symmetry

MDPI

Article

Research on Topic Evolution Path Recognition Based on LDA2vec Symmetry Model

Tao Zhang [1], Wenbo Cui [1], Xiaoli Liu [2,*], Lei Jiang [3] and Jinling Li [1]

[1] School of Information Management, Heilongjiang University, Harbin 150080, China
[2] School of Economics and Management, Northeast Agricultural University, Harbin 150030, China
[3] Information and Network Center, Heilongjiang University, Harbin 150080, China
* Correspondence: liuxiaoli@neau.edu.cn

Abstract: Topic extraction and evolution analysis became a research hotspot in the academic community due to its ability to reveal the development trend of a certain field and discover the evolution law of topic content in different development stages of the field. However, current research methods still face challenges, such as inaccurate topic recognition and unclear evolution paths, which can seriously compromise the comprehensiveness and accuracy of the analysis. To address the problem, the paper proposes a topic evolution path recognition method based on the LDA2vec symmetry model. Under given conditions, both the LDA and Word2vec used in the model conform to the structural symmetry of their datasets in high-dimensional space, and the fused LDA2vec method improves the accuracy of the analysis results. Firstly, we recognize the topics based on the LDA model, which uses Gibbs symmetric sampling and obeys the symmetric Dirichlet distribution to ensure data convergence. Secondly, Word2vec is used to learn the contextual information of the topic words in the document collection, and the words in the corpus are projected as vectors in the high-dimensional space so that the computed pairs of words with similar semantics have symmetry in the hyperplane of the high-dimensional space. Subsequently, the word vector is used as a weight, and the LDA topic word probability value is weighted to generate a new topic vector. Thirdly, the vector similarity index is employed to calculate the semantic similarity among topics at adjacent stages, and evolution paths that directly reflect the topic relationships are constructed. Finally, an empirical study is conducted in the field of data security to demonstrate the effectiveness of the proposed approach for topic evolution analysis. The results show that the proposed approach can accurately recognize the topic content and construct clear evolution paths, which contribute to the comprehensive and accurate analysis of topic evolution in a specific research field.

Keywords: high-dimensional symmetry; topic evolution; LDA; LDA2vec

check for updates

Citation: Zhang, T.; Cui, W.; Liu, X.; Jiang, L.; Li, J. Research on Topic Evolution Path Recognition Based on LDA2vec Symmetry Model. *Symmetry* **2023**, *15*, 820. https://doi.org/10.3390/sym15040820

Academic Editor: José Carlos R. Alcantud

Received: 24 February 2023
Revised: 23 March 2023
Accepted: 26 March 2023
Published: 29 March 2023

1. Introduction

With the integrated development of technology and the universality of interdisciplinary research, every discipline field has a process of dynamic change over time. The research of each discipline field must grasp the frontier hotspots in time and find the breakthrough point of innovation to promote the development of the discipline. Topic evolution analysis is the key to identifying current and potential hotspots. It reveals the knowledge venation of each discipline and predicts the future development trend of each discipline, providing support for the direction choice of strategic decision-makers and scientific researchers, and further promoting the three-dimensional and in-depth development of innovation in each field [1]. From the perspective of structure, topic evolution analysis belongs to microstructure and usually represents the research direction of a field. From the content level, topics are collections of domain knowledge units and are usually implicit semantic structures. However, topic evolution in different fields contains different era background factors, such as the science and technology development degree, data volume growth mode, topic diversification, and policy support. It not only needs to identify topics from massive information, but also needs to determine clues and knowledge between topics in different fields and periods. Therefore, the evolution analysis of topics became a hot topic in the academic field, and the innovation of its research methods aroused widespread attention among scholars. To solve the problem of the traditional latent Dirichlet allocation (LDA) model, which ignores the semantic association between text contexts [2], this paper proposes a recognition method of topic evolution paths based on the LDA2vec symmetry model. It is noteworthy that the use of Bayesian symmetric sampling and Dirichlet symmetric distribution in LDA, as well as the symmetric mapping of data in high-dimensional space in Word2vec, all represent manifestations of symmetry. Therefore, the merged LDA2vec model exhibits favorable symmetry properties. This method makes it better able to reflect the correlation relationships and evolution patterns between topics at adjacent stages, and then can accurately identify topic features and construct topic evolution paths. The following are the main contributions of our work:

We propose a new unsupervised learning method, which is a topic evolution path recognition method based on the LDA2vec symmetry model, to solve the problem of accurately calculating vector similarity indicators to measure semantic relations among topics. The experimental results show that the method can correctly display the development process and evolution types of academic disciplines.

We further analyzed and classified the clustering results using analytical softwares, such as Ucinet and VOSviewer to extract important evolutionary paths of academic discipline topics. The experimental results show that this novel approach overcomes the difficulty of selecting important evolution paths among numerous possibilities compared to previous studies that randomly extracted topic evolution paths, thereby improving the reliability of the experimental results and reflecting more realistic feature of disciplinary topic evolution.

The remaining parts of the study are structured as follows: Section 2 discusses the current state and limitations of topic evolution analysis; Section 3 describes the system design process and technical aspects of the topic evolution path recognition method based on the LDA2vec symmetry model; Section 4 presents experimental results and content analysis; and Section 5 outlines the research content, limitations, as well as the future work of the study.

2. Related Work

2.1. Topic Recognition

Extracting topics is the premise of constructing topic evolution paths. The methods for topic extraction include those that are specific to particular disciplines, as well as those that apply to general disciplines. Prasanna et al. [3] proposed a method called "doubleton pattern mining" for discovering colossal patterns from biological datasets. Kottapalle et al. [4] discussed a new method, D-Mine, for discovering super-large gene pattern sequences from biological data sets. Among them, the common topic extraction method can be divided into two categories: The first one is traditional topic extraction methods, including co-citation analysis, co-word analysis, and social network analysis. Regarding co-citation analysis, Zhu QS et al. [5] proposed this method to analyze topic evolution and took the field of carbon nanotube fiber as an example to analyze the co-citation of key papers in main citation paths, revealing the topic evolution of the discipline field. About co-word analysis, Cobo et al. [6] took two of the most important journals in the field of fuzzy set theory (FST) as examples to analyze the evolution of the number of keywords and shared keywords in different sub-periods. Regarding social network analysis, Santos et al. [7] proposed a method for extracting patterns and knowledge from scientific publications of research groups, which combines complex network analysis and other techniques. Nan Yan et al. [8] used the method of social network analysis to discuss the evolution of China's aging policy from 1978 to 2019 and put forward constructive suggestions for strengthening cooperation between government agencies. These traditional methods can quickly identify topics with a high degree of credibility. However, they only use the topic frequency as an index for determining the topic evolution path, which not only fails to accurately extract topics according to the relation of synonyms, invalid words, and potential semantics, but also ignores the potential hot spots. The other is the method of topic extraction based on machine learning, in which the LDA model is more mature. In 2003, Blei et al. [9,10] proposed LDA, an unsupervised machine learning method that can dig out potential topics in initial documents. Due to its symmetric property, LDA has the advantage of improving the generalization performance of its model [11]. Additionally, its high versatility and ease of implementation made it widely applicable across various fields [12]. Choi et al. [13] used LDA to identify the topics from the abstracts of 2356 literature about "privacy" and "personal information" in the Scopus database, revealing the research trend of personal information privacy. Zhang et al. [14] proposed a policy text clustering method based on LDA to improve the accuracy of policy text clustering. Xue et al. [15] used LDA to identify prominent topics in the collected tweets from Twitter users to study the evolving public discussion and sentiment during COVID-19. Therefore, this paper chose LDA as the method of topic extraction.

