Building Scalable Large-scale Distributed Function-dependent Discovery Platform (SmartFD) of Education Short Videos

Ying Sun[1], Guofeng Ma[2] and Shuwei Feng[3]

[1]School of Artificial Intelligence, Zhengzhou Railway Vocational and technical college, Zhengzhou, Henan 451460, China, 10330@zzrvtc.edu.cn
[2]School of Artificial Intelligence, Zhengzhou Railway Vocational and technical college, Zhengzhou, Henan 451460, China, maguofeng@zzrvtc.edu.cn
[3]School of Artificial Intelligence, Zhengzhou Railway Vocational and technical college, Zhengzhou, Henan 451460, China, 10994@zzrvtc.edu.cn

Corresponding author: Ying Sun, 10330@zzrvtc.edu.cn

Abstract. In the era of big data, the fragmentation of communication has facilitated the birth of short videos. With its "simple and concise" audio-visual language and unique narrative, short videos cater to the audience's viewing needs and thus have developed rapidly and accumulated huge data information. Big data technology is the underlying technology module of short video platform algorithm recommendation and is also the technical basis of short video communication. Therefore, this paper proposes an efficient and scalable large-scale distributed function-dependent discovery algorithm SmartFD, which provides a technical basis for the application of big data technology in short video communication from the underlying logic of the algorithm.

Keywords: Short video communication; big data technology; evaluation method; SmartFD
DOI: https://doi.org/10.14733/cadaps.2023.S10.24-33

1 INTRODUCTION

In 2011, the first mobile short video application was released by Viddy in the United States, followed by Twitter in 2013 also launched a short video application because the short video is still in the process of rapid evolution, so the academic community on the definition of the short video is divided, did not produce a precise definition, the discussion is mainly focused on its playing time, content-related and playing platform, for short video is more generally recognized by the definition of a short video. Short videos' more commonly accepted definition ranges from a few seconds to a few minutes in length [1]. They are distributed on PC and mobile devices, with aesthetic characteristics of low threshold, fragmentation, and interactivity in creation. At the early stage of the development of short-form video media, the media value of the short-form video industry has not been fully explored, the audience scope is small, only as a supplement to the traditional media, and there are great IP, presentation form, copyright content restrictions, early Internet technology, and broadband hardware is also difficult to support the short-form video media to expand the market demand, short-form video media in general presents a more sluggish development trend. However, in 2016, with the tumultuous changes in China's news communication industry, the short-form video industry entered a brand-new development stage based on the fractured transformation of China's news communication ecology. As Keshavarzi and Rahmani-Asl [2] stated, With the changes and convergence of technology, audience, capital, and communication ecology, short-form video media entered a comprehensive adjustment and rapid development stage due to the times and changes. During this period, China's news communication industry is undergoing a crucial transformation period of comprehensive deconstruction and integration and reorganization. Panarello et al. [3] also proposed that with its unique media nature of fragmentation, mobile, visualization, and interaction, short video media has gradually grown into a pivotal part of the entire media communication matrix, thus opening the intro of the short video era.

The rapid evolution of information technology has transformed the mode of information dissemination from early graphics to short video iterations. Short videos, fueled by mobile and social media, have become an indispensable tool for information dissemination in the age of fragmented reading. Over the past decade, short video media has emerged as a potent force, catalyzing national creation and interaction, and promoting media integration. It has redefined the discourse form of information dissemination and stands at the forefront of communication field fission, reconstructing the distance, content, and interaction in communication. Zheng et al. [4] emphasized that the short video industry has become the new engine of growth in the network audiovisual industry. The disruptive development of short video media is driven by technology renewal, and the in-depth use of internet technology and fission has enabled short videos to be widely accessible, contributing to the realization of the Internet of everything. However, Zheng et al. [5] pointed that the rapid development of short video media has also raised concerns about issues such as entertainment addiction, lack of innovation, information cocooning, and circle solidification. Addressing these concerns and maintaining the advantages of short video communication has become a focus of academic circles.

