

# Using Text Mining to Evaluate the Made in China 2025 Regional Action Plan

Kong Dejing<sup>1</sup>, Dong Fang<sup>2</sup>, Li Zhaofu<sup>2</sup>, Qu Xianming<sup>3</sup>

1. School of Modern Post, Beijing University of Posts and Telecommunications, Beijing 100876, China

2. School of Mechanical Science & Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

3. The CAE Centre for Strategic Studies, Beijing 100088, China

**Abstract:** The guidelines for the implementation of the *Made in China 2025 Regional Action Plan* focus on the development of major areas and sub-areas in various provinces and cities. To comprehensively understand the provincial and municipal action plans in China, consistency analyses for major development areas and difference analyses for major development sub-areas are fundamental. This paper proposes a method of analyzing policy documents based on Chinese word segmentation, feature extraction, and similarity calculation using text mining, thereby helping analyze a large number of policy documents efficiently. The results indicate the consistencies and differences in the action guidelines for various provinces and cities.

**Keywords:** Made in China 2025; regional action plan; text mining; similarity calculation

## 1 Research background

The “Research on Manufacturing Power Strategy (Phase I)” is a key advisory project jointly organized by the Chinese Academy of Engineering, Ministry of Industry and Information Technology, and the General Administration of Quality Supervision, Inspection, and Quarantine from 2013 to 2014. After conducting a systematic, in-depth survey and research, this project has clearly indicated the three-step strategic goal of transforming China into a manufacturing power and proposed the development guidelines necessary for achieving this strategic goal and also eight strategic countermeasures for implementing this manufacturing-driven strategy. The development guidelines are being innovation driven, emphasizing quality over quantity, achieving green development, and optimizing manufacturing industrial structure. Additionally, the project has emphasized the adherence to the development of the manufacturing industry, thus facilitating the establishment and release of *Made in China 2025*.

In addition to the goal of realizing a manufacturing power

through the three-step development strategy, *Made in China 2025* has indicated the emphasis on the breakthrough in the ten key fields, namely new materials, the new-generation information technology industry, energy-saving and new-energy vehicles, biological medicines and high-performance medical instruments, high-end computer numerical control (CNC) machine tools and robots, aerospace equipment, marine engineering equipment and high-technology ships, advanced rail transit equipment, agricultural machinery, and power equipment. To avoid the repeated efforts of developing low-end sectors, the Ministry of Industry and Information Technology delegates the research group to release the technical roadmap for guiding the development of these ten fields. To fulfill the central government’s request, local governments have also proposed their plans to develop the manufacturing industry and chosen the key development fields. According to relevant surveys, among the 34 provinces (autonomous regions, municipalities, and special administrative regions) in China, 30 have released the local action plans for *Made in China 2025* except Tibet, Taiwan, Hong Kong,

**Received date:** April 25, 2017; **Revised date:** May 23, 2017

**Corresponding author:** Qu Xianming, the CAE Center for Strategic Studies, Professor of Engineering. Major research fields include advanced manufacturing technology, and industry and technology development strategy. E-mail: qu.xianming@163.com

**Funding program:** CAE Advisory Project “Research on Manufacturing Power Strategy (Phase II)” (2015-ZD-15)

**Chinese version:** Strategic Study of CAE 2017, 19 (3): 149–158

**Cited item:** Kong Dejing et al. Using Text Mining to Evaluate the Made in China 2025 Regional Action Plan. *Strategic Study of CAE*, <https://doi.org/10.15302/J-SSCAE-2017.03.021>

and Macao. The Heilongjiang Province, Liaoning Province, Inner Mongolia autonomous region, and Ningxia Hui autonomous region have not yet released a complete version of their action plans.

However, when drafting their action plans, some local governments did not consider their development stage and development advantage. They blindly chose all of the ten key fields in *Made in China 2025* as the focus of development. Owing to the limited human resource and material resource, this will inevitably increase the difficulty in simultaneously developing the advantaged fields and disadvantaged fields. Consequently, they will not only fail to achieve the development goals but also cause overcapacity on a national scale.

Using text mining, we analyzed the local action plans of *Made in China 2025* in those provinces and cities. First, whether the action plans of those provinces and cities are consistent with the key development fields in *Made in China 2025* will be analyzed. Whether sufficient differentiation exists between the key subfields of the key development fields will also be analyzed. Moreover, some policies and suggestions will be proposed to avoid overcapacity. A new policy document analysis method based on text mining is proposed to overcome the difficulty of relying on manual operations to analyze a large number of policy documents rapidly.

## 2 Methodology

### 2.1 Cross-province correlation analysis of key development fields

#### 2.1.1 Consistency analysis of key development fields of the provinces and cities

The local action plans for *Made in China 2025* of those provinces and cities should reflect the key development directions suitable for their own conditions. The consistency of the key development fields of those provinces and cities also reflects the consistency degree in the national action guide. To analyze whether the action plans released by those provinces are consistent with the key development fields proposed in *Made in China 2025*, we propose a text-mining-based consistency analysis method aimed at the key development fields of *Made in China 2025*, as shown in Fig. 1. Specifically, this method is based on Chinese word segmentation [1,2], part-of-speech analysis [3,4], feature selection [5,6], and similarity computation [7,8].

