Article

Technology Hotspot Tracking: Topic Discovery and Evolution of China's Blockchain Patents Based on a Dynamic LDA Model

Jinli Wang [1], Yong Fan [2], Hui Zhang [3] and Libo Feng [4,5,*]

1   Faculty of Management and Economics, Kunming University of Science and Technology, Kunming 650093, China; wjinli@stu.kust.edu.cn
2   Institute of Humanities and Social Sciences, Kunming University of Science and Technology, Kunming 650093, China; fy2004@kmust.edu.cn
3   State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China; hzhang@buaa.edu.cn
4   School of Software, Yunnan University, Kunming 650500, China
5   Engineering Research Center of Cyberspace, Yunnan University, Kunming 650500, China
*   Correspondence: fenglibo@buaa.edu.cn

Abstract: Tracking scientific and technological (S&T) research hotspots can help scholars to grasp the status of current research and develop regular patterns in the field over time. It contributes to the generation of new ideas and plays an important role in promoting the writing of scientific research projects and scientific papers. Patents are important S&T resources, which can reflect the development status of the field. In this paper, we use topic modeling, topic intensity, and evolutionary computing models to discover research hotspots and development trends in the field of blockchain patents. First, we propose a time-based dynamic latent Dirichlet allocation (TDLDA) modeling method based on a probabilistic graph model and knowledge representation learning for patent text mining. Second, we present a computational model, topic intensity (TI), that expresses the topic strength and evolution. Finally, the point-wise mutual information (PMI) value is used to evaluate topic quality. We obtain 20 hot topics through TDLDA experiments and rank them according to the strength calculation model. The topic evolution model is used to analyze the topic evolution trend from the perspectives of rising, falling, and stable. From the experiments we found that 8 topics showed an upward trend, 6 topics showed a downward trend, and 6 topics became stable or fluctuated. Compared with the baseline method, TDLDA can have the best effect when *K* is 40 or less. TDLDA is an effective topic model that can extract hot topics and evolution trends of blockchain patent texts, which helps researchers to more accurately grasp the research direction and improves the quality of project application and paper writing in the blockchain technology domain.

Keywords: blockchain; patents; TDLDA; topic intensity; evolutionary trend

1. Introduction

Patents, papers, S&T projects are important scientific and technological resources and occupy an important position in social progress. Patents can reflect the highlights of research hotspot earlier. Therefore, patent's topic mining can help researchers generate new ideas, which has important scientific significance for paper writing and project applications.

Patents are the most important achievements in the development of science and technology. The quantity and quality of patents are important in enterprise development and technological progress [1]. The development status and research hotspots in a particular field can be effectively found by analyzing the patent text. Patent analysis can promote the rapid development of research and academic fields [2]. In recent years, blockchain technology has developed rapidly and has received the attention of the business community and academia, and has achieved certain results in the publication of papers and patent applications. Blockchain technology integrates distributed computing [3], encryption and

decryption, consensus mechanism [4], hash mapping and other technologies. Blockchain is a very innovative technology that has emerged in recent years. It has become the main driving force of the next generation of the information technology revolution [5].

Based on a search of the China National Knowledge Infrastructure (CNKI)website, we obtained 8245 patents with "blockchain" in the title since the beginning of blockchain technology, of which 7151 patents (87%) have been approved by enterprises or individuals. The companies with the most approvals are Alibaba Group Holdings Co., Ltd., which has 447 approved blockchain patents, followed by Beijing Emory Technology Co., Ltd., and Tencent Technology (Shenzhen) Co., Ltd., with more than 200 items approved. Universities and research institutes have obtained a total of 1094 patents. The University of Electronic Science and Technology of China, Guangdong University of Technology, Xi'an University of Electronic Technology, Beijing University of Aeronautics and Astronautics, Nanjing University of Posts and Telecommunications, and others have shown outstanding performance. In general, large enterprises are the main force in blockchain patent applications, and have played a role in promoting the development of blockchain technology. Colleges and universities have conducted many explorations into the theory and application of blockchain, which is an effective supplement to blockchain patent applications.

Latent Dirichlet allocation (LDA) was proposed by Blei in 2003 [6]. LDA is a document topic generation model that is suitable for short text processing [7]. It can also be applied to fields such as signal processing, image processing, and natural language processing. LDA plays an important role in short text topic mining [8]. In this paper, we obtained patent data from CNKI. We took the blockchain patent title as the research object, then proposed a time-based dynamic latent Dirichlet allocation (TDLDA) modeling method to mine the patent text. The proposed TDLDA modeling method can discover hot topics. Finally, we used TDLDA and a topic intensity model to find the evolution patterns of blockchain patents. Thus we can discover the research situations and potential research trends of each sub-field contained in the patent text.

The main contributions of this paper are as follows:

(1) We propose a time-based dynamic latent Dirichlet allocation (TDLDA) modeling method based on a probabilistic graph model and knowledge representation learning for patent text mining.
(2) We present a computational model, topic intensity (TI), that expresses the topic strength and evolution.
(3) In order to evaluate the topic quality, we used the point-wise mutual information (PMI) [9] value which can measure the word association to test the effectiveness of our proposed TDLDA model.
(4) TDLDA is an effective model to extract hot topics and evolution trends of blockchain patent texts, which can help researchers more accurately grasp the research direction of blockchain technology.

The remainder of the paper is organized as follows. Section 2 briefly summarizes the related technology and theoretical research status of topic discovery and evolution for technology hotspot tracking. Section 3 describes the foundation theory and construction of the architecture for mining blockchain patent text based on the time-based dynamic LDA model. Section 4 presents the data source and experimental process in detail. Section 5 describes the extensive comparative experiments conducted to evaluate the effectiveness of our proposed TDLDA model on patent datasets. Section 6 presents the conclusions of this paper and discusses future research directions.

2. Related Works

Topic extraction and evolution analysis of patent documents can determine the status of research in this field and bring new ideas to scientific and technological (S&T) researchers. Many experts and scholars have conducted research on patent data and found valuable knowledge, which provides good technical support for knowledge dissemination and expert decision-making. Table 1 summarizes the latest research situation from three aspects.

Table 1. A brief summarization of patent mining and topic evolution.

Fields	Methods	Contexts
Patent text mining	[10] Convolutional neural networks	Patent information management and knowledge mining
	[11] Consumer-driven product function	Identify customer needs in real time
	[12] Preposition-based semantic analysis	Overcomes the limitations of keyword-based network
	[13] Feature vector space model(FVSM)	Patents related to Internet of Things (IoT) technology
	[14] Social network analysis	Identify underlying topics inhumanoid robot technology
	[15] Bayesian network modeling	Analyze patent documents related to artificial intelligence
Technical level	[16] WELQLC-QR	Improve the similar semanticsfor question retrieval
	[17] Valid and reliable LDA	More accessible to communication researchers
	[18] AC-LDA	Capture the relationships hidden in local sentences
	[19] NN-LDA	Create anintrusion detection systems
	[20] Intelligent LDA	Explored their associated trends over time
	[21] NMF-LDA	Provide users with suggestions and selections
	[22] SwLDA	Mimics human perception behavior
Topic evolution trend	[23] Manifold learning-based model	Explore topic-sentiment associate online news
	[24] dDeep learning language model	How keyword semantics can invoketopic evolution
	[25] Semantic-aware dynamic model	Monitors the evolution of author interest
	[26] Graph-based theory	Research front detection and topic evolution
	[27] Computational content analysis	CSR-related conversations in the Twitter-sphere
	[28] Functional count data model	The patent data of Applecompany

In terms of patent text mining, many researchers have done a lot of work to discover the knowledge contained in granted patents. Li proposed convolutional neural networks and word embedding to classify patent text, which is an essential task in patent information management and knowledge mining [10]. Trappey used social media and patent mining to deploy a consumer-driven product technology function that can identify customer needs in real time [11]. An proposed a preposition-based semantic analysis method to derive technological intelligence from patents that overcomes the limitations of the existing keyword-based network analysis and demonstrated its potential through an application [12]. Lei proposed a feature vector space model (FVSM) to analyze data that used patents related to Internet of Things (IoT) technology to demonstrate the performance and effectiveness of the model [13]. Kumari proposed a topic modelling and social network analysis method for humanoid robot technology publications and patents that used topic modelling based on latent Dirichlet allocation analysis to identify underlying topics in sub-areas in the field and social network analysis to detect important and influential sub-areas [14]. Sangsung proposed a Bayesian network modeling and factor analysis method to analyze patent documents related to disaster artificial intelligence technology [15].