The explosive growth of short video application data has provided an opportunity to analyze short video communication paths and effects using big data technology. This paper aims to address the computational complexity and system complexity challenges of distributed data mining and machine learning algorithms. It selects a series of data mining and machine learning algorithms that are frequently used on a fundamental basis, have high complexity and great computational efficiency problems, and are challenging to design distributed algorithms. The study aims to develop efficient large-scale distributed parallelized data mining and machine learning methods and algorithms to improve the computational efficiency of big data intelligent analysis. The ultimate goal is to enhance the application potential of big data technology in the evaluation of short video communication effects.

2 RELATED WORK

2.1 The Spread of Short Videos

With the rapid development of the Internet and the continuous change of equipment, mobile short videos have rapidly gained attention and created a new way of media communication. First, regardless of the platform, the expansion of influence and value realization happens after reaching the audience [6]. To increase awareness, get the audience in a short period, and attract potential audiences who are not yet in the short video field to it, short video platforms have interwoven various initiatives in parallel to cover the areas that the public may be exposed to comprehensively, such as placing ads on other platforms - Weibo, Zhihu, community forums, etc. For example, by placing advertisements on other platforms - Weibo, Zhihu, community forums, etc. - and intensively pushing promotions to mobilize the curiosity of the public and create enthusiasm, cooperating with famous TV shows, placing materials across terminals, naming variety shows, and TV parties, etc., they constantly appear in the public's view, subconsciously increasing their awareness in the public's mind and creating a universal effect, and. Because of the influence of a short video, its elements can generate a wide range of appeal. Many single-content platforms will also extract a certain component of a short video to promote themselves and achieve the interchange of traffic, including music, images, copy, etc. [7].

In addition, the freedom of crossing media makes opinion leaders no longer fixed in a certain media but can develop across media, and many content producers who originally used text or pictures as the expression form in microblogs, WeChat public numbers, and Zhihu platforms combine their content with short videos to spread in the way of video. It is worth noting that since many short video platforms also include live streaming, many short video bloggers will live stream simultaneously, and the original live stream anchors will also create short videos. With the cross-platform flow of opinion leaders and the interactive chain between platforms, the information initially confined to one platform is spread among different media through the interactive chain [8]. For example, building social platforms such as QQ, WeChat, Weibo, and Zhihu provides a complete social relationship chain for people's virtual interaction, so that information from different platforms can be smoothly migrated and interest circles can be integrated. The platform's internal algorithm will calculate related users based on users' WeChat, QQ, and address book channels and make recommendations in the form of "people you may know" or address book friends.

In this way, we try to migrate the friend relationship network and support the weak relationship social chain of short video with a strong relationship social chain, forming a social network combining acquaintance social and stranger social on the short video platform. In this regard, while media such as Jitterbug, Racer, and Xiaojiaxiu ask whether to access their address books, WeChat-based short videos are automatically associated with each other, with the video number portal embedded in WeChat, and automatically associated with friends without asking. At the same time, it will be marked under the video which friends have liked it and associated with the friends' other liked videos, which further strengthens the social density of short videos and makes them a circle of friends, and adds the pop-up culture to it, so the audience can see others' comments without opening the comments.

2.2 Short Video Communication and Big Data Technology

Accurate data analysis is a necessary technical means for a short video to calculate user preferences intelligently. It is challenging to make timely adjustments to communication strategies according to the subtle changes in users and the market by traditional means. First of all, for the short video platform operation, from the time users enter the short video platform, data is constantly generated; the initial entry can count the basic information of users, such as age, region, gender, equipment used, etc., and then through the user use log, search, stay time and friends related to the recommendation of accurate data analysis to calculate the user's needs to carry out relevant push and search correlation, to enhance user stickiness. When users first enter the short video platform, the platform will recommend the recent popular videos for new users according to the algorithm. At the same time, according to the new users' friends, it will presume the common interests and push the contents of interest reflected in the previous behavior of friends. It will let users choose the content labels of goods and recommend vertical ranges according to the refined labels. At the same time, users can get the information they need with lower time consumption. Secondly, for some users who have changed from simple receivers to active disseminators, the data can gradually reduce their self-inquiry time and clarify the effectiveness of each video through the number of plays, viewing time, likes and comments, etc. In contrast, Hoy [9] argued that cloud-based big data, which can realize the integration of communication, caching, and computing capabilities, can further Optimize the corresponding data to complete more accurate data distribution.

