DS 504 Project 1 Final Report

Counting and Analyzing Bilibili Videos via Sampling

Group 3

Wei Wang, Dongyu Zhang, Yuchen Shen

ABSTRACT

Leveraging the characteristics of Bilibili video website and exploiting properties of Bilibili search API, in the paper, we sample a subset of ids from the entire video ground truth and use the proportion of available in the subset to estimate the whole. However, there are 30 million videos in our dataset, we perform stratified sampling on the total videos data. We separate entire id space into 100 groups and in each group random sample 3000 ids, then renumber them in sequence. Further approach and analysis are performed on this new dataset. Then we estimate the proportion of available videos. In each group, we sample different percentage of videos which can be from 1% to 20%, and to show the stability we will sample and analysis 10 times for every kind of sampling. After knowing the number of available videos, we calculate the average view counts, estimate the statistical property of videos, and predict number of videos will be uploaded in the future. Then the goal is to apply our estimation to the 30 million whole datasets which will get a more reliable estimation of how many available videos in Bilibili website now. Finally, for advanced sampling, we apply a bootstrap sampling, to show the estimation against the other two.

Keyword

Videos, Time-series, Sampling, Bilibili

1. INTRODUCTION

1.1 Bilibili Website

Bilibili (Chinese: 哔哩哔哩), nicknamed as B 站, literally means “the B site”, is a popular video sharing website based in China. The theme of Bilibili is focusing on animation, comic and game (ACG). There are huge number of users submit, view and add commentary subtitles on videos each day.

The video id on Bilibili is generated in numerical order, which starts from 1. For now, the ids have come to about 32,000,000. While many ids refer to non-existing videos, which means these videos do not pass the censorship or have been deleted. For example, from 0 to 10, only video 2, 7 and 9 are still accessible. Unfortunately, the total number of existing videos and view counts are not made available publicly by Bilibili. Brute-force survey of entire Bilibili video population would be very costly. Hence, in this project, we are motivated to sample the video id space to estimate the total number of existing video.

Figure 1. Website main interface of Bilibili.

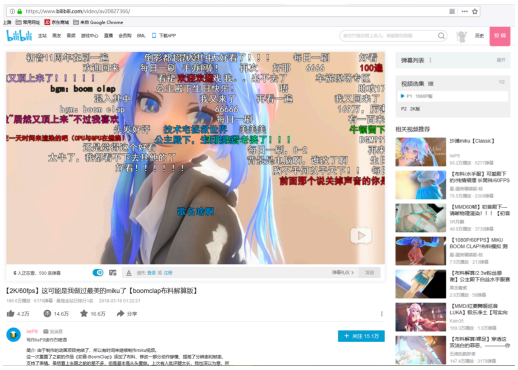


Figure 2. Videos interface of Bilibili.

For each video, there are many important statistics, such like the date of publication, duration, view count, danmaku count (danmaku is a real-time commentary subtitle), coin count ('coin' in Bilibili is for supporting the video you like, if the video get more coins, the system will recommend this video to more people). We could use these data to estimate the total view count and some other important indexes. According to the date of publication, we could build a model to predict the change of total number of videos on Bilibili. Such statistics could also shed light on the research about the popularity of Bilibili and such kinds of video websites.

1.2 Motivation and Benefits

Knowing the statistic of the numbers of videos being deleted or not passing the censorship is important from both technical and social perspective. For instance, with the results of the statistic, people can estimate the total amount of storage more accurately, compare with estimating total amount of the all videos based on video id. Since the traffic of this kind of video website contributes to a significant portion of inter-domain network traffic, the statistic is useful to estimate the network capacity needed to delivery Bilibili videos.

Furthermore, knowing the statistic of the video upload date can shed light on the Bilibili traffic on different season and time period, which can support to determine the amount of storage needed to prepare on different season, such as on period of school time or the break. And also, it's beneficial to estimate network capacity on different time period.

1.3 Previous Work

There are many studies on estimating the size and other properties on online social networks. Most of them are focusing on estimating the total number of videos based on different method. However, in order to get a better estimation for the storage or network capacity from the statistics on online social website, it would be much accurate if we take the deleted or not pass videos into account, particularly in some countries who have strict internet censorship policy. Furthermore, since most of the audience of the online website we are analyzing is young students, it's important to do the study depending on different time period. The traffic on term period and the break, even in the start of one term or around the end of one term, would be significantly different.

2. Methodology

In this section, we will introduce our API and

sampling methods to try to get accurate estimation.

