Classificatory Analysis of Disputes on the Right of New Plant Varieties Based on LDA Model

Mingyue Zhou and Yu Yang *

1School of Law, Shenzhen University, Shenzhen, China

*Corresponding author e-mail: 2018031054@email.szu.edu.cn

Abstract. This paper takes the judgments between 2011 and 2020 on disputes on the right of new plant varieties from China Judgments Online as the object of research and uses LDA (Latent Dirichlet Allocation) subject model algorithm to analyze the causes and content of these infringement disputes with keyword extraction. With the help of LDA, it finds that China’s disputes over the right of new plant varieties are mainly concentrated in four aspects: the lack of consistent standards for the identification of "new varieties"; the difficulty in responsibility division between the stakeholders involved; the ambiguity in the identification of "propagating materials" and unclear causes of exemption system.

1. Introduction

As the old saying goes, "Food is heaven for the people." The cultivation of new plant varieties provides humans with a richer source of food. In the 50 years since the founding of the People’s Republic of China, Chinese agricultural scientists and technicians have cultivated more than 6,000 new varieties of major crops, and the major food and cash crops have been replaced 4-6 times. The implementation of intellectual property rights protection for new plant varieties is becoming an emerging trend in major agricultural countries.

The promulgation of the Regulations on the Protection of New Plant Varieties of the People's Republic of China in 1997 not only marked the establishment of the protection system on the right of new plant varieties, and also marked a higher level of intellectual property protection. As the youngest member of the intellectual property family, the protection of the rights of new plant varieties is still in the stage of continuous exploration and improvement. In recent years, with more prosperous technological innovation, the number of applications and authorizations for new plant varieties has been increasing year after year, and therefore the related infringement disputes have also been increasing.

To better analyze the causes of these infringement disputes, we need to systematically classify the judgments of these cases. Traditional patent text analysis methods are mainly manual ones, which are inefficient and costly, and the classification is not accurate enough. These problems can be improved by using the LDA topic model.

The LDA subject model, also known as Latent Dirichlet Allocation, is an unsupervised machine learning technology that can gather documents, extract their subjects, give the probability distribution of each subject, and classified these subjects into different groups. Here goes the document generation method in the LDA model [1]:

\[ P(z) = \sum_{i=1}^{k} \alpha_i P(z|\phi_i) \]

\[ P(\phi_i) = \beta \prod_{w=1}^{V} \theta_{w,i} \]

\[ P(d|\phi) = \sum_{z=1}^{Z} \sum_{i=1}^{k} \frac{n_{d,i} \phi_i}{\sum_{j=1}^{k} n_{d,j} \phi_j} \]

where:

- \( P(z) \) is the Dirichlet distribution of the document topic proportion.
- \( P(\phi_i) \) is the Dirichlet distribution of the word distribution in topic \( i \).
- \( P(d|\phi) \) is the Dirichlet distribution of the document generated by topic \( d \).
- \( n_{d,i} \) is the number of words in document \( d \) that are generated by topic \( i \).
- \( \theta_{w,i} \) is the probability of word \( w \) in topic \( i \).
- \( \alpha \) and \( \beta \) are the parameters of the prior distribution of Dirichlet distribution.
α: The prior probability of each topic;  
β: The prior probability of each word;  
θ: The distribution probability of the topic of the i-th document;  
Z_{ij}: The subject of the j-th word of the i-th document  
W_{ij}: The jth word in the i-th document;  
Φ: The distribution probability of the z-th topic.  
① For document i, sample \( \theta(i) \) through Dirichlet(\( \alpha \));  
② For a subject z, through sampling from Dirichlet(\( \beta \)) to \( \phi(z) \)  
③ For each word W and the topic Z to which it belongs, abstract from the polynomial \( \theta \) to obtain \( Z_n=P(Z|\theta) \), and sample from the polynomial \( \phi \) to obtain \( W_n=P(W_n|Z_n, \phi) \)  

Here follows the formulas involved[3]:
\[
\theta(i)=\text{Dirichlet}(\alpha^*) \quad \theta(d)=\text{Dirichlet}(\alpha->)  
\beta(k)=\text{Dirichlet}(\eta^*) \quad \beta(k)=\text{Dirichlet}(\eta->)  
Z_{dn}=\text{multi}(\theta_d) Z_{dn}=\text{multi}(\theta_d)  
\]

LDA model can give us systematic classification for target texts, and balancing the velocity and quality of text classification while eliminating interferences significantly. In this thesis, we will use this model to analyze China’s courts’ judgments on the infringement disputes of the rights of new plant varieties from 2011 to 2020, and effectively analyze the core focuses of these infringement disputes, and provide data support and suggestions for the improvement of legislation in the right of new plant varieties.
2. The process of the experiment

We completed the data collection by gathering all the judgments in the name of “the Right of New Plant Varieties” from China Judgments Online from 2011 to 2020, whose sample size is 286.

