Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3632679, 12 pages
https://doi.org/10.1155/2022/3632679

Research Article

Internet Rumor Audience Response Prediction Algorithm Based on Machine Learning in Big Data Environment

Suhong Yang[1], **Shenghui Wang**[2] **and Y. Yiwen**[3]

[1]*Hangzhou Dianzi University Information Engineering College, Hangzhou, 310000 Zhejiang, China*
[2]*Shaanxi Jiatong Electronic Technology Co., LTD, Xi'an, 710075 Shaanxi, China*
[3]*Business School, University of Strathclyde, G1 1XQ, Glasgow, UK*

Correspondence should be addressed to Shenghui Wang; kk743@wfd.edu.ug

Received 20 January 2022; Revised 13 March 2022; Accepted 5 April 2022; Published 30 April 2022

Academic Editor: Mohamed Elhoseny

Rumors are an important factor affecting social stability in some special times. Therefore, the dissemination and prevention and control mechanisms of rumors have always been issues of concern to the academic community and have long been highly valued and widely discussed by experts and scholars. However, in combination with the Internet as a new type of media, although people have begun to pay attention to online rumors, research on it is still relatively fragmented, especially in the cross-domain research specific to the social influence of online rumors, and there is no clear indication of online rumors. The specific definition also did not analyze in detail the internal connection between its influence and group behavior. Therefore, this article will combine actual cases to explore and analyze the spread and influence process of online rumors and show its social influence, hoping to enrich the research of online rumors. Nowadays, the Internet has become the most important carrier to reflect the public grievances. Internet users have expressed their opinions on hot issues such as enterprises, people's livelihood, and government management, which has formed a powerful public opinion pressure, which has far exceeded the traditional media. The hidden dangers of security cannot be ignored. Therefore, how to monitor network public opinion from a large amount of network data is a difficult problem that needs to be solved urgently. Firstly, this consists of four modules: information collection, web page preprocessing, public opinion analysis, and public information report. Secondly, text clustering, the core technology of network public opinion, is optimized, and single-pass algorithm based on double threshold is proposed. Then the dual-threshold single-pass algorithm is optimized based on the MapReduce parallel computing model, and finally a network public opinion collection technology is formed under the background of big data. Simulation results can greatly improve the performance of text clustering and can effectively optimize the design using the parallel computing model based on MapReduce. The average miss rate after optimization is 0.7569 times, the average false alarm rate is 0.5556 times, and C_{det} is 0.5714 times. It proves that the collection technology based on machine learning under the background of big data is effective and has good performance.

1. Introduction

Internet has a positive impact on the economy, politics, culture and people's lives. At the same time, it should be noted that bad and false information still exists on the Internet, which affects the healthy development of society. Rumors caused panic among the people, endangered public safety, and harmed public interests. Due to the convenience and development of the Internet, rumors spread quickly and spread widely, which will cause adverse effects. Rumors disturb people's thinking, psychology, and behavior, destroying the social trust system. In particular, the spread of Internet rumors has become a major social nuisance, severely infringing on citizens' rights and interests, harming public interests, and endangering. The rapid development of Internet technology, on the one hand, is reflected in the Internet technology and, on the other hand, in the development of software and hardware, such as a variety of APP, 3G network upgrade to 4G network. The rapid development of technology has not only facilitated people's lifestyle and enriched people's participation in society, but also profoundly influenced the behavior of each subject in society. By December 2013, the

netizens reached 618 million. In 2013 alone, the netizens reached 53.58 million. The development of network society has expanded the channels for people to participate in society, express their feelings and opinions, and greatly stimulated the people's willingness to express their opinions.

Rumors have existed for a long time, and the limitations of traditional media have determined that traditional rumors are confined in a relatively narrow range, and their influence is far less than today's online rumors. The most direct cause of the spread of rumors is the lack of real information in society. Why is it missing? Because the flow of information is hindered. At the level of public life, because information is relatively more open and transparent and circulates more quickly, people can compare information from all parties, and the space for rumors to survive is much narrower. Content control is the most difficult problem. Anyone can move anything to the Internet. This kind of thing often happens. The Internet poses some rare problems for anyone who wants to clear this trash on the information superhighway. Internet rumors are undoubtedly one of the important issues. When traditional rumors enter the Internet, rumors are like "winged on." They are likely to appear first on the Internet as a "hotbed" and cause rumors to spread. They appear on the silver screen in a radial manner. Either personal curiosity, or hype by a certain person or organization, or simple unintentional action, or purposeful misleading, no matter what mentality, once rumors go through the Internet, it may cause the spread of rumors to increase exponentially. The development of self-media represented by mobile phones has greatly facilitated and speeded up the means by which people express their opinions. By December 2013, the number of mobile phone users has reached 500 million, and the number of users using mobile phones to log on to the Internet has accounted for 73.3% of the new Internet users in China, much more than the proportion using other terminals. This means that the rapid popularization of mobile terminals has promoted the increase of Internet users in China.

The Internet from 2G network to 3G network and from 3G network to 4G network, the combination of massive mobile terminals, followed by the generation of massive data. China produced more than 0.8 ZB (800 million TB) of data in 2013, twice as much as in 2012, equivalent to the total global data in 2009.

It is based on the Internet, with the expression of the people's opinions and the behavior of the government as the manager as the basic content, with the government, netizens, and network media as the three main participants, around the development process of specific events in society, and the process of mutual game. Internet technology, the popularity of mobile terminals and the improvement of people's awareness of participating in hot social events, the traditional channels of expression of public opinion, and the network has gradually become the most important position for people to express their emotions.

In this process of development, due to the rapid spread of the Internet and the combination of online or wireless communication tools such as forums, blogs, Weibo, WeChat, and self-media, the traffic information data has become diversified, and it has become more and more difficult for the government to control and monitor public opinion. In the process of the beginning, fermentation and outbreak of social public opinion, the network plays a great role, as one of the management contents of the government.