2.1 API

Now we are getting the information from the Bilibili Website. Using API providing from Bilibili (https://api.Bilibili.com/x/web-interface/view?aid =id). we can gather features of each video like views, or we can get nothing if the video is not available anymore.


{"code":0,"message":"0","ttl":1,"data":{"aid":3037,"videos":1,"tid":174,"tname":"其
他","copyright":2,"pic":"http://i1.hdslb.com/bfs/archive/4f2f7f31c141081f9543325fd24d04fe0a955f1a.jpg
","title":"【福利】听歌赏图。","pubdate":1266846509,"ctime":1497349968,"desc":"原创，听歌赏图有福
利。","state":0,"attribute":1097811,"duration":99,"rights":
{"bp":0,"elec":0,"download":1,"movie":0,"pay":0,"hd5":0,"no_reprint":0,"autoplay":1},"owner":
{"mid":12591,"name":"阿姨说第一
次","face":"http://i0.hdslb.com/bfs/face/e4dc7c33869e8e2cd95c0dd1fa4ce9a0282eb9d7.jpg"},"stat":
{"aid":3037,"view":1942,"danmaku":85,"reply":6,"favorite":5,"coin":0,"share":7,"now_rank":0,"his_rank
":0,"like":0,"dislike":0},"dynamic":"","cid":2171,"dimension":
{"width":0,"height":0,"rotate":0},"no_cache":false,"pages":
[{"cid":2171,"page":1,"from":"vupload","part":"","duration":99,"vid":"","weblink":"","dimension":
{"width":0,"height":0,"rotate":0}}]}}

Figure 3. Result from API. Containing information of each videos.

2.2 Sampling Methods

Using sampling approach, we try to sample a subset of ids from the entire video ground truth and use the proportion of available in the subset to estimate the whole. Unfortunately, due to the computation power, we are not able to have all 30 million videos be our dataset. To solve this problem, we are performing stratified sampling on the total videos data. We will separate entire id space into 100 groups and in each group random sample 3,000 ids, then renumber them in sequence. Now, we have 300,000 video ids having id from 1 to 300,000 as our experiment data (ground truth). Further approach and analysis will be performed on this new dataset.

Then we will do further sampling to estimate the proportion of available videos. First, we are planning to try simple random/uniform sampling as our baseline to be compared. Second, consider that some *id*s deleted may relate to time period, we will try stratified sampling, which will reduce the bias. In our stratified sampling, we will stratify the entire experiment data into 100 groups. In each group, we are going to sample different a percentage of videos which can be from 1% to 20%, and to show the stability we will sample and analysis 10 times for every kind of sampling. All these will also be performed to random sampling.

Notation	Description
N	The number of total *id*s
k	The number of stratified groups
N_i	The number *id*s in i^{th} group
n_i	The number of sampled *id*s in i^{th} group
e_i	The number of existed sampled *id*s in i^{th} group
$\hat{p}$	The proportion of existed *id*s (Estimator)

Table1. Notations and Terminologies.

Estimators:

For Random Sampling, we assume the estimator is:

$$\hat{p} = \frac{e}{n}$$

For Stratified Sampling, we assume the estimator is:

$$\hat{p} = \sum_{i=1}^{k} \frac{e_i}{n_i} \times \frac{N_i}{N}$$

After getting the knowledge of the number of available videos, we can calculate the average view counts. We will also estimate the statistical property of videos by different features. In order to predict number of videos uploaded in the future.

We also use bootstrap to make a more reliable result but also with some disadvantage.

2.3 Times-Series Model

We will split a year into months, and estimate the total number of videos from the sample set and predict the trend using ARIMA time series approaches which should follow a sea-

sonal model having incremental and seasonal trend.

Autoregressive integrated moving average (ARIMA) model is a generalization of an autoregr- essive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrat- ed" part of the model) can be applied one or more times to eliminate the non-stationarity.

All those analyses will be done on the experiment dataset. Having the true information of all videos, the performance will be evaluated by the true proportion of available videos and our estimation. For other statistical estimation, the performance will be evaluated by the real data. The final step of our project is to apply our estimation to the 30 million whole datasets which will get a reliable estimation of how many available videos in Bilibili website now.

3. Experiments and Result

In this section, we will do all the sampling on our ground truth data, and statistical method with time series to predict the development of Bilibili website.

3.1 Percentage of available videos

In our experiments, we are going to estimate the proportion of available videos from total. We first tried two sampling method, uniform sampling and stratified sampling. For both methods, we set six different sampling rates from 5% to 30%, and for each sampling rate, we will sample 100 times to get a distribution of the result. For stratified sampling, we split the entire data set into 100 equal groups by ids.