First, the Java programming language was used to extract and analyze the Chinese word segmentation and part-of-speech in the local action plans for the key development fields of *Made in China 2025* in those provinces and cities. The screening based on part-of-speech has also been conducted to gain the initial feature set [9]. Because of the difference in the word usage to express the same key development field in the local action plans, we adopted manual categorization and generalization of the fea-

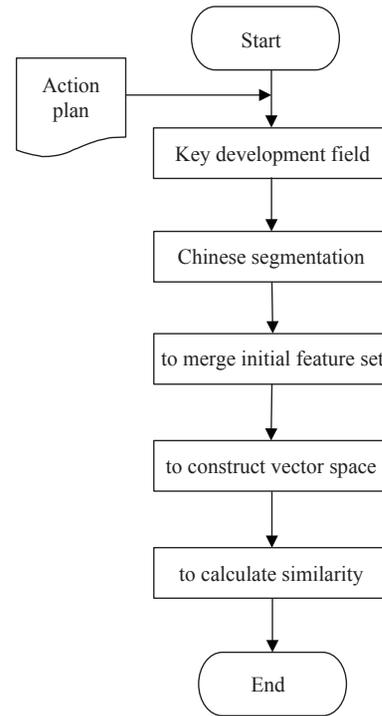


Fig. 1. Selection process of consistency analysis for the key development fields of the provinces and cities.

ture words. The initial feature words with the same meaning are classified into the same category. The feature words categorized are included in the final feature set [10].

All key development fields of those provinces and cities will be matched with the final feature set, based on which the vector space model for the key development fields of each province and city will be made. The number of dimensions of the vector space model is equal to the number of the final feature sets. The weight is equal to the number of feature words in the final feature set for the key development fields in those provinces and cities [11,12].

A similarity computation will be conducted to analyze the consistency between the key development fields of those provinces and cities. Herein, the Euclidean distance is adopted to measure the degree of similarity between the vector space models [13–16]. The Euclidean distance represents the degree of similarity of the signal. The shorter the Euclidean distance, the higher is the degree of the Euclidean distance. The computation of the Euclidean distance is as follows:

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2}$$

$d_{12}$  means the distance between two points in the Euclidean space.  $x_{1k}$  means the  $k$ th-dimensional coordinate of the first point.  $x_{2k}$  means the  $k$ th-dimensional coordinate of the second point.

#### 2.1.2 Difference analysis of the subfields of the key development fields of those provinces and cities

In addition to the consistency between the key development

fields of the local action plans for *Made in China 2025*, their action plans should also reflect their own development features. In other words, the subfields of the key development fields of those provinces and cities should be different from each other to avoid a new round of overcapacity. We herein propose a text-mining-based difference analysis method aimed at the key development subfields of those provinces and cities, as shown in Fig. 2. The difference analysis method is also based on the result of cross-province correlation analysis, Chinese word segmentation, feature selection, and similarity computation.

First, the key development fields with a high degree of similarity between those provinces will be selected. The Java programming language is used to extract the key development subfields of those provinces. The Chinese word segmentation and part-of-speech analysis of the subfield content will be conducted to obtain the initial feature set. Moreover, the feature words will be categorized and generalized. Specifically, the initial feature words with the same meaning will be classified into the same category. The feature words categorized will be included in the final feature set. All the key development fields of those provinces and cities will be matched with the final feature set, based on which a vector space model for the key development fields of each province and city will be created. A similarity computation will be conducted to analyze the consistency between the key development fields of those provinces and cities. Herein, the Euclidean distance has been adopted to measure the degree of similarity between the vector space models.

## 2.2 Research framework

First, the features of the key development fields of the local action plans aimed at *Made in China 2025* are extracted to build the vector space model. The Euclidean distance is used to measure the degree of similarity between the vector space models of those provinces. Based on the degree of similarity, the consistency analysis of the key development fields of those provinces and cities will be performed [17–19]. Next, the key development fields encompassing many provinces will be selected. The key development subfields of those provinces concerned will be also analyzed. The features of the key development subfields are extracted to build the vector space model. The Euclidean distance

is also adopted to measure the degree of similarity between those vector space models. Based on the degree of similarity, the analysis of the difference between key development subfields in those provinces and cities will be performed [17,20]. The research framework for the cross-province correlation analysis of key development fields is shown in Fig. 3.

## 3 Analysis results of local action guide

### 3.1 Consistency analysis of the key development fields in the provinces and cities

First, the Java language is used to extract the key development fields of those provinces and cities. The Chinese word segmentation and part-of-speech analysis of the key development fields will be also conducted. The screening is based on the part-

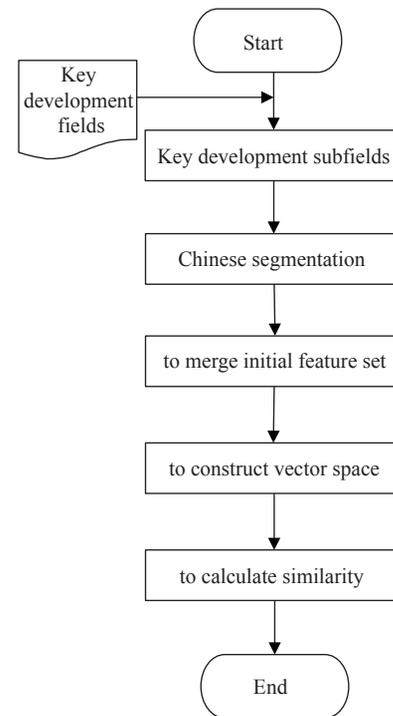


Fig. 2. Process of difference analysis of key development subfields of the provinces and cities.