Under the background of the big data era, the frequent innovation and application of the emerging network media, especially the popular new media, has created a completely different network public opinion environment, and it has significant characteristics different from any previous period. In such a brand-new network public opinion environment, the era when the dominant power of discourse is in the hands of a certain party or organization has become a history. Every individual member can make his own voice anytime and anywhere by means of the emerging network media or receive and understand the outside world actively and passively. However, it is in such an "open and partially closed" environment that most individual members are at the bottom of the information chain, unable to identify the true and false information, or passively accept a certain point of view, and become "followers" of certain public opinion. It is precisely this reason that exacerbates the uncertainty and uncontrollability of network public opinion, and in reality, people tend to use the network to express all kinds of contradictions and emotions; once become the focus of network hot discussion, it will form a storm of public opinion. If handled improperly, it is likely to form a crisis of network public opinion, which is not only harmful to the stability of social order, but also has a profound impact on the government's social governance and credibility.

Machine learning (ML) is also the core research field of AI. Machine learning is a highly interdisciplinary research field. Machine learning algorithms can be used to solve computer vision, biology, robotics, data analysis, natural language processing, and other fields. Machine learning has been, is, and will bring great impetus to many disciplines. Linear regression is terrible at dealing with nonlinear relationships, inflexible in identifying complex patterns, and extremely tricky and time-consuming to add the correct interaction terms or polynomials.

In short, machine learning is a science that studies, the acquisition of knowledge or skills, and the improvement of performance. ML science first appeared in the 1950s. It is a subject about computer, data construction, model, and simulation of human activities. Its application has covered all aspects of our lives, for example, robotic chess programs, speech recognition, unmanned driving, robots, and other fields. In addition, machine learning theory and methods have also played a huge role in this field. This paper mainly includes the introduction, related work, related methods, experiments discussion, and conclusions.

2. Related Work

It is generally believed that machine learning is a kind of knowledge that studies how computers acquire new knowledge and skills through experience and discover existing knowledge [1–3]. A common definition is to measure the performance of a computer program to accomplish learning

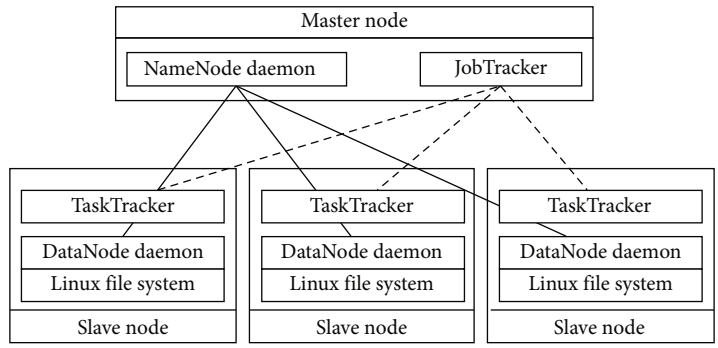

Figure 1: Basic structure of Hadoop.

Table 1: Topic quantity table.

Name	Michel's visit to China	Malaysia airlines plane missing	Poisoning in kindergarten	Balance treasure income	Coal mine accident	Other	In total
Number	178	184	135	88	123	92	800

Table 2: Double-threshold single-pass algorithm for processing result data.

Statistical unit (unit)	Michel's visit to China	Malaysia airlines plane missing	Poisoning in kindergarten	Balance treasure income	Coal mine accident
Correct documentation in cluster	145	148	121	78	109
p_{miss}	0.1713	0.2134	0.0916	0.1063	0.2339
p_{false}	0.0032	0	0.0015	0.0014	0.0029
Mean of p_{miss}			0.1633		
Mean of p_{false}			0.0018		
C_{det}			0.0021		

Table 3: Single-pass algorithm processes result data.

Statistical unit (unit)	Michel's visit to China	Malaysia airlines plane missing	Poisoning in kindergarten	Balance treasure income	Coal mine accident
Correct documentation in cluster	130	128	107	58	83
p_{miss}	0.3122	0.4207	0.2916	0.3211	0.4279
p_{false}	0.0071	0.0069	0.0039	0.0044	0.0042
Mean of p_{miss}			0.3547		
Mean of p_{false}			0.0053		
C_{det}			0.0059		

in experience (E), which requires the program to improve the performance of a tasks (T), and whether the performance of this program can be measured by performance (P).

Since its emergence, machine learning has been entered a bottleneck period. In recent years, there has not been much substantial progress in various branches of AI. It is in this context that machine learning is gradually revalued and has become one of the cores of artificial intelligence. The theory has been put into practice and successfully applied to such fields as pattern recognition, intelligent search, stock market analysis, DNA sequence research, and intelligent robots.

At present, domestic and foreign institutions that use machine learning technology to analyze social networks have achieved more results and are at the forefront level: Stanford University in the United States, IBM (released professional social network analysis software), SAS company, and Microsoft. Researchers from Tsinghua University and other universities and institutions have done more in-depth research in this field in China, according to the

Table 4: Experimental data before and after single-pass improvement.

	Double-threshold single-pass	Classic single-pass	Ratio
Mean of p_{miss}	0.1633	0.3547	0.4604
Mean of p_{false}	0.0018	0.0053	0.3396
C_{det}	0.0021	0.0059	0.3559

learning form and the experience contained in the training data set. Among them, unsupervised learning is inexperienced learning, which can recognize the data without class labels directly and predict the new sample categories using the results of unsupervised algorithms such as clustering.

Based on the advantages machine learning, which can effectively monitor network public opinion from data, the specific contributions of this article are as follows:

(1) This consists of four modules: information collection, web page preprocessing, public opinion analysis, and public information report

(2) The single-pass clustering algorithm in text clustering technology is optimized, and a single-pass algorithm based on double threshold is formed, which makes the public opinion system have more clustering performance. A single (global) threshold that applies to the entire image can be used when the grayscale distribution of object and background pixels is very pronounced. Global thresholding can then be used

(3) The dual-threshold single-pass algorithm is optimized based on the MapReduce parallel computing model so that the algorithm can carry out public opinion analysis in big data environment

However, none of the above studies on big data technology is comprehensive enough and practical enough.

3. Proposed Method

3.1. Framework Design of Public Opinion Analysis System. The public opinion analysis system designed in this paper is divided into four major functional modules, including information acquisition and web page preprocessing. Next, the design of each functional module is explained in detail.

3.1.1. Information Acquisition. The information acquisition module consists of four submodules: crawler module, update module, filter module, black-and-white list management module, and many configuration files, such as entrance address and black-and-white list.

