Identifying Main IPCs in Patents of China's Epidemic Prevention Products and Analysing Spatial Diversity of IPCs Types

Student Number: ucfnxz0

Research Question

COVID-19 started to spread all over the world at the beginning of 2020. Office of State Intellectual Property published the patents of China's epidemic prevention products (EPP) from 1993 to 2019. This paper used this dataset to try to answer these questions: 1. What kinds of main IPCs do EPP patents products have, when considering the whole EPP dataset, not just patents corresponding to different products? The general IPCs that are composed in the different area of products could be identified. And these groups of IPCs may have significant effects on the innovation of EPP products. 2. What is the IPCs types composition in the cities where patents were published? The result of question tries to measure the diversity of IPCs types on the city level in China. The structure of diversity may find on the state level. These two questions may help the researcher understand the scale of main areas of knowledge in EPP and their influences on innovation in cities.

Literature Review

Patent Analysis

Patent analysis is the data science of analysing a large amount of patent information, to discover relationships, trends and patterns in the data for decision making. (Aristodemou and Tietze, 2018) Patent documents could reflect the embedded technologies the inventors use and the usage rights they claim. (Park, Yoon and Lee, 2005) Multinational corporation applied a large number of patents in other countries to acquire massive profit on technology authorisation. For the chemical industry and the pharmaceutical industry, patents have a significant influence on protecting firms' prospects(Simonetti *et al.*, 2007). When a drug loses its patent rights, the firm's sales of that drug usually drop 50–85% in the next year (Fred Gebhart, 2006). For countries, the

patent analysis could establish public policy(Kim and Bae, 2017), such as detecting the emerging technologies and their trends to improve the accuracy of the government's policies on a specific industry. Patent analysis of industry and technologies has become a significant and useful way to understand and layout the core technologies in advance.

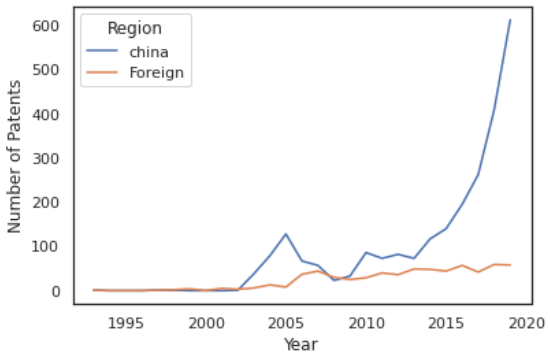
Methods used on Patent Analysis

Various methods have been used for analysing patent characteristics, such as text mining, artificial intelligence, deep learning, networks and clustering. Assad Abbas (Abbas, Zhang and Khan, 2014) summarised these tools into two categories: patent information mining and visualisation. When doing patent information mining, the first step is to define the classification of patents. Some paper used IPC or CPC classification system to cluster patent documents for a similar technology(Kim and Bae, 2017). Methods to decrease patent documents' dimension were used, such as principal component analysis (PCA), multidimensional scaling (MDS) and other nonlinear methods. K-means(Jun, Park and Jang, 2014) and self organising map(Chen and Chang, 2010) are primary methods to cluster the patents based on the distance among patents.

However, these studies lacked spatial elements. Some scholars argued that localised organisations' interactions were the internal mechanism of cluster innovation(Bagchi-Sen, Smith and Hall, 2004). Geographical proximity and the diversity of knowledge types in one region positively impact the innovation of patents. Thus, the characteristics of patent technology in the context of geography need to discuss more in the future.

Data

The dataset contained 4156 patents in February 2020. The annual number of EPP patent opened for public enlarged from 1995 to 2019. From Fig.1, the number expanded every year, despite the period between 2005-2012, which showed a slight decline and stable character. It was apparent that after 2013, the number of annual publications grew faster than any other stage. In 2019, more than 670 patents were published. On the contrary, the growth of patents applied from foreign countries was much more stable. What is more, the scales of EPP patent technologies and their ranks followed power-law distribution, as was shown in Fig.2. Ln(rank) and Ln (scales of technologies) followed this equation: $y = -1.573 x + 8.796$. R^2 was 0.94 and P-value was 1.28e-128.



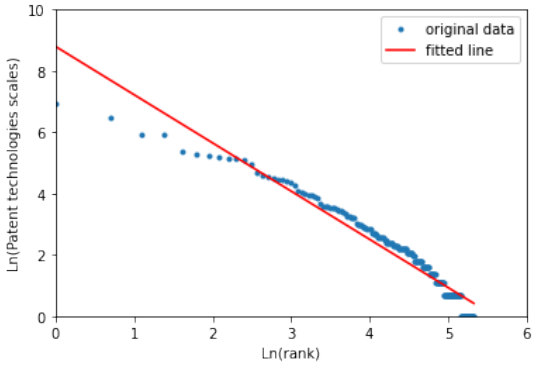


Fig.1.(Left) Trends of number of patents

Fig.2.(Right) The distribution of scales of patent technology

Table 1 shows the number of patents in each type of EPP. Medical apparatus and instruments cover the largest percentage of EPP patents in China, up to 26.6%. The drug used for treatment is the second one. This category could be used to analyse the composition of IPCs in each kind of product. However, when considering EPP's whole image, clustering the patents based on their similarity is necessary and then main IPCs could be found in each cluster.

