Hindawi
Complexity
Volume 2021, Article ID 3471168, 11 pages
https://doi.org/10.1155/2021/3471168

Research Article

Digital Media Hotspot Mining Algorithm Implementation with Complex Systems in the Mobile Internet Environment

Yufeng Jia[1] **and Sang-Bing Tsai**[2]

[1]*School of Design, Dalian Minzu University, Dalian, Liaoning 116600, China*
[2]*Regional Green Economy Development Research Center, School of Business, WUYI University, Nanping, China*

Correspondence should be addressed to Yufeng Jia; jiayufeng811001@163.com

Received 29 October 2021; Revised 19 November 2021; Accepted 4 December 2021; Published 17 December 2021

Academic Editor: Zaoli Yang

With the development of the Internet, the amount of information present on the network has grown rapidly, leading to increased difficulty in obtaining effective information. Especially for individuals, enterprises, and institutions with a large amount of information, it is an almost impossible task to integrate and analyze Internet information with great difficulty just by human resources. Internet hot events mining and analysis technology can effectively solve the above problems by alleviating information overload, integrating redundant information, and refining core information. In this paper, we address the above problems and research hot event topic sentence generation techniques in the field of hot event mining and design a hybrid event candidate set construction algorithm based on topic core word mapping and event triad selection. The algorithm uses the PAT-Tree technique to extract high-frequency core words in topic hotspots and maps the high-frequency words into sentences to generate a part of event core sentences. The other part of event core sentences is extracted from the topic hotspots by making event triples as candidate elements, and sentences containing event elements are extracted from the topic hotspots. The sets of event core sentences generated by the two methods are mixed and filtered and sorted to obtain the candidate set, which can be used to build a word graph-based main service channel (MSC) model. In this paper, we also propose an improved word graph-based MSC model and use it for the extraction of event topic sentences. Based on the above research, a hot event analysis system is implemented. The system analyzes the existing topic data and uses the event topic sentence generation algorithm studied in this paper to generate the titles of hot spots, that is, hot events. At the same time, the topics are displayed from different dimensions, and data visualization is completed. The visualization includes the trend change of event hotness, trend change of event sentiment polarity, and distribution of event article sources.

1. Introduction

Internet hot events mining and analysis technology can effectively solve the above problems by alleviating information overload, integrating redundant information, and extracting core information. Internet hot events refer to a series of news hot topics with sudden growth on the Internet within a short period, which are characterized by huge quantity, fast spreading, and wide spreading. Generally speaking, Internet hot events are hot spots that people are concerned about and contain a lot of effective and useful information, which is of great significance to enterprises and government regulatory departments, and are also generally the Internet hot spots that Internet users are concerned about. Most of the current research on hot events is mostly based on clustering technology for hot topic discovery, but topics cannot be equated with events, which are aggregated from multiple hotspots describing the same events, while events are a phrase that can highly summarize the main content of a topic, and we can understand events as the title of a topic. The research of web hot events mining technology involves natural language processing techniques, such as topic detection and tracking (TDT), hotspot clustering technology, and title generation technology [1]. This technology is to mine information that is valuable to people according to specific needs from big data information. The

technology is often cross-fused with other disciplines to form new applications. Among them, hotspot mining is a derivative product of data mining, and hotspot mining is usually in the hotspot natural language. Based on processing, the purpose of mining important information is achieved through algorithms. Topic detection and tracking technology can automatically identify new topics and continuously track known topics for news media information flow, which can effectively discover hot events generated in a period and is the theoretical basis of event mining. Hotspot clustering technology can cluster and analyze a large number of news hotspots on the Internet. Through unguided clustering of news hotspots, articles with similar hotspot contents can be clustered together to form a preliminary cluster of hotspot events. Title generation technology can generate topic titles, hot events, by analyzing the news hotspots within the topic by extracting the core content of the topic and using sentence compression technology.

Generally speaking, topics are more focused on an academic embodiment of clustering, while events are brief expressions of topic themes. Internet hot events are a linguistic lexical phenomenon, reflecting the hot issues and hot events that people are generally concerned about in a period in the international, national, or regional context. Hot events are closely related to social phenomena, and the function of expressing a public opinion and monitoring public opinion is prominent. By discovering online events, we can have a holistic understanding of the direction of current online public opinion from point to point, which is important for timely detection of negative public opinion and prevention of the large spread of sudden public opinion [2]. Based on the existing topics, we use multidocument title generation technology to analyze topic hotspots and thus generate event topic sentences of topics. The event topic sentence can reflect the topic content well, which makes the user know the topic clearly and saves the user's time to read the topic hotspot. Therefore, the study of event topic sentence generation technology under hot topic mining technology has very important practical significance and value, and the topic is less studied, which is very meaningful and challenging research.

Hotspot mining technology was developed at a high speed in the 1990s, and this technology is used to mine valuable information for people from big data information according to specific needs. This technology is often cross-fertilized with other disciplines to form new applications, of which hotspot mining is a derivative of data mining, and hotspot mining is usually based on natural language processing of hotspots, and the purpose of mining important information is achieved by algorithmic means [3]. Among them, the hotspot clustering technique belongs to one of the important means of data mining; using this method can discover the potential patterns among hotspots from a large and complicated number of hotspots by the idea of clustering similar hotspots into one category and analyzing hot topics based on intracluster and intercluster hotspot information to achieve the purpose of data mining.

2. Related Work

The hot data mining method refers to a computer processing technique to extract valuable and effective information and knowledge from hot data, and it is an important branch of data mining. Unlike other big data analysis methods, it is generally used to deal with unstructured data. Traditional data analysis methods are difficult to be directly applied to the analysis of hotspots, and Chinese is more difficult to be analyzed and utilized because of its special characteristics. All along, the academic community has been dominated by scholars in the field of mathematics and computers in the research methods of this type of data. Topic detection and tracking technology can automatically identify new topics and keep track of known topics in news media information streams. It can effectively discover hot events generated within a period and is the theoretical basis for event mining. The hotspot clustering technology can perform cluster analysis on a large number of news hotspots on the Internet. Through unguided clustering of news hotspots, articles with similar hotspot content can be clustered together to form a preliminary hotspot event cluster. The headline generation technology can extract the core content of the topic, analyze the news hotspots within the topic, and use sentence compression technology to generate topic headlines, that is, hot events. However, with the development of the method and the needs of society, many scholars at home and abroad have made great progress in the research of hotspot mining in recent years, which has involved many fields such as management, medicine, politics, and finance.