At the technical level, LDA is an effective text mining tool that uses a probabilistic graph model to mine potential topics in the text. Many scholars have carried out research on LDA technology, which has been effectively applied to various fields, and achieved good results. Liu proposed an integrated retrieval framework for similar questions, named word-semantic embedded label clustering—LDA with question life cycle (WELQLC-QR), to improve the similar semantics and high popularity for question retrieval [16]. Maier proposed a valid and reliable methodology to make LDA topic modeling more accessible to communication researchers and to ensure compliance with disciplinary standards, and developed a brief hands-on user guide for applying LDA topic modeling [17]. Wan pro-

posed an association constrained LDA (AC-LDA) for effectively capturing co-occurrence relationships; the model can effectively capture the relationships hidden in local sentences and further increase the extraction rate of fine-grained aspects and opinion words [18]. Elkhadir proposed a new robust median NN-LDA based on the generalized mean to create an intrusion detection systems (IDS) that was superior to the approaches of many LDA variants [19]. Bastani proposed an intelligent approach based on latent Dirichlet allocation (LDA) to analyze CFPB consumer complaints and extract latent topics in the narratives, and explored their associated trends over time [20].Cao proposed a probabilistic matrix factorization recommendation approach fusing neighborhood selection based on latent Dirichlet allocation to provide users with suggestions and selections that was effective at improving recommendation performance and solving the data sparsity problem [21]. Jeon proposed a saliency-weighted LDA (swLDA) model that mimics human perception behavior, remarkably outperforming previous LDA models in terms of image categorization [22].

The topic evolution trend can effectively track the development status of research hotspots. Through the evolution trend, researchers can understand the current research status of the field from the time dimension. This can help researchers get inspiration and provide a basis for paper writing and project application. Scholars have conducted various studies on the technology and application of trend prediction and evolution, and achieved good results. For example, Xu proposed a manifold learning-based model to explore topic-sentiment associations and their evolution over time in the online news domain that can visualize the hidden sentiment dynamics of topics in a low-dimensional space [23]. Hu applied a deep learning language model, Google Word2Vec, to find whether and how keyword semantics can invoke or affect topic evolution [24]. Yang proposed an ordering-sensitive and semantic-aware dynamic author topic model that monitors the evolution of author interest in timestamped documents [25]. Xu proposed a new research front detection and topic evolution approach based on graph theory utilizing topological structure and the PageRank algorithm [26]. Chae used computational content analysis to understand topics from CSR-related conversations in the Twitter-sphere to find directions for future research [27]. Kim proposed a functional count data model to analyze patent data; using the patent data of Apple, the authors investigated the company's technological structure and evolution through high-dimensional visualization using harmonic components generated by functional data analysis [28].

From the related work, we can see that the research on topic discovery and evolution has made some progress, but there is less research on the use of topic models to analyze the development of patent technology. This paper uses topic model technology to analyze the evolution of the topic of blockchain, which has certain guiding significance for research on blockchain patents.

3. Construction of Blockchain Patent Text Based on Dynamic LDA Model

3.1. Foundation Theory for LDA

Latent Dirichlet allocation (LDA) is an unsupervised machine learning technique based on a probabilistic graph model [29]. It can discover potential topics in large-scale document sets, and it works well in text mining, especially short text processing [30]. Supposing the set of documents is D, the set of topics is T, and the vocabulary in the document is denoted by w, then the relationship between the topic, document, and vocabulary of LDA can be expressed by the following formula [31]:

$$p(w|\mathrm{d}) = p(w|\mathrm{t}) * p(t|\mathrm{d}) \tag{1}$$

where $p(w|\mathrm{d})$ represents the probability of a certain word in a document, $p(t|\mathrm{d})$ indicates the probability of a certain topic t in the document, and $p(w|\mathrm{t})$ indicates the probability of a certain word w in topic t. It can be found from the formula that topic t is the intermediate layer between the document and vocabulary layers. We can get the relationship between topic and vocabulary by calculating the probability of each word in the document. Document-topic and topic-topic words satisfy a multinomial distribution, which

can be represented by a Dirichlet distribution [32]. Its probability density function can be expressed as:

$$f(p|\alpha) = \begin{cases} \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, & p_k \in (0, 1) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where α indicates the parameters of the multinomial distribution and p_k represents the probability of a topic.

Suppose that there are *K* topics in m documents, each of which has its own topic distribution and is subject to a Dirichlet distribution with a parameter of α. Each topic has its own distribution of topic words and is subject to a Dirichlet distribution with a parameter of β. Each word in the document has a corresponding topic, and its probability graph model is shown in Figure 1 [33].
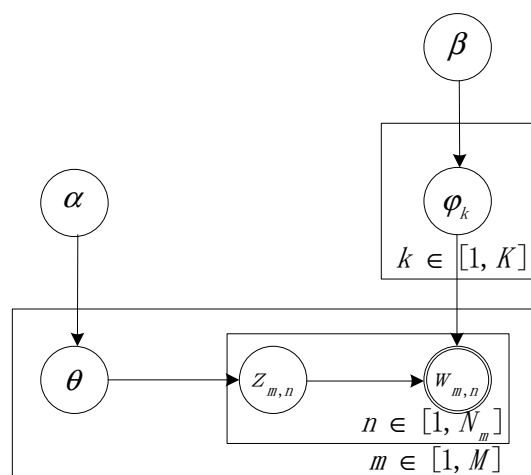

Figure 1. Probability graph model for LDA.

In Figure 1, θ indicates the topic, and the topic of the *i*-th document d_i can be expressed as $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik})$. φ_k indicates the distribution of the words corresponding to the *k*-th topic, where $k \in [1, K]$. For the *k*-th topic, the distribution of the topic words can be expressed as $\varphi_k = (\varphi_{k1}, \varphi_{k2}, \ldots, \varphi_{kj})$. $z_{m,n}$ represents the *n* topics of the *m*-th document. We finally get the required observation value $w_{m,n}$ by selecting φ_k from $z_{i,j}$.

3.2. Time-Based Dynamic LDA Blockchain Patent Text Topic Construction Process

This section introduces the construction process of TDLDA. Set D of the patent document can be represented as $D = \{d_1, d_2, \ldots d_n\}$, the vocabulary in the document can be represented as $w = \{w_1, w_2, \ldots w_n\}$, the dynamic time period can be represented as $Y = \{y_1, y_2, \ldots, y_m\}$, the generated *K* topics can be represented as $T = \{t_1, t_2, \ldots, t_k\}$, and the k topics obtained in time period y_i can be represented as $T_k^{y_i}$. The process of generating topics and corresponding topic words by the TDLDA topic model can be expressed by the following Algorithm 1:

Algorithm 1. Blockchain patent text subject mining process based on TDLDA model.

Input: Patent text data set $D = \{d_1, d_2, \ldots d_n\}$, where D is a set of documents.
Output: Topic set $T = \{t_1, t_2, \ldots, t_k\}$ within each time period and the first n topic words
according to probability.
1 initialize D, θ_i, φ_k;
2 initialize α, β;
3 $D^\xi \leftarrow jieba(d_1, d_2, \ldots, d_n)$;
4 repeat
5   for $n = 1$ to N
6 $\theta_{m,n}^y \leftarrow w_i^y$;
7   for $i = 1$ to K
8      $\theta_{m,n}^y = \theta_{m,n}^y \exp(\varphi_k)$;
9 $\varphi_k^y = \alpha \exp \varphi_{k-1}^y$;
10 until convergence

The above process can be summarized as:

(1) Traverse the vocabulary w_i^y in the patent document in each time period, and randomly assign atopic number to the vocabulary.