The crawler module interacts directly with the Internet through the initial configuration of the entry address to scrape pages from the Python web crawler. However, because the content on the network is too large, if all pages are crawled, the capacity of the local machine will be a big challenge first, and the huge amount of information crawled is not necessarily the concern of users; at the same time, for the latter, data analysis will also cause a great burden. Therefore, in the design of the system information acquisition module, a filtering module is specially designed. The module filters the crawled web pages through the black-and-white list of URL and the black-and-white list of keywords to determine whether the web page is a page of user concern, thereby reducing the crawling of meaningless web pages.

When the original web pages on the Internet are modified or deleted, if the web crawler does not track their changes and updates in time, there will be a large number of expired pages or invalid links in the local web library. The captured web page is actually a mirror and backup of the Internet content. The Internet changes dynamically, and some content on the Internet has changed. At this time, this part of the captured web page has expired. Updating module is to analyze the pages in the local web database to determine which pages need to be updated. If there is a need to update, it submits an update application to the crawler module, and the data is crawled again by the crawler module.

The entrance address management module and the black-and-white list management module manage the entrance address and the black-and-white list, respectively.

3.1.2. Web Page Preprocessing. Information preprocessing module is mainly to extract the content of the collected web page information, which needs to extract important information such as title, text, links, time, and clicks. In this system, web page information is extracted based on DOM tree.

Web pages are made up of hypertext text markup language (HTML) [4], and document object model (DOM) [5] is a common way to represent and process an HTML or XML document. It can transform semistructured HTML pages into structured DOM tree structure, study the layout structure of web pages through tree structure, and extract the content of web pages. When the DOM parses, it treats the HTML document as a tree, the < HTML > tag as the root of the tree, and other components in the document as nodes in the tree. A node can be a parent node containing child nodes or a brotherhood of the same layer.

In the information preprocessing module, the basic idea of web information extraction based on DOM is to store the web templates of each web site in the form of XML configuration files on the server, where the node content of XML files is the path of nodes in DOM. Then, the paths of each node are obtained by reading these configuration files, and the information is extracted from the DOM nodes according to each path and stored in the database.

3.1.3. Public Opinion Analysis. It can respond to the massive data distributed by the preprocessing module in time. Secondly, it can accurately classify the massive text streams, and then cannot pollute the preprocessed text. Finally, the text orientation analysis aims to analyze the emotional color of each document from the mass of public opinion data

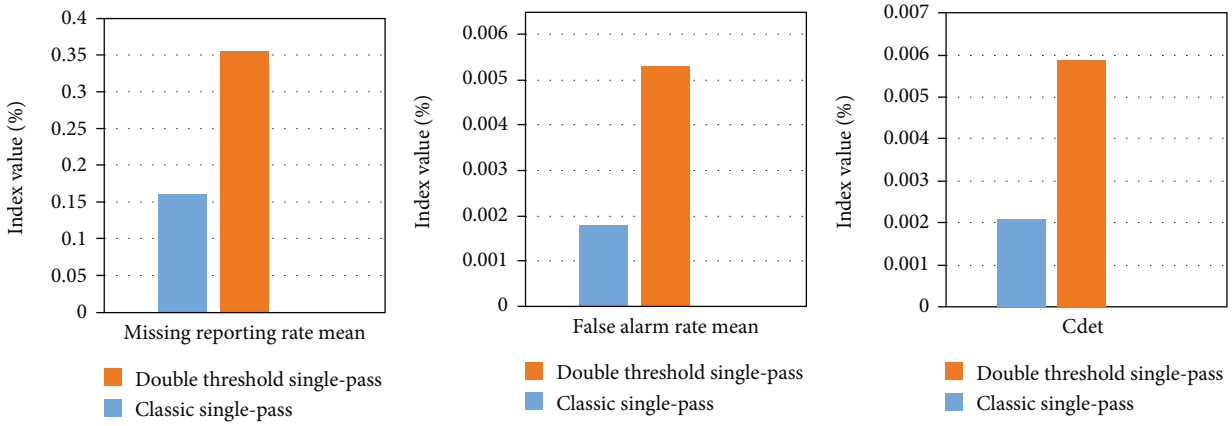

Figure 2: Histogram comparison of experimental data before and after single-pass improvement.

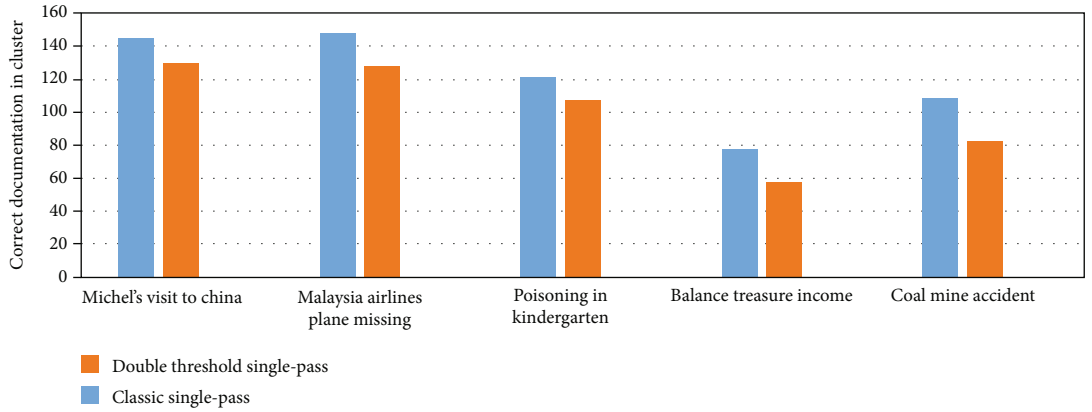

Figure 3: A comparison histogram of correctly identified documents in clusters before and after a single-pass improvement.

Table 5: The result data of double-threshold single-pass algorithm optimized by MapReduce.

Statistical unit (unit)	Michel's visit to China	Malaysia airlines plane missing	Poisoning in kindergarten	Balance treasure income	Coal mine accident
Correct documentation in cluster	156	161	129	83	117
p_{miss}	0.1031	0.1479	0.0876	0.1265	0.1529
p_{false}	0.0032	0	0.0015	0.0014	0.0029
Mean of p_{miss}			0.1236		
Mean of p_{false}			0.0010		
C_{det}			0.0012		

collected so that the users can efficiently browse to the network of emergencies, to understand the direction of public opinion. The core steps of network public opinion analysis are sorted out. According to the characteristics of network public opinion information data, a hybrid word segmentation model based on manual adjustment is proposed to process Chinese texts. The feature extraction technology is integrated into the calculation of the weight of emotional words, and a classification algorithm based on machine learning is proposed. The text sentiment orientation classifier constructed by this algorithm can be used to judge the sentiment orientation of text data. In this paper, a semantic-based text orientation analysis method is used to analyze text orientation systematically, so as to discriminate the emotion of the input document and annotate the emotion category of the document.