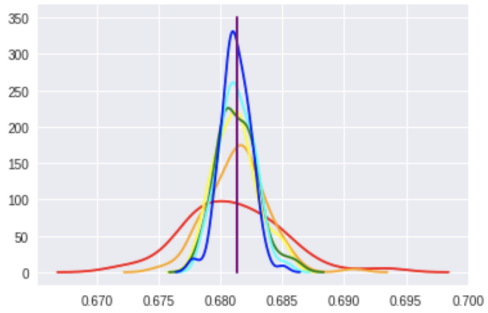


Figure 4. Result of Random Sampling. Lines with different color is the distribution of different sample amount.

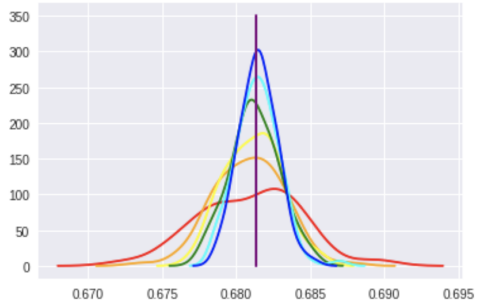


Figure 5. Result of Stratified Sampling. Lines with different color is the distribution of different sample amount.

There are 6 different distributions close to Gaussian distribution representing 6 sampling rates. The vertical line is the real proportion of available videos (Ground Truth). As the result, we can see that both sampling methods perform well on this dataset. The more data we sample, the less variance we get. The blue distribution is the 30% sample, it is closer to Gaussian distribution and the mean is closer to the ground truth.

The real proportion for the ground truth is 0.681316, our uniform and stratified sampling give average proportion of 0.681261 and 0.681178. Both of them gives a good result, and as supposed, with the increasing of sam-

pling rate, the standard deviation decreases and the mean gets close to ground truth.

3.2 Statistical Information

Since stratified sampling will be more stable and reduce the standard deviation, we choose stratified sampling with 20% sampling rate to do further research. Similarly, using stratified sampling, we analysis the statistical information of the sample, from which we can estimate the entire dataset.

	Estimation	Truth	Difference
View	7407.3422	7412.7667	-0.000731778
Danmaku	133.0314	131.3530	0.012777782
Reply	31.1622	31.1991	-0.001182726
Favorite	127.1124	127.2606	-0.00116454
Coin	55.7084	55.9315	-0.003988808
Share	23.5126	23.6497	-0.005797114
Like	33.6970	33.9477	-0.007384889
Dislike	0.6840	0.6811	0.004257818

Table 2. Comparison of statistical information on estimation and ground truth (0.3 million).

Now, we have knowledge about the estimation of several statistical information. Combining this information and the proportion of available videos, we can calculate the estimation of total view number and other features through.

$$total\ number = average\ estimation \times (\#videos \times percentage\ of\ available)$$

Therefore, we can estimate the statistical information of all videos in Bilibili (30 million), not only our ground truth through the function above.

	Estimation (million)
Available Videos	21.802112
View	161495.7043
Danmaku	2900.365482
Reply	679.4017746
Favorite	2771.318781
Coin	1214.560776
Share	512.6243386
Like	734.6657681
Dislike	14.91264461

Table 3. statistical information of videos on Bilibili website.

We can see total available videos reaches 22 million and the total times of video watched by Bilibili's audience reaches 161 billion, although it is not comparable with YouTube, Bilibili is developing fast which is a new company mainly focus on Chinese young adult.

To estimate the store space for videos Bilibili needed, we record the average duration in our sample set which is 1307s. The usual size of a video in Bilibili is 10MB per minute, so we can estimate the average size is 220MB, then we can know the store space needed is 4615 TBs.

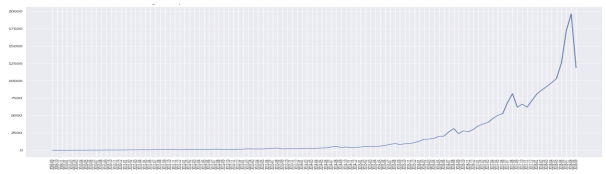


Figure 6. number of videos and uploaded with time.

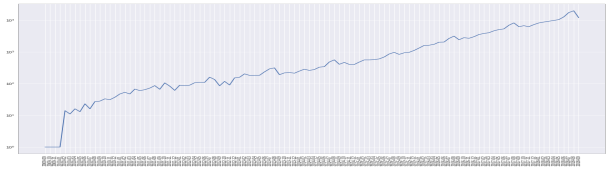


Figure 7. log of number of videos (base 10) uploaded with time.

