Article

Bert-Based Latent Semantic Analysis (Bert-LSA): A Case Study on Geospatial Data Technology and Application Trend Analysis

Quanying Cheng^{1,2} , Yunqiang Zhu^{1,3,*}, Jia Song^{1,3}, Hongyun Zeng⁴ , Shu Wang¹ , Kai Sun¹
and Jinqu Zhang⁵

- ¹ State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; chengqy.18b@igsrr.ac.cn (Q.C.); songji@reis.ac.cn (J.S.); wangshu@igsrr.ac.cn (S.W.); sunk@reis.ac.cn (K.S.)
- ² College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
- ³ Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
- ⁴ School of Earth Sciences, Yunnan University, Kunming 650500, China; hy_zeng@ynu.edu.cn
- ⁵ School of Computer Science, South China Normal University, Guangzhou 510000, China; zjq@scnu.edu.cn
- * Correspondence: zhuyq@reis.ac.cn

Citation: Cheng, Q.; Zhu, Y.; Song, J.; Zeng, H.; Wang, S.; Sun, K.; Zhang, J. Bert-Based Latent Semantic Analysis (Bert-LSA): A Case Study on Geospatial Data Technology and Application Trend Analysis. *Appl. Sci.* **2021**, *11*, 11897. <https://doi.org/10.3390/app112411897>

Academic Editors: Arturo Montejo-Ráez, Salud María Jiménez-Zafra and Rafael Valencia-García

Received: 7 November 2021

Accepted: 12 December 2021

Published: 14 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Geospatial data is an indispensable data resource for research and applications in many fields. The technologies and applications related to geospatial data are constantly advancing and updating, so identifying the technologies and applications among them will help foster and fund further innovation. Through topic analysis, new research hotspots can be discovered by understanding the whole development process of a topic. At present, the main methods to determine topics are peer review and bibliometrics, however they just review relevant literature or perform simple frequency analysis. This paper proposes a new topic discovery method, which combines a word embedding method, based on a pre-trained model, Bert, and a spherical k-means clustering algorithm, and applies the similarity between literature and topics to assign literature to different topics. The proposed method was applied to 266 pieces of literature related to geospatial data over the past five years. First, according to the number of publications, the trend analysis of technologies and applications related to geospatial data in several leading countries was conducted. Then, the consistency of the proposed method and the existing method PLSA (Probabilistic Latent Semantic Analysis) was evaluated by using two similar consistency evaluation indicators (i.e., U-Mass and NMPI). The results show that the method proposed in this paper can well reveal text content, determine development trends, and produce more coherent topics, and that the overall performance of Bert-LSA is better than PLSA using *NPMI* and U-Mass. This method is not limited to trend analysis using the data in this paper; it can also be used for the topic analysis of other types of texts.

Keywords: trend analysis; topic modeling; Bert; geospatial data technology and application

1. Introduction

Geographical data describes a location and its spatial characteristics attributes. With the rapid development of information technology, geospatial data has become an indispensable data resource for research and application in many fields, such as natural resource management, disaster emergency management, climate change and precision agriculture, etc. [1]. The technologies and applications related to geospatial data are also constantly advancing and upgrading, making new ways of thinking possible, so identifying technologies and applications is helpful to foster and fund further innovation. Through topic analysis, we can identify new research hotspots, acquire knowledge transfer processes [2], and quickly analyze the entire development process of research areas, thus benefiting researchers who are interested in a topic. In addition, it can also provide signals for paradigm shifts in discipline development [3]. For individuals, the results of topic analysis provide

an overview of the evolution of the research field and are helpful to us in grasping research trends, keeping up-to-date with the latest research trends in the field, and seeking scientific collaborators [4].

The extensive scientific literature provides researchers with a wealth of information, which is also an important data resource for analyzing the development trends. However, the time and cost for understanding and analyzing the complex dynamics of current technical approaches related to geospatial data are increasing [5]. Therefore, researchers try to save time and reduce costs by seeking automated analysis methods, which allow them to quickly find the most important information, in order to make critical decisions without consulting voluminous literature [6]. At present, there are two main methods for identifying topics in texts. One is a qualitative appraisal method used by the academia, which is known as expert overview. The other is a scientometrics-based approach. Expert overview is a comprehensive and effective method for topic identification, but it is highly dependent on expert opinion, which is time- and energy-consuming. In addition, expert overview is becoming an increasingly inefficient means, due to the explosive growth of the scientific literature. In comparison, the bibliometric approach uses related papers for statistical frequency analysis, and simply captures information such as citation statistics to identify topics. An article with a high citation count is considered as a high-value one [7]. The structural and geospatial developments of industrial symbiosis as subfields of industrial ecology have been explored by using bibliometrics [8]. A statistical approach to bibliometric data from U.S. institutions has also been used to identify institutional hotspots on a map where many high-impact papers are published. The bibliometrics-based approach plays a role in identifying the development of trends, but it lacks consideration of the content of the literature texts themselves.