Big data technology can transfer various data collected from the Internet to the analysis background of big data in time and combine with artificial intelligence to make strategic decisions and distribute different contents to different users. And, when users use a certain amount of time, the platform's post-algorithm will gradually improve the user's preference judgment based on the user's usage path and then push many videos that meet the user's preference [10]. The user has been surrounded by his preferred information. In other words, artificial intelligence algorithm based on big data is the technical basis of short video communication and an essential driving force in the era of short video communication.

With the rapid growth of big data in the modern era, intelligent analysis applications have become more complex, involving the fusion of big data, big models, and extensive computation. This has resulted in new challenges and problems in researching the basic theoretical methods and key technologies of big data intelligent analysis. One of the most significant issues is the computational efficiency problem that big data intelligent analysis faces in practical applications. Therefore, there is a need to develop efficient large-scale distributed parallel data mining and machine learning methods and algorithms to address this problem. However, designing such algorithms requires the consideration of both the inherent computational complexity of stand-alone serial algorithms and the system complexity of parallelizability, storage, I/O, and network communication of distributed algorithms in a distributed parallel environment.

In addition to the computational efficiency problem, big data intelligent analytics also faces challenges related to the high technical threshold of modeling and programming methods, which are not easy to use. To overcome this challenge, there is a need to study automated machine learning methods and algorithms and flexible, easy-to-use programming methods and platforms. However, most machine learning algorithms are associated with complex technical issues, such as the effectiveness of search and modeling methods and the optimization of search and computation efficiency. Furthermore, big data intelligent analysis is a problematic cross-cutting research problem that requires the consideration of many complex factors in both machine learning and big data processing.

To promote the application and use of big data algorithms in the short video field, this paper focuses on the application needs and research objectives of efficient big data intelligent analysis calculation methods and algorithms. Specifically, this work aims to analyze the complexity of distributed data mining and machine learning algorithms in big data scenarios, design and optimize distributed algorithms based on the fundamental theories and methods of big data distributed parallel computing, and study related data mining and machine learning. To achieve this objective, the paper selects a series of data mining and machine learning algorithms with high complexity and computational efficiency, and the technical difficulty of distributed algorithm design, and studies the methods and algorithms to achieve efficient large-scale distributed parallelization. This research aims to improve the computational efficiency of big data intelligent analysis and its application to the field of short video communication.

3 METHODOLOGY

Function dependency is a fundamental data structure in big data analysis and mining. However, function-dependent tasks' computational complexity and memory complexity is high, especially in large-scale data scenarios. The algorithm running time and memory overhead is enormous, and there are significant computational efficiency problems. To efficiently handle large-scale data sets, we propose an efficient and scalable large-scale distributed function dependency discovery algorithm SmartFD based on attribute reordering, which can achieve tens or even hundreds of times performance improvement compared with existing function dependency discovery algorithms. It also has good row scalability, column scalability, and near-linear node scalability, thus providing an algorithmic basis for evaluating the effectiveness of short video communication.

3.1 Function Dependency

Functional dependencies are used to express dependencies between attributes in a relational table. A functional dependency X → A is valid if the value of attribute set X uniquely determines the value of attribute A. If any two records in a relational table have the same value on attribute set X, they must have the same value on attribute A. If any two records in a relational table R have the same value in the set x, then they should necessarily have the same value on the set A as well. Due to the generality of attribute sets, function dependencies have been widely used in many big data algorithms and tasks. Therefore, function dependency discovery has also become a hot research problem in the field of big data. Function dependency discovery is a computationally intensive task. Its computational complexity is $O(n^2(m^2)^2 2^m)$. Moreover, the function dependency discovery process will generate a large amount of data. Moreover, the function-dependent process will generate intermediate data, which may result in an overhead of huge memory. Therefore, the existing single-computer function-dependent algorithms cannot efficiently handle large-scale data. Figure 1 shows that the current optimal single-machine function-dependent discovery algorithm HyFD still has shortcomings regarding row scalability and column scalability.