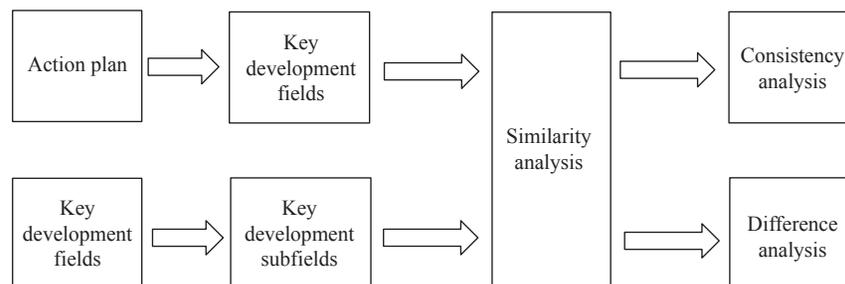


Fig. 3. Research framework for the cross-province correlation analysis of key development fields.

of-speech. After the manual categorization and generalization of feature words, the final feature set is shown in Table 1.

The Euclidean distance has been adopted to measure the degree of similarity between the vector space models of the key development fields of those provinces. The computation of the similarity and consistency between key development fields is shown in Fig. 4.

As shown in Fig. 4, the black solid circle in each lattice represents the textural similarity of the key development fields for *Made in China 2025* between two provinces. The bigger the black solid circle, the higher is the degree of similarity. For example, the degree of similarity in the same province can be represented by a relatively large black solid circle because of the same key development fields.

As shown, a relatively high degree of textual similarity

exists for the key development fields of *Made in China 2025* between Beijing, Guangdong, Guizhou, Hainan, Hebei, Shanxi, and Yunnan. According to the segmentation results, the key development fields of those provinces or cities are concentrated in the emerging fields, such as the Internet, cloud computing, new material, pharmacy, and manufacturing. They are also the key development fields repeatedly advocated by *Made in China 2025*. Therefore, their action guides are consistent with *Made in China 2025* and thus are critical in guiding the strategy of *Made in China 2025*.

### 3.2 Difference analysis of the key development fields of the provinces and cities

The ten key development fields, namely new materials, the

Table 1. Final feature set.

No. Field	No. Field	No. Field	No. Field	No. Field
1. Power	7. Aerospace	13. New energy	19. Pharmacy	25. New-type
2. Ship	8. CNC	14. New material	20. Manufacturing	26. High performance
3. Agricultural machinery	9. Instrument	15. Maine	21. Energy saving	27. Rail transit
4. Information technology	10. Textile	16. Robot	22. Electronic	28. New-energy vehicle
5. Biology	11. Material	17. Vehicle	23. Medical care	29. Integrated circuit
6. Internet of things	12. Cloud computing	18. Machine tools	24. Intelligence	30. Equipment

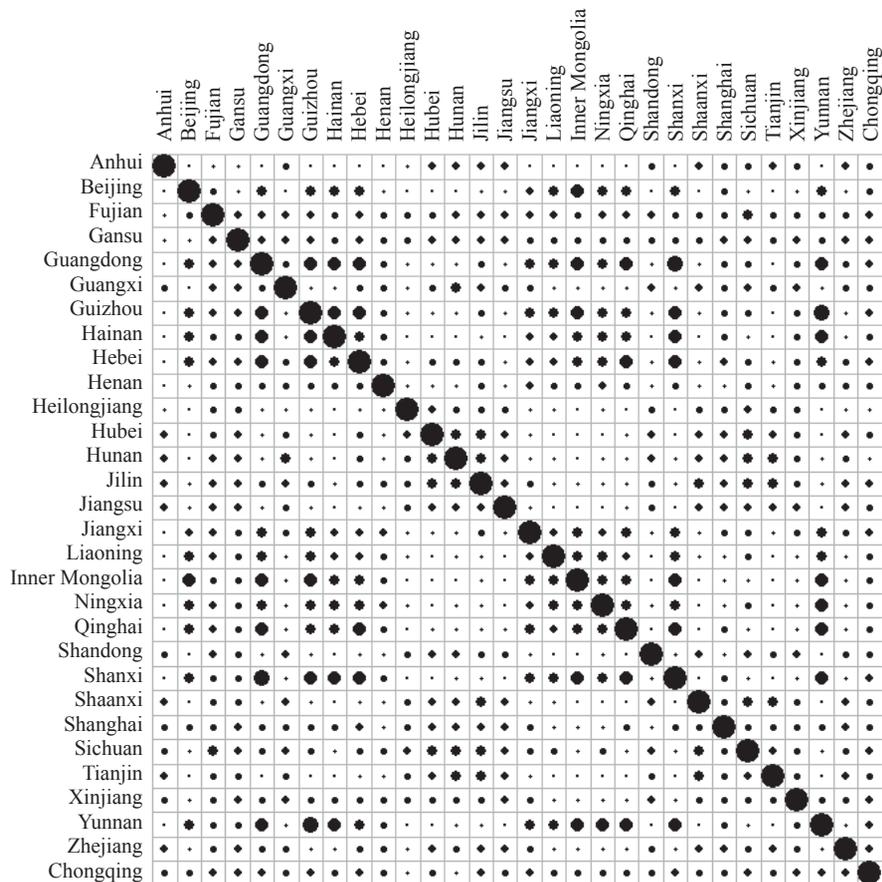


Fig. 4. Degree of similarity of key development fields in the provinces and cities.

new-generation information technology industry, energy-saving and new-energy vehicles, biological medicines and high-performance medical instruments, high-end computer numerical control (CNC) machine tools and robots, aerospace equipment, marine engineering equipment and high-technology ships, advanced rail transit equipment, agricultural machinery, and power equipment, are chosen as the demonstrative research fields. Their subfields will also be analyzed. The features of the key development subfields of those provinces will be extracted to build the vector space model. The Euclidean distance has been adopted to measure the degree of similarity between the vector space models of the key development fields of those provinces. Based on the degree of similarity, the analysis of the difference in the key development subfields between different provinces will be performed. The degree of similarity between the development subfields of those provinces and cities is shown in Fig. 5–14. The black solid circle in each lattice represents the textural similarity of the key subfields of the key development fields between two provinces. The larger the black solid circle, the higher is the degree of similarity. For example, the degree of similarity in the same province can be represented by a relatively large black solid circle because of the same key development fields.