Previous studies are mainly based on traditional methods, which merely review relevant literature or conduct simple frequency analysis without providing insights beyond revealing information about the contents of literature texts [9]. Therefore, it is urgent to conduct comprehensive and in-depth trend analysis of literature texts. Recently, a popular method involves text analysis techniques to identify the main viewpoints and trends of the research [10], for example, by using textual data such as user comments, papers, and patents to analyze keywords or social networks [11–16]. In particular, topic modeling has recently attracted the attention of trend analysis researchers, since the main purpose of trend analysis based on textual data is to detect the upward and downward trends in the frequency of each topic in the target document [17]. Topic modeling originates from early latent semantic analysis (LSA), which aims to discover meaningful semantic structures in the corpus [18], with a focus on keyword extraction. The representative approaches are through the use of TF-IDF, which is based on statistical features [19,20], TextRank, based on word graph models [21,22], and Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA), based on topic models [23]. PLSA and LDA are the most widely used probabilistic techniques in topic modeling [24]. PLSA is a latent variable model based on co-occurrence data item-document matrices, also known as the ASPECT model [25]. The superiority of PLSA is demonstrated by its comparison with k-means and LSA [26]. As a variant or extension of PLSA, LDA uses Bayesian methods for parameter estimation to compensate for the incompleteness of PLSA, in terms of topic probability distribution. However, it is difficult to explain LDA without prior knowledge of the underlying topics and hyperparameters. All these approaches mentioned above ignore the most important semantic features of words and the semantic associations between words. Although pre-trained word embeddings are widely used in classification tasks, its application in topic modeling mainly focuses on probability techniques, such as in LDA [27–31], and there is also preliminary work using these embeddings to evaluate the consistency of topic models [32,33].

The probability-based statistical topic modeling methods aforementioned are unable to capture the whole context of a document, as they usually consider only a single graph representation of a word [1]. Alternatively, the n-gram representation that considers multi-

ple words simultaneously can be used, but the efficiency of the model rapidly decreases due to the dimension disaster [34]. Therefore, Bert quantifies words as a vector, which takes into account the context and locates similar words in a similar space to address the limitations of this representation. Although this representation, based on pre-trained models, is widely used and its performance has been validated in recent text analysis, few attempts have yet been made to develop new topic models that are based on Bert. At present, only a few studies have adopted semantic embedding in topic analysis. A recent study uses Bert to generate text semantics as the input for topic classification [35]. While literature abstracts are a core corpus that reveal the distribution of research topics [36,37], these classic scientific journals can extract topic terms to analyze the trends in the research fields [38,39].

In conclusion, both the existing types of studies on geospatial data trend analysis have their limitations. Studies based on screening reviews require a lot of time and energy to screen and summarize all the literature. Methods based on bibliometric analysis are not suitable for discovering potential patterns in fields related to geospatial data. In addition, topic models, which are often used for trend analysis in other fields, are usually based on single-word vector representations that are non-contextual and sparse. In order to overcome these problems, this paper proposes a new topic modeling approach, which applies a new word embedding method in the field of computer linguistics to topic models and can help extract textual topics, namely the Bert-based Latent Semantic Analysis (Bert-LSA) topic modeling approach. It utilizes the Bert contextual word embedding algorithm and spherical k-means clustering to combine context embedding and clustering in a coordinated way, and finally assigns topics to documents. The proposed method is used to conduct a specific trend analysis of the technologies related to geospatial data, which can serve as an advanced and useful alternative method to extract meaningful topics involved in the current trend of geospatial data.

The structure of this paper is as follows. In Section 2, the textual data sources and data pre-processing are introduced. A new topic modeling approach is proposed to compensate for the limitations of existing technologies, which will be discussed in detail in Section 3. Section 4 presents the results of the trend analysis. Section 5 evaluates the proposed method in contrast with existing methods for topic consistency. Section 6 discusses the results and conclusions of this study.