Linguistic topic detection and tracking system were designed in the literature [4]. Literature [5] designed an incremental TF-IDF (term frequency-inverse document frequency) based topic event detection system, which was verified to work well. Literature [6] calculates the weight of words by selecting the log-likelihood test, by which it can handle documents with different languages, different sources, and different categories. Literature [7] analyzes hotspot mining and proposes a word frequency statistics method, which is effective in automatic classification operations, and this processing idea allows automated machine processing of hotspots. Literature [8] proposed probabilistic indexing methods and a probabilistic model based on automatic classification requirements. Literature [9] investigates vectorized representation, standardized processing, and classification methods for hotspots. Literature [10] gives a complete research framework for hotspot data from preprocessing to data analysis to results. Most of the previous research studies focus on algorithms and model effects, and although they do not delve into the hidden meanings of the data and their applications in various fields, they all promote the popularization and application of natural language processing methods, such as keyword extraction, coword analysis, and sentiment analysis well, and lay a solid foundation for further research to follow.

With the development of hotspot mining, hotspot clustering technology also began to develop rapidly.

Hotspots belong to unstructured data, and because of the special characteristics of Chinese, China's research on hotspot mining started relatively late, and so far there is no very mature method, mainly relying on the study of foreign theories, and itself is still in the process of exploration. However, the Chinese language is very profound, with multiple meanings and no space separating words from each other, and there are obvious differences with western languages such as English, so it is not possible to apply foreign methods directly. Literature [11] investigates hotspot splitting and proposes related splitting methods. Subsequent research in this technique also began to develop rapidly and other related disciplines where techniques were introduced. In recent years, the theory of hotspot mining in China has also been well developed. Literature [12] established a lexical-based feature selection method in this study, combining lexicality with TF-IDF. In the process of semantic graph structure description of hotspots, literature [13] applied semantic similarity matrix and performed such similarity calculation based on the corresponding maximum common subgraph and performed clustering analysis based on k-means algorithm, and the results showed that the accuracy of hotspot similarity degree was significantly improved in this processing mode, which can effectively meet the relevant application requirements. Literature [14] first selects the two most distant points in the data set as the initial clustering centers and then divides the other data points into the clusters closest to them until the total number of data points in the clusters reaches the set maximum value, calculates their centers of mass, obtains new cluster centers according to certain rules, and performs the above process cyclically to reduce the influence of the initial clustering center settings on the clustering results. Literature [15] improves the problem of the high computational complexity of the traditional K-means algorithm in dealing with massive data sets and speeds up the convergence of cluster centers by using a batch clustering method and updating the cluster centers using stochastic gradient descent. Literature [16] improved the traditional hierarchical clustering algorithm based on group average distance, solved the problem that its hierarchy could not be modified once it was determined, and improved the operation speed of the algorithm to some extent. In addition, literature [17] used how net and wordnet expand the semantics, which can achieve a greater improvement of clustering effect. The single-pass clustering algorithm in literature [18] is a typical representative of the incremental clustering algorithm, which has the advantages of simple principle and fast running speed and is often applied in online topic detection tasks, but the algorithm is influenced by the document input order, and different clustering results may be obtained due to the different document input order when dealing with the same document collection. To solve this problem, literature [19] introduces the concept of "generation" in the operation of the single-pass algorithm, inputting the document set in batches, clustering each batch of documents first, and then clustering the initial clustering results with the existing topic clusters, which effectively alleviates the order-sensitive problem of the single-pass algorithm's order-sensitive problem but makes the clustering results affected by the preliminary clustering process. In literature [20], based on the K-means clustering algorithm, the canopy algorithm is introduced to initialize the data, and the results of the algorithm are continuously updated by combining the hood center in the canopy algorithm and the class cluster center in the K-means algorithm, while the parallelized operation of the canopy-k-means algorithm is realized based on the Hadoop platform. The topic clustering results obtained in literature [21] based on this scheme are less affected by the input order of news data but still need to set the number of topics in advance, which is difficult to predict accurately in the complex Internet environment. In literature [22], improved single-pass algorithm was designed and implemented to make the operation results of the algorithm independent of the data processing order by introducing strategies such as double-pass clustering at the first clustering and adding a time slice setting at the center of the class cluster, while the stages of word separation, hotspot feature extraction, and topic discovery were optimized based on the Hadoop platform to improve the operation efficiency of the algorithm, respectively. However, in the massive hotspot processing task, the Hadoop platform still has certain shortcomings, as it needs to constantly read and write to the disk file system, which is lower than the memory-based spark platform in terms of processing efficiency and performance.

3. Hotspot Mining Algorithm Implementation for Digital Media in the Mobile Internet Environment

3.1. Hotspot Mining Algorithm. The implementation of hot topic mining and tracking firstly requires processing news into a digital form understandable by computer using hot feature representation, then realizing the division of topic clusters by the clustering algorithm, and finally displaying and tracking hot topics based on topic hotness evaluation method. This chapter investigates the related techniques in the above process, which mainly includes four aspects. (1) In this paper, we propose a hot topic feature representation method combining NE-LDA and woodstoves. (2) Secondly, we use a single-pass clustering algorithm. The single-pass clustering algorithm is used to achieve the discovery of news hot topics, and the parallelized implementation scheme of the single-pass algorithm is designed based on the spark platform. (3) The entropy weight method is introduced in the topic hotness assessment, and the topic hotness is objectively assessed based on three perspectives: time, media, and users. (4) Based on the results of hot topic mining and the location attributes of users, there is hot topic recommendation.

The main function of hotspot feature representation is to extract the features in each news report and convert them into a digital form that can be understood by computers, which is the basic work in the hot topic mining and tracking task and has an important impact on the subsequent processing process such as dividing topic clusters [23]. Commonly used hotspot feature representation methods include the LDA topic model and woodstoves word vector model, which focus on different aspects of hotspot features,

respectively. LDA portrays the topic of a hotspot, while word2vec focuses on describing the semantic information of hotspot. However, in the news, there may be multiple reports on the same events appearing in different locations, when the LDA descriptions of their topics will be very similar, resulting in their being classified in the same topic. To solve the problem, currently, academia and industry unify keywords such as locations and people in hotspots as named entities and combine named entity recognition technology with LDA technology to build NE-LDA models, which can effectively improve the topic recognition performance. Although the NE-LDA model can improve news recognition performance, the semantic information among hotspot contexts is still ignored in the NE-LDA model, so this paper proposes a model that fuses NE-LDA and word2vec for a comprehensive feature representation of news reports [24]. Based on the above analysis, this section first introduces the basic principles of the LDA topic model, named entity recognition technique and word2vec model, and finally provides a detailed description of the fused NE-LDA and word2vec approach used in this paper.