(2) Traverse the patent document dataset D again, perform Gibbs sampling on each document, find the subject corresponding to each w_i^y and update the number, and modify the number of words in the patent document set. Gibbs sampling can obtain a sample with certain probability. In this method, we can obtain a specific word through Gibbs sampling to further determine the probability that the word belongs to a certain topic.

(3) Repeat step 2 until Gibbs sampling converges.

(4) Finally, get K patent topics in each time period and the corresponding keywords of each topic.

4. Data Source and Experimental Process

4.1. Data Source

The dataset in this paper comes from the patent literature database of CNKI. We entered the keyword "blockchain" under the entry where the search condition is "patent name", and the search date was 27 February 2020. We obtained 8245 patents related to blockchain, including 67 items in 2016, 573 items in 2017, 2170 items in 2018, 5355 items in 2019, and 80 items in 2020. We selected all patents to export references, and got documents containing relevant fields such as inventor, applicant, patent name, address, publication date, publication number, abstract, etc. Finally, we obtained the titles of the patents in the data set as our research object.

4.2. TDLDA-Based Patent Text Topic Extraction Process

Based on the acquired data, we first preprocessed data document D, and further conducted experiments using the TDLDA algorithm. The specific process is shown in Figure 2.

First, we obtained patent documents on the subject of "blockchain" on CNKI, and constructed corresponding document sets in which the datasets were held before being processed. A patent contains many fields such as title, applicant, inventor, and granted time. The title can reflect the connotation and significance of the invention, so we used the patent titles as the research target. The patent title of each document set was regarded as a document.
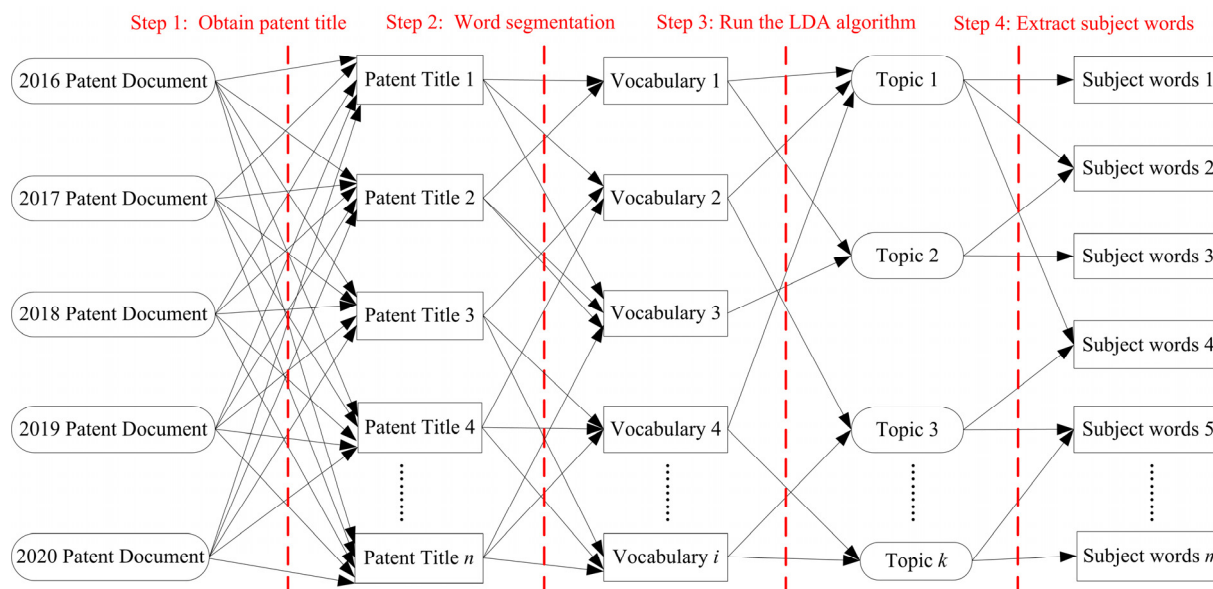
Figure 2. Topic mapping for blockchain patents in China, 2016 to 2020.

Second, we called the third-party library jieba in the python IDE development environment to segment the downloaded blockchain patent titles, remove irrelevant adjectives, adverbs, etc., extract professional vocabulary, and obtain a document composed of i words.

Finally, we used the TDLDA algorithm to perform multiple iterations on the vocabulary document to obtain *K* topics, select m topic words with similar semantics for each topic as representatives, and then summarize the topics names according to the corresponding vocabulary manually.
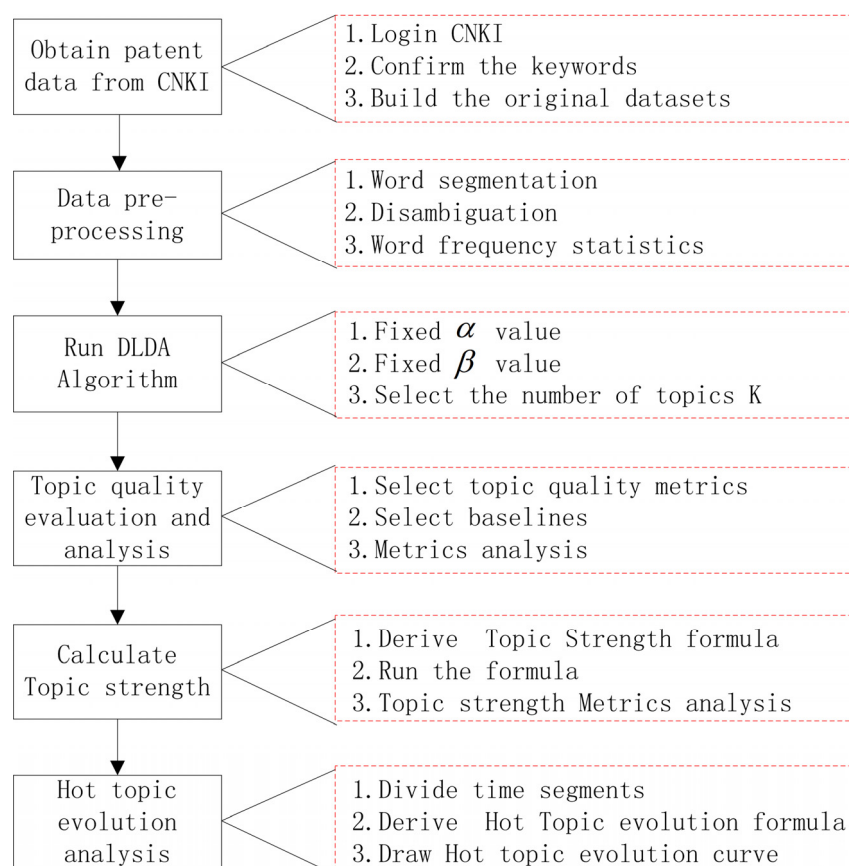
In the actual experiment, the amount of patent data in 2020 was relatively small. Therefore, although these data were obtained, the TDLDA topic model was not processed, thus This paper does not analyze patent data for 2020.

4.3. Topic Strength Calculation and Evolution Analysis Process

We first evaluated the quality of the topics based on TDLDA, and then selected 20 typical hot topics. In order to better know the popularity of topics, we calculated the strength of each topic and sorted the hot topics. Finally, evolution analysis was carried out according to yearly changes in topic intensity. Figure 3 shows the flowchart of the entire experiment.

The experimental process of this paper can be divided into six steps:

(1) Obtain patent data on blockchain in recent years.

(2) Pre-process blockchain data set to form a document set D on which to run the TDLDA algorithm.

(3) Choose the appropriate parameters to run TDLDA, and get the topic and topic words.

(4) Select principal component analysis (PCA) [34], singular value decomposition (SVD) [35], and other topic model algorithms as baseline to evaluate and analyze the topic quality of TDLDA; PCA and SVD are effective data dimensionality reduction methods, which are widely used in text mining and image noise reduction scenarios.

