Micro-video Data-Acquisition System Design

Zhengzheng Liu, Hai Ji & Sanxing Cao
School of Communication University of China, Beijing 100024, China

ABSTRACT: In recent years, more and more studies about micro-video emerged with the increasingly rapid development of micro-video. This Micro-video Data-Acquisition System gives a method that how to obtain the structured data and the unstructured data of the micro video, and the system has got data of micro-video from about twenty Chinese Internet video platforms. Finally we illustrate the micro-video data acquisition process and its key technology through examples. The data acquired by the Micro-video Data-Acquisition System provides a convenient for the micro-videos' Big Data analysis, It has great significance to the analysis of the micro-videos' effect, users' behavior, and the micro-videos' creation.
KEYWORD: Micro-video; data; acquisition; analysis

1 INTRODUCTION

With the development of Internet technology, communication technology and terminal equipment technology, High-quality Internet video-transmission has become not only an important communication channel of medium information, but also an Internet services with more and more users. The Internet micro-videos is widely accepted with its characteristics of extensive content, the integrity of the plot and short duration, which meet the living status of people's high-pressure life and fragmentation of the time. Meanwhile, Big Data technology has played a catalytic role for Internet micro-videos' development, data become an important carrier for users and Internet video platform to mining videos' transmission effects, guidance videos' creation, analysis users' habits. How to use new technology to get micro-videos' core data, such as playing number, comment number, the number of support and oppose, video introduction, actors and other information. It is an important research topic in the field of information technology.

Based on the above, this text studies on the key technologies of micro-videos' data-acquisition process. The system is mainly use the PHP + MySQL + Apache and based on the CodeIgniter framework for the back-end support. Users get micro-videos all data through the system and grasp the status of the videos' spread. On the one hand, the data can help make prediction and decision to the micro-videos' creation and propagation, on the other hand, users' feedback help to improve the system functions.

2 DATA INFORMATION

2.1 *Data sources*

This text focuses on domestic Internet video platform for data acquisition, so the system will obtain data on almost all the influential Internet video platforms of China, which are divided into large Internet video platform and professional micro-video internet platform. As shown in Table I.

Table I The target Internet platforms

Large Internet video platform	Youku, Tudou, Sohu video, Tencent video, iQIYI, Funshion, Ku6, 56, mango TV, Sina Video, Baidu video, HUASHU TV, Thunder look, Phoenix video, Letv, CNTV ...
Professional micro-video internet platform	V film, CUCTV, Maxtv...

2.2 *Data types*

Internet video resource page is mainly composing two types data, one is the structured data, which reflect the level of users' concern and recognition to its content or not, structured data is typically embodied in digital form, such as playing numbers, comment number, support number, oppose numbers, score and

so on. One is the unstructured data, which reflect the theme of the video content and other relevant information. There are many forms of unstructured data, the majority are in the form of a text message, such as video title, video introduce, actor and director information, video type, release time and so on. The system will obtain the data types are shown in Table II.

Table II Data types obtained by the system

Structured data	Playing number, comment number, support number, oppose numbers, score...
Unstructured data	Video title, video introduce, actors, director, release time, Video type, video time-length, area...

3 SYSTEM DESIGN PROCESS

The system is divided into two parts, namely the structured data acquisition process and the unstructured data acquisition process, corresponding to the two kinds of data types.

3.1 *Structured data acquisition process*

Structured data acquisition flow chart is shown in Figure 1.

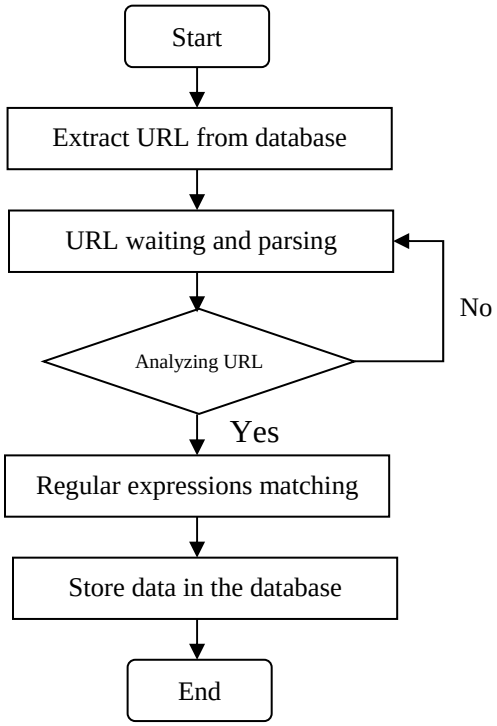

Figure 1 structured data acquisition flow chart

1) Extract an URL corresponding with the every video's structured data from Internet video platforms.
2) Divide the data into different URL queue, according to different data content acquired (such as playing number, comment number, support number, oppose number, etc.), and waiting to be analyzed one by one until the resolution.
3) Using the function of file_get_contents () in the PHP to parse the URL, extracting the string it contains if the URL is successfully resolved,
4) Use the regular expressions with the function of preg_match( ) and preg_match_all ( ) in PHP to match the string and obtain the data it contains.
5) Store the date in the database, complete the operation.

