Multimed Tools Appl (2016) 75:1543–1556
DOI 10.1007/s11042-014-2245-2

On application-unbiased benchmarking of web videos from a social network perspective

Juan Cao · Yongdong Zhang · Rongrong Ji · Xin Li

Received: 29 November 2013 / Revised: 21 August 2014 / Accepted: 22 August 2014 /
Published online: 13 September 2014

Abstract Along with the emerging focus of community-contributed videos on the web, there is a strong demand of a well-designed web video benchmark for the research of social network based video content analysis. The existing video datasets are challenged in two aspects: (1) as the data resource, most of them are narrowed for a specific task, either focusing on one content analysis task with limited scales, or focusing on the pure social network analysis without downloading video content. (2) as the evaluation platform, few of them pay attention to the potential bias introduced by the sampling criteria, therefore cannot fairly measure the task performance. In this paper, we release a large-scale web video benchmark named MCG-WEBV 2.0, which crawls 248,887 YouTube videos and their corresponding social network structure with 123,063 video contributors. MCG-WEBV 2.0 can be used to explore the fusion between content and network for several web video analysis tasks. Based on MCG-WEBV 2.0, we further explore the sampling bias lies in web video benchmark construction. While sampling a completely unbiased video benchmark from million-scale collection is unpractical, we propose a task-dependent measurement of such bias, which minimizes the correlation between the potential video sampling bias and the corresponding content analysis task, if such bias is unavoidable. Following this principle, we have shown several exemplar application scenarios in MCG-WEBV 2.0.

J. Cao · Y. Zhang (✉) · R. Ji
The Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, China
e-mail: zhyd@ict.ac.cn

J. Cao
e-mail: caojuan@ict.ac.cn

R. Ji
e-mail: rrji@xmu.edu.cn

X. Li
Department of Computer Science and Technology, Tsinghua University, Beijing, China
e-mail: lx_sy@126.com

Keywords Web video benchmark · Social network · Sampling bias · MCG-WEBV 2.0

1 Introduction

With the ever increasing research focuses on community contributed videos on the web [3, 8, 13, 15, 16, 20, 25–27], there is an emergent demand on establishing a well designed web video benchmark containing web video contents and their community structure. Comparing to the traditional video benchmarks, the unique characters of community contributed videos have brought new challenges in two aspects: (1) the new benchmark should have sufficient coverage of rich contents and contexts for multidisciplinary tasks; (2) Since any benchmark is a sample from massive user-sharing websites, the resulted potential sampling bias (if unavoidable) should be independent to the subsequent content or context analysis tasks, so as to provide fair comparison between different algorithms.

Accordingly, we summarize the existing video benchmarks from above perspectives as shown in Table 1:

From the perspective of applications scenario, the existing benchmarks can be classified as (1) "for social network analysis", (2) "for video content analysis" and (3) "for social video analysis". Most benchmarks in Class (1) only crawled the statistics of videos and users, without downloading and analyzing the video content; Most benchmarks in Class (2) crawled limited-scale videos contents, but without maintaining the social network structures of users.

The above existing video benchmarks can either analyze the characteristics of video social network without content, or only carry out specific content analysis without social network context. As this context information is becoming increasingly important for video content analysis [15, 19], we believe a large scale video database containing both video contents and their social network structures is of fundamentally important to push forward the research of both community-contributed videos and their social network. in Class (3), we construct a large-scale web video benchmark named MCG-WEBV 2.0,[1] which contains in total 248,887 popular videos and their corresponding social network structures from 123,063 owners.

From the perspective of sampling strategy, we classify the existing benchmarks to (1) "from video" and (2) "from user". The entire web video collection can be regarded as a large-scale heterogeneous network with user and video, and they are collected by the user interactions over videos. So the sampling over this network can be directly based on a video attribute, or indirectly based on a user profile, and then further collect the videos corresponding to their interactions. Most of existing benchmarks sample videos based on their view counts, age or response numbers. On the other hand, a small part of benchmarks sample videos based on the user requests. The MCG-WEBV 2.0 follows this "from video" principle and is collected from most viewed video lists and their related videos lists.

However, both sampling criteria might introduce potential bias to some specific videos if it is not performed uniformly over the entire web videos. We refer to this effect as **Sample Bias**. This data bias has been recently reported by computer vision researchers [2, 21]. For instance, Torralba and Efros disclosed the bias of object recognition benchmarks in [21] from three-fold, including: *Selection Bias*, *Capture Bias* and *Negative Set Bias*. By running the same algorithm on different datasets, the sampling bias is reported in some

[1] http://mcg.ict.ac.cn/mcg-webv.htm

Table 1 Classification of the existing video benchmarks

Applications/Sampling strategies	From Video	From User
For Video Network Analysis	[1, 2, 5–7]	[9, 10, 24]
For Video Content Analysis	[11, 12, 14, 23]	
For Social Video Analysis	[18], $MCG-WEBV$	

outdate benchmarks such as Caltech-101 and MSRC. On the other hand, Borghol et al. in [2] analyzed the popularity bias of web video datasets and verified that the sampling based on keywords search would bias to more popular content, while the sampling based on recently-uploaded videos is likely unbiased.