3.1.4. Public Opinion Report. The main function of the public opinion reporting module is to push public opinion hotspots. The public opinion hotspot push is mainly used to meet different user needs, automatically push popular public opinion for customers in the near future, provide

Table 6: Simulation data of double-threshold single-pass before and after optimization.

	Double-threshold single-pass optimized by MapReduce	Double-threshold single-pass	Ratio
Mean of p_{miss}	0.1236	0.1633	0.7569
Mean of p_{false}	0.0010	0.0018	0.5556
$C_{\det}$	0.0012	0.0021	0.5714

search keywords for customers, and collect public opinion recently pushed. The public opinion early warning is aimed at the public opinion exceeding the threshold of public opinion early warning, the system through short messages and e-mail and other real-time communication, to achieve automatic alarm, to provide early warning information for customers, and to prevent dangers in the future.

3.2. Optimization Design of Single-Pass Algorithm. Text clustering unsupervised (the number of data classes to be analyzed is unknown) [6]. Its goal is to divide the data into several classes of data clusters so that the data of each class is as different as possible, and the differences within each data cluster are as small as possible. Unsupervised text clustering hopes to discover the laws and patterns of the data itself. Compared with supervised text clustering Q, unsupervised text clustering does not need to label the data. This can save a lot of manpower and material costs. Clustering analysis [7–9] refers to the mathematical method of studying and processing the classification of a given object. Clustering is an important human behavior. A person's growth process is to learn to distinguish things by constantly improving the subconscious clustering pattern. Clustering can help us to discover the implicit association between data and identify the community structure which is closely related.

Single-pass clustering algorithm [10–12] is a simple incremental clustering algorithm, very intuitive, and easy to understand; it is obvious that it uses a greedy strategy; whenever a new document comes in, it will be allocated to a cluster, the order of the document samples has a great impact on the results of this clustering algorithm, and this algorithm is also sensitive to thresholds. The single-pass algorithm is an incremental algorithm, suitable for mining streaming data, and the algorithm has high a time efficiency; the main disadvantage of this method is that the method has the characteristic of input order dependence, that is, for the same clustering object input in different orders, different clustering results will appear.

Algorithm is sensitive to the determination of document order. A double-threshold-based single-pass algorithm is formed:

Step 1: Receive data$t_i$by processing the data features and constructing a space vector model oft_i:

$$M = \frac{1}{T}\sum_{i=1}^{T}\frac{F(H,O)}{F(Y,O)},$$
$$N^* = \ln F(\beta/J), \tag{1}$$
$$G(M,N) = \sum_{x,y} R(M,N)\left[H(X+u, Y+v) + \phi\right]^2.$$

Step 2: Calculate the similarity between the data t_i and all existing topic documents

Step 3: Find the topic of the document with the greatest similarity to the data t_i:

$$H = \sum_{M,N} G(M,N),$$
$$G(K_1, M_2) = \frac{1}{H} e^{-\frac{M^2+N^2}{2N^2}}, \tag{2}$$
$$T = \frac{1}{B} T(M,N).$$

Step 4: Set the threshold T_c. If in the process of allocating the data t_i, the data t_i is assigned to the topic having the greatest similarity. If the similarity is less than the threshold T_c, indicating that the data t_i does not belong to any existing topic, you can use t_i as a document for the new topic and create a new topic category

Step 5: Repeat Steps 2 through 4 until all samples have been assigned to a topic

Aiming at the disadvantage that single-pass algorithm is sensitive to the order of data, in order to improve the clustering effect of single-pass algorithm. The algorithm flow is as follows:

Step 1: Receive data t_i by processing the data features and constructing a space vector model of t_i

Step 2: Determine if there is a topic cluster. If there is no topic cluster, create a topic cluster K_j, and make $t_i \in K_j$, and then go to the eighth step. If there is a topic cluster, go to the third step:

Step 1: Receive data t_i by processing the data features and constructing a space vector model of t_i

$$T_{MIB} = \sqrt{m_x^2 + n_y^2},$$
$$K_{\mathrm{j}} = \sum\left(\frac{q}{t}\right), \tag{3}$$

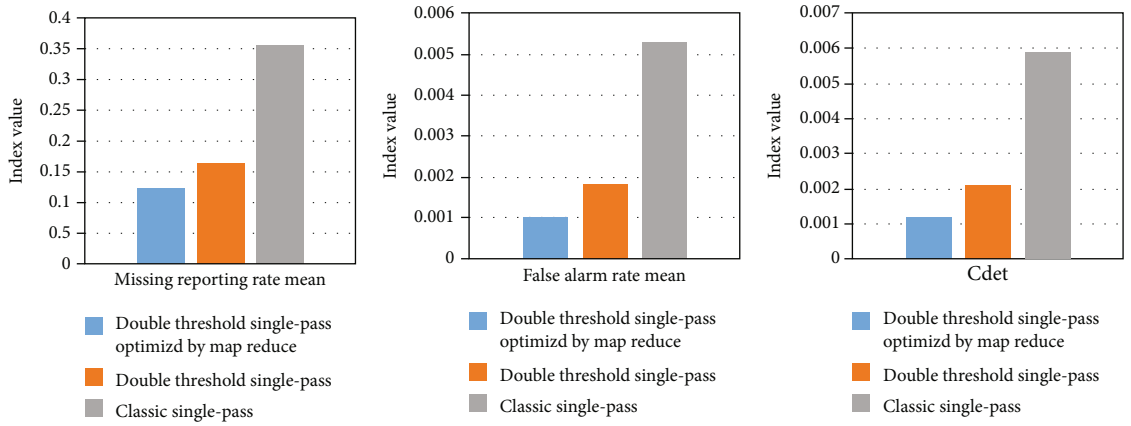

Figure 4: Three algorithms' simulation data contrast histogram.