2. Materials

Figure 1 shows the process of data collection and pre-processing used to conduct topic modeling about geospatial data. For the analysis of geospatial data technologies and application trends, abstracts of papers related to geospatial data were collected from two paper databases, namely, ScienceDirect and Scopus. A total of 609 abstracts of papers were collected, which contained terms such as “geospatial data” from 2016 to 2020. In the ScienceDirect database, the query statement was “TITLE-ABS-KEY (geospatial AND data)”, and in the Scopus database, the query statement was based on keywords (geospatial data). In the query result, only that were those connected with two words: “geospatial” and “data” were selected. In order to ensure that each abstract contains rich information, only those abstracts with more than 180 characters were selected, and 266 abstracts were finally analyzed. For those collected data, each abstract was used as an input to the Bert model, and obtained the corresponding word vector.

Figure 2 represents the number of papers related to geospatial data published per year and by country. The number of published papers has been consistently increasing from 2016 to 2019, then with a significant decline in 2020 (Figure 2a). The number of papers in the top seven countries, namely, the USA, China, India, Germany, UK, RF (Russian Federation), and Italy, accounts for about 56% of the total number of papers, with the USA having the largest number of published papers (Figure 2b).

Figure 1. Process of data collection and pre-processing.

Figure 2. (a) Number of papers related to geospatial data published per year; (b) Number of papers related to geospatial data published by country.

3. Methodology

3.1. Overall Framework

In this paper, the Bert-LSA topic model is proposed, which combines Bert and spherical k-mean clustering. The model is featured due to its ability to fully take into account the context of documents and to overcome the shortcomings of existing statistical models. Figure 3 depicts the whole process of document topic generation, which is mainly divided into four steps as follows.

- Step 1: All documents are taken as corpus, and the m-dimensional word vector corresponding to the documents is obtained by the Bert model, which is denoted as $v_i \in V^m$, where v_i is the word vector and V^m is the m-dimensional vector space. Note that here the word vectors are obtained after the documents are processed as inputs into the Bert model, rather than being directly obtained from the pre-trained model.

- Step 2: All vectorized words undergo spherical k-means clustering, which first initializes the centroid according to the K value, and then calculates the spherical distance from each word vector v_i to the centroid. According to the distance value, v_i will be assigned to different categories, which will be iterated until convergence. Finally, K clusters are obtained, each of which is called a topic.
- Step 3: The graphical representation of the generation method for each particular document vector d_j , $j = 1, 2, \dots, D$, is shown in Figure 4, which is obtained by multiplying the m-dimensional vector v_i of all words in the corpus with the term document matrix $Num \times D$, where Num is the number of words in the corpus and D is the number of documents. See Section 3.4 for details.
- Step 4: Figure 5 depicts the process of document topic generation. The cosine distance between each document and the word vector contained in each topic in Step 2 is calculated in turn, and each document is assigned to a different topic by using a topic assignment method. See Section 3.5 for details.

Figure 3. Document topic generation method.

3.2. Word Vector Generation Based on Bert

Bert (Bidirectional Encoder Representations from Transformers) [40] is a pre-trained language model released by Google that has occupied a state-of-the-art position in 11 tasks in the NLP (natural language processing) field. It is based on a multi-layer bidirectional transformer [41], and the framework consists of two steps: pretraining and fine-tuning. In the pretraining stage, it is trained on existing unlabeled text in advance and is released as a general language model. In the fine-tuning stage, it can be fine-tuned using learning data, according to the task to be performed [42,43].

In this paper, the pre-trained Bert model called “Bert-Base, Uncased” [44] is used to generate word vectors and represent the semantics of words. The reason for choosing this model is that the Bert-Base model is smaller than the Bert-Large model, and the language used in the research is only English and does not need to be case-sensitive. By entering sentences of each document, we obtain the word vector corresponding to each word in the sentence, which can accurately represent the semantic meaning of the word in its context. Python executes all the word vector generations mentioned in this paper by using the API released by [45]. It is an open-source Bert service, which allows users to use the Bert model by calling the service without paying attention to the details of Bert implementation. The important parameters `max_seq_len` and `pooling_strategy` are set to 512 and NONE respectively.

Bert takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence that has one or two segments. The first token of the sequence is always [CLS] which contains the special classification embedding, and the other special token [SEP] is used for separating segments. Bert takes the final hidden state, h , of the first token [CLS] as the representation of the whole sequence. WordPiece embedding [46] is used and split word pieces is denoted with ##. So, the statistics of the length of the documents in the datasets are based on the word pieces [47].