The two figures above are the trend of videos uploaded and view number with time. Similar to our hypothesis, these two number will in-

crease obviously from Jun. to Aug. due to the holiday time. The video number can be predicted by a time series model.

3.3 Time-Series Analysis

As shown in figure 3, the number of videos uploaded can probably fit a time series model.

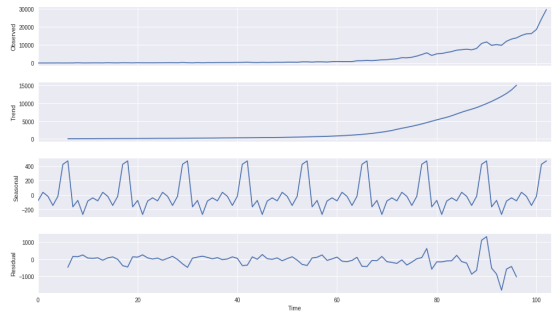


Figure 8. Exponential and seasonal trend

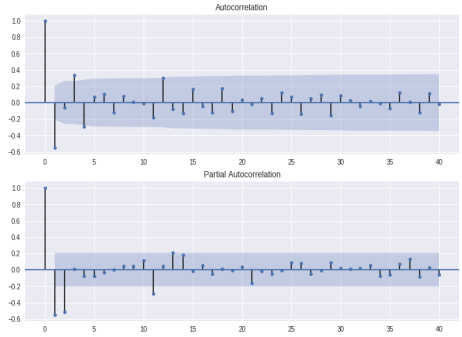
To know the trend of the data, we first decompose it. From the second and third graph, we can find the data have an exponential and seasonal trend, from which we know how to process the data to be stationary.

Since the data have exponential trend, we first do a log transformation, which is the left graph below, it still not stationary; we then do additional a first difference transformation due to the seasonal trend, which is the right graph below, it shows randomness and stationary. We can also know the stationarity from the p-value which is 5.855548e-25 < 0.05.

Figure 9. Log Transformation.

Figure 10. log + first difference transformation

Since the sequence is stationary, we will calculate the ACF and PACF to determine the model and parameter we shall use. The ACF is one order trailing and the PACF is 12 order censored. Apply the model and parameter to our data, we will get the final model with quite low AIC and BIC.



Statespace Model Results

Dep. Variable:	videos_log	No. Observations:	103
Model:	SARIMAX(1, 1, 0)x(1, 1, 0, 12)	Log Likelihood	-25.028
Date:	Thu, 04 Oct 2018	AIC	56.056
Time:	00:42:37	BIC	63.960
Sample:	02-01-2010 - 08-01-2018	HQIC	59.257
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5565	0.064	-8.701	0.000	-0.682	-0.431
ar.S.L12	-0.3426	0.081	-4.245	0.000	-0.501	-0.184
sigma2	0.1000	0.010	9.880	0.000	0.080	0.120

Ljung-Box (Q):	48.20	Jarque-Bera (JB):	32.88
Prob(Q):	0.18	Prob(JB):	0.00
Heteroskedasticity (H):	0.04	Skew:	-0.83
Prob(H) (two-sided):	0.00	Kurtosis:	5.45

Figure 11. ACF and PACF results and model result.

The first plot in figure 12 is the logarithm of the number of video. The second plot is the original number of video. The green line is the forecast (from Feb 2010 to Dec 2018) and the blue line is the real data (from Feb 2010 to Aug 2018), the forecast can form the trend quite well with both incremental and seasonal

trend. The RMSE of our forecast is 492.235, which is the average difference between the estimator and real number of videos uploaded each month, illustrates that our estimation is very close to the real data.

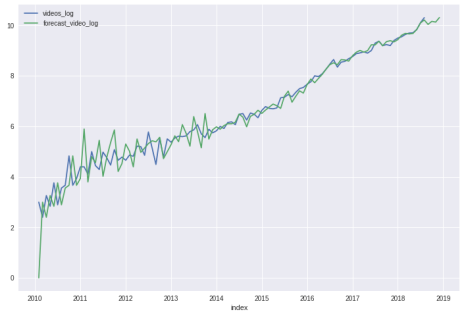


Figure 12. Forecasting of log videos uploaded using time-series model.

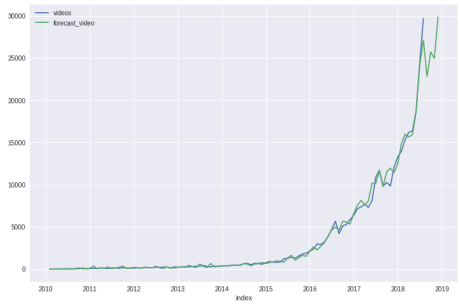


Figure 13. Forecasting of videos uploaded using time-series model.

