Research Paper

Hot Topic Mining and Evolution Analysis of Express Policy Based on LDA

Chunbei Fang, Lihua Wu

(School of Information Management, Beijing Information Science & Technology University)
Corresponding Author: Lihua Wu

ABSTRACT: *The paper conducts a text mining on the national-level express policies in China and identifies seven topics implied in the express policy texts by using the Latent Dirichlet Allocation (LDA) topic model. By dividing the policy texts into different time phases, the paper further analyzes the evolution of the express policy topics with the topic intensity as the measuring index. The research shows that operation mechanism, appeal handling, express packing, credit management, epidemic defending, service optimization, as well as cooperation and development are the key areas of the express policy layout and the policy topic evolution takes on a development characteristic of the express delivery market from the basic operation to the business expansion and then to the service improvement.*
KEYWORDS: *Express Policy Text, Topic Mining, Topic Evolution, LDA model*

I. INTRODUCTION

At present, the express industry is becoming a new growth point enabling the global economy. According to the 2022 Global Economic Impact Report released by FedEx, under the influence of such factors as the COVID-19 and supply chain tension on the global economy, FedEx has an important economic impact on global countries, regional economies, and core local hub markets from six dimensions including direct impact, employment, facility investment, indirect impact, impact through the supply chain and extension of the supply chain.

The rapid development of the express industry is inseparable from the support of the express policy. From the perspective of the time dimension, policy documents are updated according to the continuous changes of domestic and foreign environment, rather than remain unchanged. Each session before the release of the express policy presents a dynamic cycle process. The scientific and systematic study on the express policy not only can expand the theoretical study, but also has practical significance for building and improving the policy system of the express delivery market.

Taking the express policy texts at the national level in China as the research object, this paper explores the express policy topics and their evolution. The remaining parts of the paper are organized as follows: First the text data of express policy are introduced. Then the extraction process of policy topics is executed by using the LDA topic model. Next, the topic evolution is analyzed according to the temporal division of policy texts. And finally, the conclusion is presented.

II. TEXT DATA OF EXPRESS POLICY

The policy text normatively records the documents generated by the government due to the needs of policy activities [1]. It is the basis for studying the content and system of the government policy. This paper collects 161 effective express policy texts at the national level between 1985 and 2021 with Python crawling from the official website of State Post Bureau of The People's Republic of China. The time distribution of the texts is shown in Figure 1.
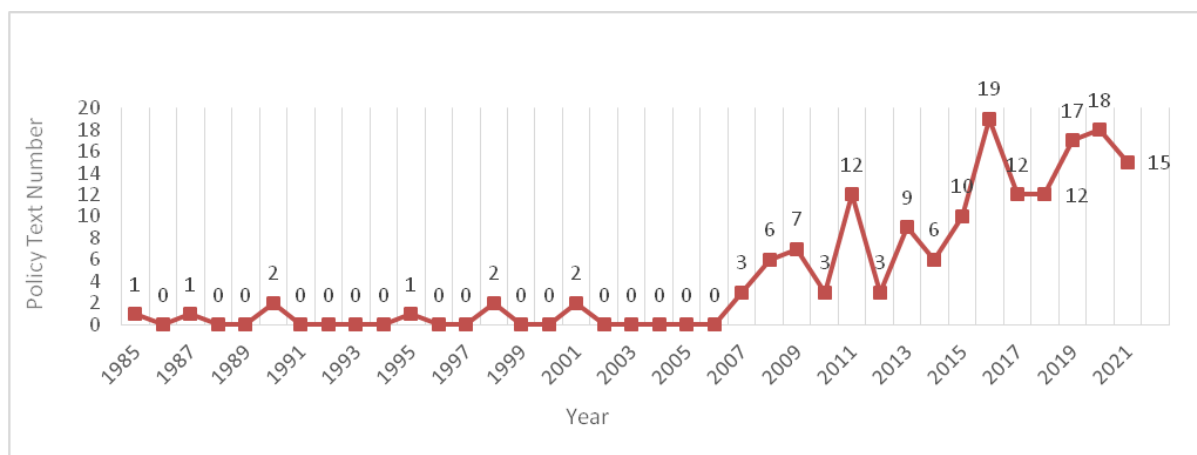

Figure 1: time distribution of express policy text

It can be seen from Figure 1 that the number of express policy texts was scarce before 2006, with only 0-2 each year. It is the stage that China's express industry started and slowly developed. The support from the government is insufficient. After 2006, the number of policy texts overall increased in a volatile manner. Especially in 2011, 2016 and after 2019, the number of policy texts presented three peaks, namely12, 19, and more than 15 respectively. The demand of e-commerce market has driven the development of the express delivery industry in China since 2011. Correspondingly, the government has given more support for the express industry.

