Big Data Analysis in Video Websites: Evidence from Bilibili and YouTube

Siyu Wang

School of Economics and Management, Tianjin University of Science and Technology, Tianjin, China

s.wang.64@student.scu.edu.au

Abstract. Contemporarily, society enters the era of high-tech information technology, and big data is the product of the rapid development of information technology. Big data with the characteristics of Volume, Variety, Velocity, Veracity and Value plays an important role in various industries. It optimizes business processes, improves business quality, guides industry innovation, and ultimately provides convenience and better services for mankind. The development of Internet technology, communication technology, data transmission technology and smart terminal technology has contributed to the rise of video websites. More and more people choose to watch video in the websites to entertainment (e.g., YouTube, Tik Tok and Bilibili). Personal recommendation is one way to get viewers to stay and enhance viewers stickiness. In this paper, big data and video websites are linked. The role played by big data techniques in personalized push and provide differentiated services for viewers of video websites will be discussed. In addition, the limitations of the state-of-art big data technologies combined with the development status of video websites are analyzed. Overall, these results shed light on guiding further exploration of video websites using big data techniques to improve the business and user experience and find future directions.

Keywords: Video Websites, Bilibili, YouTube, Big Data.

1. Introduction

Big data, also known as massive data, is a data collection composed of huge amount, complex structure, and numerous types of data. The amount of data involved in it is too huge to be captured, managed, processed and sorted into information. The analysis of data can help government agencies and enterprises to manage and make decisions in a reasonable time through current mainstream software tools [1]. Data has always existed in human life and developed with the development of human society. Techniques for acquiring and analyzing data are also constantly being updated and improved. The first recorded statistical data analysis check came from John Grant, who in 1660 collected mortality data to understand how the plague was spreading and to create an early warning system to protect people from this disease [2]. In 1965, the United States developed the first National Data Center, an early federal cloud computing solution designed to manage government records from a central location. Thus, the path to the modern data analytics market can be traced back to 1970, just after ARPANET sent its first "email" from UCLA to Stanford University [2]. Most scholars believe that the concept of "Big Data" first appeared publicly in 1998, when John Mashey, chief scientist of SGI, an American high-performance computing company, pointed out in an international conference report that with the rapid growth of data volume, there would be four challenges, i.e., difficult to understand, difficult to access, difficult to process, and difficult to organize data. The term "Big Data" is used to describe this challenge, which causes thinking in the field of computing [3]. Nowadays, big data has played a huge role in medical care, commerce, industry, finance, transportation, education, and other industries. These industries have been shaped by big data analysis to become top industries.

Ushapreethi and Lakshmipriya summarizes and describes the background of common methods in large video storage and retrieval based on visual content. It additionally provides a systematic classification of the latest technical work on video big data [4]. Feng and Xiao put forward a scalable video and data monitoring and management platform for interactive video websites to data management and marketing methods in the era of big data, and discuss the platform to realize the key

technology involved [5]. Reddy and Murali explain in detail how YouTube use HDFS (Hadoop Distributed File System) and MapReduce to realize big data analytics [6]. Rahman et al. founds big data applied by YouTube can feedback about data related to the channel, intuitively reflect data performance, provide a basis for related research around the channel [7].

In the era of big data, big data technology has helped many industries to collect relevant data and provide better services to users and provide guidance about directions of future development of the industry. On this basis, this study is aimed to research the role of big data technology applied by video websites in user service. This paper firstly introduces big data and big data analytics, and then takes YouTube and Bilibili as cases to study the use of big data in user services. Finally, it analyzes limitations of big data analysis and give the outlook.

2. Description of big data analysis

Big Data refers to data sets that are huge in volume and cannot be captured, managed, and processed with conventional software tools within a certain time range [8]. Big data has five characteristics which are also known as 5V's, i.e., Volume, Variety, Velocity, Veracity and Value. Volume refers to the amount of data collected, stored, and analyzed by big data is large. It generally takes P, E and Z as units of measurement. Variety refers to the sources of data and the types of data are various. Data collected by big data consists of structured, semi-structured and unstructured data. Velocity means the speed of processing data is fast. The efficient processing of data is convenient for people to quickly find useful information from many complex and scattered information. This is one of the advantages of big data [8]. Veracity refers to the reliability of data. As the unreliability inherent in some data, big data needs to use tools and analytics developed for management and mining of uncertain data to deal with impreciseness and uncertainty, which is another facet of big data [8]. For the value, the meaning needs to be analyzed from two aspects. Based on Oracle's definition, big data has a characteristic of relatively "low value density", since only a small fraction of huge-volume data collected by big data is useful. Therefore, the data received in the original form usually has a low value relative to its volume [9]. However, the value will be high for the users if using big data to analyze data efficiently.