Table.1. The number of each types of EPP patents and its percentage

Category of EPP	**Number**	**Percentage**
Care Products	450	10.97%
Detection and diagnostic kits	335	8.17%
Drug used for treatment	867	21.14%
Medical apparatus and instruments	1089	26.55%
Medical sterilization	68	1.66%
Waste disposal methods	172	4.19%
Method for wastewater treatment	186	4.53%
Prophylactic	88	2.15%
Other	847	20.65%

Methodology

This paper uses the K-means algorithm to find the similarity among patents, and then the specific technologies in each cluster could be found. K-means divides a group of data into k groups, based on the similarity of data itself. The number of clusters (k) and distance matrix needs to be input by researchers (Hartigan and Wong, 1979). The

silhouette scores were calculated to determine the number of K (whose silhouette score is the largest one).

This paper calculates the similarity among patents based on their IPC first four numbers to get a similarity matrix first. Table 2 shows how patent documents change to patent relation matrix. The transformation was conducted by calculating Pearson's correlation coefficient between patents. The coefficient whose P-value is upper than 0.01 will be set to 0.

Table.2. The way of changing patent documents to patent relation matrix.

Patent	**A61K**	**G06F**	**C12Q**
Patent 1	1	0	1
Patent 2	0	1	1
Patent 3	1	0	0

→

Patent	**Patent 1**	**Patent 2**	**Patent 3**
Patent 1	0	0.66	0.37
Patent 2	0.66	0	0.30
Patent 3	0.37	0.30	0

Next, the distance matrix was calculated by 1- coefficient. To change distance matrix into 2-demensions scale, the paper uses Multidimensional scaling (MDS) to reduce its dimension (Cox, 1994). When finishing this process, the result could be used as the input of K-means.

What is more, the paper also calculates the entropy of patent technology. The entropy could enable the researchers to detect the diversity of regions. (Alan Wilson, 2013; Batty et al., 2014) The entropy of technologies area in each city is calculated by the following functions. The function (1) calculated probability of IPCs types (P_i) in each city:

$$P_i = \frac{A_i}{\sum_i^k A_i} \tag{1}$$

Where A_i refers to the number of IPCs in type i and k represents the number of IPCs types. IPC types in this paper were defined by WIPO highest categories, which is from A to H. The entropy is denoted as G and it is calculated as follow function (2):

$$G = -\sum_i^k P_i * \ln P_i \tag{2}$$

Results

The experiment result of Silhouette scores (Fig.3) shows that when the number of clusters k is set to 3, the silhouette score is the highest, up to 0.797. On the right, the result of K-means cluster is shown in Fig.4. Cluster 1,2,3 contains 2720 (65%) patents, 986 (24%) patents and 449 (11%) patents, respectively.

This paper also summarises the top 10 IPCs of each cluster in table 3. Cluster 1 shows more heterogeneous clustering of various areas of patent. The technologies in “A61” area emerged much more in Cluster 2 and cluster 3 than cluster 1. According to IPC definition, “A61” is related to medical or veterinary science; hygiene. This area of technologies is mainly related to medical instruments, pharmaceutical inventions, and sterilisation.