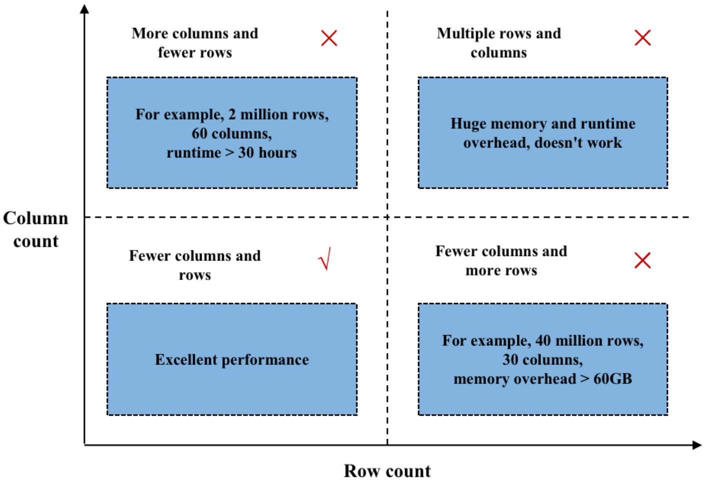
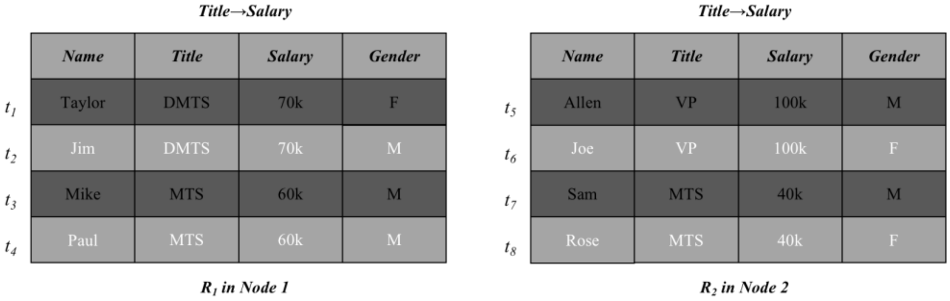

More columns and fewer rows
×
For example, 2 million rows, 60 columns, runtime > 30 hours
Multiple rows and columns
×
Huge memory and runtime overhead, doesn't work
Column count
Fewer columns and rows
√
Excellent performance
Fewer columns and more rows
×
For example, 40 million rows, 30 columns, memory overhead > 60GB
Row count

Figure 1: Performance of the current optimal single-machine function-dependent discovery algorithm at different data sizes.

Several emerging approximation algorithms, to some extent, can reduce the computational complexity of function dependence. However, in the real world, function dependencies normally have many restriction conditions. For example, the optimal solution must be found using complete functional dependencies in recommendation algorithm of short video communication. Moreover, query optimization requires using all function dependencies to construct database access constraints. Therefore, there is an urgent need for an efficient and scalable function dependency discovery algorithm that can handle extensive scale data in large-scale data scenarios.

In recent years, some big data computing frameworks have become the standard for big data processing. However, distributed function dependencies face the problem of correctness. In a distributed environment, data is often divided into multiple partitions. Function dependencies on each partition may not hold on to the total amount of data. For example, as shown in Figure 2, the relational table R is horizontally partitioned into R_1 and R_2. Name, Title, Gender, and Salary are attributes of the relational table R. In both R_1 and R_2, the relational table R is partitioned into R_1 and R_2. In R_1 and R_2, Title → Salary can be easily deduced. As shown in equations (3.1)-(3.2), Title → Salary does not hold in the whole relational table R.



