Mobile Terminal Advertising Video Acquisition and Classification

Yingshuo Jiang[1,a], Li Shi[2,b], Zhaoxin Zhang[1,c], Yongdong Xu[1,d,*]

[1]School of Computer Science Harbin Institute of Technology Weihai, China
[2]National Computer Network Emergency Response Technical Team Coordination Center of China Beijing, China
[a]jiangyingshuo@163.com, [b]shili@cert.org.cns, [c]heart@hit.edu.cn, [d]ydxu@hit.edu.cn
** Corresponding author*

Abstract: *As the network era develops rapidly, the information, obtained from the network is becoming more and more abundant, mainly by the web crawler. Information in addition to the fixed Internet, mobile Internet also contains enormous information. Recently, mobile short video carries less value than traditional one, which is called a new boom. Compared with traditional network video, short video carries less information value, while its advertising and marketing value are higher through further split propagation of social media. Many enterprises choose to publish advertisements on short video platforms for higher revenue with lower cost in order to lure users into using their products, but these ads might be risky. In this study, the crawler on the mobile Internet is realized through the network packet capture technology. The short video platform advertisements are collected and classified to provide data support for the follow-up advertising monitoring work. Finally, the ads are divided into two kinds, regular and risky through discrimination, which provides data support for the follow-up supervision.*

Keywords: *Web crawling, mobile Internet, short video, advertising*

1. Introduction

With the rapid development of the Internet, the information is getting more and more massive. Computer programs are needed to conduct any large-scale processing of web pages, requiring the use of a web crawler at some stage to fetch the pages to be analyzed[1]. A web crawler, robot or spider is a program or suite of programs that is capable of iteratively and automatically downloading web pages, extracting URLs from their HTML and fetching them[2]. Information not only occurs in lots of webpages but also the mobile devices, such as APP used in our daily life. The mobile Internet, defined as wireless access to the digitized contents of the Internet via mobile devices[3]. The number of people using the mobile Internet already exceeds those using the stationary Internet in Japan[4]. In South Korea, the number of people owning a mobile phone is 29 million (64% of the total population), the number of mobile Internet subscribers is estimated to be about 18 million (39% of the total population), and more than 3.5 million people are already using a 2.5G mobile Internet service, CDMA-1x, with a speed of 2.4Mbps[5].

In the mobile Internet age, short videos have become an important carrier for the generation and dissemination of network consensus with the help of mobile APP platforms. The most prominent feature of a short video is that the video length is no more than 20 minutes, which can be shot and uploaded by myself for sharing. Compared with TV series and movies, short videos satisfy users' need more intuitively and are concise, exciting and vivid[6]. These apps, typified by Douyin and Kwai, have become an essential medium for people to obtain information and pay attention to hot social spots[7]. The platforms include two kinds of advertisements, one is hard ads such as screen opening advertisement and information flow advertisement, the other is implanted advertisements from the perspective of commercial ads. The creators embed them in the video contents to impact users' consumption intention through the information obtained by video display. Short video ads have become a new traffic pool by relying on fragmentation and immersive user experience, precise push notifications based on users’ preferences, and social trust by interactive sharing[8]. However, some advertisements are for public welfare, but others release some illegal advertisements which exist problems such as exaggerating interests, concealing risks, tricking users and the public issues, inveigling users into downloading and using by the eye-catching advertising content. The acquisition of advertising video is essential so as to make better analysis and judgment of themselves.

Mobile information is mainly acquired through packet grabbing, which means that the computer intercepts and edits the data sent and received by the mobile devices. In addition, it can also be used to check network security. The current packet grabbing software mainly includes Fiddler of Windows and Charles of MacOS. The Fiddler used in this paper is a typical desktop tool with powerful functions, which can be applied to both Web browser client and mobile phone APP[9-11]. The data can be obtained browsed continuously so as to turn the pages of the APP by simulating the operation of phones. Then the judgement that whether the obtained video is an ordinary one or advertisement through the received field is conducted. Subsequently, information is extracted for getting the extracted AD title, link, type, and MP4 file of AD video.

The obtained text information of the advertisement video, which contains relevant information, may not be enough for the subsequent identification. Therefore, we’d better also extract the words for next analysis. Firstly, we built a liable lexicon by getting the sensitive thesaurus in the text information of ads artificially, and then continuously expanded it according to the subsequent data acquisition. We determined the category of the ad in the process analysis, based on the standard that whether it is a risky or regular, which is a matchmaking between the title and texts of the advertisement and the lexicon

2. Materials and Methods

The overall framework proposed in this paper mainly includes five processing stages: 1) configuration of the packet grabbing configuration on the computer and mobile; 2) Capture the video of the mobile and save the data packet to computer through the software; 3) Write scripts to process the packet and extract advertising information so that save it into the database; 4) Extract text information from the video; 5) Classify advertisements; 6) Problem analysis. The overall process is shown in Figure 1, The gray areas circled in the figure represent all the processes from capturing to sorting.