3.3. Spherical k -Means Clustering

The spherical k -means method is introduced for clustering sets of sparse text data. This method is based on a vector space model, whose basic principle is to describe the degree to which two vectors point in the same direction by their similarity, rather than their length [48]. For example, in the vector space model V^m , for each word vector $w_i \in V^m, i = 1, 2, \dots, N$, the inner product (Formula (1)) of two vectors is used to express the semantic similarity, where the column vectors are normalized (Formula (2)) to the unit length of the Euclidean norm, with the aim of assigning equal weights to each of the n points in the data set. Of course, we obtained these vectors after entering the text into the Bert model, rather than directly from the Bert model.

$$\cos(\theta_{x, y}) =: x^T y \quad (1)$$

$$\cos(\theta_{x, y}) = ||x||_2 ||y||_2 \quad (2)$$

In Formula (1), it describes the result of the normalization of x and y . In Formula (2), it is the definition of the standard inner product.

Finding clustering centers is also very important. For the clustering vector $v(i) \in 1, 2, \dots, v$ and w_i , the center of clustering is to find the minimum cosine value between w_i and $c_v, v = 1, 2, \dots, v$ [49]. To find the number of clusters in the dataset used in the experimentation, we ran the spherical clustering algorithm for a range of multiple values and compared the results obtained for each value.

3.4. Example of Document Vector Generation

The document vector is generated by multiplying the word vector matrix (A) and the term document matrix (B) in the figure below, i.e., $A \times B$. The word vector matrix (A) is composed of m -dimensional word vectors obtained from all the words contained in the document according to the method in Section 3.2. The term document matrix (B) is obtained by combining the word frequencies of the words contained in a single document, provided that the order of the words in a single document (i.e., the columns in B) needs to be the same as the position of corresponding words in matrix (A) (i.e., the rows in A), and if the words contained in matrix A do not appear in a single document (e.g., *DOC 1*), the value of that position is set to 0.

3.5. Method of Document Topic Determination

Section 3.3 describes how we obtained multiple topics after clustering all documents, including Topic 1, Topic 2, Topic 3, and Topic 4 in Figure 5. Section 3.4 describes how we obtained the vector of each document, Doc N , shown in Figure 5. The method of assigning documents to topics is shown in Figure 5. Taking the first document as an example, firstly, the average value of the five words (boldface in Topic 1) with the largest cosine distance between the document and the first topic (Topic 1) was obtained, which were taken as the similarity between the document and Topic 1. By analogy, the similarity values between the document and other topics were then calculated. Finally, the similarity values between the document and all topics were compared, and the document was assigned to its corresponding topic with the maximum similarity value. Other documents were calculated in the same way, and finally, all documents were assigned to different topics.

Figure 4. Example of document vector generation.

Figure 5. Method of document topic determination.

4. Trend Analysis Based on Bert_LSA

4.1. Topic Selection

In the trend analysis based on Bert_LSA, first, Bert is used to obtain the vector of words contained in the document, and the acquisition method and parameter setting are detailed in Section 3.2. Then, the spherical k -means clustering algorithm is applied for clustering, and the optimal number of clusters k is determined by the elbow method. The core index of the elbow method is Sum of the Squares Errors (SSE), and the formula is as follows:

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2 \quad (3)$$

where C_i is the i -th cluster, p is the sample point in C_i , m_i is the centroid of C_i , and SSE is the clustering error of all samples, which represents the clustering effect.

However, when the effect of the elbow method is not obvious, it is combined with the Silhouette Coefficient method to jointly determine the number of clusters. The Silhouette Coefficient is an index to evaluate the degree of density and dispersion of the class. The calculation method is listed as follows, and its value ranges between $[-1, 1]$. The larger the value is, the more reasonable it is [50].

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

where $a(i)$ represents the average value of the dissimilarity of the i vector to other points within the same cluster, $b(i)$ represents the minimum value of the average dissimilarity of the i vector to other clusters.

4.2. The Result of Trend Analysis

According to the topic number selection method in Section 4.1, the cluster numbers of USA, China, India, Germany, UK, Russian Federation, Italy, and Others were finally set to 6, 5, 6, 5, 7, 6, 7, and 4, respectively. Table 1 shows the results of topic modeling using Bert_LSA.

Table 1. The topic analysis results based on Bert_LSA model, and the percentage indicates the proportion of the country in all documents.