(5) Establish the topic strength model, calculate the intensity of each hot topic, and analyze hot topics.

(6) Build a hot topic evolution model and analyze the evolution trend of each hot topic.

```
┌─────────────────┐      ┌─────────────────────────────────────┐
│ Obtain patent   │─────▶│ 1. Login CNKI                       │
│ data from CNKI  │      │ 2. Confirm the keywords             │
└─────────────────┘      │ 3. Build the original datasets      │
         │               └─────────────────────────────────────┘
         ▼
┌─────────────────┐      ┌─────────────────────────────────────┐
│ Data pre-       │─────▶│ 1. Word segmentation                │
│ processing      │      │ 2. Disambiguation                   │
└─────────────────┘      │ 3. Word frequency statistics        │
         │               └─────────────────────────────────────┘
         ▼
┌─────────────────┐      ┌─────────────────────────────────────┐
│ Run DLDA        │─────▶│ 1. Fixed α value                    │
│ Algorithm       │      │ 2. Fixed β value                    │
└─────────────────┘      │ 3. Select the number of topics K    │
         │               └─────────────────────────────────────┘
         ▼
┌─────────────────┐      ┌─────────────────────────────────────┐
│ Topic quality   │─────▶│ 1. Select topic quality metrics     │
│ evaluation and  │      │ 2. Select baselines                 │
│ analysis        │      │ 3. Metrics analysis                 │
└─────────────────┘      └─────────────────────────────────────┘
         │
         ▼
┌─────────────────┐      ┌─────────────────────────────────────┐
│ Calculate       │─────▶│ 1. Derive Topic Strength formula    │
│ Topic strength  │      │ 2. Run the formula                  │
└─────────────────┘      │ 3. Topic strength Metrics analysis  │
         │               └─────────────────────────────────────┘
         ▼
┌─────────────────┐      ┌─────────────────────────────────────┐
│ Hot topic       │─────▶│ 1. Divide time segments             │
│ evolution       │      │ 2. Derive Hot Topic evolution formula│
│ analysis        │      │ 3. Draw Hot topic evolution curve   │
└─────────────────┘      └─────────────────────────────────────┘
```

Figure 3. Experimental process.

After the above process is completed, we will get *K* hot topics. At the same time, according to the topic intensity model, the intensity value of each topic in different years is calculated, and the evolution trend of a topic in recent years is obtained. According to the above steps, we conducted experiments on actual data sets. The specific experimental results can be seen in Section 5.

4.4. Experimental Environment and Related Parameter Settings

This experiment was conducted in the Windows 10 64-bit operating system, the development environment was JetBrains pycharm 2018, and python version 2.7.15 was used. The CPU hardware environment is a dual-core i7 6700 processor, the main frequency is 3.40GHz, and the RAM capacity is 8GB. Based on the obtained patent data, the patent data were divided into 5 datasets: BC_patents_2016, BC_patents_2017, BC_patents_2018, BC_patents_2019, and the total data set, BC_patents_16–19. The experimental parameter settings are shown in Table 2.

We set the Dirichlet distribution parameter α of document-topic to 0.2, the Dirichlet distribution parameter β of topic-vocabulary to 0.1, and the Gibbs sampling iteration number to 999.

Alpha and beta are important parameters of Dirichlet distribution. The larger for the value, the smoother for the distribution. Appropriate values for alpha and beta depend on the number of topics and the number of words in vocabulary. In this application, good results can be obtained by setting alpha = 0.2 and beta = 0.1.

Table 2. Experimental running data set and related parameter settings.

Dataset	Alpha	Beta	Number of Topics	Number of Documents	Dictionary Length	Number of Iterations
BC_patents_2016	0.2	0.1	10	67	11	999
BC_patents_2017	0.2	0.1	20	573	64	999
BC_patents_2018	0.2	0.1	20	2170	976	999
BC_patents_2019	0.2	0.1	20	5355	1570	999
BC_patents_16-19	0.2	0.1	20	8245	1985	999

Due to the small number of patent data sets in 2016, theBC_patents_2016 topic number was set to 10. The number of topics in the other datasets was set to 20. The number of documents in each dataset was 67, 573, 2170, 5355, and 8245, respectively. The dictionary length of the documents was 11, 64, 976, 1570, and 1985, respectively.

5. Experimental Results and Analysis

This section first analyzes hot topics in different time periods and puts forward the concept of hot spot strength. Then, in order to evaluate the effect of the TDLDA model in short text patent processing, PMI, a method for evaluating the quality of topics, is proposed. The indicators are compared and analyzed with PCA, SVD and other topic models. Finally, the theoretical basis of topic evolution in different time periods is proposed, and the evolution experiment results are analyzed.

5.1. Hot Topic Discovery and Analysis of Topic Quality

5.1.1. Analysis of Hot Topics and Keywords

In our experiments, we set the number of topics K to (20,40,60,80,100). Thus, we obtained K topics according to the array. Through the topic comparison, we found that when $K = 20$, the topic effect is optimal. We selected 20 topics for analysis, and summarized the corresponding topic names based on the content of the vocabulary. The specific results are shown in Table 3.

Table 3. Twenty topics and corresponding topic words related to blockchain patents, 2016 to 2019.

Serial Number	Topic	Topic Words
1	Public service	Scan code, logistics, mobile phone, computer room, rail traffic, public transport system, sell ticket, swiping card, access control system, face recognition
2	Data security	Encryption, transaction, account, secret key, data security, wallet, information, model, homomorphic, authority
3	Digital copyright	Terminal, identity, user, server, networking, information management, network systems, advertisement, trust, secret key
4	Network storage	Data processing, storage medium, server, relation, account number, network system, product, video
5	Medical insurance	Electronic, vote, Web page, case, authentication, evidence, identification photo, hospital, drug, medical treatment, prescription
6	Supply chain	Networking, framework, supply chain, finance, certificate, car, agricultural products, information security, cross-border, industry
7	Cross-chain technology	Cross-chain, business, domain name, partition, on chain, distributed, storage device, media, object, flow
8	Bill management	Electronic equipment, medium, data management, business, invoice, bill, verification, resource management, transaction, customer
9	Consensus algorithm	Consensus, mechanism, node, agreement, model, credibility, plan, Byzantium, vote, center, computing power

Table 3. *Cont.*

Serial Number	Topic	Topic Words
10	Intelligent contract	Intelligent, contract, resources, game, terminal, bracelet, communication, power grid, commodity trading, long range, insurance
11	Intelligent transportation	Recording, logistics, vehicle, coordination, drug, food, electric car, image, network, security, video, parking space
12	Books and materials	Data processing, electronic device, account book, authentication, digital certificate, resource allocation, book, work
13	Digital currency	Network, node, mining, fragmentation, account book, cluster, label, agreement, rule
14	Copyright protection	Assets, transaction, currency, copyright, content, platform, value, enterprise, bank, token, copyright protection
15	Intelligent trading	Keep accounts, agent, deploy, authority, account book, lottery, platform, product, client
16	Smart e-commerce	Evaluation, credit, private key, event, feature, e-commerce, community, user, signature
17	Energy internet	Distributed, trading System, energy, electricity, Internet, task, mode, dispatch, trading platform, power grid
18	Risk supervision	Electronic equipment, terminal equipment, transaction, public welfare, goods, digital certificate, client, insurance, risk, government affairs
19	Distributed storage	Storage system, file, distributed, business, structure, index, time, risk, client, certificate
20	Electronic file	Management system, query method, record, state, check, service platform, file, charging, positioning

After obtaining the topic terms through experiments, the focus becomes how to evaluate the strength of the subject and further analyze its evolution trend. In order to measure the proportion of one topic among all topics, we use its strength to test its topic intensity. Let $S\theta_k$ be the topic strength of the kth topic, which can be expressed as:

$$S\theta_k = \frac{\sum\limits_{j=1}^{N} Fre(\theta_k, w_j)}{\sum\limits_{i=1}^{v} Fre(w_i)} \qquad (3)$$

where V represents the length of the dictionary that is the total number of all words, $Fre(w_i)$ represents the frequency of all words, N represents the number of words in each topic, and $Fre(\theta_k, w_j)$ represents the frequency of words in the k-th topic.

