Research of Topics Discovery and Tech Evolution Based on Text Preprocessed Latent Dirichlet Allocation Model

Research Topic Analysis in GaN Tech Field

Wang Li[1, 2], Shen Xiang[1,2], Liu Xiwen[1, 2]
[1]National Science Library, Chinese Academy of Sciences
[2] Department of Library, Information and Archives Management,
University of Chinese Academy of Sciences
Beijing, China
E-mail: {wangle, shenx, liuxw}@mail.las.ac.cn

Abstract—**Computational Science and Data Science are inspiring the intelligent analysis and information service today. Machine learning text analysis is changing the traditional analysis methods. This article discusses the benefits of unsupervised learning approaches in patent text mining. Patent data of GaN industry were preprocessed by filter model based on NLTK Toolkit to identify the tech terms and then clustered them based on Latent Dirichlet Allocation model to find the latent topics which were visualized. Based on group operation, new emerging terms ranked by TFIDF through every year were used to reveal the research and development focused evolution. This research offers a demonstration of the proposed method based on 26,854 GaN patents. The results show 20 Research and Development topics with tech terms in GaN industry and present a Research and Development focus evolution based on new emerging terms every year, which provides a clue for more detaied analyses later. Our results show an efficent way to find technology focused evolution from a large scale text data.**

Keywords- LDA; automatic term identification; preprocessed text; visualization.

I. INTRODUCTION

As an unsupervised learning method, the Latent Dirichlet Allocation (LDA) is widely used for topics finding in large text analysis. Topic model is a generative model for documents which are mixtures of topics comprising words over probability distribution. Traditionally, words were used to construct an LDA model, which resulted in quite a lot of general words on top of each topic. Herein, the noun terms are utilized instead of words to discover patterns of term-use and the documents relationship.

In the Derwent Innovation Index (DII) database, original patent titles and abstracts are rewritten in English and the technology details including patent novelty, use, advantage and so on from patent full text are extracted. In this paper, based on preprocessed text dataset of 26,854 patent titles and abstracts about GaN technology field from DII, the research topics were discovered, and R&D focus changes were detected and visualized.

Researches about R&D changes or evolution based on LDA have focuses at topic level. T. L.Griffiths et al. write about a method identifying 'hot topics' or 'cold topics' [1]. D. Choi et al. explore technological trends based on patent share and their change at the topic level [2]. X. C. Gong et al. detect topic splitting and merging based on the LDA Model [3]. J. B. Qu et al. analyze topic evolution with topic relevance from adjacent time intervals [4]. Many researches have improved and practiced methods detecting R&D changes or evolution at topic level, while few have discussed finer granularity analyzing at term level.

II. TEXT PREPROCESSING

Since most terms have the syntactic form of a noun phrase [5], identifying the noun phrases in the text was executed during text preprocessing. Part-Of-Speech Tagging in Python NLTK was used to construct language filter and identify noun phrases as following:

1. The sequence consists of nouns, v-ing form and adjectives, such as the phrase 'device comprising virtual display system'.
2. The sequence ends with a noun or a v-ing form, such as the phrase 'distributing workflow' or 'business computing.

Additionally, stop contents were manipulated in the Python script from three different levels: sentence, phrase and word. For example, the publisher information sentence such as '(C) 2018 Elsevier B.V. All rights reserved' and the patent text description phrases such as 'independent claim' were stopped. Basically, uppercase and lowercase, singular and plural nouns and so on are preprocessed on word level.

After text preprocessing, the terms were prepared for LDA model.

III. RESEARCH TOPICS FINDING AND VISUALIZING

A. Research Topics Finding

The Gibbs sampling algorithm was used, with $\beta=0.1$, $\alpha=50/T$, (T is the number of topics) [6]. In practical application, β is relatively small and words can be expected into a specific research topic [1]. Since GaN field is already a specific area, fewer topics are involved in this case. Because the value of T in is very small, less than 30, topics for different T were discriminated manually to avoid overlap between topics in macro level. Finally, 20 topics were suitable for GaN patent data, as shown in Table 1.