Using the time-series model, we get above, we will accurately forecast number of videos uploaded in the future on our ground truth dataset (300 thousand *id*s). Combining the forecast number and the estimation proportion of available videos, we can further forecast number of available videos uploaded for the entire Bilibili website. Example, for

4. Other Sampling Method

Since Random Sampling and Stratified Sampling only gives an estimation of the estimator, we try to use Bootstrap Sampling method to further estimate and get a confidence interval for our estimation.

The bootstrap sampling is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples. Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This is resampling method.

For our case, we use bootstrap to resample our ground truth dataset, with 300 thousand *id*s for 1000 times; then we calculate the estimation for each sample, and sort them from small to large; the estimation on the 2.5% position is the lower boundary and estimation on the 97.5% position is the upper boundary, this will form the confidence interval.

we estimate the proportion of available videos among the bootstrap samples. Since this is based on the ground truth dataset, it will estimate information about the entire website.

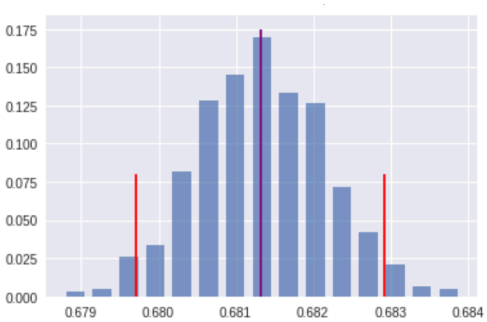


Figure 14. Bootstrap sampling. Result of confidence interval.

As shown in the figure above, the purple line is the average estimation proportion which is 0.681331; and with the two red line is the 95% confidence interval which is [0.679702, 0.682914]. A result with confidence interval is more reliable, it gives an interval in which the estimation will most possible lies in. Meanwhile, we need to resample the 300 thousand samples 1000 times which will

spend much more time than the other methods, and for larger data, it may not be possible.

5. Conclusion

In our experiments, we estimate the proportion of available videos from total. We use two sampling method, uniform sampling and stratified sampling. For both methods, we set six different sampling rates from 5% to 30%, and for each sampling rate, we sample 100 times to get a distribution of the result. For stratified sampling, we split the entire data set into 100 equal groups by ids. We analyze the statistical information of the sample, and calculate the estimation of total view number and other features. We also estimate the store space for videos Bilibili needed. We record the average duration in our sample set and estimate the average size, then calculate the total store space needed. For the Bilibili company, it is a waste of id space although infinite, since almost 35% of ids are unavailable. The company should have better rules for *id* distribution, such as assigning video an *id* after the audit rather than before.

Using time series model, we have already generated a well performance model to forecast the number of videos uploaded with time. Knowing the number will help to arrange new store space and reduce waste. We also estimate the same information for the website beside our ground truth dataset. From the estimation, we can image the quick development of Bilibili website, the huge number of videos continuously being uploaded. It is really important to develop store space for further usage.

6. Future Research

For future work, we are interested in using this method to further study how the statistics, such as the total number of danmaku of different videos, number of upload distribution during semesters or vocations, which would give us a dynamic view of traffic by Bilibili.

We can also use API to collect user information from Bilibili website, it is an interesting problem to know the influential users which random walk could be useful. Also, recommend system is another analysis to know links between different videos, company can recommend similar or potentially user interested videos to different users for experience improvement.

7. References

[1] What can we get from Bilibili API.
https://www.jianshu.com/p/1c263f39e68a

[2] Bilibili API collection, integration and developing.
https://github.com/Vespa314/bilibili-api/blob/ma-ster

[3] Getting videos data on Bilibili in one hour: A design of high speed web-crawling.
https://zhuanlan.zhihu.com/p/35359905

[4] How to crawl videos information on Bilibili using Python.
http://www.yunweipai.com/archives/23619.html

[5] Analysis of 20 million users on Bilibili website.
https://zhuanlan.zhihu.com/p/24434456

[6] Time-Series Analysis using Python: Seasonal ARIMA
https://zhuanlan.zhihu.com/p/35282988

[7] Python: Statistical Estimation: Interval Estimation.
https://www.jianshu.com/p/6cfce4cc2f7f

[8] ARIMA model Introduction
https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average

[8] Bootstrap Sampling Introduction
https://en.wikipedia.org/wiki/Bootstrapping_(statistics)