Patent research front mining of GaN semiconductor based on the LDA model

Ruijiao Ma, *Yuntao* Zhang, *Jiakuan* Chen, and *Haoyu* Wen[*]

Department of Economics & Management, Xidian University, Xi'an, 710119, China

Abstract. This paper aims to identify and analyze research front topics of GaN semiconductor. The research method was to mine and identify topics based on GaN patent data by using the LDA model and topic intensity index. Finally, through experiments, we obtained and analyzed 4 research front topics. The results provided informatics support for revealing the research status and trend of GaN semiconductor.

1 Introduction

As one of the representatives of the third-generation semiconductor, Gallium Nitride(GaN) has good electrical characteristics, so it is considered the best material for studying short-wavelength optoelectronic devices and high-temperature and high-frequency and high-power devices. It has attracted wide attention in the optoelectronic field and microwave devices. At present, GaN semiconductor technology is in the ascendant and has become the strategic commanding heights of competition around the world. Therefore, the excavation of the research front in the GaN field, which has an extremely important scientific support and reference value for the strategic layout and scientific development of the third-generation semiconductor technology.

Patent is an important carrier of technological innovation and development information, and the patent data reflects the technological innovation and application achievements in the research field. This paper uses the LDA topic model of informatics to mine the research front in the patent data of GaN semiconductor, and to provide strong intelligence support for revealing the development status and trend in this field.

2 Correlational research

Conceptually, in 1965, Price, who was called "Father of Sciometrics", published in Science proposing that "Research Front is the area of research represented by frequently cited in scientific citation networks and recently published literature sets "[1]. In 1994, Persso believed that research front consisted of citation literature, which is an important knowledge basis for research front[2]. In 2010, Upham proposed that the forefront of research is a science and technology field that is highly concerned by scientists, and a high-tech technology that the government and investors believe has potential[3].

[*] Corresponding author: hywen@xidian.edu.cn

In terms of research methods, Blei et al. made Bayesian improvements on PLSI(Probabilistic Latent Semantic Indexing) model, then proposed LDA (Latent Dirichlet Allocation) topic model, and was widely recognized and used[4]. However, the LDA model cannot explain the evolution of research topics, then Blei have proposed a dynamic topic model to realize the detection and tracking of dynamic scientific research topics[5].

At present, many scholars use relevant methods to research the application of frontier identification. This paper uses LDA to mine the topics, and use the topic intensity as an index to measure the attention of research front topics, and then dig into the research front topics with high value.

3 Research method

LDA(Latent Dirichlet Allocation) is an unsupervised learning based on word bag model and a probabilistic topic model that is able to give the topic of each document in the document dataset as a probability distribution. The core idea is to represent each document as a probability distribution composed of a set of topics, while each topic is a probability distribution composed of a series of words, thus forming a three-layer Bayesian network model of "document-topic-word"[4].

In the LDA topic model, the determination of the optimal topic number K has a crucial impact on the effect of model topic recognition. The optimal number of topics needs to be determined by calculating the perplexity index. In theory, the lower the perplexity means, the better the topic model is. The number of topics that is too large or too small will have a bad impact on the outcome. Generally, the number of topics corresponding to the low perplexity or inflection point is the optimal number of topics.

The specific steps of research methods are as follows:

- Obtain the patent data about the GaN semiconductor.
- Data preprocessing, including text format conversion, punctuation removal, number removal, stop word removal, stem parsing, word bag creation, etc.
- Set the number of hyperparameters, and optimal topics of the LDA model.
- Extraction of the LDA topics was performed using the KNIME platform.
- Topic intensity is calculated and the number of patents included in each topic identified by the LDA model can reflect the study strength of each topic. If the topic intensity of the research topic is higher than the mean, the topic is highly attention.
- Analyze the research front topics.

4 Experimental process and results

4.1 Data acquisition and preprocessing

Patent data were obtained from the incoPat Global Patent database, with the retrieval time range of 2017~2020. Retrieval formula is TIAB= (GaN or gallium nitride) and the retrieval country was selected as the United States. Finally, 1,052 patent data were downloaded, with the title and abstract as the research corpus.

In this paper, KNIME platform and Python were used to preprocess the summary text data of the above data, including punctuation removal, digital removal, stop word removal and other processes. Related common high-frequency stop words in the domain were removed by Python, such as materials, semiconductor, etc.

4.2 LDA topic identification

When the gensim package of Python was used to calculate the perplexity, and the change result of k value of 1~20 is calculated iteratively. As shown in Fig 1, it is known that with the increase of the k value, the K value at the inflection point with low perplexity should be selected as the optimal number of topics, so 10 should be selected as the number of topics.

