III Data Collection

Adequate evidence is necessary to analyze culture around Kichiku and do further popularity measure and analysis. Thus, video data collection is the fundamental step to support our discussions. One of the most direct method to get correlated video data is fetching video data records from the website of online video platforms. This section mainly discusses about data fetching technique we applied to get Kichiku video meta-data from Bilibili, which has developed as the largest youth video-sharing platform, as well as one of the origin platform of Kichiku in mainland China in the last decade.

In this collection step, aiming on further analyze, besides of basic video ID (state as *AV-number* in Bilibili), we focus on three functionalities to decide the table format in MySQL database and items of collected data.

1) **Popularity measure metadata**, including count of *view, comment, danmaku, favorite, coin, share, like and duration*. There are multiple norms to define the popularity of a video. To setup a measure standard specialized on Kichiku video (details will be discussed in the *Popularity Analysis* part), these statistical data should be taken in our consideration.

2) **Producer-related metadata**, including *auhor id, author name, fan count*[1]. We find that the group of video producer has a significantly impact on popularity of Kichiku videos. These data records are used in our survey design and popularity analysis.

3) **Content-related metadata,** including *title, description, keywords, video subtype and upload-time*. Relationship between video content and popularity is the core part of this subculture study. Hot themes (tags from keywords) of a video change by time is a epitome of Kichiku changes in the past ten years.

We use Web data extraction technique to fetch data items mentioned above. The overall pipeline of the data collection process is shown in the figure in the next page. Comparing with traditional method, this process could be divided into 3 sub-steps in general. Firstly, directed search on each Kichiku subtype is applied to get a list of video ID in one page and save them into array-like files, this step is completed by

[1] The fan count data is fetched in a separate step, which will be discussed in the data analysis part.

sending request API which will return the news list of a video displayed on the website. Secondly, the video information is fetched by requests with JSON and HTML replies, in which JSON response returns the keyword data. Thirdly, arranged data items are sent to MySQL database management system, stored there for invalid data filtering and analysis.

We have defined three general rules to filter out invalid data items. 1) View count is zero; 2) data missed in some fields, e.g. upload-time is null; 3) not a Kichiku video, by double check the video type index.

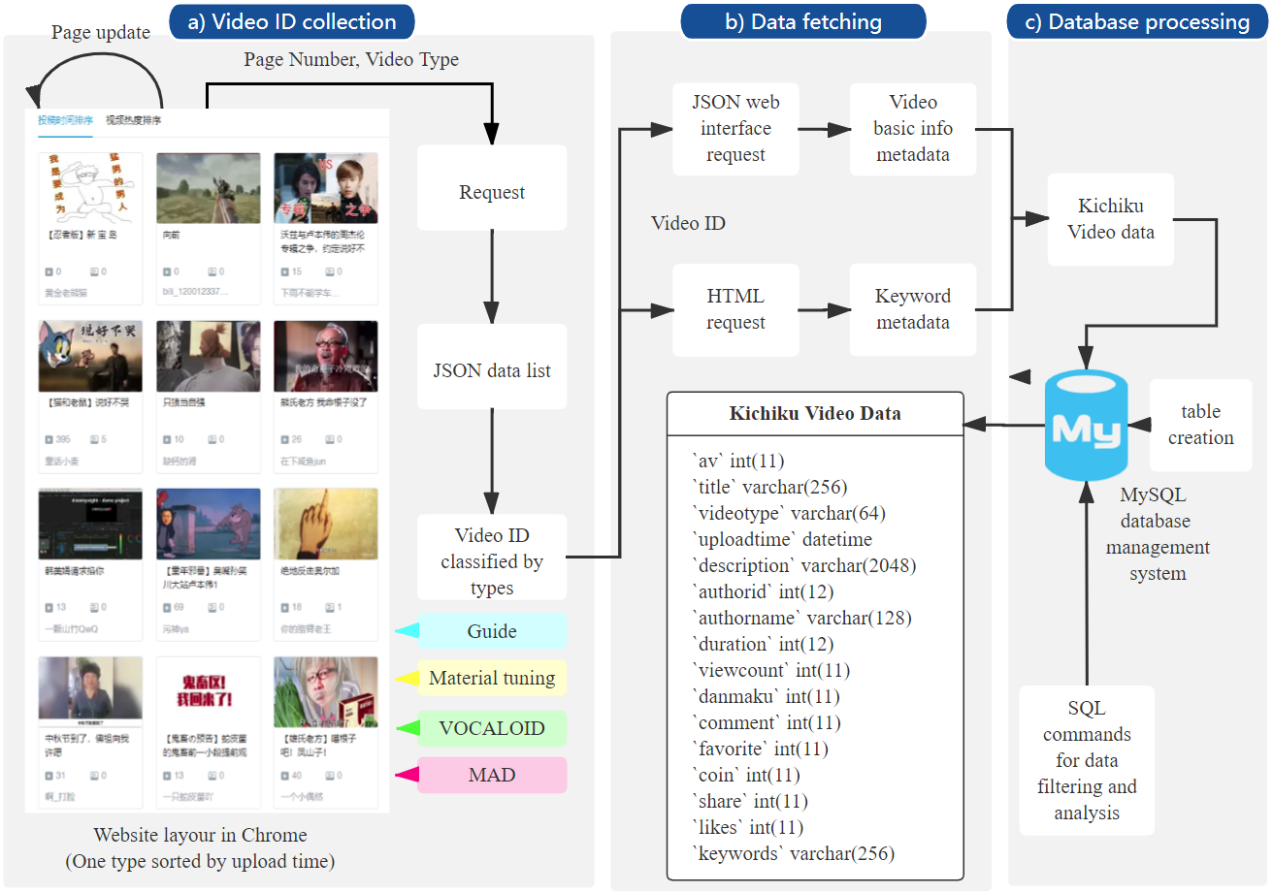

Fig. The pipeline of the data collection part. a) Video ID collection collects all ID of Kichiku video in web pages, classified them by subtypes defined by Bilibili platform. b) Data fetching step sends two requests based on video ID to get metadata. c) Create a connection to MySQL database, initialize a table, and insert fetched data into the database system by SQL queries for further analysis.

With our approach, from the timespan in September 2009 to May 2019, 165380 valid data records are returned and stored in the database. The data was captured in July 3rd, 2019. Two-month gap was set to avoid real-time growth of data for popularity analysis which is measured accumulated, such as view count and share count.