2.1. Environment Configuration of Packet Grabbing

This paper adopts Fiddler for packet grabbing. Android emulator is adopted on the mobile, and short video software such as Douyin is installed on it. The research is started by configuring Fiddler on the computer. In the Options section of Tools, Ignore Server Authentication Errors is selected and then Actions is clicked. When configuring remote links, the set should be selected with Allow Monitoring, and then the port can be set arbitrarily and the default is 8888 which needn’t be set. Subsequently, the mobile terminal and the computer terminal should be placed under the same LAN: obtaining the IP address of the computer, configuring the connected Wi-Fi on the mobile, selecting the manual agent, and then inputting the IP of the computer and setting the port number as 8888. After the agent is finally set up, enter "PC IP:8888" in the browser to open the Fiddler page, and then click the Fiddler Root Certificate to install it. The mobile phone will consider the environment unsafe if the certificate is not installed.

On top of that, short video software on the mobile phone could show network connection errors that could result in data failure with such environment configuration mentioned above. SSL Pinning, a technology for attempting to prevent a Man-in-the-middle attack (MITM), causes this case. Its primary mechanism is validating the certificate from the server on a client. If a client does not trust the certificate that it receives, it simply disconnects and does not continue the request. So when it comes to an APP that has SSL Pinning enabled, you could see a disconnection to the network or a reminder of request error. You can install the Xposed framework and JustTrustMe plug-in on the Android simulator to solve this problem. It will go well after the installation of the grab package. The above completes all the relevant environment configurations of the computer and mobile.

2.2. Capture Packets

First, the Douyin in the simulator is opened, and then we return to the computer to observe all of the packages access by Fiddler, where has a package which is JSON (data returned by the webpages). It can determine whether the packet is captured or not by observing field format of the URL, address and the size of the package (because size is quite big commonly).

By clicking on JSON, it is seen that all the data in the packet through decoding in the window is on the right of Fiddler. The data exists in JSON format, and the video data is in a field called data, in which there are several braces Each brace represents the information of a video

The number of videos in a package is fixed. Fiddler can constantly grab the video package using the simulator's simulation mouse to turn the page continuously in order to get more videos but ordinary movement cannot realize it. The simulator can not only simulate human operation, also can control the speed and time of page-turning without affecting the regular use of the device (background operating).

The next problem to be solved is how to make these video packages automatically save in computer so that you can write scripts later to extract the video information. One way to do this is to copy and paste it manually, but this is too cumbersome. Therefore you can use Fiddler's script with adding rules so that automatically save the JSON package when the video is spawn

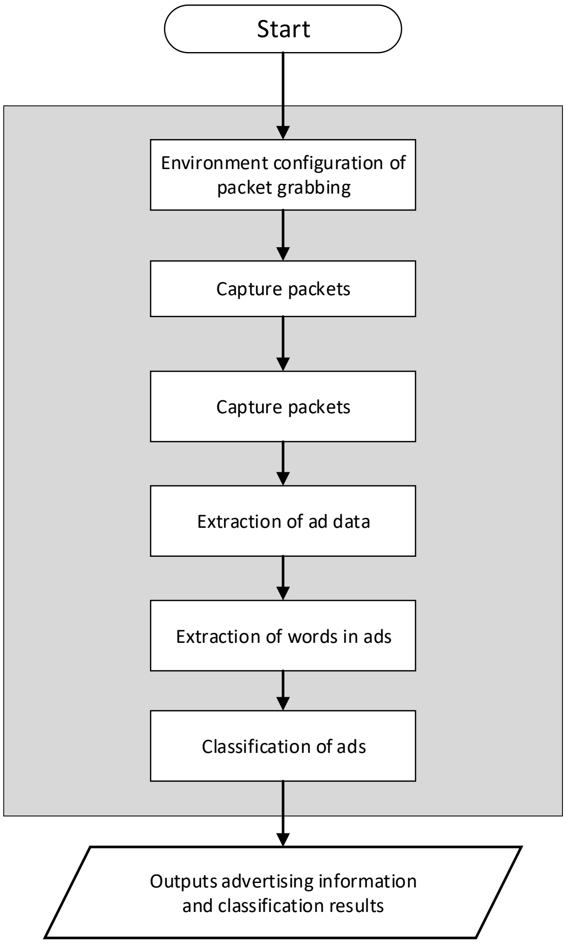
Start
Environment configuration of packet grabbing
Capture packets
Capture packets
Extraction of ad data
Extraction of words in ads
Classification of ads
Outputs advertising information and classification results

Figure 1: Overall functional flow chart.