3.2 *Unstructured data acquisition process*

Unstructured data acquisition flow chart is shown in Figure 2.

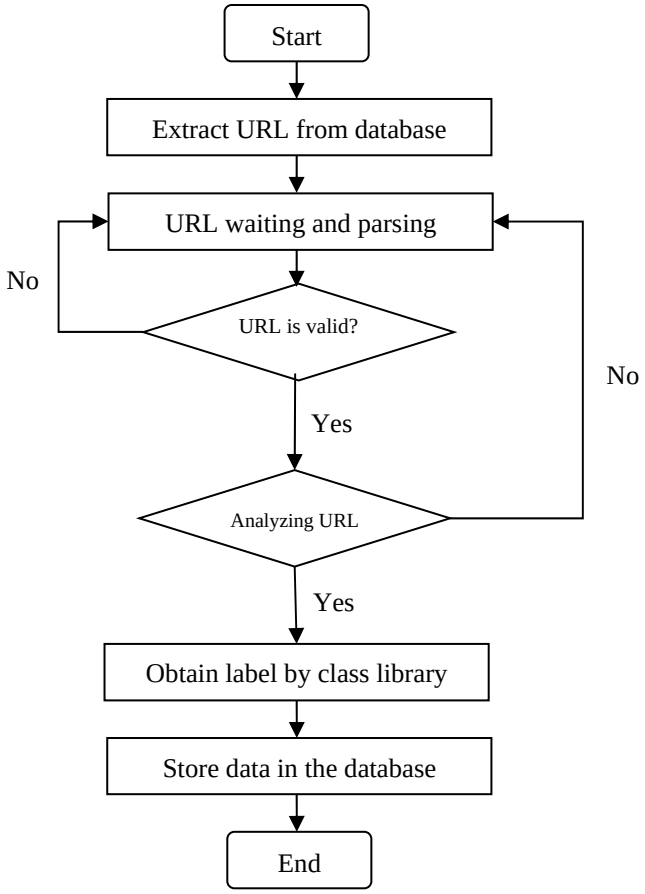

Figure 2 Unstructured data acquisition flow chart

1) Extract URLS of the video resource page corresponding with the "status=0" in the data table.
2) URL in the Video Resource Page is waitting to be analyzed one by one until the resolution.
3) Before analyzing URL, judge the URL is valid or not by cURL. If it is valid, parse it; If not, return URL queue to operate the next URL. cURL is a tool to transfer files and data by the URL grammar rules, which supports the protocol of

HTTP, FTP, TELNET, and some others. It can accomplish the remote data acquisition and collection, simulated login, the docking interface, analog Cookies and other functions.

4) Parse the valid URL by the function of file_get_html ( ) in Simple HTML DOM parsing library, the function of file_get_html ( ) can parses the page in accordance with the W3C standard DOM model.
5) Locate the HTML label and get the information between label pairs by the function of find ( ) in Simple HTML DOM parsing library and the selectors like id, class, tag and so on.
6) Store the string acquired in the corresponding data table. Then update the value of the status to 1 in the data table, which indicates that the video resource page text message has been completed acquisition.

4 KEY TECHNOLOGY & CASE-STUDY

4.1 *Structured data acquisition*

Through analyzing the architecture of the major Internet video platforms, structured data generally does not exist in HMTL source of video resource page, but mostly in the form of interface files to pass data to the video resource page, data always is strings in JSON format. The structured data come from the URL synthesized by keywords which are obtained by PHP procedure analyzing the URL of the video resources page or extracted directly.