The existing practices have shown that sampling all views from the entire website such as Flickr or YouTube is indeed infeasible. Furthermore, considering the fact that the web videos are extremely diverse in many aspects, running a task on a completely unbiased dataset is also unnecessary in many practical settings. Taking the "hot topic discovery" for instance, if it is performed over a completely unbiased benchmark where the majority of videos are unpopular at all, the quantitative comparisons would be less useful. Therefore, to guarantee a fair comparison, the corresponding benchmark should be sampled with independent bias (if unavailable) to its corresponding application, which is referred as *application-unbias* in our subsequent discussion. To this end, we propose a strategy to measure the correlation between sampling criteria and task to be evaluated. The proposed principle can guide the subsequent benchmark construction, enabling as minimal potential bias as possible to the corresponding task.

Our contributions can be summarized as follows:

- A large-scale web video benchmark called MCG-WEBV 2.0, which contains in total 248,887 videos and their corresponding social network structures from 123,063 owners. The content information in MCG-WEBV includes the keyframes and 11 visual features for every video, and the contextual information includes web profiles of videos and video contributors, 4 user interactions and the ground-truth labels of 73 hot topics (details in http://mcg.ict.ac.cn/mcg-webv.htm).
- Rather than building a totally unbiased benchmark, we first define the concept of *application-unbias* to ensure a relatively fair design of video benchmarks for a specific task. It measures the correlation between the sampling criterion and the task, aiming to select application-independent bias (if unavoidable) to influent the corresponding analysis tasks as minimal as possible. i.e., sampling videos based on user properties is shown to be less bias for the purpose of popularity analysis.
- Based on MCG-WEBV 2.0, we carry out case study about the sampling bias, and have given several important conclusions such as user behavior analysis is appropriate to be evaluated on this benchmark with less bias, while the video comment analysis will be greatly influenced by the popularity-bias embedded in the dataset.

2 Video benchmarks revisit

As shown in Table 1, the existing video benchmarks can be summarized into the following three categories based on their applications scenarios: e.g., video network analysis task, video content analysis task, and social video analysis.

2.1 Benchmarks for video network analysis

In general, benchmarks for network analysis share three common characteristics: (1) scalability, which are typically at million-level. (2) most videos provide the social network structure of their contributors, e.g., user profile information and their interactions based on comments and sharing. (3) without visual contents (e.g., the original videos and their visual or textual features), and thus unable to be used for content analysis and retrieval.

Cheng et al. [6] constructed a Web video dataset with 3,269,030 popular YouTube videos from February 22 to May 18 2007. By analyzing the related links between videos in this dataset, the authors verified the "small-world phenomenon in YouTube social network. Gill et al. [9] collected 323,677 YouTube videos based on the user request in a campus during 85 days, and investigated the traffic characterization of YouTube on this dataset. In [1], Benevenuto et al. crawled around 223,851 top responded YouTube videos and 417,759 video responses from September 21 to 26, 2007, and explored the users' video response behavior patterns. Recently, to model the popularity distribution of Web videos, Borghol et al. [2] constructed two benchmarks based on YouTube, one with 29,791 recently uploaded videos and the other with 1,135,253 videos searched by keywords. Besides, different from the above User Generated Content (UGC), some databases are constructed based on the video content in the standard Video-on-Demand (VoD) systems, whose videos are supplied by a limited number of media producers such as licensed broadcasters and production companies, and their popularities are generally controlled by a professional way. For instance, to study the user behaviors, Yu et al. [24] collected 6,700 videos based on the request logs of 150,000 users during 219 days on Powerinfo VoD system, which is deployed by China Telecom. Similarly, Huang et al. [10] analyzed a 9-month trace of MSN VoD videos deployed by Microsoft.

All these databases only downloaded the metadata for every video such as "ID", "Uploader", "Category" and "Related videos", where the original videos as well as their textual information (given by either video contributors or reviewers) are excluded. Besides the characteristics of the video social network, the above works also analyzed and compared the basic properties of their video benchmarks, such as the video lengthes, ages, and categories. These investigations provide a fundamental insight for designing suitable Web video research systems.

2.2 Benchmarks for video content analysis

The benchmarks for video content analysis are designed to explore and evaluate some special tasks such as video retrieval, topic discovery and concept detection. Such datasets in general follow three characteristics: (1) the size of datasets is limited; (2) the visual contents are provided such as keyframes and extracted visual features, as well as ground truth for a specific application, e.g., objects or instance labels; (3) the video selection criteria are indeed biased, which selects videos with special characteristics, e.g., the videos are related to one topic or contain a specific semantic concept. Therefore, it is difficult to extend to the other applications.