As shown in Fig. 5, the key development subfield of new materials in those provinces and cities does not have a high degree of similarity as a whole. This means that the focus of the key

development subfield varies from one province to another. Considering that the different provinces have focused on different aspects of the new material development, a new round of overcapacity in the field of new material will not occur, thus contributing to a more rational resource allocation. Moreover, a relatively high degree of similarity exists between Hebei and Gansu, Gansu and Chongqing, as well as Guangdong and Chongqing. In the field of new material, Hebei, Gansu, Chongqing, and Guangdong have focused on the metallurgy and chemicals of new materials.

As shown in Fig. 6, the key development subfield of the new-generation technology in those provinces and cities has some similarities. Hainan, Hunan, and Xinjiang has a relatively low degree of similarity compared with the other provinces. The development of the new-generation technology industry is faced with the potential risk of causing a new round of overcapacity. Specifically, a relatively high degree of similarity exists between Anhui and Hubei, Fujian and Hubei, Fujian and Shanxi, as well as Hubei and Shanxi. In the field of new-generation technology, Anhui, Hubei, Fujian, and Shanxi have focused on integrated circuits and intelligent terminals.

As shown in Fig. 7, the key development subfield of new-energy vehicle in those provinces and cities does not have a high degree of similarity as a whole. Compared with the other provinces, Tianjin, Shanghai, and Jilin have an obviously high degree of similarity. All of them have focused on the development of

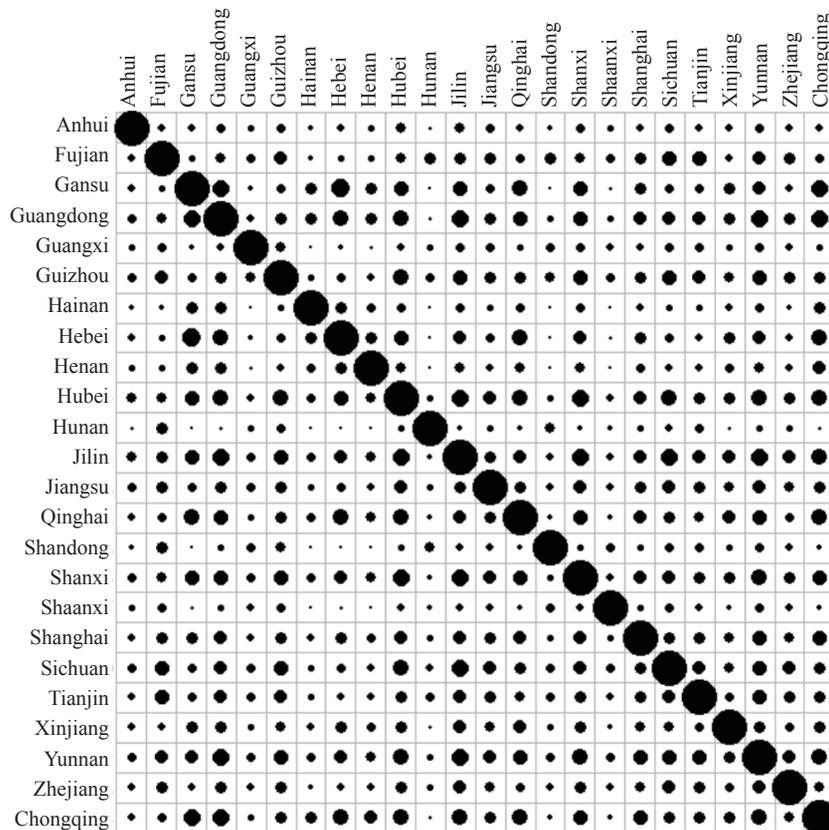


Fig. 5. Similarity of the key development subfield of new material in the provinces and cities.