2.2. Topic Evolution Analysis

As the LDA model matured in the application, plenty of scholars gradually studied the evolution trend of topics based on LDA. Some scholars introduced the time dimension into the LDA model, making it possible to reflect the dynamic change process of the topics over time [16–18]. Xu S et al. [19] proposed a dynamic model that treats the temporal dimension as an endogenous variable of the LDA model and combines author attributes to investigate the patterns of potential topics and users' interests evolving. Zhu H et al. [20] defined the topic transition probability based on LDA to analyze the evolution path of topics. Based on introducing the temporal dimension, some scholars explored the relationships among topics through similarity, including division, merging, inheritance, generation, and disappearance. Therefore, accurately calculating the similarity among topics is a key step to determining the evolution type of topics and constructing topic evolution paths. With the development and application of deep learning, some advanced models, represented by Word2vec [21] and BERT [22], can transform textual information from the unstructured form to vectorized form, making the generated word vectors related to semantics. These models solve the problem of the LDA model, which ignores the potential semantic association between topic words due to the lack of semantic embeddings [23]. Xie Q et al. [24] used BERT and LDA to analyze the similarity of monolingual and multilingual research topic evolution. Xi XW et al. [25] researched the technical similarity visualization based on Word2vec and LDA and conducted an empirical study in the field of NEDD, proving that the model had good results in the analysis of technical similarity measurement.

Although BERT contains more textual semantic information, the vector in which BERT encodes sentences is anisotropic. The vector values are influenced by the frequency of the words in all the training corpora, resulting in the vectors encoded by high-frequency words being closer together and more concentrated near the origin. This will cause that even if the semantics of a high-frequency word and a low-frequency word are equivalent; the difference in word frequency will also bring a large distance bias [26]. As a result, the distance of the word vector does not represent the semantic correlation well [27]. In contrast, Word2vec, as a shallow neural network model based on probability, concentrates on contextual logic and therefore has a natural advantage in topic feature extraction.

Based on the above-related work, combining the advantages of both LDA and Word2vec, and considering the importance of the topic words, this paper proposed a recognition method of a literature topic evolution path based on LDA2vec symmetry model, aiming to track the evolution process of topics in the temporal dimension by improving the accuracy of calculating the similarity among topics at adjacent stages. Finally, the topic evolution paths were visualized and analyzed to reveal the topics and evolution process of the field at different development stages, which helps scholars understand the overall development, evolutionary venation, and research priority of the field.

3. Methodology

To explore the evolution patterns of the topic content and judge the future development trend of the field, we propose a recognition method of the literature topic evolution path based on the LDA2vec symmetry model to get the overall evolution path and analyze the correlation relationship between the topics at adjacent stages from the perspective of topic evolution. The specific steps are as follows:

Firstly, we need to obtain the papers from the corpus and preprocess the papers to construct the feature word list and the stop word list. Secondly, we need to identify valid global topics and staged topics with corresponding topic words. The perplexity is used to determine the optimal number of global topics and staged topics. Then, the similarity between the staged topics and the global topics is calculated and compared for topic consistency verification to verify the number of valid stage topics. Moreover, LDA is next used to extract the topic words corresponding to valid global and staged topics. Thirdly, we obtain the weighted summation value by multiplying the LDA probability values corresponding to the topic words by the word vectors obtained from Word2vec training to obtain a vectorized representation of the topics and then calculate the similarity between topics at adjacent stages. Finally, we set a similarity threshold to determine the type of topic evolution paths and construct visual evolution paths for analysis. The overall analysis framework is shown in Figure 1.

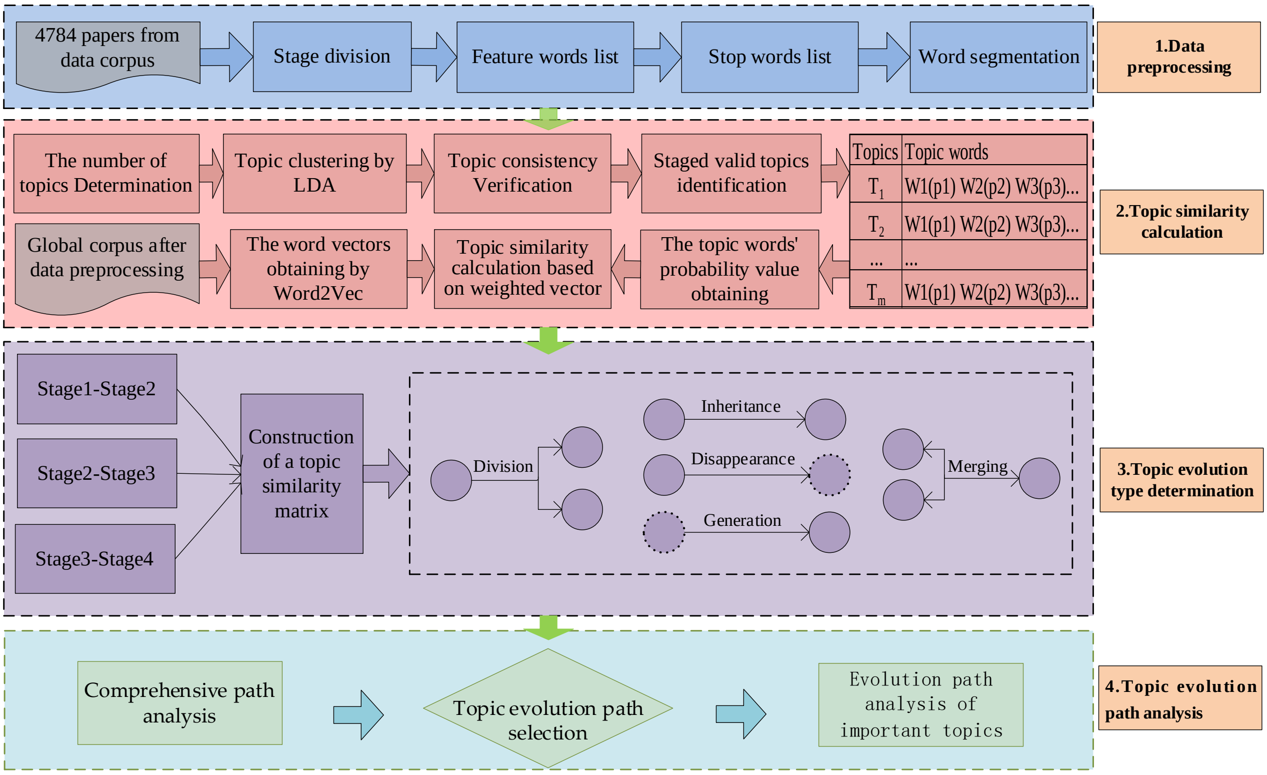

4784 papers from data corpus
Stage division
Feature words list
Stop words list
Word segmentation
1.Data preprocessing
The number of topics Determination
Topic clustering by LDA
Topic consistency Verification
Staged valid topics identification
Topics
Topic words
T_1
W1(p1) W2(p2) W3(p3)...
T_2
W1(p1) W2(p2) W3(p3)...
...
...
T_m
W1(p1) W2(p2) W3(p3)...
Global corpus after data preprocessing
The word vectors obtaining by Word2Vec
Topic similarity calculation based on weighted vector
The topic words' probability value obtaining
2.Topic similarity calculation
Stage1-Stage2
Stage2-Stage3
Stage3-Stage4
Construction of a topic similarity matrix
Division
Inheritance
Disappearance
Generation
Merging
3.Topic evolution type determination
Comprehensive path analysis
Topic evolution path selection
Evolution path analysis of important topics
4.Topic evolution path analysis

Figure 1. The framework of topic evolution analysis based on the LDA2vec symmetry model.