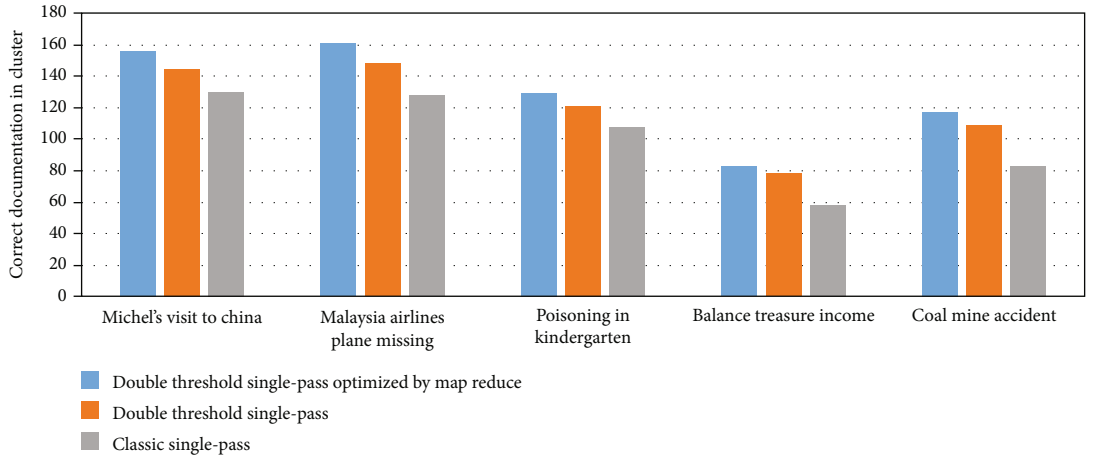

Figure 5: Contrast histogram of correct document in three algorithms' cluster.

$$S_D = S(T|D) = S/(P+Q) \tag{4}$$

Step 3: The data t_i is calculated by similarity with all previous topic documents, and the maximum similarity T is obtained:

$$t_i = \sqrt{y/(x+b)} \tag{5}$$

$$g_1 = n(y|u) = oy/(rm+p) \tag{6}$$

$$t_1 = M(1\text{-}M) \tag{7}$$

Step 4: Unlike the classical single-pass algorithm, this paper uses two thresholds: T_c and T_{mid} (where $T_c > T_{mid} > 0$). If $T > T_c$, the data t_i is merged into the cluster K_j and jumps to the fifth step. If $T_c > T > T_{mid}$, the data t_i is merged into the buffer stack of cluster K_j to wait and jump to the eighth step. If $T < T_{mid}$, the data t_i is separately attributed to create a new cluster, to the eighth step. The global threshold image segmentation region segmentation region growth and split-merge methods are two typical serial region techniques. The processing of the subsequent steps of the segmentation process should be determined according to the results of the previous steps

Step 5: Put the data t_i in the topic cluster K_j, and update the vector model of topic cluster K_j, n++, and determine the relationship between n and N. If $n > N$, then go to the fourth step; if not, turn to the sixth step

Step 6: Update the vector model of topic cluster K_j, and go to the eighth step:

$$\xi_t(I,J) = P\phi \tag{8}$$

$$R_1 = Rm \tag{9}$$

$$\varphi = \lambda \tag{10}$$

Step 7: All the data in the buffer heap of topic K_j are computed and compared with the new central vector after n updates of the central vector of K_j. Let the value of similarity be T. If $T > T_c$, go to cluster K_j, and go to the sixth step. If $T_c > T > T_{mid}$, the data remains unchanged in the buffer area and goes to the eighth step. If $T < T_{mid}$, the data t_i is separately attributed to create a new cluster, to the eighth step

$$t_2 = g[\ln Y(Y)] + g[\ln(\varphi + x(x))], \tag{11}$$

$$f_p(x) = \lambda \prod_{m=1}^{l} \phi_{pm}(\lambda) \tag{12}$$

$$f_1 = v[\|c - x\|] \tag{13}$$

Step 8: End, waiting for new data to arrive

3.3. Improvement of Dual Threshold Single-Pass Algorithm Based on MapReduce. MapReduce [13–15] parallel computing framework is a parallel computing model running on HDFS distributed storage system [16]. It can process large PB-level data in parallel in a high fault-tolerant way and realize the parallel task processing function of Hadoop platform [17, 18]. The core idea of this design is to divide and conquer the problem, not to push the data to the calculation, but to push the calculation to the data, which can greatly reduce the communication overhead.

The MapReduce execution process mainly includes the following five steps [19–21]:

(1) The input large data set slices are decomposed into hundreds of small data sets splits, which are handled by different machines

(2) Each (or several) small data set executes the Map task in parallel by an ordinary computer in the cluster, converting input into an intermediate set of key value pairs

(3) A large number of nodes are sorted and aggregated in the middle form of key value pairs according to the same principle of key value

(4) The value sets with different key values are allocated to different machines for processing, and Reduce computing tasks are performed

(5) Output Reduce calculation results

Double thresholds are improved on the MapReduce parallel framework, which can solve the problem of long iteration time for a large number of data. Based on MapReduce, it divides the improved single-pass clustering algorithm based on double thresholds into three stages: map stage, combine stage, and reduce stage.

Table 7: The analysis dimension and framework of audience response of Internet rumor.

Behavior attitude	Believe	Neutral	Unconvinced
Spread	Disseminator	Witness	Refuting rumor
Not spread	Silent supporters	Silent waiters	Silent doubters

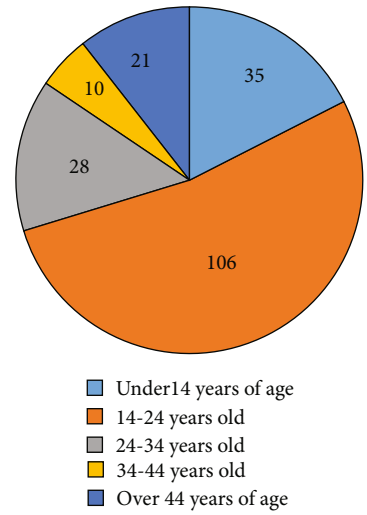

Figure 6: Age pie chart of persuasive communicator.