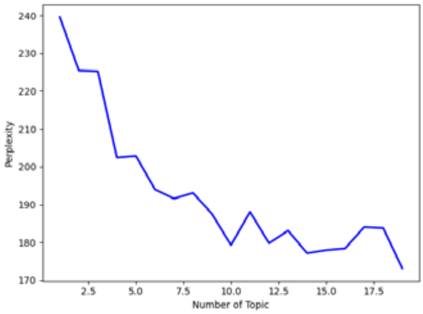

Fig. 1. Perplexity curve.

The LDA module of the KNIME platform was identified for topics, with specific parameters set to: No of topic: 10, No of words per topic: 10, Alpha=5, and Beta=0.01. The final obtained motif results are shown in Table 1.

Table 1. Patent topic identification results.

Topic	Topic words
topic_0	n-type| p-type| substrate| light| conduct| impure| algan| emit| light-emitting| diode
topic_1	substrate| film| metal| bond| epitaxy| wafer| surface| circuit| growth| thin| ga| conduct
topic_2	sensor| compound| transduce| barrier| crystal| two-dimensional| channel| 2deg
topic_3	structure| transistor| channel| silicon| stack| barrier| pattern| p-gan| oxide| polar
topic_4	transistor| switch| circuit| signal| termin| connect| driver| configure| drain| drive
topic_5	substrate| epitaxy| fabric| buffer| mask| crystal| sapphir| polar| pattern| thermal
topic_6	image| generet| network| data| train| model| adversari| learn| discrimin| neural
topic_7	light| led| emit| quantum| diode| wavelength| color| optic| photon| light-emitting
topic_8	electrode| barrier| substrate| contact| drain| channel| metal| surface| n-type| dope
topic_9	crystal| surface| silicon| aluminum| elem| composit| deposit| atom| nitrogen| indium

4.3 Topic intensity calculation

According to the topic-document distribution, the results of the topic intensity were obtained, as shown in Table 2. Taking average topic intensity 105 as a threshold, above 105 was considered high topic intensity.

4.4 Analysis of research front topics

According to the above methods, topic_0, topic_1, topic_4, and topic_7 are judged as front topics. The corresponding research contents of each topic are: topic_0-light-emitting diode, topic_1-crystal epitaxial growth, topic_4-switch device, and topic_7-LED lighting.

Light-emitting diodes and switching devices are both power electronics. GaN materials are currently widely used in a variety of electrical electronics and optoelectronic devices, and It has broad application prospects in high-performance servers, wireless base stations, solar power generation, new energy vehicles, smart grid and other aspects. Research in the field will always be a hot topic; The performance of the current GaN-based devices is far below the theoretical values. Therefore, high-quality GaN epitaxial materials are the technical core of high-performance GaN-based devices. The research in this field will always be a key technology in the GaN materials field; The semiconductor lighting industry has grown most rapidly in recent years, and form an industrial scale of ten billion dollars. At present, GaN optoelectronic devices have obvious competitive advantages in the application fields of optical storage, laser printing, high-brightness LED and wireless base stations. Among them, high-brightness LED is one of the most interested and concerned technology in the current field of device manufacturing.

Table 2. Results of the topic intensity calculation

Topic	Topic intensity
topic_0	121
topic_1	194
topic_2	47
topic_3	89
topic_4	127
topic_5	94
topic_6	80
topic_7	123
topic_8	83
topic_9	94
Average value	105.2

5 Conclusion

Taking the data of GaN semiconductor as the research object, the LDA model is used to mine the research topics in this field, and we identify the research front topics with high research heat by calculating the topic intensity index. The analytical experimental results show that the front topics such as LED lighting are one of the main research problems of GaN, which proves that this paper can effectively excavate the research front in the GaN field. At the same time, the topics excavated in this paper reveal the status and trend of GaN research, and provide a reference value for the development of scientific research in GaN.

This paper only uses the topic intensity index, and lacks the consideration of the topic novelty. The next step is to consider the topic novelty from the time dimension and improve the prospective exploration of research topics.

References

1. Price D J D S. Networks of Scientific Papers. J. Science, **149**(3683), 510-515 (1965)
2. Persson O. The Intellectual Base and Research Fronts of JASIS 1986-1990. J. J MACH LEARN RES, **45**(1), 31-38 (1994)
3. Upham S P, Small H. Emerging research fronts in science and technology: patterns of new knowledge development. J. Scientometrics, **83**(1), 15-38 (2010)
4. Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. J. J MACH LEARN RES, **3**, 993-1022 (2003)
5. Blei D M, Lafferty J D. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning,* New York: ACM Press, 113-120 (2006)