For instance, there are several widely used video datasets for action recognition, e.g., Hollywood database [12] with 1,707 professional movie video clips for 12 action classes; *UCF50* [22] with more than 5,000 realistic YouTube videos for 50 action categories; and the recently released CCV database by Columbia University, with 9,317 un-edited consumer videos for over 20 semantic categories. In addition, *CC WEB Video* [23] is a Web video benchmark designed for near duplicate video detection. Based on 24 pre-defined queries,

it collected 12,790 videos by retrieving keywords on YouTube, Google Video and Yahoo! respectively. This dataset has high near duplicate ratios. In addition to the above mentioned benchmarks openly available, there are also some expert-controlled video corpuses collected for specific applications. For example, to explore the Web video topic discovery and tracking, Liu et al. collected more than 20,000 videos uploaded during 15 days from YouKu [14] for 4 pre-defined topics.

2.3 Benchmarks for social video analysis

Different from the benchmarks for traditional video content analysis, benchmarks for social video analysis encourage researchers mining videos based on the rich social features, such as the labels and the user relationships. As a result, the dataset should crawl the original videos as well as their social networks.

The widely used general video benchmark is the video collection of TRECVID [18]. Before 2010, all the videos are professional resources including news magazines, science news, news reports, documentaries, educational programming, and archival videos in MPEG-1. Meanwhile, the dataset provides the shot segmentation results, automatic speech recognition results, as well as human annotations for multiple video analysis tasks such as feature extraction, search and copy detection. From 2010, TRECVID database has been changed from static resources to the Web video resources, containing approximately 160,00 Web videos (100 GB, 400 hours) from Internet Archive website.[2] The selection criterion is general enough to include all the videos whose duration is shorter than 4 minutes. Unfortunately, since this website is far from popular comparing to some well known social networks such as Flickr, YouTube and Facebook, its videos are less participated by the Web users and are therefore "*socialless*". As a result, there are limited contextual and social network information.

3 MCG-WEBV 2.0

MCG-WEBV 2.0 has a considerably large scale with 248,887 unique videos (12,588 hours) and 123,063 users. Moreover, it has extracted sufficient content information, including the keyframes, the pre-extracted visual and textual features, as well as the sufficient context information of the video social network created by the video contributors and reviewers. Following we'll introduce the details of this dataset.

3.1 Data crawling mechanism

The crawling of MCG-WEBV 2.0 includes three steps:

- The first crawling starts from a set of seed videos including the "Most Viewed" videos of "This Month" for 15 YouTube categories. It has an 11-month duration from Dec. 2008 to Nov. 2009, except Aug. 2009 due to Internet blocks. We have downloaded 14,473 videos in total, which is named as *Core Dataset*.
- The second crawling further expands the *Core Dataset* by downloading their 1-depth "Related Videos". It gets 234,414 distinct videos in total, named as *Expand Dataset*. By combining both datasets, our MCG-WEBV 2.0 contains totally 248,887 videos.

[2] http://www.archive.org/index.php

Table 2 The detail information downloaded and extracted by MCG-WEBV 2.0

Video Items	User Items	Video Features
ID	UserName	Shots and Keyframes
Category	Age	Textual Vector Space Feature
Video Length	Gender	36-D Audio Feature Vector
Owner	Location	2-D Video Slice
Date uploaded	Hobbies	4-D shot-level Face Feature
Number of views	Occupation	166-D Color Histogram
Number of comments	Movies	225-D Color Moments
Rating	Music	166-D Color Auto-Correlogram
Video Title	Favorite count	320-D Edge Histogram
Video Tag	Subscriber count	108-D Haar Wavelet Texture
Video Description	View count	96-D Co-occurrence Texture
Related video list	Watch count	1000-D Visual Keywords
Commented user list	Favorite video list	
	Uploaded video list	
	Subscribed user list	
	Used tag list	

– In the third crawling, we index the owners (in total 123,063 YouTube users) of all the videos in database to download their social network structure, containing the YouTube user IDs, their profile information, as well as their interactions available such as the favorite video list, uploaded video list, subscribed user list and used tag list from the YouTube API.[3]

Based on the above crawling mechanism, we downloaded the original videos, and further extracted the shots, keyframes, and 11 features as listed in Table 2, which provides the information for general video content analysis, covering textual, audio and visual content. All these data have been released online except the original video for the copyright issue. In May 2009, we released Version 1.0 [4] including the data of the first three months from Dec. 2008 to Feb. 2009, with in total 80,031 videos and their corresponding features, which has been widely used in both academic and industry researches (**downloaded for over 200 times**). In 2012, we released Version 2.0, not only expanding the video data to 11 months from Dec. 2008 to Nov. 2009, but also adding the social network information of the whole dataset. The details of database can be found at http://mcg.ict.ac.cn/mcg-webv.htm.