3.1. Data Preprocessing

In this paper, we obtain the papers with titles, keywords, abstracts, and publication dates from the Web of Science to form a document set. Then we pre-process the text, including constructing word lists, segmenting words, removing stop words, dividing stages, etc. Among these steps, word list construction is the key to ensuring the quality of the final dataset, including the following two steps:

(1) Domain feature word list construction: Given the real dilemma that the target domain is highly knowledgeable, specialized in vocabulary, and unsupervised, we extract the keywords of literature as the initial feature word list. Then, we use the TF-IDF algorithm to extract keywords in the target domain from the literature and add them to the initial feature word list. Finally, we perform de-duplication and filtering work to obtain the final feature word list.

(2) Domain stop word list construction: To improve the accuracy of domain-oriented topic recognition and prevent the interference of high-frequency invalid words (such as "analysis") appearing in each domain, this paper proposes to construct a target domain-oriented stop word list based on the universal stop word list. In the specific case application, we check the clustering results after the first LDA topic clustering, extract the meaningless words of the domain into the universal stop word list, and iterate the above process no more than five times to form the final stop word list.

3.2. Calculation of Similarity Based on LDA2vec Symmetry Model

In this paper, we use the LDA topic model to extract valid topics from the pre-processed dataset to generate topic word distributions and then select the topic words with higher probability values for each stage topic. Finally, we determine the relationships between topics by calculating the similarity between the topics at adjacent stages. Specific steps are as follows:

(1) K value determination: The perplexity, which indicates the uncertainty about the topic to which the literature belongs, is used to determine the number of global topics and staged topics. The lower the perplexity, the better the clustering effect is, and the number of topics is optimal [28]. The calculation of perplexity is shown in Equation (1).

$$\mathrm{Perplexity(D)} = \exp\frac{\sum_{\mathrm{d}=1}^{\mathrm{M}} \mathrm{logP(W_d)}}{\sum_{\mathrm{d}=1}^{\mathrm{M}} \mathrm{N_d}} \tag{1}$$

In Equation (1), D represents the set of all words in the document, m represents the number of documents, $\mathrm{W_d}$ represents the words in document z, $\mathrm{N_d}$ represents the number of words in each document, and $\mathrm{p(W_{\ d})}$ represents the probability of words in the document.

(2) Topic consistency verification: We respectively calculate the cosine similarity between the four-stage topics and the global topics, and compare the average of the similarity values named C with the threshold value of 0.5 to derive the result. If $\mathrm{C} \geq 0.5$, it proves that there is consistency among topics. If C < 0.5, it proves that there is no consistency among topics, which means these topics, named invalid topics, are not highly related to global topics and should be filtered out.

(3) Similarity calculation based on the weighted vector: Word2vec provides two classical language models, which are the skip-gram model and the CBOW model, for training. Skip-gram predicts the nearby words by the central word, while CBOW predicts the current value by the context. The purpose of this paper is to perform the topic evolution path analysis based on similarity calculation for the topics (central words), similar to the principle of the skip-gram model. As a result, the skip-gram model was chosen. First of all, we use the preprocessed global dataset as word vectors to train data and use the functions in the gensim package of the Python language to train word vectors, constructing word vector models. Then, we use the trained Word2vec model to vectorize the topic words obtained by LDA model clustering and generate the word vectors corresponding to each of the topic words. Since the probability of a topic word represents the importance of a topic, to better measure the importance of topic–topic words, we multiply the probability of a topic word extracted by LDA as a weight by the corresponding word vector of Word2vec and perform a weighted sum to obtain a vectorized representation of the topic. Then, we use cosine similarity to calculate the similarity values between topics at adjacent stages and construct a stage inter-topic similarity matrix to determine the evolution

type among topics. However, there are many hidden variables in LDA, and the multi-dimensional Gibbs sampling algorithm is used in the calculation process. To find the n-dimensional stable probability distribution during sampling, a new sample is obtained by rotating sampling on n-coordinate axes. The idea is to fix the random variable in $n - 1$ dimensions first, sample in the remaining dimension, and then rotate the sampling dimensions until the sampling converges to a certain high-dimensional space. The rotation sampling method adopted in the algorithm is to consider the status of each random variable to be equivalent, and the abstract shape of the sampled data in the high-dimensional space is symmetrical under the given conditions. For example, in a sphere with the center of the sphere at the origin of coordinates in a three-dimensional space, the coordinate variable of each dimension and its value are equivalent, and the structure of the space is symmetrical. The specific calculation is shown in Equations (2)–(4).

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1} p(z_n|\theta) p(w_n|z_n, \beta) \tag{2}$$

Formula (2) represents the concrete expression of the joint probability of the LDA model, which reflects the text generation process based on the LDA model. In Equation (2), θ is the topic vector, $p(\theta)$ represents the probability that the topic appears in the text. N represents the number of words to generate the text, z_n represents the selected topic, and $p(z|\theta)$ represents the probability distribution of topic z given θ; $p(w|z)$ represents the probability distribution of w was given topic z.

$$W_z = p(w_1|z) \times w_{k_1} + p(w_2|z) \times w_{k_2} + \ldots + p(w_j|z) \times w_{k_j} \tag{3}$$

Formula (3) refers to the calculation of the topic vector by multiplying the probability of extracted topic words extracted through LDA as weight by the Word2vec word vector. This approach enables the computed vectors to better represent the topics, which in turn improves the accuracy of the computed topic similarity.

In Equation (3), W_z represents the Word2vec trained word vector of topic w_{k_j}, z represents the topic, k_j represents the topic word, and $p(w_j|z)$ is the weight of the word vector w_{k_j}, also the probability value of the topic word w_{k_j}.

$$\text{Similar}(W_{z_1}, W_{z_2}) = \cos\text{ine}(W_{z_1}, W_{z_2}) \tag{4}$$

Formula (4) refers to the degree of similarity measured by calculating the cosine of the angle between two vectors. Cosine similarity can essentially distinguish textual differences in terms of direction, and the closer the cosine of the angle is to 1, the more similar the two vectors are.

3.3. Topic Evolution Type Determination

At present, there is no unified paradigm for the types of topic evolution. This paper divides the pattern of topic evolution into five types: division, merging, inheritance, disappearance, and generation, synthesizing the studies of Ilhan [29] and Li H [30]. Then, the similarity value between stage topics is compared with the set threshold σ to determine the topic evolution type at each stage, as shown in Table 1.

Table 1. The criteria for topic evolution type determination.

Topic Types	Description of Determination Criteria
Division	When $s > \sigma$, a topic at the previous stage can be divided into two or more topics at the current stage.
Merging	When $s > \sigma$, two or more topics at the previous can be merged into one topic at the current stage.
Inheritance	When $s > \sigma$, it means that a topic at the previous stage is relevant to a topic at the current period.
Disappearance	When $s < \sigma$, the topic at the previous stage does not exist at the current stage.
Generation	When $s < \sigma$, the topic at the current stage does not exist at the previous stage.

3.4. Topic Evolution Path Analysis

Considering that there may be lots of topic evolution paths, to recognize the important evolution paths, this paper proposes to perform a co-occurrence analysis of the global topics identified by LDA, then to extract the topic words with higher relevance and influence as the basis for selecting important evolution paths. Meanwhile, we use the Sankey visualization in the pyecharts package of the Python language to show the types of evolution between the topic contents at each stage and construct clear evolution paths. In Sankey, the element blocks represent topics, where the lines indicate the flow direction and connection of energy generated by the topics, using this feature to visualize the updates and increases generated by the topic content over time [31].