Big data analytics examines large amounts of data to uncover hidden patterns, correlations, and other insights. Nowadays, many organizations collect data in real time and use big data analytics to mine significant value from data they get and apply data to make instant and advantageous decisions. Meanwhile, they have chosen big data analytics to collect and analyze data form their clients about the behavior pattern so that they can provide their clients with more personalized services or products, which is beneficial to attract and retain clients. The common methods of big data analytics are the following four big types, i.e., descriptive, diagnostic, predictive and prescriptive. Descriptive big data analytics explains what happened in the past by using existing data from the reports or the diagrams. It's a direct acquisition of data. Nevertheless, it doesn't explain the reason of past events or what will occur in the future. Diagnostic big data analytics has a further step based on descriptive analysis. It evaluates data captured by descriptive analysis and find out the why the past events happened [2]. It helps to analyze data deeply and approach the core of data. Predictive big data analytics uses predictive models to forecast the possibility of an event that may occur in the future, a quantifiable value, the time point when the event occur by analyzing a set of co-dependent variables [2]. It plays an important role when the future has uncertainty because it helps to make better decisions. Prescriptive big data analytics can't be inseparable from the previous three analyses. It based on the analyses of what happened, why it happened and what may happen in the future, to guide what actions should be taken and what decisions should be made. Typically, this is carried out last after the previous three analyses have been completed. With the development of information technology and cloud computing, big data emerges and prospers. It represents a new stage of information process. Big data has been applied in many fields such as media, business, and health care. Especially, it drives

e-commerce and new retail in business. It plays a significant role in the prosperity of digital economy and Internet Plus economy.

3. Bilibili

Bilibili is a famous video website in China. It's a platform where users can watch various videos including copyrighted OGV likes anime, TV series, movies, UGC and PUGV. People who upload the videos with UGC and PUGV are called "uploader" in Bilibili. It uses big data analytics in personalized push to analyze the information of videos, uploaders, and video viewers so that it can connect the users and videos they prefer. The connections can enhance user stickiness as users' interests are well conformed. To achieve personalized push, big data analytics focus on two angles, i.e., viewers and videos.

Analyses on viewers uses data sets related to the basic personal information and browsing behaviors. According to the 2020 Q3 financial report of Bilibili, the average daily active users of Station B reached 90.3 million and the average monthly active users reached 333 million in the third quarter of 2022. The volume of data from viewers are huge. The basic personal information is collected when users register in Bilibili, as they need to fill in basic personal information such as age, gender, and location. When viewers watch various videos, the collection of their browsing behaviors simultaneously starts. It records the videos viewers watch, duration of viewing of each video and the actions the viewers take. Then the information is sent to the analytic center. These videos are be classified to different zones based on their theme. Viewers' preference about zones can be analyzed through the number of videos and duration of viewing. Besides, it can know the preference by the actions. If viewers take positive actions to the videos such as like, good comment and bookmark, the analytic center will push more similar videos on the homepage. If viewers take negative actions such as choose "unlike" and block the keywords or the zone of the videos, Bilibili will reduce relevant push to avoid damaging viewers' interests.