Take the Micro-film "Bosom Friend" in iQIYI (*http://www.iqiyi.com/v_19rrh3h65g.html#vfrm=2-4-0-1*) for example to explain how to obtain structured data. The micro-videos' URL of structured data can be obtained according to the URL injection module, such as the URL of playing number *http://cache.video.qiyi.com/jp/pc/231937800/* and the URL of the number of support and oppose *http://up.video.iqiyi.com/ugc-updown/quud.do?type=2&dataid=231937800*. Then analyze it respectively by the function of file_get_contents ( ) in PHP, and we can get the corresponding strings.

```
var tvInfoJs=[{"231937800":3201084}]
```

To obtain the string above by analyzing the URL *http://cache.video.qiyi.com/jp/pc/231937800/*. Among them, "231937800" is the video's tvId, "3201084" is the video's playing number. According to the character of playing number data before and after, use the regular expressions combined with the function of preg_match ( ) in PHP, to match and acquire it.

```
try{null({"code":"A00000","timestamp":"201505
07154115","data":{"albumId":231937800,"dataId":2
31937800,"action":0,"score":8.3,"down":1266,"type
":2,"voters":7642,"up":6376}})}catch(e){}
```

To obtain the string above by analyzing the URL *http://up.video.iqiyi.com/ugc-updown/quud.do?type=2&dataid=231937800*. Among them,"" up ": 6376" is the video's support number, "" down ": 1266" is the video's oppose number. Then according to the character of playing number data before and after, use the regular expressions combined with the function of preg_match ( ) in PHP, to match and acquire it.

4.2 *Unstructured data acquisition*

The internet video resources pages are generally using HTML language, and its encoding based on the document object model (DOM). The main characteristic is all the useful information is contained in the structured keywords. HTML language has a lot of structured keywords, such as <head> </ head>, <title> </ title>, <body> </ body>, etc., all the structured mark is in pairs in a standard page, which provides a convenience to obtain the static information, such as the video's actor, description, title and type and so on.

Take the Micro-film "Bosom Friend" in iQIYI (*http://www.iqiyi.com/v_19rrh3h65g.html#vfrm=2-4-0-1*) for example to explain how to obtain text information. Analyze the URL of the video resource page by the function of file_get_html ( ) in Simple HTML DOM parsing library, we can get the HMTL source code of the video resource page. Through analyzing the resource page URL of "Bosom Friend" *http://www.iqiyi.com/v_19rrh3h65g.html#vfrm=2-4-0-1*, we can get resolved to give the following code (in view of the limited space, only excerpted part of the code to illustrate the problem).

```
<!DOCTYPE html>
<html>
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<title> Bosom Friend (Micro film) - Micro film - HD genuine online viewing - iQIYI </title>
…
</html>
```

Obviously, the text information of the Micro-video "Bosom Friend" are included between the corresponding HTML tags. The title is located between <title> and </ title>; the uploader is located between and </ span> in "class =" type-con ""; director's information is located between and </ span> in "id =" datainfo-director-list ""; the video introduce is located between and </ span> in "id =" datainfo-desc-text "", and so on. Locate the tags and extract the information between the paired tags by the function of find ( ) in Simple HTML DOM parsing library, then use the function of innerText to remove the HTML tags, finally we can complete the acquisition of information.

5 CONCLUSION

The data of micro-video acquired by the Micro-video Data-Acquisition System includes the videos of their own information and users interaction information, which laid the foundation for subsequent video data visualization, analysis of users' favorite video types and users' emotion, and other a lot of work. It provide convenience for micro-videos' big data analysis, and it is of great significance.

An complete Micro-video Data-Acquisition System also relates to the establishment and optimization of the database, the visualization techniques in data injection module design. When the system is in actual operation, it will be further improved and perfected. With data mining and processing technology, It may provide users with richer data and a better visual experience.

REFERENCES

[1] Brin S, Page L. The Anatomy of a Large-scale Hypertextual web Search Engine[J]. Computer Networks, 1998, (30):107-117.

[2] Ghemawat S, Gobioff H, Leung Shun-Tak. The Google File System [A]. //Proceedings of the 19th ACM Symposium Oil-Operating Systems Principles[C]. 2003:20-43.

[3] Heydon A, Najork M. Mercator: A scalable, extensible Web crawler[J]. World Wide Web, 1999,2(4):219-229.

[4] Cho, Junghoo; Hector Garcia-Molina (2003).“Estimating frequency of change”. ACM Trans. Interet Technol. 3 (3): 256–290.

[5] Hai Ji, San Xing Cao. Research on Network Video Data Acquisition and Analysis Based on Big Data[J]. Machine Tool Technology, Mechatronics and Information Engineering[C], 2014: 3116-3119.

[6] Bin Zhang, Jia Lun Song, Zheng Zheng Liu, Hai Ji. The Micro-Video Label Classification System Design Based on Network Data Acquisition[J]. Applied Mechanics and Materials[C], 2015: 556-559.