According to the experimental results and Formula (3), we obtained the intensity of 20 topics, as shown in Figure 4.

We sorted the 20 topics through the topic strength model and analyzed the top 6: Topic4 network storage; Topic14, copyright protection; Topic2, data security; Topic18, risk supervision; Topic13, digital Currency; Topic9, consensus algorithm.
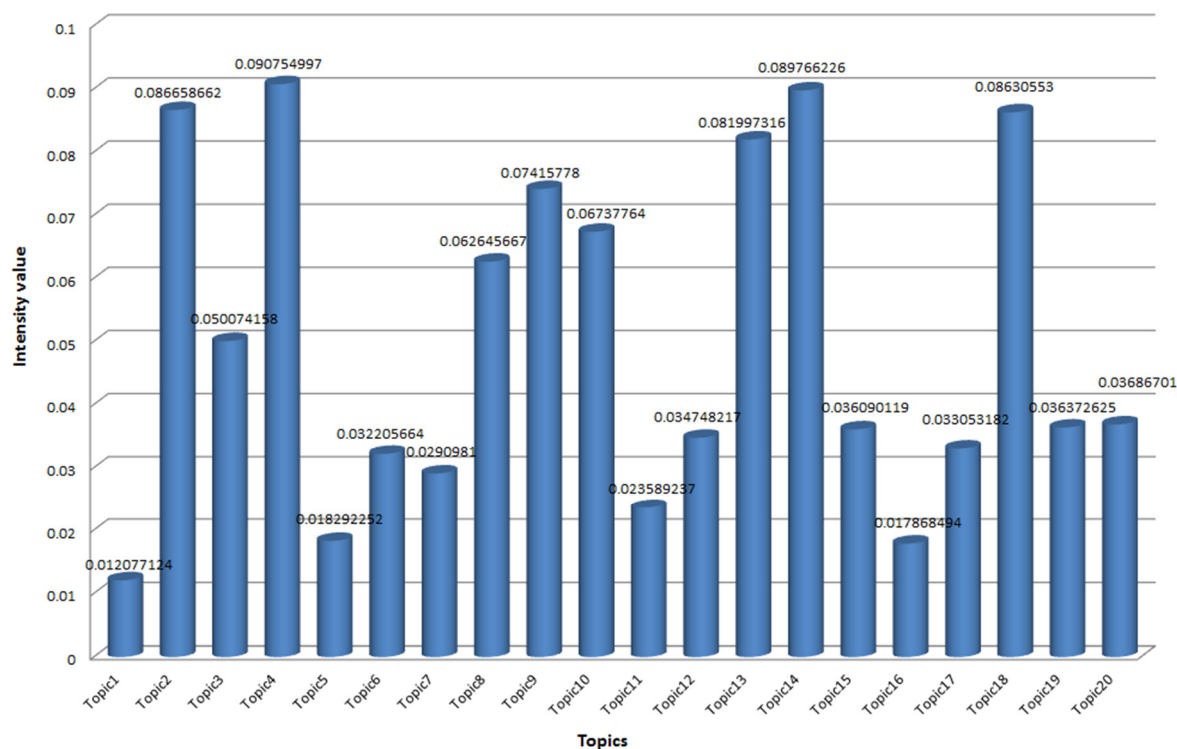
Figure 4. Intensity of 20 hot topics.

(1) Network storage. Network storage is one of the earliest areas of blockchain applications. As a distributed database, the blockchain itself provides functions such as security and reliability, multiple backups, and data traceability. Therefore, many scenarios have appeared in which network storage is a typical application. As society's demand for big data storage increases, network storage based on blockchain will occupy an important position. The main keywords of network storage include data processing, storage medium, server, network system, etc.

(2) Copyright protection. As the main application scenario in the early stage of the blockchain, copyright protection plays an important role in protecting intellectual property. The blockchain system can provide an important traceability function for various digital products, effectively protecting intellectual property and ensuring the product's sovereignty. The main keywords corresponding to this topic include assets, transactions, copyright, etc.

(3) Data security. Because of the distributed and multiple backup characteristics of the blockchain system, the blockchain can provide more secure technology than traditional databases. Blockchain technology provides important technical support for information security and sharing, and promotes the integration and development of emerging technologies such as big data, cloud computing, and artificial intelligence. The key words corresponding to data security mainly include encryption and decryption, secret key, wallet, model, homomorphism, etc.

(4) Risk supervision. The blockchain system has the characteristics of being tamper proof and having data traceability, making risk supervision an important application area, which can monitor the sources of transaction data. Risk monitoring can be used in many fields, such as securities and border trade. The keywords corresponding to the topic mainly include transactions, articles, digital certificates, risks, and certificates.

(5) Digital currency. Digital currency is the main application area of blockchain technology. Blockchain technology is derived from Bitcoin, and several digital currencies have been derived based on it, and it has become a digital currency technology strongly supported by the National Central Bank. The key words corresponding to this topic mainly include node, mining, sharding, ledger, etc.

(6) Consensus algorithm. The consensus algorithm is the core technology of the blockchain, and its execution efficiency is the main factor that affects the output, security, and expansion performance of the blockchain system. Terms such as consensus protocol, consensus model, consensus mechanism, and Byzantine are the main vocabulary of the consensus algorithm and occupy a very important position in authorized patents.

Judging from the hot patent topics, authorized patents are mainly concentrated in some application fields, and the patents for the core technologies of blockchain (such as consensus algorithms, smart contracts, cross-chain technologies, and algorithms) do not account for much. This shows that patent research and property rights protection for blockchain technology need to be strengthened. Blockchain technology patents lay a foundation for China's sustainable development in the fields of economy, finance, engineering technology, and social science.

5.1.2. Topic Performance Metrics Evaluation

After obtaining the topic and corresponding vocabulary through TDLDA, we need to evaluate the quality of the topic. In this paper, we use the PMI value to measure topic quality [36]. The PMI value was first proposed by Newman at the ADCS conference in 2009. Now it has become the main indicator for measuring the quality of a topic [37]. The PMI value can be expressed as:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{4}$$

where, w_i, w_j represent the first i and j words in a topic, $p(w_i, w_j)$ represents the probability that w_i and w_j will appear together in the same topic, and $p(w_i)$ and $p(w_j)$ represent the probability that the words w_i and w_j will appear in a certain topic. The larger the PMI value, the better the topic effect.

The baselines methods of this paper are SVD and PCA, two commonly used methods for topic modeling. PCA, SVD, and TDLDA experiments were carried out on the blockchain patent datasets BC_patents_2016, BC_patents_2017, BC_patents_2018, BC_patents_2019, and BC_patents_16-19. We can get the topic vocabulary with different K values using the three methods. We calculated the PMI value of the respective topic sets, and performed statistics according to the Top$N$ value under different K value topics. The experimental results are shown in Table 4.

Table 4. PMI values of TDLDA, SVD and PCA under different K and Top$N$ values.

Baseline	$K = 20$		$K = 40$		$K = 60$		$K = 80$	
	Top10	Top20	Top10	Top20	Top10	Top20	Top10	Top20
PCA	0.1328	0.51765	0.39395	0.78558	0.77763	1.09517	1.11799	1.41249
SVD	0.4252	0.68815	0.43065	0.74698	0.78218	1.08732	1.09908	1.355
TDLDA	0.4634	1.36685	0.58643	1.35635	0.61522	1.43057	0.6032	1.3434

In this experiment, the number of topics K was 20, 40, 60, and 80, and Top10 and Top20 refer to the first 10 and 20 words of the topic, respectively. Then their PMI values were calculated. It can be seen from the table that when the TDLDA model has $K = 20$, 40, and Top$N = 10$, 20, the PMI values are 0.4634, 1.36685, 0.58643, and 1.35635, respectively, which are significantly greater than the PMI values under the corresponding PCA and SVD models. This shows that when the number of topics K is less than 40, the TDLDA model is better than the PCA and SVD models. When $K = 60$ and Top$N = 10$, the PMI value of SVD is greater than that of the other two models. When $K = 80$, Top$N = 10$ and 20, the PMI value corresponding to the PCA model is larger than that of the other two, which indicates that the topic effect is best when the number of topics is 80. From the experimental results, it can be seen that when performing topic mining on the blockchain patent short text dataset, TDLDA can obtain very good results when the number of topics K is less than 40.