2.3. Extraction of ad Data

The extraction of advertising data is mainly based on Python language and PyCharm platform. First of all, import packets are needed. It is needed to import the OS module for operating on local folders, along with the JSON module because what needs to parse is JSON. In addition, Requests library needs to import because one can only get to video link address when downloading, so please ask for this address and write the data returned in MP4 files.

Then camouflage header should be obtained, or you may not be able to get the data when you request the link. There are many ways to get the camouflage head. The first is that arbitrarily copy a URL address of Douyin from Figure 2, and paste it into the browser. If it can pass the inspection of the browser, please finish these steps as follows, clicking the Tools of Network, clearing the page, refreshing the browser, and clicking the first request. Then on the right side of Request Headers, the item User-Agent could be found that is the camouflage header.

You can write a script that parses the JSON package with all these preparation, determine whether the data item in the package is an ad by a field (a label field in statistics in this case), and construct the Requests to download the video from the URL of the ad video. Next step is to save the local storage address and other fields of the ad into the database

2.4. Extraction of Words in Ads

The fields retrieved from the JSON package may not fully describe the ad, so the text information is needed to be extracted. The extraction steps are as follows:

First, capture the video into images by frames, put every ten frames into one image, and then save them to the local folder.

Second, read images, utilize Baidu API recognition text which makes use of the POST; the Content-Type is “application/x-www-form-urlencoded,” and then format the request body through URL encode. The requested image should be encoded by base64 and input by the URL encode: Base64 encoding of the image refers to the encoding of an image data into a string, which is used to replace the image address. The final identified result is returned in JSON format.

Third, the identified text is inserted into the database.

2.5. Classification of Ads

The first two sections have prepared all the data. This section will classify the ads according to the statistics and determine whether the ads are risky or regular.

Firstly, the risk lexicon is constructed according to the initial data crawled from the database: the words with risks are saved into the risk lexicon by observing advertising titles and video texts artificially. After that, the risk lexicon can be continuously expanded according to the subsequent crawling data.

After the construction of the lexicon, the text information of the advertisement is firstly processed by word segmentation and word de-stop and then matches with the lexicon. If some words are contained in it, the advertisement will be judged to be risky, or, it is a regular advertisement.

2.6. Problems Analysis

The video should be downloaded to add the "verify = False" field in the request, otherwise it can appear phenomena such as request congestion or refusal, resulting in unable to download the video correctly.

3. Results

First of all, holding a video grabbing every other week is supposed to do and the result will be saved to the local system for acquisition of advertising information, and reads local folders from the python scripts obtained the packet list, after which, advertisements information is extracted from every packet, and saves advertising video to the local, places all what you finished to the database at the same time. The data summarization is shown in Figure 2, and the database table is shown in Table 1. The data is updated once a week.

Table 1: Database Table.

Field	Attribute		
	Type	Primary-key	Note
id	varchar	yes	AD id
title	varchar	no	the title of the advertisement
url	text	no	video download link
type_words	varchar	no	keywords
label	varchar	no	video type
type	varchar	no	AD types
type_name	varchar	no	type name
video_path	varchar	no	video data
video_text	text	no	video text
ad_type	int	no	AD Classification Results

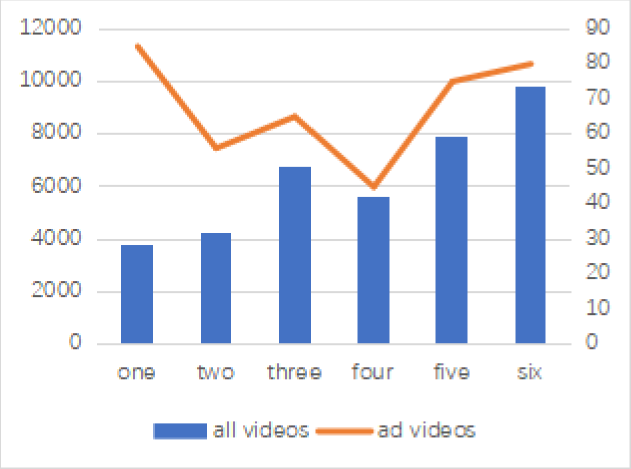

Figure 2: Data summary: the horizontal axis represents the number of weeks, the left side of the vertical axis represents the total number of videos caught, and the right side represents the number of advertisements.

