ISSN: 2688-9323

Research on Domain Patent Innovation Topic Forecast based on Structure Hole Theory

Hongsheng Xuan

School of Economics and Management, Xidian University, Xi'an 710126,China

Abstract

Domain patent data is an effective source of information for identifying technological innovation topics. By digging the subject of patent innovation in the field, it can provide reference for the prediction of the direction of technological innovation in the field. This paper combines LDA (Latent Dirichlet Allocation) topic model with structural hole theory, and proposes a quantitative method for mining patent innovation topics in the field and predicting the direction of technology development in the field.First, TF-IDF, means of Perplexity and quartile method were used to construct the LDA topic model of the domain patent. Secondly, the Quartile topic-feature word probability distribution matrix is used to associate the same feature words between topics to generate a topic-associated word adjacency matrix, and then binarize the adjacency matrix. Then, draw a domain patent subject network diagram, introduce the structural hole theory to mine innovative topics, and use the PPMCC to make relevance judgments. Finally, according to the number of structural holes in each topic's patent domain network, the topic of technological innovation in this field is mined, and the trend of technological innovation in the field is predicted. The experimental data selected chip patents in the past 5 years, combined with the results of the experimental analysis, and used the structural hole theory to realize the quantitative mining of chip patent innovation topics. The experts in the field of microelectronics were invited to complete the prediction of chip technology development based on the mining topics.

Keywords

Patent , LDA, topic network, structural hole, Innovation topic.

1. Introduction

With the proposal of speeding up the construction of an innovation-oriented country in China, enterprises and scientists pay more attention to patent as an important information source of technological innovation.Each patent contains all the information related to the invention and has become an effective source of information to identify the subject of technological innovation and to support the decision-making and research direction of enterprises and research institutions.However, due to the continuous emergence of patent resources in the field, it is increasingly difficult to mine technical topics. How to simplify the process of mining technical topics is an important issue for information intelligence personnel to pay attention to.

At present, scholars at home and abroad have proposed a variety of algorithms and implementation frameworks for patent technology subject mining. For example, wang lingyan and fang shu proposed a framework to identify emerging technology topics from patent literature in 2010 [1].In 2015, Hayoung Choi et al. proposed an algorithm to identify potential technological innovation themes in patents[2]. In 2017, Chen wei and Lin chao-ran proposed to mine the evolution trend of patent text technology subject based on lda-hmm[3]. In recent years, some scholars have focused on the evolution of patent technology themes in the field

ISSN: 2688-9323

[4-7]. scholars have mined meaningful technical topics from domain patents, which has gained wide attention and provided new ideas for the mining of domain patent technical topics. Relevant scholars have made a lot of effective research results on the application of structural hole theory[8]. Based on patent data, Wang et al. investigated the influence of structural hole and dot centrality on exploratory innovation of developers[9]. Ingawale et al. used network analysis to examine the quality of user-generated content on Wikipedia and found that high-quality articles clustered across structural holes[10]. Li and Liu studied the impact of structural holes on innovation in academic networks and found that when structural holes are in the middle level, they will have the greatest impact on the quantity and quality of innovation output[11].Information behavior is also an important object of study in the field of library and information. Researchers have conducted diversified studies on the influence of structure holes on information behavior[12-16]. Based on the application of structural hole, we draw the conclusion that structural hole is an important concept of structural analysis in network structure theory, and its theory and method is a powerful supplement to the method of information analysis.

Through the analysis of the above literatures, we found that the current research on the application of structural hole theory [8] to explore patent innovation in the field has not been in-depth.In order to make up for the shortage of current research, this paper proposes a method of mining patent innovation subject based on structure hole theory.First of all, in order to better optimize the discovery process of technical topics, tf-idf is used in this paper to weight feature words[17], and the optimal number of topics in LDA is determinedby using the degree of confusion[18] , and the quartile method is introduced to optimize the probability distribution matrix of feature words [19] . Then, according to the theme-feature matrix output by LDA, and based on the same feature words among topics, the adjacency matrix of all topic association relations is constructed, and the appropriate threshold is selected for binarization processing[20]. Finally, the paper draws the network of patent subject in the field, introduces the structure hole theory, and excavates the theme with innovative value, thus completing the development prediction of chip technology.

2. Relevant Theoretical Background

2.1. Latent Dirichlet Allocation(LDA)

2.1.1. Sub-section Headings

Blei et al. proposed the LDA(Latent Dirichlet Allocation) thematic model based on statistics in 2003, which was composed of a three-layer generative bayesian network [21].The graphical representation of the LDA theme model is shown in figure 1.