5.2. Blockchain Patent Hot Topic Evolution Analysis

5.2.1. Theoretical Basis of Topic Evolution

Topic evolution is the change patterns of topic words in a period series [38], which can be expressed by the topic intensity within a certain period. Scholars have studied topic intensity for detailed applications. For example, Cui [39] proposed the definition of topic intensity that He used, θ_k^d, to represent the ratio of the kth topic in document d, and then the kth topic in period t. The topic strength in time t is θ_k^t. It can be expressed as:

$$\theta_k^t = \frac{\sum\limits_{d=1}^{M} \theta_k^d}{M} \tag{5}$$

According to the needs of topic evolution for patent text mining, we propose a calculation model of topic strength, that is, the proportion of the frequency of topic words in a topic among all topic word frequencies. The modeling process can be expressed as follows. It will generate K topics after the model run, and there are J keywords in each topic. The number of occurrences of topics in the kth topic in period i can be expressed as:

$$T_k^i = \sum\limits_{j=1}^{J} count_k(w_j) \tag{6}$$

where w_j is the jth word, $count_k(w_j)$ is the number of the keywords in the kth topic, T_k^i means the number of occurrences of topics in the kth topic in period i.

According to the frequency of occurrence of subject words, the total frequency of occurrence of all topic words is recorded as M, which can be expressed as:

$$M^i = \sum\limits_{j=1}^{N} count(w_j) \tag{7}$$

where N means the number of all topics, M^i means the total frequency of occurrence of all topic words.

Then the topic intensity of the kth topic in a certain period can be expressed as:

$$\theta_k^i = \frac{T_k^i}{M^i} = \frac{\sum\limits_{j=1}^{J} count_k(w_j)}{\sum\limits_{j=1}^{N} count(w_j)} \tag{8}$$

where θ_k^i means the topic intensity of the kth topic in period i.

According to Formula (8), we can obtain the topic intensity of the blockchain topic in each year through experiments, and analyze the evolution trend of each hot topic according to the topic intensity, including rising, falling, and steady.

5.2.2. Topic Evolution Trend Analysis

This section analyzes and discusses the evolution trend of blockchain patent topics from the three perspectives of ascent, descent, and stability.

(1) Analysis of the rising trend of blockchain patent topics

According to the experimental results of Formula (8), we obtained the blockchain patent topic that showed an upward trend from 2016 to 2019, which is shown in Figure 5.
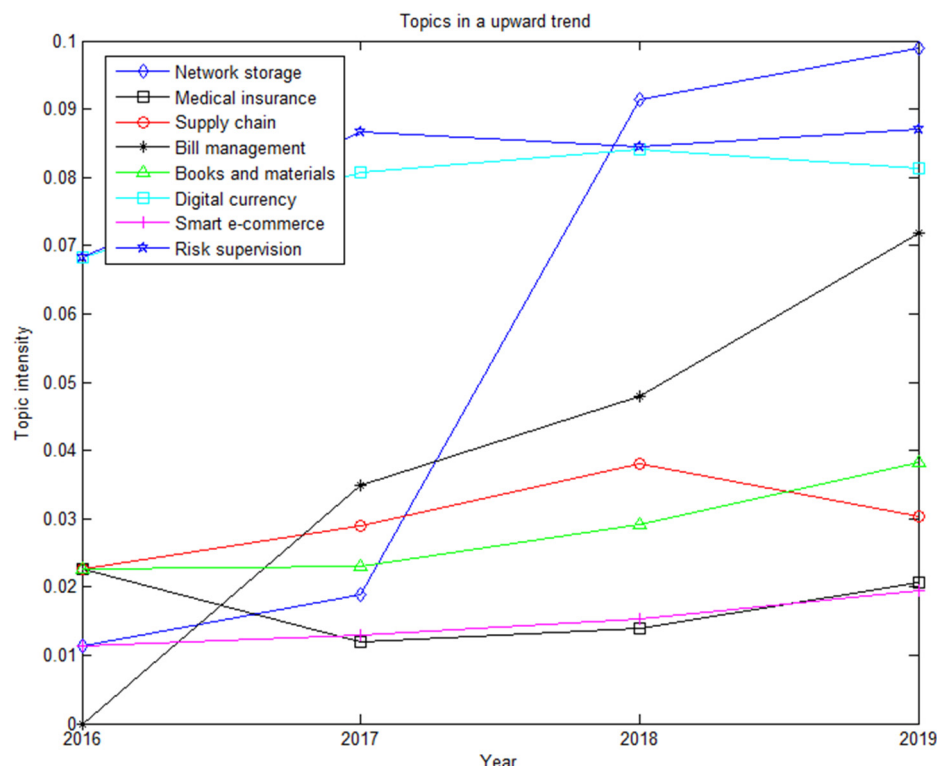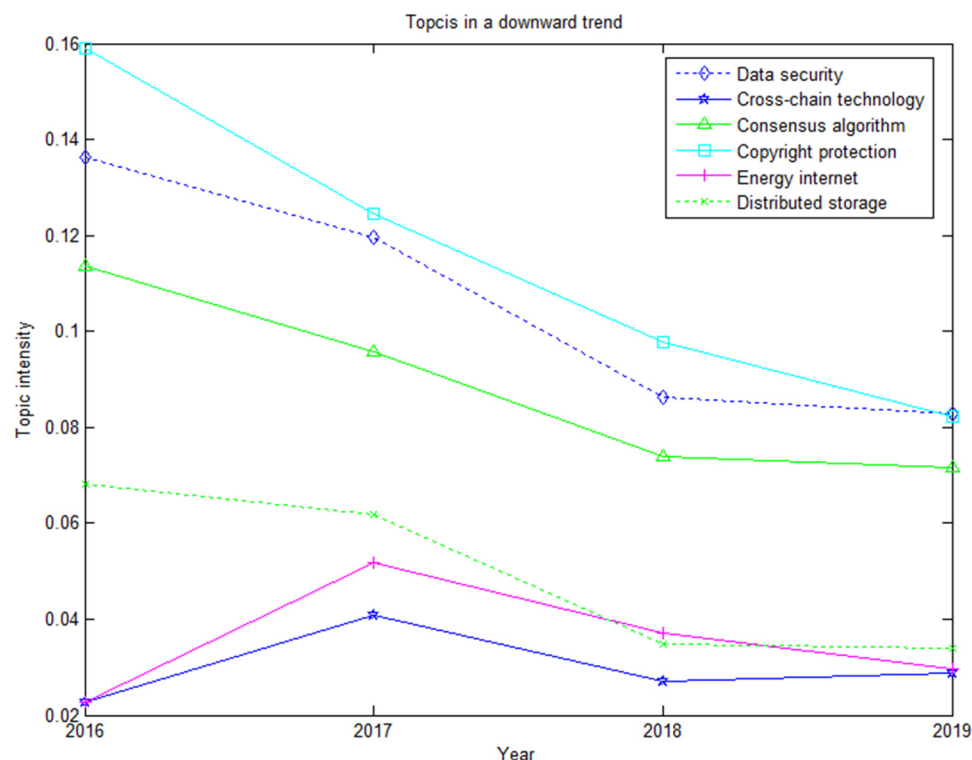
Figure 5. Topics showing an upward trend.