The crawling mechanism is inspired by the quantitative validations reported by the YouTube authority, as well as the research conclusion of Gill et al. in [9], both of which have shown that the popular videos generally have higher ratings, which indicate that they have more interesting contents and higher video qualities. In the meanwhile, the popular videos are widely spread among the web users, and thus are active enough to represent the characteristics of the real-world network [6]. Subsequently, the corresponding research exploring on these videos is more useful in practice. Finally, our motivation

[3] http://code.google.com/intl/en/apis/youtube/overview.html.

also partially comes from the empirical evidence of the Pareto principle, which states that in the website referencing behaviors, 20 % (10 %) of the pages on a web server accounts for 80 % (or 90 %) of the requests [9]. As a result, we would like to expect that the most viewed videos also play a dominant role in the entire YouTube video collection.

3.2 Insights into social network for MCG-WEBV 2.0

The social network of MCG-WEBV 2.0 is constructed based on the owners of popular videos, referred as "video social network" in this paper. Besides the basic statistics which have been widely studied in the related work of social network analysis [7, 9], we also investigated two specific characteristics for popular video collection:

First, we crawled the uploaded video lists of 105,460 owners, and gave the detailed statistics in Fig. 1. Although there are some over-productive contributors (Max=50,004), the majority is less productive with mean value 91.76 and median value 18. Moreover, the scores in Fig. 1b show that only 1.17 % owners uploaded more than 1,000 videos, while 36.48 % owners uploaded less than 10 videos. All the above evidences imply that most of the contributors for popular videos are common users with limited uploading: The uploading frequency does not directly result in the user popularity. In the meanwhile, we further explore the small part of over-productive users, and find that most of them are official accounts such as the online TV channels and organizations, as shown in Fig. 1c. These productive accounts do not correspond to specific person. As a result, they don't meet the law of general web users and should be removed from the related researches on the the user behavior analysis.

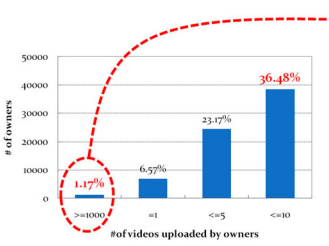
Second, by analyzing the users who give comments to popular videos, we disclose that there are only 2.6 % overlaps between the collections of video contributor and video reviewer. By excluding the self-comment, this overlap decreases to 1.5 %.

4 Sampling criteria from a social network perspective

As introduced in Section 1, since completely unbiased sampling is infeasible, being aware of the potential sample bias can help researchers to construct or select a more suitable

Statistics for 105,460 owners	# of uploaded videos
Mean	91.76
Median	**18**
Max	50004
Min	1
Std	597.90

(a)



(b)

User ID	# of uploaded videos
yellowbook	50,004
expertvillage	50,004
HSN	50,000
ChinaTimes	50,000
AssociatedPress	42,416
imagiers	41,253
CBS	40,174
europapress	27,782
FashionTV	26,545
CNN	25,015

(c)

Fig. 1 The statistics of the uploading action for owners. **a** is the global statistics. **b** is the amounts of owners who uploaded videos in YouTube >=1,000, =1, <=5 and <=10 respectively. **c** is the top 10 most productive owners. It displays that these over-productive users are not the specific person, but the official accounts of the online TV channels or organizations

benchmark for their specific tasks. In the following, we summarize the existing sampling techniques and their corresponding bias, and propose to measure the correlation between potential bias and tasks from a video social network perspective.

4.1 Sampling techniques and biases

As shown in Fig. 2, the entire web video collection can be regarded as a scale-free heterogeneous network with user U and video V, collected by the user interactions I over videos. We can sample a subset of videos either directly based on a video profile or indirectly based on a user profile (by further collecting videos corresponding to their interactions). We refer the above two kinds of sampling techniques as "*from videos*" and "*from users*" in Table 1, respectively. To the best of our knowledge, most existing video benchmarks sample videos based on the most popular video lists ranked by video view counts [4, 6], most responsive video lists ranked by the number of video responses [1], video search results by video tags [11, 14, 23], video categories [5, 12] and the video uploading times [2], which can be categorized as "*from videos*". On the other hand, work in [9, 10, 24] collected the videos based on the request logs of users from a specific region, which can be categorized as "*from users*".

The above sampling techniques mainly contain two kinds of sampling biases:

Application-Bias The sampling techniques are explicitly designed to collect the videos with a specific attribute, and the resulting databases can only fit to the corresponding application.

Popularity-Bias The sampling techniques have implicitly favors towards popular video contents.

