Neural Computing and Applications
https://doi.org/10.1007/s00521-018-03983-z

MULTI-SOURCE DATA UNDERSTANDING (MSDU)

Big data analytics for MOOC video watching behavior based on Spark

Hui Hu[1,2] · Guofeng Zhang[1,2] · Wanlin Gao[1,2] · Minjuan Wang[1,2]

Received: 24 February 2018 / Accepted: 27 December 2018

Abstract
The purpose of this study is to measure the effectiveness of courses delivered using MOOCs in China Agricultural University. Video watching is considered to be the most important way to disseminate knowledge in Massive Open Online Course (MOOC). Its mission is to understand the degree of students' learning engagement and to provide suggestions for teachers to construct courses. This paper proposes the analysis methods of students' video watching behavior in MOOCs platform and verifies it with the data of the cauX platform. Initially, a detailed statistical analysis of video watching data and behavior was performed. Later, data preprocessing algorithms based on Spark platform were developed and used to calculate the number of video watching behaviors in every hour and every minute. Then, the entropy weight method was used to calculate the weight of pause video, seek video and speed change video. Finally, we analyze and discuss the results of experiment. The results show that the proposed method based on Spark platform can quickly and accurately analyze the characteristics of video watching behavior.

Keywords MOOC · Big data · Video watching behavior · Spark

1 Introduction

In the past few years, a new model of network course teaching, Massive Open Online Courses (MOOCs), has drastically risen in popularity in education [1]. MOOC platforms such as Coursera, edX and Udacity have offered courses with enrollments reaching hundreds of thousands and have become subjects of intensive debate [2]. Unlike traditional courses, MOOCs have the characteristic of learning by everybody in anytime and anywhere, which is easy and free to use. The courses in MOOC platform are mainly for higher education. Many universities and institutions in China are offering various online courses via different platforms like Xuetangx of Tsinghua University, Caux of China Agricultural University [3] and icourse of NetEase, etc. And now it has accumulated massive log data in the process of using vast teaching resources [4]. However, due to the lack of supervision in the teaching process, it is hard to known the participation, enthusiasm and learning effect of MOOCs. Therefore, it is important for MOOC platform owners and course instructors to deduce student engagement and adjust the course structure to attain optimum pedagogical effectiveness.

The tracking logs in MOOC platform provide detailed records of the students' interaction with course content, including video watching, discussion forums, assignments and additional course content [5]. The results of these log analyses can be used to improve educational effectiveness and support basic research on learning. Therefore, most researches focus on big data analysis of learning behavior and performance prediction in MOOC platform.

In order to understand students and improve the curriculum resources, many researchers at home and abroad try to establish learning behavior models to study the students' learning behavior in the MOOC platform, including

✉ Wanlin Gao
gaowlin@cau.edu.cn

✉ Minjuan Wang
minjuan@cau.edu.cn

Hui Hu
cauhuhui@cau.edu.cn

Guofeng Zhang
cauzhgf@cau.edu.cn

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China

[2] Key Laboratory of Agricultural Informatization Standardization, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100083, China

Published online: 07 January 2019

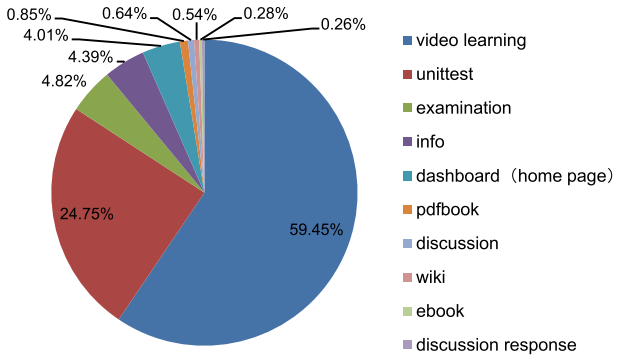

Fig. 1 Percentage of hits in the MOOC platform from 2014 to 2017

studying on the characteristics and classification, analyzing and predicting learning behavior, and the evaluation methods of MOOC learning, etc. In the research of student engagement, video watching is regarded as the most important way to disseminate knowledge [6–8]. As shown in Fig. 1, video watching has the highest popularity ratio compared with other learning behaviors. Typical measures for computing video watching data and MOOC analysis include the length of the video watched (LVW) and the retreat rate during video watching [9]. The length of the video watched means how much knowledge being get. The retreat rate focuses on how many students finished the video, that is, the audience coverage of the video. There are still some researches based on the questionnaire related to the students' self-evaluation [10, 11].