As can be seen from Figure 5, eight topics (Topic4, Topic5, Topic6, Topic8, Topic12, Topic13, Topic16, and Topic18) were on the rise, and they represent network storage, healthcare, supply chain, bill management, library materials, smart e-commerce, risk supervision, etc. Among them, Topic18 (risk supervision) declined in 2018 and rebounded immediately in 2019, generally showing an upward trend; and the proportion of all topics is relatively large which indicates that blockchain was used more in e-government. Topic4 (network storage) and Topic8 (bill management) were on an upward trend, and the rise was relatively large, accounting for more topics, which indicates that the blockchain was increasingly used in network storage and bill management. Topic6 (supply chain) and Topic13 (digital currency) were on an upward trend from 2016 to 2018 and slightly decreased in 2019, but the overall trend was upward, which indicates that blockchain was inseparable from the supply chain and digital currency. Blockchain is widely used in all aspects of the supply chain, which improves the efficiency of the supply chain operation and saves a lot of resources. The combination of blockchain and digital currency provides technical support for the financial sector. Topic12 (book materials) and Topic16 (smart business) were on an upward trend, but the proportion is relatively small, indicating that the blockchain has been applied to the field of book management and e-commerce, greatly improving management and operation efficiency. Topic5 (health care) declined in 2017 and then rose in 2018–2019, which indicates that blockchain was widely used in the field of healthcare, providing safe and reliable fast query and other functions for healthcare information.

(2) Analysis of the downward trend of blockchain patent topics

According to the experimental results of Formula (8), we obtained the blockchain patent topics that showed a downward trend from 2016 to 2019, shown in Figure 6.

Figure 6. Topics showing a downward trend.

As can be seen from Figure 6, six topics (Topic2, Topic7, Topic9, Topic14, Topic17, and Topic19) showed a downward trend, and they represent data security, cross-chain technology, consensus algorithm, copyright protection, energy Internet, and distributed storage. Among them, Topic14 (copyright protection), Topic2 (data security), and Topic9 (consensus algorithm) were on a downward trend, but they accounted for a large proportion of all topics, with a ratio of 0.07–0.16, which indicates that copyright protection, data security, and consensus algorithm have always been important core technology or application field in blockchain. The decline in the ratio is mainly because blockchain was applied to more fields. Looking at the frequency of the corresponding topic words every year, we can see that its absolute value always increased which indicates that the ratio of copyright protection, data security, and consensus algorithm declined, but continued to develop and was a research hotspot. Topic19 (distributed storage) was in a downward trend. Looking at the frequency of Topic19's corresponding topic words every year, the absolute value increased, but the growth rate was slower compared to other topics. Topic7 (cross-chain technology) and Topic17 (energy Internet) both rose first and then fell, and their share in 2017 reached a peak, which indicates that these two topics were research hotspots that year. Combined with reality, cross-chain technology and energy Internet were granted multiple patents that year and many papers were published.

(3) Blockchain patents tend to be stable or fluctuate

According to the experimental results of Equation (8), we obtained the blockchain patent topics that were stable or fluctuated from 2016 to 2019, shown in Figure 7.
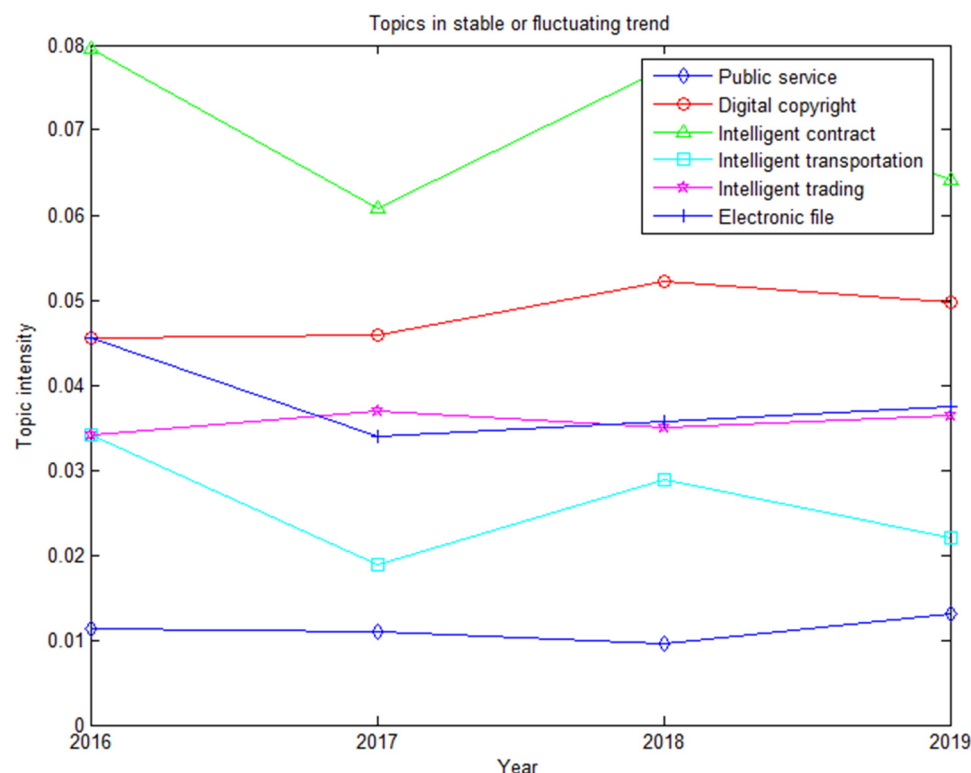
Figure 7. Topics showing stable or fluctuating trend.

As can be seen from Figure 7, six topics (Topic1, Topic3, Topic10, Topic11, Topic15, and Topic20) showed steady development or a fluctuation trend, and they represent public service, digital copyright, smart contract, smart transportation, digital currency, intelligent transactions, and electronic files. Among them, Topic10 (smart contract) and Topic11 (smart transportation) decreased in 2017, increased in 2018, and decreased in 2019. The fluctuations were relatively large. The ratio of smart contracts was 0.06–0.08, and the proportion was large. The ratio of intelligent transportation was 0.02–0.03, which indicates that smart contracts were a research hotspot in the early stage of blockchain development, and intelligent transportation was less popular during that stage. The fluctuation was due to the emergence of new hot topics in certain years. Topic1 (public service), Topic3 (digital copyright), and Topic15 (intelligent trading) rose and fell, but the rise and fall were not large. The overall trend was relatively stable. Topic20 (electronic file) first fell and then rose, reached the lowest in 2017, and began to rise in 2018, with a ratio of 0.037–0.045. The trend tended to be stable, indicating that blockchain has always been a concern of scholars in these fields and is in a stable development state, showing characteristics of refinement, extensiveness, and intelligence.

6. Conclusions

Tracking scientific and technological (S&T) research hotspots can help scholars to grasp the current research status and regular development pattern of the field over time. This paper proposes a short text topic acquisition method, the time-based dynamic LDA (TDLDA) model. We took Chinese blockchain patent datasets from 2016 to 2019 from CNKI as the research subject. First, we called the third-party library jieba to obtain the professional terminology of the blockchain patent literature. Second, we used the TDLDA, an unsupervised representation method, to capture potential topics in blockchain patents, in order to obtain hot topics and corresponding topic words. We proposed a topic intensity model to calculate the intensity of hot topics, then sorted the *K* hot topics and calculated their intensity in different time periods, and discussed the feature distribution and development trend of each hot topic. Finally, we used the PMI evaluation index to compare the

three modeling methods of PCA, SVD, and TDLDA, and analyzed the topic quality in these models. of the three models. The experimental results show the following: (1) Blockchain patent topics were processed by TDLDA to get *K* topics from 2016 to 2019. When *K* = 20, the topic effect was the best; eight topics showed an upward trend, six topics showed a downward trend, and six topics tended to be stable or fluctuate slightly. (2) We sorted the 20 topics according to intensity, and analyzed the first six topics: network storage, copyright protection, data security, risk supervision, digital currency, and consensus algorithm. We found that these topics were the main areas for blockchain patent applications. (3) When the number of topics *K* was less than or equal to 40, the PMI value of TDLDA is the best among the three topic models, which indicates that TDLDA was a suitable method for short-text topic data mining of patents. The research in this paper can help researchers more accurately grasp the research direction and improve the quality of project applications and paper writing.

