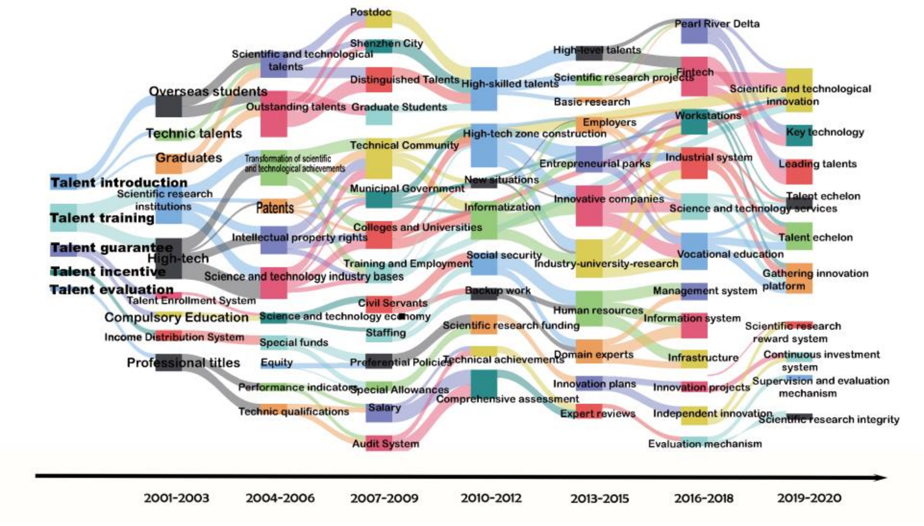

Fig. 5. Depicting topic content evolution through Sankey diagram.

As is shown in Fig. 5, the first column on the left is represented by 5 topics. Some keywords will appear repeatedly in different time slices, but we choose keywords only when they first appear. The keywords of adjacent time slices are connected by the same color line segment. The size of the color block represents the proportion of the topic word at a certain time. The thickness of the line segment represents the degree of evolution between the two keywords in the adjacent time. The paper identifies five evolutionary paths of the same marks: in the early stage, we find that the Guangdong Province mainly introduced overseas talents and skilled talents. With the development of big data, it began to focus on introducing high-level talents, such as highly skilled leaders who have made outstanding contributions to economic development and effective strategies. Furthermore, in the context of policy discourse of breaking the "five-only", professional titles are no longer the only evaluation criteria, and more attention is paid to improving the evaluation mechanism and implementing comprehensive assessment of talents.

3.3 Visualization of topic intensity evolution

We used equations 4 and 5 to calculate the topic intensity (see Fig. 6). As shown in Fig. 6, the topic intensity of 5 topics is unevenly distributed and widely disparate in 2001-2003, 2016-2018 and 2019-2020, among which "talent evaluation" reaches the highest value. 2001-2003 and 2004-2006, 2013-2015 and 2016-2018 adjacent time slices have the most considerable fluctuation. This corresponds to the talent strategy in 2002 and the innovation-driven strategy in 2013 issued by the Chinese government. We can conclude by observing that after the central government proposed an effective strategy, local ministries and commissions released various policies in response to it. As a result, the number of policies reached a small peak.