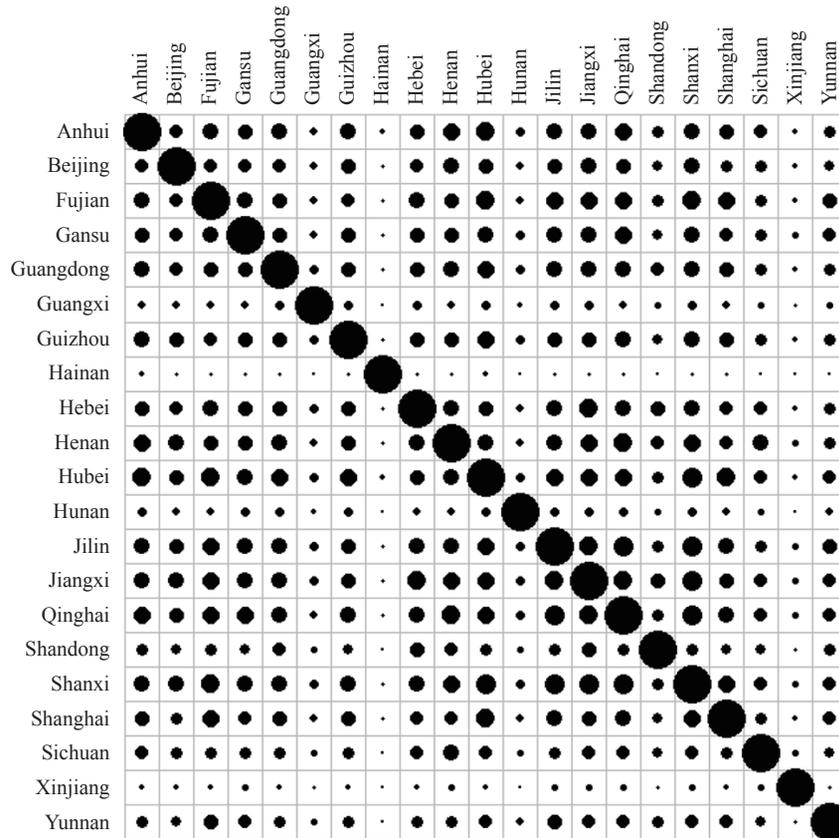


Fig. 6. Similarity of the key development subfield of new-generation information technology in the provinces and cities.

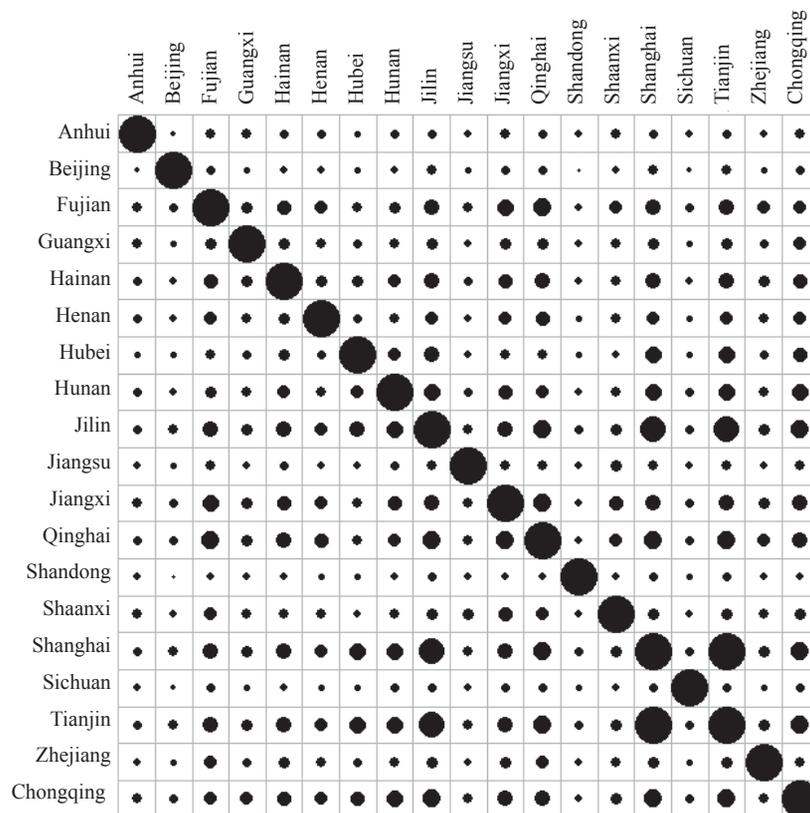


Fig. 7. Similarity of the key development subfield of energy-saving and new-energy vehicle in the provinces and cities.

intelligently connected vehicles. With a low degree of similarity, other provinces have overly concentrated their resources in a certain area. Despite the low risk of causing a new round of overcapacity, they should focus on the repeated investment in a certain area.

As shown in Fig. 8, the key development subfield of biological medicine and high-performance medical instrument in those provinces and cities has a high degree of similarity as a whole. Only Shandong and Zhejiang have a relatively low degree of similarity compared with the other provinces. This means that this field will easily cause the excessive concentration of resources and repeated investment. Because the resource allocation is not based on their features, it will involve a high risk of causing a new round of overcapacity. With the highest degree of similarity, Jilin, Gansu, Hebei, and Shanghai have focused on the development of high-performance medical instrument and biological pharmacy.

As shown in the Fig. 9, the key development subfields of high-end CNC machine tools and robot in those provinces and cities are different from each other as a whole. A relatively high degree of similarity exists between Sichuan and Henan, Sichuan and Gansu, as well as Henan and Hunan. In the field of high-end CNC machine tools and robot, Sichuan, Henan, Gansu, and Hunan have focused on the development of CNC machine and robot, thus leading to a high risk of causing a new round of over-

capacity. Zhejiang and Jilin have a relatively low degree of similarity compared with the other provinces. Considering the actual conditions and features of the provinces, the risk of causing a new round of capacity is low in this field.

As shown in Fig. 10, the key development subfield of aerospace equipment in those provinces and cities has a low degree of similarity as a whole. This means that those provinces have focused on different aspects of aerospace equipment development. To a certain extent, it can avoid the risk of causing a new round of overcapacity in this field and also contribute to a more rational resource allocation. A relatively high degree of similarity exists between Anhui and Sichuan, Shanghai and Jilin, as well as Sichuan and Shanghai. In the field of aerospace equipment, Anhui, Sichuan, Shanghai, and Jilin have focused on the development of aerospace, aviation, and unmanned aerial vehicle.