There are certain limitations in the research of student engagement based on the length of the video watched and the rate of finish watching. Unfortunately, the MOOC platform cannot log whether or not a student has left the video in the middle, resulting in the true video engagement time unknown. Moreover, the data collected also cannot show the behavior that the student did not pay attention to the video such as left the computer or do other things while the video is playing. Besides, the current researches of student engagement are less based on the background of big data and MOOCs.

Despite the challenges that MOOCs present, the data that these platforms collect bring substantial opportunities for studying the process of student learning. The backend infrastructures are driving them to capture detailed measurements on students as they interact with the different forms of learning integrated into the courses. For video watching, individual clickstream events are captured, with a click event generated and stored each time a student interacts with a video, specifying the particular action (e.g., pause, speed change), position and time at which it occurred [12]. The clickstream events are more detail to evaluate the student engagement and predict learning effect [13].

Based on these richly logged interactions of students' video watching behavior, this paper develops computational methods that answer critical questions about when and how long will students grapple with the video material. Detailed preprocessing of the raw data is introduced. The analysis algorithms based on Spark platform are given, which include the time distribution of video watching behavior, the video length distribution while the student pause the video, seek or change the playing speed. At last, we discuss and explore how to make video that meets the needs to improve the student engagement level of video watching.

2 Methods

The proposed structure for big data analysis framework in MOOC is presented in Fig. 2. The log models of the MOOC platform capture detailed measurements on students as they interact with the different forms of learning integrated into the courses by visiting the Web site on computer or using the phone app. Then, the Spark gets the log data from the database of MOOC platform. After preprocessing, calculation and analysis with Spark streaming, the results are put into the data warehouse for further research.

2.1 Data collection

The benchmark data used in this paper are exported from the database of cauX. The cauX dataset has 29,059,818 log data and contains 11 attributes, but association rule analysis only selects students' video watching data of the course

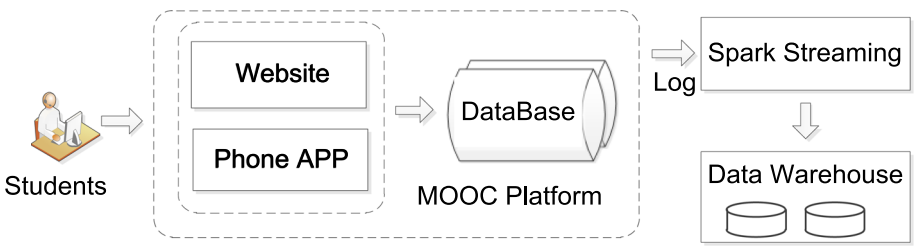

Fig. 2 Big data analysis framework

Table 1 Example data of the students' video watching behavior

ID	Username	Date	Page_adr	Page_action
21378121,,, 2015308080116,,, 2016/9/17 19:15:35,,, /courses/ciee/08112230_2016/2016_fal/courseware/fa5a9d4f8c844 5e82fc4421f4b02fbc/6b204e0b4d5848dfa140c0 f435d09c54/undefined/event,,, {"POST":{"event_type":["seek_video"], "event":["{\\"id\\":\\"i4x-ciee-08112230_2016-vi				
deo-ad49599ad793454cbbdd395892ef634c\\",\\"code\\":\\"html5\\",\\"old_time\\":416.38,\\"new_time\\":468,\\"type\\":\\"onSlideSeek\\ "}"],"page":["http://caux.cn/courses/ciee/08112230_2016/2016_fal/courseware/fa5a9d4f80c 8445e82fc4421f4b02fbc/6b204e0b4d5848dfa140c0f435d09c54/"]}, "GET": {}}				

named computer graphics set up in 2016 as a case study. Table 1 gives an example of data collected from MOOC platform. It includes the information of id, username, date, page address and page action which is in a comma-separated values format.

As for video watching behavior, page action includes hide transcript, load video, play video, pause video, speed change, sequence video, seek video and page close, as shown in Fig. 3.