USA		China	
Topic	Ratio (%)	Topic	Ratio (%)
Water/Polarhub/Enviroatlas	21%	Landscape/Livability/Government	29.4%
Building/Air/BIM	19.4%	Extraction/Metadata/Information	23.5%
Fire/Risk/Precipitation	17.7%	Soybean/Crop/Area/Policy	17.6%
GEE/Framework/Model	17.7%	Geohazards/Landslide/Anomaly	14.8%
Stream/Land/Temperature	12.9%	Multisource/Search/Metadata	14.7%
Greenery/Heat	11.3%		
India		Germany	
Topic	Ratio (%)	Topic	Ratio (%)
Cloud/computing/Hadoop/Share	26.3%	Navigation/Prediction/Street	25%
Flood/Distribution/Coastline	21.1%	Visualization/Database/Datasets	25%
SDI/WPS/Framework	21.1%	Change/Land/Observation	18.8%
Stormwater/Groundwater/Conserve	10.5%	Stress/Life/M Measurement	18.8%
Land/Investor/Vicinity	10.5%	Demand/Heat/Supply	12.4%
School/Platform/location	10.5%		
UK		Russian Federation	
Topic	Ratio (%)	Topic	Ratio (%)
Geohazards/Household/Landslide	16.7%	Risk/Environment/Management	27.3%
Point cloud/Framework	16.7%	Network/Generation/Transport	18.2%
Feature/Attribute/Database	16.7%	Customer/Bank/Transaction	18.2%
Mangrove/Fishing/Intensity	16.7%	Monitoring/Change/Climate/Season	18.2%
BIM/Project/Evaluation	16.7%	Client/Cloud/computing/Device	9.1%
Weather/MCSA (Multi-Channel Sequences Analysis)/Condition	8.3%	Image/Anomaly/Validation	9%
Network/Source/Accessibility	8.2%		
Italy		Others	
Topic	Ratio (%)	Topic	Ratio (%)
Geo/Disaster/Cluster	30%	Land/Housing/City/Water	41.2%
Challenge/Spiral/OpenGIS	20%	Village/Fire/model/System	23.5%
Crop/Precision/Classification	10%	Datasets/Soil/Accuracy	23.5%
Landslide/Hazard/Flood	10%	SDI/Web/Collection	11.8%
Map/Territory/Accessment	10%		
Location/Behavior/Category	10%		
GNSS/Radar/Remote/sensing	10%		

The results of the trend analysis show that the focus of each country's concern is different. USA, for example, focuses on the environment, buildings, and fires. However, China pays attention to livability, information extraction, and crops, while words like government and policy also appear. India focuses on information technology, such as cloud computing, SDI and WPS, etc., as well as focuses on disaster events such as floods. Italy and others also pay attention to disaster-related content. Germany, UK, and the Russian Federation all focus on content related to climate change. In general, countries pay more attention to disaster events (e.g., fires and floods), and related information technologies, such as cloud computing, Hadoop and GEE, etc., have also received higher attention. This is also a good indication that our proposed method can successfully identify the current technologies and application trends related to the use of geospatial data, which can quickly provide research hotspots for relevant researchers, especially those who are not specialized in GIS. In this way, it considerably saves the time needed to read a large amount of literature, which is of practical significance.

5. Quantitative Evaluation

5.1. Evaluation Method

The methods of evaluating topic models mainly include perplexity and topic consistency. The perplexity has its own merits, as it can evaluate probability-based topic models well, whereas in non-probability-based topic models, these methods do not capture semantic consistency between words [51]. Topic consistency can be used to measure whether words within a topic are coherent, i.e., if a group of terms are consistent with each other, then these terms are coherent. For a specific topic, the semantic similarity between words in the topic determines the degree of coherence of the topic, so topic consistency can be measured by the semantic similarity between words in the topic [52–55]. The greater the consistency value of the topic is, the more coherent the words of each topic will be. To evaluate the non-probabilistic topic model proposed in this paper, the following two consistency measures were used: (1) University of Massachusetts (U_Mass) [56], (2) Normalized Pointwise Mutual Information ($NPMI$) [57].

U_Mass is defined as:

$$U_Mass = \frac{2}{N \times (N - 1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \quad (5)$$

where $P(w_i, w_j)$ is the joint probability of two words w_i and w_j . A small value for ϵ is chosen to avoid calculating the logarithm of 0.

$NPMI$ is defined as:

$$NPMI = \frac{1}{K} \sum_K \frac{2}{T(T-1)} \sum_{1 \leq i \leq j \leq T} \frac{\log_2 \left(\frac{p(w_i, w_j)}{p(w_i)p(w_j)} \right)}{-\log_2 p(w_i, w_j)} \quad (6)$$

where K is the number of topics, and each topic consists of the T most relevant word. $p(w_i, w_j)$ is the probability that the word pair (w_i, w_j) co-occurs in a document, and $p(w_i)$ is the probability that the word w_i appears in the document.