The LDA topic model is mainly based on the Bayesian principle to model the topic information described by hotspots. The main function of hot feature representation is to extract the features of each news report and convert them into a digital form that the computer can understand. It is the basic work in the task of hot topic mining and tracking and has an important impact on the subsequent process of dividing topic clusters. In practical application, the topic distribution of each article and the word distribution under each topic are calculated based on the input document collection and the given number of topics, and its specific operation process and principle are elaborated next by the LDA probability graph model shown in Figure 1.

After decades of development, the discipline of hotspot mining has been evolving and evolving day by day. The types of hotspots it deals with are getting richer and richer, the technologies used are rapidly changing, and the application scenarios it implements are expanding. It can be summarized as the following characteristics. As an application-driven field, hotspot mining incorporates a large number of technologies from multiple fields, and Figure 2 shows examples of disciplines that have had a significant impact on the development of hotspot mining. This characteristic dictates that it is not practical to discuss hotspot mining in isolation from its closely related disciplines, either in theoretical research or in engineering applications. A specific task often requires a clever combination of different techniques; for example, to mine natural language hotspot data, it is more popular to combine hotspot mining with web crawlers and natural language processing techniques, as detailed in the thesis for specific applications. Hotspot mining is the integration and application of various disciplines with the goal of knowledge discovery and can be significantly enhanced by integrating new methods from multiple disciplines. See Figure 2.

Based on the training results of the LDA topic model, the topic distribution of all documents and the word distribution of several topics can be obtained. Thus, the document-topic matrix and the topic-word matrix can be obtained, and the topic information can be summarized and analyzed by these two matrices. The LDA document-topic distribution is the Dirichlet distribution, as shown in (1); that is, for any document, its topic distribution γ_d is as follows, where β is the hyperparameter of the distribution, which is set as the default value in this experiment because there is no more a priori information, and it is an n-dimensional vector, and n represents the predetermined number of topics K, which is also the hyperparameter of the model:

$$\gamma_d = \int \beta^{\mathrm{Dirichlet}(\overrightarrow{m})} \cdot \eta^r \, \mathrm{d}m. \tag{1}$$

The evolution of topic strength can be indicated based on the heat results of the segmentation model after time slicing. There are two general ways to measure the hotness: the first one is based on the number of documents under the topic, but this method is crude, and the other one is based on the probability value generated by the LDA model. In this paper, we choose the average probability calculation method based on the characteristics of journal abstracts and the relationship between the number of topics and the sample size, combined with the threshold setting method of probability difference value, which gives the probability value of a document belonging to several topics. In this way, the average probability calculation method can clearly show the intensity evolution of the topic. The specific intensity calculation formula is shown in the following equation:

$$p(\theta) = \frac{\mathrm{np}!}{\sum r!(n-r)!}. \tag{2}$$

KL scatter is one of the common measures of similarity. The measure of KL scatters distance is given by the following equation:

$$\mathrm{KL}(x, y) = \frac{\delta y}{\delta x} \cdot \sum_{i=1}^{n} X_i Y_i + C_x + C_y. \tag{3}$$

The formula represents the difference between these 2 topics on set V. If this difference is smaller, then these two topics are more similar. However, the measure of this similarity should be symmetric, and the KL difference distance has asymmetry, so the measure of similarity can be performed by using the JS distance. The formulas are as follows:

$$\mathrm{sigma}(\alpha, \beta) = \sigma_\alpha^2 \cdot \mu_\beta \sum_{i,j=1}^{n} \alpha_{ij} \beta_{ij}, \tag{4}$$

$$\overline{Y} = \sum_{i,j=1}^{n} \left[(x_i - \overline{x})^2 \cdot \lambda \sqrt{x^2 - \alpha\beta}. \right. \tag{5}$$

The data results under different period dimensions are subjected to similarity measure and matching, and the probability changes of the same hot topic words can reflect

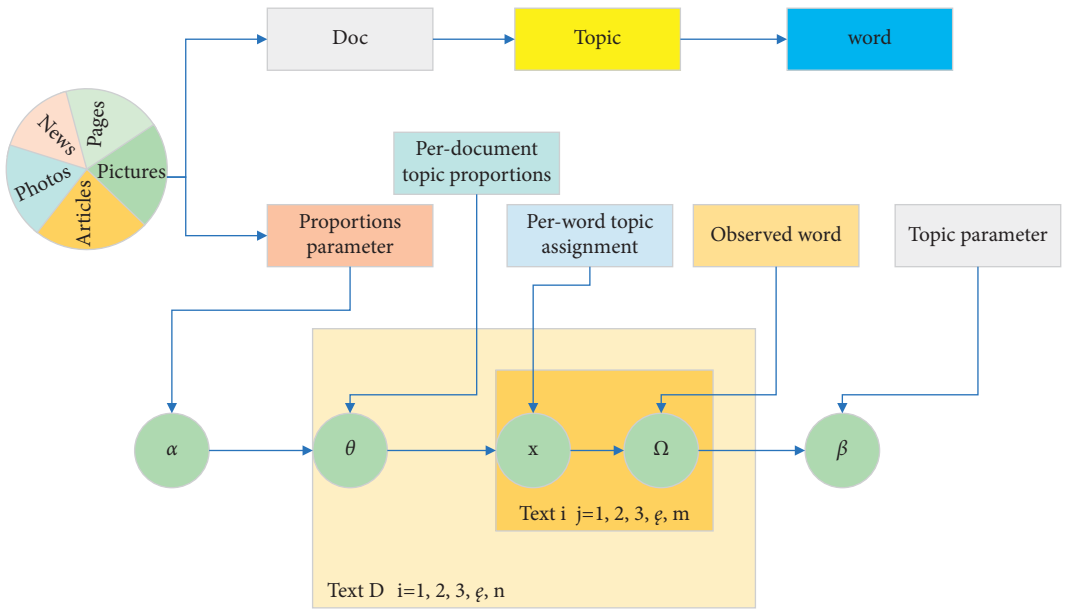

Figure 1: LDA probability model.