4. Empirical Research

4.1. Data Sources

To test the effectiveness of the topic evolution method based on the LDA2vec symmetry model, this paper carried out an empirical study. The experimental data were collected from the field of data security in the Web of Science from 2017 to 2022, and the retrieval time was October 2022. The search formula is determined to be "TI = 'data safety' OR 'data security'". We filtered out the literature with incomplete information, such as conference proceedings and conference reports. As a result, we collected a total of 4784 pieces of literature. After preprocessing the collected literature, we took the Snowden incident in 2013, COVID-19, and the introduction of data security laws by countries in 2021 as the basis for dividing the data into four stages, the budding stage (2007–2012), growth stage (2013–2018), development stage (2019–2020), and accelerated development stage (2021–2022), as shown in Table 2. The number of publications increases annually; especially in the years 2013, 2019, and 2021, we can see a significant increment of literature, which reflects the synergistic development of scientific research with major emergencies and national policies.

Table 2. Description of time division.

Stages	Time Division	Description of the Characteristics at Each Stage	Number of Literature
Budding Stage	2007–2012	Related literature began to be published, but the number of literature per year was sporadic without a specific pattern.	405
Growing Stage	2013–2018	The number of literature increased slowly year by year.	1225
Development Stage	2019–2020	The number of literature increased rapidly year by year.	1295
Accelerated Development Stage	2021–2022	The number of literature increased year by year, and the growth rate maintained at a high level.	1859

4.2. Topic Identification

4.2.1. Determination of the Number of Topics Based on the Perplexity

We calculated the perplexity of the global and four-stage datasets to obtain the optimal number of topics. To prevent the overfitting phenomenon, we selected the values where the decrease in perplexity was not obvious or at the inflection point. The trend of perplexity is shown in Figure 2. We finally determined the optimal number of global topics and the four-stage topics were, respectively, 21, 10, 12, 15, and 14. Additionally, we calculated that C was 0.79, which met the topic consistency criterion.

4.2.2. Topic Extraction

After determining the optimal number of valid topics at four stages, we continued to determine the number of corresponding topic words under the global and four-stage valid topics. Taking the interpretability and expressive effect of the topic words as the measure, we continuously adjusted the number of topic words, and finally extracted the first 15 topic words. Due to the restriction of space, we show the first 3 topic words under the global and four-stage topics in Table 3.

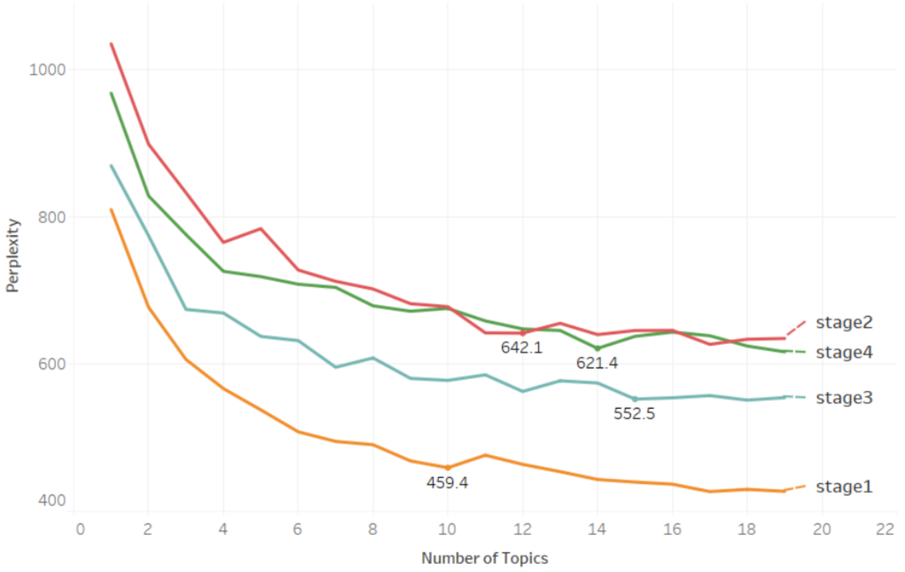


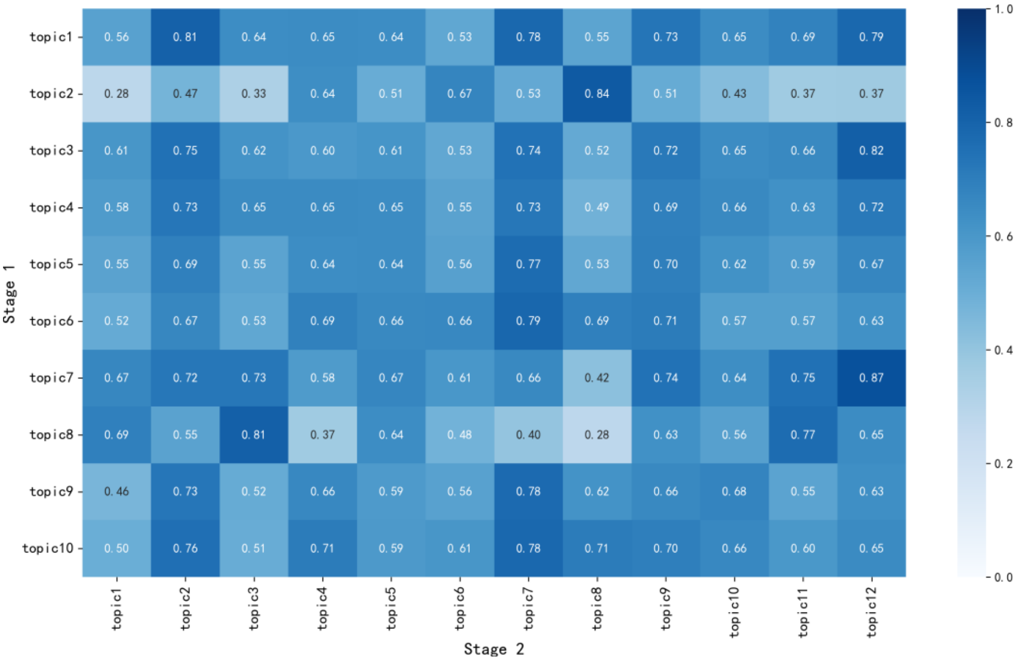
Figure 2. The trend of perplexity with the number of topics.

Table 3. The identification results by LDA.

Stages	Topics
Stage 1	1_1 (services business confidentiality); 1_2 (patients trial safety); 1_3 (protocol data_security service cloud_computing); 1_4 (data_protection shows scalable); 1_5 (technology data_security_and_privacy employing); 1_6(cancer ethical digital); 1_7 (storage service environment); 1_8 (image algorithm mechanism); 1_9 (privacy purpose clinical_trials); 1_10 (concerns healthcare benefit)
Stage 2	2_1 (protocol radio channels); 2_2 (privacy data_sharing make); 2_3 (storage algorithm trusted); 2_4 (intervention migration interest); 2_5 (big_data line choice); 2_6 (routing route patients); 2_7 (trust management review); 2_8 (patients clinical hiv); 2_9 (detection shared algorithms); 2_10 (multi apps effective); 2_11 (image carried sensor); 2_12 (service mechanism cloud_computing)
Stage 3	3_1 (technique evidence secret); 3_2 (speed service community); 3_3 (design tools apps); 3_4 (mechanism anti challenge); 3_5 (privacy big_data technology); 3_6 (patients health healthcare); 3_7 (blockchain supply_chain blockchain_technology); 3_8 (smart_devices merging packets); 3_9 (database hierarchical threat); 3_10 (encryption cryptography algorithm); 3_11 (selected usage university); 3_12 (attribute access_control abe); 3_13 (safety adoption human); 3_14 (image encrypted sensitive); 3_15 (cloud_computing file cloud_storage)
Stage 4	4_1 (blockchain blockchain_technology sharing); 4_2 (distributed supply_chain blockchain); 4_3 (video suggestions home); 4_4 (privacy technology public); 4_5 (big_data algorithm encryption); 4_6 (image encryption images); 4_7 (covid factors china); 4_8 (attacks resources models); 4_9 (healthcare medical patient); 4_10 (energy bit built); 4_11 (market influence setting); 4_12 (patients healthcare health); 4_13 (clinical apps workers); 4_14 (points trading joint)
Global	T_1 (technologies factors china); T_2(attacks supply_chain dynamic); T_3 (local including privacy); T_4 (patients care treatment); T_5 (covid participants education); T_6 (learning transaction share); T_7 (algorithm encryption compared); T_8 (edge attack detection); T_9 (healthcare digital_economy throughput); T_10 (optimization power lightweight); T_11 (machine_learning energy training); T_12 (authentication flow authenticated); T_13 (sharing distributed reduce mechanism); T_14 (big_data management cloud_computing); T_15 (ciphertext attribute cloud_storage); T_16 (patient medical health); T_17 (face delay obtained); T_18 (cyber growth reporting); T_19 (apps acceptance multimodal); T_20 (blockchain technology blockchain_technology); T_21 (image secured images)