5.2. Evaluation Result

The method proposed in this paper was compared with PLSA in terms of its topic consistency, where the PLSA implementation uses open-source code (<https://github.com/yedivanseven/PLSA>) (accessed on 7 November 2021). Figures 6 and 7 show the average topic consistency calculated by using PLSA and Bert-LSA, respectively, where the abscissa is the number of words N selected in each topic, with values of N ranging from 3 to 13, and the ordinate is the topic consistency value. Here, the value of topic consistency is the average of the corresponding topic consistency values for all countries when different numbers of topics are selected. When evaluated with the U-mass method, the topic

consistency of the PLSA model remains almost constant as N increases, and its value is generally low. Similarly, the topic consistency of the Bert-LSA model gradually decreases as N increases, but its value is generally higher than that of PLSA model, which means that the Bert-LSA model performs better than the PLSA model.

Figure 6. Topic consistency values of PLSA and Bert-LSA models obtained by the U-Mass method.

Figure 7. Topic consistency values of PLSA and Bert-LSA models obtained with the *NPMI* method.

When evaluated with the *NPMI* method, the topic consistency of the PLSA model decreases when the value of N is from three to five, increases when the value of N is from

five to seven, and then remains basically unchanged thereafter. For the Bert-LSA model, the topic consistency keeps decreasing when the value of N is from three to seven, increases when the value of N is from seven to nine, and then keeps decreasing. On the whole, the Bert-LSA model still outperforms PLSA model.

6. Conclusions and Discussion

In this paper, a new method of topic identification has been proposed. First, a word embedding algorithm was adopted that was based on a pre-trained model, which generates a word representation that can capture the context of a document. After that, we used a spherical k-means clustering algorithm to construct topic clusters. Finally, a topic assignment method was used to assign documents to different topics. The assignment process was in order to calculate the similarity between documents and topics.

The method proposed in this paper was applied to the literature abstracts related to geospatial data. First, it shows the characteristics of geospatial data technology and application development trends in related research in several leading countries. Second, the topic coherence of this method was evaluated by using U-Mass and $NPMI$, and its performance was compared with that of the existing method, PLSA. The results show that the proposed method can produce highly coherent topics. The research in this paper provides new ideas for the trend analysis of technologies and applications related to geospatial data, and helps professionals engaged in research related to geospatial data to identify their future research directions at any time. In addition, this method captures the development trends of related technical fields through text, which can be used as an information tool for anyone who is responsible for strategic decision-making in sectors related to geospatial data, to determine the prospect and market of the fields. This paper also has some shortcomings, for example, it is unable to successfully identify the topic when the number of texts is extremely large. In the future, we will work hard on topic modeling for a large number of texts.

Author Contributions: Conceptualization, methodology, validation, formal analysis, Q.C. and Y.Z.; Software, Q.C.; Supervision, J.S., H.Z., J.Z., K.S. and S.W.; Funding Acquisition, Y.Z.; Writing—original draft preparation, Q.C. and Y.Z.; Writing—review and editing, Q.C. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (grant numbers: 42050101, 41771430, 41631177) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number: XDA23100100).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in Section 2.

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

References

1. Lu, Y.; Zhai, C.X. Opinion Integration through Semi-Supervised Topic Modeling. In Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 21–25 April 2008; pp. 121–130.
2. Li, F.; Li, M.; Guan, P.; Ma, S.; Cui, L. Mapping publication trends and identifying hot spots of research on Internet health information-seeking behavior: A quantitative and co-word biclustering analysis. *J. Med. Internet Res.* **2015**, *17*, e81. [[CrossRef](#)] [[PubMed](#)]
3. Ying, D. Community detection: Topological vs. Topical. *J. Informetr.* **2011**, *5*, 498–514.
4. Chen, X.; Wang, S.; Tang, Y.; Hao, T. A bibliometric analysis of event detection in social media. *Online Inf. Rev.* **2019**, *43*, 29–52. [[CrossRef](#)]
5. Jacobi, C.; Atteveldt, W.V.; Welbers, K. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digit. J.* **2016**, *4*, 89–106. [[CrossRef](#)]
6. Alami, N.; Meknassi, M.; En-Nahni, N.; Adlouni, Y.E.; Ammor, O. Unsupervised Neural Networks for Automatic Arabic Text Summarization Using Document Clustering and Topic modeling. *Expert Syst. Appl.* **2021**, *172*, 114652. [[CrossRef](#)]