TABLE I. GAN RESEARCH TOPICS BASED ON LDA MODEL

Topic1	**Score**	**Topic 2**	**Score**	**Topic 3**	**Score**
layer	0.0919	substrate	0.0167	gate electrode	0.0258
gallium	0.0428	material	0.0097	drain electrode	0.0192
buffer layer	0.0399	diode	0.0077	source electrode	0.0177
substrate	0.0344	array	0.0066	source	0.0165
aluminum	0.0192	device	0.0061	barrier layer	0.0163
Topic 4	**score**	**Topic 5**	**score**	**Topic 6**	**score**
active layer	0.0632	substrate	0.0676	quantum dot	0.0066
light emitting device	0.0241	growing	0.0161	gallium arsenide	0.0057
emitting device	0.0197	layer	0.0133	indium	0.0054
semiconductor layer	0.0167	gallium	0.0111	composition	0.0051
p-type semiconductor layer	0.0150	epitaxial layer	0.0105	indium phosphide	0.0049
Topic 7	**score**	**Topic 8**	**score**	**Topic 9**	**score**
substrate	0.0185	layer	0.0188	light	0.0168
temperature	0.0162	manufacture	0.0165	wavelength	0.0104
growing	0.0115	nitride semiconductor layer	0.0148	light source	0.0074
nitrogen	0.0098	group	0.0107	light-emitting device	0.0071
heating	0.0089	thickness	0.0094	phosphor	0.0052
Topic 10	**score**	**Topic 11**	**score**	**Topic 12**	**score**
group	0.0507	forming	0.0407	aluminum	0.0277
crystal	0.0205	substrate	0.0337	silicon	0.0245
manufacture	0.0192	surface	0.0180	titanium	0.0142
gallium	0.0140	etching	0.0152	silicon carbide	0.0138
single crystal	0.0108	removing	0.0099	zinc	0.0137
Topic 13	**score**	**Topic 14**	**score**	**Topic 15**	**score**
substrate	0.0276	device	0.0216	substrate	0.0242
active layer	0.0111	diode	0.0095	second electrode	0.0104
semiconductor laser	0.0109	semiconductor element	0.0079	material	0.0102
surface	0.0105	circuit	0.0070	first electrode	0.0093
direction	0.0097	anode	0.0060	electrode	0.0092
Topic 16	**score**	**Topic 17**	**score**	**Topic 18**	**score**
substrate	0.0405	layer	0.0515	semiconductor layer	0.0258
surface	0.0344	chip	0.0231	light emitting element	0.0189
wafer	0.0226	p-type layer	0.0164	electrode	0.0179
gallium	0.0200	sapphire substrate	0.0159	light emitting diode	0.0136
laser beam	0.0061	n-type layer	0.0148	compound semiconductor	0.0097
Topic 19	**score**	**Topic 20**	**score**		
substrate	0.0370	layer	0.0258		
gallium	0.0272	active region	0.0152		
manufacturing	0.0263	device	0.0138		
surface	0.0175	first layer	0.0138		
thin film	0.0152	second layer	0.0120		

B. Research Topics Visualization

Based on LDA model, the metric for terms and topics was measured and used to calculate the similarities between terms. A visualization map was constructed by applying Multidimensional Scaling to the similarities [6]. 20 topics in the GaN field were visualized, as shown in Figure 1. The threshold value for terms showed in the map was 0.001 in this case.

IV. R&D FOCUS EVOLUTING

The metric θ of topics and documents was used to find the topic contributing the most to every document. $\theta_{i,j}$ can reveal the degree to which topic i is referred to in the document j, (1). p(topic=i | θ) according to Dirichlet distribution ($\theta_{i,j} \geq 0, \sum_{i} \theta_{i,j} = 1$) [7]. The most contributed topic was assigned for every document in this case.