As shown in Fig. 11, the key development subfield of marine engineering equipment and high-technology ship in those provinces and cities has a low degree of similarity as a whole. This means that those provinces have focused on different aspects of marine engineering equipment and high-technology ship development. To a certain extent, it can avoid the risk of causing a new round of overcapacity in this field. A relatively high degree of similarity exists between Anhui and Hubei, Sichuan and Shanghai, as well as Anhui and Shanghai. They have focused on the development of marine engineering equipment and

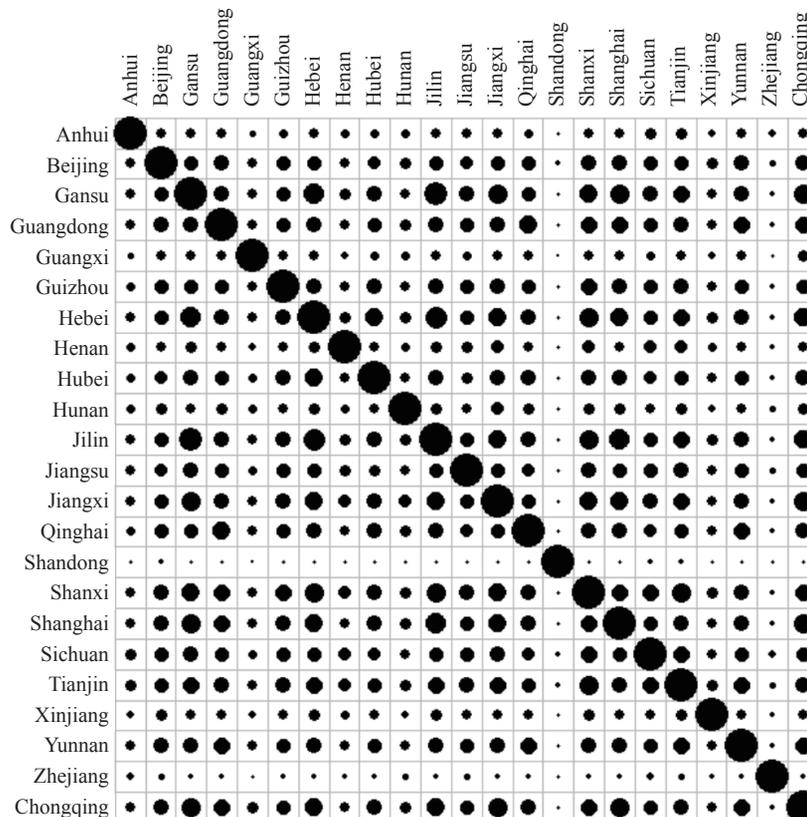


Fig. 8. Similarity of the key development subfield of biological medicine and high-performance medical instrument in the provinces and cities.

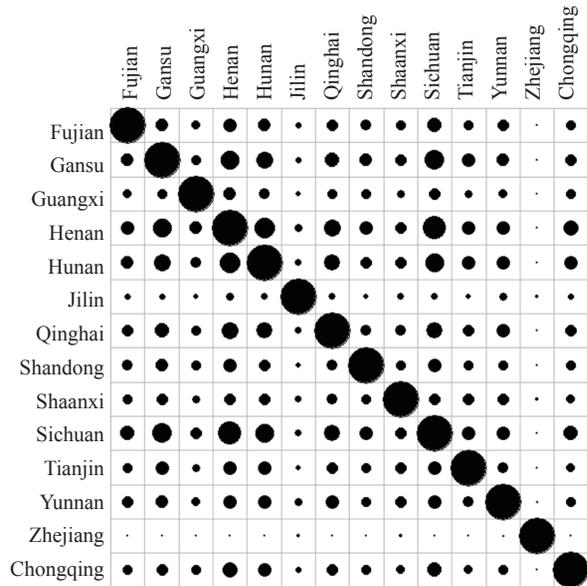


Fig. 9. Similarity of the key development subfield of high-end CNC machine tools and robot in the provinces and cities.

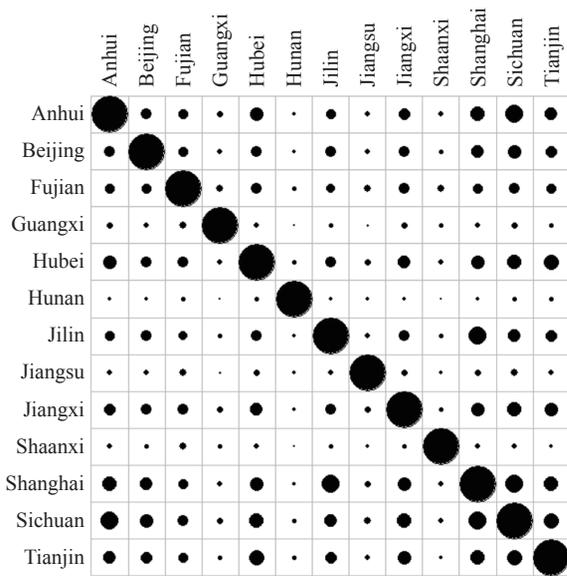


Fig. 10. Similarity of the key development subfield of aerospace equipment in the provinces and cities.

high-technology ship.

As shown in Fig. 12, the key development subfield of advanced rail transit equipment in those provinces and cities has a low degree of similarity as a whole. This means that those provinces have focused on different aspects of advanced rail transit equipment. Owing to the low level of technical overlapping, it can reduce the risk of causing a new round of overcapacity in this field and also contribute to a more rational resource allocation. With a relatively high degree of similarity, Tianjin and Hubei have focused on the fields of manufacturing rail transit vehicles and key components.