The specific implementation steps are as follows:

(1) In the cluster client system configuration, start a new task, specify the specific location of input and output content, and configure map and reduce functions to run the required classes, and then commit the new task to execution

(2) The Input Fonnat phase splits the vector value vector of the sample points in the input HDFS distributed file system into 64 MB or 128 MB blocks. Each block is parsed into a key value pair in the URL, and the value represents the vector value obtained from the calculation of the sample points

(3) Map process. The key value pairs in the standard format input at this stage can be expressed as <id, vector>, where id denotes the number of the sample points and vector denotes the vector values formed by the sample points. In this process, the map function is designed to be different from the traditional binary cyclic function. The inner function is mainly used for n updates of the K_j center vector, and the outer function is mainly used for cyclic traversal of the text vector itself

(4) Partitioner process. A hash calculation is performed on the index value, i.e., the value represented by

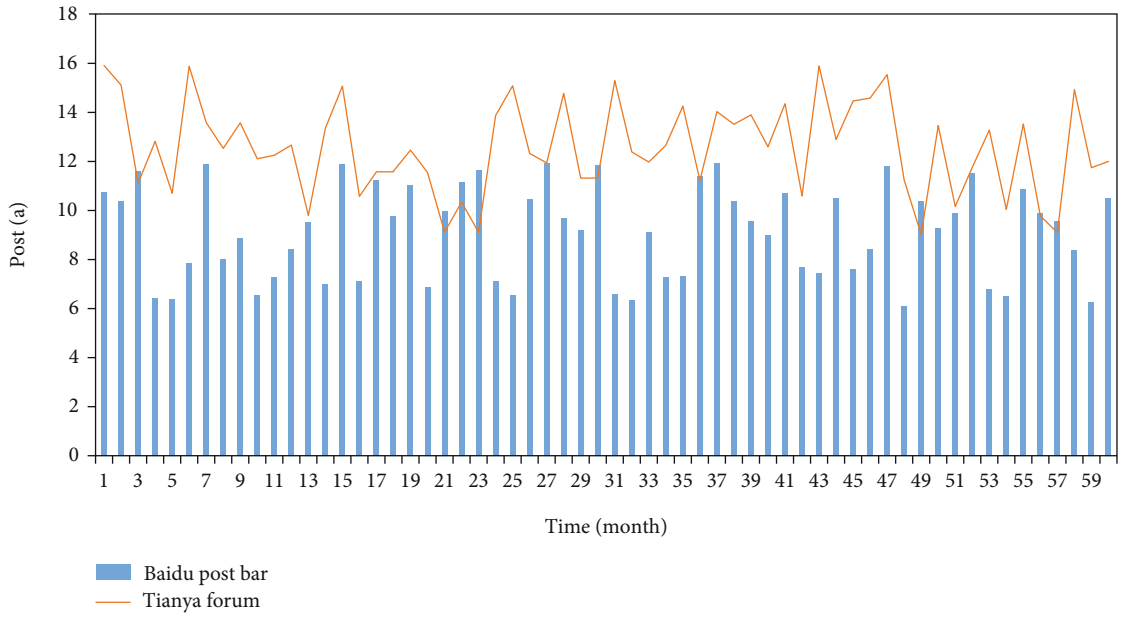

Figure 7: The result of the search.

the key in the key pair. This procedure unifies the sample points with the same index value on the same system node for calculation

(5) Combiner process. The design of this stage is mainly used data output from each map program is normalized locally, and the sample points which are divided into the same category are calculated once their average value

(6) Reduce process. The input of the process is the normalized key value pair with <URL," "> format. The average value obtained from the previous stage is used as the new condensation point or center of mass in the re-classified category

(7) Out format process. This process mainly implements the output of the reduce function. The key is the specific identifier of each class, and the value is the vector writer class used to write content after the final encapsulation of the system. At the same time, the class also stores information about other dimensions of the sample points

4. Experiments

Collection technology machine learning is proposed, in order to adapt to the explosive growth of network collection technology based on machine learning under the background of large data. Among them, the MapReduce parallel computing framework is a parallel computing model running on HDFS distributed storage system. It can process big data at PB level in a high fault-tolerant way and realize the parallel task processing function of Hadoop platform.

The basic structure of Hadoop system is shown in Figure 1. Logically, the basic structure of Hadoop system includes two parts: distributed storage and parallel computing. It uses NameNode as the master node of distributed storage to store and dominate metadata over the entire distributed file system and DataNode as the slave node of large-scale data storage. Each slave node stores real data on its own node based on the underlying Linux system. In parallel computing architecture, Hadoop uses JobTracker as the main control node of MapReduce parallel computing framework to schedule and manage the execution of jobs and TaskTracker.

In order to implement the principle of localized computing in Hadoop system design, data storage node DataNode and computing node TaskTracker will be merged and set up so that each slave node runs simultaneously as DataNode and TaskTracker so that each TaskTracker can process the data stored on the local DataNode as much as possible.

When the cluster is large or the two master nodes are overloaded, they will affect each other.

In this paper, the experimental data is searched through the network search engine, and then the 800 topics are obtained by manually swiping the page. The specific data of the topic are shown in Table 1.

In this paper, the simulation performance indicators are evaluated by TDT2004. The method is mainly through the missed detection rate, false alarm rate, and detection cost [22–25] to evaluate the algorithm designed in this paper.

The missed detection rate is the ratio of the data belonging to a topic to the actual document of the topic in the topic database. The formula is as follows:

$$p_{miss} = \frac{B}{A+B}. \tag{14}$$

The false alarm rate is the ratio of the number of documents on a topic to the number of documents in the topic database that do not describe the topic. The formula is as