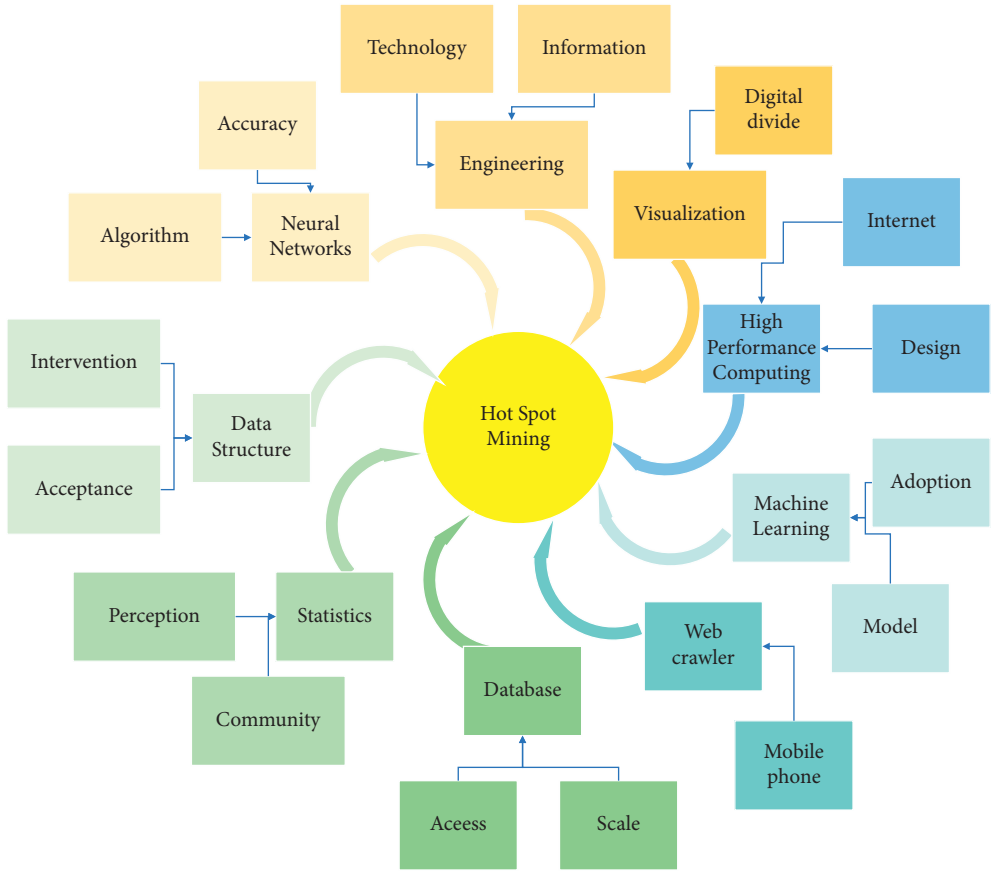

Figure 2: Hotspot mining draws technology from various fields.

the content evolution trend of the topic, and the results calculated according to the topic intensity can reflect the intensity relationship of the same topic in different time regions [25]. The algorithm in this paper makes full use of the advantages of the shortest path algorithm based on word graphs and uses reasonable point weights and edge weights to score. The final cluster search also implements dynamic scoring of the end nodes. Therefore, it has both information

volume and language coherence. All aspects performed well and scored the highest. The analysis process of topic evolution analysis is shown in Figure 3, and the data selection by sliding time window for segmentation experiment is combined with a similarity measure for topic alignment. See Figure 3.

3.2. Implementation of Digital Media Hotspot Mining Algorithm in Mobile Internet Environment. Internet group communication is a special form around content, channels, and structure in cyberspace, a kind of communication behavior based on meaning production and information gathering, which not only demonstrates unique formation conditions and diffusion process but also has the basic characteristics of Internet communication. There are three main reasons for the formation of Internet group communication. ① The first reason is social causes. Sociological theory shows that the imbalance of social structure is the primary cause of Internet group communication. When these social contradictions accumulate to a certain level and there is no place to disseminate them, social media becomes a public platform for these contradictions and conflicts to be expressed and ventilated. Thus, the behavioral effect level of Internet group communication is filled with a large number of intertwined and conflicting social contradictions and hidden problems. ② The second reason is psychological causes. Social psychology research shows that overall social satisfaction is an intrinsic cause of Internet group communication. The "Social Psychological Map" developed by the Institute of Psychology of the Chinese Academy of Sciences includes life satisfaction (LS), income satisfaction (IS), social status satisfaction (SPS), local economic satisfaction (LES), national economic satisfaction (NES), and social justice satisfaction (SJS). With the downward shift of the center of gravity of the Internet application, a large number of youth groups are hiding the psychological state of "small loss" and the Internet participation of the disadvantaged groups and the underprivileged society brought about by the imbalance of social structure, and the sense of relative deprivation has stimulated the generation of mass incidents. ③ The third reason is technical causes. The essence of clustering is the process of dividing samples into different categories according to the degree of similarity between the sample features in the data set, and it is required that the similarity between samples in the same cluster should be as large as possible, and the similarity of samples in different categories should be as large as possible. It may be small. The development and maturity of mobile Internet technology will provide model innovation for Internet group communication at the basic level and become the technological motive for shifting Internet communication from "individual fragmentation" to "group circling." The "mimetic environment" formed by social media is an important field for people to break free from social norms, present themselves, and seek group identity on the one hand, and various elements of the real society are amplified and fermented by the network on the other. On the other hand, various elements of real society are amplified and fermented by the Internet. Therefore, although Internet technology brings technical advantages to the dissemination of group information, it also brings the possibility of the formation and proliferation of negative information, such as social conflicts, civil pressure, and group polarization.

The change of topic hotness is the topic hotness in different time slices; this paper introduces the concept of "topic index" proposed by literature [26] and uses "topic index" to express the process of change of hotness of this topic, to express more graphically the change of topic, and to express the change of topic T with time; the line graph of the change of topic index with time is depicted. From the trend graph of the topic index, you can see how long the process of topic T has gone from generation to climax, and you can understand the current development status of the topic, and so on. The process of constructing a line graph of topic index changes is shown in Figure 4.

With the development of events and feedback from Internet users, a certain news topic will extend many related topics, that is, the change of topic content, so this section focuses on the correlation relationship of each subtopic within different time slices. There may be correlations between topics on different time slices, and it is also the correlation changes between these subtopics that make up the whole life cycle of the whole topic development. The subtopics on each time slice are mined by using the composite model proposed above, and the subtopics are represented as weighted word vectors. Mining the correlations between these subtopics usually requires the use of a similarity measure, which is calculated based on the similarity between the subtopic and the subtopics preceding and following it within a certain time frame, to track the development of the hot topic throughout its lifecycle. Each hot topic has the following four stages. The first stage is germination stage: the emergence of a topic, that is, a topic has just been created and has not yet been reported extensively. The second stage is spreading stage: the topic is noticed, and as the online media reports, netizens gradually pay attention to the topic, which is the climbing period for the topic to become a hot topic. The third stage is climax stage: the peak of the topic, with the information released by various informants, or an important event of the topic. After the climax stage, the topic will drop from the peak because of the emergence of new topics, but the heat of this stage is still very high. The fourth stage is the decline stage: the topic will show a decreasing trend over time until it disappears. In the whole life cycle of hot topics, along with the development process of topic generation, spread, expansion, sublimation, and extinction, the development stages and changes of topics are analyzed.