2.2 Data preprocessing

This paper targets on the logs of students' video watching behavior in MOOC platform in the course of computer graphics set up in 2016 which from September 1st to December 31st. The raw data maybe have the mistakes caused by the anomaly. If these exceptions and errors are applied directly without analysis and filtering, it will directly or indirectly affect the accuracy of the results. Therefore, it is necessary to preprocess the raw data to ensure the accuracy of the results.

2.2.1 Computational tools—Spark

Researchers developed a specialized framework called Apache Spark to deal with the problem of slow data sharing in Hadoop MapReduce due to the replication, serialization and disk IO [14]. The resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines, is the architectural foundation of Spark [15]. Spark is the computational tools for data analysis in this paper. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS). HDFS is a distributed file system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.

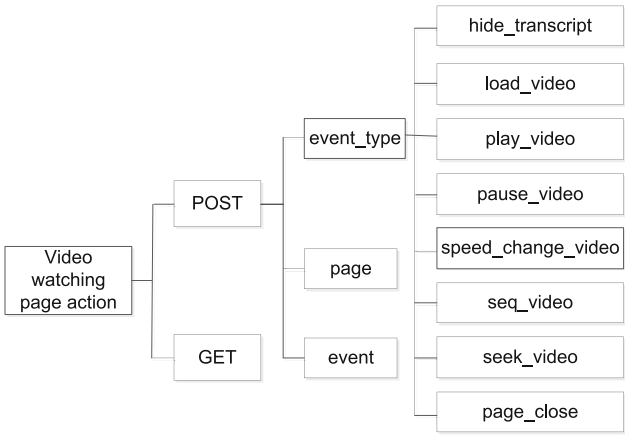

Fig. 3 Page actions in video watching

Figure 4 presents the data analysis process in Spark. Spark will automatically publish RDD to the cluster and execute RDD in parallel during the program running. Firstly, the program creates a series of RDDs by inputting the data of students' video watching behavior in MOOC Platform that upload to HDFS before. Secondly, it is the RDD transformation by calling the function of Spark to deal with RDD, such as map, flatmap and filter. Due to the read-only characteristic of RDD, every calculation makes a new RDD. Finally, the results are saved back to HDFS when the program is finished. Spark allows developers to write programs with Scala, Python and Java quickly. All the programs in this study are based on the Python language.

2.2.2 Standardize data

In this paper, the data of page action are stored as a string in the database which translated by JSON (JavaScript Object Notation). It should be formatted back to JSON before analysis, as shown in Fig. 5. Figure 5a shows the string format of the raw data; (b) JSON format after processed; and (c) a table for improvement. Firstly, reduce the unnecessary escape character, such as the double slash or quotation mark. Secondly, identify the type of elements using the function of type in Python. It can learn that the data are a nested structure of list and dictionary in Python. Then, it is easy for us to visit the value by key; for example, the value of old time is equal to page_action ['POST']['event'][0] ['old_time'].

2.2.3 Feature selection

Feature selection can drop the irrelevant or redundant features to reduce the number and select the optimal

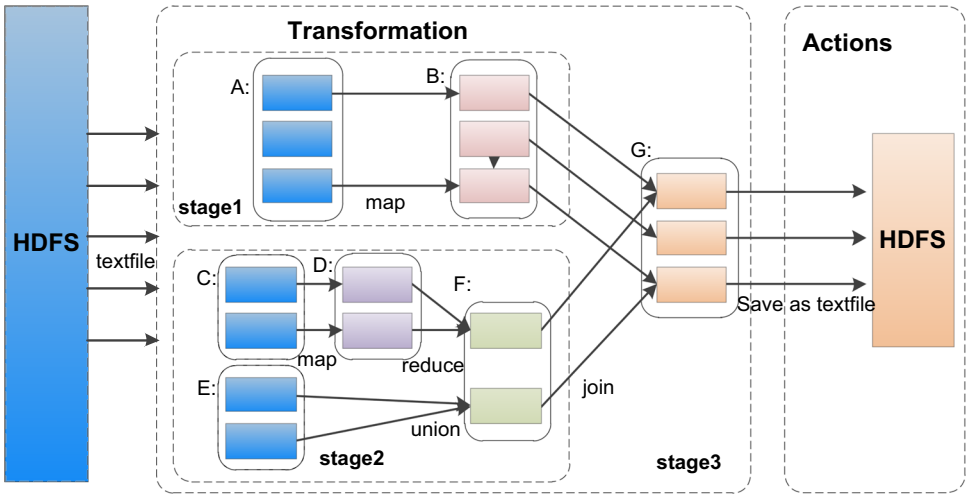

Fig. 4 Data analysis process in Spark

feature subset [16–19]. If there are N kinds of video watching behaviors then using the set of $\{x_1, x_2, \ldots x_n\}$ to describe. After feature selection, select K behaviors to optimize the analysis results and using the set of $\{x_1, x_2, \ldots x_k\}$ to describe; here k is smaller than n.