Title→Salary

	Name	*Title*	*Salary*	*Gender*
t_1	Taylor	DMTS	70k	F
t_2	Jim	DMTS	70k	M
t_3	Mike	MTS	60k	M
t_4	Paul	MTS	60k	M

R_1 in Node 1

Title→Salary

	Name	*Title*	*Salary*	*Gender*
t_5	Allen	VP	100k	M
t_6	Joe	VP	100k	F
t_7	Sam	MTS	40k	M
t_8	Rose	MTS	40k	F

R_2 in Node 2

Figure 2: The relationship table R of horizontal distribution.

$$t_3[Title] = t_7[Title] \tag{3.1}$$

$$t_3[Salary] \neq t_7[Salary] \tag{3.2}$$

$$Title \neq Salary \tag{3.3}$$

To ensure the correctness of distributed function dependencies, researchers have proposed an approach based on data redistribution. Let attribute A be the key in data redistribution. By performing data redistribution on the relational table R, tuples with the same value on attribute A is able to be shuffled to the same partition. The Left-Hand Side (LHS) contains a function dependency on attribute A that holds globally if it holds on all data partitions. The data redistribution-based approach ensures that all functional dependencies for the Left-Hand Side containing attribute A are correctly discovered. Other attributes can be processed sequentially, and all functional dependencies can eventually be found. However, the existing algorithms based on data redistribution are less scalable and computationally inefficient. To address these problems, this study proposes a novel distributed function dependency algorithm, SmartFD, which supports the processing of large-scale distributed data sets.

3.2 An Algorithmic Framework for Attribute-Based Reordering

SmartFD can solve the high load problem of function dependencies in big data scenarios. Specifically, function-dependent load reduction can be addressed by a data distribution strategy. The number of data redistributions is equal to the number of mediators. After the first few data redistributions, the complexity of most of the function dependencies will be substantially reduced. The relationship table R is shown in equation (4):

$$G_c(A_i) = \{cFD|\}A_i \in lhs(cFD), A_j \notin lhs(cFD), \forall j < i \tag{3.4}$$

where *cFD* refers to candidate function dependence. Let $N_c(A_i)$ be the size of $G_c(A_i)$. Taking the first attribute A_1 as an example, the number of candidate function dependencies in $G_c(A_1)$ containing k attributes in the left part is $C_{m-1}^{k-1}C^{1}_{m-1-(k-1)}$, $k \in [1, m-1]$. $N_c(A_1)$ is computed exactly as follows:

$$C^1_{m-1} + C^1_{m-1}C^1_{m-1-1} + \cdots + C^{m-3}_{m-1}C^1_{m-1-(m-3)} + C^{m-2}_{m-1}C^1_{m-1-(m-2)} = \sum_{j=1}^{m-1} j \times C^j_{m-1} \tag{3.5}$$

$$\sum_{j=1}^{m-1} j \times C^j_{m-1} = (m-1) \times 2^{m-2} \tag{3.6}$$

After the first data distribution, the load of all dependencies containing A_1 will be greatly reduced. Then, $N_c(A_2)$ represents the number of candidate dependencies for which the left part contains A_2 but not A_1. Specifically, $N_c(A_2)$ is equal to:

$$\sum_{j=1}^{m-2} j \times C_{m-2}^{j} + \sum_{j=0}^{m-2} C_{m-2}^{j} = (m-2) \times 2^{m-3} + 2^{m-2} \tag{3.7}$$

Similarly, when $i > 2$, we can set T to be the sum of all functional dependencies equal to $m \times 2^{m-1} - m$. Further introducing the cumulative ratio function (CRF) to the model, the CRF is the ratio of all function dependencies that have undergone negative reduction to the data redistribution and can be expressed as:

$$CRF(k) = \frac{\sum_{i=1}^{k} N_c(A_i)}{T} \tag{3.8}$$

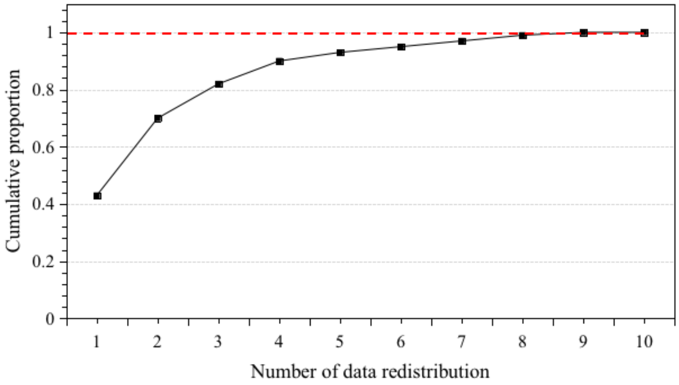


Figure 3: Paradigm of the cumulative ratio function.