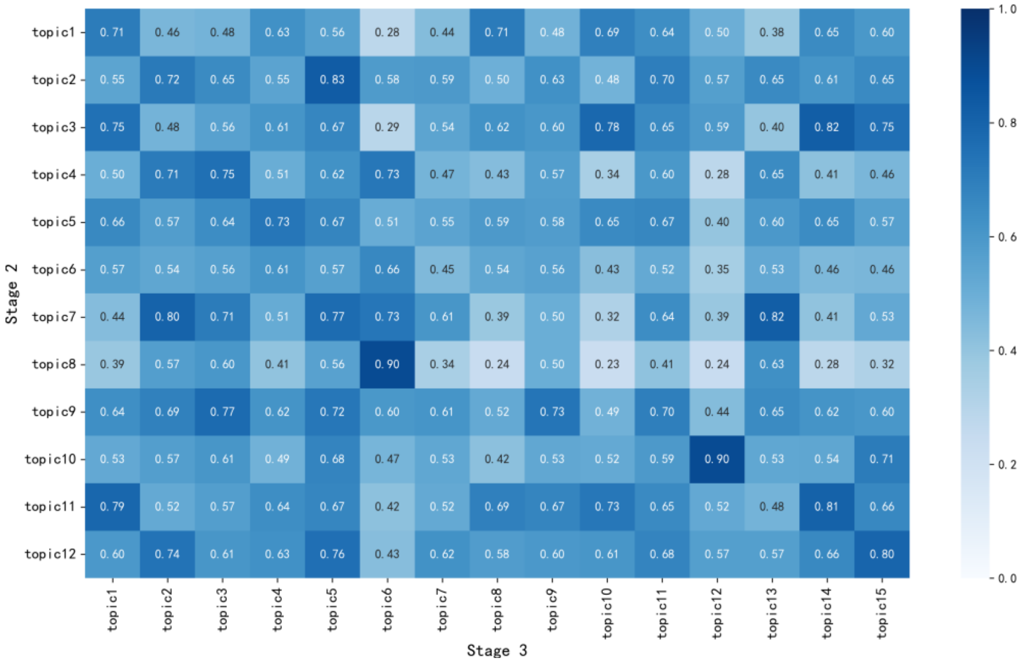
4.3. Topic Evolution Type Analysis

After extracting topics by LDA, we calculated the similarity using Equations (2)–(4) and formed similarity matrices between adjacent stage topics (see Figures 3–5). According to the matrices, we calculated the average of the topic similarity between adjacent stages. When we took the value 0.67 as the threshold, which was the maximum similarity average value of the selected stages based on the experiments, most topics showed a clear evolution type among topics.



Stage 1 \ Stage 2	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10	topic11	topic12
topic1	0.56	0.81	0.64	0.65	0.64	0.53	0.78	0.55	0.73	0.65	0.69	0.79
topic2	0.28	0.47	0.33	0.64	0.51	0.67	0.53	0.84	0.51	0.43	0.37	0.37
topic3	0.61	0.75	0.62	0.60	0.61	0.53	0.74	0.52	0.72	0.65	0.66	0.82
topic4	0.58	0.73	0.65	0.65	0.65	0.55	0.73	0.49	0.69	0.66	0.63	0.72
topic5	0.55	0.69	0.55	0.64	0.64	0.56	0.77	0.53	0.70	0.62	0.59	0.67
topic6	0.52	0.67	0.53	0.69	0.66	0.66	0.79	0.69	0.71	0.57	0.57	0.63
topic7	0.67	0.72	0.73	0.58	0.67	0.61	0.66	0.42	0.74	0.64	0.75	0.87
topic8	0.69	0.55	0.81	0.37	0.64	0.48	0.40	0.28	0.63	0.56	0.77	0.65
topic9	0.46	0.73	0.52	0.66	0.59	0.56	0.78	0.62	0.66	0.68	0.55	0.63
topic10	0.50	0.76	0.51	0.71	0.59	0.61	0.78	0.71	0.70	0.66	0.60	0.65

Figure 3. Similarity matrix between stage 1 and stage 2.



Stage 2 \ Stage 3	topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic8	topic9	topic10	topic11	topic12	topic13	topic14	topic15
topic1	0.71	0.46	0.48	0.63	0.56	0.28	0.44	0.71	0.48	0.69	0.64	0.50	0.38	0.65	0.60
topic2	0.55	0.72	0.65	0.55	0.83	0.58	0.59	0.50	0.63	0.48	0.70	0.57	0.65	0.61	0.65
topic3	0.75	0.48	0.56	0.61	0.67	0.29	0.54	0.62	0.60	0.78	0.65	0.59	0.40	0.82	0.75
topic4	0.50	0.71	0.75	0.51	0.62	0.73	0.47	0.43	0.57	0.34	0.60	0.28	0.65	0.41	0.46
topic5	0.66	0.57	0.64	0.73	0.67	0.51	0.55	0.59	0.58	0.65	0.67	0.40	0.60	0.65	0.57
topic6	0.57	0.54	0.56	0.61	0.57	0.66	0.45	0.54	0.56	0.43	0.52	0.35	0.53	0.46	0.46
topic7	0.44	0.80	0.71	0.51	0.77	0.73	0.61	0.39	0.50	0.32	0.64	0.39	0.82	0.41	0.53
topic8	0.39	0.57	0.60	0.41	0.56	0.90	0.34	0.24	0.50	0.23	0.41	0.24	0.63	0.28	0.32
topic9	0.64	0.69	0.77	0.62	0.72	0.60	0.61	0.52	0.73	0.49	0.70	0.44	0.65	0.62	0.60
topic10	0.53	0.57	0.61	0.49	0.68	0.47	0.53	0.42	0.53	0.52	0.59	0.90	0.53	0.54	0.71
topic11	0.79	0.52	0.57	0.64	0.67	0.42	0.52	0.69	0.67	0.73	0.65	0.52	0.48	0.81	0.66
topic12	0.60	0.74	0.61	0.63	0.76	0.43	0.62	0.58	0.60	0.61	0.68	0.57	0.57	0.66	0.80

Figure 4. Similarity matrix between stage 2 and stage 3.