4. Experimental Design and Validation

The experiments were conducted on 20 websites with news content published within a week as the test corpus, and 500 news articles were used as labeled data. The experiments include two aspects of algorithm complexity and model performance, in which the algorithm complexity test mainly measures the time consumption of the proposed label-vec clustering model and

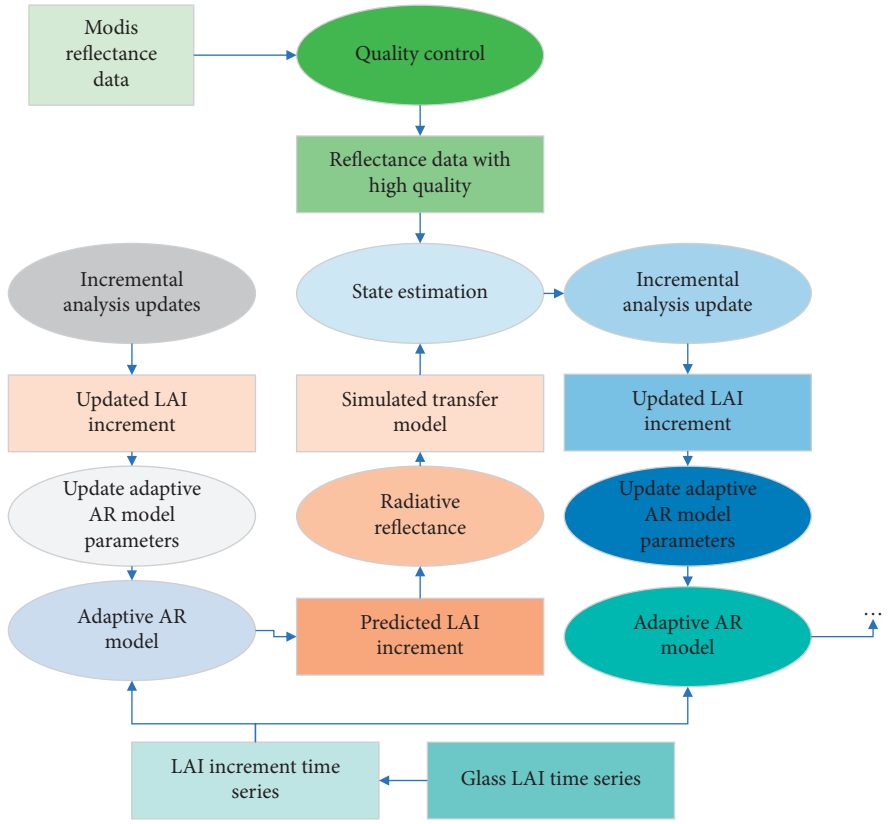

Figure 3: Hotspot evolution analysis process.

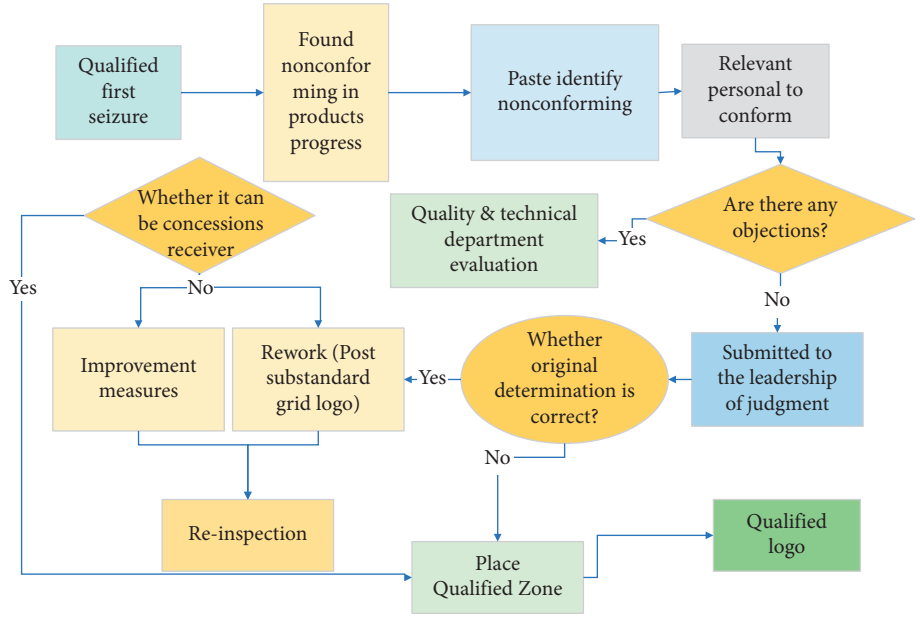

Figure 4: The process of constructing a topic hot index graph.

baseline under different sizes of the corpus. The algorithm complexity test measures the time consumption of the proposed label-vec clustering model and baseline under the different sizes of the corpus. The performance test includes two parts: intracluster metrics and intercluster metrics.

The test corpus consisted of 500 pieces of news content in one week, divided into ten test corpora of increasing size. The results of the experimental comparison are shown in Figure 5.

The time consumption of the label-vec algorithm is significantly better than that of the k-means and single-pass