Figure 3 shows the CRFs paradigm. We can find that the function dependencies are concentrated in the first half of the computational process, and after data redistribution, the function dependencies can be verified.

The above observations show that processing of the first half of the data accounts for most of the workload of function-dependencies. Therefore, the optimized algorithm should process the first half of the more complex processes first to achieve load balancing. Figure 4 illustrates the overall processing flow of the SmartFD algorithm, which consists of two phases. In the first stage, we propose a combined algorithm based on AFDD and Batch AFDD. Next, we will show the data preprocessing phase of the SmartFD algorithm.

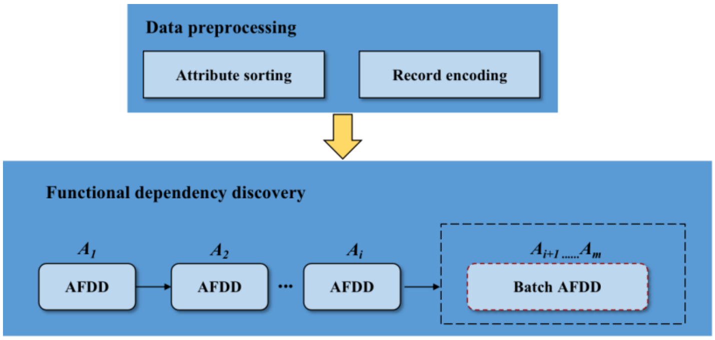


Figure 4: General flow of the SmartFD algorithm.

4 RESULTS

KReal datasets were used for the experiments. The DATA-1 was derived from the short video data related to Huangshan Mountain in 2021 on the TikTok, containing 4 million records and 71 attributes. The DATA-2 is derived from the short video data about Jiuhua Mountain in 2021 on the TikTok, which includes 3.2 million records and 42 attributes. HyFD is the current optimal centralized function-dependent discovery algorithm. A multi-threaded version of HyFD was used for the experiments. To discover all minimal function dependencies, the experiments disable the memory protection mechanism of HyFD. In addition, HFDD is the currently available distributed function dependency algorithm. The experiments are based on the Apache Spark distributed parallel computing platform and evaluated HFDD using the same configuration as SmartFD.

SmartFD outperforms HyFD and achieves a speedup ratio of 3.2 to 44.9 times. As the size of the data set increases, SmartFD still achieves good performance. HyFD, on the other hand, fails due to running out of time or memory constraints. In addition, the function dependency results found by SmartFD are identical to those of HyFD, so the correctness of SmartFD can be guaranteed (Figure 5).

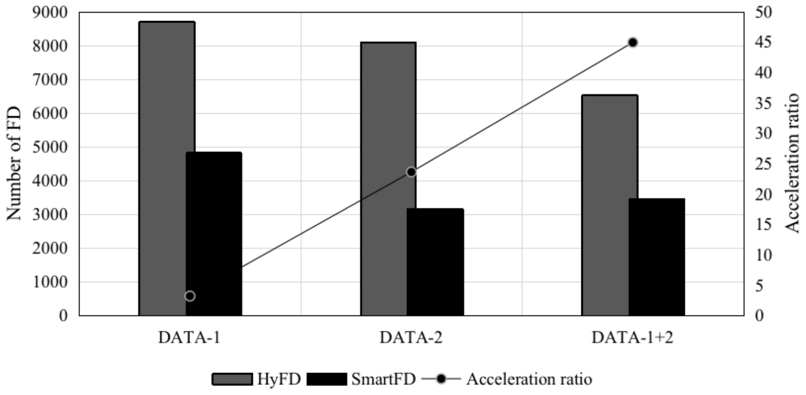


Figure 5: Comparison between SmartFD and HyFD.