$$\theta_{m\times k} = \begin{array}{c} \begin{array}{cccc} topic_0 & topic_1 & \dots & topic_k \end{array} \\ \begin{bmatrix} \theta_{0,0} & \theta_{0,1} & \cdots & \theta_{0,k} \\ \theta_{1,0} & \theta_{1,1} & \cdots & \theta_{1,0} \\ \vdots & & \ddots & \vdots \\ \theta_{m,0} & \theta_{m,1} & \cdots & \theta_{m,k} \end{bmatrix} \end{array} \begin{array}{c} \\ doc_0 \\ doc_1 \\ \vdots \\ doc_k \end{array} \quad (1)$$

The evolution of R&D focus through new terms emerging in every year was observed. All documents were grouped by year, and terms in a year's documents were counted. $Terms_y$ means terms in year. Then, (2) was used to extract new emerging terms in year y, E_y , ranked by sum of TF-IDF scores.

$$E_y = Terms_y - \sum_{n=y0}^{y-1} Terms_n. \quad (2)$$

In practice, the top technical terms are ranked and identified by term frequency and TF-IDF value. But there are a large number of high frequency general terms by term frequency rank while the technical terms obtained by TF-IDF are more meaningful. Based on (2), new emerging terms were counted from 2011 to 2017 every year in GaN field, as shown in Figure 2.

 ISBN: 978-1-61208-700-9

V. CONCLUSIONS AND FUTURE WORK

The analysis model proposed in this paper gives an efficient way to find technology focus evolution from a large scale text data, such as patent information in this case. Compared with topic level analysis, the tech evolution provides a breakthrough point for finer granularity analyzing to discover hot tech researches on a timescale which could be a clue for finding more information about these tech focuses. In the future, we will continue to optimize the analysis model in practice, especially the processing of synonyms.

REFERENCES

[1] T. L. Griffiths and M. Steyvers, "Finding scientific topics," PNAS, vol. 101, pp. 5228-5235, 2004.

[2] D. Choi, B. Song. "Exploring Technological Trends in Logistics: Topic Modeling-Based Patent Analysis," Sustainability, vol. 10, pp. 2810-2835, 2018.

[3] X. C. Gong and X. Y. An, "A Research of Topic Splitting and Merging Detecting in the Medical Field Based on the LDA Model," Library and Information Service, vol. 61, pp. 64-74, 2017.

[4] J. B. Qu and S. Y. Ou, "Analyzing Topic Evolution with Topic Filtering and Relevance". Data Analysis and Knowledge Discovery, vol. 2, pp. 64-75, 2018.

[5] D. Bourigault, "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases," In Proc. of COLING the 14th conference on Computational linguistics, pp. 977-981, 1992.

[6] L. Wang, L. X. Zou and X. W. Liu, "Research of Correlation Information Mining and Visualizing Based on the LAD model," Data Analysis and Knowledge Discovery, vol. 2, pp. 98-106, 2018.

[7] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research. vol. 3, pp. 993-1022, 2003.