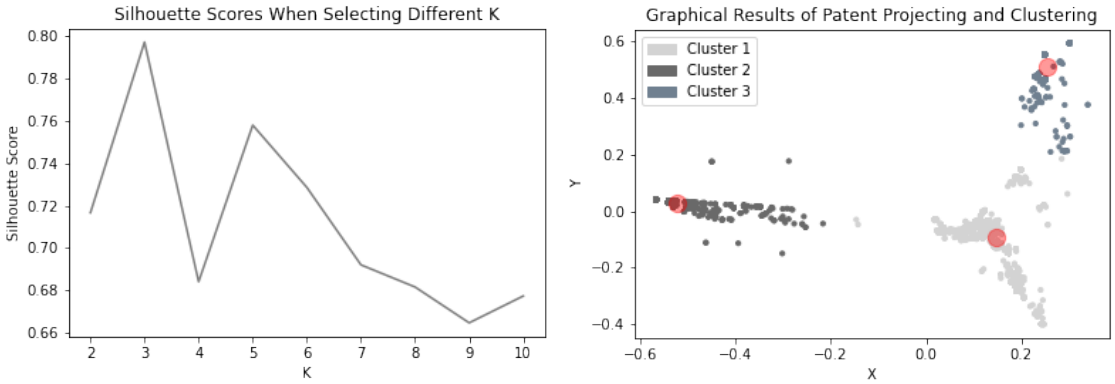


Fig.3.(Left) Silhouette scores when selecting different K

Fig.4.(Right) Graphical results of patent projecting and clustering

Also, the result of the entropy of IPCs’ types was shown in Fig.5. Various cities located at the northwest of Hu Line, which was proposed by Hu huanyong in 1935 to compare the disparity of population density between north and south regions in China, did not have statistics. Thus, these cities are categorized into 0. The experiment result showed that some megacities, such as Beijing, Shanghai, Guangzhou and Shenzhen, which contained many patent innovations, could also have high diversity on IPCs types.

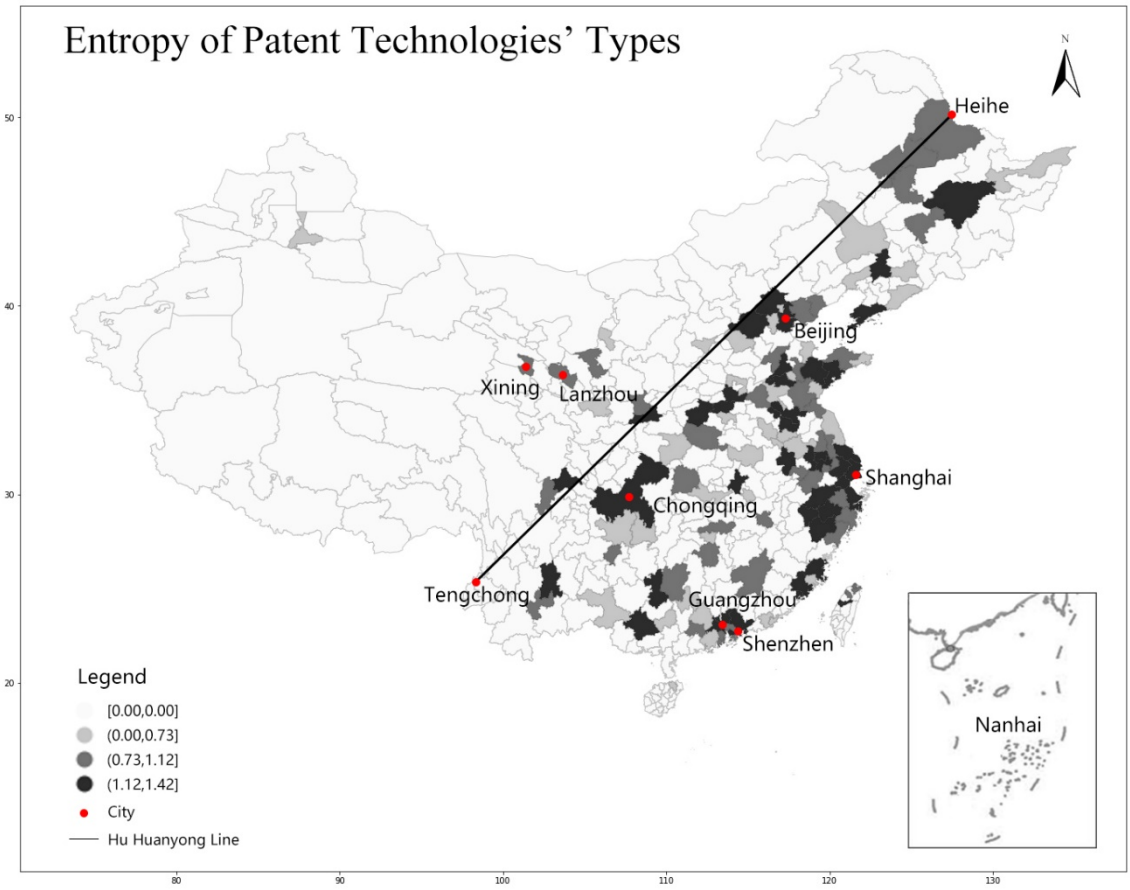


Fig.5. Entropy of Patent Technologies' Types

Table.3. Top 10 IPCs for each cluster.