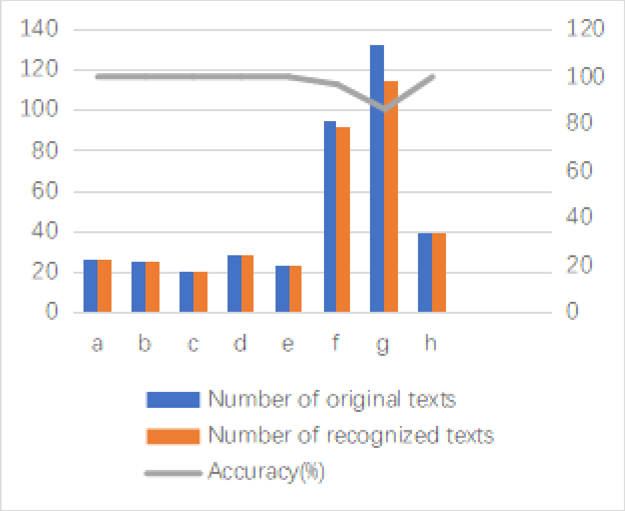

Figure 3: Video text extraction results and analysis.

Then, the text in the image is extracted through Baidu API, and the extraction results and analysis examples are shown in Figure 3.

It can be seen from the results in the table that text recognition has high accuracy. The texts of some pictures are overlapped due to the complex structure, so the recognition accuracy is relatively low.

Next, it is the distinction and classification of advertisements, which determine whether the advertisement belongs to the risk or normal filed. First step is risky lexicon establishment, which mainly contains the following words: [' free ', 'embodiment', 'red envelope', 'reward' and 'entry',' money ', 'no fee', 'no payment', 'send' and 'quota' and' welfare ', 'received', 'fee', 'VX', 'no charge', 'permanent', 'gifting red envelope'..., and then we need to classify the advertising data stored in the database, conduct data statistics once a week, and the results of judging the classification are shown in Table 2.

Table 2: Sample results of advertising classification.

AD Title	Risk Word	Result of decision
Turn on the watermelon video and search for whatever you want to watch. There are tons of them for you to watch!	no	Normal ad
Again to the account, play the strongest elimination can make money, do not believe to try	money	Risk ad
Man, god, demon, spirit, immortals, demons, Buddha, which profession will you practice?	no	Normal ad
Sou: anytime, anywhere, looking for someone to chat with!	no	Normal ad
Boss please pay attention! This fishing makes no payment also can be straight, a blast machine 8.8 billion gold coins!	no payment	Risk ad

4. Conclusions

Advertising has the characteristics of fast transmission speed, comprehensive coverage, and difficulty in tracing and traceability, making it the focus of relevant regulatory departments.This study realizes Internet crawling through the technique of network packet grabbing, which collects and classifies the advertisements on the short video platforms and provide data supporting for the follow-up advertising monitoring work. A thesaurus will eventually advertise and be divided into two categories: regular and risk.

References

[1] Thelwall M. A web crawler design for data mining[J]. Journal of Information Science, 2001, 27(5): 319-325.
[2] T.Y. Chun, World wide web robots: an overview, Online & CD-ROM Review 23 (3) (1999) 135–142.
[3] Minhee Chae,Jinwoo Kim. What's so different about the mobile Internet ?[J]. Communications of the ACM,2003,46(12).
[4] Business 2.0. Wireless Internet is more (Jan. 11, 2001).
[5] KMIC, Korean Ministry of Information and Communication (December 2001).
[6] Wang Xiaohong, Ren Yaoti. New Features and Problems of Short Video Production in China [J]. The Press, 2016(17):76-79.
[7] Liu Chao. Research on the Situation and Development Trend of Short Video Network Transmission [J]. News Culture Construction, 2021(01): 163-164.
[8] Liu Yinyi, Huang Hongzhen. Short Video Advertising Infringement Chaos and Its Governance [J]. Press Outpost, 2021(03):124-125.
[9] Wang Fen. Application of Fiddler Tool in Interface Test [J]. Wireless Internet Technology, 2021,18(02):113-114.
[10] Hu Chunmei. Web Project Interface Test [J]. Electronic Technology & Software Engineering, 2018(20):45.
[11] Xiao Jia. HTTP Packet Capture [M]. Beijing: Posts and Telecom Press, 2019.