Author Contributions: The paper is the result if the intellectual collaboration of all listed authors. J.W. designed the setup, performed the experiments and wrote the manuscript. Y.F. analyzed the data and revised the manuscript. H.Z. analyzed the simulation result and revised the paper. L.F. proposed many useful suggestions in the process of theoretical analysis. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lee, H.J.; Oh, H. A Study on the Deduction and Diffusion of Promising Artificial Intelligence Technology for Sustainable Industrial Development. *Sustainability* **2020**, *12*, 5609. [CrossRef]
2. Ampornphan, P.; Tongngam, S. Exploring Technology Influencers from Patent Data Using Association Rule Mining and Social Network Analysis. *Information* **2020**, *11*, 333. [CrossRef]
3. Feng, L.; Zhang, H.; Chen, Y.; Lou, L. Scalable Dynamic Multi-Agent Practical Byzantine Fault-Tolerant Consensus in Permissioned Blockchain. *Appl. Sci.* **2018**, *8*, 1919. [CrossRef]
4. Yang, H.; Liang, Y.; Yuan, J. Distributed Blockchain-Based Trusted Multidomain Collaboration for Mobile Edge Computing in 5G and Beyond. *IEEE Trans. Ind. Inform.* **2020**, *16*, 7094–7104. [CrossRef]
5. Feng, L.; Zhang, H.; Tsai, W.; Sun, S. System architecturefor high-performance permissioned blockchains. *Front. Comput. Sci.* **2019**, *13*, 1151–1165. [CrossRef]
6. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
7. Zuo, Y.; Zhao, J.; Xu, K. Word network topic model: A simple but general solution for short and imbalanced texts. *Knowl. Inf. Syst.* **2016**, *48*, 379–398. [CrossRef]
8. Zuo, Y.; Wu, J.; Zhang, H. Topic Modeling of Short Texts: A Pseudo-Document View. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 2105–2114.
9. Church, K.W.; Hanks, P. Word association norms, mutual information, and lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
10. Li, S.; Hu, J.; Cui, Y.; Hu, J. DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics* **2018**, *117*, 721–744. [CrossRef]
11. Trappey, A.J.; Trappey, C.V.; Fan, C.Y.; Lee, I.J. Consumer driven product technology function deployment using social media and patent mining. *Adv. Eng. Inform.* **2018**, *36*, 120–129. [CrossRef]
12. An, J.; Kim, K.; Mortara, L.; Lee, S. Deriving technology intelligence from patents: Preposition-based semantic analysis. *J. Informetr.* **2018**, *12*, 217–236. [CrossRef]
13. Lei, L.; Qi, J.; Zheng, K. Patent Analytics Based on Feature Vector Space Model: A Case of IoT. *IEEE Access* **2019**, *7*, 45705–45715. [CrossRef]

14. Kumari, R.; Jeong, J.Y.; Lee, B.H.; Choi, K.N.; Choi, K. Topic modelling and social network analysis of publications and patents in humanoid robot technology. *J. Inf. Sci.* **2019**. [CrossRef]

15. Park, S.; Jun, S. Patent Keyword Analysis of Disaster Artificial Intelligence Using Bayesian Network Modeling and Factor Analysis. *Sustainability* **2020**, *12*, 505. [CrossRef]

16. Liu, Y.; Tang, A.; Sun, Z.; Tang, W. An integrated retrieval framework for similar questions: Word-semantic embedded label clustering—LDA with question life cycle. *Inf. Sci.* **2020**, *537*, 227–245. [CrossRef]

17. Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; et al. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Commun. Methods Meas.* **2018**, *12*, 93–118. [CrossRef]

18. Wan, C.; Peng, Y.; Xiao, K.; Liu, X.; Jiang, T.; Liu, D. An association-constrained LDA model for joint extraction of product aspects and opinions. *Inf. Sci.* **2020**, *519*, 243–259. [CrossRef]

19. Elkhadir, Z.; Mohammed, B. A cyber network attack detection based on GM Median Nearest Neighbors LDA. *Comput. Secur.* **2019**, *86*, 63–74. [CrossRef]

20. Bastani, K.; Namavari, H.; Shaffer, J. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Syst. Appl.* **2019**, *127*, 256–271. [CrossRef]

21. Cao, Y.; Li, W.; Zheng, D. A hybrid recommendation approach using LDA and probabilistic matrix factorization. *Clust. Comput. J. Netw. Softw. Tools Appl.* **2019**, *22*, S8811–S8822. [CrossRef]

22. Jeon, J.; Kim, M. Discovering Latent Topics with Saliency-Weighted LDA for Image Scene Understanding. *IEEE Multimed.* **2019**, *26*, 56–68. [CrossRef]

23. Xu, Y.; Li, Y.; Liang, Y.; Cai, L. Topic-sentiment evolution over time: A manifold learning-based model for online news. *J. Intell. Inf. Syst.* **2019**, *55*, 27–49. [CrossRef]

24. Hu, K.; Luo, Q.; Qi, K.; Yang, S.; Mao, J.; Fu, X.; Zheng, J.; Wu, H.; Guo, Y.; Zhu, Q. Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Inf. Process. Manag.* **2019**, *56*, 1185–1203. [CrossRef]

25. Yang, M.; Qu, Q.; Chen, X.; Tu, W.; Shen, Y.; Zhu, J. Discovering author interest evolution in order-sensitive and Semantic-aware topic modeling. *Inf. Sci.* **2019**, *486*, 271–286. [CrossRef]

26. Xu, Y.; Zhang, S.; Zhang, W.; Yang, S.; Shen, Y. Research Front Detection and Topic Evolution Based on Topological Structure and the PageRank Algorithm. *Symmetry* **2019**, *11*, 310. [CrossRef]

27. Chae, B.K.; Park, E.O. Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. *Sustainability* **2018**, *10*, 2231. [CrossRef]

28. Kim, J.M.; Kim, N.K.; Jung, Y.; Jun, S. Patent data analysis using functional count data model. *Soft Comput.* **2019**, *23*, 8815–8826. [CrossRef]

29. Ren, Y.; Wang, Y.; Zhu, J. Spectral Learning for Supervised Topic Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 726–739. [CrossRef]

30. Zhang, C.; Lu, S.; Zhang, C. A Novel Hot Topic Detection Framework with Integration of Image and Short Text Information from Twitter. *IEEE Access* **2019**, *7*, 9225–9231. [CrossRef]

31. Chen, Y.; Zhang, H.; Liu, R. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowl. Based Syst.* **2019**, *163*, 1–13. [CrossRef]

32. Martinez, A.M.; Kak, A.C. PCA versus LDA. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 228–233. [CrossRef]

33. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]

34. Shang, F.; Cheng, J.; Liu, Y. Bilinear Factor Matrix Norm Minimization for Robust PCA: Algorithms and Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2066–2080. [CrossRef]

35. Jiang, Z.; Lin, Z.; Davis, L. Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2651–2664. [CrossRef]

36. Wang, J.; Fan, Y.; Feng, L. Research Hotspot Prediction and Regular Evolutionary Pattern Identification Based on NSFC Grants Using NMF and Semantic Retrieval. *IEEE Access* **2019**, *7*, 1–12. [CrossRef]

37. Newman, D.; Karimi, S.; Cavedon, L. External evaluation of topic models. In Proceedings of the ADCS, Sydney, Australia, 4 December 2009; pp. 11–18.

38. Rosario, J.; Mohamed, A.; Saverio, N. A Vocabulary for Growth: Topic Modeling of Content Popularity Evolution. *IEEE Trans. Multimed.* **2018**, *20*, 2683–2692.

39. Cui, K. The Research and Implementation of Topic Evolution Based on LDA. Ph.D. Thesis, National University of Defense Technology, Changsha, China, 2010; pp. 18–26.