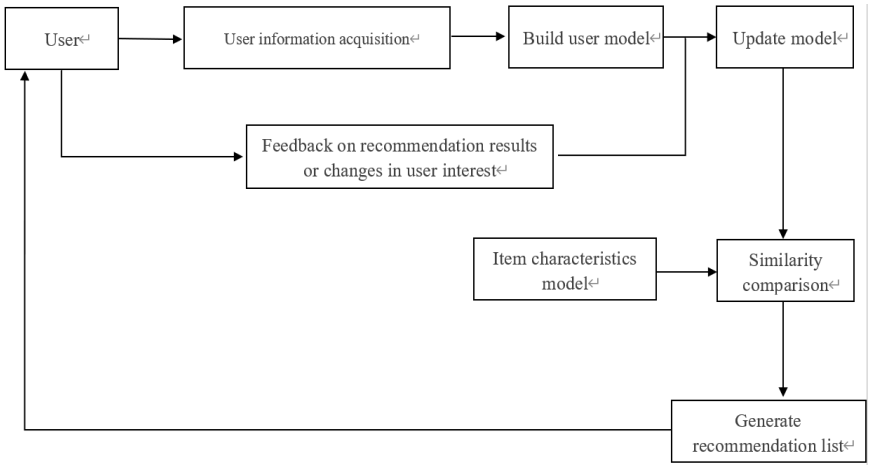


Figure 1. Content-based recommendation algorithm

Analyses on UGC videos are concentrated on the contents, the labels and zones and the data performances. When uploaders upload the videos, they need to choose related labels for their videos and decide which zone the video should be posted. Subsequently, the analytic center group individual videos together according to the labels and zones, which is convenient to push them to specific viewers. The data performances are collected as well. Types of data often collected is the number of likes, bookmarks, and coins. The coin is not real money. It's a prop that expresses approval of the video. Videos with better data performance are more likely to get pushed by the center because they tend to be of higher quality. The base method of video analysis is applied, which is called "video indexing". Indexing is the process of sorting the video content for dynamic and efficient video storage and retrieval. Conventional video indexing methods are utilized for creating video indexing guidelines for the data set CORPORA [4]. Both static background region identification and dynamic

motion of objects are considered for video indexing. factors such as viewer response, variations among viewers and data set creation are best for the effective indexing [4]. Big data connect UGC, PUGV videos and copyrighted OGV videos. When viewers search and watch copyrighted OGV videos, related UGC and PUGV videos will be push as they contain the keywords of the OGV videos (a systematic push). A sketch of content-based recommendation algorithm the is shown in Fig. 1.

After a thorough analysis of the video and the viewers, perspective big data analytics starts to help the completion of the link between videos and viewers. There are two ways to achieve the link: collaborative filtering between viewers and videos and collaborative filtering between viewers and viewers. Collaborative filtering-based recommendation algorithm is utilized to implement the link. The schematic diagram of this algorithm is illustrated in Fig. 2. The collaborative filtering between viewers and videos is a two-way selection. Individual videos in the same zone which is preferred by relevant users will be pushes to these viewers. The coordination between viewers and viewers is a result of group analysis. Some videos will be pushes because they are watched by a group of viewers who has similar basic personal information or browsing preference. The link between the videos and viewers is helpful to increase users' loyalty because their can find videos that satisfy their interests. As a result, users have always maintained strong interest and enjoy personal service in Bilibili, which shows the characteristic "value" for big data.

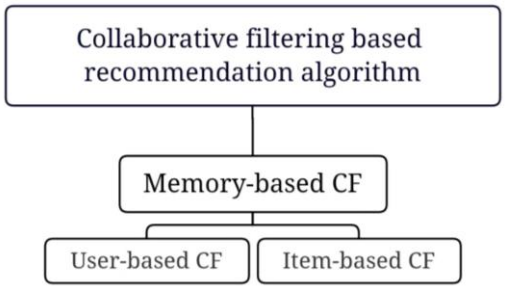


Figure 2. Collaborative filtering-based recommendation algorithm

4. YouTube

YouTube is a video sharing service that allows users to watch videos posted by other users and upload videos of their own. The contents of videos in YouTube are consisted of UGV, OUGV and a small number of OGV. People who upload videos are called YouTuber YouTube creates data sets by collecting variables. Variables including viewing habits, watch time, number of views, number of "likes" and "dislikes" for each video, and the number of channel "subscribers" constitutes as big data. It is used to provide more insight to help understand how the viewers use the videos and their perception to guide the design and development of future content [6].