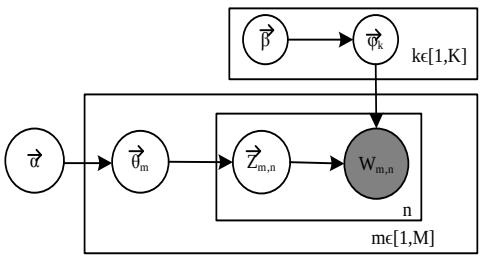

Figure 1: the directed probability graph of LDA

In the figure above, M represents the number of documents; K represents the number of topics; Represents the superparameter of the document-topic distribution; Hyperparameter representing the distribution of theme-word; M represents the parameter marker of

ISSN: 2688-9323

document m topic distribution. There are m types of "document-topic" distribution. K refers to the word item distribution parameter markers of topic k. There are k types of "theme-word" distribution.

Wm,n represents the NTH word in document m; Zm,n represents the topic to which the NTH word in document m belongs. The dark gray circle represents the observable variable, and the white circle represents the hidden variable. The conditional dependence between the two variables indicates that; The black box represents the number of repeated samples, and the letters in the box represent the number of repeated samples. Due to LDA's advantages in patent text analysis[23-28], we will generate the probability distribution matrix of patent document-topic and theme-feature words based on LDA probability topic modeling.

2.2. Structural Hole

In 1992, Ronald s. Burt (referred to as Burt) proposed the structural hole theory [8], which provides an in-depth explanation of the key position of the individual in the group. Burt pointed out that the best strategy for individuals is to find structural holes in the group's structural network, and then cross the structural holes so that groups that were not related to each other form a connection, while individuals themselves become the media of information flow. Burt reasoned that individuals in structural holes gain more competitive advantage and innovation through information filtering, which is a tool for discovering unknown horizons. In recent years, various disciplines and fields have gradually realized the application value of structural hole theory in their own academic fields. For example, a paper [29] published by Burt in 2004 on the theory of structural hole has been widely cited, and the cited literatures are distributed in the fields of business economy, computer science, library and information.

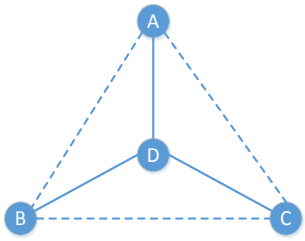

Figure 2: Structural hole concept map

As shown in the figure above, if nodes A, B, C and D are connected by each other, there is no structural hole.If nodes A, B and C are not directly connected but established through node D, if node D is removed, there is A structural hole between nodes A, B and C, and the node D at the position of the structural hole is called the structural hole node.

The three commonly used algorithms to measure the number of structural holes include the structural constraint algorithm [30], Effective size algorithm [30], and mediation centrality algorithm [31].

Structure Constraint algorithm [30]. The Network Constraint index proposed by Burt in 1992 measures Network closure and structure hole, which describes how closely a node in a Network is connected directly or indirectly with other nodes.The higher the coefficient, the higher the network closure and the fewer the structure holes.Effective size algorithm [30], the idea of which is derived from part of Burt's definition of structural holes, that is, if there is no redundancy relationship between them indirectly, then the gap between them is the structural hole.When the effective scale is larger, the redundancy of the network is lower and the possibility of structure holes is higher. Intermediary centrality algorithm [31], proposed

ISSN: 2688-9323

by freeman in 1979, is used to measure the degree of individual control over resources.The idea of this algorithm is: if an individual is in the shortest path of many other pairs of individuals, the individual has a high degree of intermediary centrality.Assumptions :(1) weight coefficients of each line are equal;(2) information always travels along the shortest path in the network.In terms of the measurement of structural holes, both the structural constraint algorithm and the mediation centrality algorithm are adopted [32]. The actors occupying the structural holes tend to have a high mediation centrality in their social networks.

3. Technical Route

The technical route of the innovation theme mining method in this paper is divided into four stages: data preparation stage, data processing stage, binarization stage and innovation theme mining stage. The writing of this chapter will be carried out in the logical order of the technical route.

3.1. Data Preparation Stage

The data preparation stage is mainly divided into two parts: data preprocessing and the construction of vector space model.Data preprocessing: firstly, word segmentation is carried out on the corpus, then stop words and stem reduction are removed, and finally punctuation marks, special symbols and Numbers are removed.

3.2. Data Processing Stage

In the data processing stage, the main parts are: the construction of LDA topic model and the quadratically processing of theme-feature matrix.Model (1) build the LDA theme: the first to use the method based on degree of confusion [21] to determine the optimal number of theme, the data set is divided into training set and test set, use the TF - IDF weighted processing of data set, the LDA model is established using the weighted after training set to model after the training, such as the test sets as corpus calculating the confusion of the LDA model under different themes, finally confused degree the most hours of topics as a model of the optimal number of topics; Then the LDA topic model is formally constructed.