Obviously, all the application-driven benchmarks contain application-bias, since videos that are sampled based on the attributes of the corresponding application, so the sampling criteria are application related. Moreover, besides the potential sampling bias, they only provide limited content for the specific task. For instance, in [11], Jiang et al. only collected videos with 20 pre-defined semantics for the video semantic detection task. Similarly, in

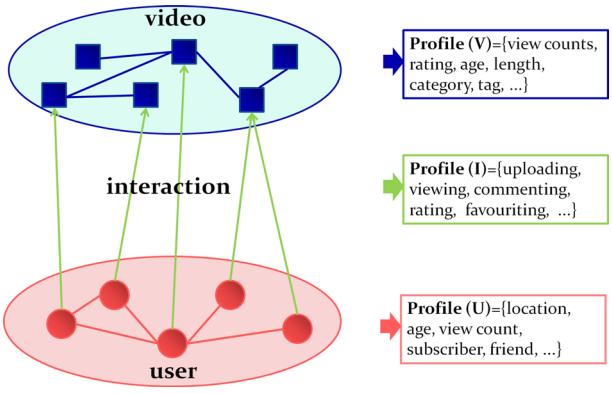


Fig. 2 The heterogenous network of web videos and users. We can sample a subset of videos from this network either by video nodes or by user ones

[14], Liu et al. only selected the videos belonging to four pre-defined topics to evaluate their web video topic discovery algorithm.

On the other hand, popularity-bias is very common in the benchmarks sampled based on the search engines and recommendation systems, which have been verified to be biased towards popular content [2]. Figure 3 shows a direct observation of popularity-bias, which displays the comparison of the video popularity distribution on different datasets. While previous works [5] assume that the unbiased video popularity might follow a Zipf distribution, so most of related works [5, 6, 9] utilize standard Zipf to fit the curves of video popularity and analysis their differences. Except Fig. 3e, all the curves have a truncated tail in different degrees. It is due to their sampling techniques biased toward popular content, and leads to less unpopular videos than the expected Zipf distribution. Figure 3e shows that, sampling videos based on the geography distribution of users has little popularity-bias, and the corresponding curve can perfectly fit Zipf distribution. The datasets in Figs. 3a and 3c are sampled from most viewed videos, so they have similar distribution curves with heavily truncated tails. The datasets in Figs. 3b and 3d are sampled separately from a specific geographic region or a specific category, whose popularity-bias is less than Figs. 3a and 3c, but more than Fig. 3e. So they have the middle heavy tails.

Figure 3 shows that different sampling strategies will lead to sampling bias at different levels. As a result, how to quantify the potential bias is an important issue, which is solved in the next section.

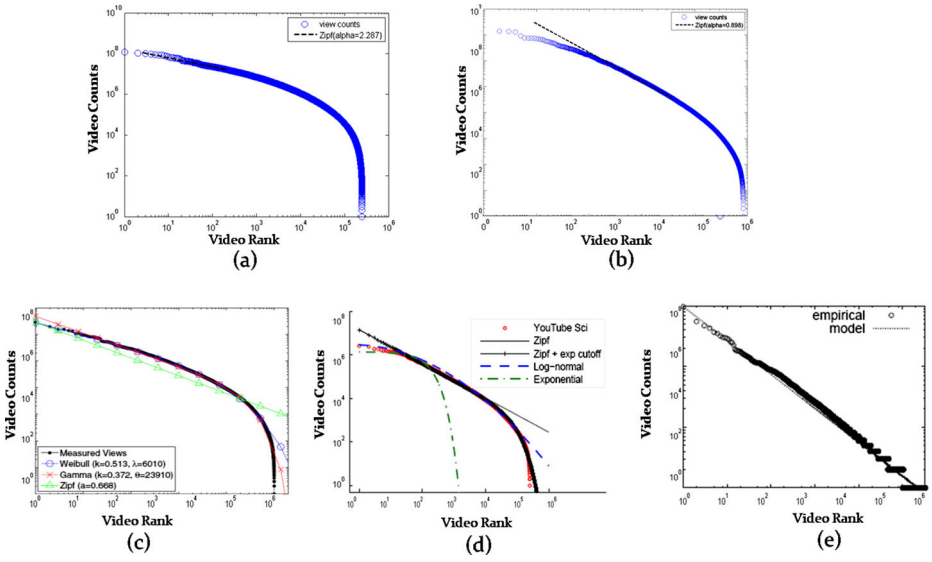


Fig. 3 the comparison of video popularity distribution on 5 datasets with different sampling criteria. **a** is MCG-WEBV 2.0 with 248,887 videos, sampled based on video view count. **b** is the MCG-WEBV-GEO with 85,117 YouTube videos uploaded From NewYork, sampled based on video location. **c** is the dataset in reference [6] with the statistics of 3,269,030 YouTube videos, sampled based on video view count and comment count. **d** is the dataset in reference [5] with the statistics of 252,255 YouTube videos, sampled from YouTube Category "Science". **e** is the dataset in reference [9] with the statistics of 323,677 Youtube videos, sampled based on the user request uploaded from a specific campus

4.2 Measuring sampling bias

Most existing video datasets are designed for different tasks. As a result, the experimental comparison between all datasets as implemented in [21] is unfeasible. Alternatively, we propose to measure the sampling bias by computing the correlation between the statistics of potential bias and the sampling criteria in the video social network.