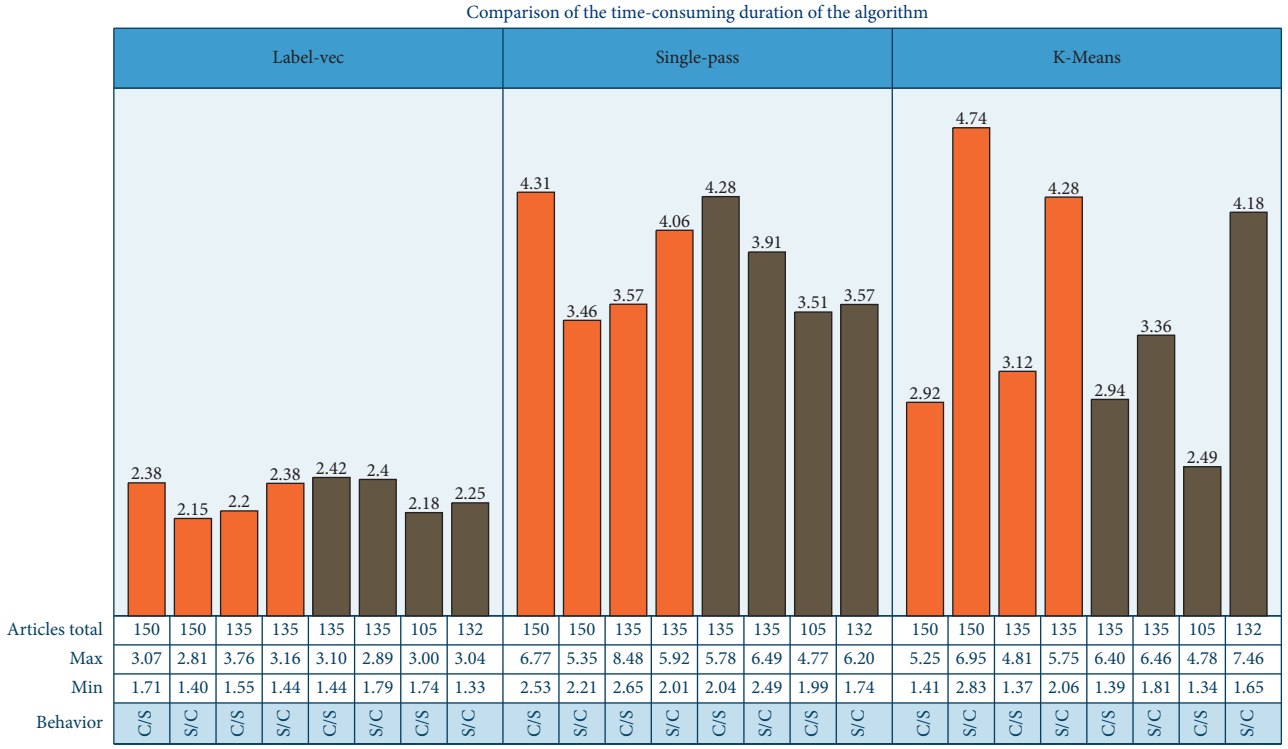

Figure 5: Comparison of the time-consuming duration of the algorithm.

algorithms. Moreover, the growth rate of label-vec time consumption is relatively slower while the corpus increases.

As seen in Figure 5, the time consumption of the model is also not linearly increasing, since for each sample, calculating whether it is a core sample requires calculating the distance for all unvisited samples in the bucket where it is located, and as the number of samples increases, the average number of samples in the bucket also increases.

The model performance includes two parts: external and internal indexes. The external metrics include Jaccard similarity, FM index, and Rand index; the internal metrics include DBI index and Dunn index. The performance analysis test corpus is a total of 8000 news data from news labs. Among them, 1000 pieces of tagged data are used for labeled indicators.

Figure 6 tests the external performance metrics of the label-vec algorithm and uses the single-pass and k-means algorithms as references. Compared with the single-pass and K-means clustering algorithms, the external performance index of the label-vec algorithm is better, especially that the FM index and RI index are significantly better than the single-pass and K-means, indicating that the label-vec algorithm can handle the intercluster clustering more effectively in the news hotspot clustering. It shows that the label-vec algorithm can handle the intercluster relationship more effectively and identify the outliers and noise data more easily. See Figure 6.

Figure 7 tests the internal performance metrics of the label-vec algorithm and uses the single-pass algorithm and k-means as references. The internal performance metrics of the label-vec algorithm are better than the single-pass

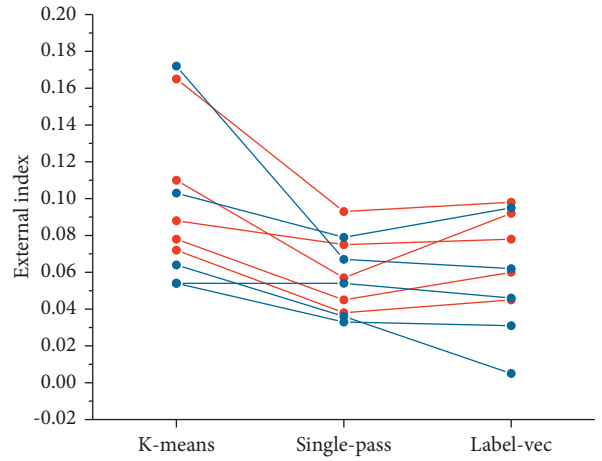

Figure 6: Comparison of external index metrics of different algorithms.

algorithm and k-means algorithm, especially the information entropy is significantly better, which indicates that the label-vec algorithm can handle the cluster shapes corresponding to news hotspots more effectively by density clustering. Sociological theoretical research shows that the imbalance of social structure is the primary motivation for the spread of Internet groups. The current society is in the process of profound changes. Due to the incoordination of the social structure, social groups are often in a state of opposition, contradiction, or conflict. When these social contradictions have accumulated to a certain extent, they have nowhere to spread. Social media has become a public

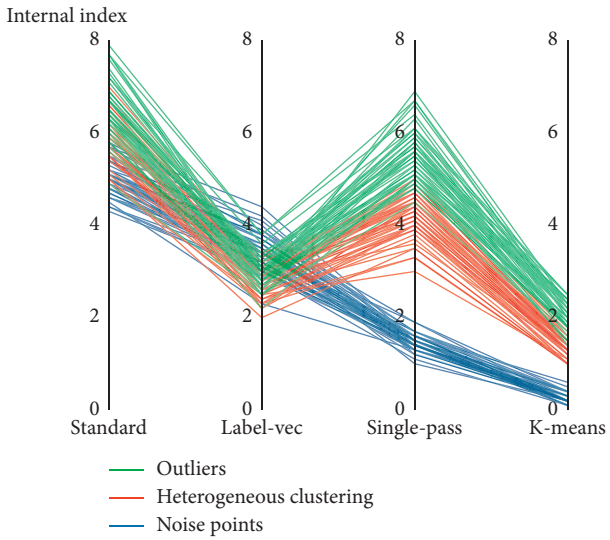

Figure 7: Comparison of internal index metrics of different algorithms.