3.3. Inarization Stage

The binarization stage is mainly divided into three parts: the correlation statistics of the same feature words between subjects, the construction of adjacency matrix, and the binarization processing of adjacency matrix.

3.4. Innovation Theme Mining Stage

The innovation theme mining stage is mainly divided into three parts: the drawing of the domain patent theme network map, the measurement of the structure hole and the mining of the domain patent innovation theme.

4. Experimental Verification

4.1. Data Acquisition and Preprocessing

The experiment verified that the corpus selected Patent documents in the chip field, and downloaded 9197 English Patent documents in the chip field from 2014 to 2018 from the Patent database Total Patent. The retrieval expression is Ti:(integrated circuit OR microcircuit OR microchip OR chip);The downloaded patent document entry includes title, abstract, IPC classification number, and claim. The corpus for constructing the LDA topic model USES the abstract from the chip patent document entry, and USES Python's natural language toolkit - NLTK[23] to complete the preprocessing of the abstract.

ISSN: 2688-9323

4.2. Build the Topic Model

4.2.1. Determine the Optimal Number of Topics

Before constructing the LDA topic model, in addition to preparing the preprocessed corpus, the optimal number of topics should also be determined. In this paper, toolkit -sklearn[28] was used to calculate the confusion value of the number of topics between 0 and 500 to determine the optimal number of topics.

4.2.2. Construct Vector Space Model

In the construction of the vector space model, toolkit-sklearn [24] was used to set the number of characteristic words as 2000, and then the word frequency (TF) matrix, inverse text (IDF) matrix and tf-idf matrix were generated. The number of columns and columns of the three matrices were 9197 and 2000.

4.2.3. Build the LDA Topic Model

Read the tf-idf matrix, set the number of LDA topics as 20, and the number of iterations as 50.The construction of the LDA model USES the LDA toolkit [28] to generate the document-topic (doc-topic) probability distribution matrix and topic-term probability distribution matrix.

4.3. Binarization Adjacency Matrix

4.3.1. Correlation Statistics of the Same Feature Words between Subjects

According to the quadratically differentiated theme-feature word probability distribution matrix, write a program using Python's data analysis library, Pandas[34].

4.3.2. Binarization Processing of Adjacency Matrix

According to binarization rules defined in section 3.3, binarization of adjacency matrix is realized.

4.4. Explore Patent Innovation Themes in the Field

4.4.1. Draw the Network Diagram of Patent Subject in the Field

Based on the data stored in the binarization adjacency matrix in section 4.3.2.

4.4.2. Structural Hole Measurement of Domain Patent Subject Network

According to the research ideas proposed in section 3.4, three network structure indexes, namely structural constraint, effective scale and intermediary centrality, are measured respectively by using the software UCINET 6[35].

4.4.3. Predict the Technical Development Direction of the Field

Five identified innovation themes are extracted from the four-differentiated theme-feature word matrix to form a summary table of innovation themes, as shown in table 4-10. Each innovation theme is explained by 32 feature words.

5. Conclusion

This paper proposes a mining area of patent innovation subject and predict technology development direction in the field of quantitative methods, the method of the core is to combine the LDA and the structure hole theory, mining provides a new way of thinking for innovative themes, its aim is to dig out the theme of the current phase of the most innovative value and then complete the development of technology in the field of prediction.

References

[1] Wang lingyan, Fang shu, Ji peipei. Research on the technical framework of identifying emerging technology topics using patent documents [J]. Competitive intelligence, 2010: 74-78.

[2] HAYOUNG CHOI, SEUNGHYUN OH, SUNGCHUL CHOI, et al. Innovation Topic Analysis of Technology:The Case of Augmented Reality Patents[J]. IEEE Access, 2018(6): 16119 - 16137.

[3] Chen wei, Lin chao-an, li jin-qiu, et al. Evolution trend analysis of patented technology subject based on lda-hmm -- a case study of Marine diesel engine technology [J]. Journal of intelligence, 2018, 37(7): 732-741.

[4] Wu feifei, Zhang yalu, Huang lucheng, et al. Multi-dimensional dynamic evolution analysis of technology subject based on AToT model - a case study of graphene technology [J]. Library and information work, 2017, 61(5): 95-102.

[5] Lioule method, le fugang. Research on patent technology evolution based on LDA model and classification number [J]. Modern intelligence, 2017, 37(5): 13-18.

[6] Yi huifang, wu hong, li chang, et al. Evolution of graphene technology based on subject life cycle and technological entropy [J]. Journal of information, 2019, 38(2): 64-70.