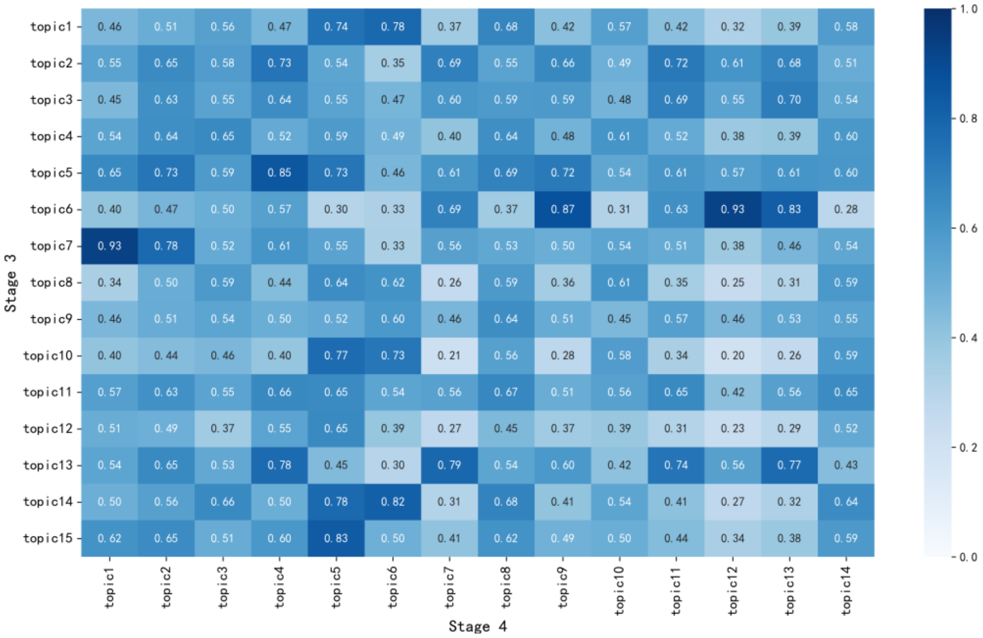


Figure 5. Similarity matrix between stage 3 and stage 4.

(1) Topics of division and merging types

Topics of division and merging types account for the largest proportion and can divide or merge into new research directions with the continuous development of science and technology, policies, and other factors, and have a strong capacity for topic evolution. Regarding the topics of division type: according to the topic evolution path criteria of division (see Section 3.3) and similarity matrix between stage 1 and stage 2 (see Figure 3), we find the topic stage 1_5 divides into stage 2_2, stage 2_7, stage 2_9, and stage 2_12, indicating that the data security and privacy issues caused by technology (stage 1_5) belong to the fundamental research in the field of data security, and are dividing into the data use and governance studies of sharing (stage 2_2), trust (stage 2_7), detection (stage 2_9), and service (stage 2_12). Regarding the topics of merging type: According to the topic evolution path criteria of merging (see Section 3.3) and similarity matrix between stage 3 and stage 4 (see Figure 2), we find the topic stage 4_7 merges from stage 3_2, stage 3_6, and stage 3_13. It indicates that the security issues of usage data (stage 3_13), such as community services (stage 3_2) and patient medical health data (stage 3_6), are merging into the new security utilization study of big data for COVID-19 (stage 4_7) represented by China.

(2) Topics of disappearance and generation types

Regarding the topics of disappearance type: their research heat is gradually declining or even disappearing. Identifying such topics can provide a reference for researchers and government governance for data security. According to the topic evolution path criteria of disappearance (see Section 3.3) and all similarity matrices (Figures 3–5), it can be seen that such topics are: stage 2_6, stage 3_4, stage 3_8, stage 3_9, stage 3_11, and stage 3_12, which mainly involve the research of information system infrastructure security and traditional data security technologies, such as route, mechanism, database, process control, etc. They mostly distribute at the development stage. Regarding the topics of generation type: they are novel and emerging hot topics, reflecting the trend of research topics in the field of data security. According to the topic evolution path criteria of generation (see Section 3.3) and all similarity matrices (Figures 3–5), it can be seen that such topics are: stage 2_6, stage 3_4, stage 3_8, stage 3_9, stage 3_11, and stage 3_12, which mainly involve the data security issues related to blockchain, video, energy, and applications. They mostly distribute at the accelerated development stage and are the most concerned and studied questions in the era of big data.

4.4. Comprehensive Path Analysis

After similarity calculation and topic evolution type determination, we constructed the stage topic evolution paths, as shown in Figure 6. The Sankey diagram visually and clearly shows the topic evolution process in the whole life cycle of the data security domain as of October 2022. The horizontal flow of the Sankey diagram depicts the type and process of topic evolution at four stages, and the vertical element blocks represent the distribution and importance of topics at each stage. We selected some of the topic words to characterize the topics in our analysis.

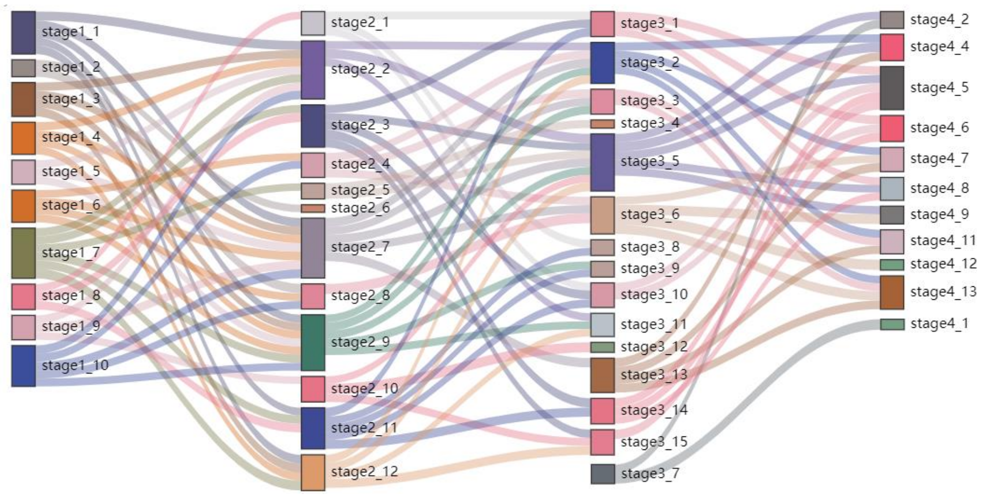


Figure 6. Topic evolution paths in data security based on the LDA2vec symmetry model.

4.4.1. Topic Evolution Path Selection

To select significant topic evolution paths, we co-occurred with the LDA clustering results of global topics from the overall perspective of topic evolution in the field of data security to obtain a co-occurrence matrix of size 39 × 39. Then, we used Ucinet to convert the co-occurrence matrix into Pajek format, which was later imported into VOSviewer to map out the topic word co-occurrence mapping of data security. We eventually extracted three important influencing factors, namely health, technology, and data application, and governance. These topics with high relevance and impact are characterized by research continuity and growth, and therefore are the focus of attention in the field of data security, and the visualization result is shown in Figure 7.

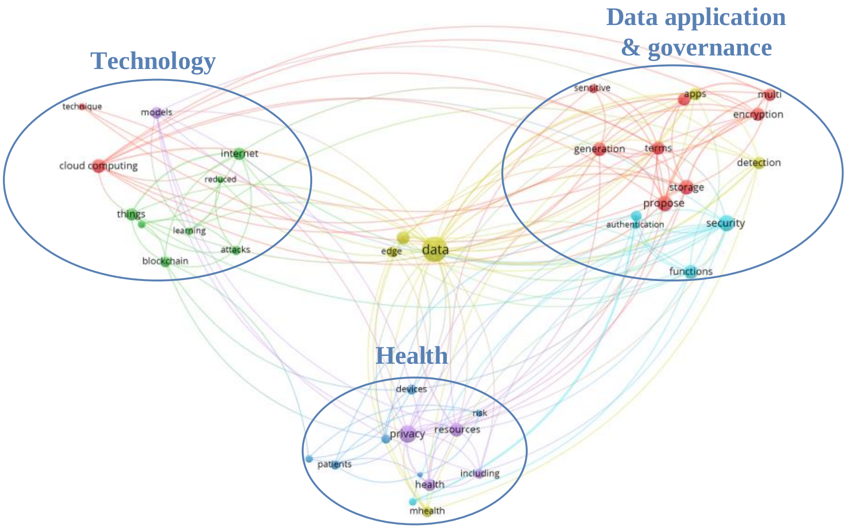


Figure 7. Visualization of global topic category division.