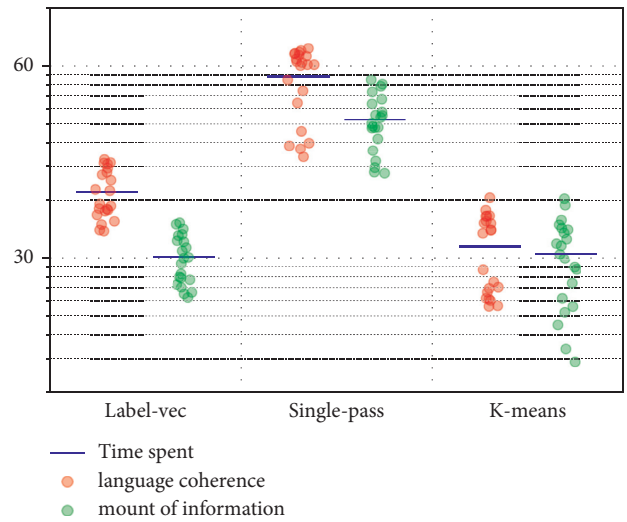

Figure 8: Comparison of the amount of information and language coherence detected by different algorithms.

platform through which these contradictions and conflicts can be expressed and vented. Therefore, the behavior effect level of Internet group communication is full of a large number of social contradictions and hidden problems that are intertwined and intertwined. This indicates that the label-vec algorithm can handle the cluster shapes corresponding to news hotspots more effectively through density clustering. From the above performance tests, it can be seen that the external and internal performance indexes of the label-vec clustering model are better than those of the K-means and single-pass models. The reason is that compared with the K-means algorithm, the label-vec algorithm does not need to decide the number of cluster centers in advance to prevent the problem of local optimum, and it has better performance in dealing with heterogeneous clusters and outliers. It also has better performance in handling heterogeneous clusters, outliers, and noise points. Compared with the single-pass algorithm, the similarity calculation model of label-vec takes into account multidimensional information, such as words and semantics, and the accuracy rate is higher. See Figure 7.

From Figure 8, we can see that the algorithm in this paper outperforms the single-pass and K-means algorithms in terms of information content and linguistic coherence. We analyzed the 98 event topic sentences generated by the three algorithms and found that the baseline method often produced "off-topic" event topic sentences; that is, the generated event topic sentences did not represent the topic of the event, but the topic sentences performed better in terms of linguistic coherence compared with the information content score. The reason for the low accuracy of the baseline algorithm is that the algorithm uses the conditional probabilities of the words provided by the language model to obtain the highest scoring sequence of words, which counts the conditional probabilities between words; that is, the higher the probability is, the easier it is for two words to be used together, so the linguistic coherence is guaranteed. However, the algorithm's consideration of information content is only helpful when the set of candidate sentences is generated, so the resulting event topic sentences have low information content scores and average linguistic coherence scores. The shortest path algorithm based on word graph, with predefined sentence length, leads to the phenomenon of "truncation"; that is, the generated sentences may be incomplete, so that the information content can be guaranteed to a certain extent, but the sentence coherence score is not high because the sentences are truncated. The algorithm in this paper takes advantage of the shortest path algorithm based on word graph and uses reasonable point weights and edge weights for scoring and also implements dynamic scoring for the end nodes in the final cluster search, so it performs well in both information content and linguistic coherence and has the highest score. However, the algorithm in this paper performs worse than the word graph-based shortest path generation algorithm in terms of sentence compressibility. See Figure 8.

For mining effective hotspot information through certain technical means for a large amount of network hotspot data, this paper proposes a research framework for digital media hotspot mining algorithm in the mobile Internet environment based on the current research status, research means, and shortcomings in the research process under multifaceted research study. Hot topic mining does not have a unified evaluation standard to judge the results. The concept of weights is incorporated into the traditional language model to train the word vectors, enhance the feature representation of important words in hotspots, facilitate the subsequent effective extraction of valuable information in feature extraction, and combine convolutional neural networks to realize the feature extraction of contextual depth semantics and explore the influence of different parameters on the experimental results by adjusting the parameters in the experiments. The experimental results show that the parameter selection of the convolutional kernel has some influence on the F-value of hotspot analysis,

and there are some differences in the parameter selection for different data sets.

The algorithm combines the advantages of the K-means algorithm and single-pass algorithm and proposes the clustering algorithm label-vec select points with the maximum density in the moving range, which realizes the optimization of the K-means algorithm at the point selection and selects the denser points as the initial target points of K-means by moving autonomously according to the density value. In the algorithm design, different from single pass, this paper proposes the clustering method of moving the concept of ring range to select points to speed up the traversal and reduce the computational effort. The experimental results show that the label-vec algorithm with the maximum density selection of points in a certain moving range has a better overall performance in terms of clustering effect and running time.

At different data set sizes, the speedup ratio increases as the number of computational nodes increases, because the increase in the number of nodes improves the operational efficiency of the parallel algorithm. At the same time, the larger the dataset size is, the more obvious the increase of the speedup ratio is. In the relatively small data size, the trend of the speedup ratio increases gradually with the increase of the number of nodes, because under the large-scale dataset, increasing the number of computation nodes can effectively share the computation volume and thus improve the execution efficiency of the algorithm, but when the dataset size is small, too many computation nodes will lead to excessive communication cost and scheduling overhead between nodes, which makes the efficiency improvement of the algorithm not obvious.

5. Conclusion

In the context of the era of information technology and big data, a large amount of unstructured descriptive information is hidden behind the Internet. Hotspot mining technology mainly extracts and mines unknown information from a large number of original unprocessed documents, which allows users to quickly obtain effective information in a large amount of cluttered information, make accurate judgments and processing for related problems, and even prevent from the future by processing in advance according to the information mined. With the further maturity of hotspot mining technology, the development and wide application of hotspot mining technology are an inevitable trend in the future, and this technology will be more and more widely used in various fields of scientific research, society, and life. In this paper, we focus on two aspects of hotspot clustering and topic extraction. First, we introduce the main theoretical knowledge of hotspot mining in detail, and after understanding hotspot mining, we learn more about cluster analysis and topic extraction. Further, based on theoretical research, this paper crawls the articles on the Internet through web crawler technology and analyzes the crawled articles through *R* language and conducts an in-depth analysis of the keywords and concludes that the focus of group attention is concentrated in a certain aspect. Analyze the principle of the agglomerated hierarchical clustering algorithm, improve the shortcomings of repeated calculation of the agglomerated hierarchical clustering algorithm based on single link method calculation, introduce triple storage, and propose an improved agglomerated hierarchical clustering algorithm. From the experimental results, it can be seen that the improved algorithm can reduce the running time of the algorithm and improve the efficiency. Finally, the processed hotspots were feature selected and represented as a document word matrix for subsequent analysis. For clustering and topic extraction, the K-means algorithm was used for clustering, and the TF-IDF model was used for topic extraction of articles, and hotspots were classified into five categories and then classified for topic extraction to better study the content of hotspots.