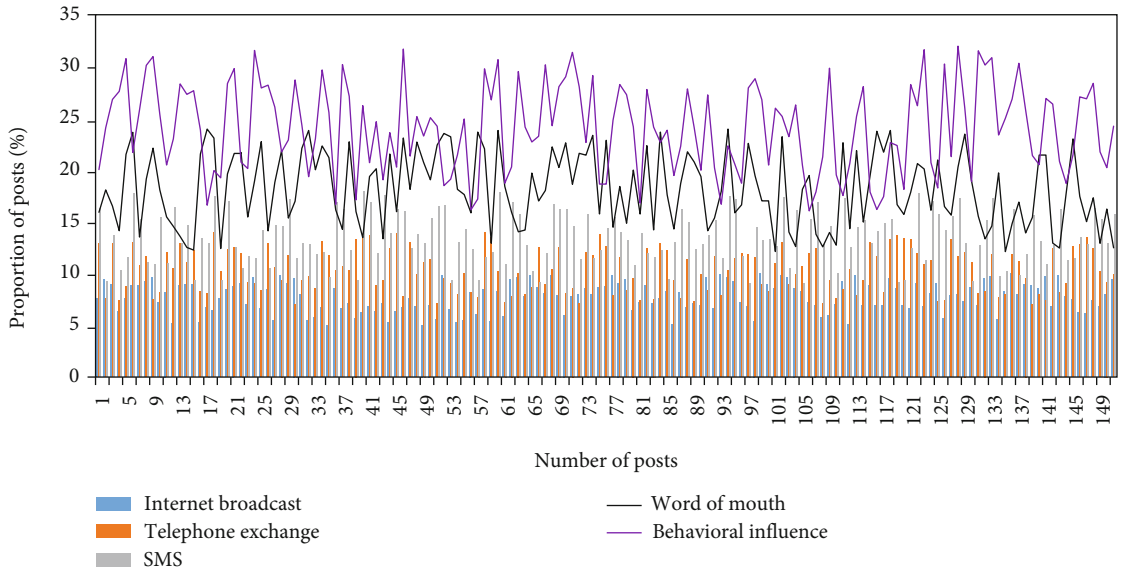

Figure 8: The specific distribution.

follows:

$$p_{false} = \frac{C}{C + D}. \tag{15}$$

A is a document that originally belongs to a topic and is correctly judged as a topic. B is a document that does not belong to a topic and is not correctly classified as a topic. C is a document that does not belong to a topic, but is considered to be a topic. D is a document that does not belong to a topic and is not mistaken for a topic.

In addition, the TDT2004 criterion also provides a test cost. The formula is as follows:

$$C_{\text{det}} = C_{\text{miss}} \cdot p_{miss} \cdot p + C_{false} \cdot p_{false} \cdot (1 - p). \tag{16}$$

Where C_{miss} is the price of missing report, C_{false} is the price of false positives, and p is the possibility of a document appearing in a topic.

5. Discussion

The results of the dual-threshold single-pass algorithm are shown in Table 2, and the data of the single-pass algorithm are shown in Table 3. The relevant experimental parameters are set as follows: $T_c = 0.6$, $T_{mid} = 0.35$, $k = 20$, $C_{miss} = 0.1$, $C_{false} = 1$, and $p = 0.02$.

In the proposed dual-threshold single-pass algorithm, the mean miss rate, the mean false alarm rate, and C_{det} of Table 2 and Table 3 are put into Table 4, and the three performance ratios of the two algorithms are calculated in Table 4. In addition, the column diagram of Table 4 is drawn as shown in Figure 2.

Combining Table 4 and Figure 2, we can see that the double-threshold single-pass algorithm has better performance than the single-pass algorithm in terms of the mean miss rate, the mean false alarm rate, and C_{det}. The average false alarm rate of single-pass algorithm with double thresholds is 0.4604 times of that of the single-pass algorithm, the average false alarm rate is 0.3396 times of that of the single-pass algorithm, and C_{det} is 0.3559 times of that of the single-pass algorithm. Compared with the classical single-pass algorithm, the performance of the proposed single-pass algorithm based on double threshold is better. In addition, Figure 3 shows the contrast histogram of the two clustering algorithms for correctly identifying documents. Similarly, it has better clustering performance than that designed in this paper.

In the big data era of network information explosion, using the big data technology MapReduce to optimize and improve the double-threshold single-pass clustering algorithm can solve the problem of long iteration time. Similarly, the dual-threshold single-pass algorithm optimized by MapReduce is used to simulate the process, and the data is shown in Table 5.

In the same way, the double-threshold single-pass algorithm is put into Table 6 with the mean of miss alarm rate, false alarm rate, and C_{det} before and after the optimization of big data MapReduce, and the three performance ratios of the two algorithms are obtained in Table 6. In addition, the three performance indicators of single-pass algorithm, double-threshold single-pass algorithm, and MapReduce optimized double-threshold single-pass algorithm are drawn as a histogram shown in Figure 4.

Combining Table 6 and Figure 4, it can be seen that the average miss detection rate, the average false alarm rate, and the performance index C_{det} of the dual-threshold single-pass algorithm have been significantly improved after the large data technology MapReduce optimization. The average miss rate after optimization is 0.7569 times, the average false alarm rate is 0.5556 times, and C_{det} is 0.5714 times. It can be seen that it is feasible to use the big data technology MapReduce to optimize the double-threshold single-pass algorithm. And since the MapReduce is a parallel computing

framework, it can improve the iterative efficiency of the double-threshold single-pass algorithm. Therefore, in terms of clustering performance, the double-threshold single-pass algorithm optimized by MapReduce is the best, the algorithm is the second, and the single-pass algorithm is the worst.

In addition, Figure 5 shows the contrast histogram of the three algorithms. It can also be seen that the clustering results of the double-threshold single-pass algorithm optimized by MapReduce are the best, the double-threshold single-pass algorithm takes the second place, and the single-pass algorithm is the worst.

Nowadays, in the network public opinion, the network rumor is more and more harmful; this article divides the network rumor audience from the attitude and the action two dimensions and may subdivide into six kinds of reaction types, such as the rumor propagator, the disclaimer, and the silent supporter, as shown in Table 7.

This paper uses the current hot micro-blogging platform to test the age of the Internet rumor audience. After publishing the topic of Internet public opinion, it analyzes the age of 200 disseminators and gets the disseminator's convincing age pie chart as shown in Figure 6.

From the Figure 6, we can see that the disseminators of online public opinion are mainly concentrated in the 14-26 years old, it is just the adolescent stage of the formation of ideas, and it can be seen that they are easy to form a trend to follow the trend, which will have a very bad impact on their thinking. Therefore, effective extraction of network public opinion information can very well inhibit the spread of network public opinion, so as to correctly guide the network public opinion audience.

Search for related posts in Baidu Tieba and Tianya forums with "earthquake" as the keyword, "in order of relevance." Due to the large number of posts, only the first 20 pages of information are selected from each of the two target websites. The result of the search is shown in Figure 7.

The ways that netizens get information are as follows: Internet communication (QQ, forums, post bars, etc.), phone calls, mobile phone text messages, word of mouth, and behavioral influences (some are learned through multiple channels). The specific distribution is shown in Figure 8.