HFDD cannot handle a dataset with 30 columns in a given time limit and has poor column expansion. In contrast, SmartFD can process datasets with 30 columns quickly and efficiently. SmartFD and HyFD are both distributed function-dependent algorithms. SmartFD is a distributed function dependency algorithm, but SmartFD outperforms HFDD mainly because HFDD directly verifies all the candidate function dependencies without any pruning operation. At the same time, SmartFD achieves efficient pruning by distributed sampling, which significantly reduces the function dependency verification overhead. SmartFD can also achieve load balancing and high resource utilization through attribute reordering (Figure 6).

We also evaluate the scalability of the SmartFD algorithm in detail in terms of node scalability. When the number of nodes is less than 10, the load and memory overhead of each node becomes larger, and the performance of the operating system decreases due to the lack of available memory. We evaluated the node scalability of SmartFD by manipulating the variation of the number of nodes in different datasets. As shown in Figure 7, the running time of SmartFD is continuously compressed as the number of computational nodes increases. In other words, SmartFD shows good node scalability and algorithm scalability.

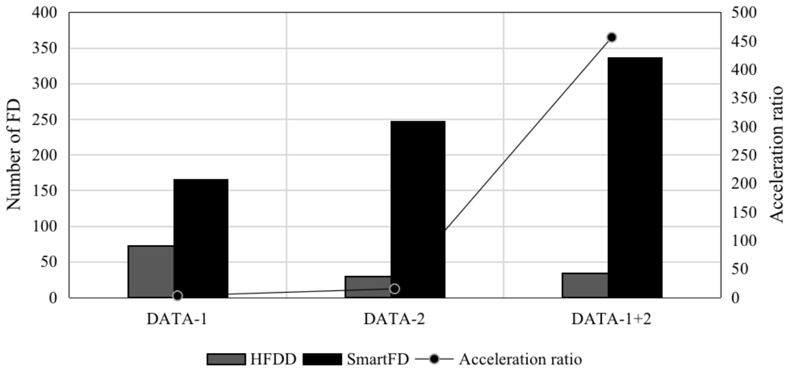


Figure 6: Comparison between SmartFD and HFFD.

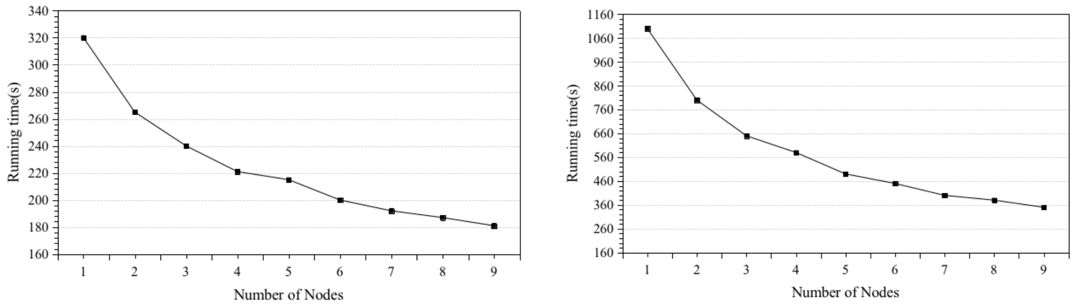


Figure 7: Node scalability evaluation: (a) DATA-1, (b) DATA-2.