III. TEXT TOPIC MINING

Text topic mining refers to mining the hidden topic information in the text from the given text data set using the relevant topic model, so as to meet the specific research needs [2]. The LDA topic model is a classic topic mining model [3]. The model assumes that in the text data set, each document contains multiple topics, and each topic is composed of words related to the topic. This model is often used in the research of topic mining in policy texts.

3.1 Policy Text Preprocessing

Text preprocessing refers to the structured processing of the original text data, the screening of effective information, and the formation of structured data through word segmentation, removal of stop words, etc [4]. Policy text preprocessing is one of the most important works for text mining methods. Through the collection and collation of the policy text data in the early stage, a policy document library with plain text is obtained. All the documents in the library only contain the policy name and the text content for the subsequent text preprocessing. The flow of text preprocessing is shown in Figure 2.
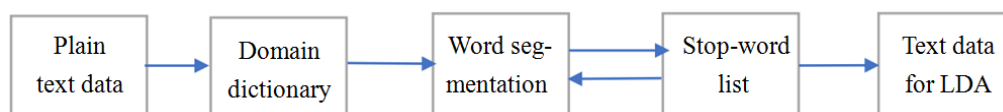

Figure 2: flow of policy text preprocessing

In Figure 2, the domain dictionary has the feature to express the key information of a professional domain through concise and short words. Its content essence is information extraction of text. In this paper, a new-word discovery algorithm based on mutual information and adjacency entropy is used to form a domain dictionary for the express delivery industry. The domain dictionary is loaded for improving the effect of word segmentation.

3.2 Parameter Estimation of LDA Model

The key step of LDA topic model is to estimate the parameters of the model and infer the topic probability distribution and polynomial distribution of feature terms. It is very difficult to directly solve the model parameters. The words in texts need to be used as observable variables. Therefore, the generation of topics is the process of solving the parameters of LDA model. Common parameter estimation methods include the variational inference algorithm, EM algorithm, Gibbs algorithm, etc. Wherein, Gibbs sampling method is a simple and widely applicable Markov chain Monte Carlo algorithm. The basic idea of the method is to construct

a Markov chain that converges to the target probability distribution, and extract samples that are close to the target distribution from the chain. Gibbs sampling method is used for the parameter estimation of LDA model in this paper.

3.3 Probability Distribution of Text Topic

The probability distribution of text topics is obtained after the topic modeling is conducted on the 161 express policy texts. The number of topics is determined to 7 through calculating the perplexity of topics by calling the related model of Python Gensim library. Table 1 gives an illustration of the probability distribution of each topic in a part of documents. The higher the probability value of a topic in a document is, the stronger the strength of the topic in the document.

Table 1: probability distribution of text topics in a part of documents

Document	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7
1	0.1961	0.0029	0.4901	0.0029	0.0029	0.1709	0.1341
2	0.1755	0.0035	0.0035	0.0035	0.0035	0.7498	0.0608
3	0.0373	0.0014	0.1566	0.0014	0.0014	0.5519	0.2502
4	0.6183	0.0023	0.0622	0.0023	0.0023	0.1735	0.1391
5	0.7503	0.0016	0.0016	0.0016	0.0016	0.1088	0.1344
6	0.1226	0.0016	0.0016	0.0016	0.0016	0.1893	0.6816
7	0.1375	0.0021	0.0021	0.0021	0.0021	0.6367	0.2175
8	0.1188	0.0015	0.0015	0.0015	0.0015	0.7834	0.0916
9	0.1511	0.0016	0.0016	0.4307	0.0016	0.1709	0.2426
10	0.0814	0.0017	0.0017	0.0017	0.5264	0.2199	0.1672

3.4 Keywords of Text Topic

The probability of each keyword in a topic is calculated further. Table 2 illustrates the top 10 keywords ranked by their probability values in each topic.

Table 2: top 10 keywords in each topic

Keyword	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7
1	mail	appeal organization	express packing	express credit	model city	smart express box	express service
2	express business	security check machine	adhesive tape	putting on file	disinfecting	terminal node	construction
3	express package	free trade	packing box	copy	epidemic	standard	development
4	operation	jurisdiction	packing material	express market	emergency event	international express	electronic commerce
5	licence	express delivery	green packing	market management	deadline	optimization	service
6	sender	complainant	recycle	credit system	place	material	cooperation
7	article	express company	postal industry emergency	emergency event	epidemic prevention	technology	manufacturing
8	branch	security check	alteration item	credit information	e-commerce platform	operation	plan
9	measure	defendant	directories	licence cancellation	accident	reform	employees
10	user	consumer	size	service quality	space	license approval	region

3.5 Policy Topic Definition

By combining the probabilities of keywords under each topic with the meanings of these words in the express industry, the seven topics are interpreted. Take the first topic as an example. The top ten keywords in the first topic are mail, express business, express package, operation, license, sender, article, branch, measure, and user. These words refer to the market subject, market object and such things related to the market access of the express industry. Therefore, the first topic is classified as "express operation mechanism". The other six topics are summarized in a similar manner. Finally, the seven topics are defined as operation mechanism, appeal handling, express packing, credit management, epidemic defending, service optimization, as well as cooperation and development. The seven topics manifest the key areas of express policy layout in China.