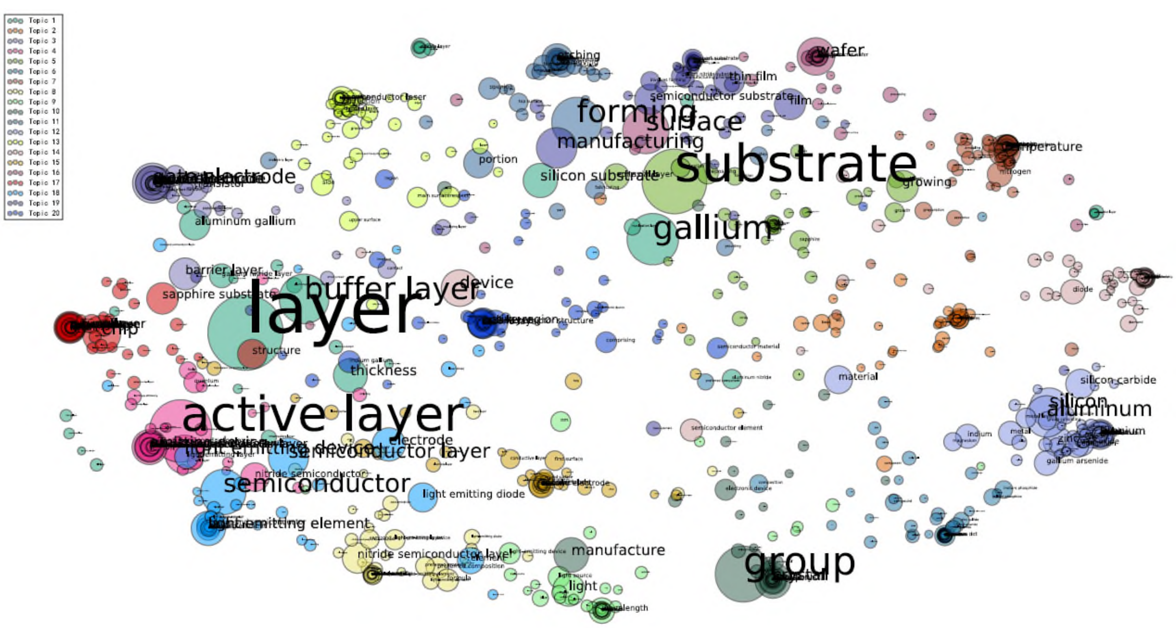

Figure 1. Visualization of GaN research topics

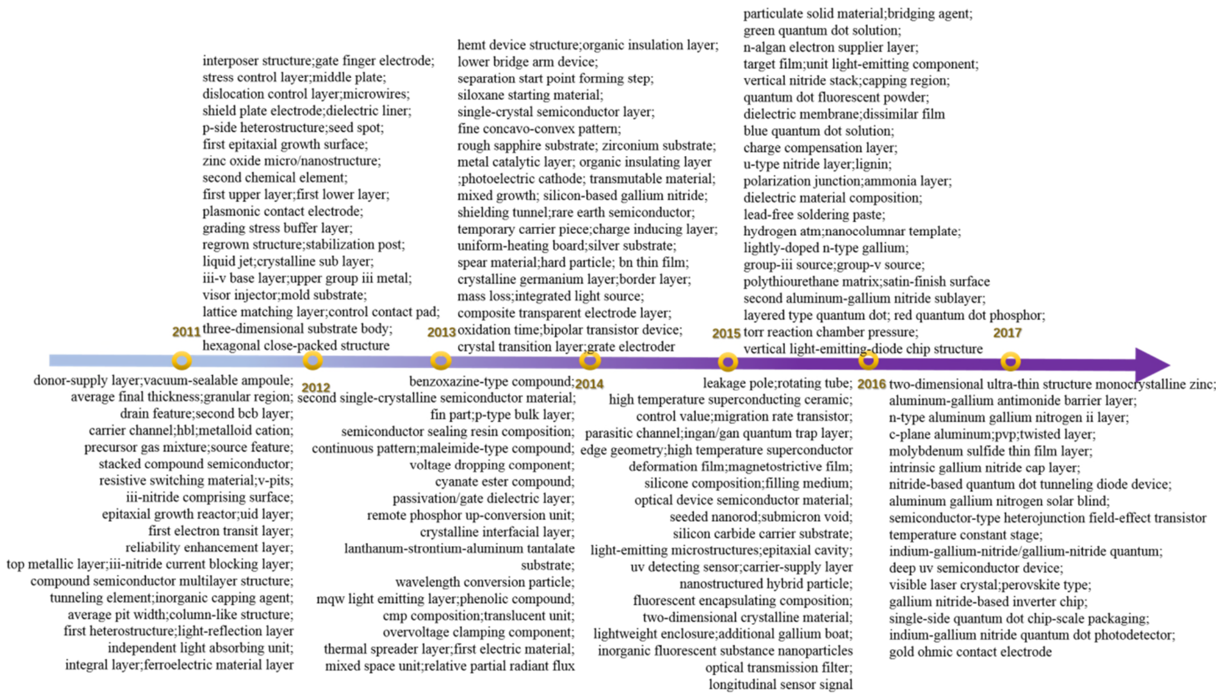

Figure 2. GaN Tech evolution based on new terms from 2011 to 2017