Table 2 describes the results after feature selection based on our data analysis. In this paper, we focus on the video watching behaviors of play, pause, seek, speed change, sequences video. The 'currentTime' attribute means the length video watched (LVW) while video watching behaviors occurred. The 'old_time' and 'new_time' are the video length before and after seek. The 'old' and 'new' represent the sequences number before and after sequences video.

{"POST":{"event_type":["seek_video"],"event":["{\\"id\\":\\"i4x-ciee-08112230_2016-video-ad49599ad793454cbbdd395892ef634c\\",\\"code\\":\\"html5\\",\\"old_time\\":416.38,\\"new_time\\":468,\\"type\\":\\"onSlideSeek\\"}"],"page":["http://caux.cn/courses/ciee/08112230_2016/2016_fal/courseware/fa5a9d4f80c8445e82fc4421f4b02fbc/6b204e0b4d5848dfa140c0f435d09c54/"]}, "GET": {}}

(a) The raw data — string format

{
"POST": {
"event_type": ["seek_video"],
"event": [
{"id": "i4x-ciee-08112230_2016-video-aD49599ad793454cbbdd395892ef634c",
"code": "html5",
"old_time": 416.38,
"new_time": 468,
"type": "onSlideSeek" }
],
...
}

(b) JSON format of (a)

event_type	id	old_time	new_time	...
seek_video	i4x-ciee-08112230_2016-video-aD49599ad793454cbbdd395892ef634c	416.38	468	...

(c) The meaning of (b)

Fig. 5 Example for the data preprocessing from string to JSON format

Table 2 Results after feature selection for video watching behavior

event_type	Attribute
play_video	ID, Username, Date, video_id, currentTime
pause_video	ID, Username, Date, video_id, currentTime
seek_video	ID, Username, Date, video_id, old_time, new_time
speed_change_video	ID,Username,Date,video_id, currentTime, old_speed, new_speed
seq_video	ID, Username, Date, video_id, old, new

2.2.4 Error analysis and processing

The log model of MOOC platform can record any information while the student triggered. However, some of the records are meaningless when analyzing video watching behavior. For example, the log model will record information when the students click the speed change button, although the speed the students choose is the same as before. In the seek video behavior, error data are obtained because the new time equals the old time, which means the length of video played is the same as before. Thus, those data described above should be removed before analysis.

Another error that affects the correct result is interference data. For the continuous seek video, the changes in every second are recorded and obtained by the log model. Therefore, we only take the first old time and the last new time to analyze in this preprocessing of continuous behavior. Table 3 shows an example of the output file after preprocessing.

Table 3 Example data of the output file

21378121,,, 2015308080116,,, 2016/9/17 19:15:35,,, seek_video,,, i4x-ciee-08112230_2016-video-ad49599ad793454cbbdd395892ef634c,,,416.38,,,468

and map to a dataset with the hour as key and set its value is 1. Then, the program calls the key value pair(Key, Value) and receives the returned dataset(Key, Value) aggregated by the reduce function.

Algorithm 1: Calculate the number of video watching behaviors in every hour

Input: text file
Output: text file
Parameters: text file path, output path

```
1). Read preprocessing data from HDFS；
2). lines = spark.read.text(sys.argv[1]).rdd.map(lambda r: r[0])
    counts = lines.flatMap(lambda x: x.split(' ')) \
                  .map(lambda x: (x, 1)) \
                  .reduceByKey(add)
3) output = counts.collect()
      for (time, count) in output:
          print("%s: %i" % (time, count))
4) spark.stop()
```

2.3 Data analysis algorithms

2.3.1 The time distribution characteristics of students' video watching

In this paper, we calculate the number of video watching behaviors in every hour. Algorithm 1 describes the detailed calculation process. We input the hour of date from HDFS

2.3.2 Video length distribution characteristics of students' video watching

Algorithm 2 describes the detailed calculation process of the video length distribution. We define 60 s as a period of time and then calculate the number of pauses, seeks and speed changes as the video is played.