4.4.2. Evolution Path Analysis of Important Topics

Combining the above global topic classification results and topic evolution paths, the paths with high topic importance in data security were extracted for visual analysis. Eventually, we formed three main evolution paths, as shown in Figure 8.

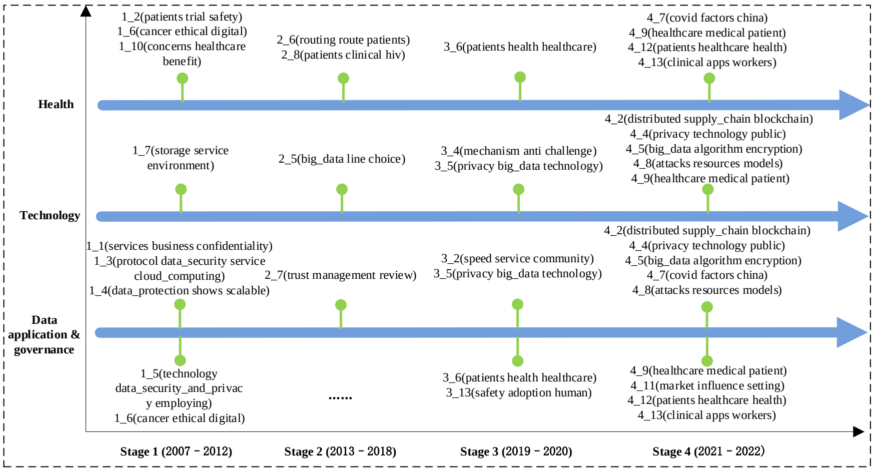

1_2(patients trial safety)
1_6(cancer ethical digital)
1_10(concerns healthcare benefit)
2_6(routing route patients)
2_8(patients clinical hiv)
3_6(patients health healthcare)
4_7(covid factors china)
4_9(healthcare medical patient)
4_12(patients healthcare health)
4_13(clinical apps workers)
Health
1_7(storage service environment)
2_5(big_data line choice)
3_4(mechanism anti challenge)
3_5(privacy big_data technology)
4_2(distributed supply_chain blockchain)
4_4(privacy technology public)
4_5(big_data algorithm encryption)
4_8(attacks resources models)
4_9(healthcare medical patient)
Technology
1_1(services business confidentiality)
1_3(protocol data_security service cloud_computing)
1_4(data_protection shows scalable)
2_7(trust management review)
3_2(speed service community)
3_5(privacy big_data technology)
4_2(distributed supply_chain blockchain)
4_4(privacy technology public)
4_5(big_data algorithm encryption)
4_7(covid factors china)
4_8(attacks resources models)
Data application & governance
1_5(technology data_security_and_privacy employing)
1_6(cancer ethical digital)
......
3_6(patients health healthcare)
3_13(safety adoption human)
4_9(healthcare medical patient)
4_11(market influence setting)
4_12(patients healthcare health)
4_13(clinical apps workers)
Stage 1 (2007 – 2012)
Stage 2 (2013 – 2018)
Stage 3 (2019 – 2020)
Stage 4 (2021 – 2022)

Figure 8. Topic evolution path map.

The first is the topic evolution path of data security in healthcare, in which division-type topics account for the majority. These topics mainly involve data security issues, such as user health data privacy, medical ethics, and electronic health data throughout the whole evolution path, reflecting the heat and importance of topic research on health data security [32]. Among them, stage 4_13 (clinical app workers) belongs to the newborn topic. In the context of big data and the COVID-19 outbreak, various algorithm-driven medical and health applications emerged in large numbers, providing people with health information services while hiding various security risks, which became a hot research issue for scholars nowadays.

The second is the topic evolution path of data security technology, in which inheritance-type topics account for the majority. These topics focus on the security requirements during the whole life cycle, which includes data collection, transmission, storage, processing, sharing, and destruction. Moreover, they are continuously upgraded and optimized with the changes in the scale, confidentiality, and sensitivity of data caused by the application of technologies, such as big data, blockchain, supply chain, encryption, and decryption algorithms. Among them, stage 3_4 (mechanism anti challenge) belongs to the disappearance-type topic. It indicates that the form of data storage and computing changed with the development and application of cloud computing technology, leading to increased data security needs and serious challenges to the coping mechanism. Meanwhile, the form of risk gradually shifts to the evolution of algorithmic black boxes caused by the application of emerging technologies [33]. To ensure the effectiveness of data security governance, how to drive the evolution of data security technology from traditional cryptography to emerging areas, such as cloud computing, big data, IoT, and blockchain in 2019–2020, is one of the important research directions in the field of data security.

Because of the wide range of data application and governance topics, we selected the third evolution path based on topic stage 2_7 (trust management review), which has a strong ability to divide and merge. The third evolution path is related to data application and governance, which mainly studied the characteristics of confidentiality, integrity, and availability of data. The topics in the third evolution path focus on data security issues, such as personal privacy protection, digital ethics, and data security monitoring at the budding and growing stages. At the development and accelerated development stages, the emergence of generation-type topics, such as topic stage 3_7 (blockchain supply_chain blockchain_technology), indicates that with the application of emerging technologies including big data, cloud computing, and blockchain, algorithmic black box arose, making data security issues gradually unvisualized [34,35]. In this case, traditional data security governance failed to meet security needs, and data security governance is in a state of response and defense passively. Meanwhile, against the background of the digital economy development strategies promulgated and the COVID-19 outbreak in many countries, the topics are more focused on the protection of big data privacy and the security of medical and health data.

In summary, we have some new findings. Regarding the types of topic evolution: In the field of data security, there are mainly inheritance, division, and merging-type topics at the budding stage (2007–2012) and growing stage (2013–2018). However, as scientific research continues to progress, disappearance and generation-type topics arise at the development stage (2019–2020) and accelerated development stage (2021–2022), injecting new vitality into scientific research in the field of data security. Regarding topic evolution paths: On the one hand, personal data security is attached to importance, especially in medical and health, which runs through data security governance all the time. On the other hand, it can be seen that the trend of data security governance is gradually shifting to synergistic governance of data and algorithmic security through the topic evolution paths. Faced with the dual challenges of rapid iterative escalation of risks brought by emerging technology applications and the inability of existing data security governance mechanisms to respond flexibly, future research topics in the field of data security will gradually evolve in the direction of "technology governance" and "collaborative governance of data and algorithms".

5. Conclusions

The paper proposes a method for identifying the evolution paths of literature topics using a symmetric model based on LDA2vec. This approach utilizes Bayesian symmetric sampling and Dirichlet symmetric distribution in LDA, as well as features such as the symmetric mapping of data in high-dimensional space in Word2vec. By leveraging structural symmetry, the proposed method analyzes the evolution paths of topics, enhancing the accuracy of topic similarity calculation and addressing the issue of the excessive number of evolution paths, which makes it difficult to quantitatively select important paths. After describing the principle of the proposed method, we selected literature in the field of data security to verify the validity of the method. The empirical results show that the proposed method could accurately display the development process and evolution types in data security. Finally, based on the empirical results, we recognized three important evolutionary paths in the field of data security, which were health, technology, and data application and governance.