Taking popularity-bias as an example, the main idea is that if the sampling criterion is independent or linearly correlated to the video popularity, the resulting dataset has no popularity-bias. Subsequently, we combine the Kullback-Leibler divergence (KL) and Pearson Correlation Coefficient (CC) to measure the correlation between sampling criteria S and video popularity C. $KL(C\|S)$ is a non-symmetric measure of the difference between two probability distributions C and S. It represents the divergence when modeling C with S. Larger score of $KL(C\|S)$, corresponds to less correlation between C with S, and vice versa. On the other hand, $CC(C, S)$ represents the degree of linear correlation between C and S, and the value ranges from -1 to 1. $CC(C, S)$ is closer to 1 or -1, means that C and S are more linear correlated, while the linear correlation implies that the sampling criterion has no bias towards popular videos than unpopular ones. On the contrary, the score of $CC(C, S)$ closed to 0 doesn't mean that C and S are independent, and we should consider the score of KL together to make a better judgement. The details of the measure algorithm is shown as Algorithm 1.

5 Case study of sampling bias on MCG-WEBV 2.0

In MCG-WEBV 2.0, we calculate KL and CC between video popularity and several widely used sampling criteria, including three video attributes i.e. age, tag number, comment number of videos, and three user attributes i.e. the uploaded video number, subscriber number and view number of owners. As shown in Fig. 4, the statistics of KL and CC are consistent with the empirical distribution curves, which verifies their reliability. Based on these statistics, we make conclusion of the following principles:

Algorithm 1 Measuring the popularity-bias of a dataset.

Input : view counts $C = \{c_1, c_2, \cdots, c_n\}$ and the statistics of sampling criteria $S = \{s_1, s_2, \cdots, s_n\}$ for all the videos in dataset, n is the size of dataset.
Output : popularity-bias evaluation

1. Normalize C and S to probability distribution $P(C)$ and $P(S)$ summed into 1.
2. $KL\left(P(C)\middle\|P(S)\right) = \sum_{i=1}^{n} P(c_i)\frac{P(c_i)}{P(s_i)}$
3. $CC(P(C), P(S)) = \dfrac{\sum P(C)P(S) - \frac{\sum P(C)\sum P(S)}{n}}{\sqrt{(\sum P(C)^2 - \frac{(\sum P(C))^2}{n})(\sum P(S)^2 - \frac{(\sum P(S))^2}{n})}}$
4. For each sampling criteria:
   (a) larger score of KL, less popularity-bias;
   (b) small score of KL and large score of CC, less popularity-bias, but linearly correlation between video popularity and sampling criteria;
   (c) smaller score of KL and CC, more popularity-bias;

Figure 4e–g show that the video popularity has little correlation with the statistics of the user social network, where the user attributes (number of subscriber, view count and number

of uploaded videos) all have large *KL* and small *CC* of 0.04, 0.04 and −0.01 respectively. However, its converse doesn't meet this rule as shown in [1], where the user popularity is shown to be relevant to the rating number and popularity of their uploaded videos with *CC* of 0.44 and 0.27 respectively. Different from the pure user social network such as Facebook, YouTube is a video social network, where the relationships between users are built mainly based on their interactions over videos. Therefore, the dependency between videos and users are asymmetric in video social network, where the video popularity is relatively independent to users. However, the user popularity is dependent to videos. In conclusion, we can build relatively popularity-unbiased video benchmarks by sampling based on users in a video social network.

Whether a video with more tags has larger popularity? On average, each popular video in MCG-WEBV 2.0 has 14.1 tags, while a general YouTube video in [17] has 9 tags, and a consumer video in [11] even has only 2.5 tags. This difference implies that the popular videos have more tags then the unpopular ones. It is reasonable since the videos with more tags have more chances to be searched in a query-by-keyword scenario. However, Fig. 4d shows a different observation, where the video popularity does not regularly increase with tag number as expected, and even displaying negative correlation with a -0.0309 correlation coefficient. By further investigating the videos with long tag list (e.g., more than 100), we find that most videos with too many tags are irresponsibly labeled with a lot of meaningless words. In general, the bad labels are associated to "bad" video contents, which results in the decrease of video popularities.