7. Chertow, M.R.; Kanaoka, K.S.; Park, J. Tracking the diffusion of industrial symbiosis scholarship using bibliometrics: Comparing across Web of Science, Scopus, and Google Scholar. *J. Ind. Ecol.* **2021**, *25*, 913–931. [\[CrossRef\]](#)
8. Bornmann, L.; Angeon, F.D.M. Hot and cold spots in the US research: A spatial analysis of bibliometric data on the institutional level. *J. Inf. Sci.* **2019**, *45*, 84–91. [\[CrossRef\]](#)
9. Kivikunnas, S. Overview of process trend analysis methods and applications. In Proceedings of the Erudit Workshop on Applications in Pulp and Paper Industry, Aachen, Germany, 9 September 1998; pp. 395–408.
10. Song, M.; Kim, S.Y.; Lee, K. Ensemble analysis of topical journal ranking in bioinformatics. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 1564–1583. [\[CrossRef\]](#)
11. Hung, J.L. Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. *Br. J. Educ. Technol.* **2012**, *43*, 5–16. [\[CrossRef\]](#)
12. Hung, J.L.; Zhang, K. Examining mobile learning trends 2003–2008: A categorical meta-trend analysis using text mining techniques. *J. Comput. High. Educ.* **2012**, *24*, 1–17. [\[CrossRef\]](#)
13. Kim, H.J.; Jo, N.O.; Shin, K.S. Text Mining-Based Emerging Trend Analysis for the Aviation Industry. *J. Intell. Inf. Syst.* **2015**, *21*, 65–82. [\[CrossRef\]](#)
14. Kim, Y.M.; Delen, D. Medical informatics research trend analysis: A text mining approach. *Health Inform. J.* **2018**, *24*, 432–452. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Terachi, M.; Saga, R.; Tsuji, H. Trends Recognition in Journal Papers by Text Mining. In Proceedings of the 2006 IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 8–11 October 2006; IEEE: Taipei, Taiwan, 2006; Volume 6, pp. 4784–4789.
16. Tseng, Y.H.; Lin, C.J.; Lin, Y.I. Text mining techniques for patent analysis. *Inf. Process. Manag.* **2007**, *43*, 1216–1247. [\[CrossRef\]](#)
17. Kang, H.J.; Kim, C.; Kang, K. Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA). *Processes* **2019**, *7*, 379. [\[CrossRef\]](#)
18. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [\[CrossRef\]](#)
19. Guo, A.Z.; Tao, Y. Research and improvement of feature words weight based on TFIDF algorithm. In Proceedings of the 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, Chongqing, China, 20–22 May 2016; pp. 415–419.
20. Li, J.Z.; Fan, Q.N.; Zhang, K. Keyword Extraction Based on tf/idf for Chinese News Document. *Wuhan Univ. J. Nat. Sci.* **2007**, *12*, 917–921. [\[CrossRef\]](#)
21. Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 25–26 July 2004; pp. 404–411.
22. Zhang, X.; Wang, Y.; Wu, L. Research on cross language text keyword extraction based on information entropy and TextRank. In Proceedings of the Information Technology, Networking, Electronic and Automation Control Conference, Chengdu, China, 15–17 March 2019; pp. 16–19.
23. Wei, H.X.; Gao, G.L.; Su, X.D. LDA-based word image representation for keyword spotting on historical Mongolian documents. In Proceedings of the International Conference on Neural Information Processing, Kyoto, Japan, 30 September 2016; pp. 432–441.
24. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
25. Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 30 July 1999; pp. 289–296.
26. Newman, D.J.; Block, S. Probabilistic topic decomposition of an eighteenth-century American newspaper. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 753–767. [\[CrossRef\]](#)
27. Xie, P.; Yang, D.; Xing, E. Incorporating word correlation knowledge into topic modeling. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 725–734.
28. Yang, Y.; Downey, D.; Boyd-Graber, J. Efficient methods for incorporating knowledge into topic models. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 308–317.
29. Das, R.; Zaheer, M.; Dyer, C. Gaussian LDA for topic models with word embeddings. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; Volume 1, pp. 795–804.
30. Nguyen, D.Q.; Billingsley, R.; Du, L.; Johnson, M. Improving topic models with latent feature word representations. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 299–313. [\[CrossRef\]](#)
31. Moody, C.E. Mixing Dirichlet topic models and word embeddings to make lda2vec. *arXiv* **2016**, arXiv:1605.02019.
32. Callaghan, D.; Greene, D.; Carthy, J.; Cunningham, P. An analysis of the coherence of descriptors in topic modeling. *Expert Syst. Appl.* **2015**, *42*, 5645–5657. [\[CrossRef\]](#)
33. Ding, R.; Nallapati, R.; Xiang, B. Coherence-aware neural topic modeling. *Comput. Sci.* **2018**, arXiv:1809.02687.
34. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Ma-Chine Learn. Res.* **2003**, *3*, 1137–1155.
35. Zhou, Y.; Li, C.; He, S.; Wang, X.; Qiu, Y. Pre-trained contextualized representation for chinese conversation topic classification. In Proceedings of the 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), Shenzhen, China, 1–3 July 2019; IEEE: Piscataway, NJ, USA; pp. 122–127.