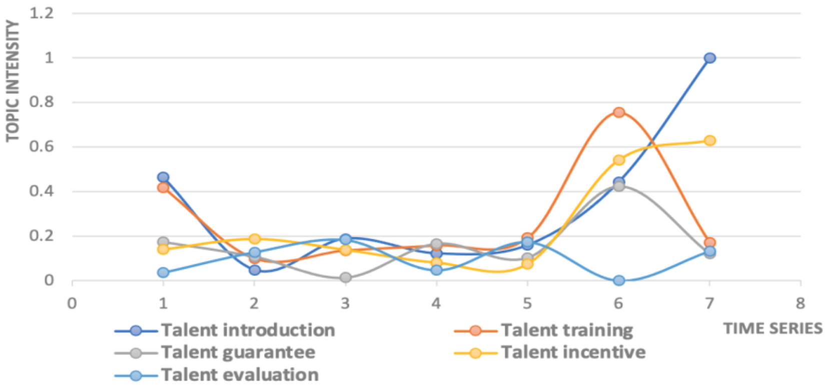

Fig. 6. Topic intensity evolution chart

4 Discussion

In the paper, we detect five evolutionary paths by LDA2Vec modeling. Our research shows that the topic content of science and technology talent policy have been more specific and innovative. It is worth noting that with the innovation-driven strategy proposed in 2013 as the dividing line, Guangdong Province's science and technology talent policy topics of talent introduction, talent training, and talent incentive have become more and more critical. Meanwhile, more relevant policies have been published. However, talent guarantee and talent evaluation show lower intensity, which needs to be strengthened. Before the Chinese government proposed the innovation-driven strategy in 2013, Guangdong, the leader of reform and opening up, focused on talent introduction and incentive. At this time, talent introduction shows the highest topic intensity. For example, Shenzhen released the Regulations on Labor Contracts in Shenzhen Special Economic Zone and the Notice on Several Provisions for Encouraging Expatriates to Come to Shenzhen for Entrepreneurship. All of these policies paid more attention to Vigorously introducing overseas students, skilled talents and graduates. Undoubtedly, as the last link of talent development policy, talent evaluation requires a long process to form a universally recognized, unified, scientific talent evaluation standard, which leads to an embryonic stage in talent evaluation. Since the Chinese government proposed an innovation-driven strategy in 2013, Guangdong Province has vigorously introduced talents again, which led to the topic content and topic intensity of talent introduction increased substantially in this period. In recent years, Guangdong Province has issued many policies such as "Opinions on Accelerating the Attraction and Cultivation of High-Level Talents" and "Outline of Medium and Long-Term Talent Development Plan of Guangdong Province (2010-2020)" to accelerate the attraction and cultivation of high-level talents, so that the topic intensity of talent training reached the highest in 2016-2018. The topic intensity of talent incentive has been increased since 2013, in which keywords show an increasing abundance. Besides, the policy focus has gradually evolved from (the system of) income distribution and preferential policies to results incubation and research reward system. After years of policy accumulation, Guangdong Province has gradually formed a more mature evaluation system of scientific and technological talents.

5 Conclusion

The paper introduces an entropy measure for topic intensity in the framework of previously developed algorithms of the Latent Dirichlet Allocation (LDA) type. Besides, we construct the policy framework of science and technology collaborative innovation based on science and technology innovation policies. The novel contribution given is represented by studying the evolution of science and technology innovation policy. In this regard, we vividly understand the space-time background of organization and talent development. In terms of topic content distribution, on the one hand, the keywords under the five relevant topics are increasingly enriched over time. On the other, in terms of topic intensity evolution, there is no apparent bias among five topics before 2013. Still, it has changed a lot after 2013, among which talent introduction, talent training and talent incentive are getting more and more attention. As a next step to move, we will consider the participation of experts in the specific fields in order to get more authoritative and accurate topic extraction results and, thus, achieve an external, more realistic validity.

Acknowledgements. This work was supported by Guangzhou Science and Technology Plan Project “Theory, method and application of burst terms detection” (grant no. 202002030384), South China Normal University with the project "Quantitative analysis method and empirical research of policy texts based on multi synergy"(grant no. ZDPY2107) and by Guangdong philosophy and social science planning project “Multi dynamic evaluation method and empirical research on scientific and technological talents” (grant no.GD21CTS05).

References

1. Collings D., Kamel M.: Strategic talent management: A review and research agenda. Human resource management review 19(4), 304-313 (2009).
2. Grimmer J., Stewart B.: Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis 21(3), 267-297 (2013).
3. Michaels, E., Helen H., Beth A.: The war for talent. Harvard Business Press (2001).
4. Moody C.: Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (2016).
5. Qinghong Z., Xiaodong Q., Yunliang Z.: Cross-media Fusion Method Based on LDA2Vec and Residual Network. Data Analysis and Knowledge Discovery 3(10),78-88 (2019).
6. Jun Y., Juhyeon L.: A Text Mining Analysis of US-Chinese Leaders on Trade Policy. Journal of International Logistics and Trade 17(3), 67-76 (2019).