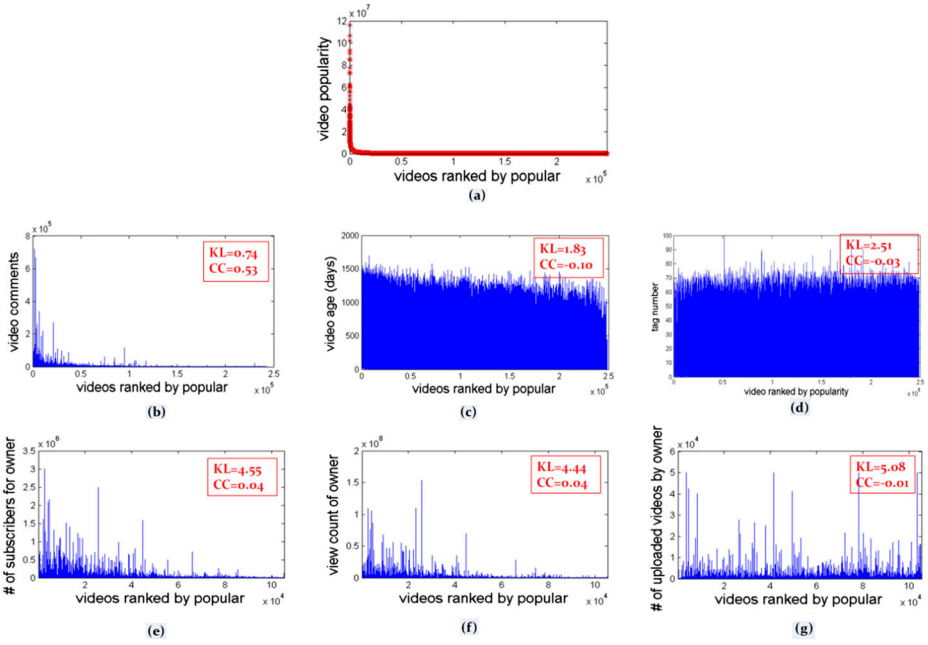


Fig. 4 The relationship between video popularity and different characteristics of videos and their contributors. All the statistics are generated based on the popularity rank. Besides the empirical data distribution displayed as figures, we also give the scores of Kullback-Leibler divergence(KL) and Pearson Correlation Coefficient(CC) between this statistic and popularity

The correlation between video popularity and video age has been widely studied in [2, 5] to predict the future popularity of a target video. In a global view, video popularity exhibits a weak correlation with the video age (with *CC* 0.18). By further exploring this correlation in different age periods including younger than 1 week, 1 week to 1 month, 1 month to 3 months, 3 months to 1 year, and older than 1 year, we get different *CC* scores with 0.118, 0.028, 0.00, 0.046 and 0.121 respectively. It shows that the popularities of the youngest (e.g., newer than 1 week) and oldest (e.g., older than 1 year) videos are weakly correlated to their video ages, but they are almost independent in the rest of the major periods. According to these observations, we conclude that sampling the videos based on their uploading time does not contain the popularity-bias, if they are neither too young (e.g., younger than 1 weeks) nor too old (e.g., older than 1 year).

According to the above analysis, we conclude that MCG-WEBV 2.0 is application-unbiased, which contains sufficient content and context data to evaluate several kinds of applications such as video categorization, video retrieval and video topic discovery [3, 15]. Our recent work in [15] is an application case on MCG-WEBV 2.0. It combines the video and user network in MCG-WEBV 2.0 into a heterogeneous video social network, and mines the heterogeneous community structures in it. These communities are then used to rerank the video ranking list to improve the retrieval performance. It is worth to mention that, MCG-WEBV 2.0 is popularity-biased, so it is not appropriate to evaluate the popularity-related tasks such as the video response and comment analysis. But it can fairly evaluate the popularity-unrelated tasks such as the video age analysis and user behavior analysis.

6 Conclusion

To cope with the ever growing web videos contributed by community users, there is an emerging demand for a related benchmark specialized for both social ngetwork analysis and video content understanding. In this paper, we have introduced a so-called MCG-WEBV 2.0 web video benchmark,sampled from YouTube with rich content and context information, and designed for social network based video content analysis. Our key novelty is to unveil the sampling bias in collecting videos from the web community, which is unexploited in the previous benchmarks. We show that a proper sampling design from a social network perspective can offer a more fair real-world evaluation platform for the state-of-the-art social network and content analysis techniques. Especially, we demonstrate that while a fully unbiased benchmark is not feasible, it is possible to isolate the sampling bias from the corresponding tasks, which offers practical suggestion for the future benchmark designs.

Acknowledgments This work is supported by the National High Technology Research and Development Program of China (2014AA015202), National Nature Science Foundation of China (61172153,61100087), National Key Technology Research and Development Program of China (2012BAH39B02).the Beijing New Star Project on Science & Technology (2007B071).