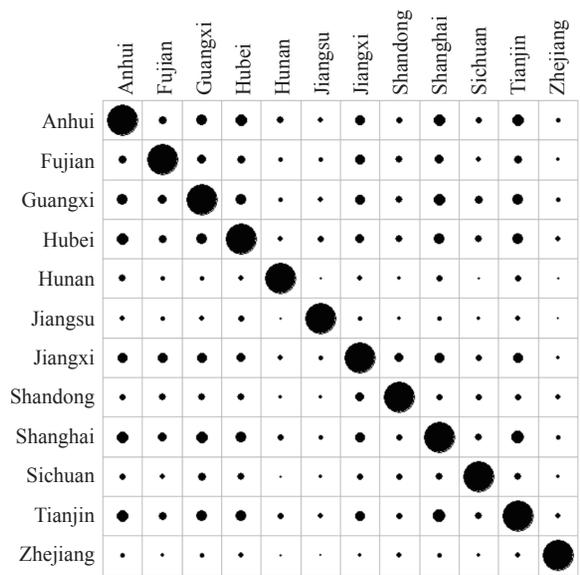


Fig. 11. Similarity of the key development subfield of marine engineering equipment and high-technology ship in the provinces and cities.

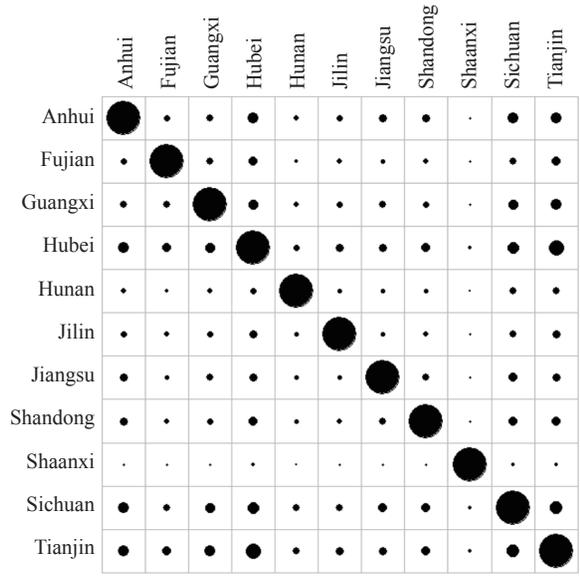


Fig. 12. Similarity of the key development subfield of advanced rail transit equipment in the provinces and cities.

As shown in Fig. 13, the key development subfield of advanced rail transit equipment in those provinces and cities has a high degree of similarity as a whole. Owing to the high level of technical overlapping, it will involve a high risk of causing a new round of overcapacity. The highest degree of similarity exists between Sichuan and Gansu, Sichuan and Jilin, as well as Jilin and Gansu. Sichuan, Gansu, and Jilin have focused on the fields of agricultural machinery, agricultural-product processing machinery, and modern precision agricultural machinery in the hilly area.

As shown in Fig. 14, the key development subfield of power

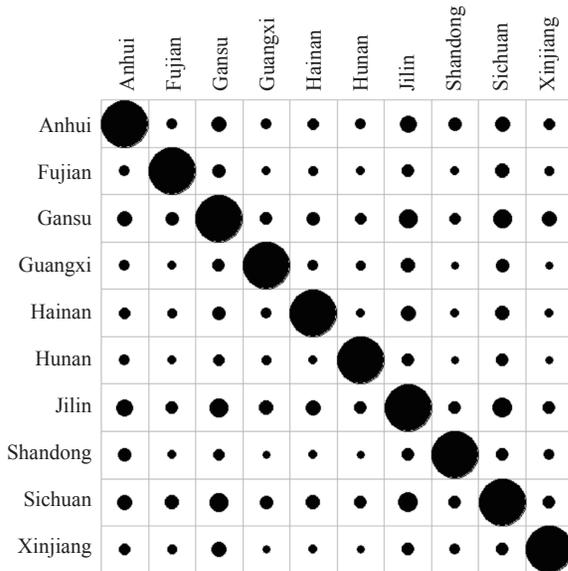


Fig. 13. Similarity of the key development subfield of agricultural machinery in the provinces and cities.

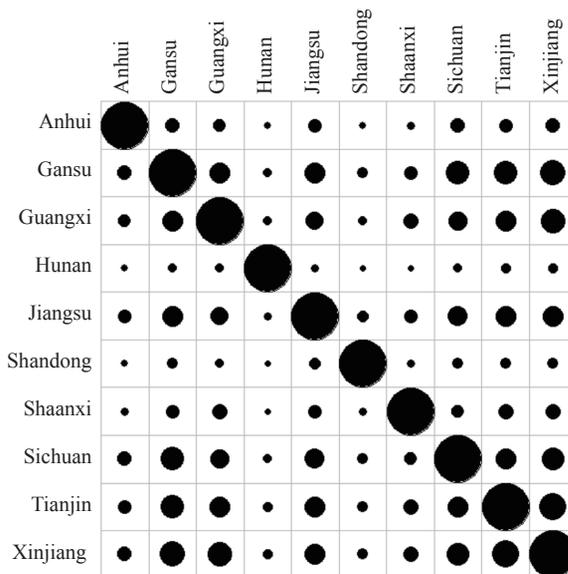


Fig. 14. Similarity of the key development subfield of power equipment in the provinces and cities.

equipment in those provinces and cities has a high or low degree of similarity as a whole. Specifically, Xinjiang has a high degree of similarity with Tianjin, Guangxi, and Gansu. They have focused on the transmission and distribution of power grids, which will have a risk of causing a new round of overcapacity in this field. Hunan has a low degree of similarity with the other provinces, which suggests the local characteristics.