36. Ji, Q.; Pang, X.; Zhao, X. A bibliometric analysis of research on Antarctica during 1993–2012. *Scientometrics* **2014**, *101*, 1925–1939. [CrossRef]
37. Natale, F.; Fiore, G.; Hofherr, J. Mapping the research on aquaculture: A bibliometric analysis of aqua-culture literature. *Scientometrics* **2012**, *90*, 983–999. [CrossRef]
38. Sung, H.Y.; Yeh, H.Y.; Lin, J.K.; Chen, S.H. A visualization tool of patent topic evolution using a growing cell structure neural network. *Scientometrics* **2017**, *111*, 1267–1285. [CrossRef]
39. Qi, Y.; Zhu, N.; Zhai, Y.; Ding, Y. The mutually beneficial relationship of patents and scientific literature: Topic evolution in nanoscience. *Scientometrics* **2018**, *115*, 893–911. [CrossRef]
40. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Comput. Sci.* **2018**, arXiv:1810.04805.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
42. Yoo, S.Y.; Jeong, O.R. Automating the expansion of a knowledge graph. *Expert Syst. Appl.* **2019**, *141*, 112965. [CrossRef]
43. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to Fine-Tune BERT for Text Classification? In Proceedings of the China National Conference on Chinese Computational Linguistics, Kunming, China, 13 October 2019; pp. 194–206.
44. Available online: https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip (accessed on 7 November 2021).
45. Available online: <https://github.com/hanxiao/bert-as-service> (accessed on 7 November 2021).
46. Wu, Y.H.; Schuster, M.; Chen, Z.F.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *Comput. Sci.* **2016**, arXiv:1609.08144.
47. Available online: <https://spacy.io/> (accessed on 7 November 2021).
48. Dhillon, I.S.; Modha, D.S. Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **2001**, *42*, 143–175. [CrossRef]
49. Buchta, C.; Kober, M.; Feinerer, I.; Hornik, K. Spherical k-means clustering. *J. Stat. Softw.* **2012**, *50*, 1–22.
50. Peter, R.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
51. Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J.L.; Blei, D.M. Reading tea leaves: How humans interpret topic models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 288–296.
52. Aletras, N.; Stevenson, M. Evaluating topic coherence using distributional semantics. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)—Long Papers, Potsdam, Germany, 19–22 March 2013; pp. 13–22.
53. Li, C.; Wang, H.; Zhang, Z.; Sun, A.; Ma, Z. Topic modeling for short texts with auxiliary word embeddings. In Proceedings of the 39th International ACM Sigir Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; ACM: New York, NY, USA, 2016; pp. 165–174.
54. Mimno, D.M.; Wallach, H.M.; Talley, E.M.; Leenders, M.; McCallum, A. Optimizing semantic coherence in topic models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, John McIntyre Conference Centre, Edinburgh, UK, 27–31 July 2011; pp. 262–272.
55. Fu, Q.; Zhuang, Y.; Gu, J.; Zhu, Y.; Guo, X. Agreeing to Disagree: Choosing Among Eight Topic-Modeling Methods. *Big Data Res.* **2021**, *23*, 100173. [CrossRef]
56. Röder, M.; Both, A. Hinneburg, Exploring the space of topic coherence measures. In Proceedings of the 8th ACM International Conference on Web Search and Data Mining, Shanghai, China, 2–6 February 2015; ACM: New York, NY, USA, 2015; pp. 399–408.
57. Verlag, G.N.; Informatik, F. Von der Form zur Bedeutung: Texte automatisch verarbeiten/From Form to Meaning: Processing Texts Automatically. 2009. Available online: <http://tubiblio.ulb.tu-darmstadt.de/98069/> (accessed on 7 November 2021).