References

1. Benevenuto F, Rodrigues T, Almeida V, Almeida J, Ross K (2009) Video interactions in online video social networks, ACM Transactions on Multimedia Computing. Commun Appl 5:1–25
2. Borghol Y, Mitra S, Ardon S, Carlsson N, Eager D, Mahanti A (2011) Characterizing and modeling popularity of user-generated videos. In: IFIP Performance

3. Cao J, Ngo CW, Zhang YD, Li JT Tracking web video topics:discovery, visualization and monitoring, IEEE Transaction on Circuits and Systems for Video Technology
4. Cao J, Zhang YD, Song YC, Chen ZN, Zhang X, Li JT (2009) Mcg-webv: A benchmark dataset for web video analysis. In: Technical Report, ICT-MCG-09-001
5. Cha M, Kwak H, Rodriguez P, Ahn YY, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: ACM SIGCOMM Conference on Internet Measurement
6. Cheng X, Dale C, Liu JQ (2008) Statistics and social network of youtube videos. In: International Workshop on Quality of Service
7. Ding Y, Du Y, Hu Y, Liu Z, Wang L, Ross K, Ghose A (2011) Broadcast yourself: understanding youtube uploaders. In: ACM SIGCOMM conference on Internet measurement conference
8. Gao Y, Wang M, Zha ZhJ, Shen Jl, Li XL, Wu XD (2013) Visual-textual joint relevance learning for tag-based social image search. In: IEEE Trans Image Process 22(1)
9. Gill P, Arlitt M, Li Z, Mahanti A (2007) Youtube traffic characterization: a view from the edge. In: ACM SIGCOMM conference on Internet Measurement
10. Huang C, Li J, Ross KW (2007) Can internet video-on-demand be profitable? SIGCOMM Comput Commun Rev 37:133–144
11. Jiang YG, Ye GN, Chang SF, Ellis D, Loui AC (2011) Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In: ACM International Conference on Multimedia Retrieval
12. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE International Conference on Computer Vision and Pattern Recognition
13. Li HJ, Liu B, Yi L, Guan Y, X Luo Zh (2014) On the Tag Localization of Web Video. In: Multimedia Systems
14. Liu L, Sun L, Rui Y, Shi Y, Yang S (2008) Web video topic discovery and tracking via bipartite graph reinforcement model. In: International Conference on World Wide Web
15. Pang L, Cao J, Zhang YD, Lin SX (2011) Leveraging collective wisdom for web video retrieval through heterogeneous community discovery. In: ACM International Conference on Multimedia
16. Song YC, Zhang YD, Cao J, Xia T, Liu W, Li JT Web video geolocation by geotagged social resources, IEEE Transaction on Multimedia
17. Sharma AS, Elidrisi M Classification of multimedia content using tags and focal points. In: Project Report of University Of Minnesota., http://www-users.cs.umn.edu/ankur/FinalReport_PR-1.pdf, 2009
18. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: ACM International Workshop on Multimedia Information Retrieval
19. Song YCh, Zhang YD, Cao J, Tang JH, Gao XY, Li JT (2014) A unified geolocation framework for web videos. ACM Trans Intell Syst Technol (TIST) 5(3):49
20. Tang JH, Yan SCh, Hong RCh, Qi GJ, Chua TS (2009) Inferring Semantic Concepts from Community-contributed Images and Noisy Tags. In: ACM International Conference on Multimedia
21. Torralba A, Efros AA (2011) Unbiased look at dataset bias. In: IEEE International Conference on Computer Vision and Pattern Recognition
22. UCF 50 Human Action Dataset., http://server.cs.ucf.edu/vision/data/UCF50.rar/, 2010
23. Wu X, Hauptmann AG, Ngo CW (2007) Practical elimination of near-duplicates from web video search. In: ACM International Conference on Multimedia
24. Yu H, Zheng D, Zhao BY, Zheng W (2006) Understanding user behavior in large-scale video-on-demand systems. SIGOPS Oper Syst Rev 40:333–344
25. Zha ZhJ, Wang M, Zheng YT, Yang Y, Hong RCh, Chua TS (2012) Interactive video indexing with statistical active learning. IEEE Trans Multimed 14(1):17–27
26. Zha ZhJ, Zhang HW et al (2013) Detecting group activities with multi-camera context. IEEE Trans Circ Syst Video Technol 23(5):856–869
27. Zha ZhJ, Yang LJ, Mei T, Wang M, Wang ZF, Chua TS, Hua XSh (2010) Visual query suggestion: Towards Capturing User Intent in Internet Image Search. In: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMMCAP) 6(3)

Juan Cao born in 1980. Associate professor of ICT CAS. Her research interests focus on large scale social media analysis.

Yongdong Zhang born in 1973, Ph.D. Professor of ICT CAS. His major field includes image processing and video processing.

Rongrong Ji born in 1979, Ph.D. Professor of Xiamen University. His major field includes Computer Vision, Multimedia, and Machine Learning.