Algorithm 2: Calculate the number of pauses, seeks and speed changes in every 60 seconds

Input: text file
Output: text file
Parameters: text file path, output path

```
1). Read preprocessing data from HDFS；
2) if event_type in [pause_video  seek_video, speed_change_video]
3).    def currentTime _res(x)
           return x%60+1
     counts = lines.flatMap(lambda x: x.split(' ')) \
                .filter(currenTime_res, currentTime)
                .map(lambda x: (x, 1)) \
                .reduceByKey(add)
4).  for (currenttime, count) in output:
           print("%s: %i" % (currenttime, count))
5). spark.stop()
```

Note that：Pause_video, seek_video, speed_change_video represent the pauses, seeks, speed changes behavior, respectively.

Note that:Pause_video, seek_video and speed_change_video represent the pauses, seeks and speed changes behavior, respectively.

2.3.3 Entropy weight method

Different video watching behaviors occurred in the different length the video watched (LVW). In this paper, the entropy weight method is used to calculate the weight of the frequency of the video watching behavior of pause video, search video and speed change video in the LVW. The flowchart of entropy weight method is shown in Fig. 6.

Step 1: Construct the relation matrix. Suppose the type of video watching behavior is m and LVW is divided into n periods, then the relationship matrix S is established. The k_{ij} indicates the frequency of i behavior occurred in j LVM.

$$S = \left[k_{ij}\right]_{n*m} \quad \text{for } i = 1, 2, 3\ldots n \text{ and } j = 1, 2, 3\ldots m \tag{1}$$

Step 2: Normalization is accomplished using the following equation recommended for x problem:

$$k'_{ij} = \frac{k_{ij} - \min\left(k_{ij}\right)}{\max\left(k_{ij}\right) - \min\left(k_{ij}\right)} \tag{2}$$

where $\max(k_{ij})$ and $\min(k_{ij})$ are the maximum and minimum values of kth sequence.

Step 3: The entropy of each video watching behavior is given by:

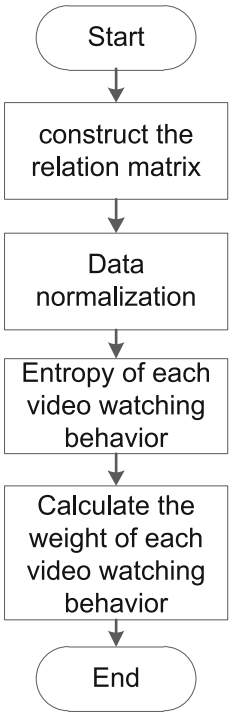

Fig. 6 Flowchart of entropy weight method

$$E_j = -\ln(n)^{-1}\sum_{i=1}^{n} p_{ij}\ln p_{ij} \tag{3}$$

where $p_{ij} = k'_{ij}/\sum_{i=1}^{n} k'_{ij}$; if $p_{ij} = 0$, then $\lim_{p_{ij}\to 0} p_{ij}\ln p_{ij} = 0$.

Step 4: The weight of each video watching behavior is then calculated as follows:

$$w_j = \frac{1 - E_j}{m - \sum E_j}\,(j = 1, 2, 3\ldots m) \tag{4}$$

Then, the frequency of different LVW is given by:

$$F_i = \sum_{j=1}^{m} k_{ij} w_j \tag{5}$$

3 Results and discussion

Figure 7 shows the experimental results of Algorithm 1. The results of time distribution characteristics of students' different video watching behaviors are illustrated in Fig. 6b–f. The experiments show that the time period in which the video watching page has the highest click rate is from 8 am to 10 am and reaches the peak at 9 am. The next rapid growth period is from 7 pm to 11 pm and then decreases after midnight. It has the same trend of change for specific video watching behavior. However, the difference is that the peak at night is greater than during daytime on the behavior of play video, pause video and sequence video.