## 4 Conclusions

Local governments should be involved in handling the overcapacity. In 2009, the Chinese government released ten industrial

revitalization plans. However, some local governments released a series of preferential policies and supporting policies without considering their actual conditions, thus causing repeated investments in various industries nationwide, leading to overcapacity.

We found that those provinces (autonomous regions, municipalities, and special administrative regions) in China have shifted their focus to the ten key fields, implying that the development of traditional industries is at the risk of being neglected. The high degree of similarity in the key development fields between those provinces (autonomous regions, municipalities, and special administrative regions) cannot fully exploit those regions. Local policies can be involved in guiding the development of the key industries. The differentiated development of key industries in those regions should be guided to avoid the overcapacity on a national scale. Some suggestions are proposed as follows:

(1) *The Provincial and Municipal Guide for Made in China 2025* should be revised to facilitate the coordinated regional development. Based on the regional advantages and the actual needs of different industries and enterprises of different scales, differentiated policy support should be provided to stimulate the vitality of regional development. The competitiveness benchmarking analysis of the key development fields in different regions should be performed. The segments of the primary industry, secondary industry, and tertiary industry should be determined. They should be highly prioritized for further development.

(2) The information regarding the capacity and demand of the key fields should be released regularly. The governments of all levels should contribute significantly to realize the industrial informatization. With the new-generation information technologies and network technologies, they should also keep track of the changing market demand of the key and popular industries as well as release the relevant information.

## References

- [1] Liang X T, Gu L. Study on word segmentation and part-of-speech tagging [J]. *Computer Technology and Development*, 2015, 25(2): 175–180. Chinese.
- [2] Cao W F. Research on key techniques of Chinese word segmentation (Master's thesis) [D]. Nanjing: Nanjing University of Science and Technology, 2009. Chinese.
- [3] Mo J W, Zheng Y, Shou Z Y. Improved Chinese word segmentation method based on dictionary [J]. *Computer Engineering and Design*, 2013, 34(5): 1802–1807. Chinese.
- [4] Yue Z Y. A study on Chinese word Segmentation combined with dictionary and statistics (Master's thesis) [D]. Wuhan: Wuhan University of Technology, 2010. Chinese.
- [5] Sun J. Robust estimations and feature screenings for some nonparametric and semiparametric models (Doctoral dissertation) [D]. Jinan: Shandong University, 2013. Chinese.
- [6] Zhang T, Wu W Z, Wan Y L. Model selection based on feature screening [J]. *Journal of Guangxi University of Science and Technology*, 2016, 27(1): 26–30. Chinese.

- [7] Li R, Wang Z Q, Li S H. Chinese sentence similarity computing based on frame semantic parsing [J]. *Journal of Computer Research and Development*, 2013, 50(8): 1728–1736. Chinese.
- [8] Shen B. Research on similarity calculation of Chinese text based on word segmentation (Master's thesis) [D]. Tianjin: Tianjin University of Finance & Economics, 2006. Chinese.
- [9] Yang T L, Zhang H L. Pattern matching for super large patterns set [J]. *Chineses Journal of Computers*, 2014, 37(5): 1147–1158. Chinese.
- [10] Zhang B F, Su J S. Feature set segmentation for collaboration in text classification [J]. *Computer Science*, 2009, 36(2): 142–145. Chinese.
- [11] Sun R, Liu Z T, Liao T. Research on reducing dimension of feature vector based on ontology [J]. *Computer Engineering and Design*, 2010, 31(17): 3864–3867. Chinese.
- [12] Ren M R. Research on word frequency extraction and automatic text classification in digital library [D]. Harbin: Heilongjiang University, 2002. Chinese.
- [13] Sun R Z. Research and implementation of text similarity calculation based on semantic understanding (Master's thesis) [D]. Shenyang: Shenyang Institute of Computing Technology, Chinese Academy of Science, 2015. Chinese.
- [14] Tang G. Text similarity calculation based on semantic domain vector space model (Master's thesis) [D]. Kunming: Yunnan University, 2013. Chinese.
- [15] Wang Z Z, He M, Du Y P. Text similarity calculation based on topic model LDA [J]. *Computer Science*, 2013, 40(12): 229–232. Chinese.
- [16] Li R B, Li A H, Cai Y P, et al. Euclidean distance based method for unclassifiable region of support vector machine [J]. *Journal of Computer Applications*, 2010, 30(2): 476–478. Chinese.
- [17] Zhou J. Intelligent manufacturing—main direction of “China Manufacturing 2025” [J]. *China Mechanical Engineering*, 2015, 26(17): 2273–2284. Chinese.
- [18] Zhang Y. Short text similarity calculation based on characteristic extension of BTM theme model (Master's thesis) [D]. Hefei: Anhui University, 2014. Chinese.
- [19] Zhan Z J, Yang X P. Text similarity calculation based on language network and semantic information [J]. *Computer Engineering and Applications*, 2014, 50(5): 33–38. Chinese.
- [20] Jin Y H. Text similarity calculation based on context framework [J]. *Computer Engineering and Applications*, 2004, 40(16): 36–39. Chinese.