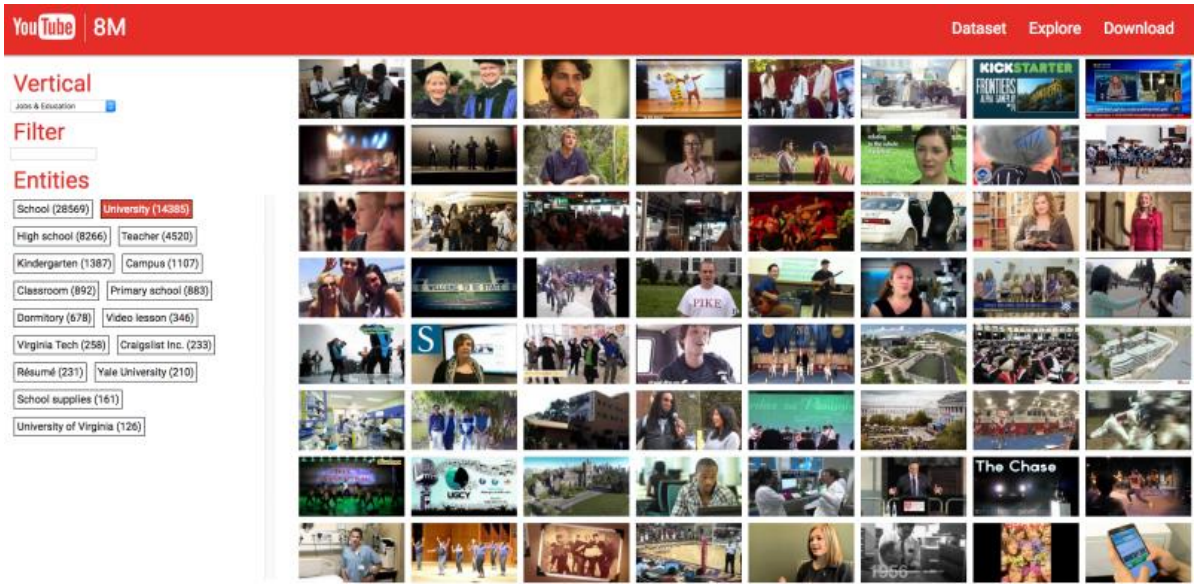


Figure 3. The application of YouTube-8M

Descriptive analyses are used to explore the data and trends in demographics and engagement on the channel. Continuous data (view duration) is presented as both medians and ranges, and means and standard deviations. Linear regression analysis was used to create a regression model to correlate impression click-through rate with the average view duration. Statistical analysis was performed on Microsoft Excel [6]. The analyses base on two types of data: personalization of viewers and performance of videos.

YouTube knows viewer's personalization through the videos they watch and keywords they put in the search boxes. It records the duration of viewing, theme and channels and viewers' interaction of videos they watch. Additionally, it continuously follows up the behavioral change of viewers, e.g., the frequency of viewing to specific theme and channels, to grasp and the change of viewer's interests and change the push mechanism. YouTube collects all data about each video. Average viewing time and number of times, number of subscriptions, number of comments and other data can help find the appropriate push mechanism for videos. Furthermore, it links videos of different channels according to the theme, which aims to reach associated recommendations. YouTube has introduced its the largest multi-label video classification dataset "YouTube-8M". It can label each video based on their topic [10]. The implementation of "YouTube-8M" on YouTube is exhibited in Fig. 3.

After analyzing all data about viewers and videos, YouTube needs to predict which videos the viewers will click in to watch. Based on the present data, videos which have themes or are in channels watched or searched frequently by viewers will be push on the homepage to attract viewers. Videos of the YouTubers followed by viewers will also be push on the homepage. Meanwhile, it modifies its push mechanism when it realizes the change viewers made through the change of data. It timely pushes related videos that conform to viewers' new interaction. The flow chart of the recommendation system is shown in Fig. 4. To satisfy the search requirements, YouTube improves its search mechanism. YouTube collects the keywords of the videos' tittle. Hence when the viewers enter words, it will respond viewers by relevant videos it has gathers. YouTube uses big data analytics to satisfy viewer's needs. The change of viewers' interaction can feedback and be collected, consequently it can change the directions of analysis on viewers' preference and design future content better. This two-way mechanism is conducive to providing better-quality and more differentiated services to each viewer.

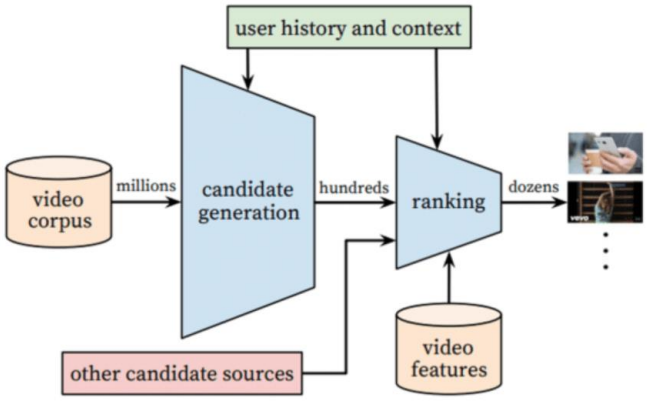


Figure 4. Recommendation system architecture demonstrating the "funnel" where candidate videos are retrieved and ranked before presenting only a few to the user

5. Limitation & Prospects

The limitation of big data is mainly in three aspects, i.e., storage, disposal, and security. As the clients of video websites has been calculated in hundreds of millions. Data from clients is not only large but also increasing in real time. It requires huge data storage space. Besides, the increase in fragmented data makes it difficult to store because it is unstructured. These unstructured data often have more detailed knowledge, but also more data noise, which makes it difficult to analyze and clean the data [11].