Rank	Patent Data Set		
	Cluster 1	**Cluster 2**	**Cluster 3**
1	A61H35/02(246) *	A61P31/16 (374)	A61B5/01(200)
2	G06F19/00(136)	A61P11/00(343)	A61B6/03(85)
3	C12Q1/68(126)	A61P31/14(290)	A61B5/00(39)
4	B25J11/00(118)	A61P31/12(184)	A61M16/00(33)
5	C12Q1/70(114)	A61K39/215(82)	A61B6/00(22)
6	G16H40/20(105)	A61K9/20(64)	G01J5/00(21)
7	G16H50/80(94)	A61P31/18(57)	A61B5/0205(20)
8	A61M16/00(93)	C07K14/165(55)	G01K13/00(19)
9	A41D13/11(88)	A61K38/21(54)	A61B5/08(18)
10	G06T7/00(79)	A61P29/00(52)	A61B5/087(16)

* Number in parentheses refers to number of IPCs emerged in EPP.

Discussion

This paper used clustering (K-means) methods to summarize the characteristics of EPP patents. Some other methods, such as Pearson's correlation to calculate the similarity among patents, multidimensional scales to decrease patent documents' dimension and entropy to define IPC types' diversity, have also been used. The limitation of these methods need to mention. MDS restricts the structure of input matrix, which needs to be a linear relation. Applying MDS in a nonlinear context could distort the original data relationship. Nevertheless, the similarity among patents may not just be described as linear relation. The invention of a patent could be more complexed when considering different factors, such as the amount of local funding and connectivity to the knowledge cluster. The definition of similarity among patents needs to be discussed more in the future. For K-means, it is sensitive to noise and the precision of input data sample need to be high. This algorithm also needs the data to be convex. What is more, entropy is much more sensitive when the scale of the IPCs in one city is relatively small.

Conclusions

This paper mainly focused on find characteristics and spatial diversity of technologies in patents of China's epidemic prevention products. By applying the clustering method, the paper identified some key IPCs in each cluster, which could further interpret the characteristics of EPP patents on the whole image. What is more, the result of entropy of IPC types in China shows that megacities which contain the high number of EPP patents also have relatively high diversity on IPCs types. (1705 words)

References

Abbas, A., Zhang, L. and Khan, S. U. (2014) 'A literature review on the state-of-the-art in patent analysis', *World Patent Information*. Elsevier Ltd, pp. 3–13. doi: 10.1016/j.wpi.2013.12.006.

Alan Wilson (2013) *Entropy in Urban and Regional Modelling (Routledge Revivals)*, *<h>Entropy</h> <h>in Urban</h> <h>and Regional</h> <h>Modelling</h> (Routledge Revivals)*. Taylor and Francis (Routledge Revivals). doi: 10.4324/9780203142608.

Aristodemou, L. and Tietze, F. (2018) 'The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data', *World Patent Information*. Elsevier Ltd, pp. 37–51. doi: 10.1016/j.wpi.2018.07.002.

Bagchi-Sen, S., Smith, H. L. and Hall, L. (2004) 'The US biotechnology industry: Industry dynamics and policy', *Environment and Planning C: Government and Policy*, 22(2), pp. 199–216. doi: 10.1068/c0345.

Batty, M. *et al.* (2014) 'Entropy, complexity, and spatial information', *Journal of geographical systems*, 16(4), pp. 363–385. doi: 10.1007/s10109-014-0202-2.

Chen, Y. S. and Chang, K. C. (2010) 'Exploring the nonlinear effects of patent citations, patent share and relative patent position on market value in the US pharmaceutical industry', *Technology Analysis and Strategic Management*, 22(2), pp. 153–169. doi: 10.1080/09537320903498496.

Cox, T. F. (1994) *Multidimensional scaling / Trevor F. Cox and Michael A. A. Cox*. Edited by M. A. A. Cox. London: Chapman & Hall (Monographs on statistics and applied probability (Series) ; 59).

Fred Gebhart (2006) 'Major drugs lose patent protection in 2006', *Drug Topics*, p. S8.

Hartigan, A. and Wong, M. A. (1979) 'A K-Means Clustering Algorithm', *Journal of the Royal Statistical Society*, 28(1).

Jun, S., Park, S. S. and Jang, D. S. (2014) 'Document clustering method using dimension reduction and support vector clustering to overcome sparseness', *Expert*

Systems with Applications, 41(7), pp. 3204–3212. doi: 10.1016/j.eswa.2013.11.018.

Kim, G. and Bae, J. (2017) 'A novel approach to forecast promising technology through patent analysis', *Technological Forecasting and Social Change*, 117, pp. 228–237. doi: 10.1016/j.techfore.2016.11.023.

Park, Y., Yoon, B. and Lee, S. (2005) 'The idiosyncrasy and dynamism of technological innovation across industries: Patent citation analysis', *Technology in Society*, 27(4), pp. 471–485. doi: 10.1016/j.techsoc.2005.08.003.

Simonetti, R. *et al.* (2007) 'The Dynamics of Pharmaceutical Patenting in India: Evidence from USPTO Data', *Technology analysis & strategic management*, 19(5), pp. 625–642. doi: 10.1080/09537320701521382.