In this paper, the background and significance of hot topic mining are explained, the current status of domestic and international research on hot topic mining is studied and analyzed in detail, and the advantages and disadvantages of the existing algorithms are summarized; text modeling is an important step in news topic mining; to improve the correctness of news topic mining, the text is introduced into word2vector model to train word vectors containing contextual semantics, and the weighted word vector algorithm is proposed. To improve the correctness of news topic mining, the text introduces the word2vector model to train word vectors containing contextual semantics and proposes a weighted word vector algorithm that combines the word weights calculated by TF-IDF algorithm and word vectors, analyzes the principle of the cohesive hierarchical clustering algorithm, improves the disadvantages of repeated calculations of cohesive hierarchical clustering algorithm based on the single link method, and introduces triadic storage to propose an improved cohesive hierarchical clustering algorithm. Because of the shortcomings of the improved cohesive hierarchical clustering algorithm that cannot change the results and the K-means algorithm that randomly selects the initial clustering centers, this paper proposes a composite model clustering algorithm, which combines the two algorithms. The improved cohesive hierarchical clustering algorithm can provide the number of news topics and initial clustering centers to the K-means algorithm, and the K-means algorithm can compensate for the shortcomings of the improved cohesive hierarchical clustering algorithm. In terms of hot topics mining, this paper carefully analyzes the characteristics of news web pages, improves the traditional TF-PDF algorithm of hotness evaluation, and introduces user engagement including the number of reads and comments. To understand the whole process of hot topics, this paper introduces "topic index" and proposes a time-slice-based topic mining method, which treats a day as a time slice and analyzes the daily hot topics to understand the changing status of old hot topics and dig out new ones. By analyzing the daily hot topics, we can understand the changing status of the old hot topics and explore the new hot topics.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that no conflicts of interest exist concerning this study.

References

[1] Y. Jin and X. Li, "Visualizing the hotspots and emerging trends of multimedia big data through scientometrics," *Multimedia Tools and Applications*, vol. 78, no. 2, pp. 1289–1313, 2019.

[2] T. Liu, F. Xue, J. Sun, and X. Sun, "A survey of event analysis and mining from social multimedia," *Multimedia Tools and Applications*, vol. 79, no. 45, pp. 33431–33448, 2020.

[3] Y. Chen, "Mining of instant messaging data in the Internet of Things based on support vector machine," *Computer Communications*, vol. 154, pp. 278–287, 2020.

[4] W. W, X. Xia, M. Wozniak, X. Fan, R. Damaševičius, and Y. Li, "Multi-sink distributed power control algorithm for cyber-physical-systems in coal mine tunnels," *Computer Networks*, vol. 161, pp. 210–219, 2019.

[5] T. Pei, C. Song, S. Guo et al., "Big geodata mining: objective, connotations and research issues," *Journal of Geographical Sciences*, vol. 30, no. 2, pp. 251–266, 2020.

[6] J. Wang, Y. Fan, H. Zhang, and L. Feng, "Technology hotspot tracking: topic discovery and evolution of China's blockchain patents based on a dynamic LDA model," *Symmetry*, vol. 13, no. 3, p. 415, 2021.

[7] X. Ding, M. Zheng, and X. Zheng, "The application of genetic algorithm in land use optimization research: a review," *Land*, vol. 10, no. 5, p. 526, 2021.

[8] J. Gao, X. G. Yue, L. Hao, C. James, and M. Crabbe, "Optimization analysis and implementation of online wisdom teaching mode in cloud classroom based on data mining and processing," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 1, pp. 205–218, 2021.

[9] T. Chen, L. Peng, J. Yang, and G. Cong, "Analysis of user needs on downloading behavior of English vocabulary APPs based on data mining for online comments," *Mathematics*, vol. 9, no. 12, p. 1341, 2021.

[10] R. Talat, M. S. Obaidat, M. Muzammal, A. H. Sodhro, Z. Luo, and S. Pirbhulal, "A decentralised approach to privacy preserving trajectory mining," *Future Generation Computer Systems*, vol. 102, pp. 382–392, 2020.

[11] Y. Deng, C. Li, and D. Wang, "An integrated approach for knowledge management in the context of product innovation," *Cluster Computing*, vol. 22, no. 4, pp. 9385–9396, 2019.

[12] H. Su, Q. Liu, and C. Mu, "Research on product reviews hot spot discovery algorithm based on mapreduce," *IEEE Access*, vol. 8, pp. 111829–111836, 2020.

[13] R. Miao, Y. Wang, and S. Li, "Analyzing urban spatial patterns and functional zones using sina weibo POI data: a case study of Beijing," *Sustainability*, vol. 13, no. 2, p. 647, 2021.

[14] X. Jiang, H. Ding, H. Shi, and C. Li, "Novel QoS optimization paradigm for IoT systems with fuzzy logic and visual information mining integration," *Neural Computing & Applications*, vol. 32, no. 21, pp. 16427–16443, 2020.

[15] C. Sun, "Research on investment decision-making model from the perspective of "Internet of Things + Big data"," *Future Generation Computer Systems*, vol. 107, pp. 286–292, 2020.

[16] I. A. Ajah and H. F. Nweke, "Big data and business analytics: trends, platforms, success factors and applications," *Big Data and Cognitive Computing*, vol. 3, no. 2, p. 32, 2019.

[17] X. Zhou, W. Liang, I. Kevin, and Y. Laurence, "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 171–178, 2020.

[18] B. J. Galli, "Application of statistical analysis tools and concepts to big data and predictive analytics to new product development," *International Journal of Strategic Engineering*, vol. 3, no. 1, pp. 17–35, 2020.

[19] G. Dong, B. Li, X. Wei, and T. Qin, "Mining key users of microblog topics based on trust model," *International Journal of Performability Engineering*, vol. 15, no. 11, p. 3024, 2019.

[20] Q. Xu and M. Li, "A new cluster computing technique for social media data analysis," *Cluster Computing*, vol. 22, no. 2, pp. 2731–2738, 2019.

[21] J. Lian, "Implementation of computer network user behavior forensic analysis system based on speech data system log," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 559–567, 2020.

[22] S. Köhler and M. Pizzol, "Life cycle assessment of bitcoin mining," *Environmental Science & Technology*, vol. 53, no. 23, pp. 13598–13606, 2019.

[23] M. Wang, "FollowMe: a mobile crowd sensing platform for spatial-temporal data sharing," *International Journal of High Performance Computing and Networking*, vol. 14, no. 4, pp. 416–424, 2019.

[24] D. Li, Z. Cai, L. Deng, X. Yao, and H. Wang, "Information security model of block chain based on intrusion sensing in the IoT environment," *Cluster Computing*, vol. 22, no. 1, pp. 451–468, 2019.

[25] R. Zhu, H. Ye, H. Sun, X. Li, Y. Duan, and J. Hou, "Construction and application of knowledge-base in telecom fraud domain," *International Journal of Intelligent Information and Database Systems*, vol. 14, no. 2, pp. 198–214, 2021.

[26] F. Al-Turjman, "5G-enabled devices and smart-spaces in social-IoT: an overview," *Future Generation Computer Systems*, vol. 92, pp. 732–744, 2019.