[7] Liu meijia. Research on the evolution of RFID technology based on patent analysis[D]. Beijing: Beijing university of technology, 2013: 4-82.

[8] Andrews S B. Structural Holes - The Social Structure of Competition-Burt, RS[J]. Ad-ministrative Science Quarterly, 1995, 40(2): 355-358.

[9] WangC L, Rodan S, Fruin M, et al. Knowledge Networks, Collaboration Networks, and Explor-atory Innovation[J]. Academy of Management Journal, 2014, 57(2): 484-514.

[10] Ingawale M, Dutta A, Roy R, et al. Network Analysis of User Generated Content Qu-alityin Wikipedia[J]. Online Information Review, 2013, 37(4): 602-619.

[11] Li Y Q, Liu C H. The Role of Network Position, Tie Strength and Knowledge Divers-ity in Tourism and Hospitality Scholars' Creativity[J]. Tourism Management Perspectives, 2018,27:136-151.

[12] Di Vincenzo F, Hemphala J, Magnusson M, et al. Exploring the Role of Structural Holes in Learning: An Empirical Study of Swedish Pharmacies[J]. Journal of Knowledge M-anagement, 2012, 16(4): 576-591.

[13] Bizzi L. The Dark Side of Structural Holes: A Multilevel Investigation[J]. Journal ofManagement, 2013, 39(6): 1554-1578.

[14] Markoczy L, SunSL, PengM W, et al. Social Network Contingency, Symbolic Manag-ement, and Boundary Stretching[J]. Strategic Management Journal, 2013, 34(11): 1367-1387.

[15] Aarstad J. Structural Holes and Entrepreneurial Decision Making [J]. EntrepreneurshipResearch Journal, 2014, 4(3): 261-276.

[16] Figueiredo C, Chen W H, Azevedo J. Central Nodes and Surprisein Content Selectionin Social Networks[J]. Computersin Human Behavior, 2015, 51: 382-392.

[17] Pang jianfeng, Bu dongbo, Bai shuo. Research and implementation of text automatic classification system based on vector space model [J]. Computer application research, 2001(9): 23-26.

[18] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[19] Paola Della Roccaa, Sabrina Senatorea, Vincenzo Loia. A semantic-grained perspective of latent knowledge modeling[J]. Information Fusion , 2017 (36): 52–67.

[20] Jiang Ming, liu hui, huang huan. Research on image binarization technology [J]. Software guide, 2009, 8(4): 175-177.

[21] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[22] Zhang han, Xu shuo, Qiao xiaodong. Review on the development of thematic models integrating the internal and external characteristics of scientific and technological literature [J]. Journal of information technology, 2014, 33(10): 1108-1120.

[23] Lioule method, Le fugang, Zhu yalan. Application of LDA model in patent text classification [J]. Modern intelligence, 2017, 37(3): 35-39.

[24] Fan yu, Fu hongguang, wen yi. Patent information clustering based on LDA model [J]. Computer application, 2013, 33(S1): 87-89, 93.

ISSN: 2688-9323

[25] Wang bo, Liu shengbo, Ding kun, ,et al. Patent content analysis method based on LDA theme model [J]. Scientific research management, 2015, 36(3): 111-116.

[26] Yang chao, zhu donghua, wang xuefeng. Analysis of patent technology subject - LDA model based on SAO structure [J]. Library and information work, 2017, 61(3): 86-96.

[27] Won Sang Lee, Eun Jin Han, So Young Sohn. Predicting the pattern of technology convergence using big-data[J]. Technological Forecasting & Social Change, 2015(100): 317-329.

[28] Mujin Kim, Youngjin Park, Janghyeok Yoon. Generating patent development maps for technology monitoring using[J]. Computers & Industrial Engineering, 2016(98): 289-299.

[29] Burt R S. Structural Holes and Good Ideas[J]. American Journal of Sociology, 2004, 110(2): 349-399.

[30] Ronald S. Burt. Structural Holes[J]. The social structure of competition,1992: 18, 55.

[31] Freeman L C. A set of measures of centrality based on betweenness. Sociometry, 1977 (1):35-41.

[32] Martin Kilduff. Social Networks and Organizations. 2003.

[33] Paola Della Roccaa, Sabrina Senatorea, Vincenzo Loia. A semantic-grained perspective of latent knowledge modeling[J]. Information Fusion , 2017 (36): 52–67.

[34] NumFOCUS. pandas: Python Data Analysis Library. https://pandas.pydata.org/.

[35] Borgatti, S. P. , M. G. Everett, and L. C. Freeman. Ucinet 6 for Windows :Software for Social Network Analysis Harvard: Analytic Technologies. 2002.