5 CONCLUSION

According to the "China Network Audio-Visual Development Research Report 2021" released by China Network Audio-Visual Program Service Association, among all network audio-visual services, a short video has the highest user usage rate, accounting for 88.3% of the total Internet users, with a user scale of 873 million people, more than half of the total national population. The short video market size is as high as 205.13 billion, up 57.5% year-on-year, accounting for 34.1% of all online video industries. In comparison, the short video market size is twice the market size of the integrated video (119.03 billion) and live webcast (113.44 billion), and the year-on-year growth percentage is 3.5 times that of integrated video (16.3%) and 1.7 times that of the live webcast (34.5%). The short video has become the leader of new media platforms. The vast market share and user scale also produce a large number of data streams, providing the basis for the penetration of big data technology into the short video field. However, the existing programming methods and platforms for intelligent analysis of big data face outstanding ease of use problems, which are difficult to be grasped and used by ordinary data analysts. This bottleneck seriously restricts the promotion and popularization of big data technology in the field of short video screens. Specific to the short video platform recommendation algorithm, it still suffers from high computational and memory complexity and huge computation time and memory overhead. Therefore, we propose an efficient and scalable large-scale distributed function-dependent discovery algorithm SmartFD. First, the study presents an algorithmic framework based on attribute reordering to achieve load balancing and improve resource utilization in distributed scenarios. Then, an efficient distributed algorithm AFDD is proposed to discover all function dependencies grouped by a given attribute. In the AFDD algorithm, the study designs a distributed sampling method based on the FSE mechanism and an index-based distributed verification method to improve the efficiency of algorithmic recommendation computation for short videos in big data scenarios. Experimental results on Apache Spark, a distributed data-parallel computing platform, show that SmartFD outperforms HyFD and HFDD, achieving speedup ratios of 3.2-44.9 times and 2.5-455.7 times, respectively. In other words, compared with existing function-dependent discovery algorithms, SmartFD can achieve performance improvements of tens or even hundreds of times and has good node scalability, which means that SmartFD has great potential for applications in the field of short video screens.

Ying Sun, https://orcid.org/0000-0002-5071-7224
*Guofeng M*a, https://orcid.org/0000-0002-9773-1509
Shuwei Feng, https://orcid.org/0009-0003-6820-0359

REFERENCES

[1] Dessì, D.; Fenu, G.; Marras, M.; & Recupero, D.-R.: Bridging learning analytics and cognitive computing for big data classification in micro-learning video collections, Computers in Human Behavior, 92, 2019, 468-477. http://doi.org/10.1016/j.chb.2018.03.004

[2] Keshavarzi M.; Rahmani-Asl M.: GenFloor: Interactive Generative Space Layout System via Encoded Tree Graphs, Frontiers of Architectural Research, 10(4), 2021, 771-786. https://doi.org/10.1016/j.foar.2021.07.003

[3] Panarello, A.; Celesti, A.; Fazio, M.; Puliafito, A., Villari, M.: A big video data transcoding service for social media over federated clouds, Multimedia Tools and Applications, 79(13), 2020, 9037-9061. http://doi.org/10.1007/s11042-019-07786-9

[4] Zheng, S.; Cui, J.; Sun, C.; Li, J.; Li, B.; Guan, W.: The effects of the type of information played in environmentally themed short videos on social media on people's willingness to protect the environment, International Journal of Environmental Research and Public Health, 19(15), 2022, 9520. http://doi.org/10.3390/ijerph19159520

[5] Zheng, L.; Liu, S.: Research on the strategy of mobile short video in product sales based on 5G network and embedded system, Microprocessors and Microsystems, 82, 2021, 103831. http://doi.org/10.1016/j.micpro.2021.103831

[6] Zhang, X.; Wu, Y.; Liu, S.: Exploring short-form video application addiction: Socio-technical and attachment perspectives, Telematics and Informatics, 42, 2019, 101243. http://doi.org/10.1016/j.tele.2019.101243

[7] Farooq, S.; Kamal, M.-A.: An investigation into adoption of digital design software in the education of interior design, Universal Journal of Educational Research, 8(11B), 2020, 6256-6262. https://doi.org/10.13189/ujer.2020.082264

[8] Ullah, W.; Ullah, A.; Hussain, T.; Muhammad, K.; Heidari, A.-A.; Del Ser, J.; De Albuquerque, V.-H.-C.: Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data, Future Generation Computer Systems, 129, 2022, 286-297. http://doi.org/10.1016/j.future.2021.10.033

[9] Hoy, R.-R.: Quantitative skills in undergraduate neuroscience education in the age of big data, Neuroscience Letters, 759, 2021, 136074. http://doi.org/10.1016/j.neulet.2021.136074

[10] Mubarak, A.-A.; Cao, H.; Zhang, W.; Zhang, W.: Visual analytics of video-clickstream data and prediction of learners' performance using deep learning models in MOOCs' courses, Computer Applications in Engineering Education, 29(4), 2021, 710-732. http://doi.org/10.1002/cae.22328