IV. TOPIC EVOLUTION ANALYSIS OF EXPRESS POLICY

4.1 Phase Division of Express Policy Text

The "five-year plan" is used as the time benchmark to divide the policy stages. The "five-year plan" is the most important public policy in China, which plays an important strategic guiding role in promoting the national development and the introduction of policies in various industries. As shown in Figure 1, the number of policies issued before 2010 was small, so the policies before 2010 were studied as a whole. The policies after 2010 were divided into five years as a cycle. Finally, the evolution period of the express policies is divided into four stages, namely 1985-2010, 2011-2015, 2016-2020 and 2021.

Table 3: phase division of express policy text

	Phase I	Phase II	Phase III	Phase IV
Time interval	1985-2010	2011-2015	2016-2020	2021
Number of policy	28	40	78	15

4.2 Policy Topic Intensity

In order to quantify the topic of the express policy more accurately and intuitively, the paper takes topic intensity as the measure index to evaluate the policy topic at each stage. Topic intensity is a probability in essence, which indicates the concerned degree of the current topic [5]. When topic intensity reaches a certain threshold, the corresponding topic can be called a hot topic. The topic intensity of each topic under different time windows is shown in Table 4.

Table 4: topic intensity of express policy in each stage

Stage Topic	Phase I (1985-2010)	Phase II (2011-2015)	Phase III (2016-2020)	Phase IV (2021)
Operation mechanism	0.3225	0.3057	0.1909	0.1104
Appeal handling	0.0455	0.0019	0.0055	0.0209
Express packing	0.0264	0.0351	0.0549	0.0712
Credit management	0.0548	0.0810	0.0620	0.0307
Epidemic defending	0.0869	0.0450	0.0444	0.0771
Service optimization	0.2084	0.1830	0.3398	0.3702
Cooperation and development	0.2556	0.3483	0.3025	0.3196

4.3 Evolution Analysis of Policy Topic

For Phase I, operation mechanism has the largest topic intensity according to Table 4 and so it is the hot topic in this stage. The express business, meanwhile, was separated from the postal business. In 2007, the reform of the postal system was implemented, and the domestic express business gradually rose. However, the previous license standard for the postal business operation was not fully applicable to the express business. Therefore, the focus of the express policy during this period was to standardize the access of the express market and manage the basic business operation of express.

For Phase II, the hot topic is cooperation and development. In this stage, the rapid development of e-commerce in China promoted the needs for express delivery. To support the rapid expansion of the express market, the Chinese government released a series of policies to strengthen the industrial cooperation, construct the express infrastructure and improve the service level. As a result, the network and informatization of express industry has significantly improved. The delivery service terminals and the professional employees have greatly increased too.

Phase III and Phase IV have the same policy topic that is the service optimization. Service optimization was carried out from aspects such as technology, operation, packaging material, international business, quality standard, etc. For example, State Post Bureau issued a guiding opinion on accelerating the construction of smart express boxes in 2020. Then, the contactless distribution service typically supported by smart express boxes has played an important role during the COVID-19 epidemic period.

V. CONCLUSION

The paper implements a topic mining and evolution analysis on the policy texts of express industry. The research shows that the topics including operation mechanism, cooperation and development, and service optimization have always been the focuses of express policies. By contrast, the policy attention to the other four topics obviously has been low. This is because appeal, packing, credit, and epidemic all are only specific aspects of business operation in express industry. They are not as broad in business coverage as the former three topics, but they are beneficial supplements to the three hot topics.

REFERENCES

[1].   Shao, Shijin, Logistics topic evolution and matching research based on LDA and FastText under the COVID-19. 2022.
[2].   Chen, Junyu, Research on online review application based on text mining. 2020.
[3].   Rashid, J., Shah, S.M.A., Irtaza, A., Fuzzy topic modeling approach for text mining over short text. Information Processing & Management, 2019, 56(6): p. 102060.1-102060.19.
[4].   Wang, Yongming, Construction of investor sentiment indicators based on text mining and its application in quantitative investment strategies. 2022.
[5].   Zhang, Zhi, Research on logistics service quality evaluation of express enterprises based on text emotion analysis. 2020.