The following two points are the research limitations of this paper: ① In the course of the experiment, we proposed a relatively comprehensive method for constructing valid and invalid word lists. However, there are inevitably some omissions that may affect the experimental results. ② The method proposed in the study is an unsupervised one, which can only be verified scientifically and effectively by means of empirical studies and comparing common word vector models, such as BERT and Word2vec. In the future, we will further explore the application of deep learning models, such as transformer, etc., in the topic evolution field to improve the accuracy of topic recognition and evolution paths.

Author Contributions: Conceptualization, T.Z. and W.C.; methodology, W.C.; software, W.C.; validation, W.C.; formal analysis, W.C.; resources, X.L.; writing—original draft preparation, W.C.; writing—review and editing, W.C. and J.L.; supervision, T.Z. and L.J.; project administration, T.Z.; funding acquisition, T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Foundation of China, grant number 22BTQ064.

Data Availability Statement: Data is available upon request to the authors.

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

References

1. Liang, S.; Liu, X. Research progress on topic evolution of scientific and technical literature based on text mining. *Libr. Inf. Serv.* **2022**, *66*, 138–149.
2. Martin, G.; Tiago, P.P.; Eduardo, G.A. A network approach to topic models. *Sci. Adv.* **2018**, *4*, eaaq1360.
3. Prasanna, K.; Seetha, M. A doubleton pattern mining approach for discovering colossal patterns from biological dataset. *Int. J. Comput. Appl.* **2015**, *119*, 41–47. [CrossRef]
4. Kottapalle, P.; Maddala, S.; Gunjan, V.K. D-mine: Accurate discovery of large pattern sequences from biological datasets. In *Proceedings of the International Conference on Soft Computing Systems: ICSCS 2015*; Springer: New Delhi, India, 2016; Volume 1, pp. 647–661.
5. Zhu, Q.; Leng, F. Analysis of topic evolution based on co-citation of documents on the main citation path. *J. China Soc. Sci. Tech. Inf.* **2014**, *33*, 498–507.
6. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Informetrics* **2011**, *5*, 146–166. [CrossRef]
7. Santos, B.S.; Silva, I.; Costa, D.G. Symmetry in Scientific Collaboration Networks: A Study Using Temporal Graph Data Science and Scientometrics. *Symmetry* **2023**, *15*, 601. [CrossRef]
8. Yan, N.; Feng, T.; Hu, Y.; Qi, X. Understanding Aging Policies in China: A Bibliometric Analysis of Policy Documents, 1978–2019. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5956.
9. Blei, D.M.; Ng, A.Y.; Jordan, M.J. Latent Dirichlet allocation. *Mach. Learn. Res.* **2003**, *3*, 993–1022.
10. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
11. Lechao, X.; Jeffrey, P. Synergy and Symmetry in Deep Learning: Interactions between the Data, Model, and Inference Algorithm. In Proceedings of the 39th International Conference on Machine Learning: ICML 2022, Baltimore, MD, USA, 17–23 July 2022; pp. 24347–24369.
12. Ning, B.; Zong, X.; He, K.; Lian, L. PREIUD: An Industrial Control Protocols Reverse Engineering Tool Based on Unsupervised Learning and Deep Neural Network Methods. *Symmetry* **2023**, *15*, 706. [CrossRef]
13. Choi, H.S.; Lee, W.S.; Sohn, S.Y. Analyzing research trends in personal information privacy using topic modeling. *Comput. Secur.* **2017**, *67*, 244–253. [CrossRef]
14. Zhang, T.; Ma, H. Clustering Policy Texts Based on LDA Topic Model. *Data Anal. Knowl. Discov.* **2018**, *2*, 59–65.
15. Xue, J.; Chen, J.; Hu, R.; Chen, C.; Zheng, C.; Su, Y.; Zhu, T. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *J. Med. Int. Res.* **2020**, *22*, e20550. [CrossRef]
16. Zhou, H.; Yu, H.; Hu, R. Topic evolution based on the probabilistic topic model: A review. *Front. Comput. Sci.* **2017**, *11*, 786–802. [CrossRef]
17. Han, X. Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. *Scientometrics* **2020**, *125*, 2561–2595. [CrossRef]
18. Han, W.; Han, X.; Zhou, S.; Zhu, Q. The Development History and Research Tendency of Medical Informatics: Topic Evolution Analysis. *JMIR Med. Inform.* **2022**, *10*, e31918. [CrossRef] [PubMed]
19. Xu, S.; Shi, Q.; Qiao, X.; Zhu, L.; Zhang, H.; Jung, H.; Lee, S.; Choi, S.P. A Dynamic Users' Interest Discovery Model with Distributed Inference Algorithm. *Int. J. Distrib. Sens. Netw.* **2014**, *10*, 239–245. [CrossRef]
20. Zhu, H.; Qian, L.; Qin, W.; Wei, J.; Shen, C. Evolution analysis of online topics based on 'word-topic' coupling network. *Scientometrics* **2022**, *127*, 3767–3792. [CrossRef]
21. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
22. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: NAACL 2019, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
23. Huang, L.; Chen, X.; Zhang, Y.; Wang, C.; Cao, X.; Liu, J. Identification of topic evolution: Network analytics with piecewise linear representation and word embedding. *Scientometrics* **2022**, *127*, 5353–5383. [CrossRef]
24. Xie, Q.; Zhang, X.; Ding, Y.; Song, M. Monolingual and multilingual topic analysis using LDA and BERT embeddings. *Informetrics* **2020**, *14*, 101055. [CrossRef]
25. Xiaowen, X.; Ying, G.; Xinna, S.; Jin, W. Research on the technical similarity visualization based on word2vec and LDA topic model. *J. China Soc. Sci. Tech. Inf.* **2021**, *40*, 974–983.
26. Baosong, Y.; Longyue, W.; Derek, W.; Lidia, C.; Zhaopeng, T. Assessing the Ability of Self-Attention Networks to Learn Word Order. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: ACL 2019, Florence, Italy, 28 July–2 August 2019; pp. 2633–2643.
27. Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; Li, L. On the Sentence Embeddings from Pre-trained Language Models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: EMNLP 2020, Online, 16–20 November 2020; pp. 9119–9130.
28. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef] [PubMed]

29. İlhan, N.; Öğüdücü, Ş.G. Predicting community evolution based on time series modeling. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: ASONAM 2015, Paris, France, 25–28 August 2015; pp. 1509–1516.
30. Li, H.; Hu, J.; Tong, Z. Subject Topic Mining and Evolution Analysis with Multi-Source Data. *Data Anal. Knowl. Discov.* **2022**, *6*, 44–55.
31. Tan, C.; Xiong, M. Contrastive analysis at home and abroad on the evolution of hot topics in the field of data mining based on the LDA Model. *Inf. Sci.* **2021**, *39*, 174–185.
32. Yigzaw, K.Y.; Olabarriaga, S.D. (Eds.) Health data security and privacy: Challenges and solutions for the future. In *Roadmap to Successful Digital Health Ecosystems: A Global Perspective*; Academic Press: Cambridge, MA, USA, 2022; pp. 335–362.
33. Djenna, A.; Bouridane, A.; Rubab, S.; Marou, I.M. Artificial Intelligence-Based Malware Detection, Analysis, and Mitigation. *Symmetry* **2023**, *15*, 677. [CrossRef]
34. Iqbal, F.; Boutaba, R. Data Security in Cloud Computing: Challenges and Solutions. *IEEE Commun. Mag.* **2021**, *59*, 88–94.
35. Pathak, R.; Soni, B.; Muppalaneni, N.B. Role of Blockchain in Health Care: A Comprehensive Study. In *Proceedings of the 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022*; Springer Nature: Singapore, 2023; pp. 137–154.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.