6. Conclusions

The reason why Internet rumors can influence people is not only related to external factors such as Internet technology, but also the "human" factor, because people choose to believe the content of the rumors. What a rumor is potentially saying is more important than the surface it provides. The stronger the factors that stimulate rumors, the more people will join in the spread of rumors. As mentioned above, in many cases, the truth of the incident also depends on whether the person himself is willing to believe it or not. Therefore, it is necessary to think of ways to refute rumors from the perspective of "people": When online rumors appear, firstly, individuals should maintain rationality and self-discipline, not forward and spread the rumors at will, and secondly actively seek authoritative information and explanations. Nowadays, the network is already a network public opinion information. Aiming at this problem, this paper proposes a machine learning-based network public opinion collection technology in the context of big data, which can effectively extract the network public opinion information in the big data environment. This method is mainly based on the classical single-pass algorithm optimization design. Firstly, it is sensitive to the data processing sequence, based on double thresholds. Secondly, based on double thresholds to big data processing, this paper uses the big data technology MapReduce to optimize it. Simulations show that the proposed double thresholds can effectively reduce the sensitivity of single-pass algorithm to data processing sequence and improve the clustering performance of single-pass algorithm. Moreover, we optimize the dual-threshold single-pass algorithm by using the large data technology MapReduce, which also improves the clustering performance of the algorithm. In addition, because MapReduce is a parallel running framework, it can improve the efficiency of the dual-threshold single-pass algorithm. These simulations show that the network public opinion collection technology is suitable for public opinion information collection in big data environment. Although the single-pass algorithm has excellent performance for topic extraction of network information, the clustering algorithm has a strong dependence on the text input order. For the same data set, different input data may lead to differences in clustering results.

Data Availability

This article does not cover data research. No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LQY20G030001.

References

[1] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[2] A. L. Buczak, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[3] A. Giusti, J. Guzzi, D. C. Cireşan et al., "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robotics & Automation Letters*, vol. 1, no. 2, pp. 661–667, 2016.

[4] Z. Gao, H. Cong, G. Jiang et al., "Computer aided design system for flat-knitted fabric based on hyper text markup language 5," *Journal of Textile Research*, 2017.

[5] H. E. Zhi-lin and L. U. Zhao, "Design and implementation of management platform of comprehensive overload control based on XML document object model," *Journal of Langfang Teachers University*, 2017.

[6] K. K. Bharti and P. K. Singh, "Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering," *Applied Soft Computing*, vol. 43, pp. 20–34, 2016.

[7] J. Meng and D. Liu, "A new method for identifying bad data of power system based on spark and clustering analysis," *Power System Protection & Control*, vol. 44, no. 3, pp. 85–91, 2016.

[8] W. Zhang, J. Yang, Y. Fang, H. Chen, Y. Mao, and M. Kumar, "Analytical fuzzy approach to biological data analysis," *Saudi Journal of Biological Sciences*, vol. 24, no. 3, pp. 563–573, 2017.

[9] F. Karaaslan, "Correlation coefficients of single-valued neutrosophic refined soft sets and their applications in clustering analysis," *Neural Computing & Applications*, vol. 28, no. 9, pp. 2781–2793, 2017.

[10] Y. Dang, X. U. Zhiwei, L. Liu, and Y. Wang, "Research on improved single-pass text clustering algorithm in public opinion," *Journal of Inner Mongolia University of Technology*, 2017.

[11] J. Wu, Q. Meng, S. Deng, H. Huang, Y. Wu, and A. Badii, "Generic, network schema agnostic sparse tensor factorization for single-pass clustering of heterogeneous information networks," *PLoS One*, vol. 12, no. 2, article e0172323, 2017.

[12] L. I. Fang, L. L. Dai, Z. Y. Jiang, and S. Li, "The combination of an autoencoder network and single-pass clustering for detection and tracking," *Journal of Beijing university of Chemical Technology*, 2017.

[13] T. R. Sree and S. M. S. Bhanu, "HADM: detection of HTTP GET flooding attacks by using analytical hierarchical process and Dempster–Shafer theory with Map Reduce," *Security & Communication Networks*, vol. 9, no. 17, pp. 4341–4357, 2016.

[14] M. R. Ghazi and D. Gangodkar, "Hadoop, MapReduce and HDFS: a developers perspective," *Procedia Computer Science*, vol. 48, pp. 45–50, 2015.

[15] J. Bai, D. Yanhui, and T. Lu, "Internet rumor reporting system based on the blockchain incentive mechanism," *Mobile Information Systems*, vol. 2021, Article ID 7500639, 7 pages, 2021.

[16] L. Wang and J. Zhai, "Research of distributed storage system based on HDFS," *Intelligent Computer & Applications*, 2016.

[17] C. Wang, H. U. Yuping, and Y. I. Yeqing, "Cross-layer parameter optimization algorithm for Hadoop cloud computing platform," *Journal of Central China Normal University*, 2016.

[18] X. Yang, H. Ma, and M. Wang, "Rumor detection with bidirectional graph attention networks," *Security and Communication Networks*, vol. 2022, Article ID 4840997, 13 pages, 2022.

[19] Q. Chen, C. Liu, and Z. Xiao, "Improving MapReduce performance using smart speculative execution strategy," *IEEE Transactions on Computers*, vol. 63, no. 4, pp. 954–967, 2014.

[20] Q. Liu, D. Jin, X. Liu, and N. Linge, "A survey of speculative execution strategy in MapReduce," in *Cloud Computing and Security*, pp. 296–307, Springer, Cham, 2016.

[21] E. C. Puig, J. A. Interrante, M. D. Osborn, and E. Pool, *Integrating execution of computing analytics within a mapreduce processing environment*, U.S. Patent Application No. 14/317,687, 2016.

[22] A. Ahmed, "Performance analysis of machine learning based Botnet detection and classification models for information security," *Journal of Cybersecurity and Information Management*, no. 1, pp. 44–53, 2019.

[23] F. Amal, "A survey on machine learning techniques for supply chain management," *American Journal of Business and Operations Research*, vol. 2, no. 1, pp. 24–38, 2021.

[24] L. I. Shun, M. O. Nanwen, L. I. Qinyong, and D. O. Orthopedics, "Analysis on curative effect and leakage rate of percutaneous vertebroplasty by different viscosities of bone cement in patients with osteoporotic vertebral compression fractures," *Laboratory Medicine & Clinic*, 2017.

[25] K. Pardee, A. A. Green, M. K. Takahashi et al., "Rapid, low-cost detection of Zika virus using programmable biomolecular components," *Cell*, vol. 165, no. 5, pp. 1255–1266, 2016.