Figure 8 shows the experimental results of Algorithm 2. According to the experimental results of Algorithm 2, it is clear that the number of video watching behaviors occurred in every minute of the video decreases with the increase in the video play time. Figure 8a shows the experimental results of the feature values of all video watching behaviors occurred in every minute of the video. Figure 8b shows the number of pause video, seek video and speed change video occurred in every minute during the video played. Overall, the largest number appears within 60 s while remaining stable from 60 to 300 s and decreases rapidly after 300 s, which is the same as the number of seek video. But, the number of pause video and speed change video starts to decrease after 60 s; especially, the speed change video is the fastest and almost zero after 60 s, as shown in Fig. 8b.

Furthermore, we have studied the frequency and accumulative frequency of the 'currentTime' after calculated the weight of pause video, seek video and speed change video based on the entropy of information principles [12], which is 0.39, 0.44 and 0.17. The result is shown in Table 4; it can be seen that every 60 s has the frequency greater than 0.1 and the accumulative frequency is up to 71% within 360 s.

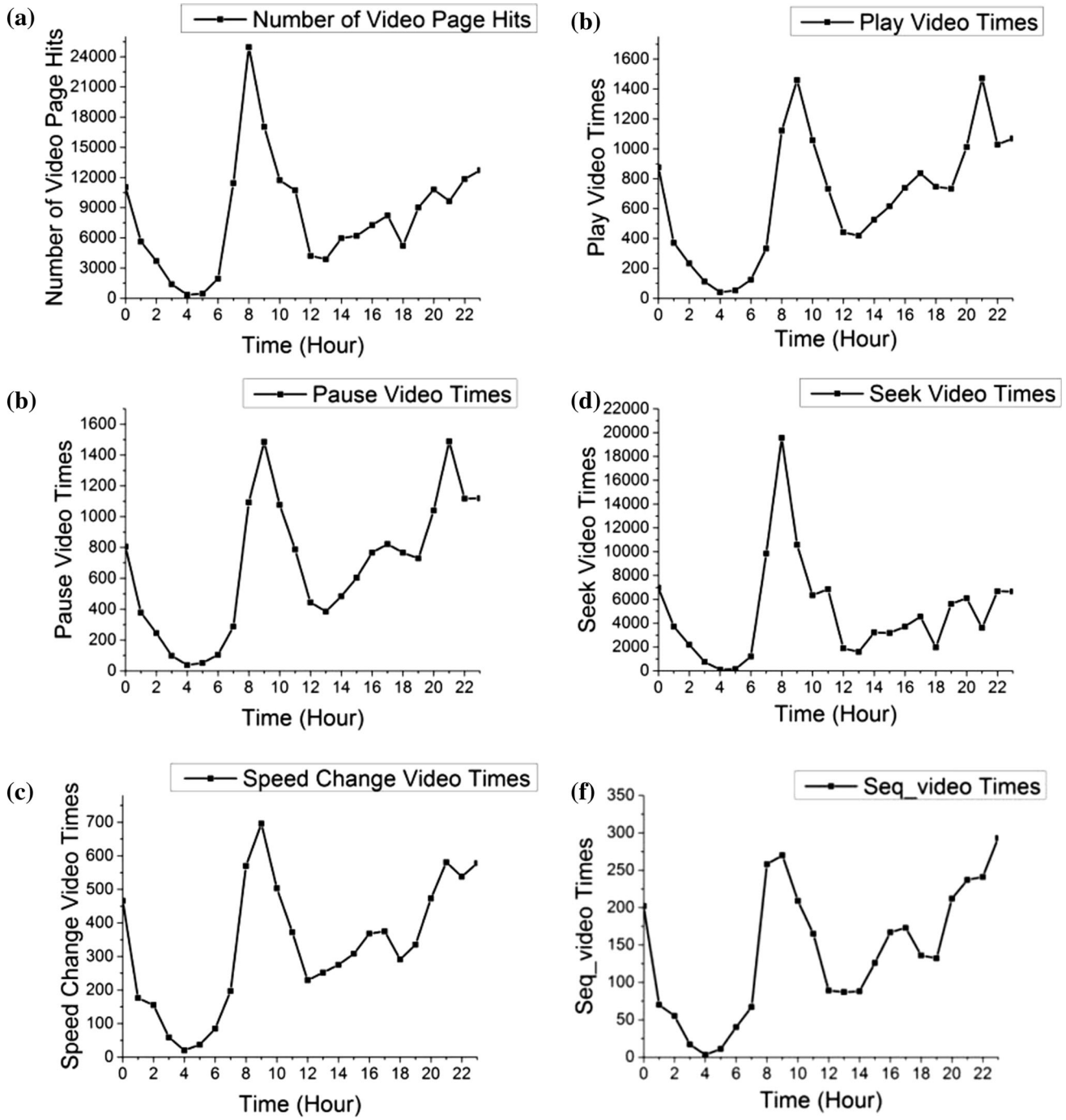

Fig. 7 Feature values of video watching behaviors in every hour

Most of the ideas presented in this work can be extended to other courses. For example, Algorithm 1 is designed to help explore and understand the video watching behavior of time distribution characteristics in MOOC. It is easy for course instructors to know each student's activity of