Since the development of the right of new plant varieties in China has just started, the sample size of the infringement dispute judgments involved is small. So we consider setting the numbers of subjects K to 3-5. After comparison, we found that when the number is set to 3, the discrimination of the subjects is not strong enough, and there are even some subjects that are not included. When the number of subjects is set to 4, there is a relatively idealistic degree of discrimination, covering all technical fields.
When the number of subjects is set to 5, a large number of similar subjects are divided into different topics. Therefore, we finally set the number of subjects to 4, and the distribution of the top 4 keywords of each subject in the judgments is shown in Table 1.

Table 1. Frequency Distribution

| Topic | Frequency | Keywords |
|-------|-----------|----------|
| Topic 0 | 30 | New varieties, identity identification, traits, number of comparison sites |
| Topic 1 | 179 | Qualified entities, invalid authorization, exclusive management rights, unlicensed production |
| Topic 2 | 16 | Propagating materials, seeds, commercial grains, grafting |
| Topic 3 | 35 | Planting, self-use, seed-breeding agreement, license |

According to Table 1, it can be seen that Topic 0 is about how to identify "new varieties", Topic 1 is about the division of responsibilities between the stakeholders involved, Topic 2 is about how to interpret "propagating materials", and Topic 3 is about the exemption of the right of new plant varieties.

3. Analysis of the outputs

The keywords extracted by the LDA subject model overlaps Top Ten Classic Cases of Protection on the Right of New Varieties of Agricultural Plants [5] issued by the Department of Seed Industry and Development Center for Science and Technology of Ministry of Agriculture and Rural Affairs in September 2019 as high as 89.73%, which proves that it is highly feasible and accurate to use LDA subject model to analyze judgment texts. From the distribution of these keywords, we can understand the causes for the infringement cases of the right of new plant varieties and the focuses of disputes over the last 10 years, to guide anti-infringement legislation and law enforcement in the field of the seed industry in China.

4. Conclusion

The seed industry is a fundamental industry with strategic significance, which bears the mission of feeding the nation with stable crop production. It is essential for a variety of innovation to protect the intellectual rights of a variety of owners.

Through data dimensionality reduction and filtering out of irrelevant words on a large scale with LDA, the thesis manages to extract subjects more distinctly from case texts with multiple keywords and complete the classificatory analysis of the infringement judgments of the right of new plant varieties, explaining the causes of the infringement. It provides domestic seed intellectual legislation and law enforcement with clear direction for improvement, helping the seed industry shift more attention to quality from quantity.

Acknowledgments

This paper is funded by Law School, Shenzhen University. We would like to thank Professor Xianwei Zhang for his suggestions in writing this paper.

References

[1] Chuhan Ai, Xin Xiong, Jiande Wu, Classificatory Analysis of Patent Text Analysis Based on LDA Subject Model, Science, technology and industry, 2019(03):77-82.

[2] Information on https://blog.csdn.net/Daycym/article/details/88876460?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommend_From_Machine_LearnPai2-1.add_param_isCf&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommend_From_Machine_LearnPai2-1.add_param_isCf

[3] Information on https://blog.csdn.net/sphinxrascal168/article/details/82721946?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommend_From_Machine_LearnPai2-4.channel_param&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommend_From_Machine_LearnPai2-4.channel_param
[4] Information on https://wenshu.court.gov.cn/website/wenshu/181217BMTKHNT2W0/index.html?pageId=751a39dbfae61ec0c2a91f486f12ed72&s1=%E6%A4%8D%E7%A7%8B%E6%84%80%E9%A9%E6%96%B0%5B0%E5%93%81%E7%A7%8D

[5] Information on http://www.agri.cn/V20/ZX/nyyw/201909/t20190912_7193561.htm