In disposing data, there are two problems: speed and non-determinacy. Velocity is one of the characteristics of big data. It requires the validity of data to be within the range of seconds. In this time range, big data processing equipment needs to calculate a large amount of data, process and analyze the data, and finally put the answers obtained into production activities [11]. However, due to the limitation of existing big data analysis technology, this requirement is difficult to achieve, which limits the application of big data to some extent. Data increment under big data is huge, but the level of data technology can't keep pace with its increment. Even though the level of data technology is improving rapidly, relative to the data growth rate is still lagging [12]. Big data has non-determinacy. The uncertainty of data itself brings difficulties to the learning and prediction application of big data [11]. Even though viewers can enjoy relatively personal push, they still find the contents of the push are not always what they want. This is because the predictive analysis has uncertainty as it can't ensure what will happen in the future will be as prediction, and there is a certain one-sidedness in the prediction method based on the existing data. In this case, viewers often find that the push doesn't exactly fit all their interests.

Data sets collected by big data are rich in content and high in value. For this reason, these data are easy targets for theft and attack. As a consequence, security of data storage is vital as it contains private information of clients. In September 2022, TikTok is attacked by hijack and the information of tens of thousands of accounts was leaked for improper profit. Besides, security of data usage needs to be paid attention. On account of the nature of big data, clients will open a lot of permissions to the platform. However, clients now are worried about the security of data usage as some data without their authorization has been collected. The data collected is used for other purposes without their knowledge. In the first half of 2018, the world's most famous social network company, Facebook, suffered the biggest crisis of trust in its history. Globally, more than 50 million Facebook user data were leaked, and third-party data analysis institutions used the data for business analysis activities, even involving the 2016 US election. If security of data storage and usage is damage, the video websites will be boycotted and suffer from a crisis of confidence.

In the future, video websites should create sufficient data storage space to handle the high-volume data brought by users. They need to keep update big data analysis technology which can help in disposal of unstructured data, huge data increment and giving full play to the timeliness of big data.

Furthermore, these website needs to build security protection measures for the existing database, take actions in preventing information leakage, correctly use the user's open rights, and do not abuse big data technology.

6. Conclusion

In summary, this paper investigates how video websites use big data to provide viewers with services. Taking two representative companies (i.e., Bilibili and YouTube) in the industry as the specific research objects, video websites use information from viewers and videos to create data sets and analyze viewers and videos separately through big data analytics. Analyses are performed in order to predict viewers' viewing trends and link videos and viewers. It's conducive for the websites to provide differentiated services such as personal push and it can guide the contents the websites should design and develop in the future. Further, it helps enhance user stickiness. Based on the analysis, it encourages video websites to refine their own data sets and constantly update big data technology so that they can bring better user experience. Nevertheless, the research in this paper mainly focuses on how video websites use big data to provide personalized recommendations in differentiated services and the significance. The way to implement big data to improve other services of video websites and service viewers better still needs further research and explorations. Overall, these results offer a guideline for video websites using big data to provide higher-quality and differentiated services and perfect business contents.

References

[1] Chao Y. Overview of Big Data development history [J]. Contemporary Economy, 2015 (8): 13-15.

[2] Wang H, Ji S. Research on the real-life application and development trend of big data analytics [J]. Information Network Security, 2021.

[3] Hong M. Big data development status and future trends [J]. Transportation Research, 2019, 5(5): 1-11.

[4] Ushapreethi P, Lakshmipriya G G. Survey on video big data: Analysis methods and applications [J]. International Journal of Applied Engineering Research, 2017, 12(10): 2221-2231.

[5] Feng Z, Xiao K. Data Management and Marketing Methods of Interactive Video Websites in the Era of Big Data [J]. Mathematical Problems in Engineering, 2022, 2022.

[6] Reddy L C S, Murali D. YouTube: big data analytics using Hadoop and map reduce [J]. International Journal of Engineering & Technology, 2018, 7(3.29): 12-15.

[7] Rahman N A, Ng H J H, Rajaratnam V. Big Data Analysis of a Dedicated YouTube Channel as an Open Educational Resource in Hand Surgery [J]. Frontiers in Applied Mathematics and Statistics, 2021, 7: 593205.

[8] Cao Y. Development of Big Data and Technology Applications [J]. Communication World, 2019, 1.

[9] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics [J]. International journal of information management, 2015, 35(2): 137-144.

[10] Abu-El-Haija S, Kothari N, Lee J, et al. Youtube-8m: A large-scale video classification benchmark [J]. arXiv preprint arXiv:1609.08675, 2016.

[11] Shi W. Limitations of Big Data [J]. Hebei Enterprises, 2018, 8.

[12] Diao S, Yao Z. On the Limitations of Big Data Thinking and beyond [J]. Natural Dialectics Research, 2017, 5: 8.