the video watching every month and every day and take it as part of the grade. Besides, MOOC platform owners can forecast workload and adjust the load balance of the platform. Algorithm 2 describes the detailed calculation process of the video length distribution. The results can help the course instructors to arrange a reasonable video length to attach more students' attention and reduce the drop rate of video watching.

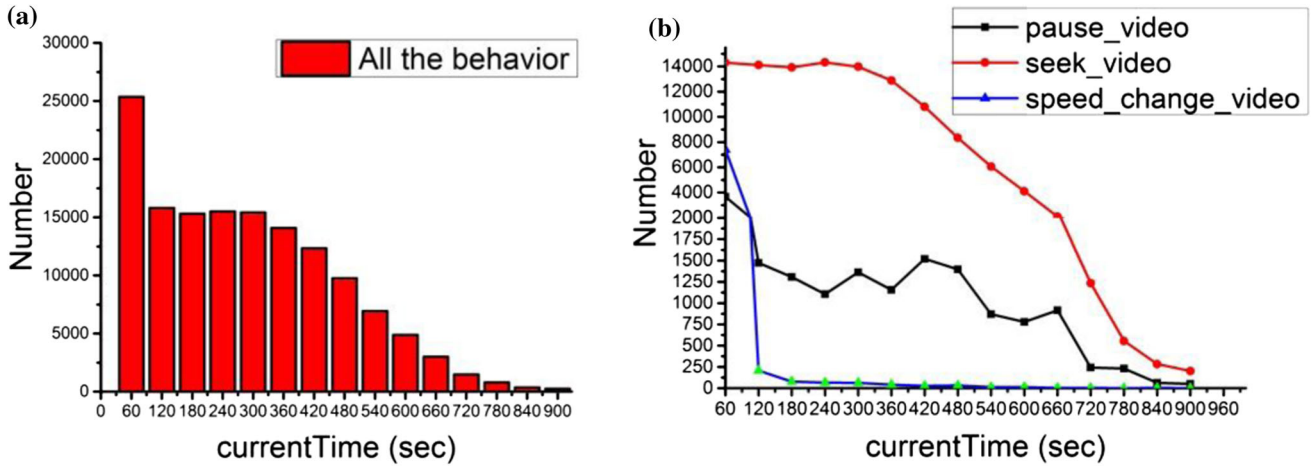

Fig. 8 Feature values of video watching behaviors occurred in every minute

Table 4 Frequency and accumulative frequency of video watching behaviors occurred in every minute of the video played

currentTime (s)	Frequency	Accumulative frequency	currentTime (s)	Frequency	Accumulative frequency
0–60	0.130538	0.130538	480–540	0.05185	0.921651
60–120	0.117992	0.24853	540–600	0.036035	0.957686
120–180	0.115694	0.364224	600–660	0.020911	0.978596
180–240	0.117812	0.482036	660–720	0.0109	0.989496
240–300	0.116426	0.598462	720–780	0.005515	0.995011
300–360	0.106787	0.705249	780–840	0.00253	0.997541
360–420	0.092215	0.797464	840–900	0.00184	0.999381
420–480	0.072337	0.869801	900 ~	0.000619	1

4 Conclusions and future work

This paper proposes the students' video watching behavior analysis methods in MOOC based on Spark platform. Contributions include the methods have been used to analyze the data and the behavior of video watching, and the algorithms have been used for data preprocessing and calculating the number of video watching behavior using Spark platform in every hour and every 60 s, and the entropy weight method has been used to calculate the weight of different video watching behaviors.

The experimental results show that there are two peaks in video watching behavior: One is from 8 am to 10 am and the other is from 7 pm to 11 pm. Besides, the number of pauses, seeks and speed changes in every minute during video watching is the most in first 60 s, then remaining stable between 60 and 300 s and rapidly decreases after 300 s.

In future, we plan to recommend development models for the video watching behavior prediction and video design.

Acknowledgements The authors would like to thank their colleagues for their support of this work. The detailed comments from the anonymous reviewers were gratefully acknowledged. This work was supported by the Beijing higher education and teaching reform project in 2014 (No. 2014-ms044).

Compliance with ethical standards

Conflict of interest The authors declared that they have no conflicts of interest to this work.

References

1. Kahl MP (2015) An overview of the world of moocs. Proc Soc Behav Sci 174(1):427–433
2. Brinton CG, Chiang M (2015) MOOC performance prediction via clickstream data and social learning networks. In: 2015 IEEE conference on computer communications (INFOCOM), pp 2299–2307. IEEE
3. Li X, Chen Y, Gong X (2017) MOOCs in China: a review of literature, 2012–2016. In: New ecology for education—communication X learning, pp 21–32. Springer, Singapore
4. Sun Xiaoyin, Zhou Wei (2017) big data analytics technology based on MOOC. Comput Mod 4:89–93

5. Hmedna B, El Mezouary A, Baz O (2017) An approach for the identification and tracking of learning styles in MOOCs. In: Europe and MENA cooperation advances in information and communication technologies, pp 125–134. Springer, Cham
6. Chen CJ, Wong VS, Teh CS, Chuah KM (2017) MOOC videos-derived emotions. J Telecommun Electr Comput Eng (JTEC) 9(2–9):137–140
7. Li Manli, Shunping Xu, Sun Mengliao (2015) Analysis of learning behaviors in MOOCs—a case study of the course "Principles of Electric Circuits". Open Educ Res 21(2):63–69
8. Johnson L, Adams Becker S, Estrada V, Freeman A (2014) The NMC horizon report: 2014 higher education edition. Austin, Texas
9. Slemmons K, Anyanwu K, Hames J, Grabski D, Mlsna J, Simkins E et al (2018) The impact of video length on learning in a middle-level flipped science setting: implications for diversity inclusion. J Sci Educ Technol 27(5):469–479
10. Zhang H, Huang T, Lv Z, Liu S, Zhou Z (2017) MCRS: a course recommendation system for MOOCs. Multimed Tools Appl 77:7051–7069
11. Agnihotri L, Mojarad S, Lewkow N, Essa A (2016) Educational data mining with Python and Apache spark: a hands-on tutorial. In: Proceedings of the sixth international conference on learning analytics and knowledge, pp 507–508. ACM
12. Chen Kan, Zhou Yaqian, Ding Yan et al (2016) Research on learning engagement of online video: analysis on the relation between MOOCs video features and seek behavior while watching. J Distance Educ 34(4):35–42
13. Sinha T, Jermann P, Li N, Dillenbourg P (2014) Your click decides your fate: inferring information processing and attrition behavior from mooc video clickstream interactions. arXiv preprint arXiv:1407.7131
14. Pandey SC (2018) Recent developments in big data analysis tools and apache spark. In: Big data processing using spark in cloud. Springer, Singapore, pp 217–236
15. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: cluster computing with working sets. In Proceedings of the 2nd USENIX conference on hot topics in cloud computing (HotCloud'10). USENIX Association, Berkeley, CA, USA
16. Zhu X, Suk HI, Wang L, Lee SW, Shen D (2017) A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. Med Image Anal 38:205–214
17. Zheng W, Zhu X, Wen G, Zhu Y, Yu H, Gan J (2018) Unsupervised feature selection by self-paced learning regularization. Pattern Recogn Lett. https://doi.org/10.1016/j.patrec.2018.06.029
18. Zhu Xiaofeng, Zhang Shichao, Rongyao Hu, Zhu Yonghua, Song Jingkuan (2018) Local and global structure preservation for robust unsupervised spectral feature selection. IEEE Trans Knowl Data Eng 30(3):517–529
19. Zheng Wei, Zhu Xiaofeng, Zhu Yonghua, Rongyao Hu, Lei Cong (2017) Dynamic graph learning for spectral feature selection. Multimed Tools Appl 11:1–17